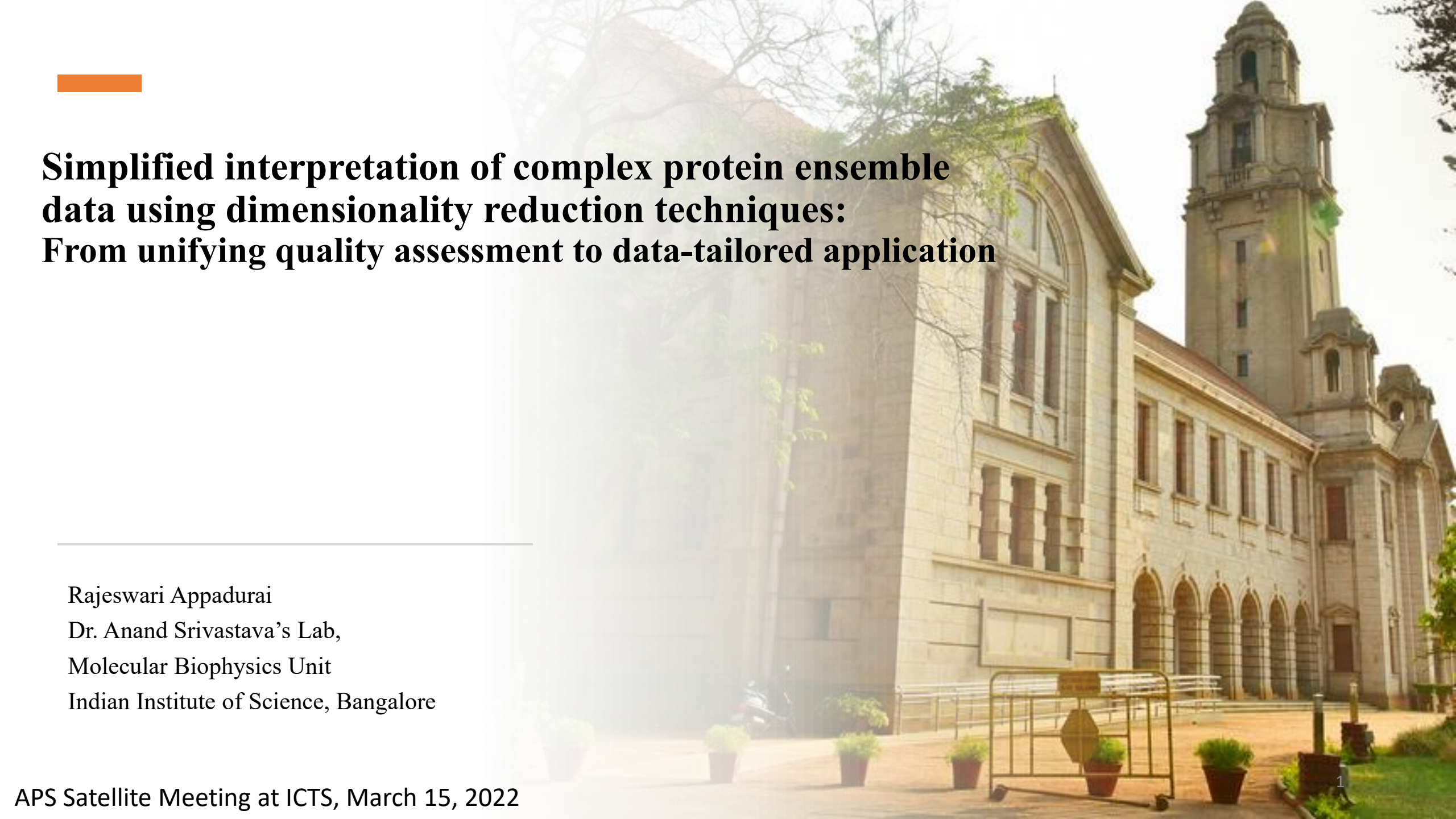
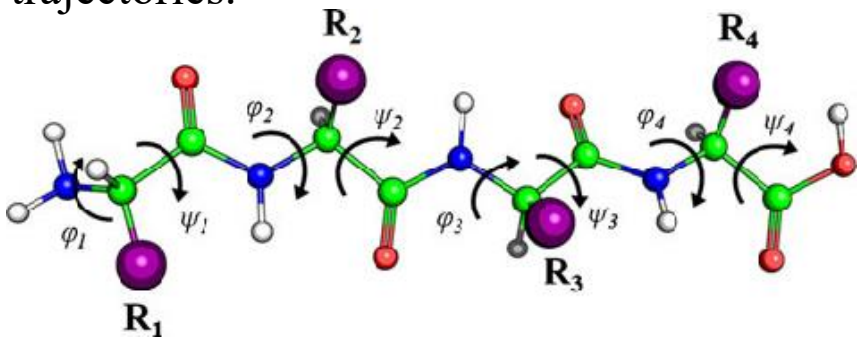
Simplified interpretation of complex protein ensemble data using dimensionality reduction techniques: From unifying quality assessment to data-tailored application



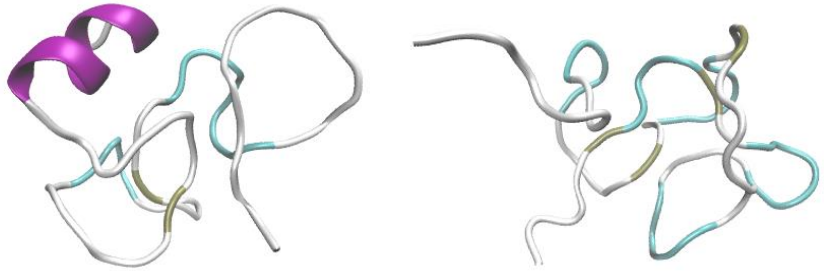
Rajeswari Appadurai
Dr. Anand Srivastava's Lab,
Molecular Biophysics Unit
Indian Institute of Science, Bangalore

Complexity in Protein structural ensemble

Proteins are inherently flexible and adopt a huge range of conformations - often explored using MD simulations as large trajectories.

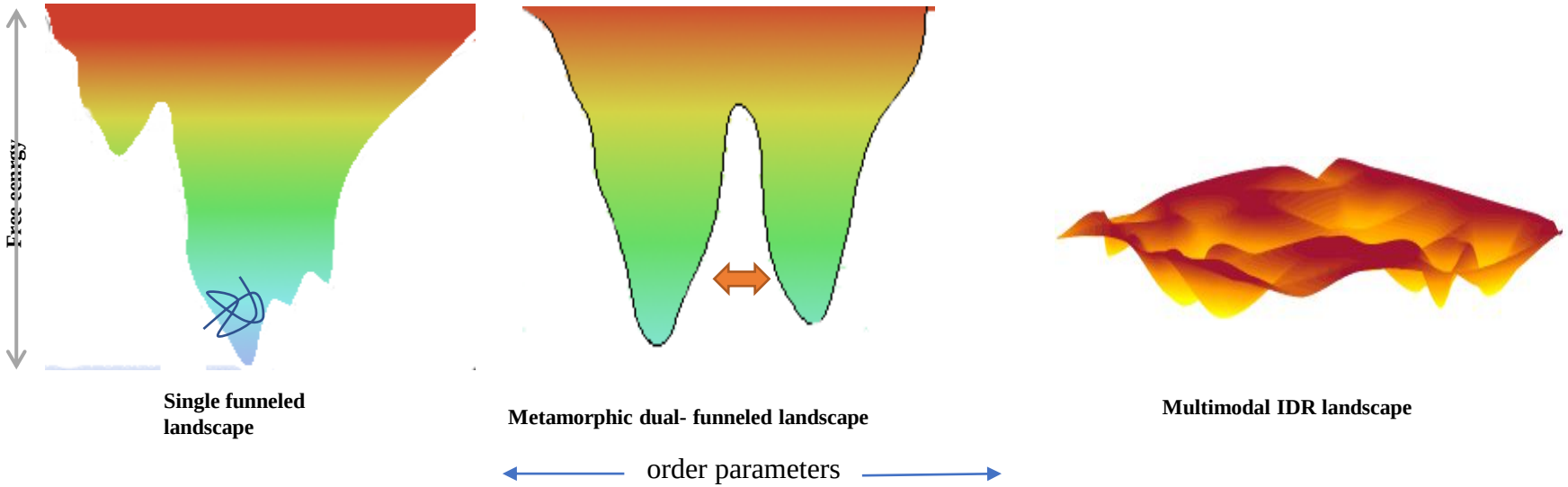


Use of classical order parameters



Rg value (1.01 nm)

In order to visualize the full landscapes, need low-d order parameters that faithfully captures the high-d data.



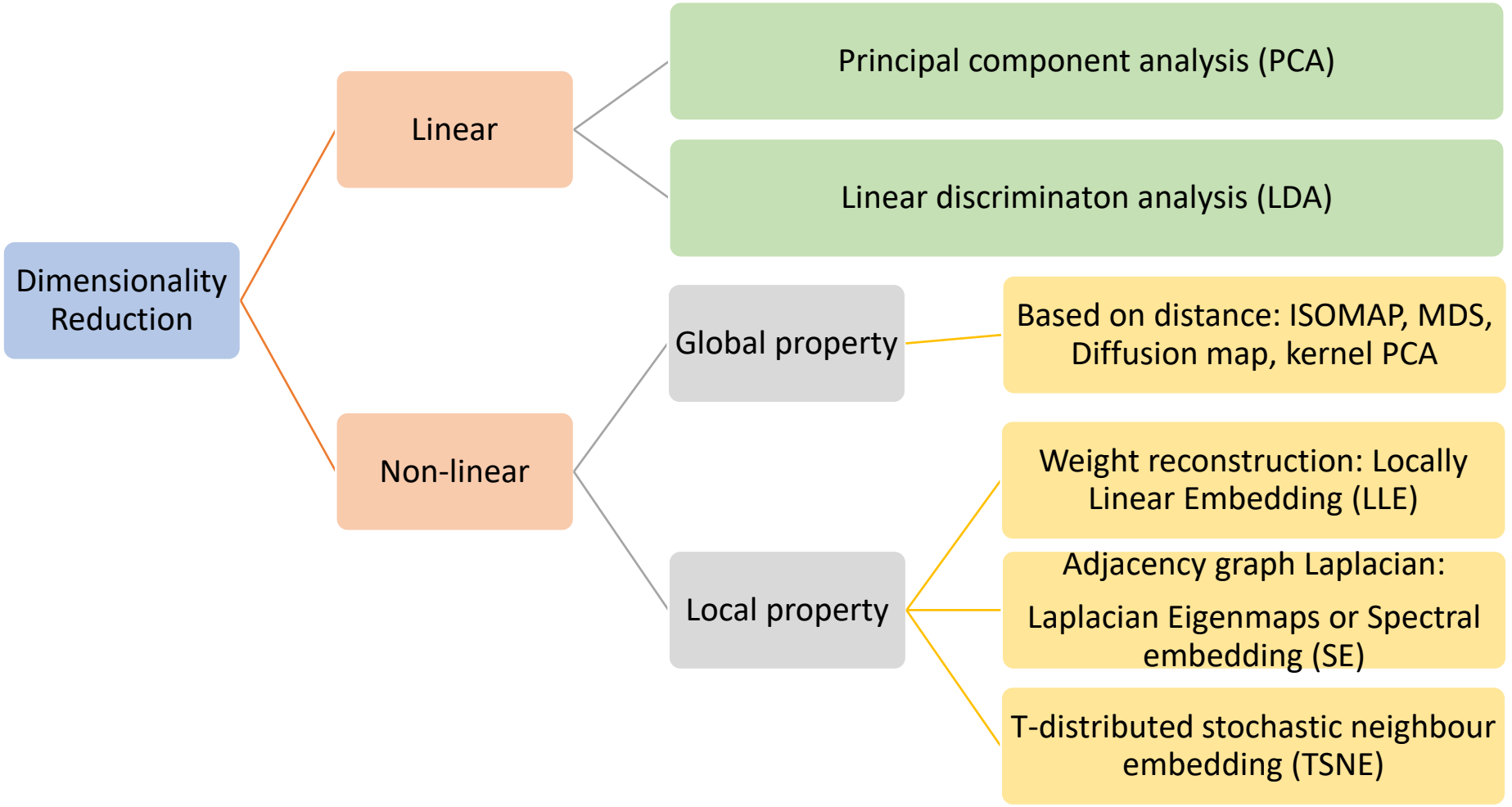
Single funneled landscape

Metamorphic dual- funneled landscape

Multimodal IDR landscape

Dimensionality reduction techniques for extracting useful low-d information

- Dimensionality reduction (DR) techniques aim to identify the underlying latent features.



