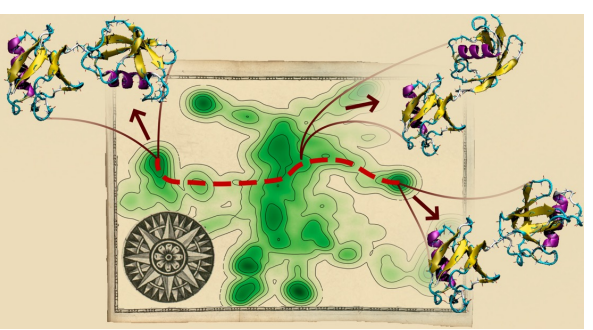
Universität Konstanz

Machine learning in biomolecular simulations: from characterizing conformational free energy landscapes to scale bridging

Christine Peter

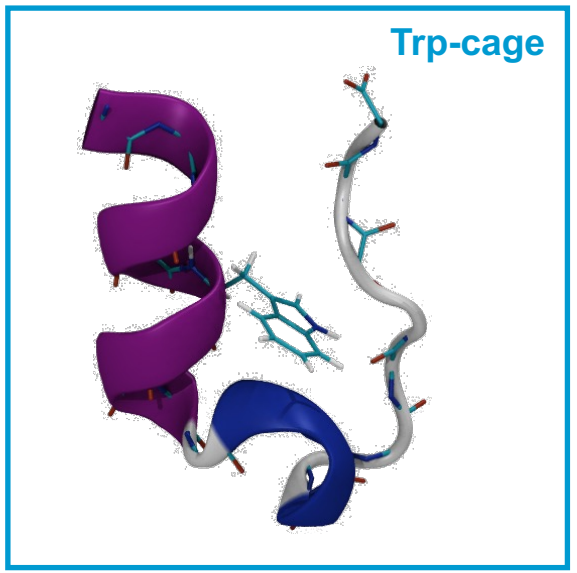
ECT* Workshop, Trento 2024

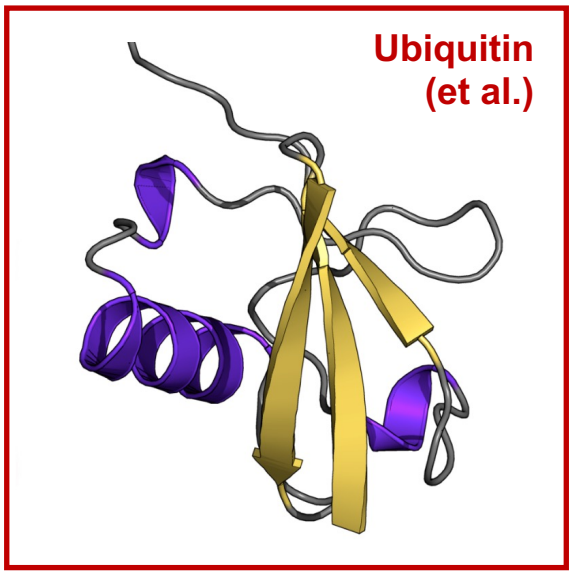


Outline

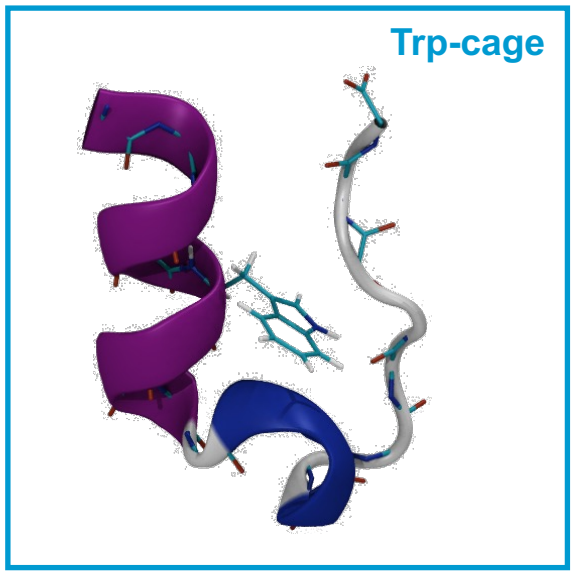
- Setting the stage
- EncoderMap
  - Learning meaningful representations of conformational phase space
  - Generating protein conformations and visualizing molecular motion
- Extracting meaningful feature sets from graph representations of proteins
  - Generating residue interaction landscapes
- Utilizing low-dimensional embeddings for clustering
  - Identifying conformational states
- Backmapping based sampling
  - Linking scales through low-dimensional representations

Two recurring systems

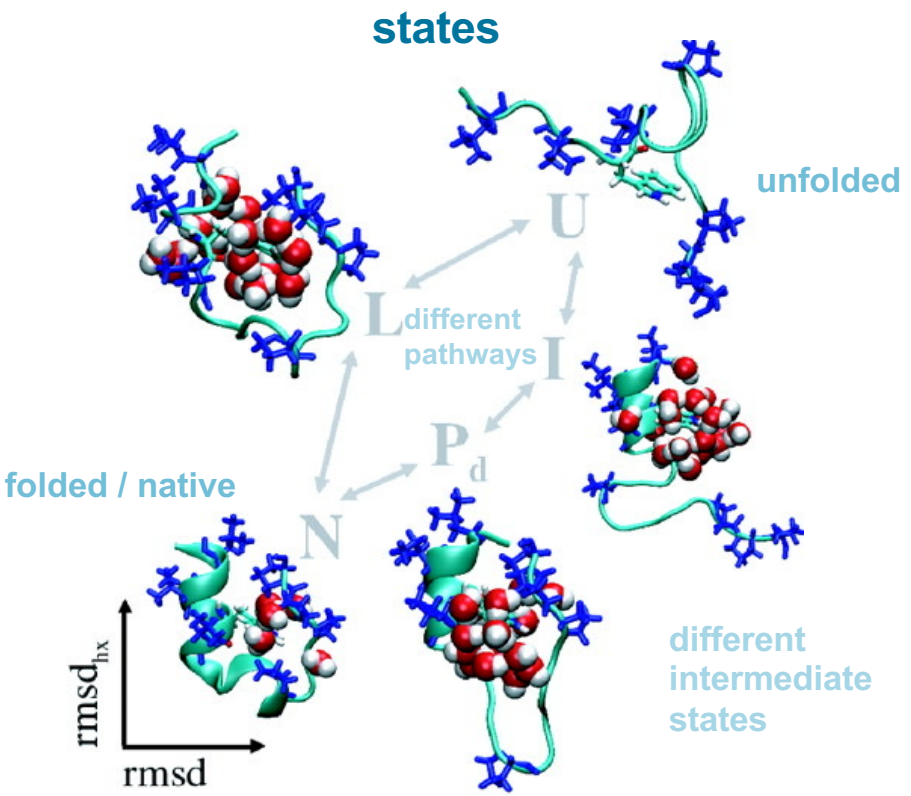




Two recurring systems

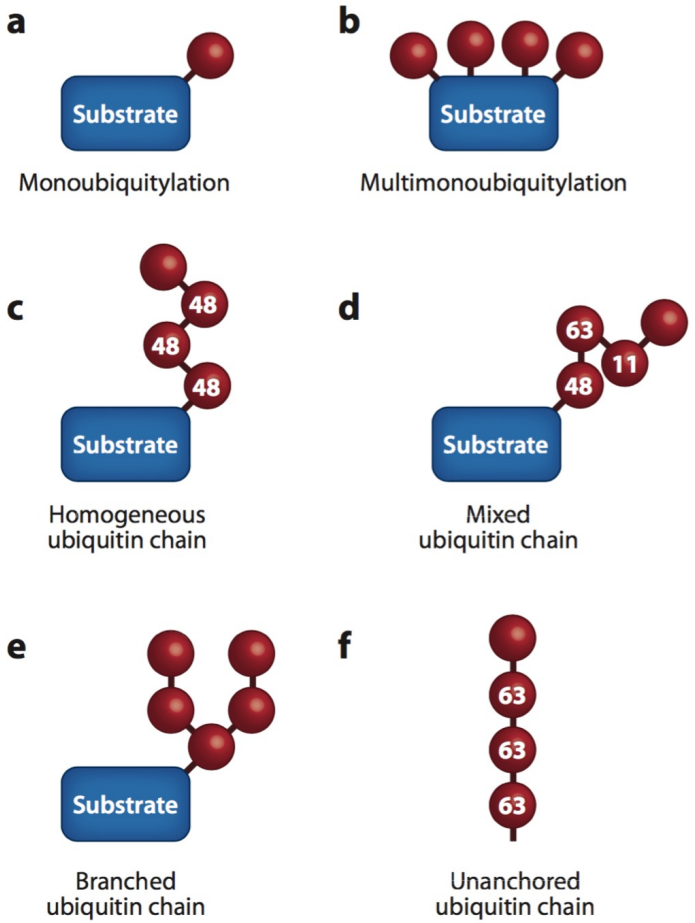


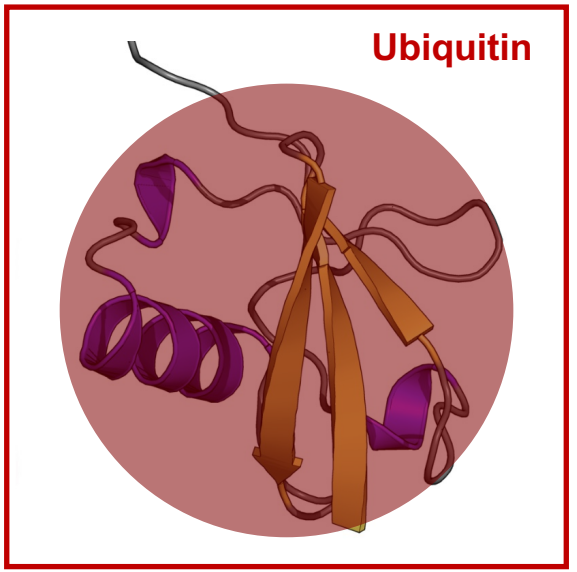
- folding
- states
- features & CVs
- low-dimensional representations



Juraszek, J. and P. G. Bolhuis.
PNAS 103.43 (2006): 15859-15864.

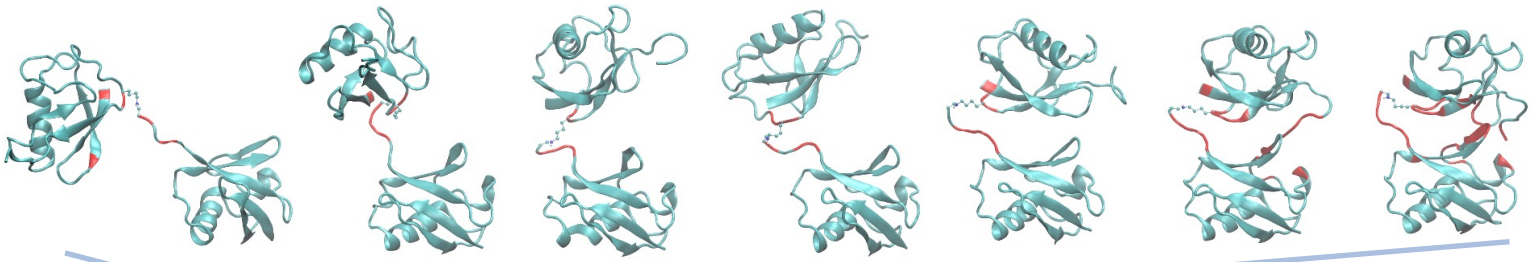
Two recurring systems: Trp-cage and Ubiquitin et al.



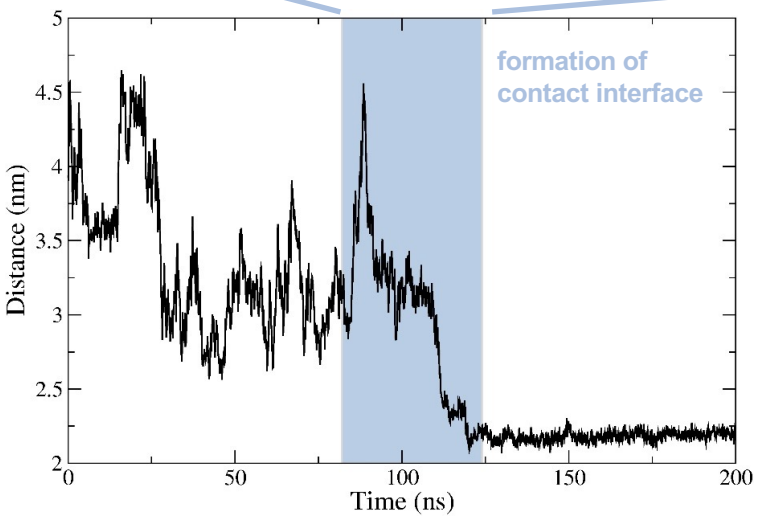


- protein-protein interactions
- protein interfaces
- sampling / scale-bridging

An atomistic simulation of K48-linked di-Ubiquitin

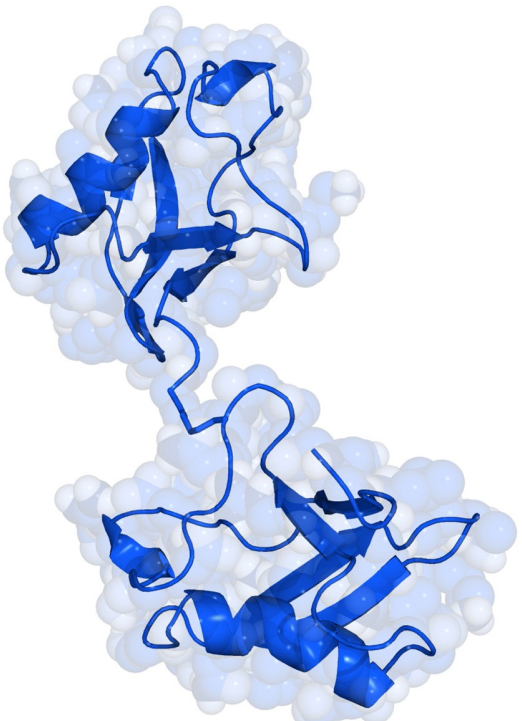


protein interfaces



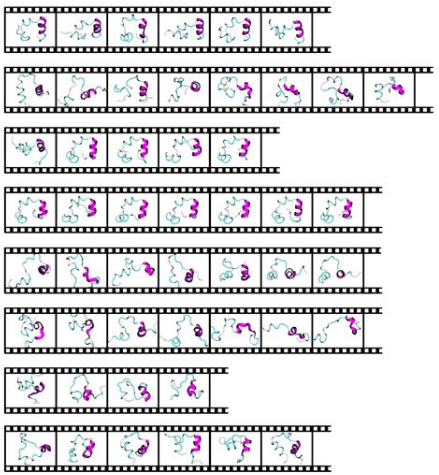
the accessible timescales are too short
➔ need for enhanced sampling

