Clustering conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding

Rajeswari Appadurai,[1] Jaya Krishna,[2] Massimiliano Bonomi,[3] Paul Robustelli,[2] and Anand Srivastava[1,a)]

[1)] *Molecular Biophysics Unit, Indian Institute of Science Bangalore, C. V. Raman Road, Bangalore, Karnataka 560012, India*

[2)] *Dartmouth College, Department of Chemistry, Hanover, NH, 03755, USA*

[3)] *Structural Bioinformatics Unit, Department of Structural Biology and Chemistry. CNRS UMR 3528, C3BI, CNRS USR 3756, Institut Pasteur, Paris, France*

[a)] Electronic mail: anand@iisc.ac.in

Abstract: Intrinsically disordered proteins (IDPs) populate a range of conformations that are best described by a heterogeneous ensemble. Grouping an IDP ensemble into "structurally similar" clusters for visualization, interpretation, and analysis purposes is a much-desired but formidable task as the conformational space of IDPs is inherently high-dimensional and reduction techniques often result in ambiguous classifications. Here, we employ the t-distributed stochastic neighbor embedding (t-SNE) technique to generate homogeneous clusters of IDP conformations from the full heterogeneous ensemble. We illustrate the utility of t-SNE by clustering conformations of two disordered proteins, Aβ42, and a c-terminal fragment of α-synuclein, in their APO states and when bound to small molecule ligands. Our results shed light on ordered sub-states within disordered ensembles and provide structural and mechanistic insights into binding modes that confer specificity and affinity in IDP ligand binding. t-SNE projections preserve the local neighborhood information and provide interpretable visualizations of the conformational heterogeneity within each ensemble and enable the quantification of cluster populations and their relative shifts upon ligand binding. Our approach provides a new framework for detailed investigations of the thermodynamics and kinetics of IDP ligand binding and will aid rational drug design for IDPs.

Significance: Grouping heterogeneous conformations of IDPs into "structurally similar" clusters facilitates a clearer understanding of the properties of IDP conformational ensembles and provides insights into "structural ensemble: function" relationships. In this work, we provide a unique approach for clustering IDP ensembles efficiently using a non-linear dimensionality reduction method, t-distributed stochastic neighbor embedding (t-SNE), to create clusters with structurally similar IDP conformations. We show how this can be used for meaningful biophysical analyses such as understanding the binding mechanisms of IDPs such as α-synuclein and Amyloid β-42 with small drug molecules.

Keywords t-distributed stochastic neighbor (t-SNE), intrinsically disordered protein (IDP), conformations clustering, drug design, machine learning

I. Introduction

In general, knowledge of the 3-dimensional structure of a protein is the first step toward a molecular-level mechanistic understanding of its biological function. This knowledge is also central to activities such as the rational design of drugs, inhibitors, and vaccines and in the broad area of protein engineering and biomolecular recognition[1-7]. With the advances made in structure determination techniques[8-16] and recent transformative leaps made in computationally predicting the structure from sequence[17-19], the science of structural biology is going through paradigmatic changes where the knowledge of structure is not the biggest bottleneck anymore[20]. However, outside the realm of these structured proteins exist a "dark" proteome of intrinsically disordered proteins (IDPs) that constitute more than 40% of all known proteins and play important roles in cellular physiology and diseases[21-27]. An IDP can populate a heterogeneous ensemble of conformations and is functional without taking a unique structure. In essence, IDPs are expanding the classical hypothesis of sequence-structure-function to the sequence-disordered ensemble-function(s) paradigm. Though solution-based experiments like NMR, FRET, and SAXS do provide structural information for IDPs, they generally report time and ensemble-averaged properties of IDP conformations[28-30]. In the absence of computational models, solution experiments are challenging to interpret in terms of individual atomic resolution structures that constitute IDP ensembles. In other words, IDPs are not directly amenable to conventional high-resolution structure determination, structure-based functional correlation, protein engineering, and drug-designing strategies that hinge upon the knowledge of a reference 3-dimensional structure.

Computational tools, particularly those that incorporate the available experimental information, can be effectively used to generate high-resolution ensemble structures of IDPs. Of late, several broad classes of different approaches have been developed for this purpose. Methods based on pre-existing random coil library and simple volume exclusions (examples: Flexible Meccano[31], TraDES[32], BEGR[33]) are often used to create an initial exhaustive pool of conformations, which are further processed to produce refined ensembles upon combining with experimental constraints[30,34-39]. These methods, though purely statistical in nature, provide a computationally efficient approach to calculating IDP conformational ensembles that are consistent with experimental data. The second set of approaches utilizes

Physics-based molecular simulations either in a coarse-grained representation (examples: SIRAH[40], ABSINTH[41], AWSEM-IDP[42], SOP-IDP[43], HPS[44] and others) or with an all-atom resolution,[45-48] to generate initial Boltzmann-weighted conformational ensembles that can be further refined with experimental restraints using various reweighing approaches[49-51]. Recently developed molecular mechanics force fields for IDPs[45-48] used in combination with parallel tempering based enhanced sampling approaches such as Replica exchange solute tempering (REST)[52-55] and hybrid tempering (REHT)[56] has also shown promise in producing atomic-resolution accurate IDP ensembles consistent with experimental solution data without any added bias in the simulations.

While significant advances have been made in generating high-resolution IDP conformational ensembles that are consistent with experimental data, the subsequent interpretation of these ensembles to address key biological questions related to the interactions of IDPs remains extremely challenging. IDP conformational ensembles are inherently extremely high-dimensional. That is, the phase space of IDPs consists of several thousands of features, which may vary relatively independently, making it extremely challenging to uncover correlations in conformational features among conformations contained in IDP ensembles. This often makes sequence-ensemble-function relationships of IDPs very difficult to understand, even when aided by relatively accurate IDP conformational ensembles. If one could efficiently identify representative conformational sub-states in IDP ensembles, and quantify their relative populations in different molecular and cellular contexts, it would become significantly easier to identify conformational features of IDPs that may be associated with specific functional roles or disease states[57-59]. Therefore, parsing the heterogeneous ensemble data into representative conformational states can be as critical as the generation of the ensemble itself as it allows one to leverage conventional structural-biology analysis tools for IDPs.

The process of dividing large abstract data set into a number of subsets (or groups) based on certain common relations such that the data points within a group are more similar to each other and the points belonging to different groups are dissimilar is called clustering. Due to its ability to provide better visualization and statistical insights, clustering is ubiquitous in the analyses of big-data biological systems with wide-ranging applications such as profiling gene expression pattern[60,61], de novo structure prediction of proteins[62,63], the quantitative structure-activity relationship of chemical entities[64], docking and binding geometry scoring[65],

and also in analyses of protein ensemble from molecular dynamics (MD) trajectory[66]. In the latter case, the similarity or dissimilarity between a pair of conformations is often defined by classifiers or collective variables (CVs) such as root mean squared displacements (RMSD), differences in native contacts, dihedral angle distribution, the radius of gyration, and the end-to-end distance. The choice of CV is generally empirical in nature but in general, CVs chosen based on domain knowledge do yield reliable clustering in structured protein ensembles. However, they do not sufficiently describe the highly heterogeneous IDP ensemble. This is primarily due to the mismatch one sees in conformations for the same CV values since low-dimension CV projections are unable to capture the finer details in the coordinate space that separates one conformation from another one. This has a severe impact on clustering since often different conformations of IDP have similar projected collective variables and are kept together in one group. To illustrate this, we present a set of conformations from a simulated IDP ensemble with the same value of R_g (Fig. S1 in Supplementary Material (SM)). It is evident from this illustration how this could lead to ambiguous classification. The problem can be circumvented to some extent by considering multi-dimensional projections. But this often leads to having too many sparsely-populated clusters, which attenuates the whole exercise of effective clustering.

Theoretically well-grounded dimensionality reduction (DR) techniques are now commonly being used in protein conformation analysis to extract the latent low dimensional features and the quantum of information lost during the projection depends heavily on the kind of data set under consideration[67–72]. For example, a highly heterogeneous data set that lies on a high-dimensional manifold as in the case of IDPs is best handled with the non-linear dimension reduction (NLDR) techniques, which generally attempt to keep the nearest neighbors close together. While methods such as ISOMAP and Local Linear Embedding are best suited to unroll or unfold a single continuous manifold, a more recent method such as t-Distributed Stochastic Neighbor Embedding (t-SNE) is more appealing as it helps to disentangle multiple manifolds in the high-dimensional data concurrently by focusing on the local structure of the data and tend to extract clustered local groups of samples. Consequently, t-SNE tends to perform better in separating clusters and avoiding crowding. We will show that t-SNE is particularly well-suited for clustering seemingly disparate IDPs conformations into homogeneous subgroups since it is designed to conserve the local neighborhood when

reducing the dimension, which ensures similar datasets remain similar and dissimilar points remain dissimilar in the low dimensional space as that of the high dimensional space[73]. Due to its ability to provide a very informative visualization of heterogeneity in the data, t-SNE is being increasingly employed in several applications such as clustering data from single cell transcriptomics[74–77], mass spectrometry imaging[78], and mass cytometry[79,80]. Lately, t-SNE has also been used for depicting the MD trajectories of folded proteins[81–87] and for interpretation of mass-spectrometry based experimental data on IDPs by juxtaposing with classical GROMOS-based conformation clusters from the corresponding molecular simulation trajectories of the IDP under consideration[88].

In this paper, we demonstrate the effectiveness of t-SNE (in combination with K-means clustering) for identifying and visualizing representative conformational substates in IDP ensembles. We investigate the small molecule binding properties of Amyloid β-42 (Aβ42) and α-synuclein (αS), proteins involved in the neurodegenerative proteinopathies like Alzheimer's and Parkinson's diseases, respectively. Therapeutic interventions by sequestering the monomeric state of these IDPs have recently been explored using state-of-the-art biophysical experiments and long timescale molecular simulations[89,90]. A set of repurposed small molecules such as the c-Myc inhibitor-G5 (benzofurazan N-([1,1-biphenyl]-2-yl)-7-nitrobenzo[c][1,2,5]oxadiazol-4-amine (10074-G5)) and a Rho kinase inhibitor - Fasudil (along with the high-affinity Fasudil variant Ligand-47) have been identified as promising agents against the monomers of Aβ42 and αS, respectively. Since the monomeric states of these IDPs are extremely heterogeneous, it is not fully understood how the different conformations form viable complexes with these small molecules and what is the molecular origin of their affinity and specificity. This is mainly owing to the inefficient clustering of the IDP structures using the classical clustering tools. Here we revisit the molecular trajectories of Aβ42 (a total of 56 μsecs) and αS (total of 573 μsecs) using t-SNE (in combination with K-Means clustering). This exercise has drastically improved our knowledge of the binding mechanism of small molecules to such large IDPs and also provides us with strategies for designing specific inhibitors with high-affinity binding. Besides, it also helps in understanding the conformational landscape of APO and ligand-bound IDPs, which are otherwise hard to obtain. We believe that the method presented here is general in nature and can be used to cluster and visualize IDP ensembles across systems and assist in detailed structural,

thermodynamics, and kinetics analyses of IDP conformations in APO and bound states.

II. Results and Discussion

We aim to cluster the heterogeneous mixture of disordered protein conformations into a subset of unique and homogeneous conformations. To do this, as a first step, we employ t-SNE that projects the large dimensional data in lower dimensions. We then apply K-means clustering on the projections to identify the clusters in the reduced space. Before we illustrate the power of this algorithm as a faithful clustering tool for realistic IDP ensembles, we use a simple alanine-dipeptide (ADP) toy model in SM to provide physical intuition into how t-SNE works. Please see Fig. S2 and the subsection titled *"Physical intuition into t-SNE-based clustering algorithm using alanine dipeptide"* in SM. We use this toy model system to also make the reader familiar with a critical hyper-parameter called perplexity used in the t-SNE algorithm, and prescribe a strategy to choose its right value for effective clustering. Going forward, we apply the method for the analyses of IDP ensembles of complex systems such as Aβ42 and αS, each in the presence and absence of small-molecule inhibitors. We have listed all the systems under consideration in Table I below. We represent the conformations within Aβ42 and αS ensembles by the inter-residue Lennard-Jones contact energies and the Cartesian coordinates of heavy atoms, respectively. These measures are kept consistent with the original work from where trajectory data is taken for Aβ42[89] and αS[91] to enable faithful comparisons. t-SNE was performed based on the distance/dissimilarity calculated between all pairs of conformations as RMSD of the above data.

TABLE I: Information on systems and trajectories used in this study

S.No	Description	Simulation scale	#snapshots	Reference
1	Alanine-dipeptide	10 (ns)	2500	[56]
2	Aβ42 APO	27.8 (μs)	35,000	[89]
3	Aβ42-G5 bound	28.2 (μs)	35,000	[89]
4	Aβ42 apo + G5 bound	-	70,000	[89]
5	αS C-terminus in apo	100 (μs)	55,545	[90]
6	αS C-ter + Fasudil	200 (μs)	55,045	[90]
7	αS C-ter + Ligand-47	200 (μs)	55,545	[90]
8	αS C-ter + Fasudil + Lig47	-	166,135	[90]
9	αS full-length APO	73 (μs)	36,562	[90]

A. Prescription for choosing optimal parameters for t-SNE clustering of IDPs

The results of t-SNE depend largely on the choice of perplexity. Since the objective criterion here is to maximize clustering, we adopt the well-known Silhouette score,[92] commonly used for optimizing the number of clusters (K) in K-means clustering, for tuning the perplexity values as well. As shown through the formulation in the method section below, the Silhouette score computes the average of every point's distance to its own cluster (cohesiveness) than to the other clusters (separateness) and is defined such that its value lies in the range of -1 to 1. A score of 1 is most desirable indicating perfectly separated clusters with clearly distinguishable features. A positive value generally indicates acceptable clustering while negative values are unacceptable for distinguishable clustering. The cohesiveness and separateness of clusters are generally measured based on Euclidean distance. Since the clusters here are identified on a reduced t-SNE space, computing the score on this space (S_{ld}) alone may be misleading. This is particularly true when using sub-optimal parameters that often clump the points randomly during the dimensional reduction step by t-SNE. Therefore, it is important to measure the goodness of clustering with respect to the original distance in the high dimensional space (S_{hd}), in addition to that in the low dimensional space. The integrated score ($S_{ld} * S_{hd}$), therefore, adds value to the estimated clustering efficiency in terms of reliability.