?

Which method suits best for the given trajectory?

Quality framework based on Ranking

- Ranking:

$$\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \; OR \; \delta_{ik} = \delta_{ij} \; AND \; k < j\}|$$
$$r_{ij} = |\{k : d_{ik} < d_{ij} \; OR \; d_{ik} = d_{ij} \; AND \; k < j\}|$$

- Co-ranking Matrix **Q**:

$$q_{kl} = |\{(i,j) : \rho_{ij} = k \; AND \; r_{ij} = l\}|$$

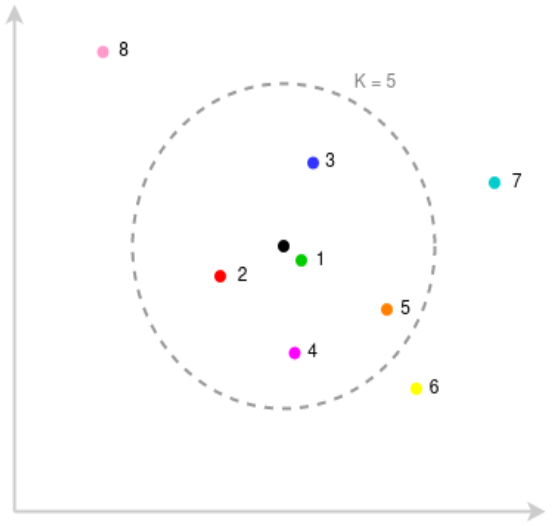
- Rank Error $= \rho_{ij} - r_{ij}$

- Evaluation Metrics Equations:

$$\mathbf{T(K)} = 1 - \frac{2}{G_K} \sum_{(k,l) \in LL_K} (k - K) q_{kl}$$

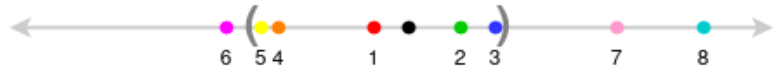
$$\mathbf{C(K)} = 1 - \frac{2}{G_K} \sum_{(k,l) \in UR_K} (l - K) q_{kl}$$

$$\mathbf{LCMC(K)} = \frac{K}{1 - N} + \frac{1}{NK} \sum_{(k,l) \in UL_K} q_{kl}$$

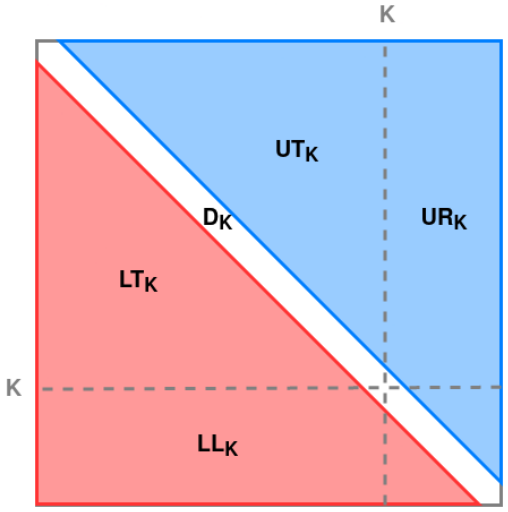


a. Original Space

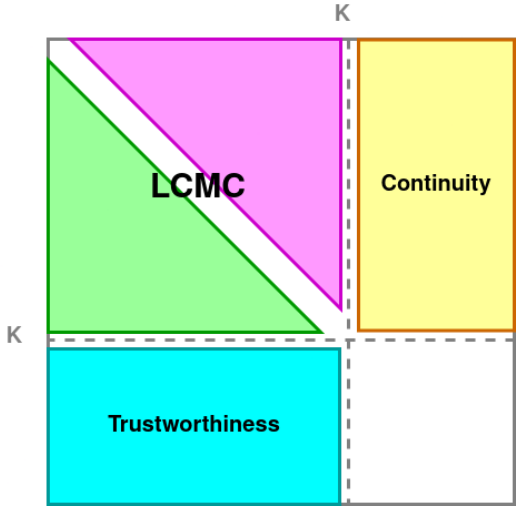
Neighbour	Original Space	Projection Space
●	1	2
●	2	1
●	3	3
●	4	6
●	5	4
●	6	5
●	7	8
●	8	7



b. Projection Space



d. Intrusion and Extrusion



c. Evaluation Metric Zones

Lee et al. Neurocomputing 2009, 72, 1431–1443.

Quality framework based on distance ratio

1. Evaluates distance proportionality

$$DRM(i,j) = \begin{cases} 1, & \text{for } d_R(i,j) \in (D_R(i,j) - \tau, D_R(i,j) + \tau) \\ 0, & \text{otherwise} \end{cases}$$

$$DR_\tau = \frac{\left(\sum_i \sum_j DRM_{(i,j)} - N\right)}{N(N-1)}$$

τ is the tolerance value

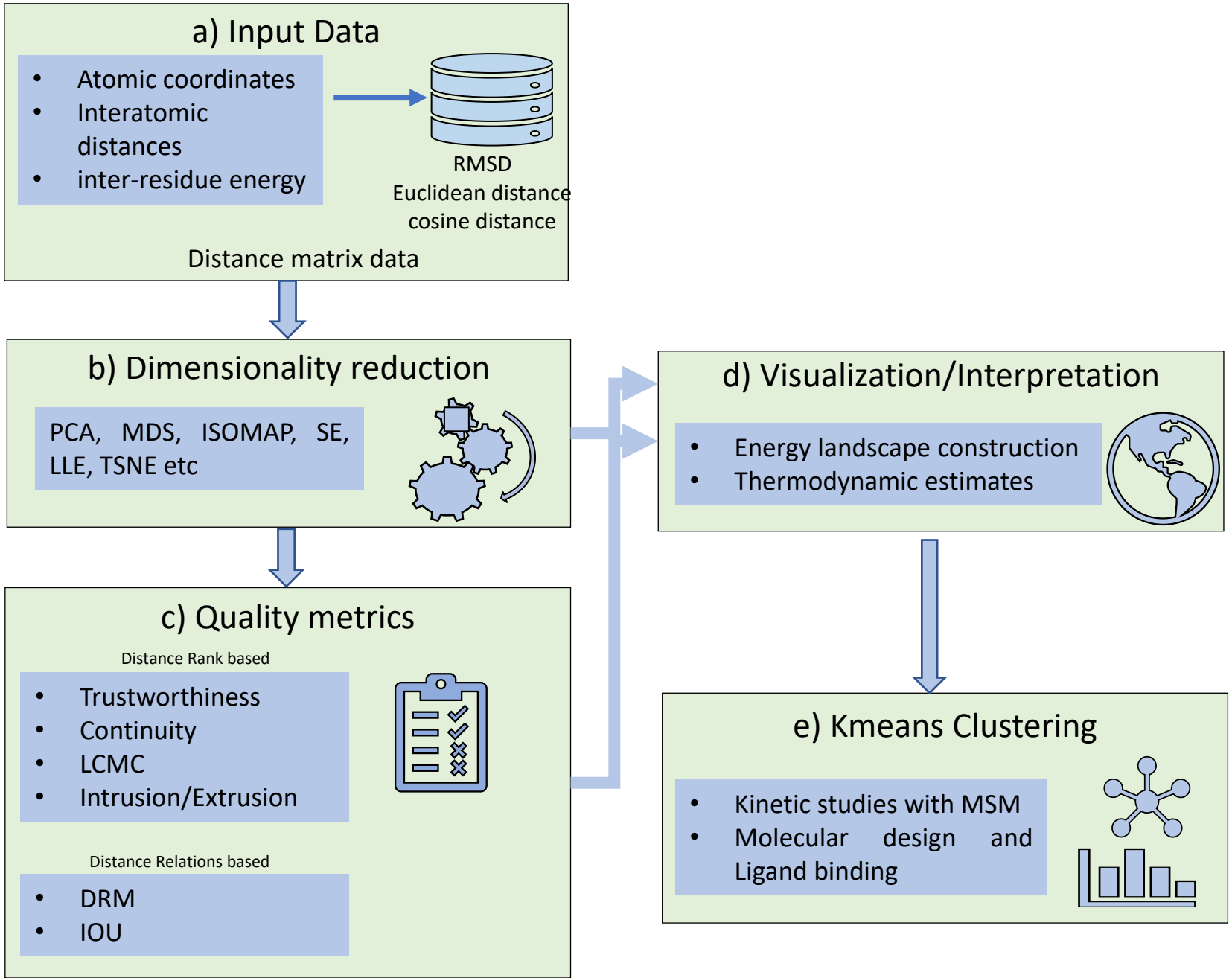
2. Evaluates neighbourhood preservation

$$IOU(x,\delta) = \frac{|n_x(\delta) \cap N_x\delta|}{|n_x(\delta) \cup N_x\delta|}$$