A coarse grained simulation of K48-linked di-Ubiquitin

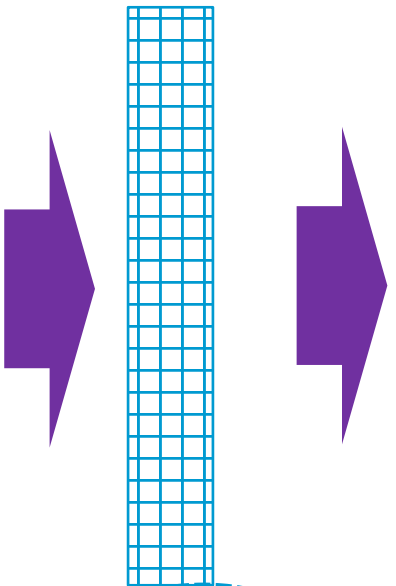


➔ Need for methods to characterize conformational phase space

Low-dimensional representations of conformational phase space



Simulation data (Cartesian coordinates)



Input data set (features)

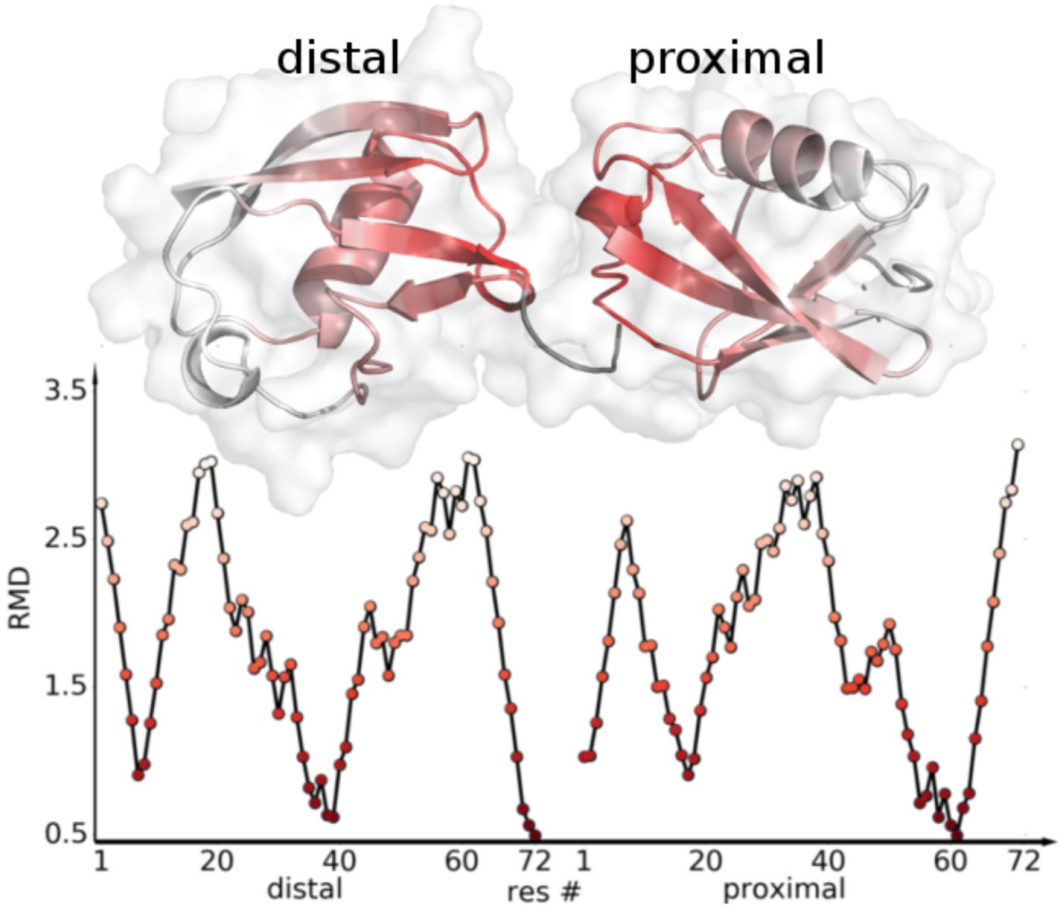
internal distances, dihedrals, ...

Dimensionality reduction
PCA, MDS, ...

Map in low-dim space

Conformational states

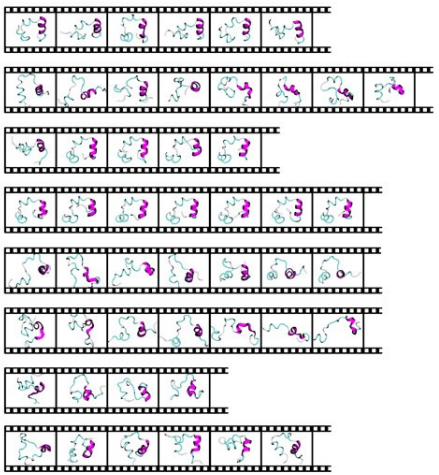
Collective variables that characterize dimer structures



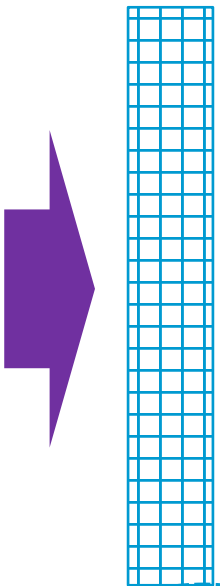
➔ 144 residue-wise minimum distances (RMD)

➔ fingerprints of each structure

Low-dimensional representations of conformational phase space



Simulation data (Cartesian coordinates)



Input data set (features)

internal distances, dihedrals, ...

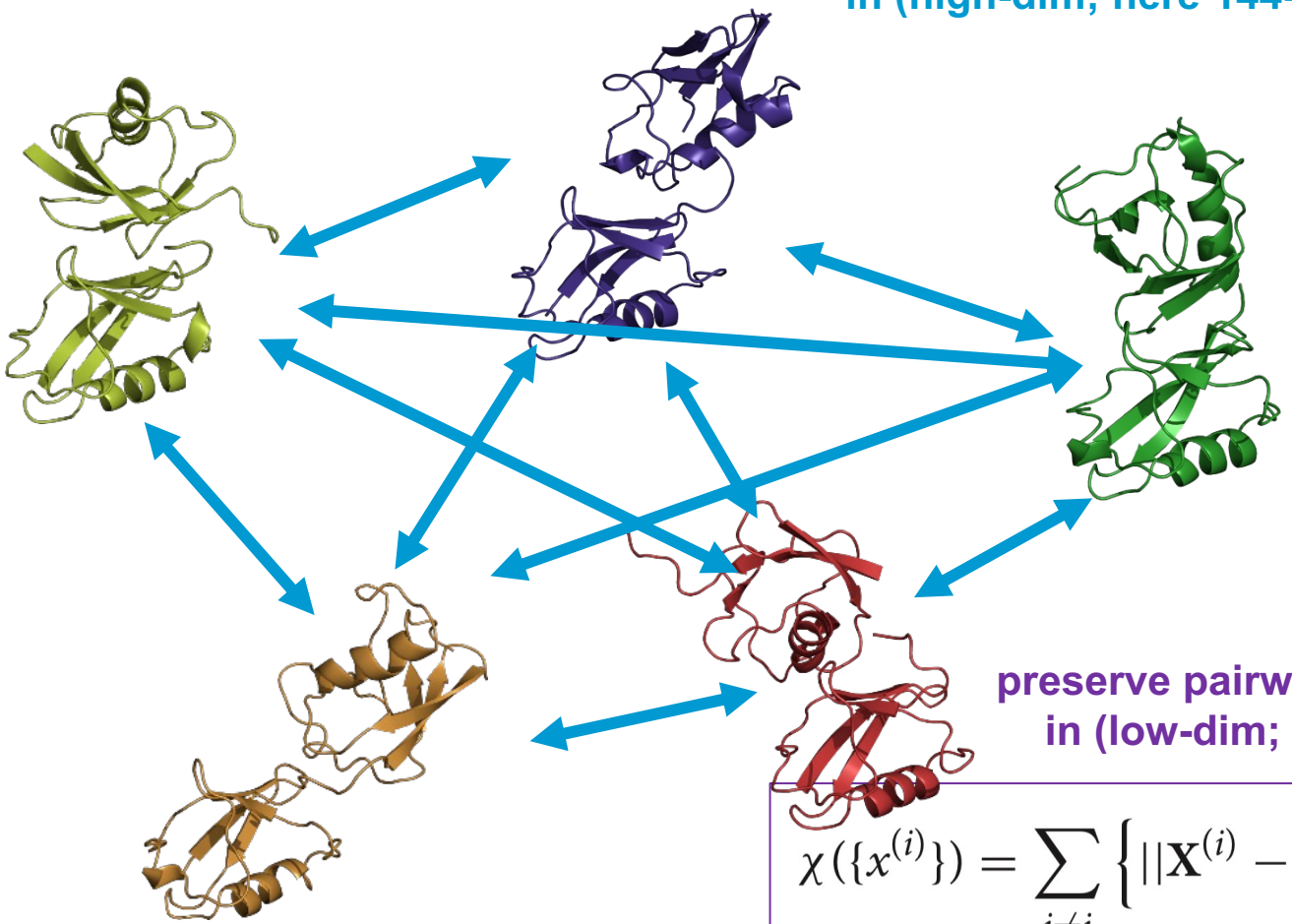
Dimensionality reduction
PCA, MDS, ...

Map in low-dim space

Conformational states

The idea behind metric multidimensional scaling

pairwise differences/distances between structures
in (high-dim; here 144-dim) feature space



preserve pairwise differences/distances
in (low-dim; here 2-dim) embedding

$$\chi(\{x^{(i)}\}) = \sum_{i \neq j} \left\{ ||\mathbf{X}^{(i)} - \mathbf{X}^{(j)}|| - ||\mathbf{x}^{(i)} - \mathbf{x}^{(j)}|| \right\}^2$$

The idea behind SketchMap (Ceriotti et al J. Chem. Theory Comput. 2013)

distances in high- and low-dimensional spaces are transformed by a **sigmoid function**

$$s(r,a,b) = 1 - (1 + (2^{a/b} - 1)(r/\sigma)^a)^{-b/a}$$

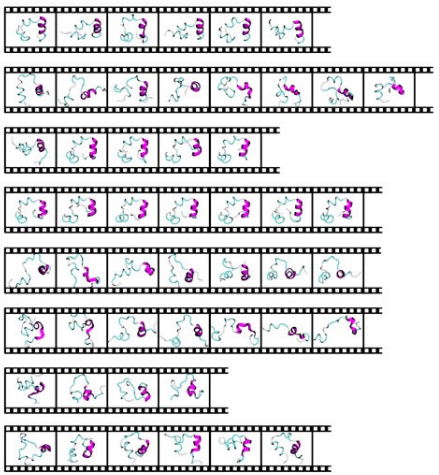
produce mapping by minimizing stress function for **landmark** structures:

$$\chi^2 = \sum_{i \neq j} \left[s(R_{ij}, A, B) - s(r_{ij}, a, b) \right]^2$$

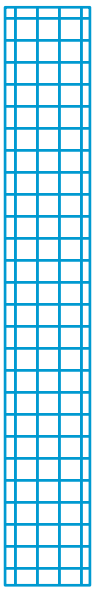
calculate projection **x** of any high-dimensional point **X** by minimizing:

$$\sigma^2(\mathbf{x}) = \sum_{i=1}^{N} \left[s(R_i(\mathbf{X}), A, B) - s(r_i(\mathbf{x}), a, b) \right]^2$$

Low-dimensional representations of conformational phase space



Simulation data (Cartesian coordinates)



Input data set (features)

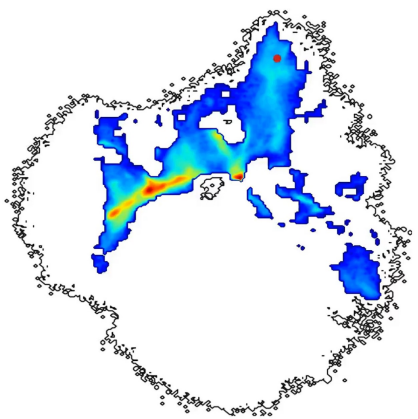
RMDs

Dimensionality reduction
here: SketchMap
- MDS-like
- sigmoid function
- uses landmark points
- projection of all data

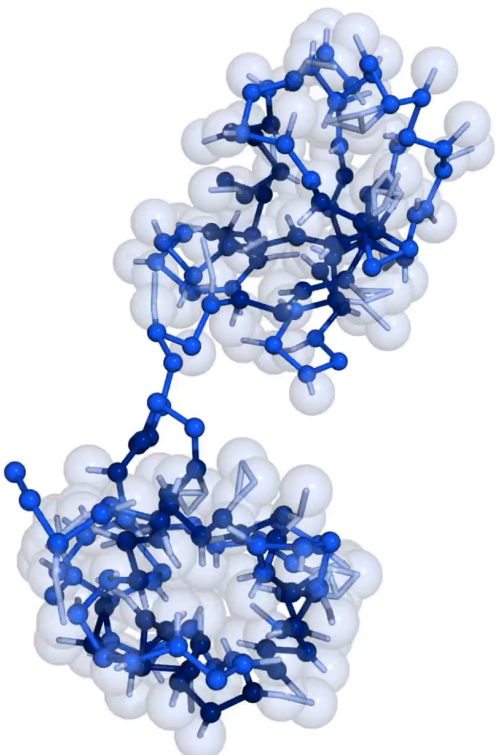
Ceriotti et al *J. Chem. Theory Comput.* **2013**

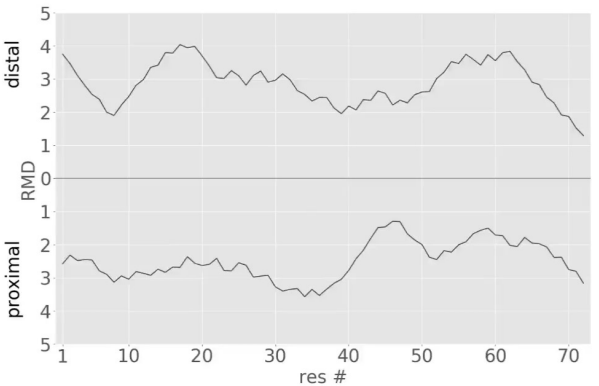
Map in low-dim space

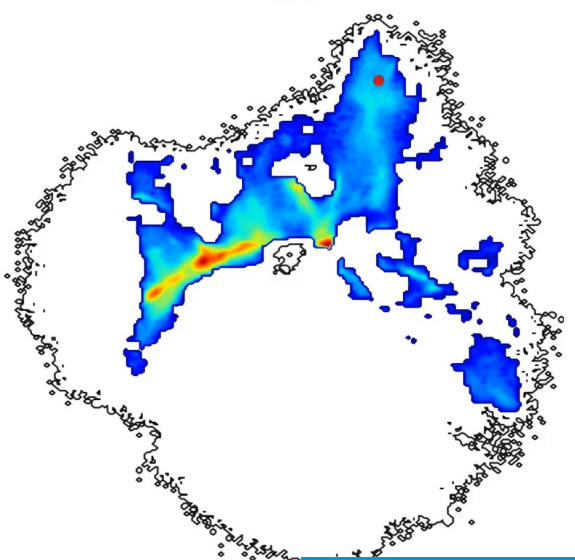
Conformational states



Residue-wise minimum distances as input features: Illustration for one simulation