B. t-SNE for clustering Aβ42 conformational ensembles

1. t-SNE identifies the clustering pattern intrinsic to the Aβ42 ensemble

We apply our algorithm on APO and G5-bound Aβ42 all-atom MD simulations trajectories obtained from Vendruscolo group[89]. We have used an identical set of representative frames for clustering as in the original work (35000 frames from each ensemble) where each system was run for 27.8 (μs). Furthermore, to be consistent, we represent the conformations similarly by inter-residue Lennard-Jones contact energies. We used the distance between all pairs of conformations from the RMSD of the contact energies and feed that into our t-SNE pipeline. In the case of Aβ42 (APO and G5-bound), the calculated Silhouette score for a range of K and perplexities indicates a positive value with respect to both the distances at

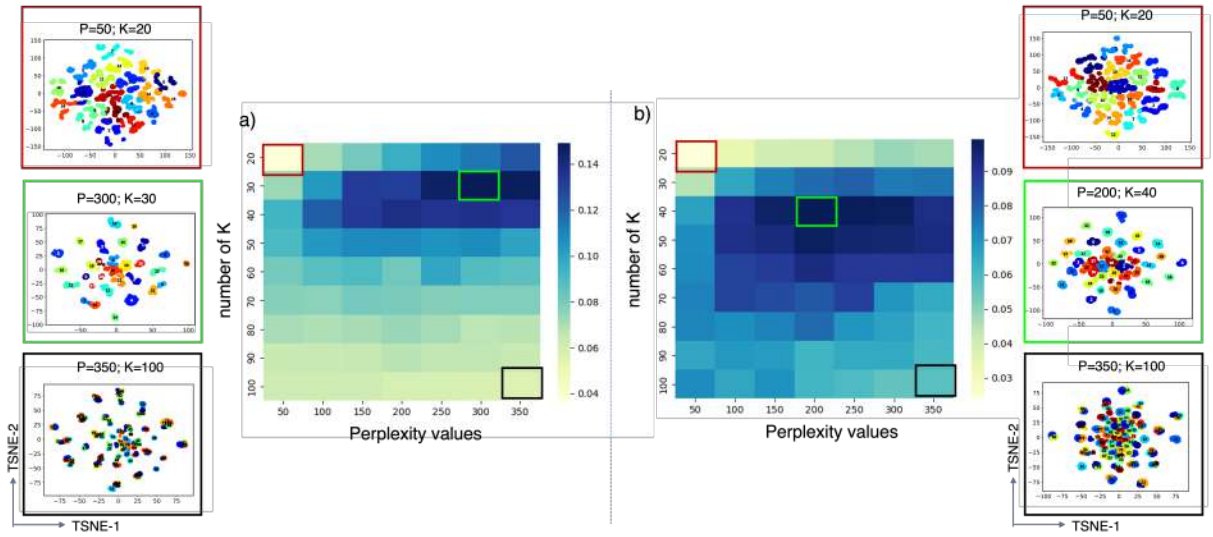


FIG. 1: Hyperparameter optimization based on integrated Silhouette score for the (a) APO, and (b) G5-bound ensembles. t-SNE map obtained with selected optimal and sub-optimal values of P and K illustrates how these parameters affect the clustering efficiency. While with fewer than an optimal number of clusters, different states are merged into one group, with too many clusters we see a breakdown of a single cluster into multiple pieces.

the low dimensional space (S_{ld}) as well as at the high dimensional space (S_{hd}) (Table S2 S3) suggesting reliable clustering. This can be compared against the large negative score (-0.6) with respect to the high dimensional distance, obtained for the classical GROMOS-based clustering, which indicates that the conformations are grouped into wrong clusters. In Fig. 1(a,b), we report the integrated score (S_{hd}*S_{ld}) as measured for the clusters in APO and G5-bound ensembles of Aβ42. The figure shows the highest Silhouette score for a range of small perplexities but only with a single optimal cluster size (30 in the case of APO trajectory and 40 in G5 bound trajectory). This consistent behavior confirms that at the optimal parameters, the identified clustering pattern is truly intrinsic to the underlying ensemble structure and corresponds roughly to the possible number of metastable structures, rather than just random noise. At these optimal values, we find that the low-dimensional t-SNE map shows discrete clusters in both APO and G5-bound ensemble. Whereas at sub-optimal values, the identified clusters either encompass different pieces together in a single cluster (for example at P=50; K=20 in APO) or break into multiple clusters of similar conformations (at P=350, K=100 in APO system).

2. *Clustering reveals ordered sub-states within disordered Aβ42 ensemble*

Once the optimal number of clusters for a given data set is decided using the prescriptions described above, we inspect the uniqueness and homogeneity of individual clusters by back-mapping to the conformations in the bound and unbound ensembles. Fig. 2 shows the conformations within each cluster of Aβ42 ensemble indicating unique topology and secondary structural architecture. To understand this quantitatively, we reordered the conformations based on the cluster indices and plotted their pairwise similarity/distances in Fig. S3 and Fig. S4 for APO and G5 bound Aβ ensemble, respectively, and compare them with the distance map before clustering. The results show that the clusters obtained with optimal parameters indeed yield better homogeneity than that obtained with sub-optimal parameters. The conformational distance is measured based on the RMSD of inter-residue LJ energies. To further accentuate the homogeneity illustrations, we have also plotted the respective RMSD of cartesian coordinates. Please note that for the clustering with t-SNE, only the RMSD of LJ energies was used. For the input maps (Fig S3a and S4a), we ordered the conformations sequentially in the X and Y axes. For the maps generated after clustering, the frames are sorted based on the cluster indices and placed from 0th cluster to N^{th} cluster (S3-S4, b-d for both the upper and lower panel in Fig. S3 and Fig. S4, respectively). As indicated by the figures, the input distance maps of these ensembles show a certain level of conformational memory across the contiguous frames (the Red blocks/grids at the diagonal band) as the trajectories are generated from a history-dependent metadynamics approach. Nevertheless, the clustering obtained with sub-optimal parameters adversely affects even this intrinsically clustered data and several off-diagonal Red patches appear in the plot indicating either wrong groupings or broken clusters. On the other hand, with the optimal parameters, the algorithm yields better clustering as indicated by the thickening of the diagonal Red blocks and reduction of off-diagonal Red patches.

More interestingly, though the G5 bound conformational ensemble was clustered only based on the similarities of protein conformations, the ligand is shown to have a specific binding orientation with the protein within each cluster (Fig. 2(b)). This is an exciting result that sheds light on the hidden ordered features in a disordered IDP ensemble, which can confer specificity for ligand binding. The ability of t-SNE to cluster a seemingly disordered

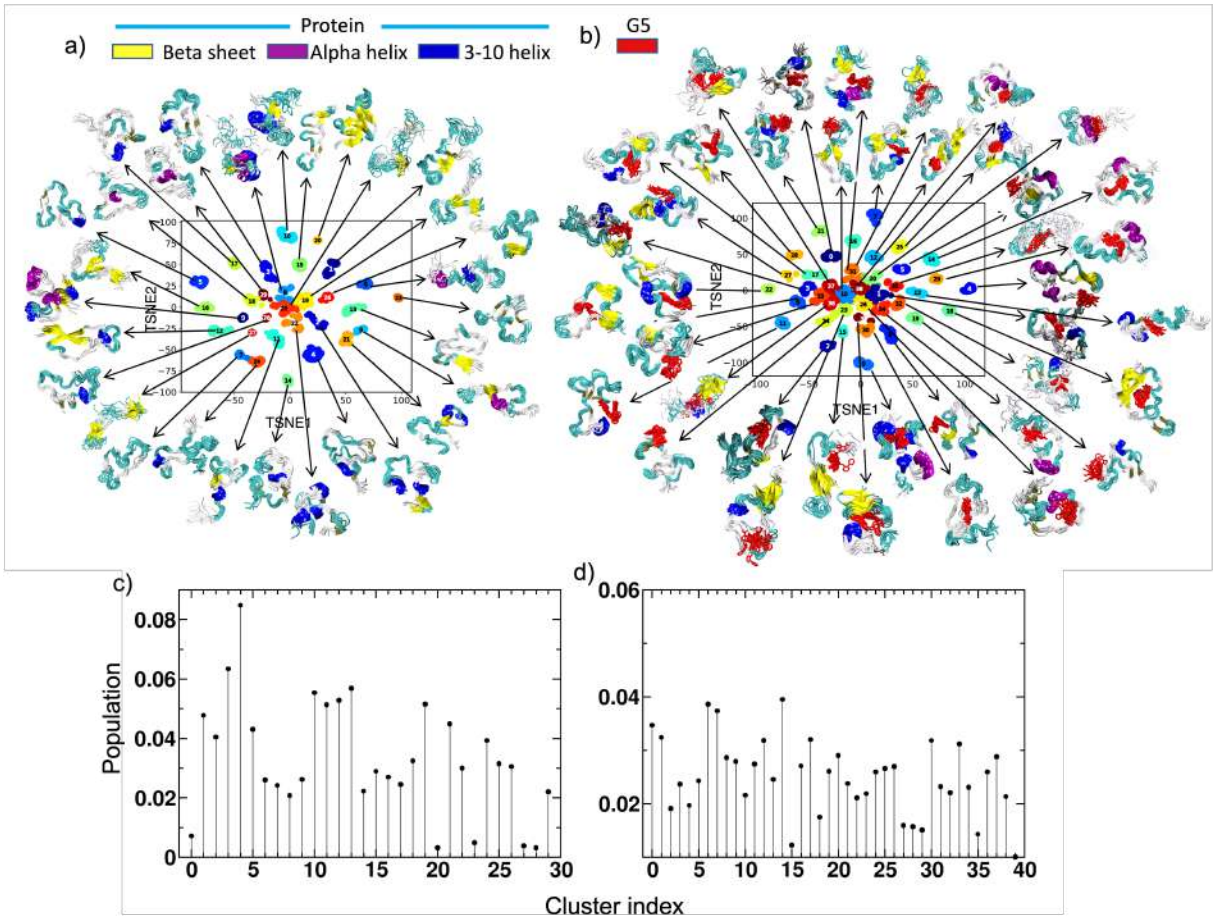


FIG. 2: t-SNE based conformational clustering of Aβ-42 ensembles in the absence and presence of G5 in a and b respectively. The cluster-wise population statistics is shown in Fig 2c and d.

ensemble into substates with distinct structural features can vastly reduce the library of conformational screening from a large ensemble into a handful of structures. This will aid in a high throughput structural and statistical analysis of IDP ensemble data and can have a huge impact on our fundamental understanding of disorder-function relationships and in the design of therapeutic drugs for IDP molecules.

3. Insights into the binding properties of Aβ42 with G5

The cluster-based population statistics of different metastable conformations have been analyzed and shown in Fig. 2(c,d). The results indicate that the distributions are more equally probable in the case of the ligand-bound ensemble than in the APO state. From this

population distribution of different metastable conformations, we have estimated that the Gibbs conformational entropy ($-\sum(p\ ln\ p)$) of the G5-bound ensemble is larger than the apo ensemble (Fig. S5 in SM). The number of optimal unique conformations (30 in APO versus 40 in G5-bound) and their respective Silhouette score (in high dimension space, 0.21 versus 0.15) (Table S1 and S2) also suggest consistent observation. Taken together, these results further corroborate the entropic expansion on ligand binding as deduced from the earlier studies[89]. Though the ligand has very specific binding geometry within each cluster, they vary significantly across the different clusters. We show the contact probabilities of G5 with individual protein residues in Fig. 3(a) for individual clusters. We also plot the residue-wise contact probabilities using the total trajectory, which provides averages without clusters (Fig. S6 in SM). As indicated by the figures, the G5 preferentially binds to aromatic residues such as Tyr/Phe (residue numbers 10, 19, 20) and hydrophobic residues such as Ile/Val/Met (residue numbers 31, 32, 35, 36). The screening of such large aromatic and hydrophobic residues by the ligand would possibly limit the transient order stabilization and likely increase the heterogeneity/entropy. To further quantify how the contacts of G5 at diverse locations affect the interaction strength, we applied a high throughput numerical technique called molecular mechanics with generalized Born and surface area solvation (MM/GBSA) to loosely estimate the free energy of the binding of ligands to proteins[93,94]. Our MMGBSA-derived binding scores are shown in Fig. 3(b). We see that the G5 binds at relatively equal strength in multiple clusters. But interestingly, we also noted a few of the clusters (cluster numbers 14, 29, and 30) that show statistically stronger binding than the others. More interestingly, these same clusters consist of a relatively larger population in the ensemble than the other conformers. The protein residues involved in binding in these selected clusters along with their energy contributions to the total energy as plotted in Fig. 3(c) and the conformational binding-geometry for the cluster that exhibits the most favorable MM/GBSA binding is shown in Fig. 3(d,e). In Fig. S7 in SM, we also show the same data (binding geometries and residue-wise interactions) for the two other clusters, which show the second and the third best MM/GBSA scores. Our analyses reveal that ligands interact with multiple favorable sites simultaneously, which indicates that even a partially collapsed or ordered state of an IDP can provide a specific binding pocket for small molecule interactions. These unique insights gained as a result of high-fidelity clustering can be lever-

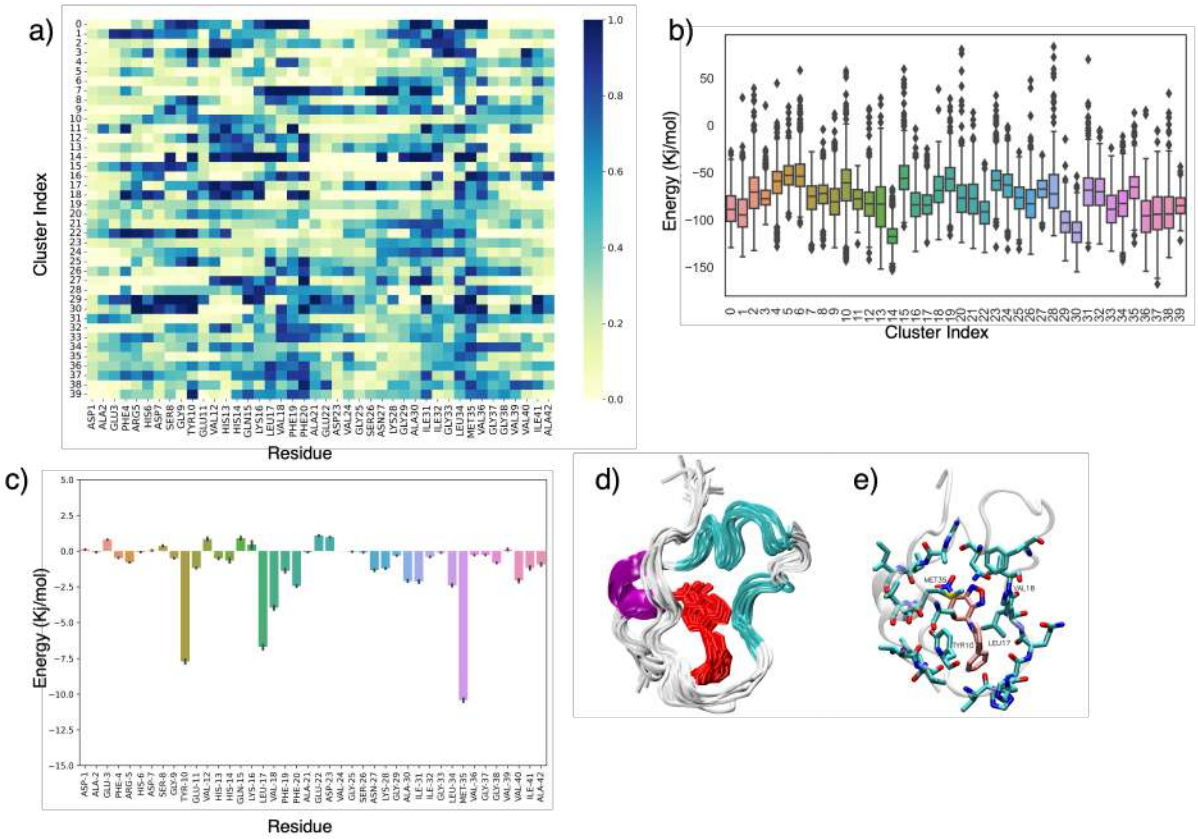


FIG. 3: Cluster-wise contact probabilities and their respective binding affinity as measured using MMGBSA analysis are shown in (a) and (b) respectively. For the cluster that shows the most favorable binding (cluster no: 14)), we have shown the residue-wise decomposed energy contribution in (c). The superposition of ten central conformations from this specific cluster is shown in (d) and the interacting residues are shown in stick representation in (e)

aged for future IDP-drug designing with time-tested strategies that are similar to the ones applied for the folded proteins.

C. t-SNE for clustering αS conformational ensembles

1. t-SNE reveals distinct conformation sub-states despite extreme plasticity

Next, we examine our clustering framework for understanding the metastable conformations of αS. This is a longer IDP consisting of 140 amino acids and has a more fuzzy conformational landscape than the Aβ42 peptides with no persistent residual secondary structure propensity. The shorter constructs consisting of only the C-terminal residues that directly bind to Fasudil and Ligand-47 have also been shown to have extremely heterogeneous confor-

mations. The t-SNE projection produces a fully crowded map for the full-length αS, except few well-segregated dense clusters (Fig. S8a). For the c-ter APO system, there is an almost complete overlap of the conformation space (Fig. S8b) suggesting extreme heterogeneity in their conformational landscape. This is very unlike what is seen in Fig. 1 for Aβ-42.