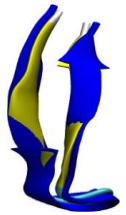
δ - Neighbourhood cut-off

Acknowledgement: Upal Bhattacharya, IISER Kolkata

Overview of the evaluation workflow

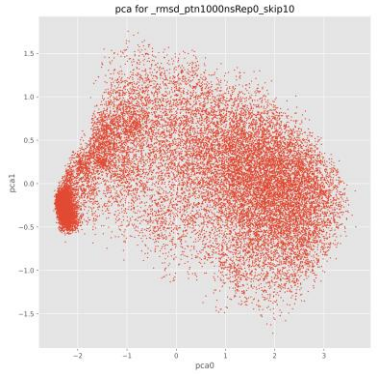


Application on β hairpin folding trajectory

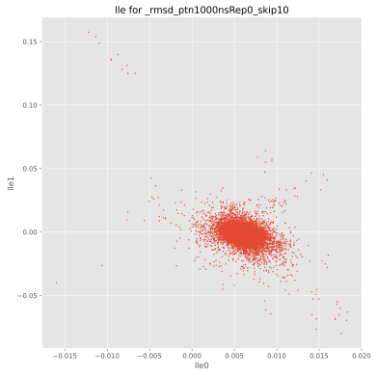


https://github.com/codesrivastavalab/ReplicaExchangeWithHybridTempering

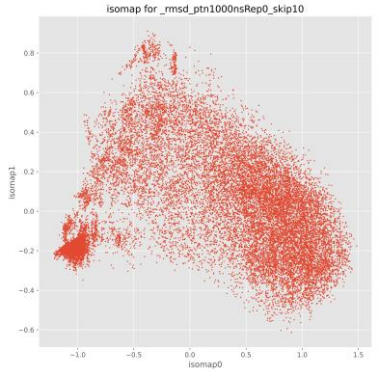
PCA



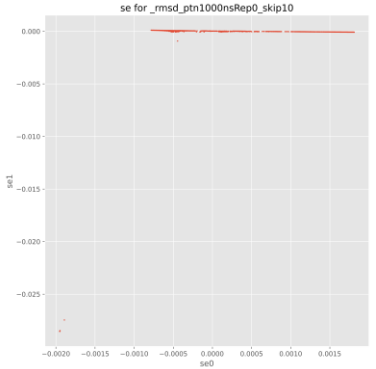
LLE



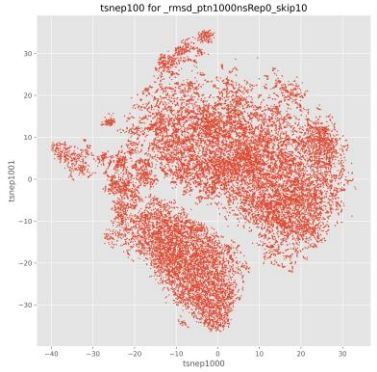
ISOMAP



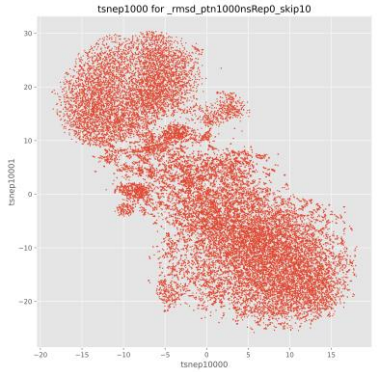
SE



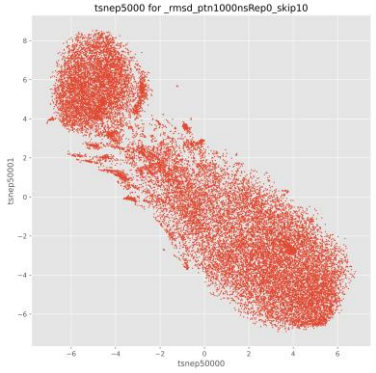
TSNE



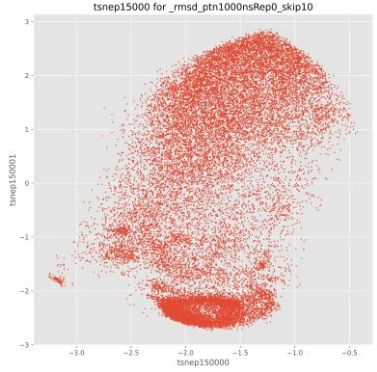
P=100



P=1000

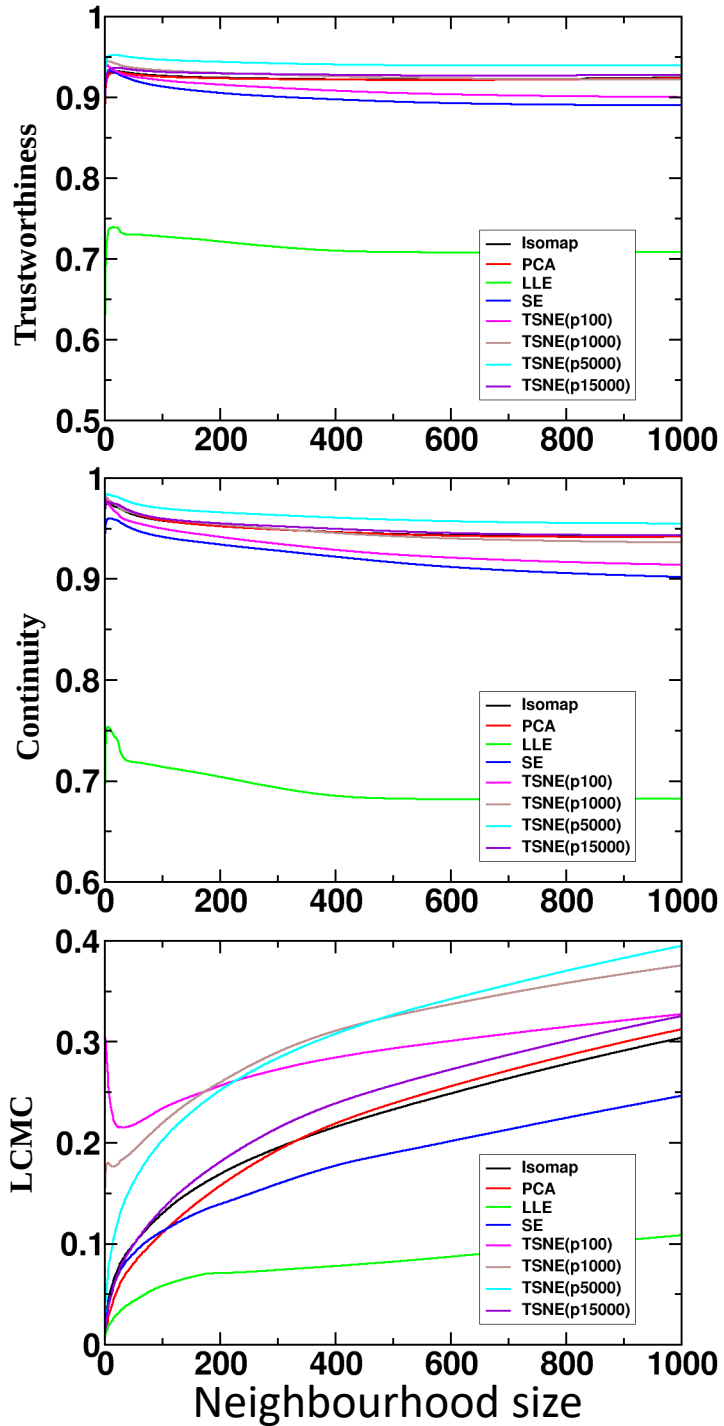


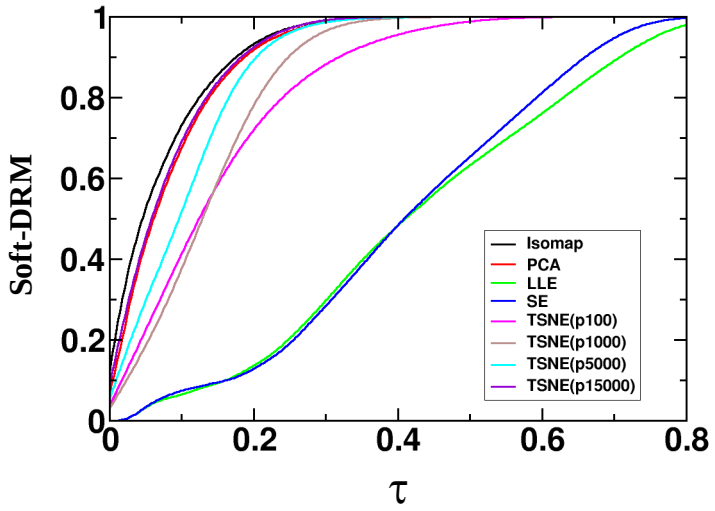
P=5000

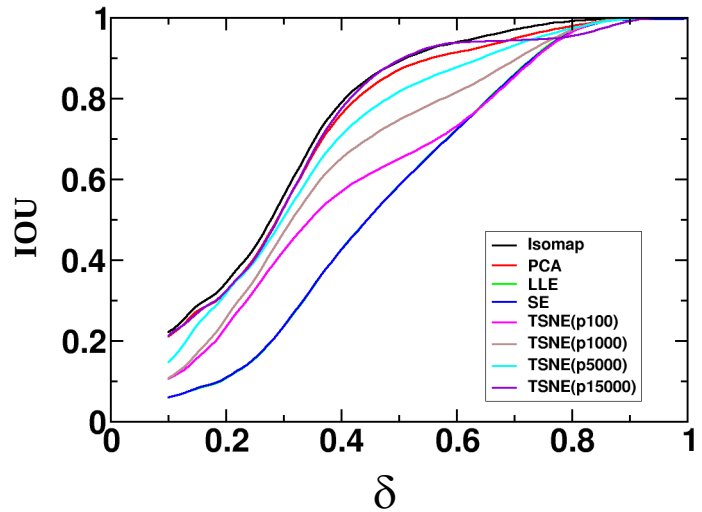