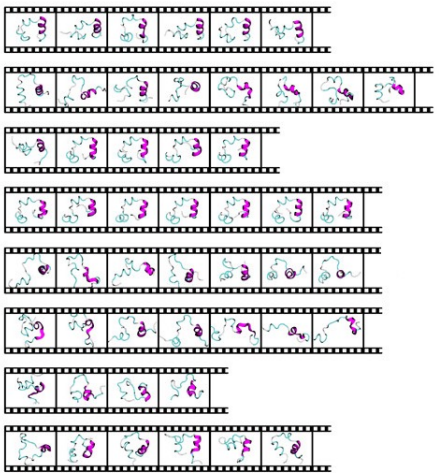
➔ need for methods to process huge datasets

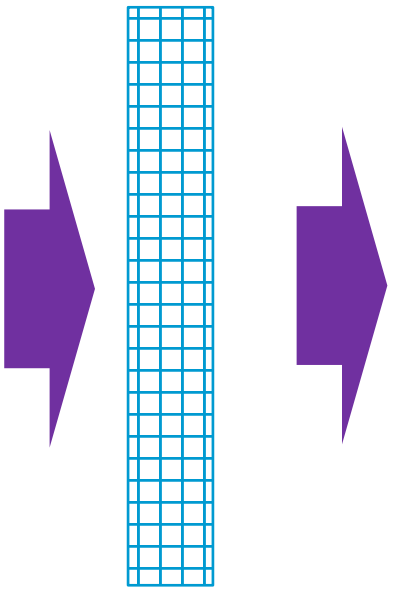
Berg, Kukharenko, Scheffner, & Peter, 2018, *PLOS Comp. Biol.* 14, e1006589

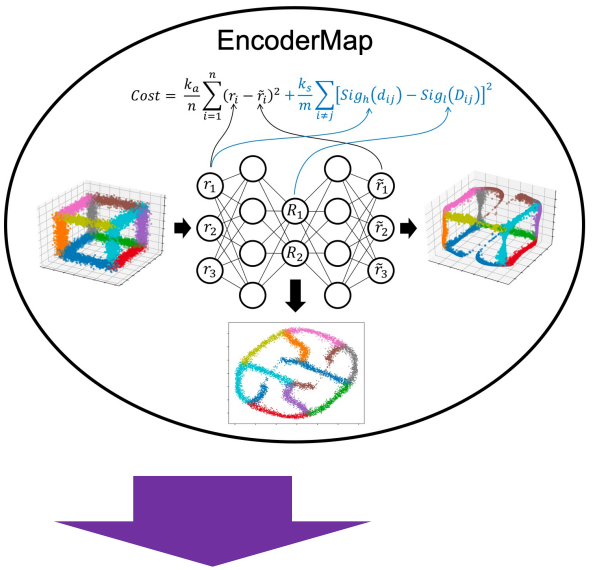
Outline

- Setting the stage
- EncoderMap
  - Learning meaningful representations of conformational phase space
  - Generating protein conformations and visualizing molecular motion
- Extracting meaningful feature sets from graph representations of proteins
  - Generating residue interaction landscapes
- Utilizing low-dimensional embeddings for clustering
  - Identifying conformational states
- Backmapping based sampling
  - Linking scales through low-dimensional representations

Low-dimensional representations of conformational phase space







Simulation data (Cartesian coordinates)

Input data set (features)

internal distances, dihedrals, ...

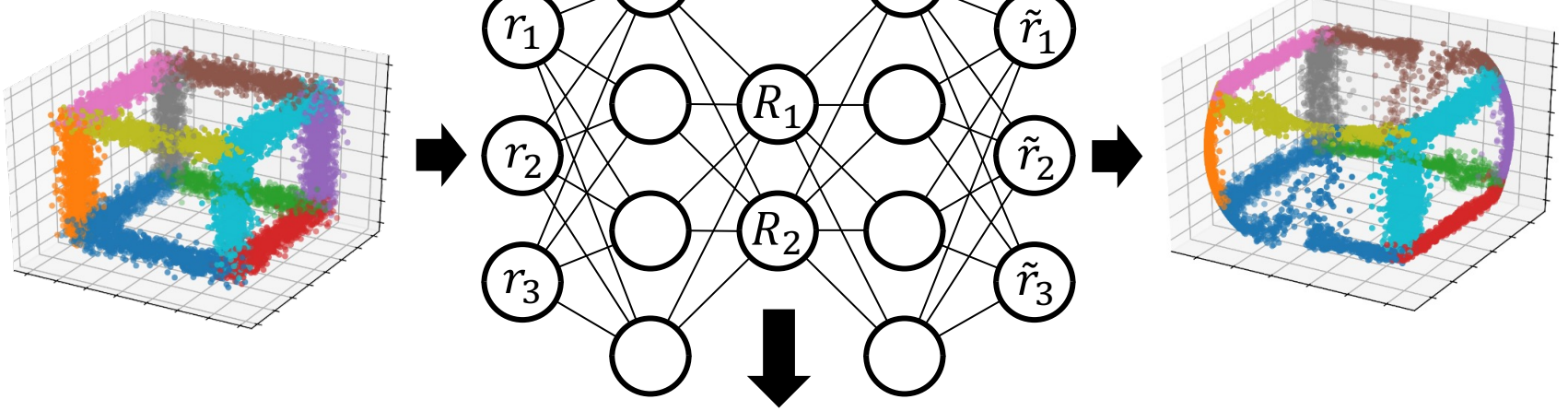
Map in low-dim space

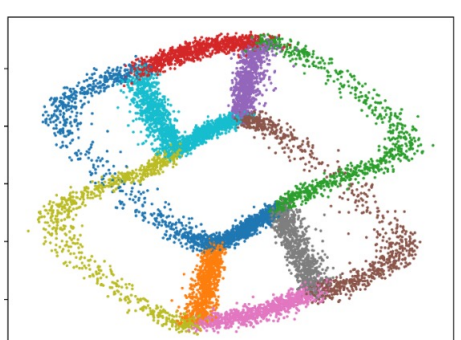
Conformational states

Lemke & Peter, 2019, JCTC, 15, 1209-1215
Lemke, Berg, Jain, Peter, 2019, *J. Chem. Inf. Model,* 59, 4550

github.com/AG-Peter/EncoderMap

$$Cost = \frac{k_a}{n}\sum_{i=1}^{n}(r_i - \tilde{r}_i)^2$$

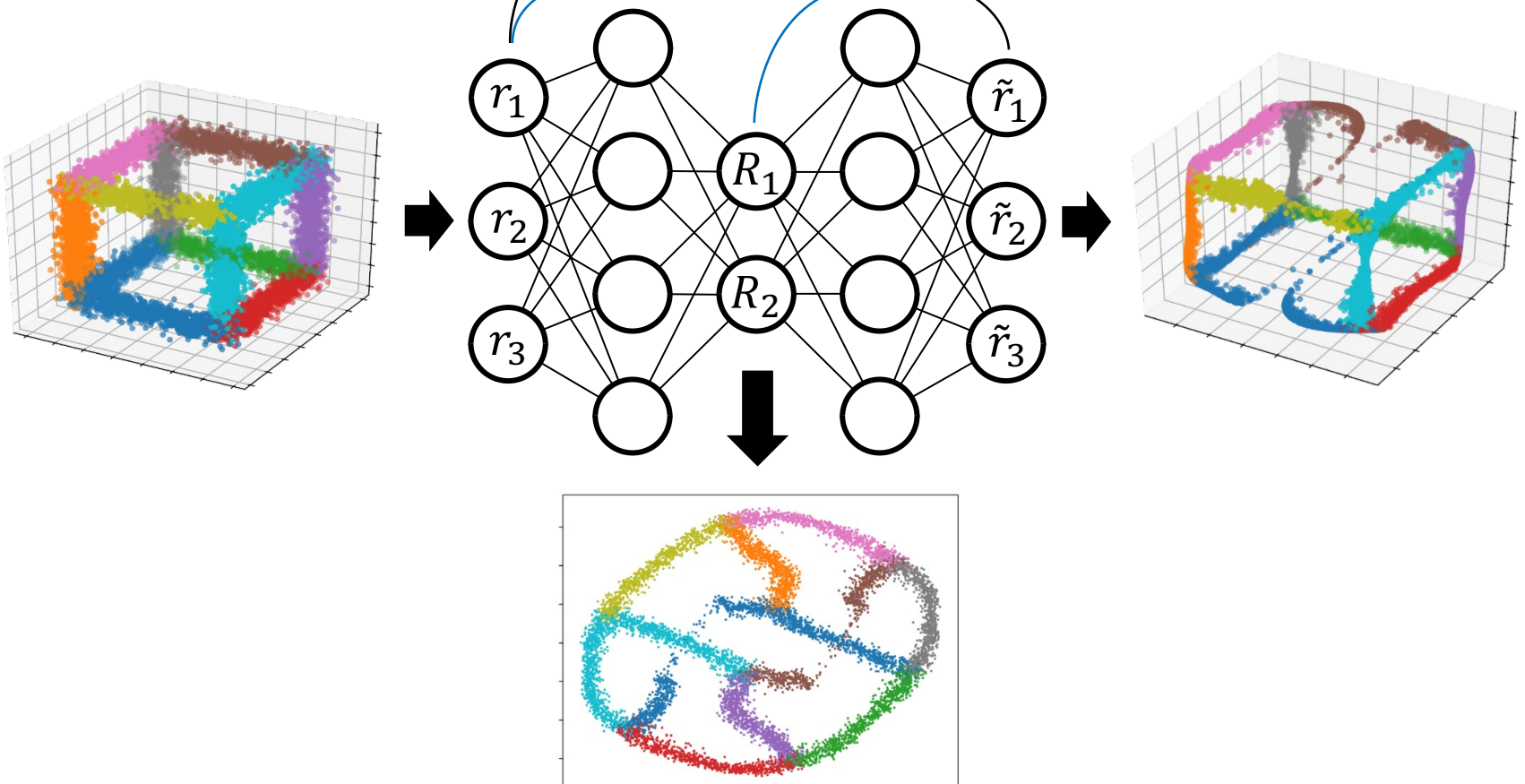




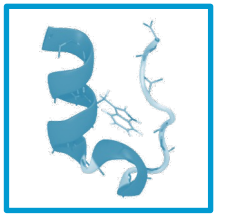
Lemke & Peter, 2019, JCTC, 15, 1209-1215
Lemke, Berg, Jain, Peter, 2019, *J. Chem. Inf. Model,* 59, 4550

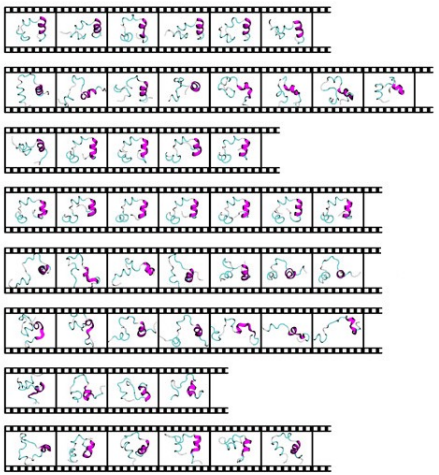
github.com/AG-Peter/EncoderMap

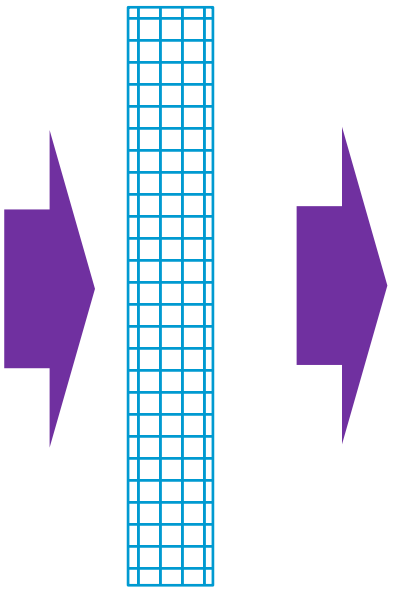
$$Cost = \frac{k_a}{n}\sum_{i=1}^{n}(r_i - \tilde{r}_i)^2 + \frac{k_s}{m}\sum_{i\neq j}\left[Sig_h(d_{ij}) - Sig_l(D_{ij})\right]^2$$

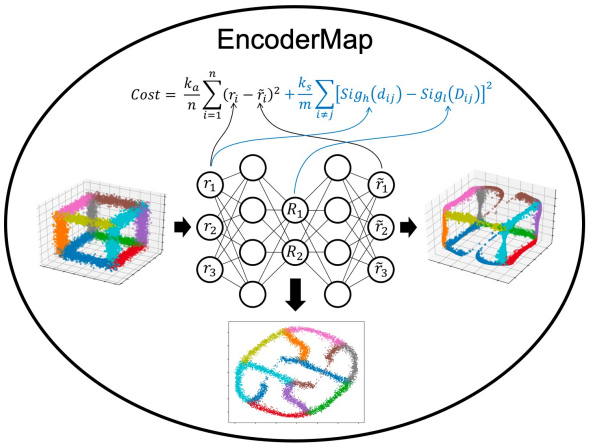


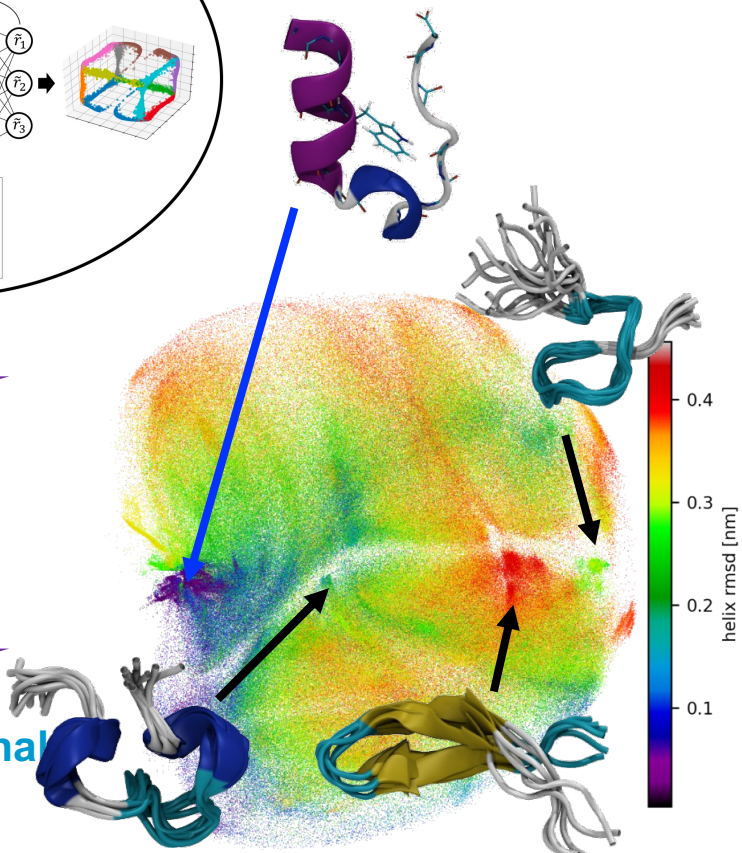
EncoderMap example 1: Trp-cage











Simulation data (Cartesian coordinates)

Input data set (features)

internal distances, dihedrals, ...