In order to get a better sense of the conformational diversity, we plotted the distance map as done before for Aβ42 peptide. Here again, due to the extreme heterogeneity, the conformational states rapidly exchange among themselves, which in turn creates a very cluttered distance map of the original trajectory. This is shown in the first subplot for full αS in Fig. S9(a) and for the c-terminal (c-ter) peptide in Fig. S9(b). This suggests there are very less intrinsic groupings of these conformations in the high dimensional space, which is consistent with the tSNE projections seen in Fig. S8. As a result, the clusterability or Silhouette score is very less though with a positive value closer to zero for S_{hd} indicating poor clusterability of these ensembles. However, we find that in highly fuzzy IDP systems with complex heterogeneous conformations, obtaining a positive clustering coefficient itself is a positive sign. A small positive score can indicate that the sample is on or very close to the decision boundary between two neighboring clusters. In such cases, the ideal number of clusters would be either too many smaller clusters (for a scattered map as in the case of full-length αS) or a single large cluster (for a continuous map as in the case of c-ter) and hence must be chosen based on the knowledge of how finer or coarser details need to be captured. Therefore, we have chosen the maps generated with K=50 and perplexity=400 for the full-length system and K=20 and perplexity=1800 for the C-terminus system for further analyses (Fig. S10). We divide the continuous blob of t-SNE projection space into 20 contiguous regions for the APO c-ter system and 50 clusters for full-length protein with the idea of obtaining a cluster-wise visualization that is more informative than the raw ensemble data. This choice of hyperparameters is further be justified by measuring the extent of homogeneity in the intra-cluster conformations in comparison to the inter-clusters. As shown by the distance maps generated after the clustering exercise shown in Fig. S9, we obtained clusters with balanced intra-cluster homogeneity and inter-cluster diversity. This is characterized by deeper Red in the diagonal and Blue in the off-diagonal region of the distance map and is particularly distinct in the case of the clustering with K=50 and K=20 for the full-length and c-terminal ensembles, respectively. The other choices of Ks do not

give such clear distinctions in clusters as the number of clusters ends up being either too high or too low. This further accentuates the ability of the clustering algorithm toward local neighborhood preservation.

Visual representation of the conformations in the 50 subgroups of APO full-length αS system and 20 subgroups of APO c-ter αS system are shown in Fig. 4(a,b). In spite of the extreme heterogeneity of the original space in the αS ensemble, the subspace represented by the t-SNE captures distinct clusters with unique and uniform conformations in both systems. The conformations neither have any secondary structure nor collapsed to form rigid pockets as in the case of the Aβ system. Yet, there seems an order within each of the clusters. Interestingly, we find that the conformations of C-ter αS peptide have a conformational range that ranges from a fully-extended rod-like shape to acutely bent hairpin conformations (4(b)) and with different degrees of bending angles between these two extremes. To bring out this feature very clearly, we have presented the clusters in a sequence, arranged based on the average bend angle measured between the C-alpha atoms of residues 121, 131, and 140 (Fig. 5(a)), which make up the c-terminal, middle and N-terminal end of the peptide. We have plotted the distribution of the bend angles in Fig. 5(b). As evident from the figure, the single-dimensional bend angle could nicely separate out the fuzzy disordered ensemble. An interesting and desired bye-product of this high-fidelity clustering is that it seems to inform a single collective variable (bend angle in this case) that uniquely defines the various conformations across clusters. And, this could be useful for further thermodynamics and kinetics studies that are often explored with a limited number of well-defined collective variables due to algorithmic and computational limitations.

2. Insights into conformation kinetics and binding of c-terminal αS

Next, we compare the Fasudil and Lig47 bound geometries of c-ter αS along with the APO state in order to elucidate their differential ligand binding behavior. The t-SNE maps for the Fasudil and Lig47 bound c-ter αS individually are shown in Fig. S11 for different values of perplexities. Like the APO c-ter projections, the t-SNE maps for Fasudil bound (Fig. S11 (a)) and Lig47-bound (Fig. S11 (b)) produce an overspread and continuous plot with no distinct groups and therefore we used a similar value of K equals to 20 to cluster

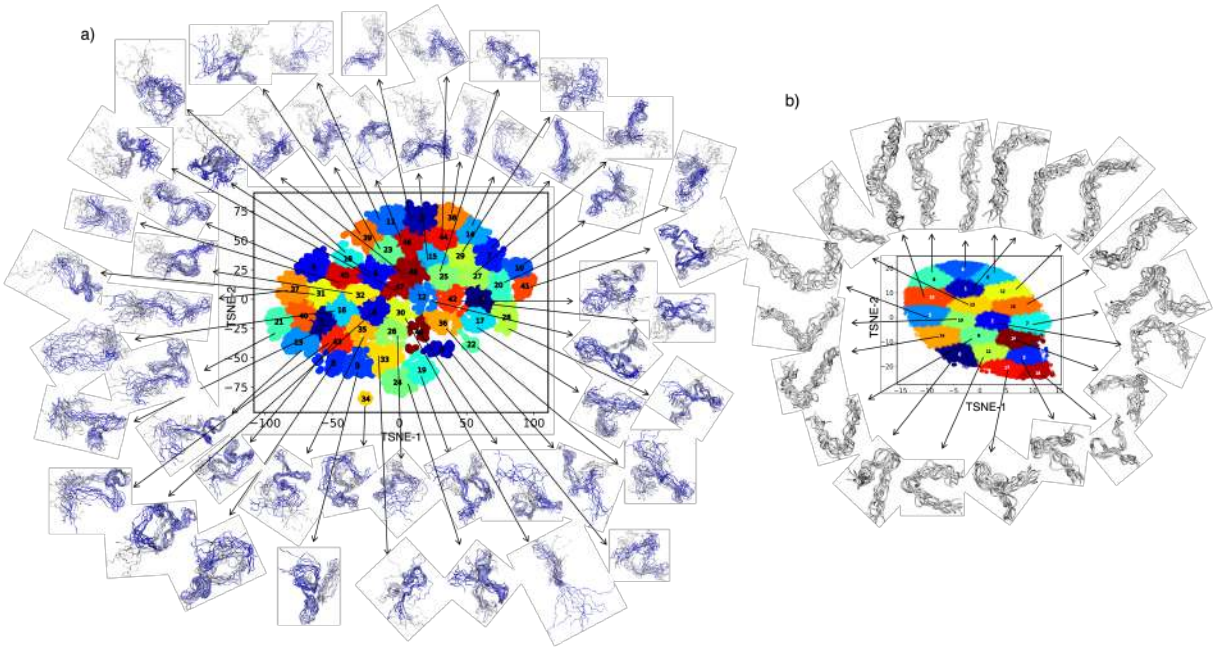


FIG. 4: t-SNE based conformational clustering of 140 amino-acids long full-length APO αS ensembles and the 20 amino-acids long c-terminal APO ensemble.

them. Also, much like the APO c-ter αS, the bound peptides exhibit a conformational space that ranges from acutely bent hairpin-like conformation to rod-like conformations as shown in Fig. 5(d,g).

Unlike the localized binding of G5 with individual metastable states of Aβ42, the binding of ligand seems to be highly non-localized with αS and occurs at scattered sites of the protein (Fig. 5(d,g). This is mainly because, as reported in one of our recent papers[91], only weak specificity is exerted between the small molecules and the protein. However, when comparing the conformations of Fasudil versus Lig-47 bound protein, we noticed that the latter exert better localization/specificity over the former, particularly with the bent states. Yet the binding does not change the bending profile or the population of states greatly, indicating that the ligands could not sufficiently lock the protein in a specific conformation(s) or induce any change in the conformational space. To show this clearly, we have shown the accessible tSNE conformational space of the collated APO and ligand-bound trajectories. As shown in Fig. S12(a), the conformational space overlaps almost fully suggesting that the binding does not seem to favor one structure over the other and is ubiquitous. This is in severe contrast to the behavior exhibited by Aβ42 APO and ligand-bound t-SNE projections as shown in Fig. S12 (b). The map in Fig. S12 (b) clearly shows that the APO and bound Aβ42 ensembles

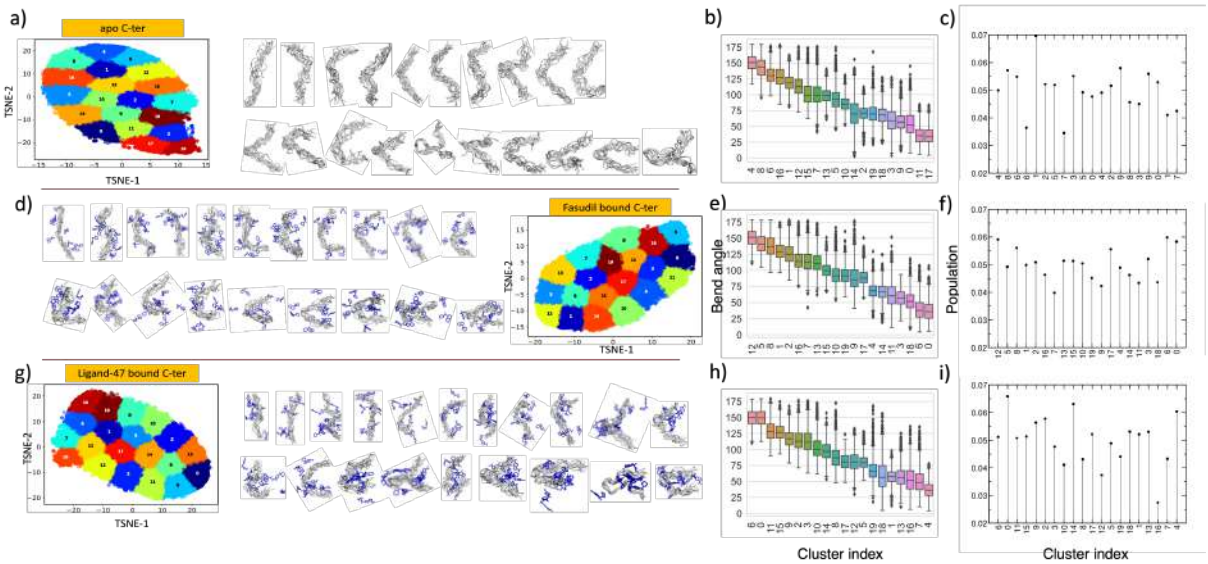


FIG. 5: T-SNE based conformational clustering of apo, Fasudil-bound and ligand47-bound C-terminal α-Synuclein ensembles in a, b and c respectively. The conformational subspace along the T-SNE axes is subdivided into 20 clusters. The structure of these conformations within each cluster shows unique architecture (a-c). The clusters of conformations are arranged based on the bend angle calculated on the CA atoms of residues 121, 131, and 140 and shown from most extended to acutely bent one. The distribution of the bend angle in each cluster is shown as box plot in (d) to (f) for the apo, Fasudil-bound, and Ligand47-bound ensembles. The corresponding population of states in the three systems is shown in g-i.

have clusters that are distinct with only a few regions showing overlapping projections.

Though the different extended and bent conformations are almost equiprobable in both APO and ligand-bound αS systems (Fig 5c,f and i), the bent states exhibits better kinetic stability than the extended states as shown by their transition probability in Fig 6(a-c). As noted before, the same bent states were shown to localize ligand-47 in specific sites than the Fasudil. Hence to understand the differential binding of Fasudil and Ligand-47 across the different states, we looked at the probability of contact of each residue on the C-ter αS (Fig. 6(d,e). We find that the probability of the contacts made by Lig-47 is much stronger than Fasudil. In particular, the bent states favor the strong and simultaneous binding of three Tyr residues (residue 5, 13, and 16) with Lig-47 (Fig 6 d-h).

Taken together, the results of t-SNE based clustering exercise on α-S reveal one of the most interesting and non-obvious learning about the emergence of the possible role of peptide local curvatures as sites of ligand binding. We see that besides the weak chemical specificity

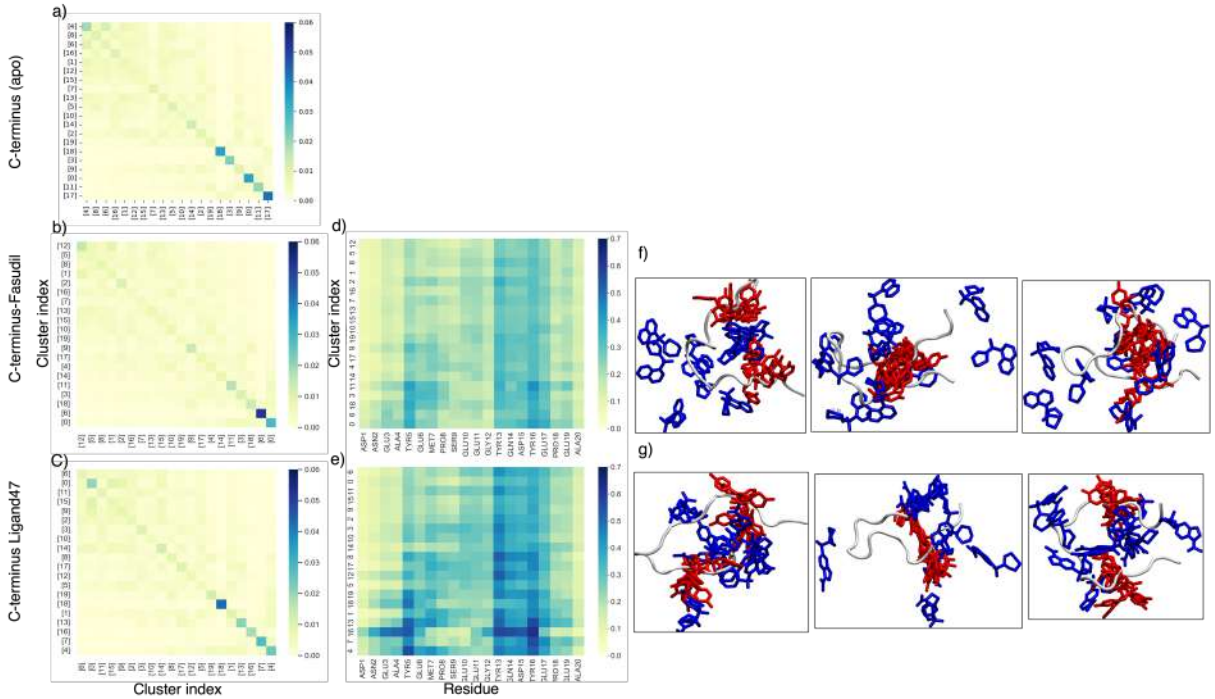


FIG. 6: The cluster-wise transition probability is shown for the apo (a), Fasudil-bound (b) and Ligand47-bound (c) αS_{c-ter} ensembles. Residue-wise inter-molecular contact probabilities between protein and the two ligands across different clusters individually are shown in (d) and (e). The inter-molecular contacts are shown for the three acutely bent αS_{C-ter} conformations with Fasudil (cluster no: - 18, 6, 0) in (f) and with ligand-47 (cluster no: - 16, 7, 4) in (g).

through the aromatic-pi interactions, the small molecules recognize certain local curvature in the protein. Such a physical complementarity is more prominent in the case of Lig-47 which clutches to the bent conformations and as a result, possibly exerts a stronger affinity with the protein as compared to the Fasudil molecule. This result possibly explains the experimental observation as to why the Lig47 binds more strongly than the Fasudil to αS protein. In addition, the study hints at a novel route of targeting the sites on the IDP that may take up high local curvatures/bent as a potential therapeutic intervention strategy against the surging IDP-driven pathologies.

III. Conclusion

In spite of the well-established knowledge of the inherent conformational heterogeneity in an IDP ensemble and despite advances made in accurately determining the ensemble conformations using integrative approaches, successful application of IDPs to drug targeting is limited. The main reason behind this is the lack of accurate classifications of the conformational ensemble. Our algorithm provides that tool where several thousands of structures can be grouped into representative sets of the distinct and tractable conformational library, which invariably will aid while carrying out *in silico* functional and drug screening studies in a rational manner. This tool also makes it very convenient to generate sub-groups of similar conformations for long IDPs, whose full conformational ensemble is highly intractable for structural biophysical analyses. For example, we applied our algorithm to study how the long disordered regions of FUS protein interact with RNA molecules[95] and this t-SNE tool allowed us to illustrate the complex binding behavior of FUS with RNA in an interpretable manner.