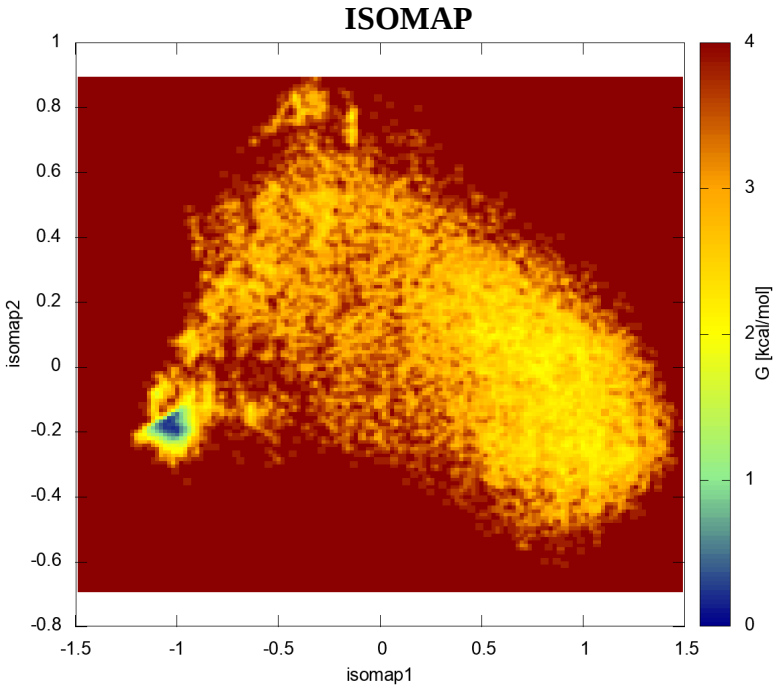
P=15000

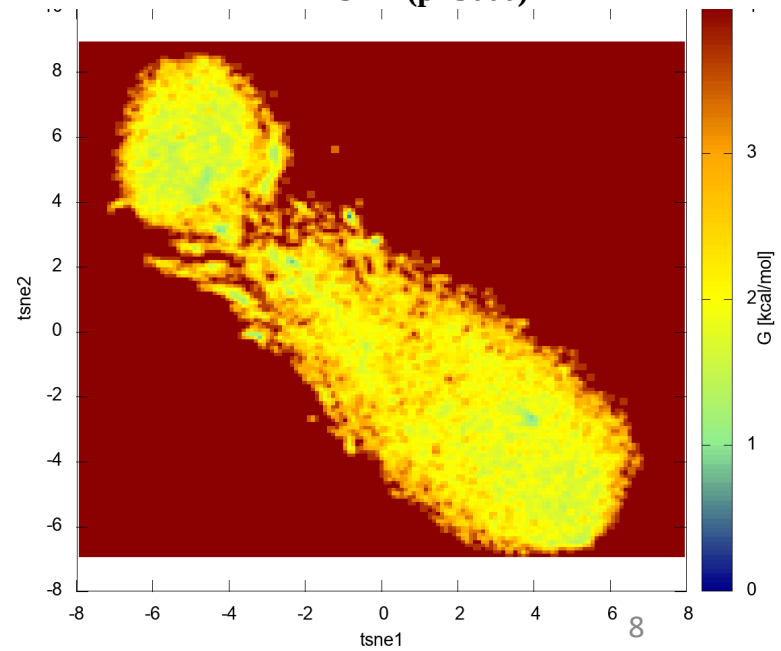
β hairpin folding



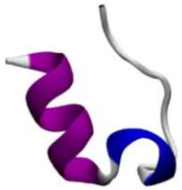


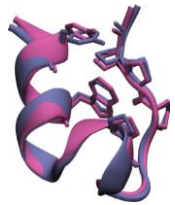


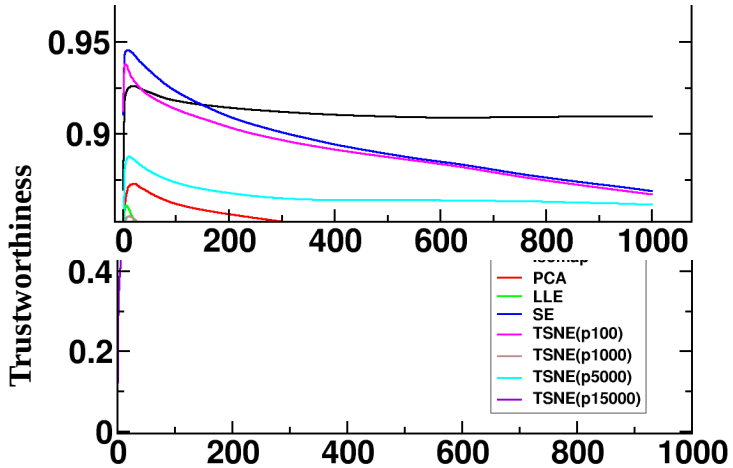


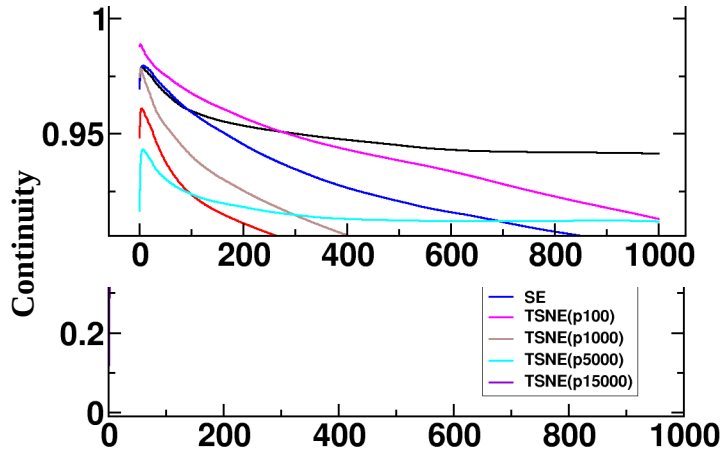


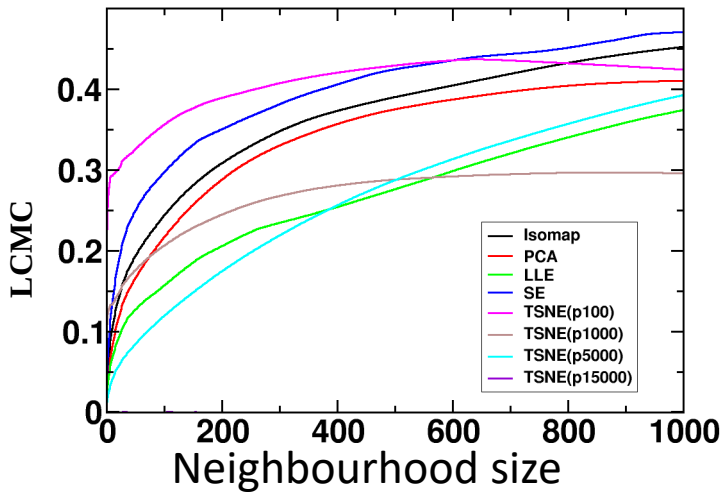
Trp-cage folding

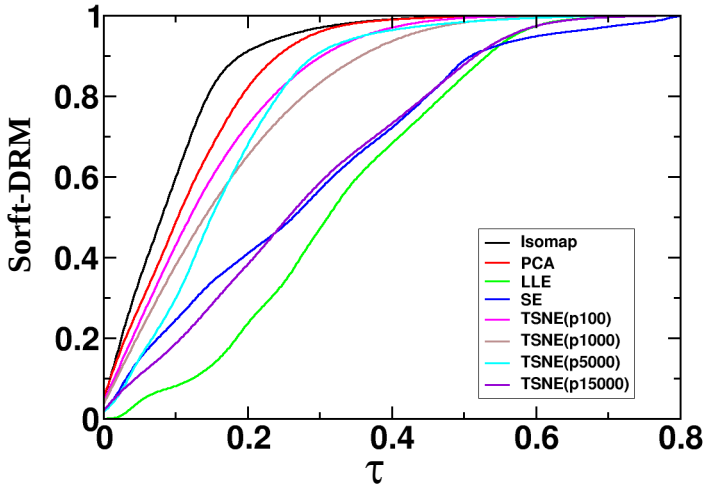


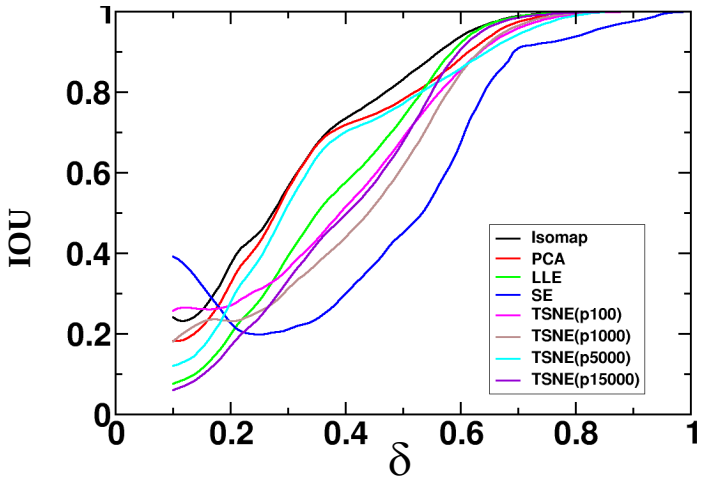


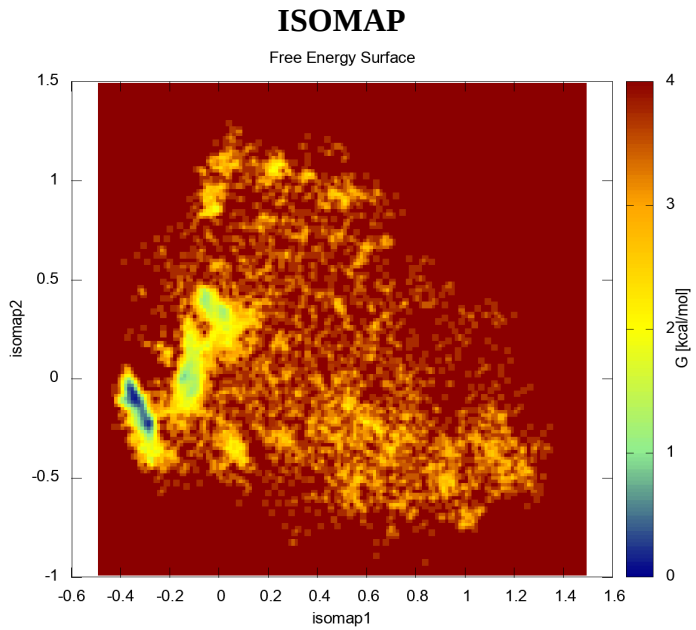


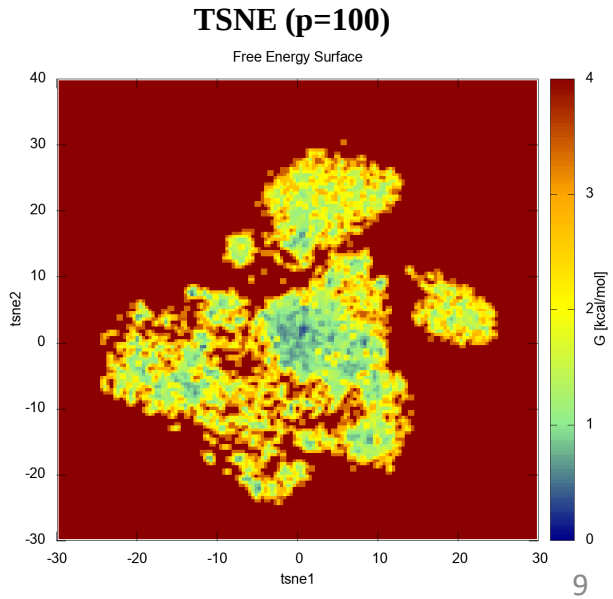




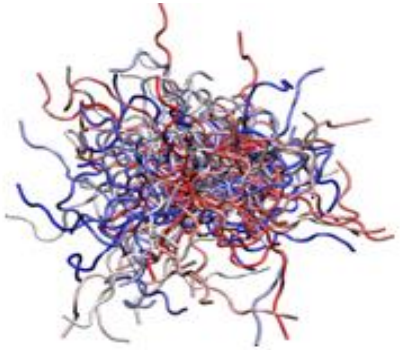


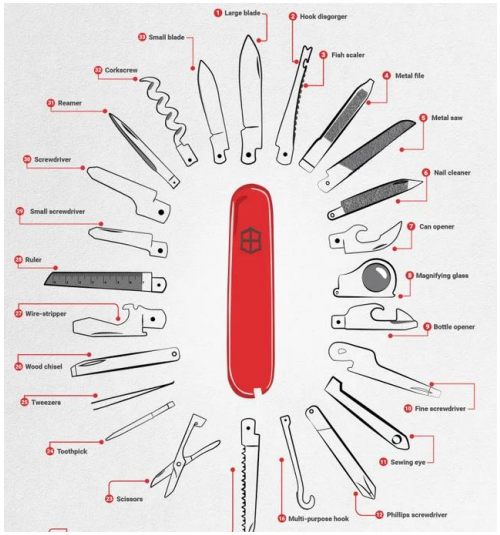


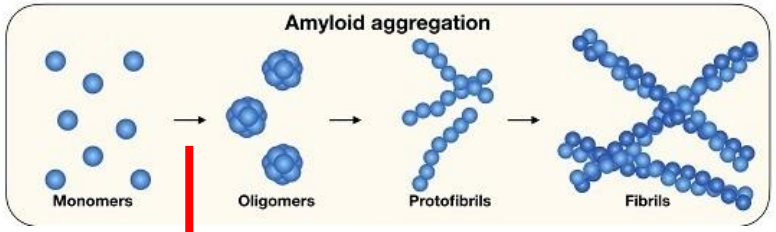




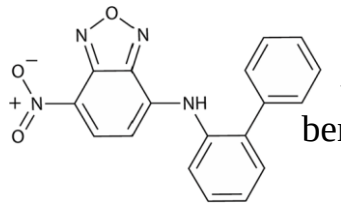
Clustering Intrinsically disordered protein ensemble



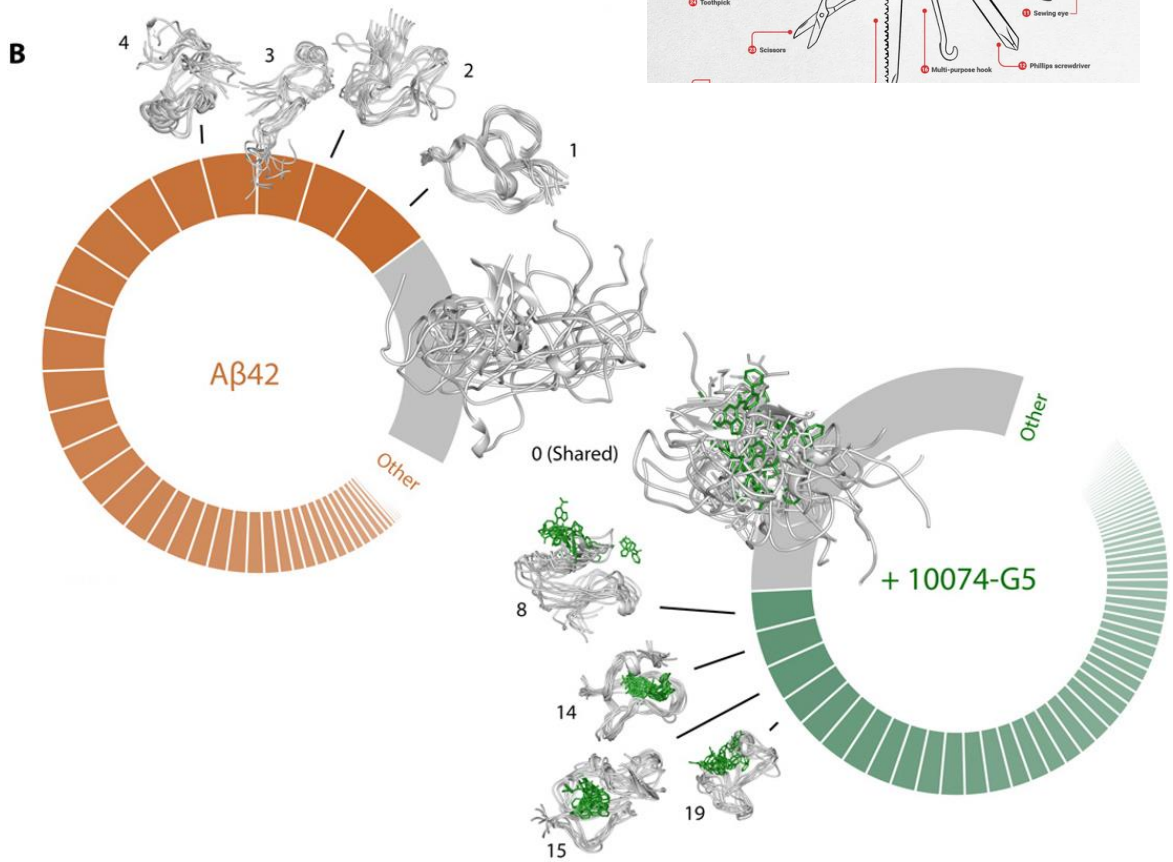




G5 sequesters Aβ in its monomeric form.