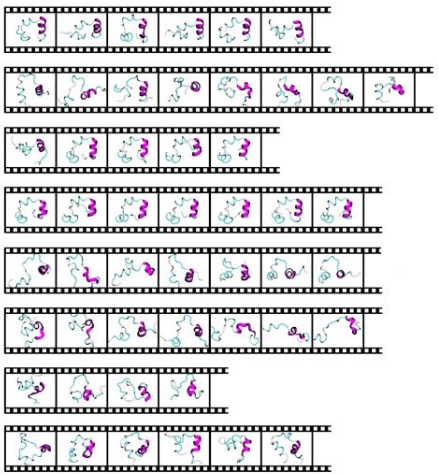
Map in low-dim space

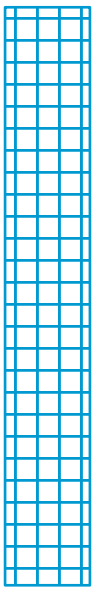
Conformational states

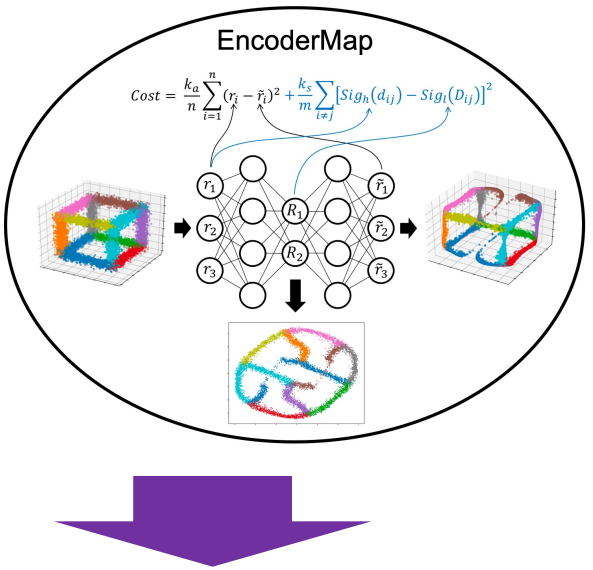
Lemke & Peter, 2019, JCTC, 15, 1209-1215
Lemke, Berg, Jain, Peter, 2019, *J. Chem. Inf. Model,* 59, 4550

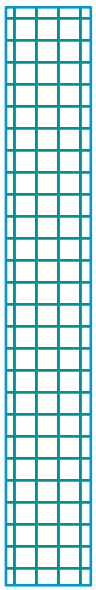
github.com/AG-Peter/EncoderMap

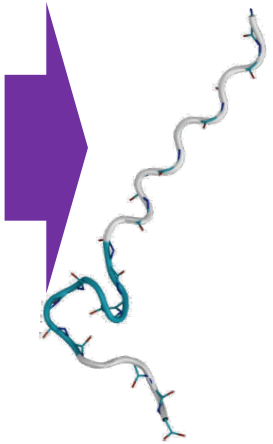
EncoderMap example 1: Trp-cage











(Cartesian) Structure reconstruction

Simulation data (Cartesian coordinates)

Input data set (features)

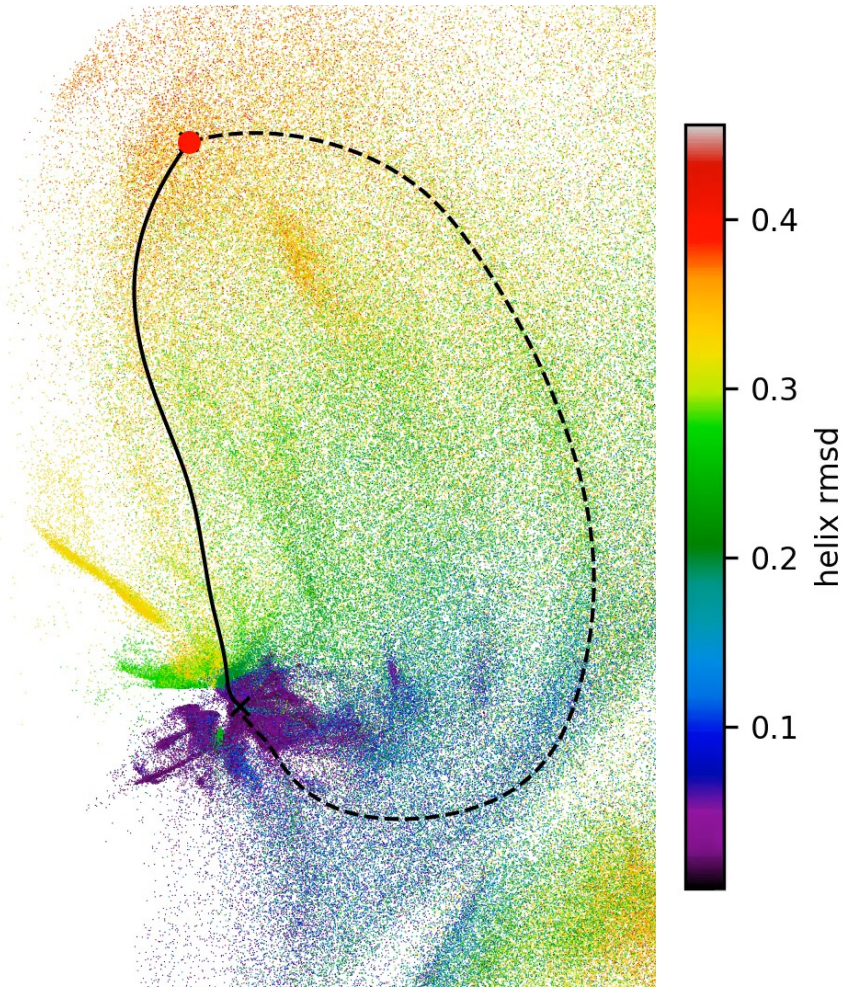
internal distances, dihedrals, ...

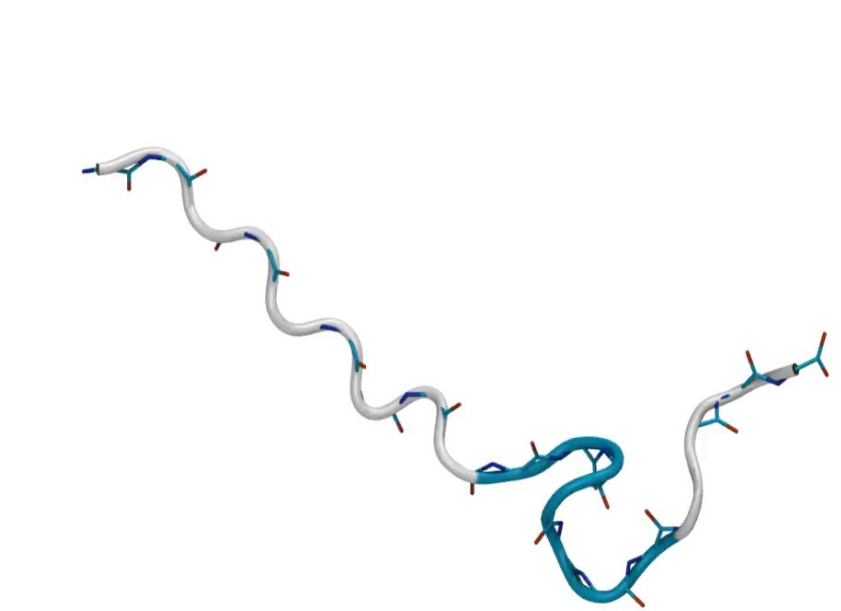
Map in low-dim space

Conformational states

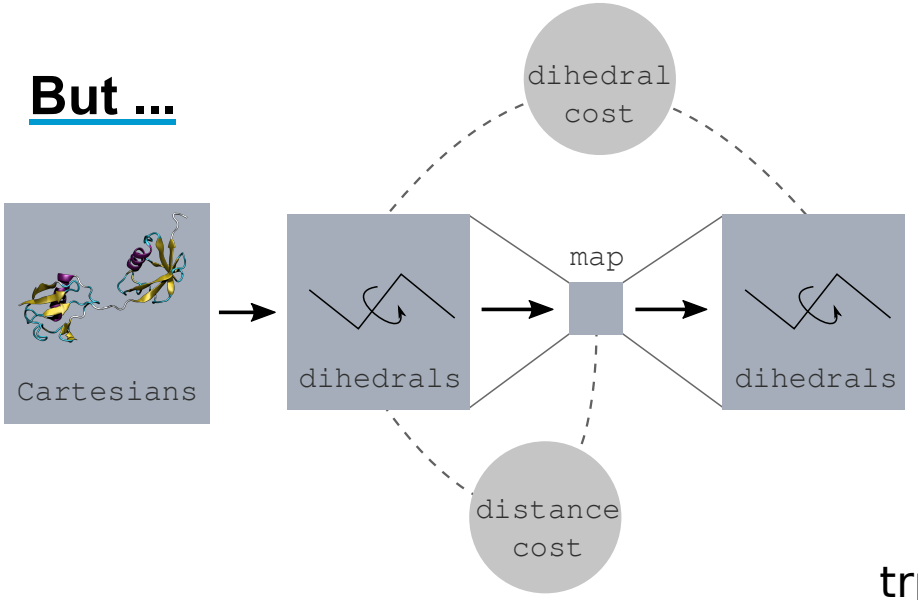
Generated data (in feature space)

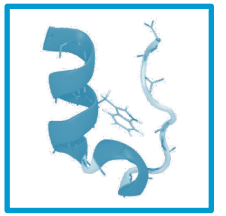
dihedrals, ...

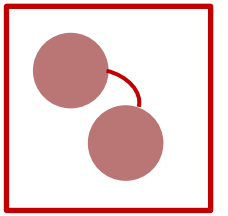




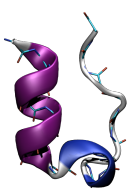
But ...