While highlighting the advantages of t-SNE, we do acknowledge that unlike projection techniques such as PCA and MDS, the interpretation of t-SNE projection is non-trivial. One would need to train and process the data set extensively to arrive at the hyperparameters that provide a clustering solution that can be meaningfully interpreted and visualized. This limitation arises since t-SNE optimization is non-convex in nature with random initialization that produces different sub-optimal visual representations at different runs. However, this affects mainly the global geometry and hierarchical positioning of the clusters and not the local clustering pattern. We illustrate the consistency in local clustering upon different runs with different random initialization by quantifying the silhouette score and mutual information of clusters in Table S4. Moreover, finding a single optimal global geometry of the IDP dataset is not often possible owing to their extreme heterogeneity with almost equal transition probability between different clusters. However, if one necessitates the global preservation, tuning the perplexity[96], and other parameters like Early exaggeration and Learning rate, initializing with PCA and Multi-scale similarities will be helpful[75,97]. In addition, some of the variations of t-SNE methods such as h-SNE can also be helpful[98].

Another factor that should be considered while using t-SNE on ultra-large datasets is the

associated computational cost. Analyzing large data sets (beyond n $\gg 10^6$) with t-SNE is not only computationally expensive (scales with $O(n^2)$), but also suffers from slow convergence and fragmented clusters. If the computational cost becomes formidable, one could use methods such as Barnes-Hut approximation[99] and the FIT-SNE method to accelerate the computation. In short, Barnes-Hut approximation considers a subset of nearest neighbors for modeling the attractive forces and the FIT-SNE method relies on a fast Fourier transformation, which reduces the computational complexity to $O(nlogn)$ and $O(n)$, respectively. To mitigate the slow convergence and fragmentation of clusters, it is often desirable to run t-SNE on a sub-sample of the trajectory that includes all unique populations and then projects the rest of the points onto the existing map.

In line with discussing the possible pitfalls of the t-SNE method, it is also understood that adding a new data point onto the existing t-SNE map can lead to erroneous interpretation as the method is essentially non-parametric and does not directly construct any mapping function between the high dimensional and low-dimensional space. Recent extensions of the method in combination with deep neural networks allow for parametric mapping[75,100] and could be tried if such a situation can not be avoided. Moreover, the possibility of out-of-sample mapping with parametric t-SNE can be explored further for driving simulations from one state to another and to match experimentally known values. For instance, in such cases, the similarities can be obtained from NMR chemical shifts or from SAXS intensities. From that perspective, t-SNE as an integrative modeling tool looks very promising.

IV. Materials and Method

A. Input for t-SNE analysis

The systems details about the trajectories of alanine-dipeptide, Aβ42 and αS ensembles are reported in Table 1. The conformations of the trajectories were represented by backbone dihedral angle, inter-residue LJ-interaction potential, and atomic coordinates of heavy atoms for alanine-dipeptide, Aβ, and αS ensembles, respectively.

1. t-SNE based dimensional reduction

Given a number of observations (conformations) n and with d dimensional input features in the original space defined as $X = \{x_1, x_2, \ldots, x_n\} \in \mathrm{R}^d$, t-SNE maps a smaller s dimensional embedding of the data that we denote here by $Y = \{y_1, y_2, \ldots, y_n\} \in \mathrm{R}^s$. Here $s \ll d$ and typically s = 2 or 3. This projection is based on the similarity and dissimilarity between conformations. The similarity or dissimilarity between the conformations in the high dimensional space is computed based on Euclidean or RMS distances. t-SNE aims to preserve the local neighborhood such that the points that are close together in the original space remain closer in the embedded space. In the original space, the likelihood of a point x_j to be the neighborhood of x_i instead of every other point x_k is modeled as a conditional probability $p_{(j|i)}$ assuming the Gaussian distribution centered at point x_i with a standard deviation of σ_i.

$$p_{(j|i)} = \frac{exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})}{\sum_{k \neq i} exp(-\frac{\|x_i - x_k\|^2}{2\sigma^2})} \tag{1}$$

Similarly, the conditional probability in the embedded space ($q_{(j|i)}$), with the same n points initialized randomly, is computed but now based on a t-distribution. Having a longer tail than Gaussian, the t-distribution moves dissimilar points farther away to ensure less crowding in the reduced space.

$$q_{(j|i)} = \frac{(1 + \| y_i - y_j \|^2)^{-1}}{\sum_{k \neq i} (\| y_i - y_k \|^2)^{-1}}$$

(2)

To ensure symmetry in the pairwise similarities, the joint probability is calculated from the conditional probability as follows:

$$p_{ij} = (p_(j|i) + p_(i|j)/2n \tag{3}$$

Finally, the difference between the two probability distributions, calculated as Kullback-Leibler (KL) divergence is then minimized by iteratively rearranging the points in the low

dimensional space using gradient descent optimization.

$$C = KL(P|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ji}}{q_{ji}} \tag{4}$$

where P and Q are the joint probability distributions in the high and low dimensional space over all the data points.

The major tunable hyperparameters in t-SNE are the perplexity, learning rate and the number of iterations. The perplexity value, P defines the Gaussian width, σ_i, in Equation 1 above such that, $log_2 P = H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$ for all i. Loosely, this parameter controls the number of nearest neighbors each point is attracted to and therefore balances the preservation of similarities at a local versus global scale. Typically, low perplexity values tend to preserve the finer local scale and high perplexity values project a global view. To optimize perplexity, we ran the algorithm with varying values of perplexities and chose the one that yields a high silhouette score. The other two parameters such as the learning rate and the number of iterations control the gradient descent optimization. While we chose the default value of 200 for the learning rate, the number of iterations was chosen to be 3500, which is large enough for avoiding random fragmentation of clusters as suggested in the literature.[97]

B. Kmeans clustering of data on the reduced space obtained from t-SNE

Kmeans clustering is the simplest unsupervised clustering algorithm that partitions the data into non-overlapping clusters. The algorithm starts by grouping data points randomly into K clusters, as specified by the user. Then it iterates through computing the cluster centroids and reassigning data points to the nearest cluster centroid until no improvements are possible. The parameter, K, is optimized by running at various values and chosen based on the maximized clustering efficiency.

C. Optimizating the hyperparameters ($Perp$ in t-SNE and K in k-means) using Silhouette score

Silhouette score for a datapoint i is measured by,

$$S_i = \frac{(b_i - a_i)}{max(b_i - a_i)} \tag{5}$$

where a_i is the intra-cluster distance defined as the average distance to all other points in the cluster to which it belongs. b_i represents the inter-cluster distance measured as the average distance to the closest cluster of datapoint i except for that it's a part of. Typically the silhouette score ranges between 1 and -1, where a high value indicates good clustering, and values closer to 0 indicate poor clustering. A negative value indicates the clustering configuration is wrong/inappropriate.

The distance between points is usually measured in terms of the Euclidean distance metric. Since the clusters, in our case are identified in a reduced representation with TSNE, computing the score based only on the distances in the reduced space ($S_l d$) may be misleading, if the points are wrongly put together during the dimensional reduction step by t-SNE. Therefore, it is important to measure the goodness of clustering with respect to the original distance in the high dimensional space ($S_h d$), in addition to that in the low dimensional space. The integrated score ($S_l d * S_h d$), therefore, adds value to the estimated clustering efficiency in terms of reliability.

D. Cluster-wise conformational analysis and visualization

The conformations corresponding to each cluster are extracted using Gromacs based on the cluster indices. All the conformations were used for estimating the contact probability, binding energy, and homogeneity within individual clusters. Whereas, for visualization purposes, we extracted ten representative conformations from each cluster that is closest to the corresponding cluster centroid (as identified using KD-tree based nearest neighbor search algorithm). The conformations are rendered using VMD.

Our current implementation of the model is available on the GitHub repository: https://github.com/codesrivastavalab/tSNE-IDPClustering.

V. Author Contributions

R.A. and A.S. conceived and designed the research; A.R. performed the calculations with help from J.K; R.A., J.K., M.B., P.R., and A.S. analyzed data; R.A.and A.S. wrote the paper together with inputs from J.K., M.B., P.R.

VI. Acknowledgments

A.R. thanks the Wellcome Trust DBT India Alliance for Early Career Fellowship (Grant number: IA/E/18/1/504308). A.S thanks the Department of Science and Technology (DST) of India for the early career grant (SERB-ECR/2016/001702). A.S. also thanks the DST for the National Supercomputing Mission grant (DST/NSM/R&D_HPC_Applications/2021/03.10). Computational support from the high-performance computing facility "Beagle" setup from grants by a partnership between the Department of Biotechnology of India and the Indian Institute of Science (IISc-DBT partnership program) is greatly acknowledged. AR and AS are grateful to the SciNet HPC Consortium, ComputeCanada for their generous computational support.

References

[1]Kevin M. Ulmer. Protein engineering. Science, 219(4585):666–671, 1983.

[2]Jeremy R. Knowles. Tinkering with enzymes: What are we learning? Science, 236(4806):1252–1258, 1987.

[3]Samuel H. Gellman. Introduction: molecular recognition. Chemical Reviews, 97(5):1231–1232, 1997.

[4]David Mobley and Ken Dill. Binding of small-molecule ligands to proteins: "what you see" is not always "what you get". Structure (London, England : 1993), 17:489–98, 05 2009.

[5]David Boehr, Ruth Nussinov, and Peter Wright. The role of conformational ensembles in biomolecular recognition. Nature chemical biology, 5:789–96, 11 2009.

[6]Jerome M. Fox, Mengxia Zhao, Michael J. Fink, Kyungtae Kang, and George M. Whitesides. The molecular origin of enthalpy/entropy compensation in biomolecular recognition.

Annual Review of Biophysics, 47(1):223–250, 2018.

[7]Petrus Jansen van Vuren, Alexander J. McAuley, Michael J. Kuiper, Nagendrakumar Balasubramanian Singanallur, Matthew P. Bruce, Shane Riddell, Sarah Goldie, Shruthi Mangalaganesh, Simran Chahal, Trevor W. Drew, Kim R. Blasdell, Mary Tachedjian, Leon Caly, Julian D. Druce, Shahbaz Ahmed, Mohammad Suhail Khan, Sameer Kumar Malladi, Randhir Singh, Suman Pandey, Raghavan Varadarajan, and Seshadri S. Vasan. Highly thermotolerant sars-cov-2 vaccine elicits neutralising antibodies against delta and omicron in mice. Viruses, 14(4), 2022.

[8]Xiao chen Bai, Greg McMullan, and Sjors H.W Scheres. How cryo-em is revolutionizing structural biology. Trends in Biochemical Sciences, 40(1):49–57, 2015.

[9]Michael J. Robertson, Justin G. Meyerowitz, and Georgios Skiniotis. Drug discovery in the era of cryo-electron microscopy. Trends in Biochemical Sciences, 47(2):124–135, 2022. Special Issue: Pushing boundaries of cryo-EM.

[10]Martin Turk and Wolfgang Baumeister. The promise and the challenges of cryo-electron tomography. FEBS Letters, 594(20):3243–3261, 2020.

[11]Vidya Mangala Prasad, Daniel P. Leaman, Klaus N. Lovendahl, Jacob T. Croft, Mark A. Benhaim, Edgar A. Hodge, Michael B. Zwick, and Kelly K. Lee. Cryo-et of env on intact hiv virions reveals structural variation and positioning on the gag lattice. Cell, 185(4):641–653.e17, 2022.

[12]Yigong Shi. A glimpse of structural biology through x-ray crystallography. Cell, 159(5):995–1014, 2014.

[13]Deepthi Joseph, Smruti Ranjan Nayak, and Aravind Penmatsa. Structural insights into gaba transport inhibition using an engineered neurotransmitter transporter. The EMBO Journal, 41(15):e110735, 2022.

[14]Ishika Pramanick, Nayanika Sengupta, Suman Mishra, Suman Pandey, Nidhi Girish, Alakta Das, and Somnath Dutta. Conformational flexibility and structural variability of sars-cov2 s protein. Structure, 29(8):834–845.e5, 2021.

[15]Sang Park, Bibhuti Das, Fabio Casagrande, Ye Tian, Henry Nothnagel, Mignon Chu, Hans Kiefer, Klaus Maier, Anna De Angelis, Francesca Marassi, and Stanley Opella. Structure of the chemokine receptor cxcr1 in phospholipid bilayers. Nature, 491, 10 2012.

[16]Ashok Sekhar and Lewis E. Kay. An nmr view of protein dynamics in health and disease. Annual Review of Biophysics, 48(1):297–319, 2019.

[17]John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon Kohl, Andrew Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. Nature, 596:1–11, 08 2021.

[18]Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 373(6557):871–876, 2021.

[19]Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp) – round xiv. Proteins: Structure, Function, and Bioinformatics, 89, 09 2021.

[20]Sriram Subramaniam and Gerard J Kleywegt. A paradigm shift in structural biology. Nat Methods, 19:20–23, 2022.

[21]P. Tompa. Intrinsically unstructured proteins. Trends Biochem. Sci., 27:527, 2002.

[22]A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky. Flexible nets. the roles of intrinsic disorder in protein interaction networks. FEBS J., 272:5129, 2005.

[23]H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol., 6:197, 2005.

[24]J. Habchi, P. Tompa, S. Longhi, and V. N. Uversky. Introducing protein intrinsic disorder. Chem. Rev., 114:6561, 2014.

[25]P. E. Wright and H. J. Dyson. Intrinsically disordered proteins in cellular signaling and regulation. Nat. Rev. Mol. Cell Biol., 16:18, 2015.

[26]Prakash Kulkarni, Vitor BP Leite, Susmita Roy, Supriyo Bhattacharyya, Atish Mohanty, Srisairam Achuthan, Divyoj Singh, Rajeswari Appadurai, Govindan Rangarajan, Keith Weninger, et al. Intrinsically disordered proteins: Ensembles at the limits of anfinsen's dogma. Biophysics Reviews, 3(1):011306, 2022.

[27]M Madan Babu, Robin van der Lee, Natalia Sanchez de Groot, and Jörg Gsponer. Intrinsically disordered proteins: regulation and disease. Current Opinion in Structural Biology, 21(3):432–440, 2011.

[28]M. R. Jensen, R. W. Ruigrok, and M. Blackledge. Describing intrinsically disordered proteins at atomic resolution by nmr. Curr. Opin. Struct. Biol., 23:426, 2013.

[29]P. Sormanni, D. Piovesan, G. T. Heller, M. Bonomi, P. Kukic, C. Camilloni, M. Fuxreiter, Z. Dosztanyi, R. V. Pappu, M. M. Babu, S. Longhi, P. Tompa, A. K. Dunker, V. N. Uversky, S. C. Tosatto, and M. Vendruscolo. Simultaneous quantification of protein order and disorder. Nat. Chem. Biol., 13:339, 2017.

[30]Gregory-Neal W Gomes, Mickaël Krzeminski, Ashley Namini, Erik W Martin, Tanja Mittag, Teresa Head-Gordon, Julie D Forman-Kay, and Claudiu C Gradinaru. Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. Journal of the American Chemical Society, 142(37):15697–15710, 2020.

[31]Valéry Ozenne, Frédéric Bauer, Loïc Salmon, Jie-rong Huang, Malene Ringkjøbing Jensen, Stéphane Segard, Pau Bernadó, Céline Charavay, and Martin Blackledge. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics, 28(11):1463–1470, 2012.

[32]Howard J. Feldman and Christopher W.V. Hogue. A fast method to sample real protein conformational space. Proteins, 39:112–131, 2000.

[33]Gary W. Daughdrill, Stepan Kashtanov, Amber Stancik, Shannon E. Hill, Gregory Helms, Martin Muschol, Véronique Receveur-Bréchot, and F. Marty Ytreberg. Understanding the structural ensembles of a highly extended disordered protein. Mol. BioSyst., 8:308–319, 2012.

[34]Wing-Yiu Choy and Julie D. Forman-Kay. Calculation of ensembles of structures representing the unfolded state of an sh3 domain. Journal of Molecular Biology, 308(5):1011–

1032, 2001.

[35]Pau Bernadó, Laurence Blanchard, Peter Timmins, Dominique Marion, Rob W. H. Ruigrok, and Martin Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proceedings of the National Academy of Sciences, 102(47):17002–17007, 2005.