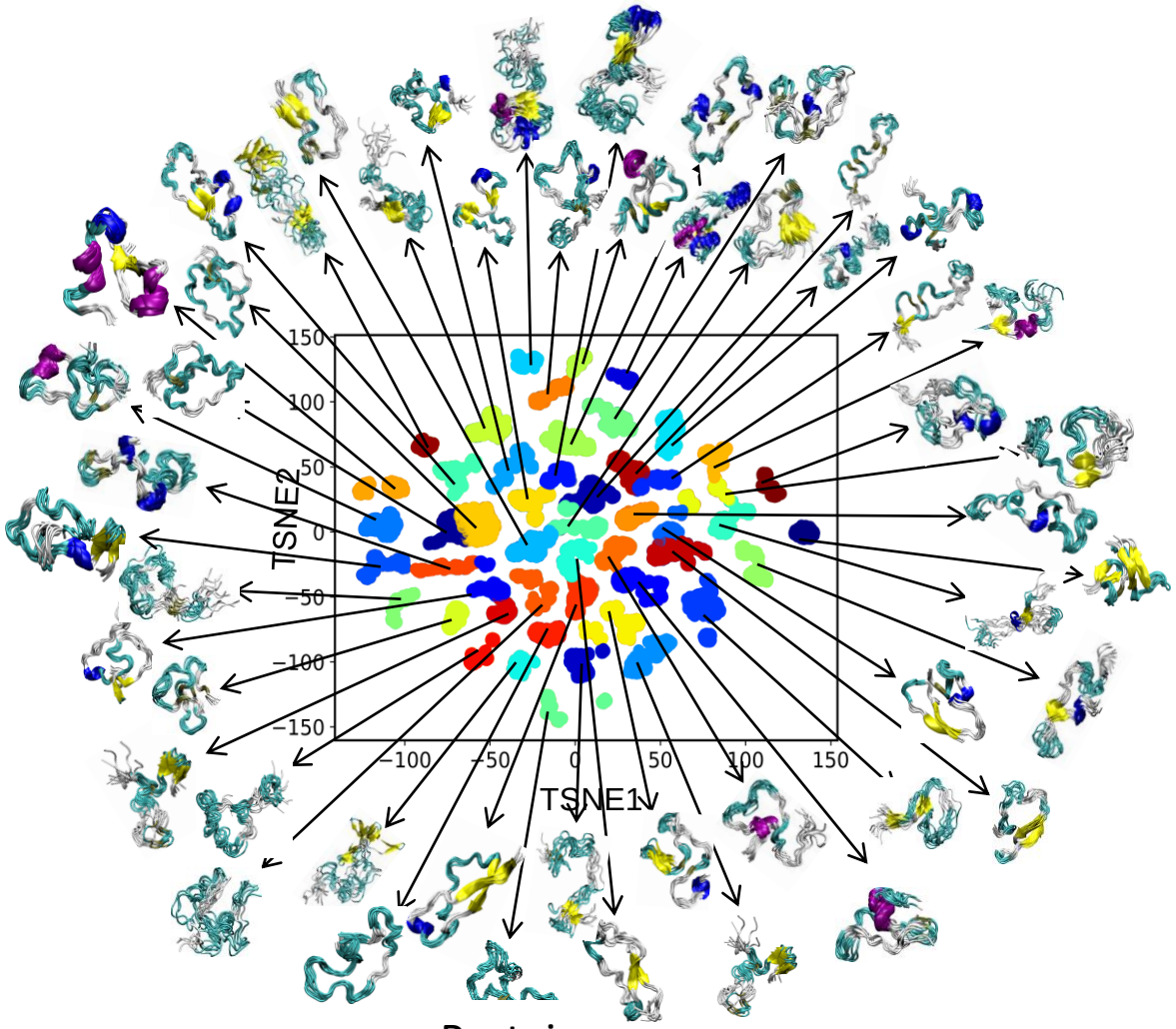
10074-G5
{biphenyl-2-yl-(7-nitro-benzo[1,2,5]oxadiazol-4-yl)-amine}



Heller et al., Sci. Adv. 2020;