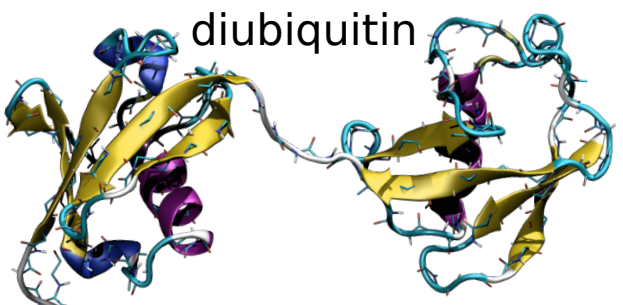




trp-cage

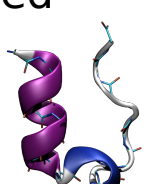


diubiquitin



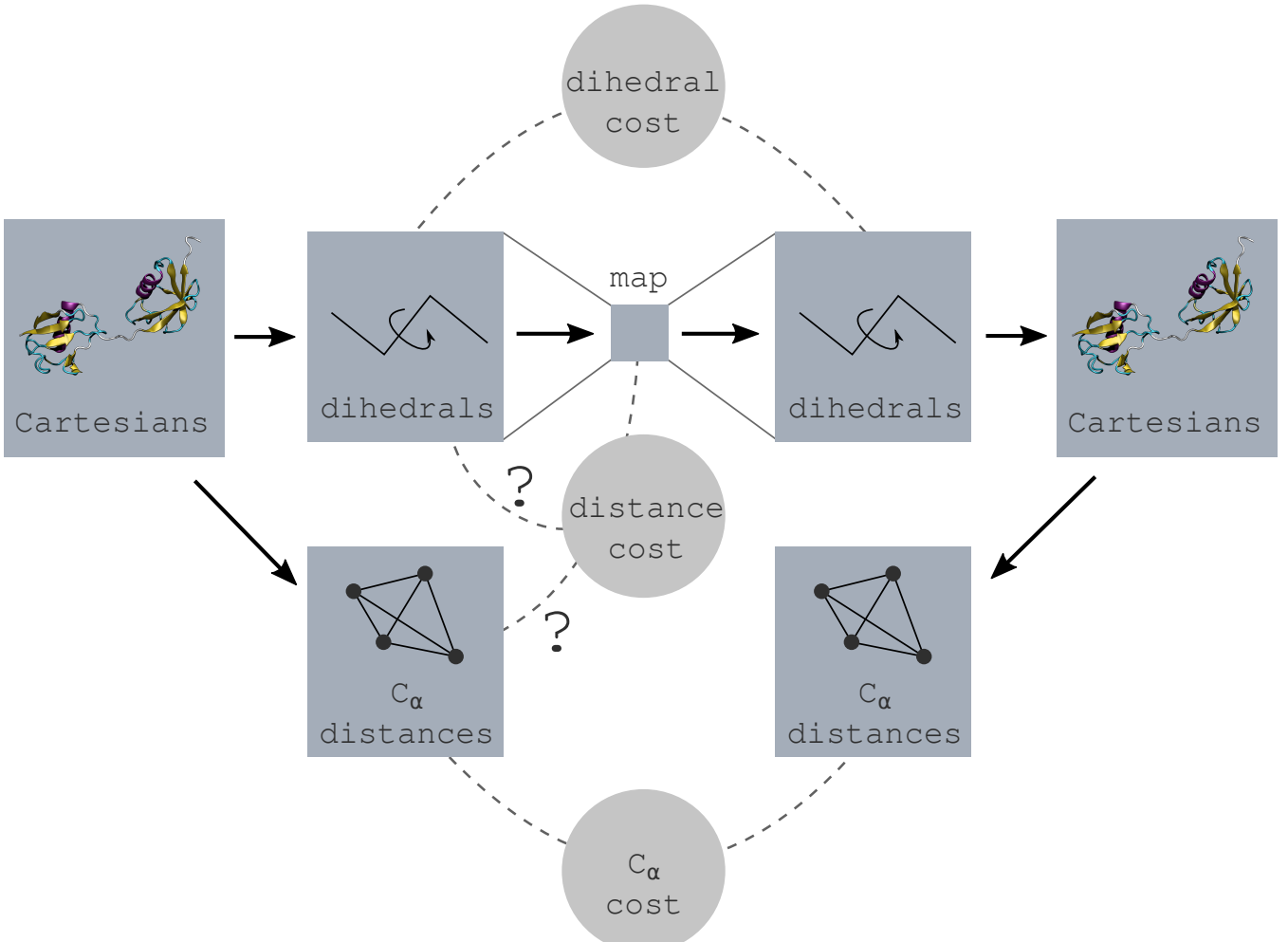
original

generated

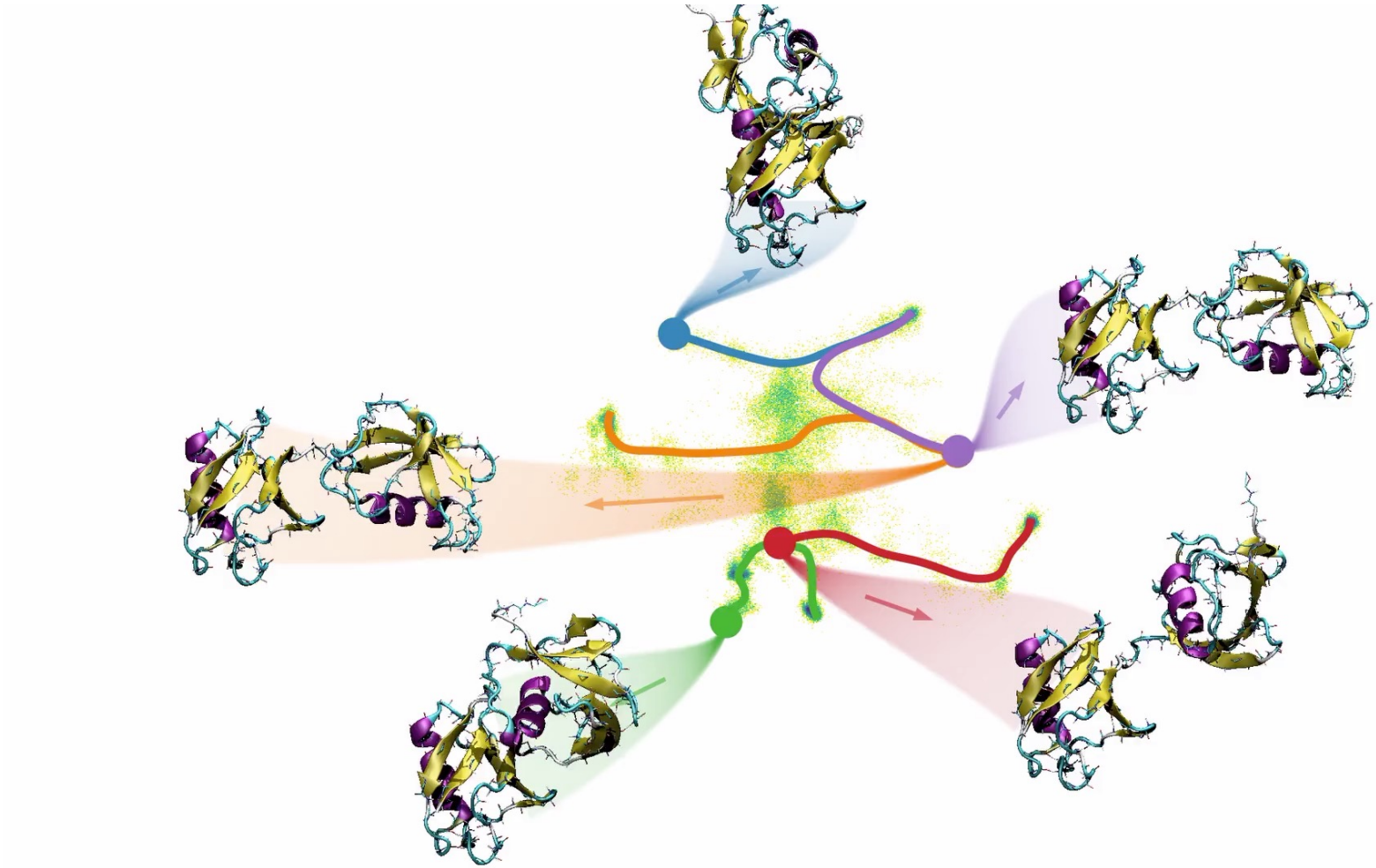


EncoderMap (II)

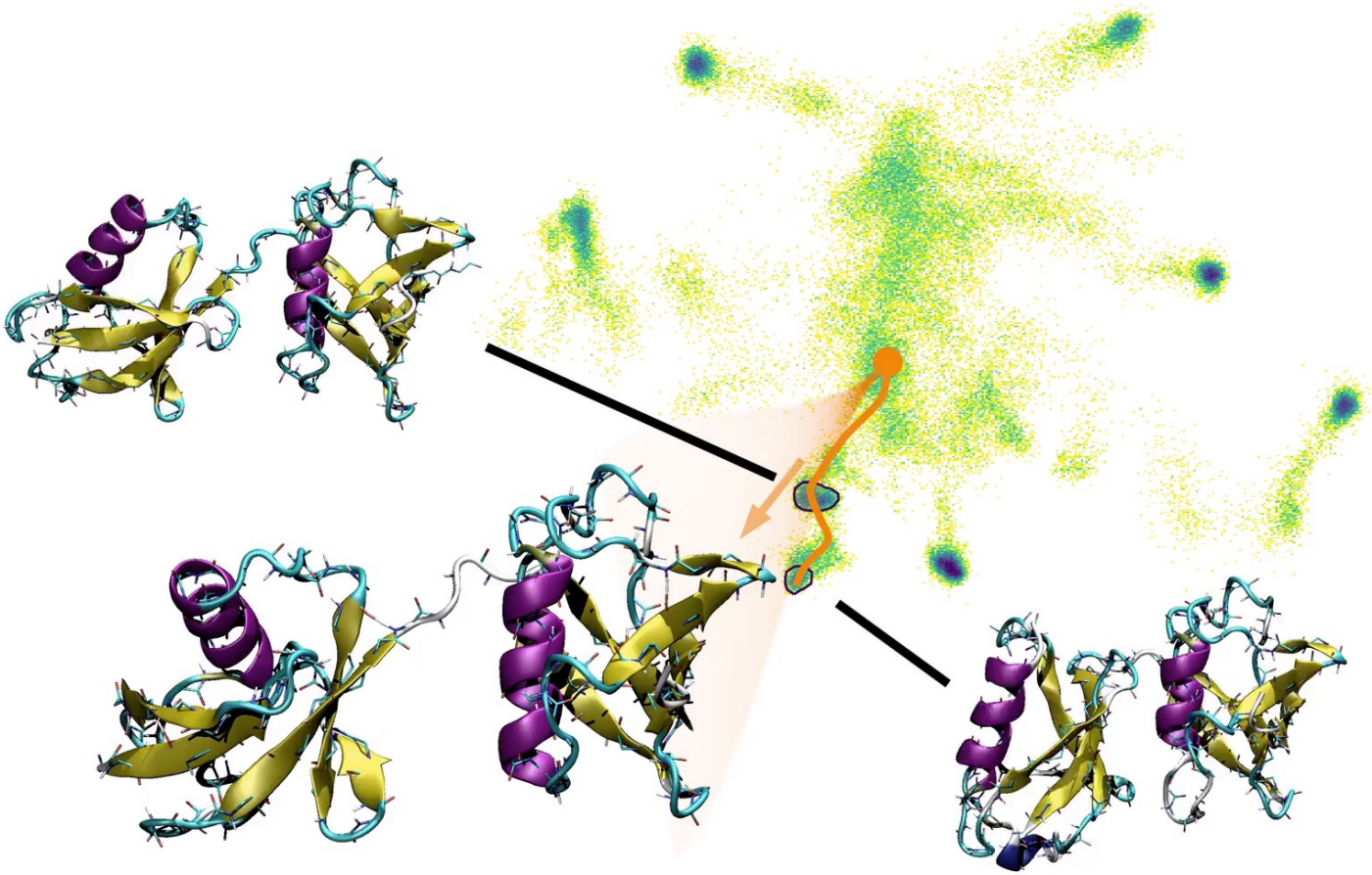
$$C = k_{\mathrm{dih.}} C_{\mathrm{dihedral}} + k_{\mathrm{C}_\alpha} C_{\mathrm{C}_\alpha} + k_{\mathrm{dist.}} C_{\mathrm{distance}}$$



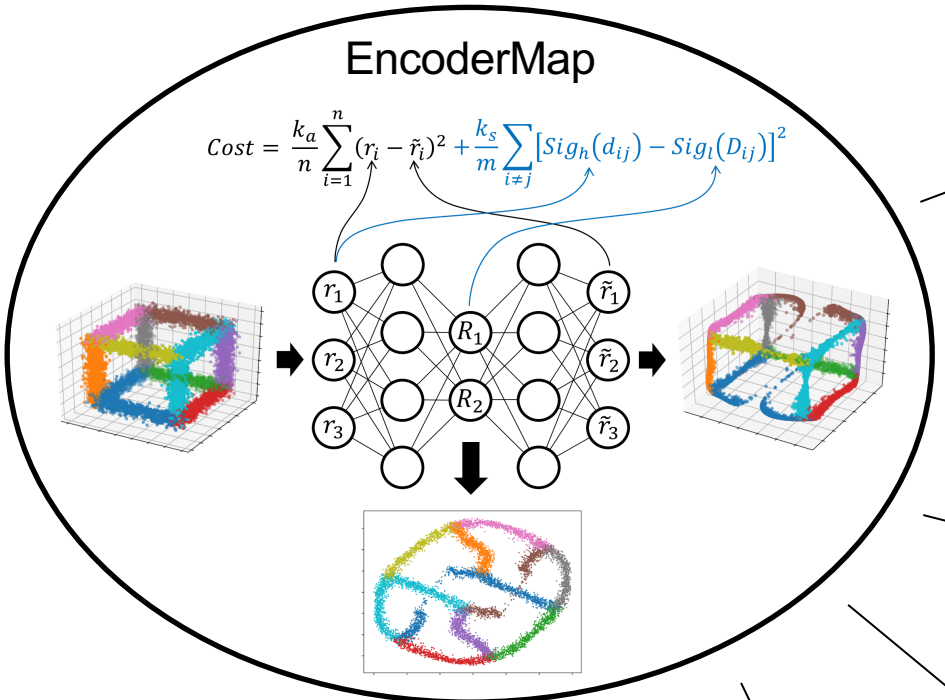
Example 2: M1-linked di-Ubiquitin – generating paths



Example 2: M1-linked di-Ubiquitin – visualizing important motions



Lemke, Berg, Jain, Peter, 2019, *J. Chem. Inf. Model,* 59, 4550



Computationally efficient

No landmarks required

Yields a function to map from high-d to low-d

Yields a function to map from low-d to high-d

Can deal with nonlinear data/structures

Preserves distances

➔ EncoderMap: combines advantages of a NN autoencoder and SketchMap/MDS

Outline

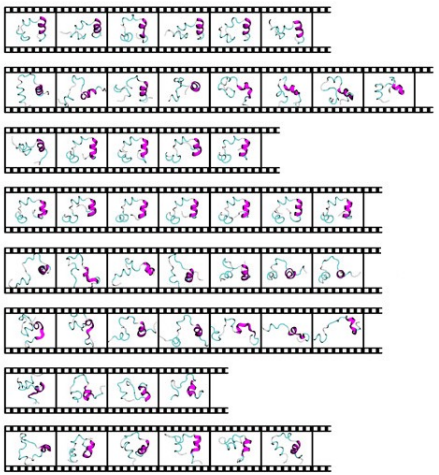
- Setting the stage
- EncoderMap
  - Learning meaningful representations of conformational phase space
  - Generating protein conformations and visualizing molecular motion
- Extracting meaningful feature sets from graph representations of proteins
  - Generating residue interaction landscapes
- Utilizing low-dimensional embeddings for clustering
  - Identifying conformational states
- Backmapping based sampling
  - Linking scales through low-dimensional representations

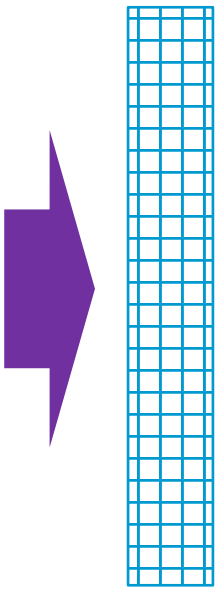
➔ the input features matter

Outline

- Setting the stage
- EncoderMap
  - Learning meaningful representations of conformational phase space
  - Generating protein conformations and visualizing molecular motion
- Extracting meaningful feature sets from graph representations of proteins
  - Generating residue interaction landscapes
- Utilizing low-dimensional embeddings for clustering
  - Identifying conformational states
- Backmapping based sampling
  - Linking scales through low-dimensional representations

Extracting meaningful feature sets from protein graph representations



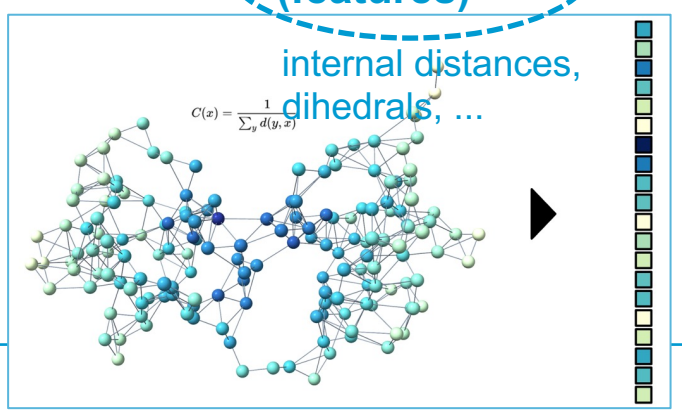


Dimensionality reduction
PCA, MDS, EncoderMap ...

Simulation data (Cartesian coordinates)

Input data set (features)

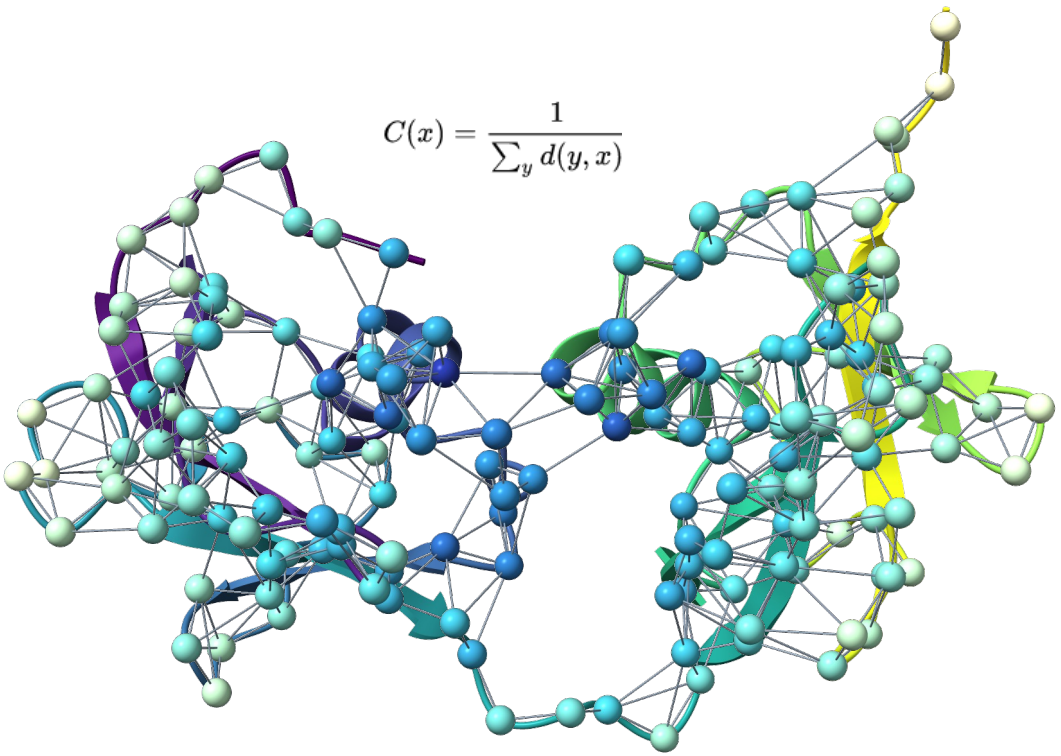
internal distances, dihedrals, ...