[36]Gabrielle Nodet, Loïc Salmon, Valéry Ozenne, Sebastian Meier, Malene Ringkjøbing Jensen, and Martin Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from nmr residual dipolar couplings. Journal of the American Chemical Society, 131(49):17908–17918, 2009.

[37]Malene Ringkjøbing Jensen, Loïc Salmon, Gabrielle Nodet, and Martin Blackledge. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. Journal of the American Chemical Society, 132(4):1270–1272, 2010.

[38]Joseph A. Marsh and Julie D. Forman-Kay. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. Journal of Molecular Biology, 391(2):359–374, 2009.

[39]Lisa M. Pietrek, Lukas S. Stelzl, and Gerhard Hummer. Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations. Journal of Chemical Theory and Computation, 16(1):725–737, 2020.

[40]Florencia Klein, Exequiel E Barrera, and Sergio Pantano. Assessing sirah's capability to simulate intrinsically disordered proteins and peptides. Journal of Chemical Theory and Computation, 17(2):599–604, 2021.

[41]Andreas Vitalis and Rohit V Pappu. Absinth: a new continuum solvation model for simulations of polypeptides in aqueous solutions. Journal of computational chemistry, 30(5):673–699, 2009.

[42]Hao Wu, Peter G Wolynes, and Garegin A Papoian. Awsem-idp: a coarse-grained force field for intrinsically disordered proteins. The Journal of Physical Chemistry B, 122(49):11115–11125, 2018.

[43]Upayan Baul, Debayan Chakraborty, Mauro L Mugnai, John E Straub, and D Thirumalai. Sequence effects on size, shape, and structural heterogeneity in intrinsically disordered proteins. The journal of physical chemistry. B, 123(16):3462—3474, 2019.

[44]Gregory L Dignon, Wenwei Zheng, Young C Kim, Robert B Best, and Jeetain Mittal. Sequence determinants of protein phase behavior from a coarse-grained model. PLoS computational biology, 14(1):e1005941, 2018.

[45]R. B. Best, W. Zheng, and J. Mittal. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. J. Chem. Theory Comput., 10:5113, 2014.

[46]P. Robustelli, S. Piana, and D. E. Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. Proc. Natl. Acad. Sci. U. S. A., 115:E4758, 2018.

[47]Gül H Zerze, Wenwei Zheng, Robert B Best, and Jeetain Mittal. Evolution of all-atom protein force fields to improve local and global properties. The journal of physical chemistry letters, 10(9):2227–2234, 2019.

[48]Wai Shing Tang, Nicolas L Fawzi, and Jeetain Mittal. Refining all-atom protein force fields for polar-rich, prion-like, low-complexity intrinsically disordered proteins. The Journal of Physical Chemistry B, 124(43):9505–9512, 2020.

[49]Massimiliano Bonomi, Carlo Camilloni, Andrea Cavalli, and Michele Vendruscolo. Metainference: A bayesian inference method for heterogeneous systems. Science Advances, 2(1):e1501177, 2016.

[50]Sandro Bottaro, Tone Bengtsen, and Kresten Lindorff-Larsen. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach, pages 219–240. Springer US, New York, NY, 2020.

[51]Lim Heo Giacomo Janson, Gilberto Valdes-Garcia. Direct generation of protein conformational ensembles via machine learning. BioRxiv.

[52]Pu Liu, Byungchan Kim, Richard A Friesner, and BJ Berne. Replica exchange with solute tempering: A method for sampling biological systems in explicit water. Proceedings of the National Academy of Sciences, 102(39):13749–13754, 2005.

[53]Lingle Wang, Richard A Friesner, and BJ Berne. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (rest2). The Journal of Physical Chemistry B, 115(30):9431–9438, 2011.

[54]Utsab Shrestha, Puneet Juneja, Qiu Zhang, Viswanathan Gurumoorthy, Jose Borreguero, Volker Urban, Xiaolin Cheng, Sai Venkatesh Pingali, Jeremy Smith, Hugh O'Neill, and

Loukas Petridis. Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. Proceedings of the National Academy of Sciences, 116:201907251, 09 2019.

[55]Utsab Shrestha, Jeremy Smith, and Loukas Petridis. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. Communications Biology, 4:243, 02 2021.

[56]Rajeswari Appadurai, Jayashree Nagesh, and Anand Srivastava. High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. Nature communications, 12(1):1–11, 2021.

[57]M. Bonomi, G. T. Heller, C. Camilloni, and M. Vendruscolo. Principles of protein structural ensemble determination. Curr. Opin. Struct. Biol., 42:106, 2017.

[58]Lei Yu and Rafael Brüschweiler. Quantitative prediction of ensemble dynamics, shapes and contact propensities of intrinsically disordered proteins. PLOS Computational Biology, 18(9):1–26, 09 2022.

[59]Stefano Gianni, María Inés Freiberger, Per Jemth, Diego U. Ferreiro, Peter G. Wolynes, and Monika Fuxreiter. Fuzziness and frustration in the energy landscape of protein folding, function, and assembly. Accounts of Chemical Research, 54(5):1251–1259, 2021. PMID: 33550810.

[60]Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25):14863–14868, 1998.

[61]Uri Alon, Naama Barkai, Daniel A Notterman, Kurt Gish, Suzanne Ybarra, Daniel Mack, and Arnold J Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 96(12):6745–6750, 1999.

[62]CA Floudas, HK Fung, SR McAllister, M Mönnigmann, and R Rajgaria. Advances in protein structure prediction and de novo protein design: A review. Chemical Engineering Science, 61(3):966–988, 2006.

[63]Robin Pearce and Yang Zhang. Deep learning techniques have significantly impacted protein structure prediction and protein design. Current opinion in structural biology, 68:194–207, 2021.

[64]Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. Molecular informatics, 29(6-7):476–488, 2010.

[65]M Michael Gromiha, K Yugandhar, , and Sherlyn Jemimah. Protein–protein interactions: scoring schemes and binding affinity. Current opinion in structural biology, 44:31–38, 2017.

[66]Jianyin Shao, Stephen W Tanner, Nephi Thompson, and Thomas E Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. Journal of chemical theory and computation, 3(6):2312–2334, 2007.

[67]Mojie Duan, Jue Fan, Minghai Li, Li Han, and Shuanghong Huo. Evaluation of dimensionality-reduction methods from peptide folding–unfolding simulations. Journal of Chemical Theory and Computation, 9(5):2490–2497, 2013.

[68]Florian Sittel and Gerhard Stock. Perspective: Identification of collective variables and metastable states of protein dynamics. The Journal of Chemical Physics, 149(15):150901, 2018.

[69]Gareth Tribello and Piero Gasparotto. Using dimensionality reduction to analyze protein trajectoriesdata$_s heet_1 .pdf. \underline{FrontiersinMolecularBiosciences}, 6 : 1 - -11, 062019$.

[70]Jane R. Allison. Computational methods for exploring protein conformations. Biochemical Society Transactions, 48(4):1707–1724, 08 2020.

[71]Francesco Trozzi, Xinlei Wang, and Peng Tao. Umap as a dimensionality reduction tool for molecular dynamics simulations of biomacromolecules: A comparison study. The Journal of Physical Chemistry B, 125(19):5022–5034, 2021.

[72]Heidi Klem, Glen M. Hocky, and Martin McCullagh. Size-and-shape space gaussian mixture models for structural clustering of molecular dynamics trajectories. Journal of Chemical Theory and Computation, 18(5):3218–3230, 2022.

[73]L. van der Maaten and G. Hinton. Visualizing data using t-sne. J. Mach. Learn. Res., 2008.

[74]Karthik Shekhar, Sylvain W Lapan, Irene E Whitney, Nicholas M Tran, Evan Z Macosko, Monika Kowalczyk, Xian Adiconis, Joshua Z Levin, James Nemesh, Melissa Goldman, et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. Cell, 166(5):1308–1323, 2016.

[75]Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. Nature communications, 10(1):1–14, 2019.

[76]George C Linderman and Stefan Steinerberger. Clustering with t-sne, provably. SIAM Journal on Mathematics of Data Science, 1(2):313–332, 2019.

[77]Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. Cell, 172(5):1091–1107, 2018.

[78]Walid M Abdelmoula, Benjamin Balluff, Sonja Englert, Jouke Dijkstra, Marcel JT Reinders, Axel Walch, Liam A McDonnell, and Boudewijn PF Lelieveldt. Data-driven identification of prognostic tumor subpopulations using spatially mapped t-sne of mass spectrometry imaging data. Proceedings of the National Academy of Sciences, 113(43):12244–12249, 2016.

[79]Qing Zhang, Wenlong Zhang, Tingsheng Lin, Wenfeng Lu, Xin He, Yuanzhen Ding, Wei Chen, Wenli Diao, Meng Ding, Pingping Shen, et al. Mass cytometry reveals immune atlas of urothelial carcinoma. BMC cancer, 22(1):1–13, 2022.

[80]Michael Thomas Wong, David Eng Hui Ong, Frances Sheau Huei Lim, Karen Wei Weng Teng, Naomi McGovern, Sriram Narayanan, Wen Qi Ho, Daniela Cerny, Henry Kun Kiaang Tan, Rosslyn Anicete, et al. A high-dimensional atlas of human t cell diversity reveals tissue-specific trafficking and cytokine signatures. Immunity, 45(2):442–456, 2016.

[81]Jakub Rydzewski and Wieslaw Nowak. Machine learning based dimensionality reduction facilitates ligand diffusion paths assessment: a case of cytochrome p450cam. Journal of Chemical Theory and Computation, 12(4):2110–2120, 2016.

[82]Oliver Fleetwood, Jens Carlsson, and Lucie Delemotte. Identification of ligand-specific g protein-coupled receptor states and prediction of downstream efficacy via data-driven modeling. 10:e60715, jan 2021.

[83]Arthur Voronin and Alexander Schug. Selection of representative structures from large biomolecular ensembles. The Journal of Chemical Physics, 156(14):144102, 2022.

[84]Vojtěch Spiwok and Pavel Kříž. Time-lagged t-distributed stochastic neighbor embedding (t-sne) of molecular simulation trajectories. Frontiers in molecular biosciences, 7:132, 2020.

[85]Anita Rácz, Levente M Mihalovits, Dávid Bajusz, Károly Héberger, and Ramón Alain Miranda-Quintana. Molecular dynamics simulations and diversity selection by extended continuous similarity indices. Journal of Chemical Information and Modeling, 62(14):3415–3425, 2022.

[86]Hongyu Zhou, Feng Wang, and Peng Tao. t-distributed stochastic neighbor embedding method with the least information loss for macromolecular simulations. Journal of chemical theory and computation, 14(11):5499–5510, 2018.

[87]Maria V Yelshanskaya, Dhilon S Patel, Christopher M Kottke, Maria G Kurnikova, and Alexander I Sobolevsky. Opening of glutamate receptor channel to subconductance levels. Nature, 605(7908):172–178, 2022.

[88]Oscar Palomino-Hernandez, Carlo Santambrogio, Giulia Rossetti, Claudio O. Fernandez, Rita Grandori, and Paolo Carloni. Molecular dynamics-assisted interpretation of experimentally determined intrinsically disordered protein conformational components: The case of human -synuclein. The Journal of Physical Chemistry B, 126(20):3632–3639, 2022.

[89]Gabriella T. Heller, Francesco A. Aprile, Thomas C. T. Michaels, Ryan Limbocker, Michele Perni, Francesco Simone Ruggeri, Benedetta Mannini, Thomas Löhr, Massimiliano Bonomi, Carlo Camilloni, Alfonso De Simone, Isabella C. Felli, Roberta Pierattelli, Tuomas P. J. Knowles, Christopher M. Dobson, and Michele Vendruscolo. Small-molecule sequestration of amyloid-β as a drug discovery strategy for alzheimer’s disease. Science Advances, 6(45):eabb5924, 2020.

[90]Paul Robustelli, Alain Ibanez-de Opakua, Cecily Campbell-Bezat, Fabrizio Giordanetto, Stefan Becker, Markus Zweckstetter, Albert C. Pan, and David E. Shaw. Molecular basis of small-molecule binding to -synuclein. Journal of the American Chemical Society, 144(6):2501–2510, 2022.

[91]P. Robustelli, S. Piana, and D. E. Shaw. Mechanism of coupled folding-upon-binding of an intrinsically disordered protein. J. Am. Chem. Soc., 142:11092, 2020.

[92]Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65, 1987.

[93]J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case. Development and testing of a general amber force field. J. Comput. Chem., 25:1157, 2004.

[94]Giulio Rastelli, Alberto Del Rio, Gianluca Degliesposti, and Miriam Sgobba. Fast and accurate predictions of binding free energies using mm-pbsa and mm-gbsa. Journal of computational chemistry, 31:797–810, 11 2009.

[95]Sangeetha Balasubramanian, Shovamayee Maharana, and Anand Srivastava. Interplay of the folded domain and disordered low-complexity domains along with rna sequence mediate

efficient binding of fus with rna. 2022.

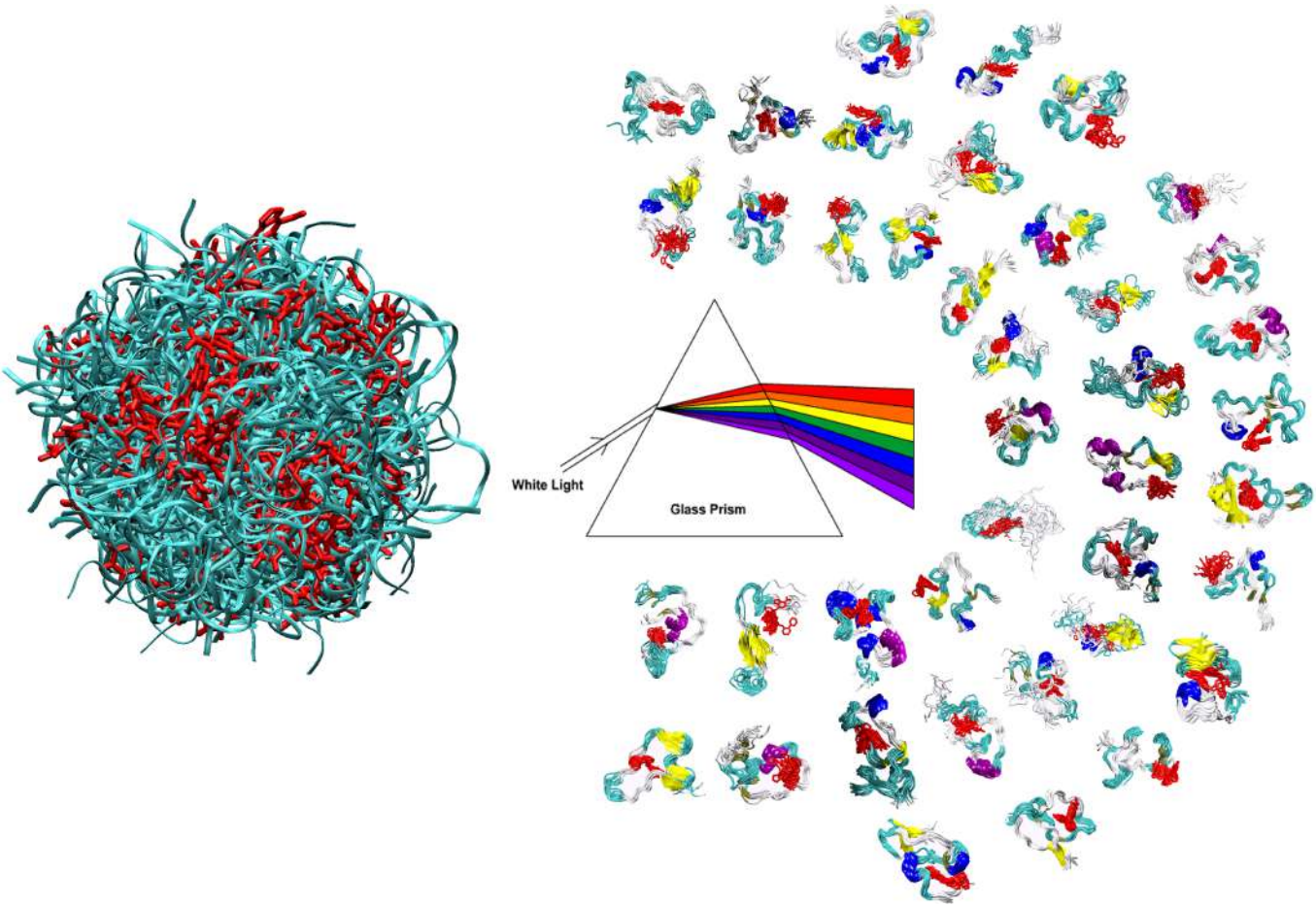
[96]Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively, 2016.