Using tSNE for clustering conformations

IDR ensemble of Aβ

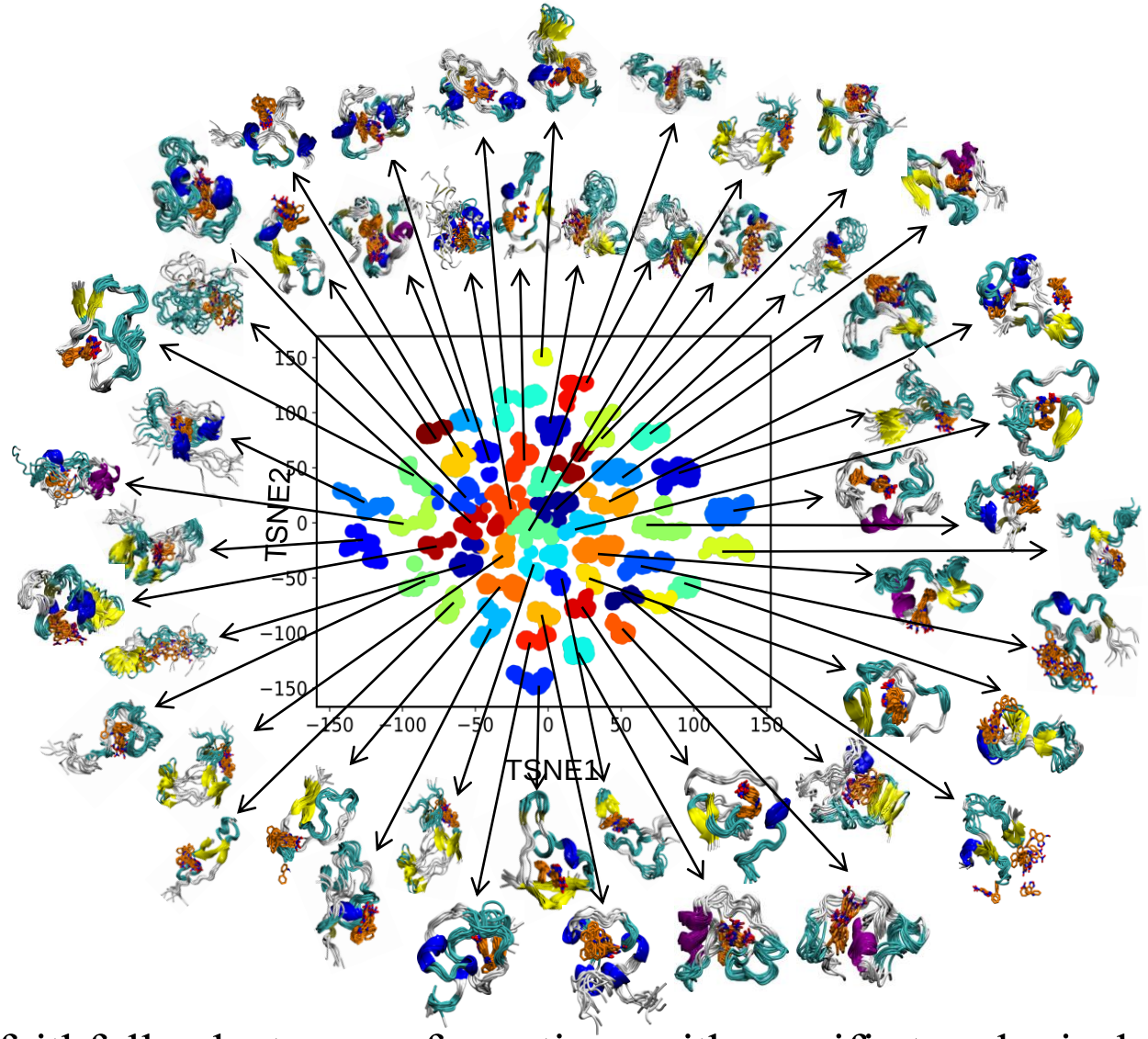


Protein

Beta sheet Alpha helix 3-10 helix

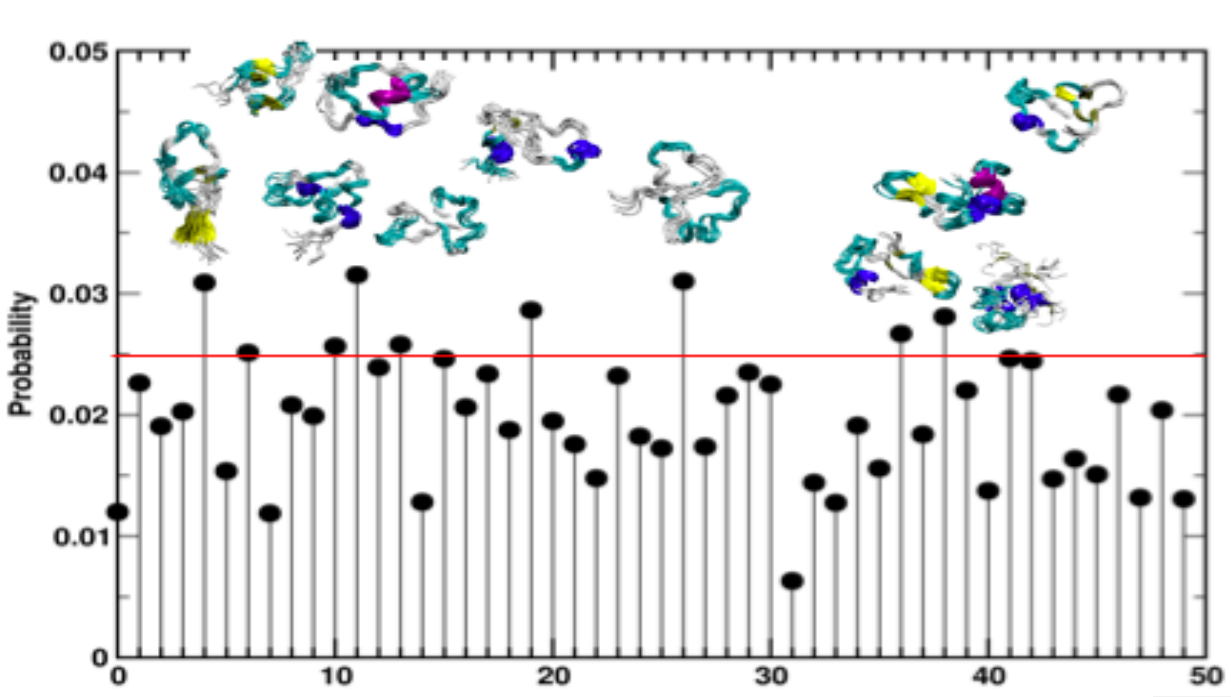
G5

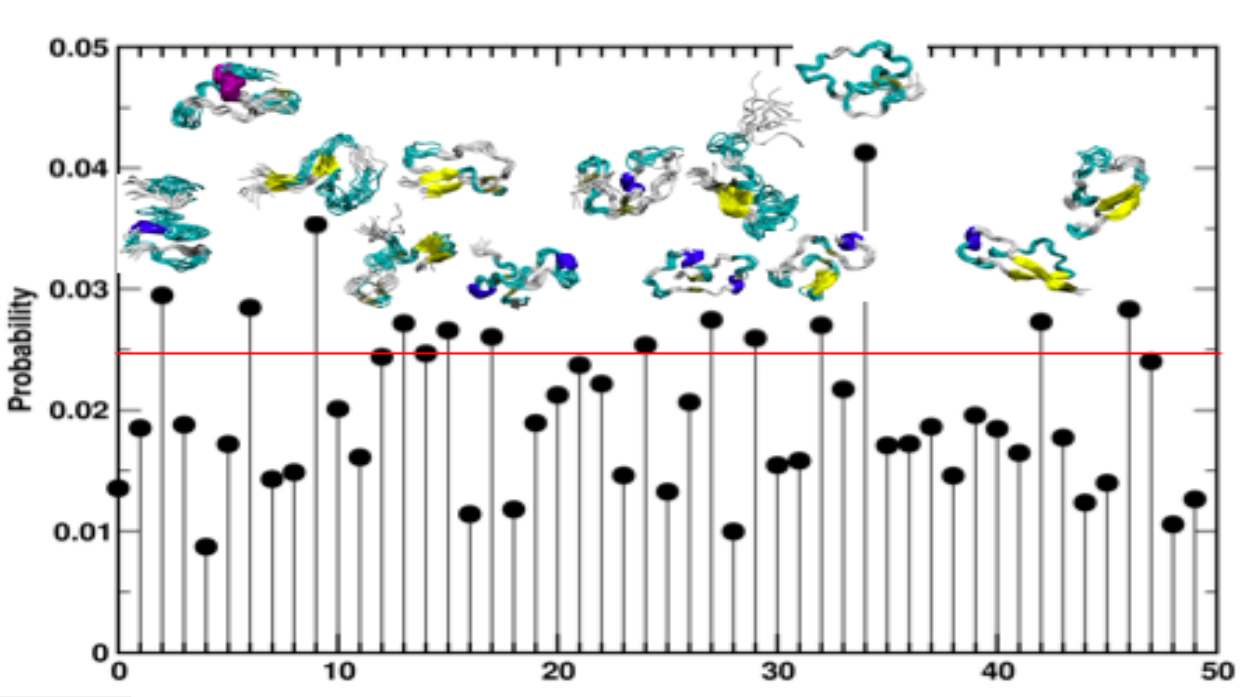
IDR ensemble of Aβ - G5 complex



TSNE faithfully clusters conformations with specific topological features. And binding motifs are better identified with tsne.

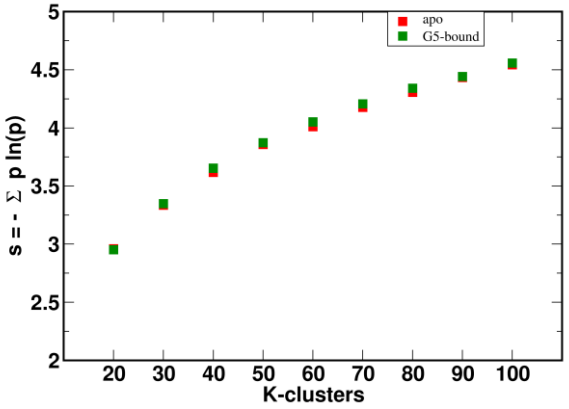
Conformational Populations: Apo & G5-bound ABeta



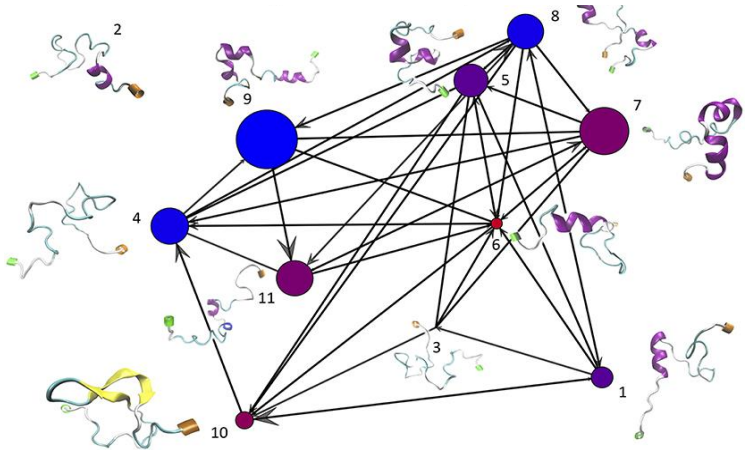


Applications

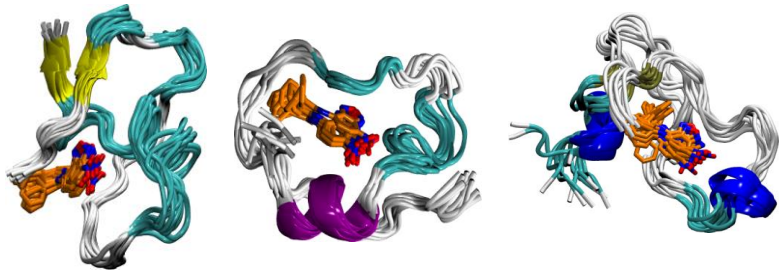
Thermodynamics estimates: Entropy



Kinetic studies with MSM



Molecular recognition and design



Summary and Conclusion

We have presented a unified framework for comparing different DR techniques.

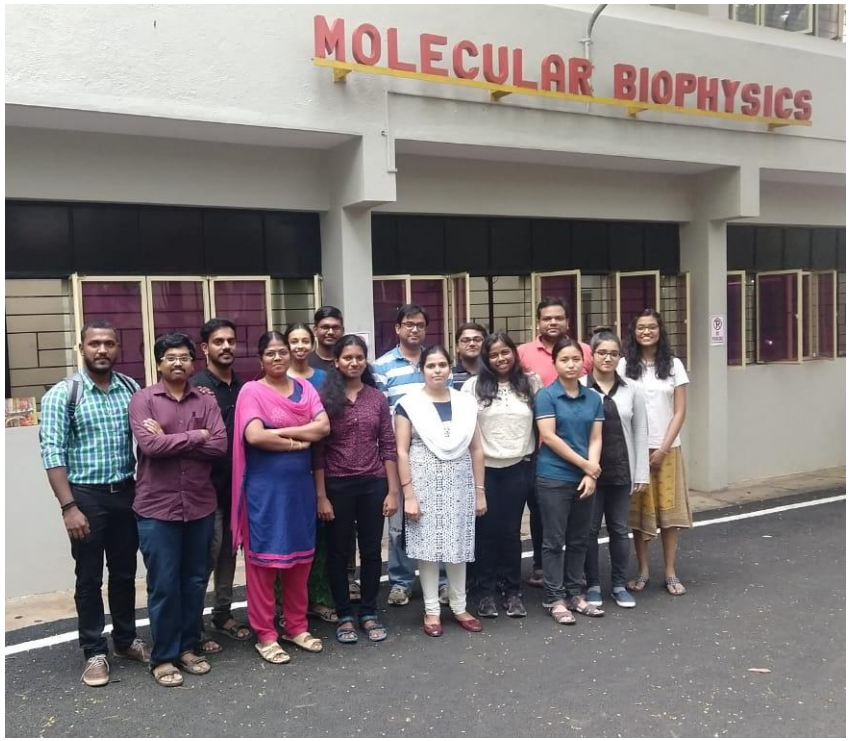
This framework can assist in choosing appropriate DR method that faithfully represent the high-d data into Low-d.

Further we showed that TSNE can be utilized for faithfully clustering heterogeneous data such as that of an IDP ensemble.