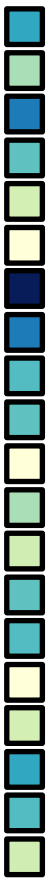


Map in low-dim space

Conformational states

Extracting meaningful feature sets from protein graph representations

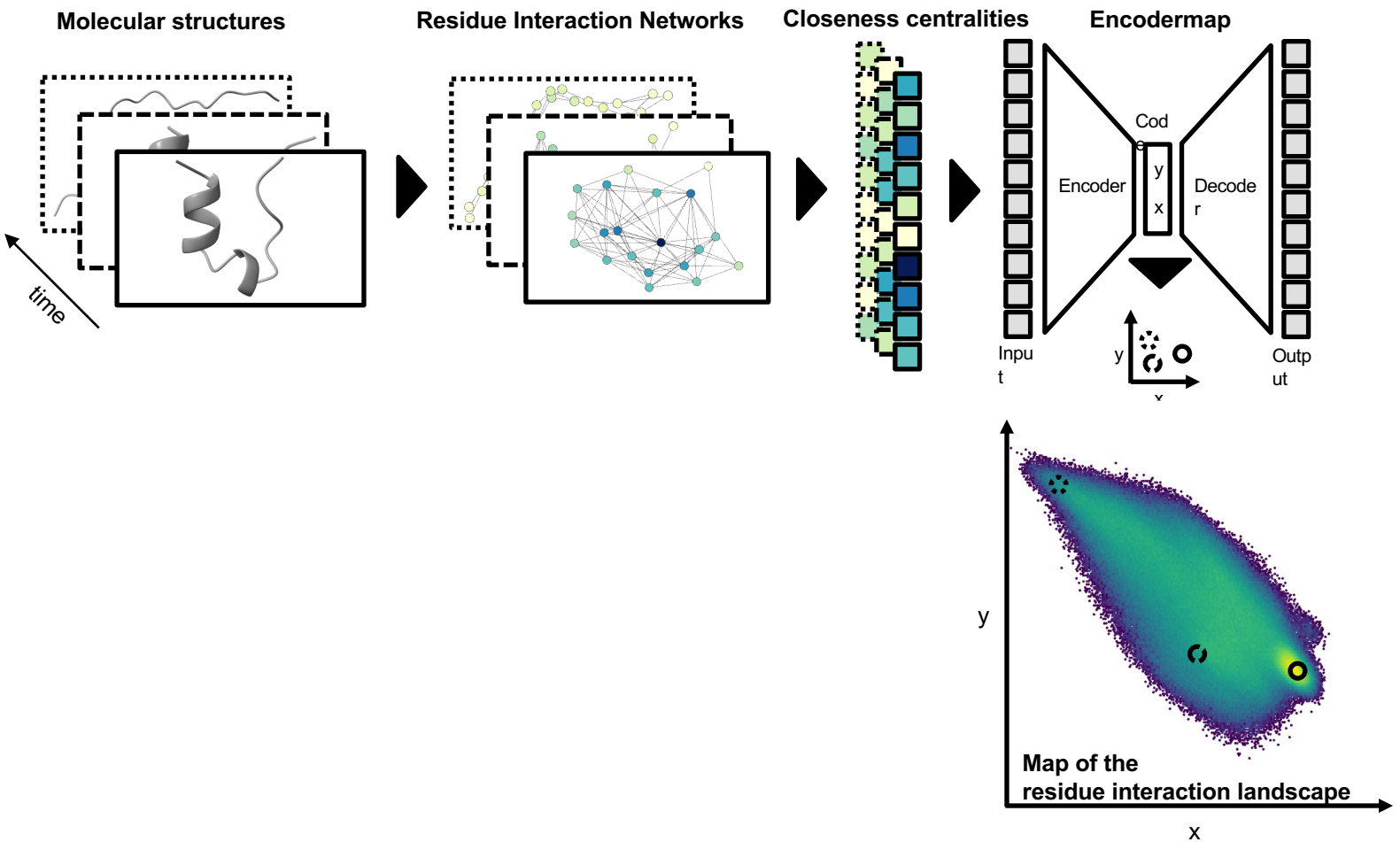




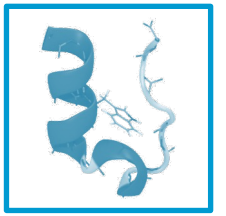
➔ Closeness centrality: „How close is a node to all other nodes in the network?“

➔ Feature set that captures role of each residue in the protein structure

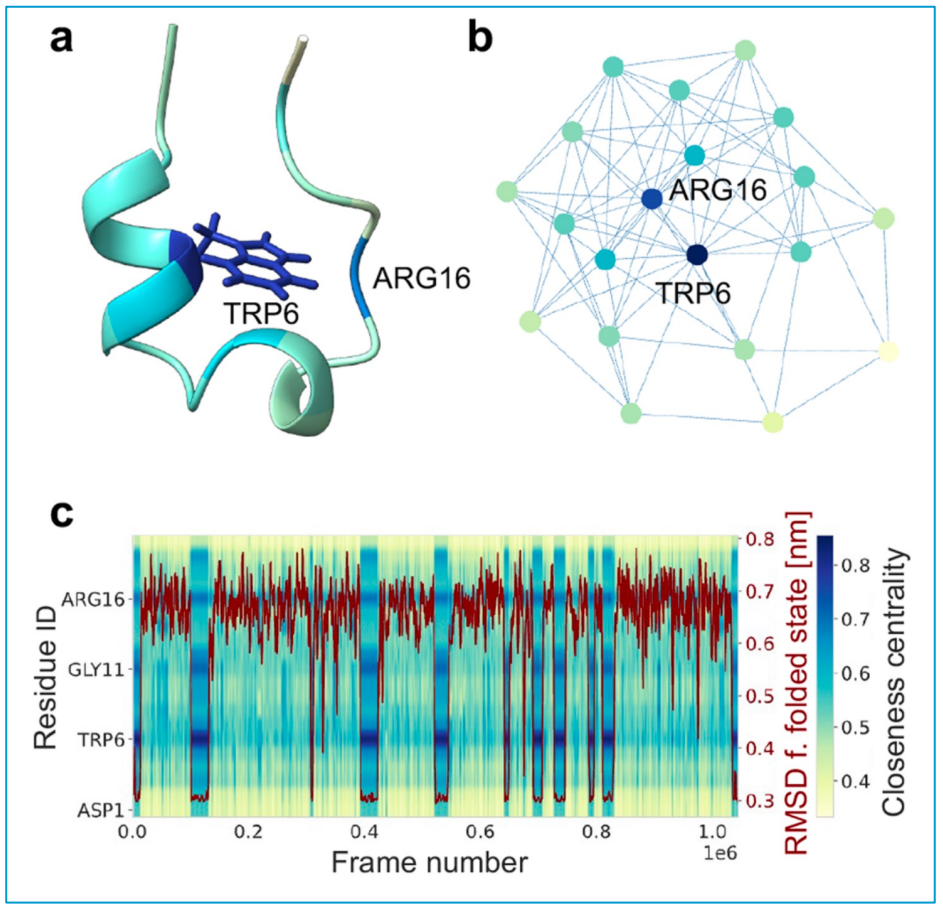
Workflow


Molecular structures
Residue Interaction Networks
Closeness centralities
Encodermap
time
Encoder
Cod
e
y
x
Decode
r
Inpu
t
Outp
ut
y
Map of the
residue interaction landscape
y
x

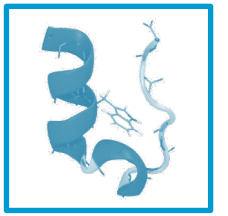
Example 1: Trp-cage

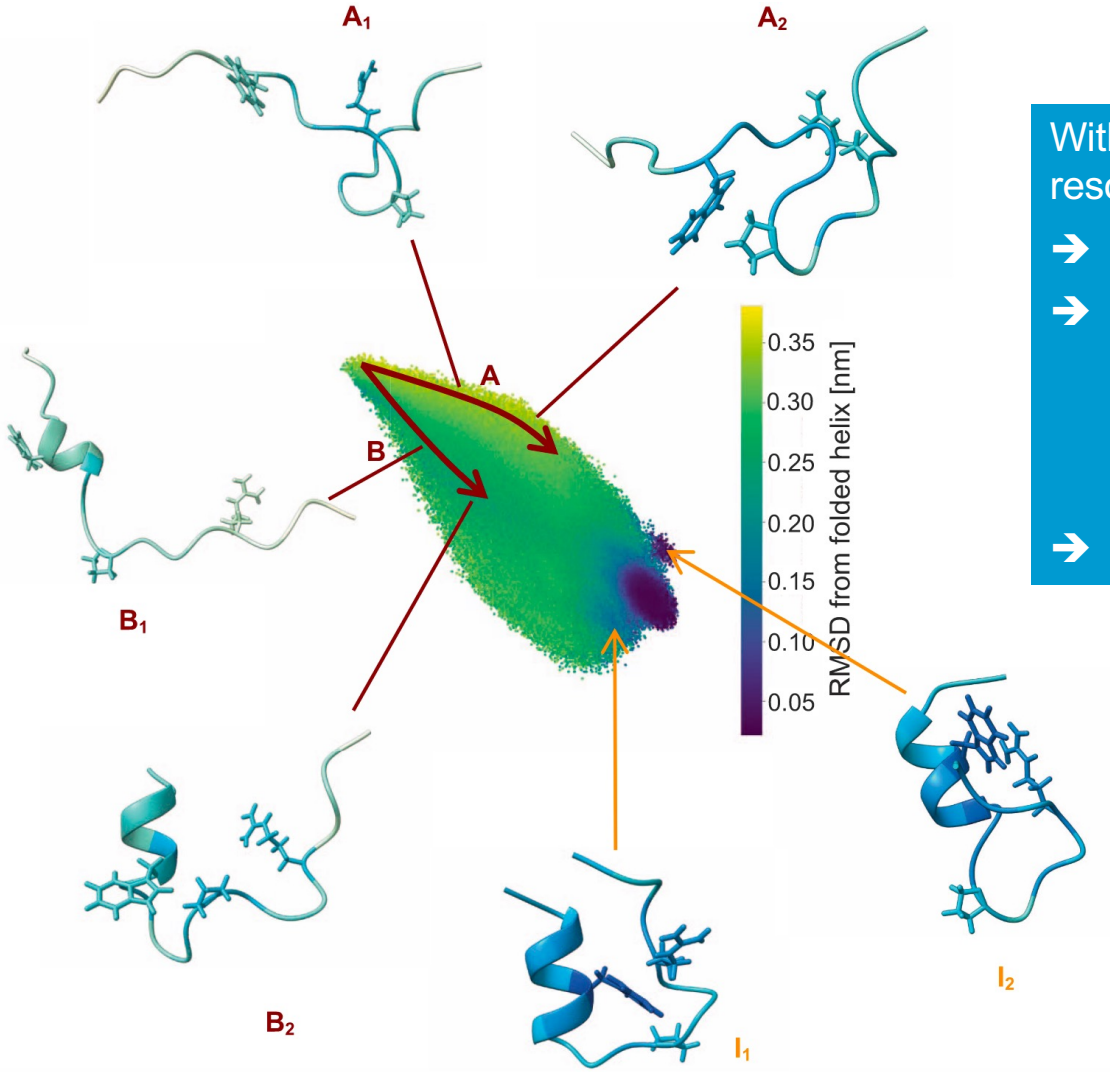


Trp-Cage



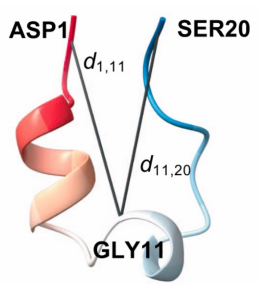
Residue interaction landscape of Trp-cage



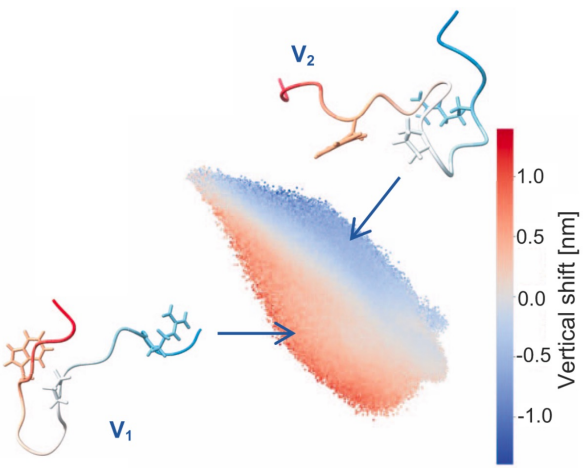


Without protein-specific prior input, the map resolves known folding behaviors

- ➔ Separation of folded and unfolded states
- ➔ Two folding pathways:
  - A: Nucleation-condensation; hydrophobic collapse first
  - B: Diffusion-collision; helix forms first
- ➔ Near-native intermediates