[97]Anna C Belkina, Christopher O Ciccolella, Rina Anno, Richard Halpert, Josef Spidlen, and Jennifer E Snyder-Cappione. Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. Nature communications, 10(1):1–12, 2019.

[98]Nicola Pezzotti, Thomas Höllt, B Lelieveldt, Elmar Eisemann, and Anna Vilanova. Hierarchical stochastic neighbor embedding. 35(3):21–30, 2016.

[99]Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. The Journal of Machine Learning Research, 15(1):3221–3245, 2014.

[100]Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In Artificial intelligence and statistics, pages 384–391. PMLR, 2009.



Graphical Abstract

Supplemental Material

Clustering conformational ensembles of intrinsically disordered proteins with t-distributed stochastic neighbor embedding

Rajeswari Appadurai,[1] Jaya Krishna,[2] Massimiliano Bonomi,[3] Paul Robustelli,[2] and Anand Srivastava[1, a)]

[1)] *Molecular Biophysics Unit, Indian Institute of Science Bangalore, C. V. Raman Road, Bangalore, Karnataka 560012, India*

[2)] *Dartmouth College, Department of Chemistry, Hanover, NH, 03755, USA*

[3)] *Structural Bioinformatics Unit, Department of Structural Biology and Chemistry. CNRS UMR 3528, C3BI, CNRS USR 3756, Institut Pasteur, Paris, France*

[a)] Electronic mail: anand@iisc.ac.in

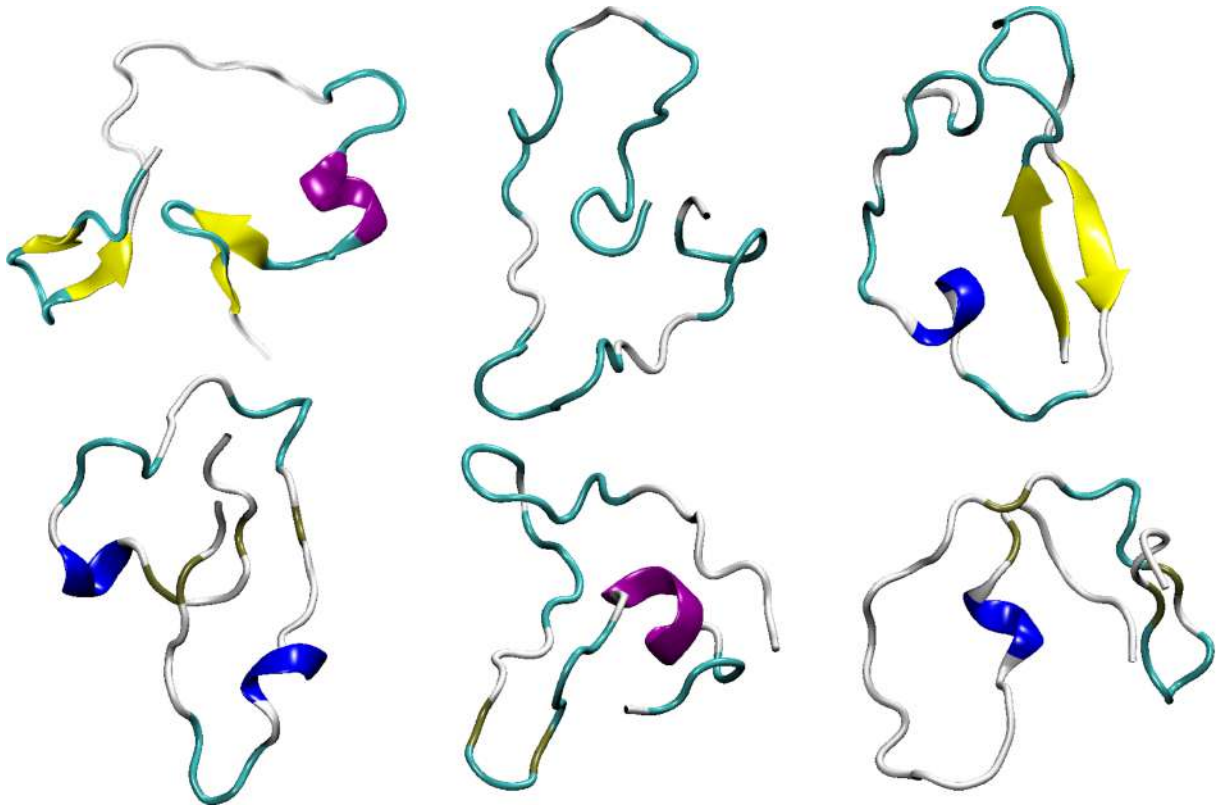


FIG. S1: Snapshots of Aβ42, showing to have same Rg values (1.10 nm) but have different structural organization altogether

A. Physical intuition into tSNE-based clustering algorithm using alanine dipeptide

We first employ the t-SNE method on the alanine dipeptide (ADP) trajectory where we compare the results with the well-known 2D Ramachandran plot using dihedral distance as dissimilarity score. Ramachandran plot is also called $\phi - \psi map$ due to the backbone dihedral angles along the peptide bond[1,2]. Here, we can fix the number of clusters to four (K = 4) based on the four known sub-regions of the Ramachandran plot namely beta-sheet, PPII, right-handed α helix, and left-handed α helix (Fig. S2(a)). The left-handed α helix region lies separately in the second half of the ϕ dihedral axis whereas the beta-sheet and the right-handed α helical regions occupy the first half of the ϕ dihedral axis. To quantify the goodness of clustering, we calculate the Silhouette score[3] on the raw data and arrive at a score of 0.55 for the 2D map. Of note, when the numbers of clusters are not known a priory unlike the ADP system, we have a prescription that makes use of silhouette score with the t-SNE perplexity values to find the optimum number of clusters.

The second feature of t-SNE is a tuneable parameter called "perplexity," which dictates (loosely) how to balance attention between local and global aspects of your data. The parameter is, in a sense, a guess about the number of close neighbors each point has. The perplexity value has a complex effect on the resulting pictures. The original paper says, "The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50." But the story is more nuanced than that. Getting the most from t-SNE may mean analyzing multiple plots with different perplexities.

In the Ramachandran plot, low dimensional projection of data along any one of the projections (ϕ or ψ), yields overlap of different conformations onto each other. We show this at the bottom and left of Fig. S2 (a) for projection along ϕ and ψ, respectively. PCA, the most common dimensional reduction method, fails to achieve clear separation and has a very low Silhouette score of 0.154 (see Fig. S2(b)). This is because PCA tries to linearly transform the data along an axis of maximal variation, which is the ψ axis in the Ramachandran plot, and hence cannot capture the distinction between L-helix and other conformers. On the other hand, the t-SNE projections provide more faithful representations of the clusters. In Fig. S2 (c), we plot the t-SNE projections for a range of perplexity values. For a certain

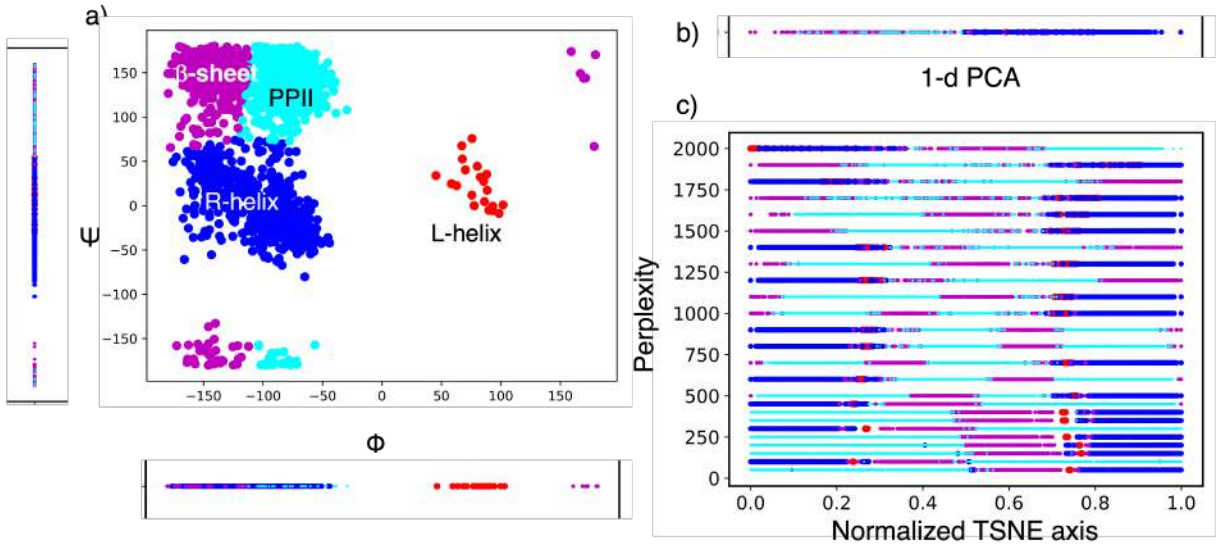


FIG. S2: Dimensional reduction of Ramachandran map of alanine di-peptide: a) The 2D Ramachandran map color-coded by the subregions. Simple 1-dimensional projections along the X and Y axes are shown. Results of 1D transformation of RC-map using b) PCA and c) TSNE with different perplexities.

perplexity value ($Perp$ =400), t-SNE clearly separates out the 4 sub-regions as in the original space with a much improved Silhouette score of 0.50. At low perplexities, t-SNE focuses on the local variations and tries to preserve the closest neighbors as much as possible in the original space. However, very low perplexity yields too many clusters with single or very few conformations per cluster, which nullifies the advantages of clustering in the first place. On the other hand, t-SNE essentially degrades to PCA at very high perplexities and leads to overcrowding as greater variations are tolerated at high perplexity scores. With perplexity as a tuneable parameter to balance the degree of local preservation on one hand and minimize the overcrowding on the other, t-SNE offers an exciting possibility to meaningfully cluster and visualize complex and heterogeneous high-dimensional IDPs datasets.

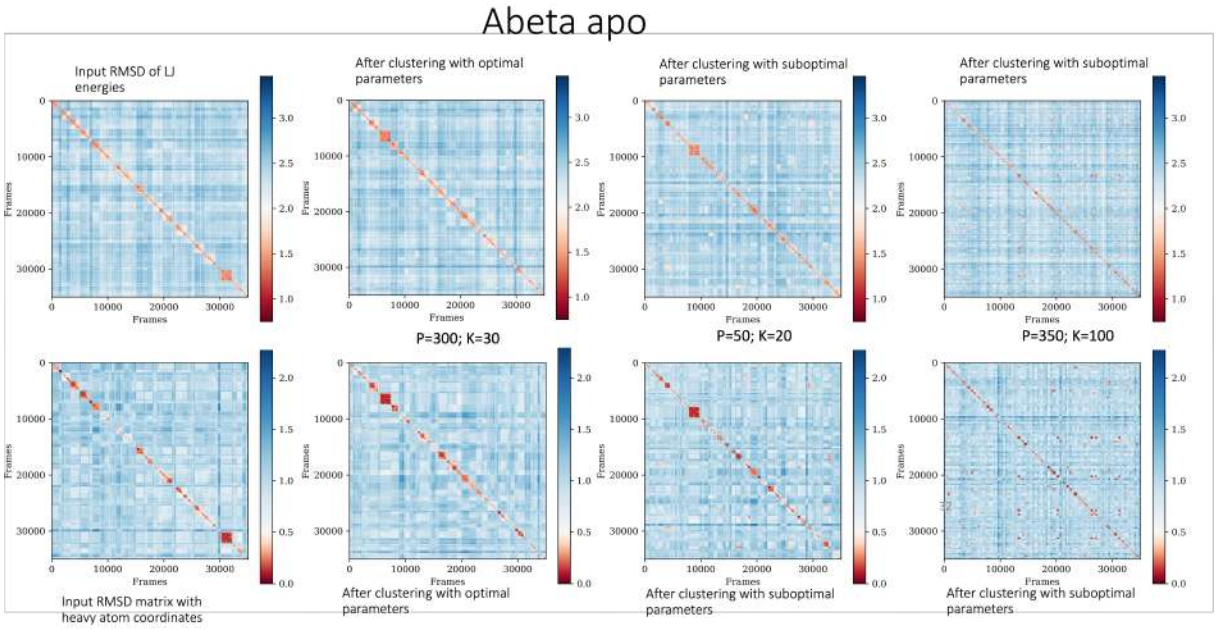


FIG. S3: Pairwise RMSD map of apo Aβ-42 conformations: Row-1 and Row-2 represents the RMSD map generated on the inter-residue LJ distances and on the heavy atom coordinates of conformations, respectively. The map generated on the raw trajectory (before clustering) is shown in a and e. The plots made after clustering is shown in b-d and f-h, where the frames are reordered according to the cluster indices (from the 0th cluster to the Nth cluster). The P and K values used for the clustering is indicated.

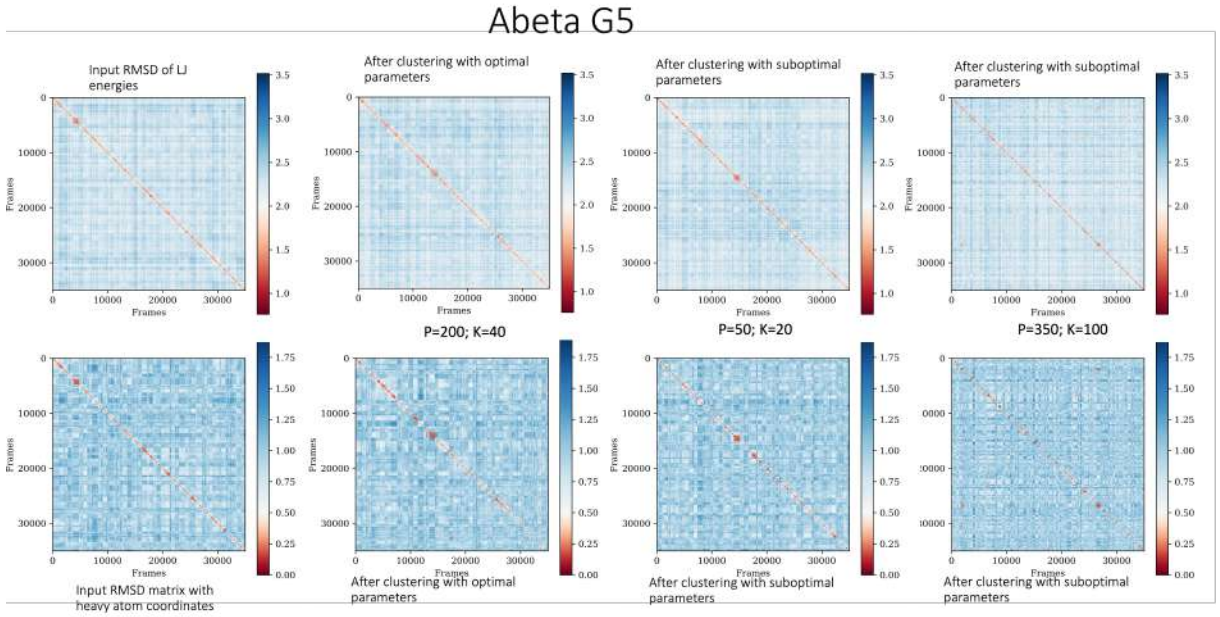


FIG. S4: Pairwise RMSD map of G5-bound Aβ-42 ensemble: Row-1 and Row-2 represents the RMSD map generated on the inter-residue LJ distances and on heavy atom coordinates of the conformations, respectively. The map generated on the raw trajectory (before clustering) is shown in a and e. The plots made after clustering is shown in b-d and f-h, where the frames are reordered according to the cluster indices (from the 0th cluster to the Nth cluster). The P and K values used for the clustering are indicated.