Acknowledgements

Upal Bhattarchaya, IISER Kolkata

Dr. Anand Srivastava's Lab,
Indian Institute of Science



Funding

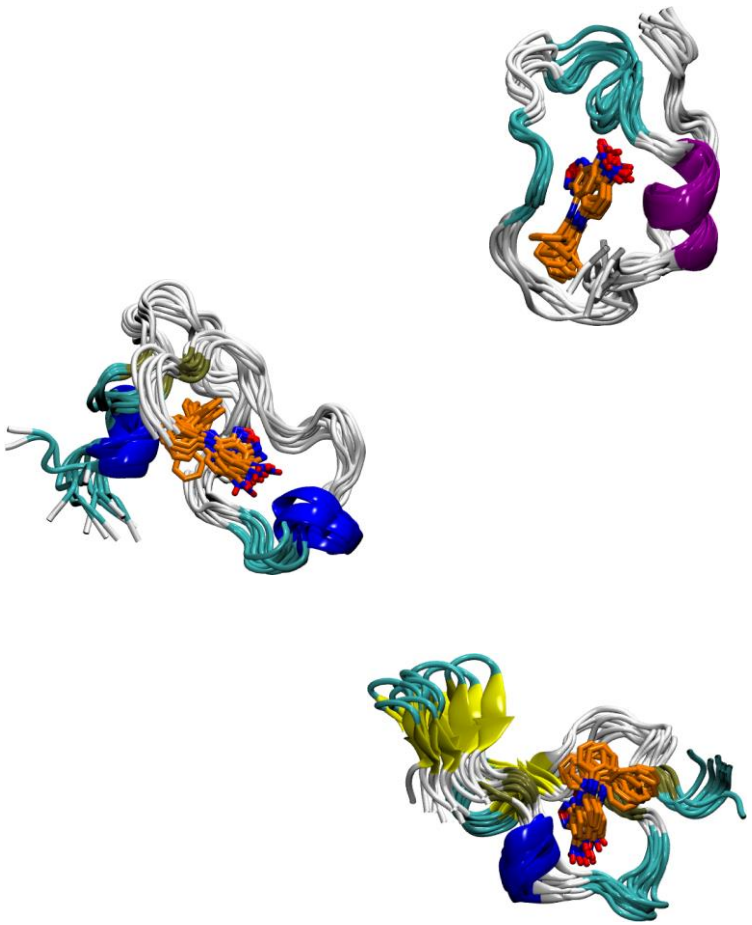
IndiaAlliance
DBT wellcome

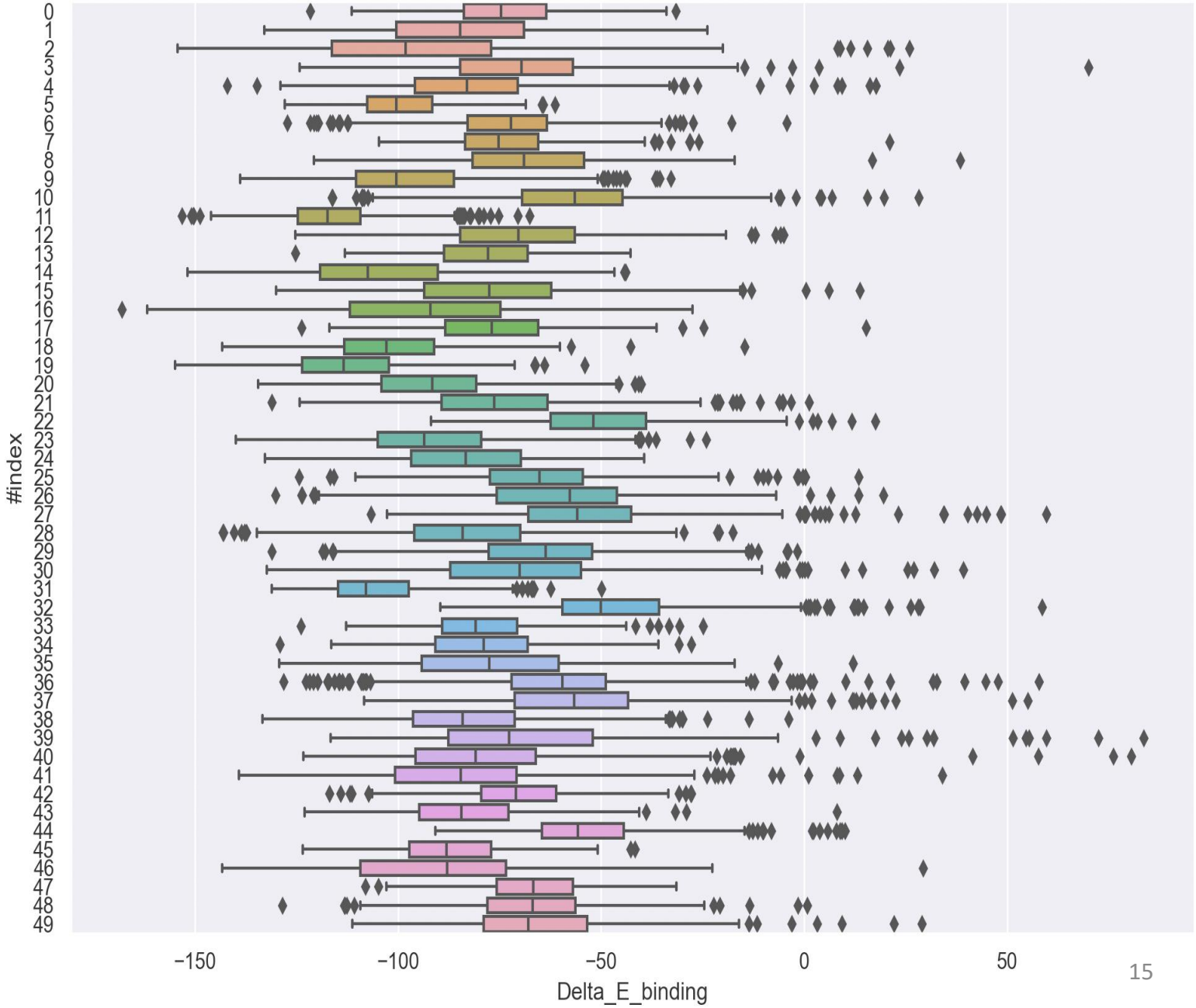
Computational Facility

- Supercomputer Education and Research Centre, Indian Institute of Science
- Compute Canada

Thank you for your attention!

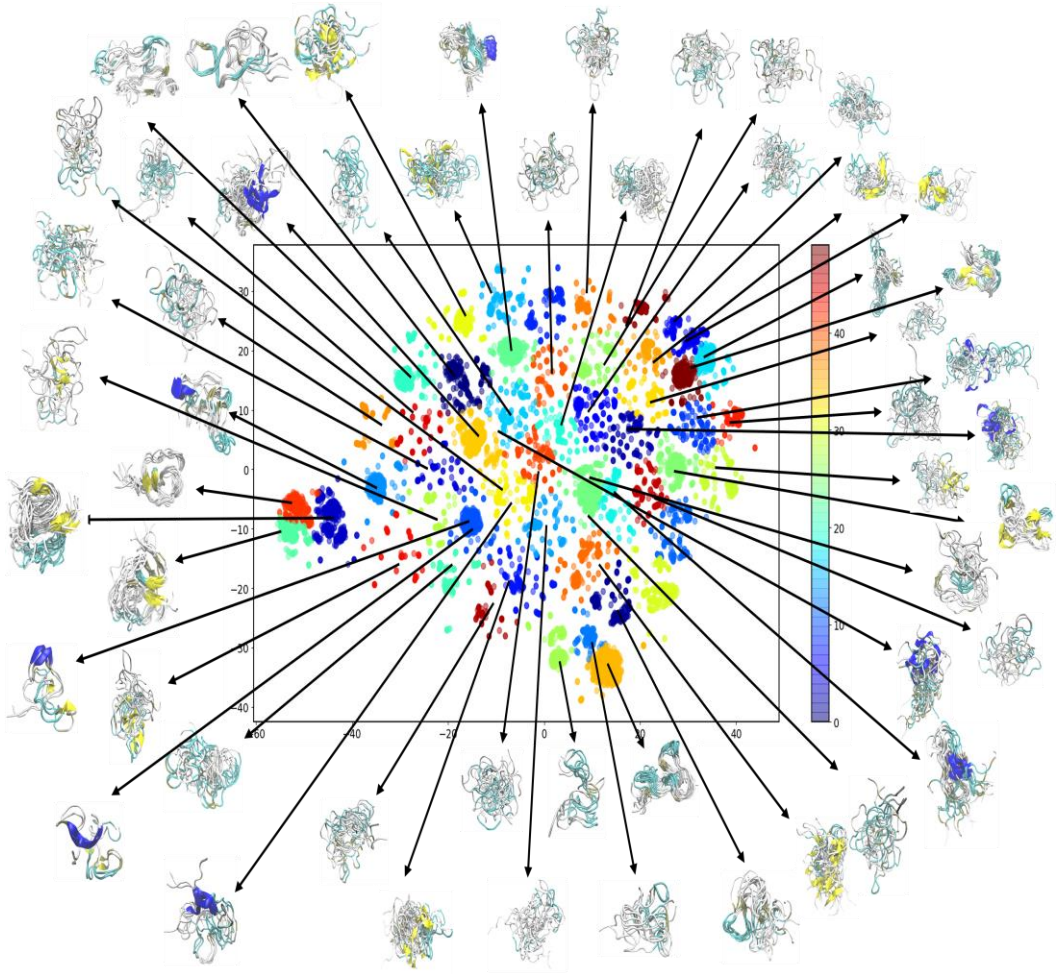
Aβ - G5 interaction





Using tSNE for clustering conformations of HNRNPA1

IDR ensemble of HnRNPA1-RGG domain



Protein

Beta sheet | Alpha helix | 3-10 helix

G5

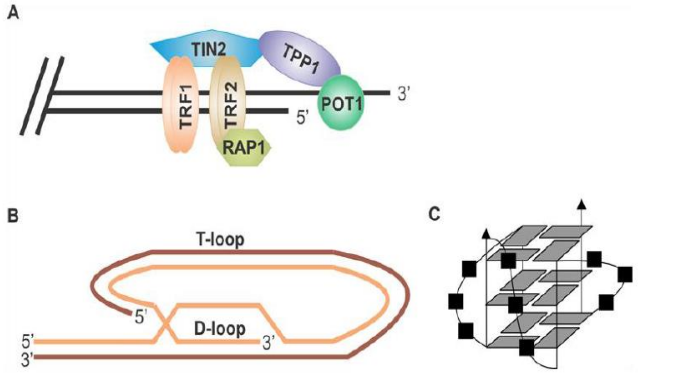
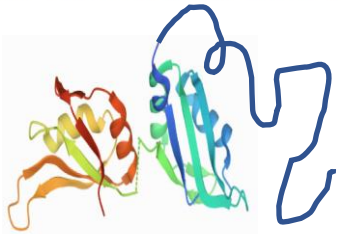


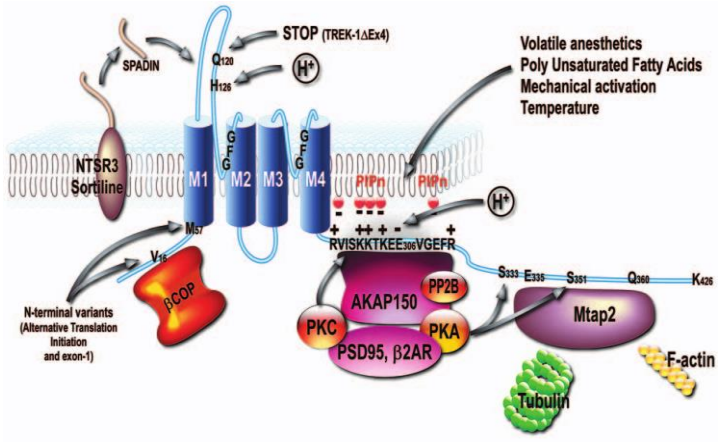
Figure 1-3: End protection problem resolved by the Shelterin complex (A) and the formation of the higher order structures, t-loop and d-loop (B) as well as G-quadruplexes (C) formed at the ends of the telomere.



Ghosh and Singh, NAR 2018, 2020
Mittag and co-workers, Science 2020, NAR 2021

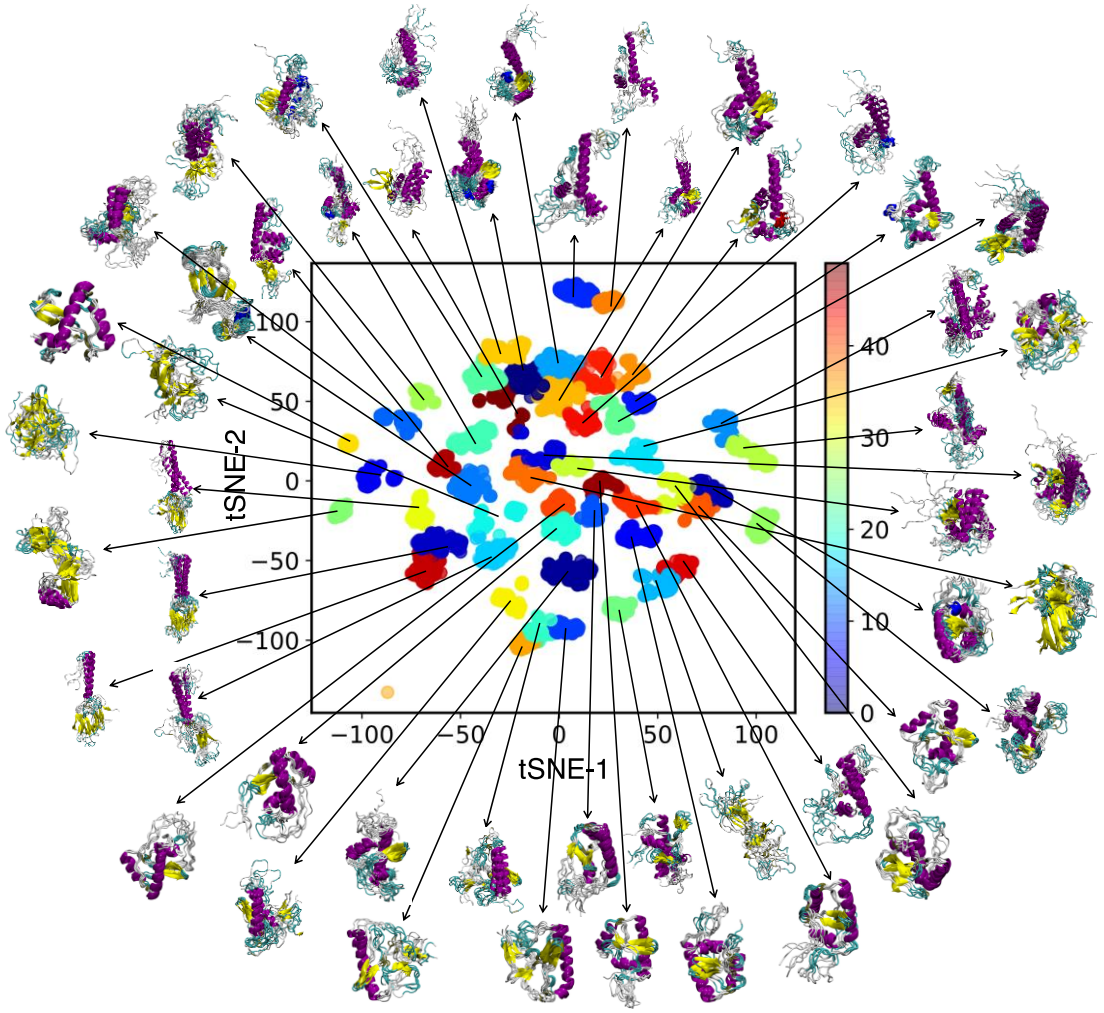
TSNE faithfully clusters conformations with specific topological features. And binding motifs are better identified with tsne.