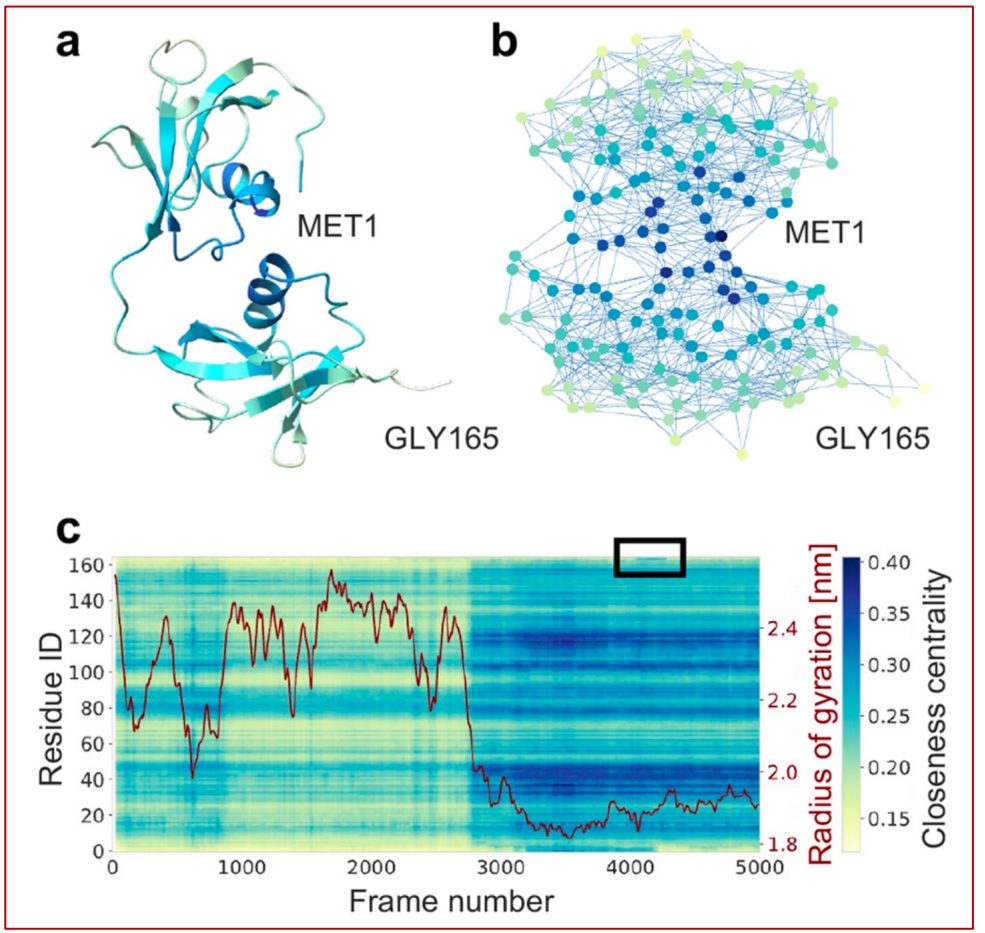
Vertical shift: $d_{11,20}$-$d_{1,10}$



Example 2: FAT10

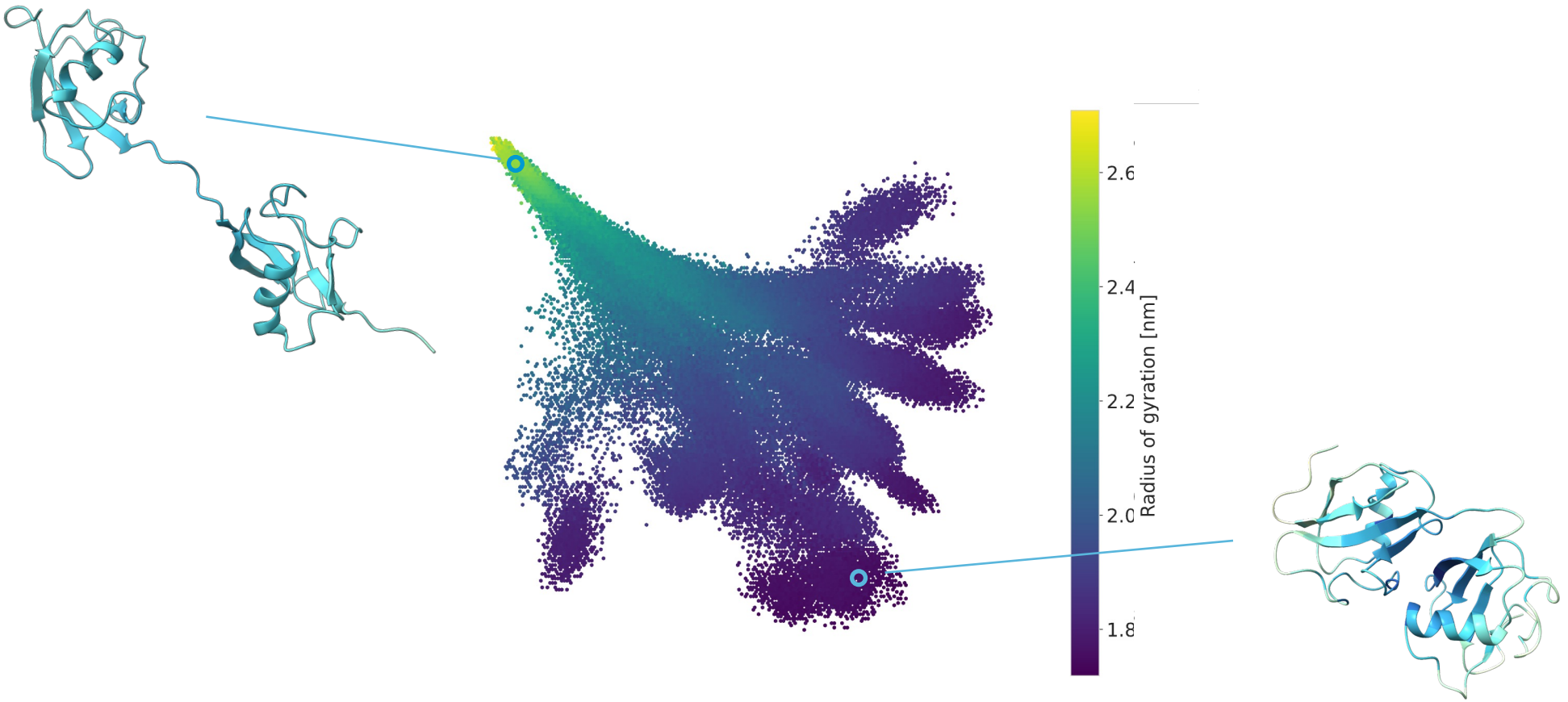
Trp-Cage

FAT10

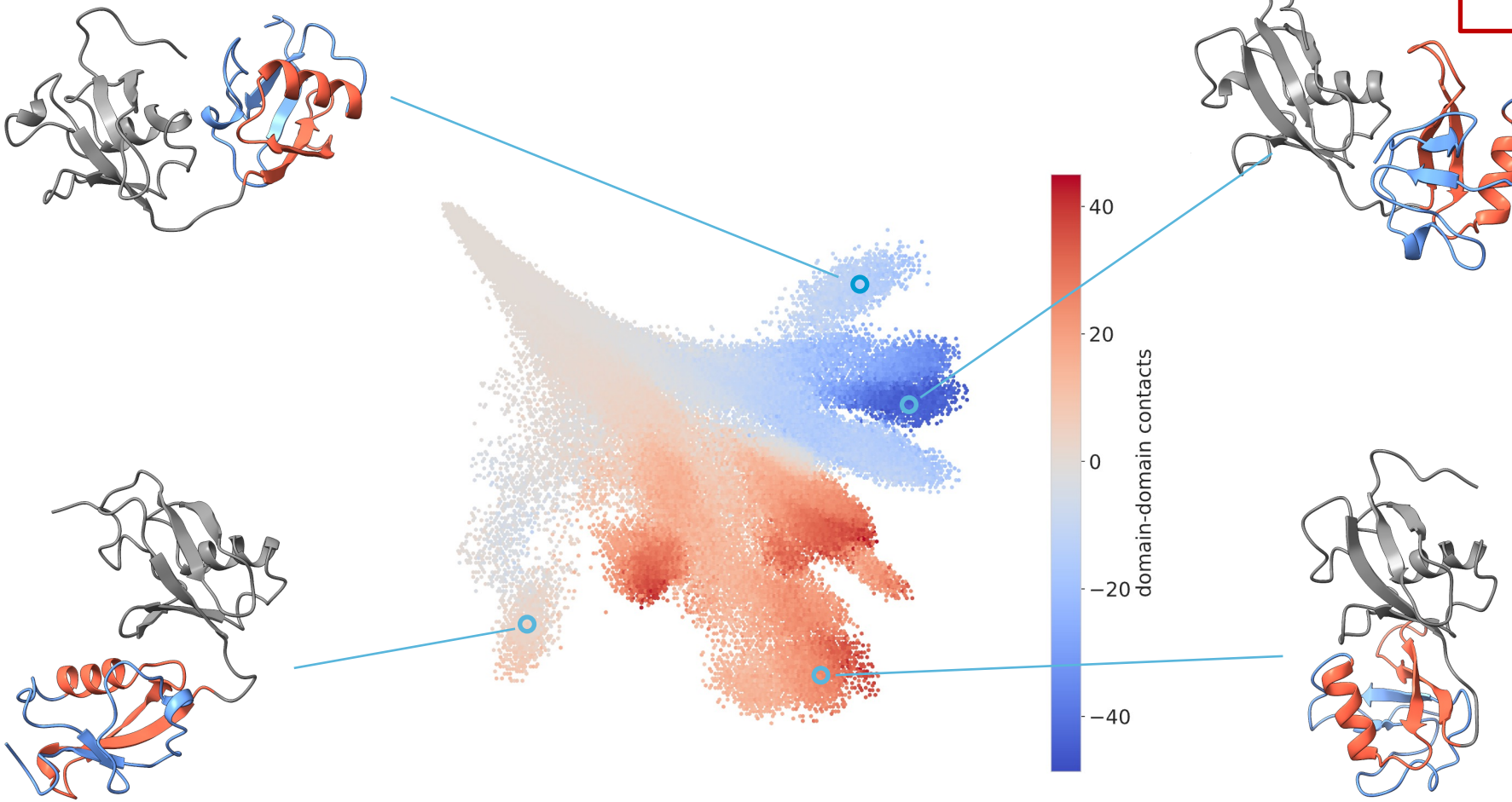


Franke & Peter, 2023, JCTC, https://doi.org/10.1021/acs.jctc.2c01228

Residue interaction landscape of FAT10

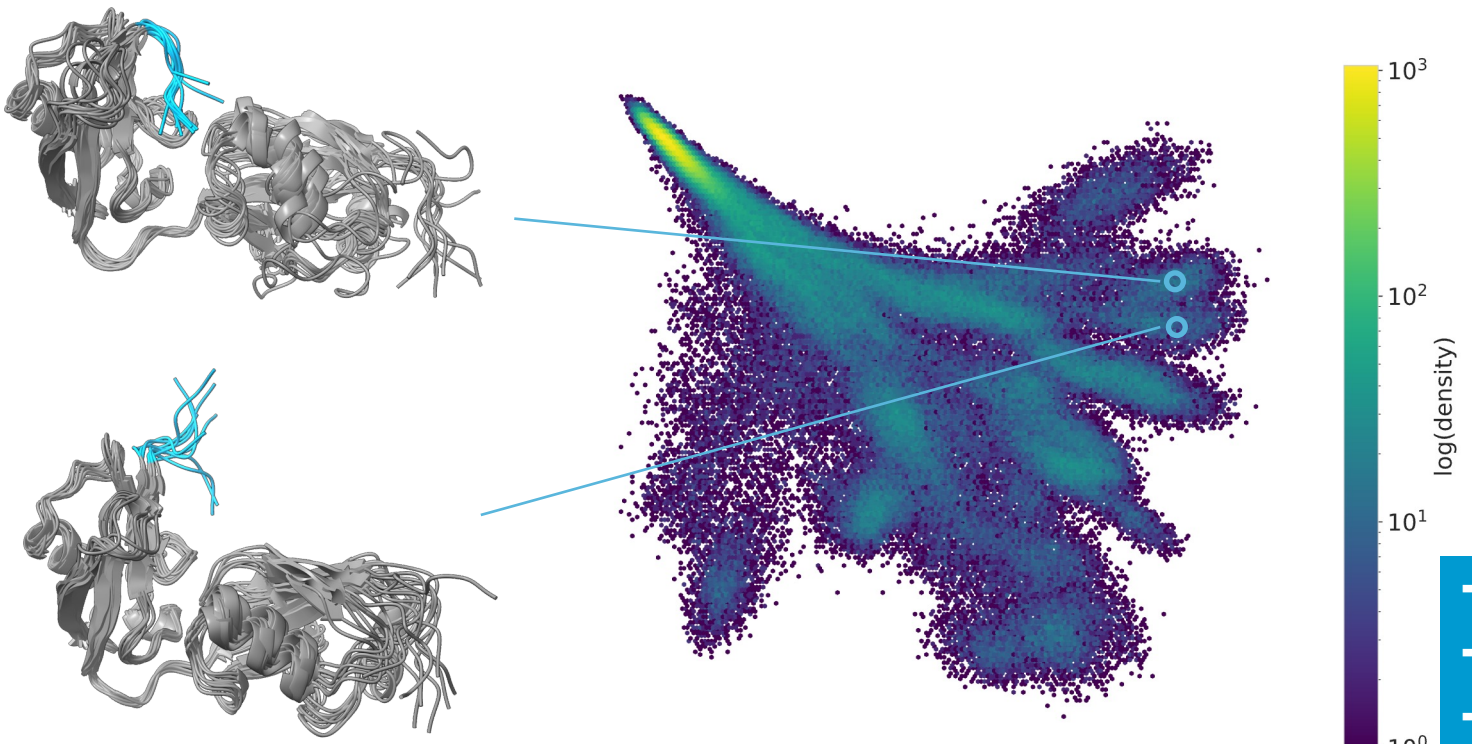


Residue interaction landscape of FAT10



Franke & Peter, 2023, JCTC, https://doi.org/10.1021/acs.jctc.2c01228

Residue interaction landscape of FAT10

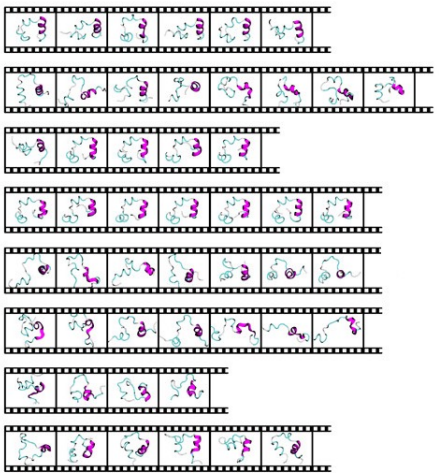


- ➔ global interpretability
- ➔ resolves meaningful details
- ➔ useful for visualization and further down-stream processing

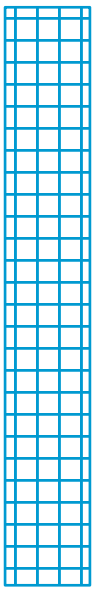
Outline

- Setting the stage
- EncoderMap
  - Learning meaningful representations of conformational phase space
  - Generating protein conformations and visualizing molecular motion
- Extracting meaningful feature sets from graph representations of proteins
  - Generating residue interaction landscapes
- Utilizing low-dimensional embeddings for clustering
  - Identifying conformational states
- Backmapping based sampling
  - Linking scales through low-dimensional representations

Clustering



Simulation data (Cartesian coordinates)



Input data set (features)

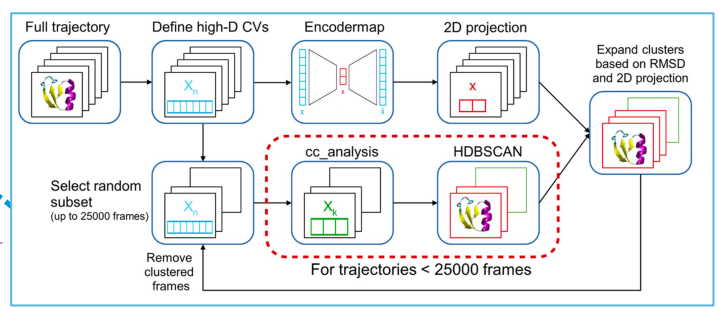
internal distances, dihedrals, ...

Dimensionality reduction
PCA, MDS, EncoderMap ...