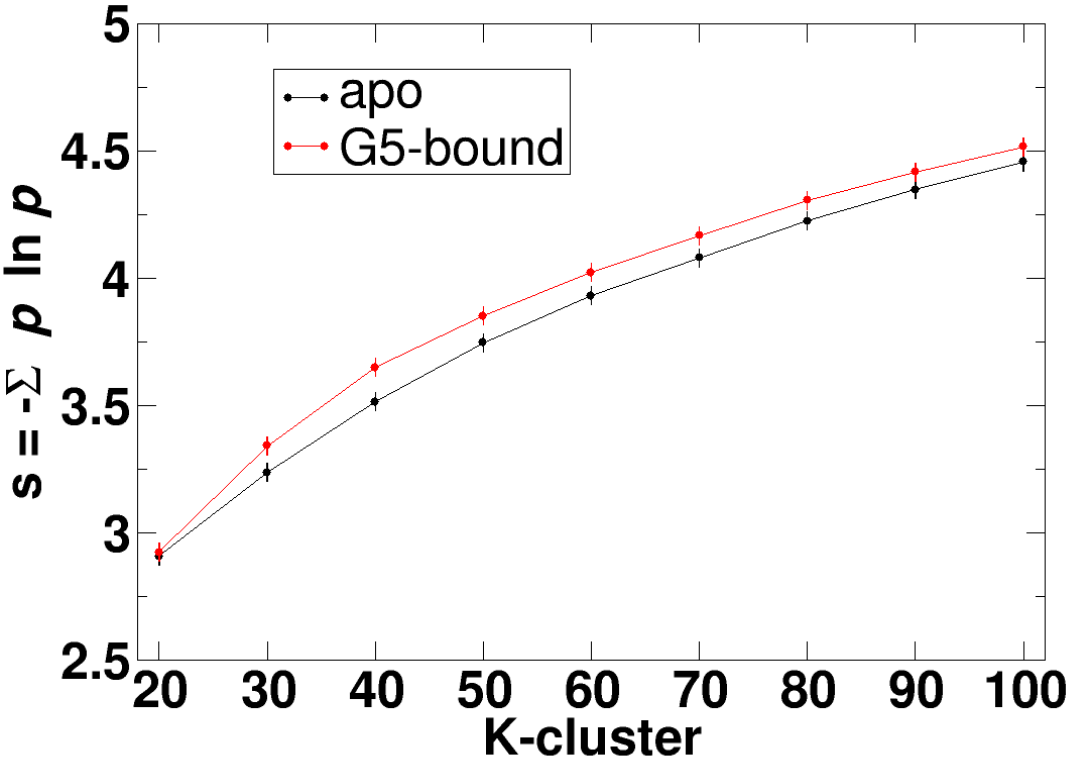


FIG. S5: Estimation of Gibbs conformational entropy ($S = -\sum(p_i \ln p_i)$) in the apo and G5-bound A*beta*-42 ensembles. p_i is the fractional occupancy of each cluster, weighted by the metadynamics weights obtained from[4].

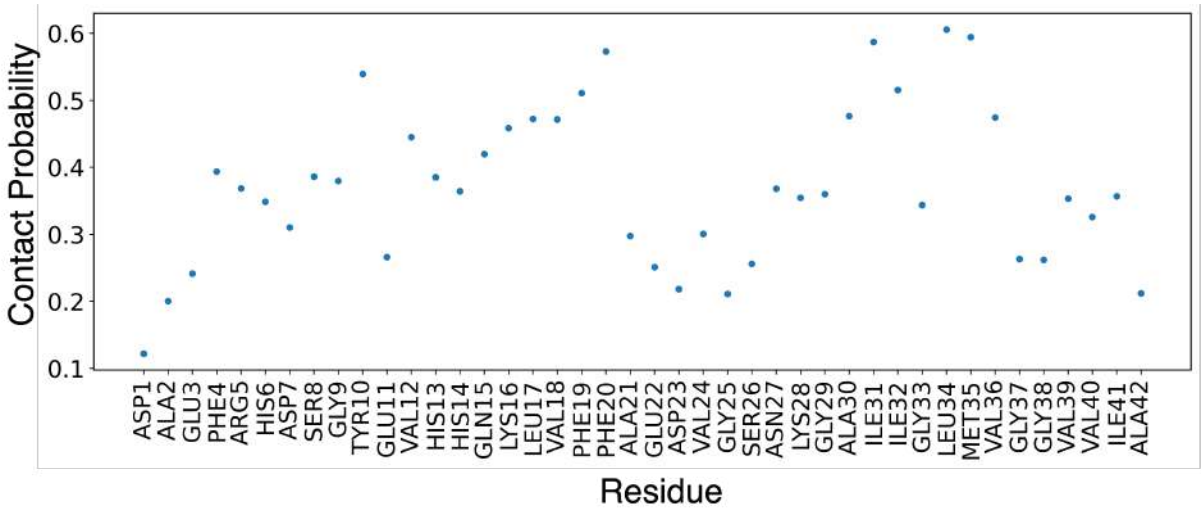


FIG. S6: Residue-wise Propensity of contacts made by Aβ42 with G5 calculated from the total trajectory.

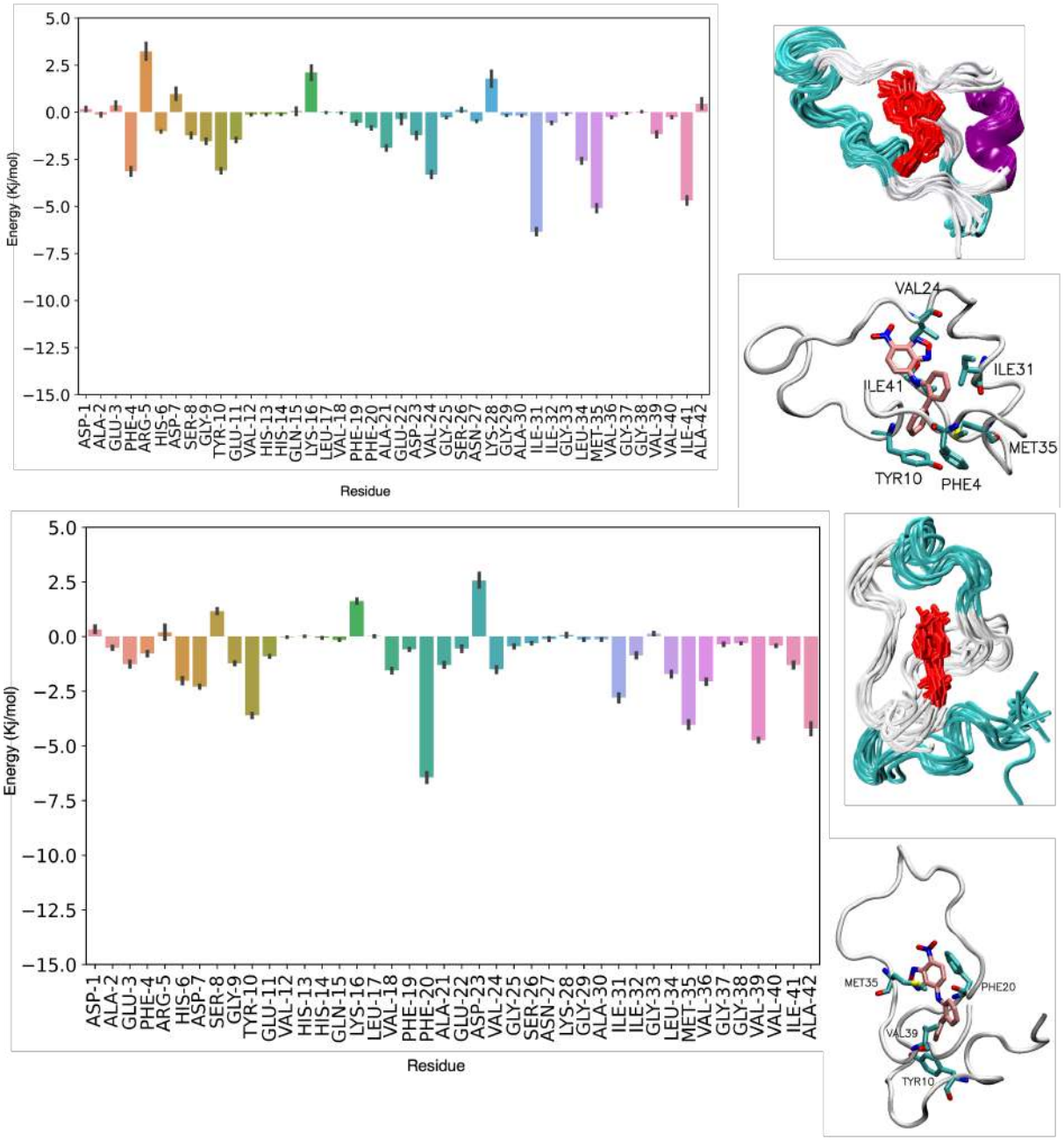
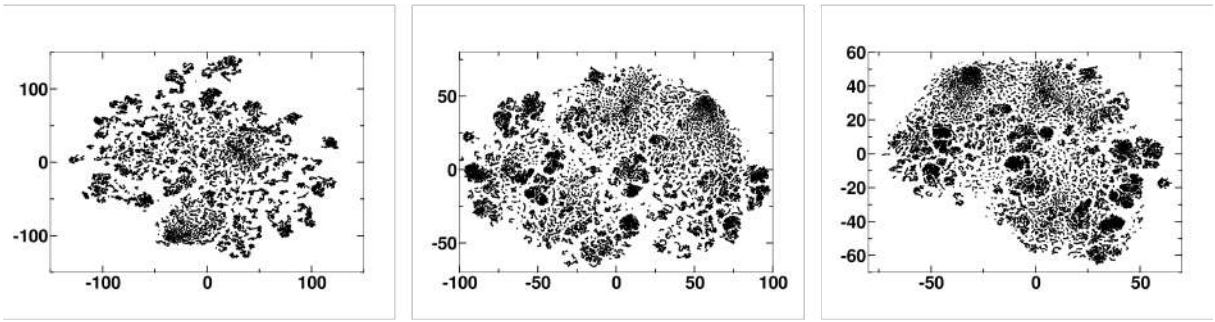
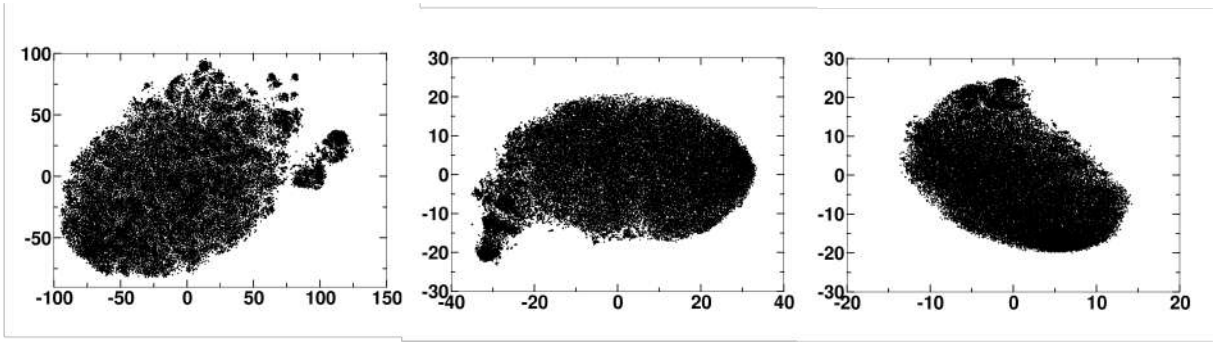


FIG. S7: Residue-wise decomposition of total binding energy from the other two favorable bound geometries (cluster29 in (a) and cluster 30 in (b))

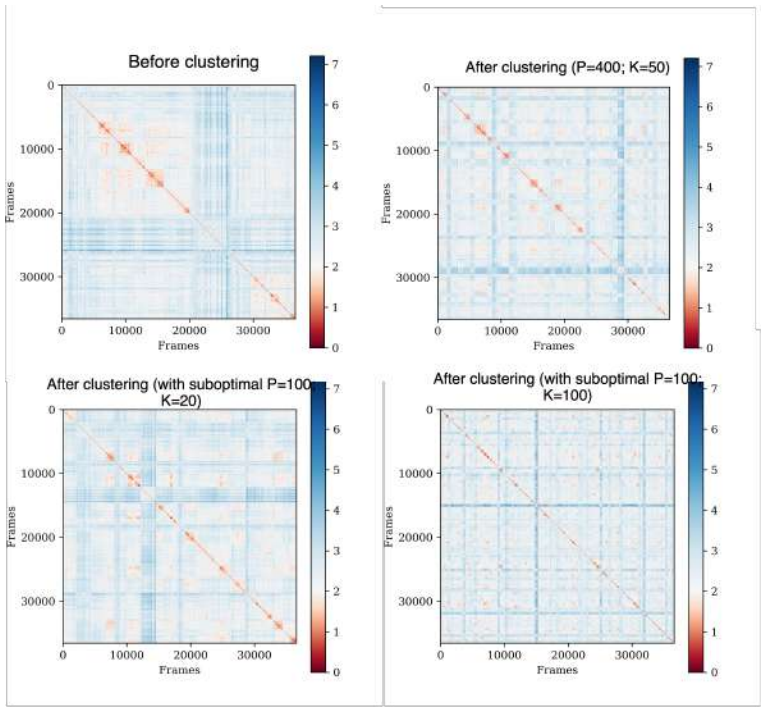


(a)

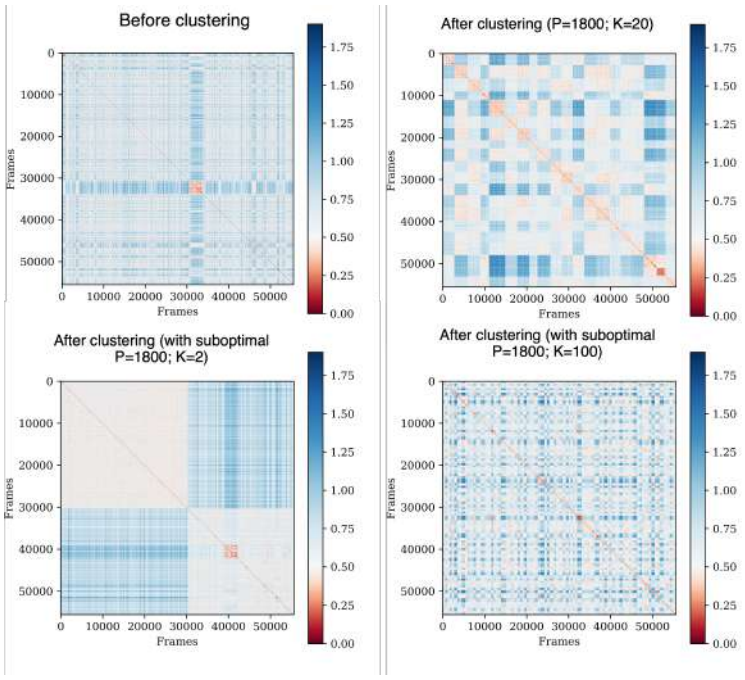


(b)

FIG. S8: (a) Representative t-SNE maps generated with three different perplexity values (100, 1000 and 2000) for the full length apo α-synuclein ensemble (b) Representative t-SNE maps generated with three different perplexity values (100, 1000 and 2000) for the apo c-terminus α-synuclein ensemble



(a)



(b)

FIG. S9: Pairwise RMSD map of (a) full-length apo and (b) c-terminusαS apo ensemble: The RMSD between each pair of conformation is measured based on the heavy atom coordinates. The map generated on the raw trajectory (before clustering) is shown in a. The plots made after clustering is shown in b-d, where the frames are reordered according to the cluster indices (from the 0th cluster to the Nth cluster). The P and K values used for the clustering are indicated.