Using tSNE for clustering conformations of TREK-CTD



Honoré, Nat Rev Neurosci. 2007

IDR ensemble of TREK-1 C-terminal domain

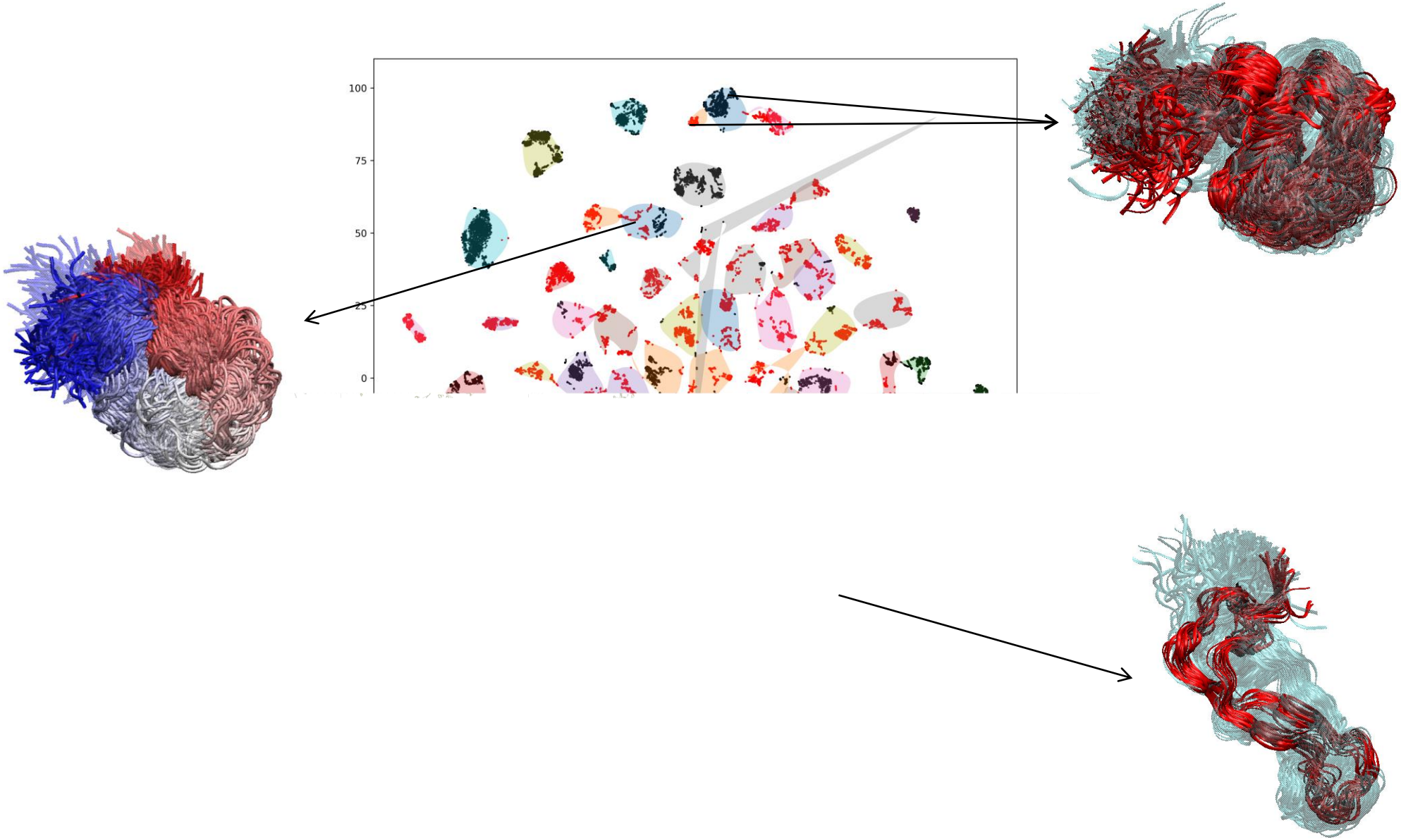


Protein

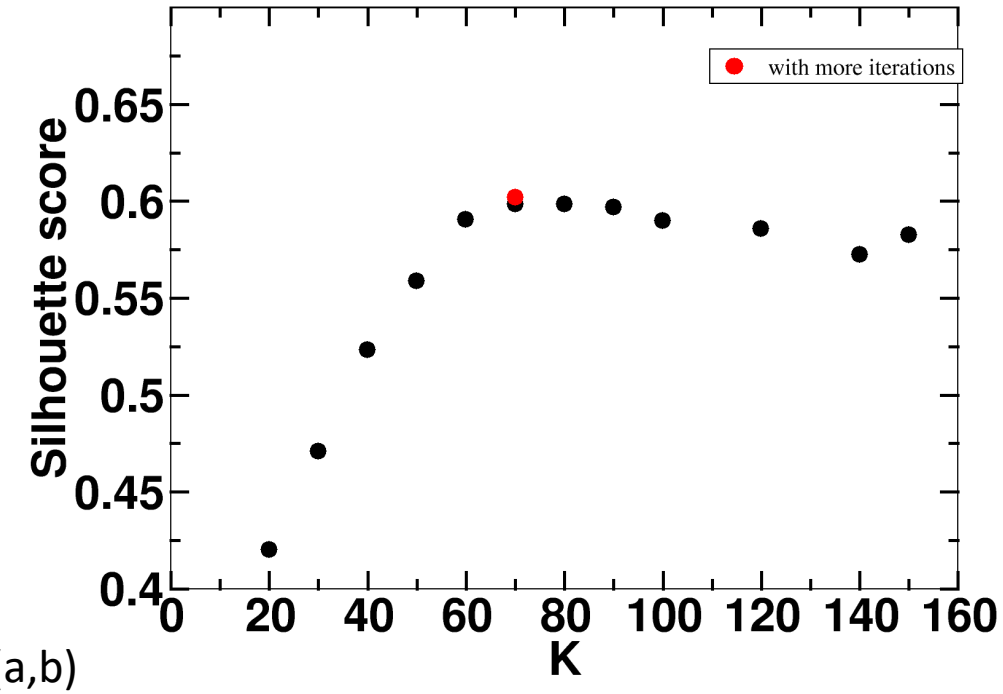
Beta sheet | Alpha helix | 3-10 helix

G5

TSNE faithfully clusters conformations with specific topological features. And binding motifs are better identified with tsne.


100
75
50
25
0

Kmeans clustering



Silhouette Score = (b-a)/max(a,b)
where a= average intra-cluster distance i.e the average distance between each point within a cluster.
b= average inter-cluster distance i.e the average distance between all clusters.

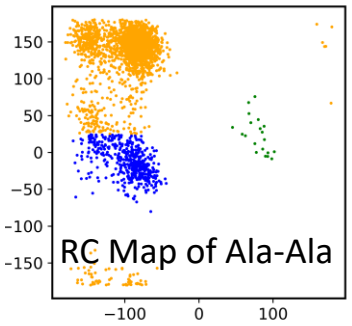
1: Means clusters are well apart from each other and clearly distinguished.
0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.
-1: Means clusters are assigned in the wrong way.

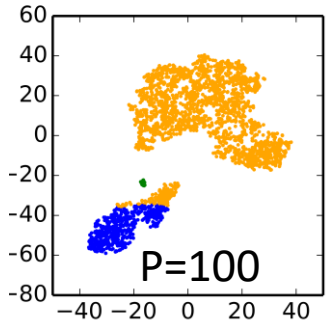
Method

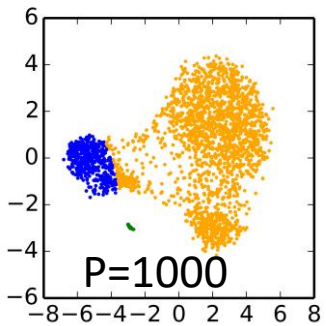
T-distributed stochastic neighbour embedding (TSNE) for clustering IDPs

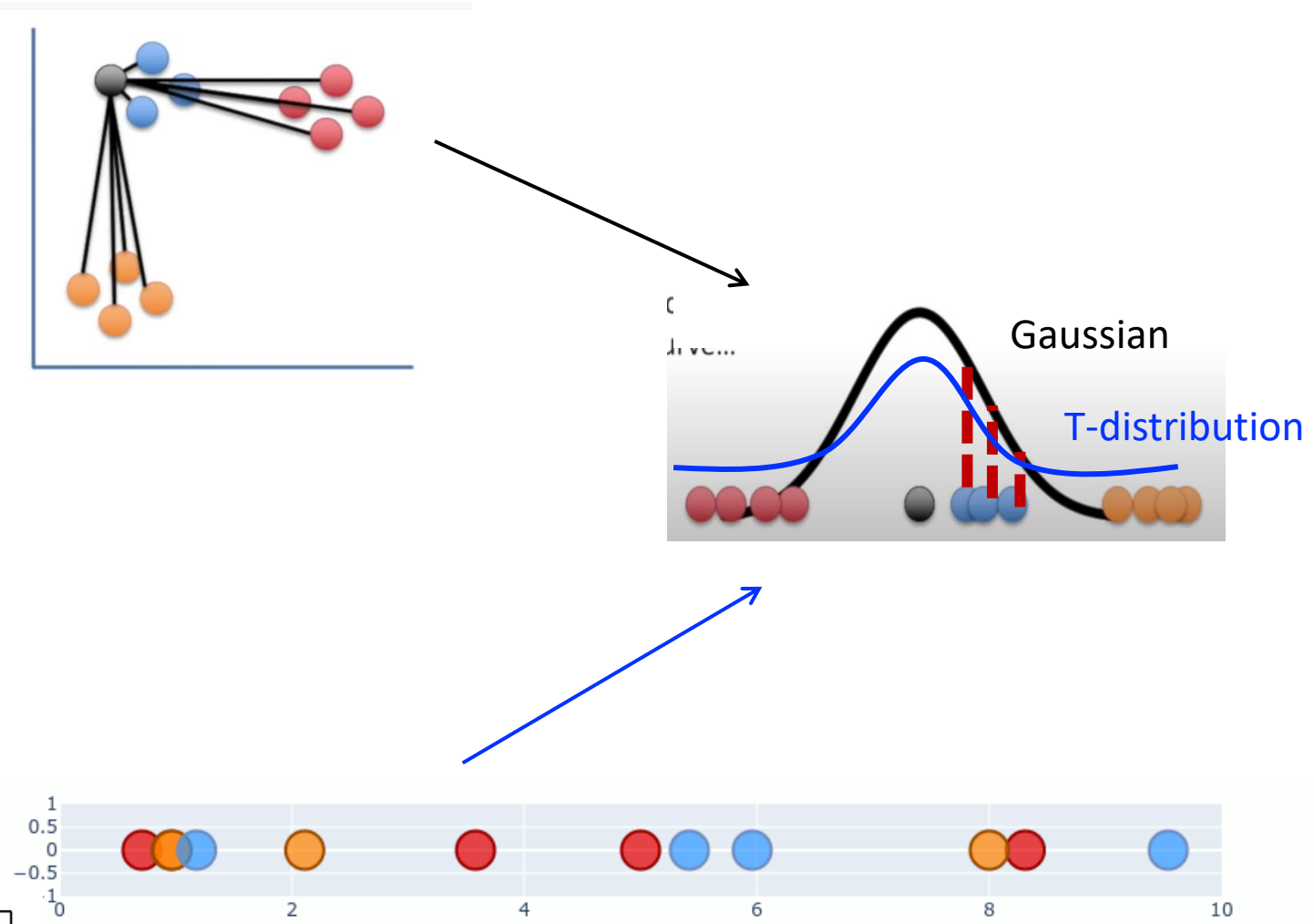
- TSNE reduces the high dimensional space into human readable low dimensional space (2d or 3d).
- While preserving original, high-dimensional structure.
- Unlike PCA it's a non-linear technique.

Hyper-parameter: Perplexity









Matten and Hinton, 2008, J. Machine Learning Research