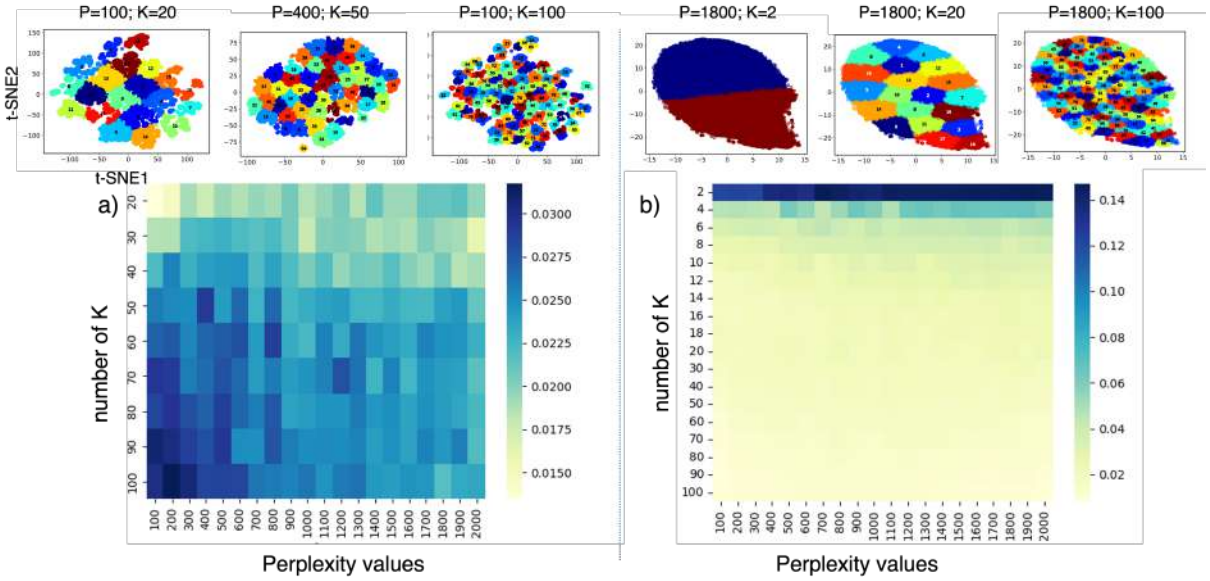
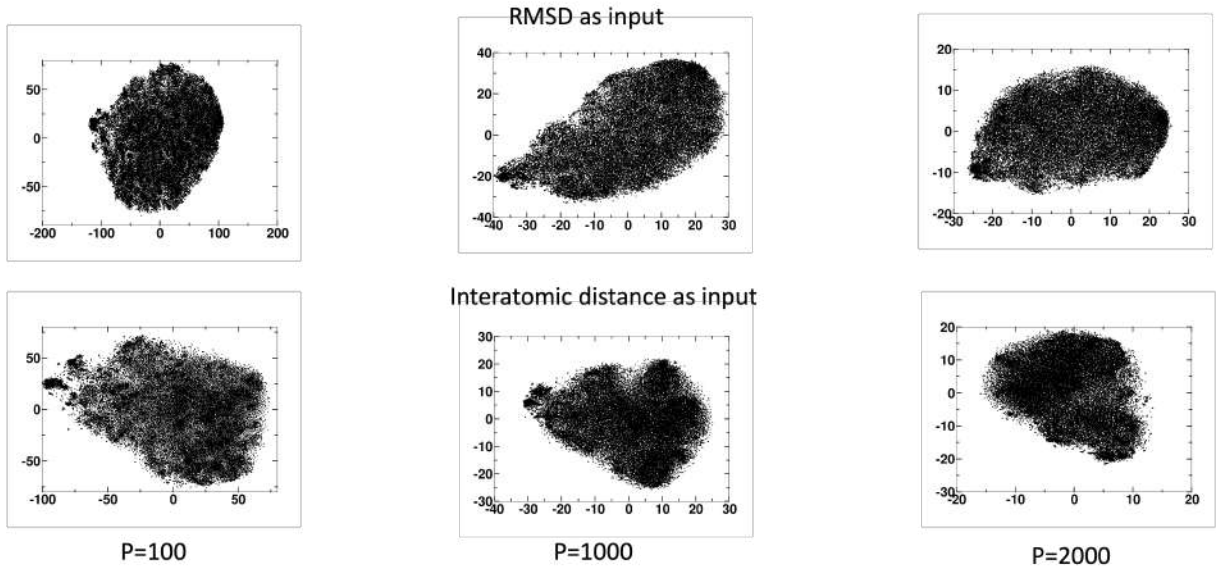
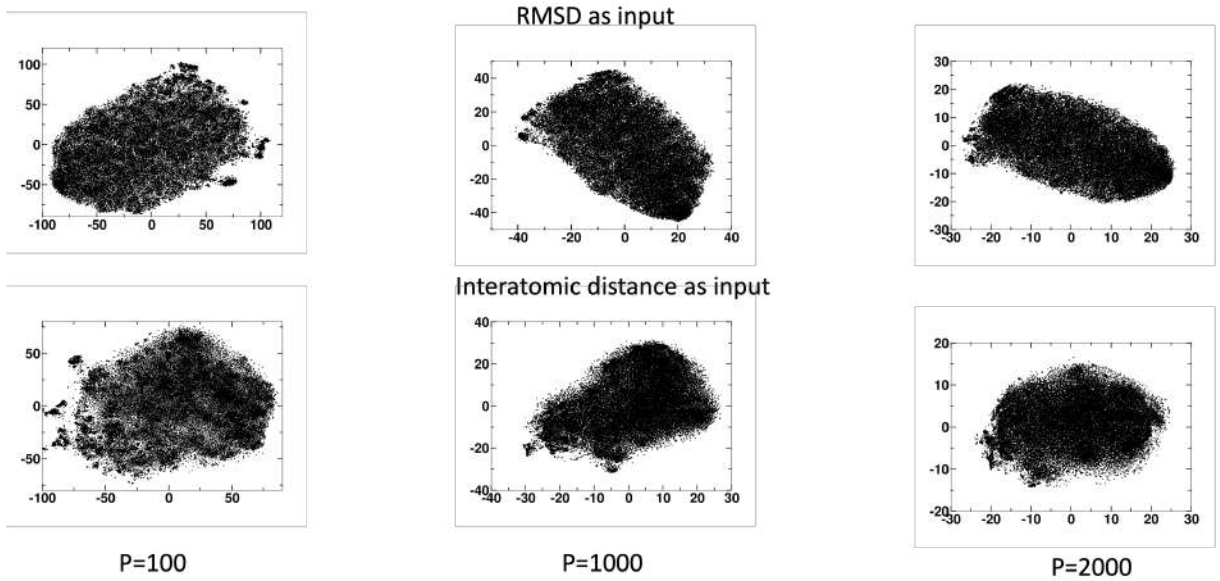


FIG. S10: Hyperparameter optimization based on integrated Silhouette score for the (a) apo full-length and (b) apo c-terminal α-synuclein ensembles.



(a)



(b)

FIG. S11: Representative t-SNE maps were generated with three different perplexity values (100, 1000, and 2000) for the (a) Fasudil-bound and (b) Lig47-bound c-terminus α-synuclein ensemble

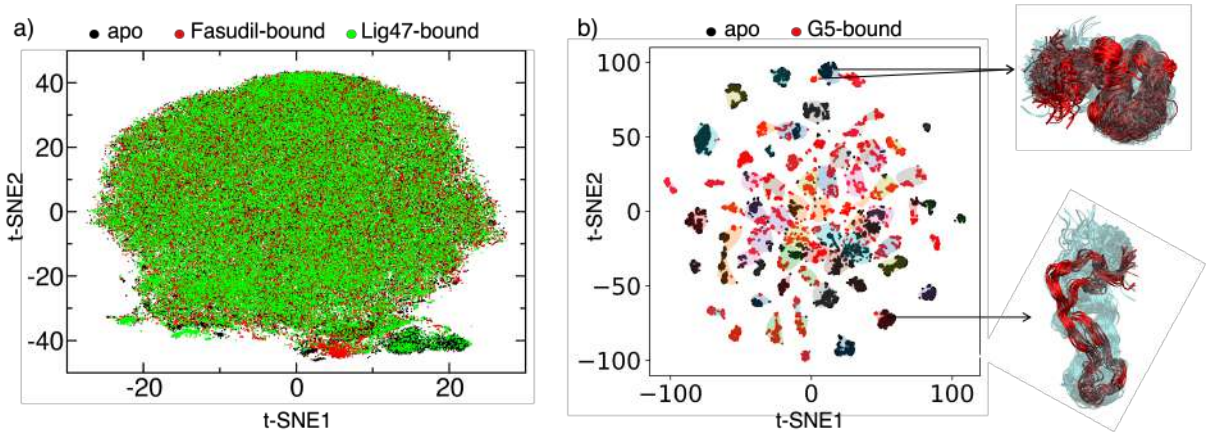


FIG. 12: a) TSNE map of collated ensembles of α-synuclein C-terminal peptide simulated in the presence and absence of Fasudil or ligand-47. The reduced coordinates of apo, Fasudil-bound, and Ligand-47 bound data are color coded in Black, Red, and Green respectively. b) t-SNE map of collated ensembles of Aβ-42 protein simulated in the presence (red dots) and absence of G5 (black dots). The map shows mostly non-overlapping clusters between the two ensembles except for a very few clusters that consist of conformations from both apo and bound ensembles. These similar and overlapped conformations from the two ensembles are rendered and superposed in the inset with cyan indicating apo and red indicating G5-bound Aβ-42 ensembles. We also mapped the conformational clusters that showed the highest binding affinity with G5 molecules (cluster numbers 14, 29, and 30) as predicted from Fig. 3 in the main text and indicate them on the collated map with blue arrows. It is evident that these high-affinity conformations from the G5-bound ensemble do not have closely related conformers in the APO state.

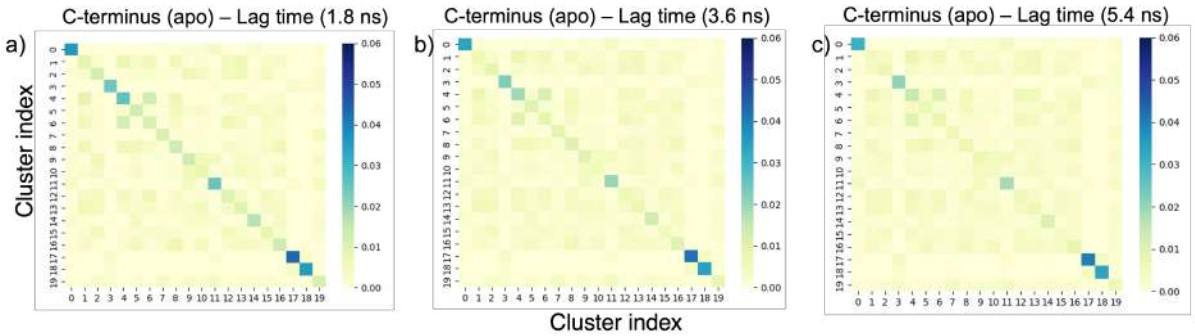


FIG. 13: Cluster-wise transition probability of apo α-synuclein C-terminal peptide conformations at different lag times - a) 1.8 ns b) 3.6 ns c) 5.4 ns.

TABLE S1: Mean Silhouette score for the apo Aβ42 ensemble calculated based on the distances in the low S_{ld} and high dimensional space S_{hd}. S_{hd} values are given in the bracket

	20	30	40	50	60	70	80	90	100
50	0.424	0.498	0.546	0.576	0.583	0.598	0.593	0.589	0.583
	(0.087)	(0.148)	(0.171)	(0.157)	(0.139)	(0.13)	(0.119)	(0.109)	(0.108)
100	0.505	0.586	0.64	0.642	0.633	0.611	0.598	0.584	0.582
	(0.141)	(0.177)	(0.189)	(0.163)	(0.139)	(0.125)	(0.116)	(0.11)	(0.107)
150	0.543	0.626	0.686	0.658	0.639	0.625	0.598	0.581	0.574
	(0.15)	(0.211)	(0.194)	(0.166)	(0.141)	(0.13)	(0.119)	(0.112)	(0.109)
200	0.579	0.666	0.711	0.665	0.62	0.616	0.599	0.581	0.558
	(0.166)	(0.197)	(0.195)	(0.164)	(0.133)	(0.13)	(0.124)	(0.116)	(0.107)
250	0.617	0.689	0.706	0.658	0.653	0.605	0.588	0.572	0.553
	(0.174)	(0.212)	(0.193)	(0.161)	(0.151)	(0.129)	(0.121)	(0.116)	(0.108)
300	0.638	0.709	0.712	0.66	0.617	0.594	0.569	0.57	0.545
	(0.178)	(0.211)	(0.193)	(0.164)	(0.144)	(0.132)	(0.117)	(0.116)	(0.107)
350	0.662	0.718	0.699	0.637	0.616	0.593	0.57	0.563	0.543
	(0.188)	(0.205)	(0.192)	(0.157)	(0.141)	(0.133)	(0.119)	(0.114)	(0.105)

TABLE S2: Mean Silhouette score for the G5-bound Aβ42 ensemble calculated based on the distances in the low S_{ld} and high dimensional space S_{hd}. S_{hd} values are given in the bracket

	20	30	40	50	60	70	80	90	100
50	0.418	0.464	0.513	0.543	0.554	0.577	0.578	0.567	0.584
	(0.057)	(0.097)	(0.131)	(0.132)	(0.13)	(0.127)	(0.127)	(0.116)	(0.123)
100	0.473	0.555	0.612	0.63	0.616	0.618	0.581	0.57	0.556
	(0.072)	(0.118)	(0.149)	(0.145)	(0.141)	(0.139)	(0.122)	(0.121)	(0.119)
150	0.487	0.587	0.652	0.645	0.627	0.604	0.577	0.553	0.554
	(0.082)	(0.126)	(0.148)	(0.145)	(0.14)	(0.14)	(0.128)	(0.124)	(0.128)
200	0.495	0.605	0.667	0.66	0.635	0.598	0.573	0.518	0.509
	(0.081)	(0.129)	(0.149)	(0.149)	(0.146)	(0.137)	(0.137)	(0.12)	(0.121)
250	0.508	0.615	0.662	0.643	0.61	0.604	0.545	0.516	0.491
	(0.089)	(0.13)	(0.149)	(0.148)	(0.146)	(0.144)	(0.128)	(0.125)	(0.121)
300	0.517	0.596	0.649	0.63	0.604	0.531	0.527	0.51	0.474
	(0.094)	(0.126)	(0.151)	(0.151)	(0.147)	(0.131)	(0.127)	(0.127)	(0.119)
350	0.509	0.591	0.633	0.622	0.592	0.522	0.492	0.489	0.48
	(0.093)	(0.129)	(0.146)	(0.15)	(0.151)	(0.124)	(0.125)	(0.124)	(0.12)

TABLE S3: Consistent local clustering in the apo Aβ42 trajectory upon different runs with different random initializations. We have noted that there is neither change in the choice of optimal perplexity and K values, nor the clustering pattern (measured by normalized mutual information score) upon rerunning.

S.No	S_{ld}	S_{hd}	Normalized mutual information
1	0.7086	0.2105	0.9755
2	0.7158	0.2102	

REFERENCES

[1]G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology, 7(1):95–99, 1963.

[2]Ashraya Ravikumar, Chandrasekharan Ramakrishnan, and Narayanaswamy Srinivasan. Stereochemical assessment of (,) outliers in protein structures using bond geometry-specific ramachandran steric-maps. Structure, 27(12):1875–1884.e2, 2019.

[3]Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20:53–65, 1987.

[4]Gabriella T. Heller, Francesco A. Aprile, Thomas C. T. Michaels, Ryan Limbocker, Michele Perni, Francesco Simone Ruggeri, Benedetta Mannini, Thomas Löhr, Massimiliano Bonomi, Carlo Camilloni, Alfonso De Simone, Isabella C. Felli, Roberta Pierattelli, Tuomas P. J. Knowles, Christopher M. Dobson, and Michele Vendruscolo. Small-molecule sequestration of amyloid-β as a drug discovery strategy for alzheimer’s disease. Science Advances, 6(45):eabb5924, 2020.