Map in low-dim space

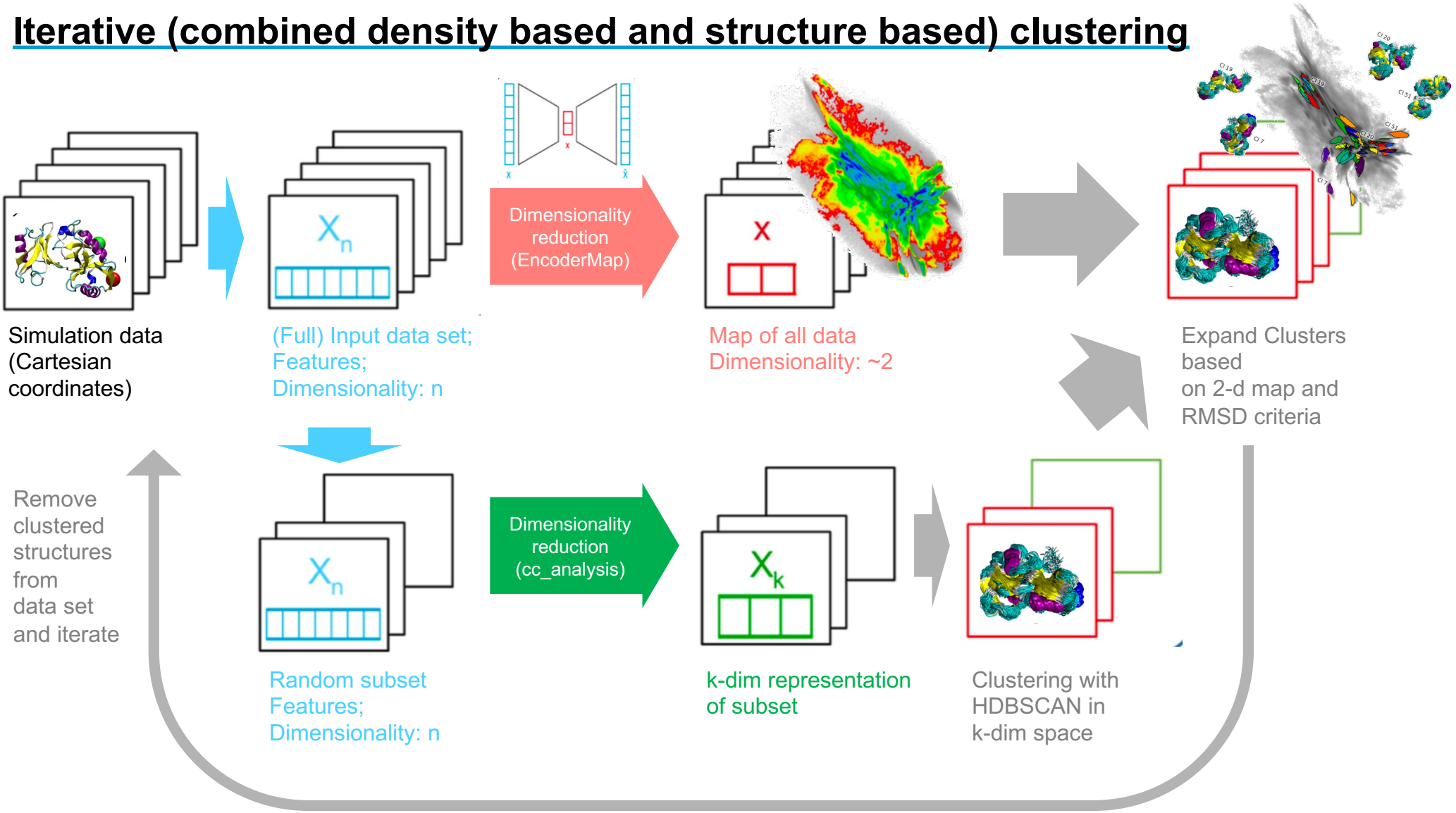
Conformational states

Efficient iterative (combined density-based and structure-based) clustering

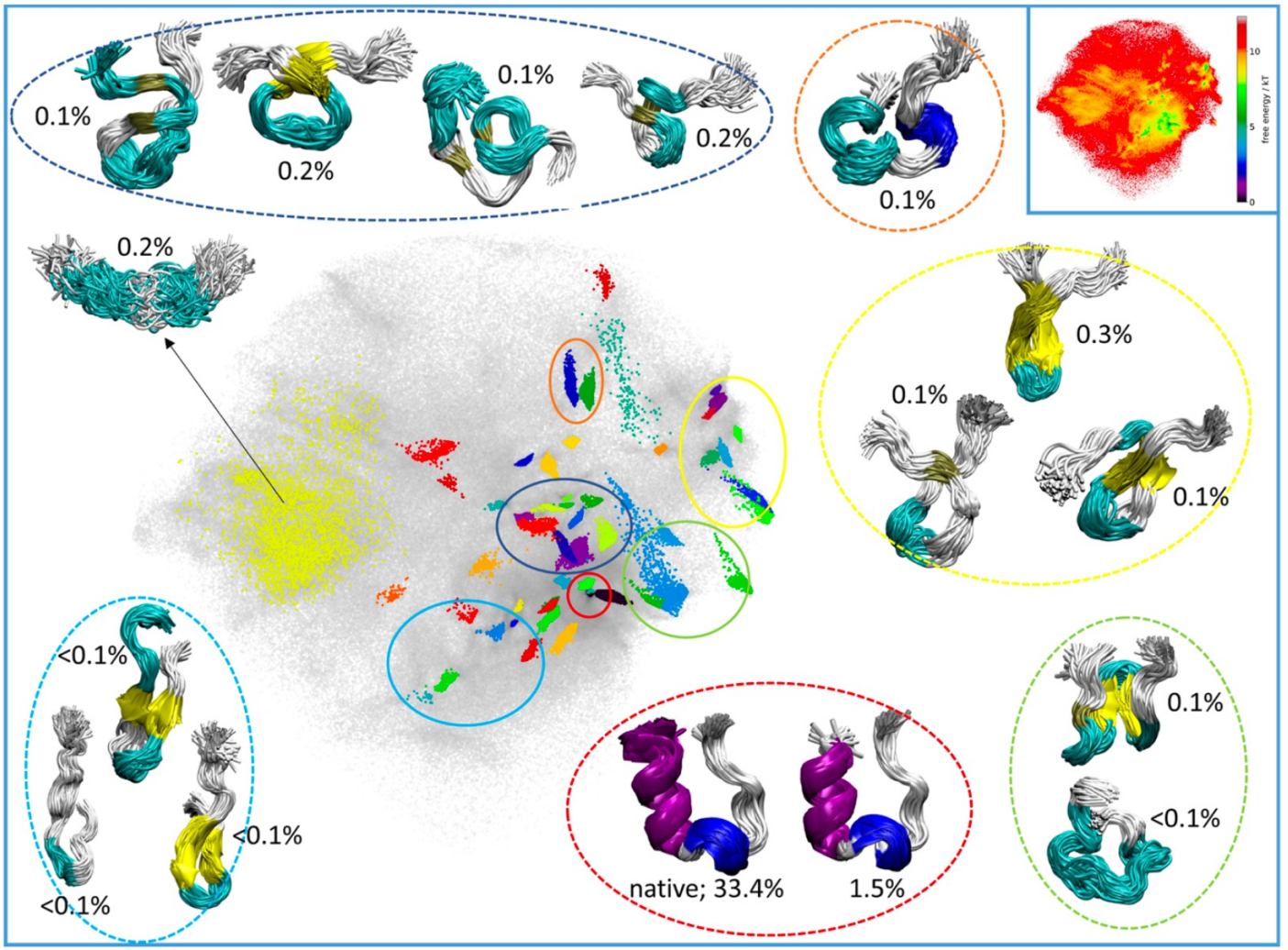


Hunkler et al
J. Chem. Phys. 158, 144109 (2023)

Iterative (combined density based and structure based) clustering

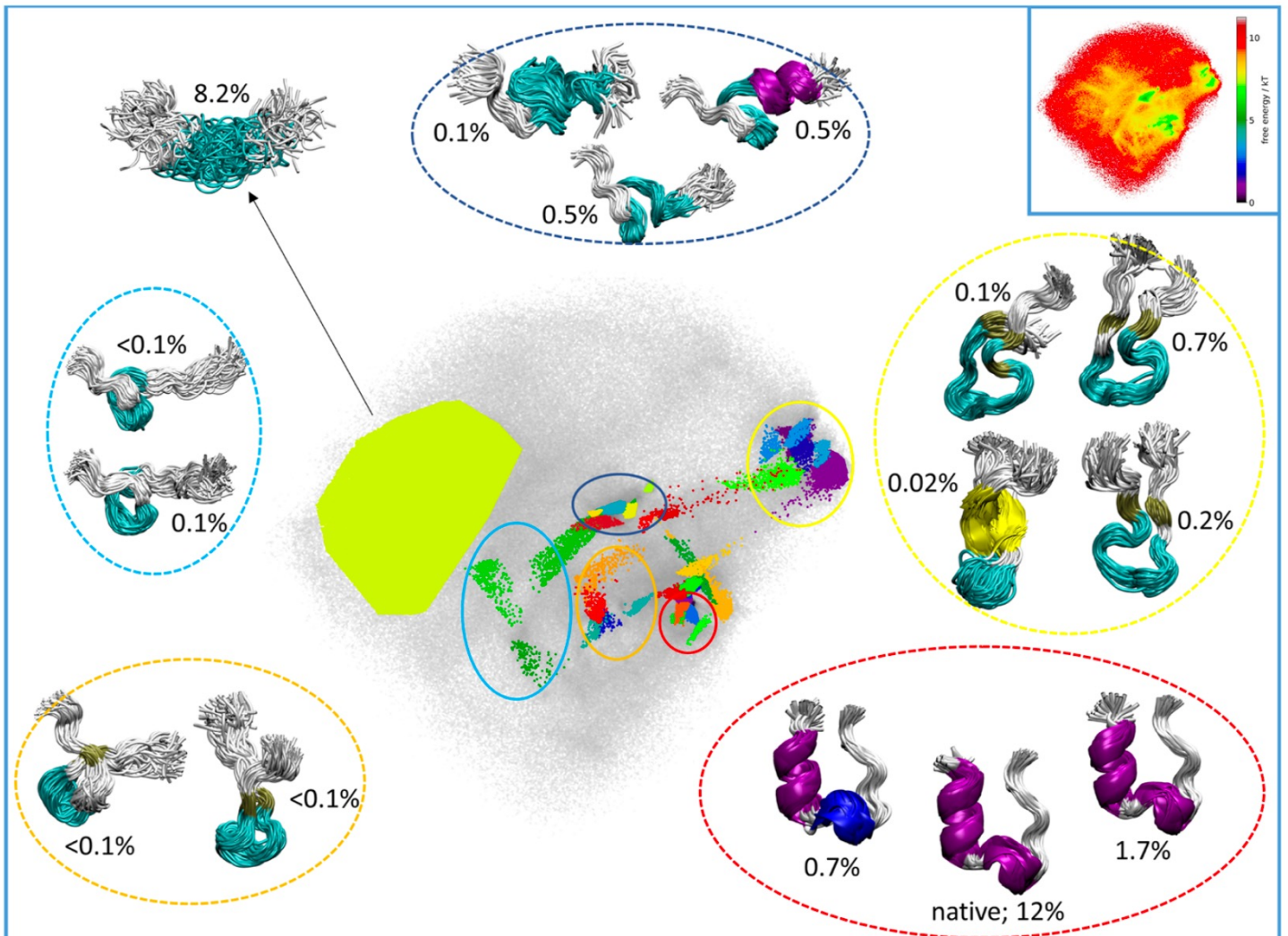


Example 1: Trp-cage



TC5b (40 temperature RE trajectories)

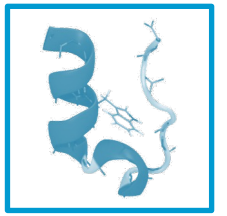
Example 1: Trp-cage



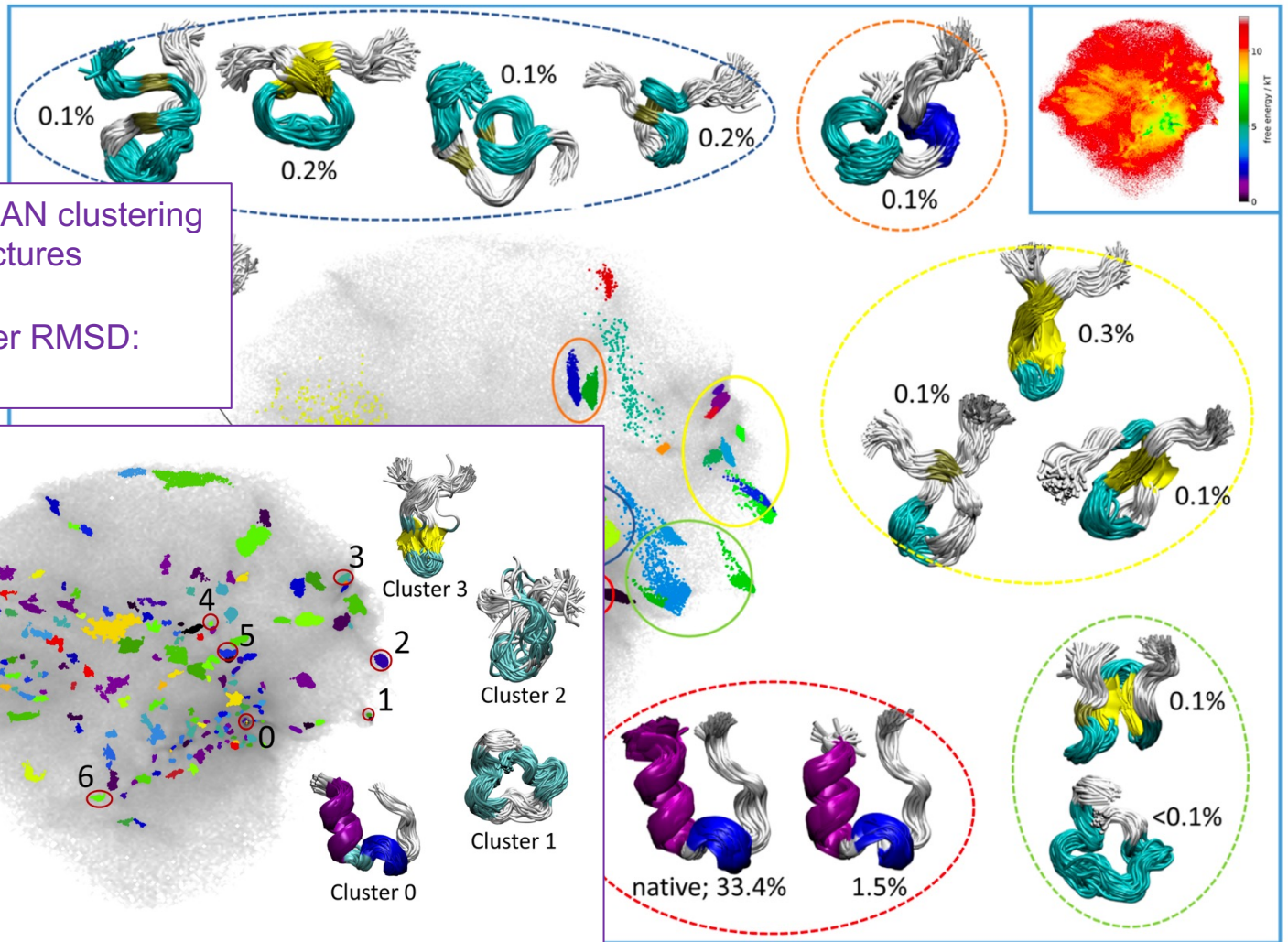
TC10b
(DE Shaw trajectories)

Hunkler et al J. Chem. Phys. 158, 144109 (2023)

Example 1: Trp-cage



direct HDBSCAN clustering
< 15% of structures assigned;
average cluster RMSD:
2.25 Å

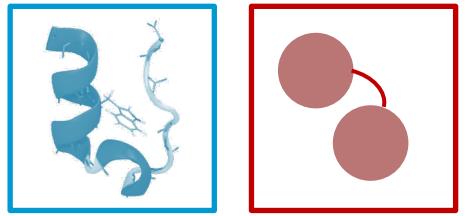


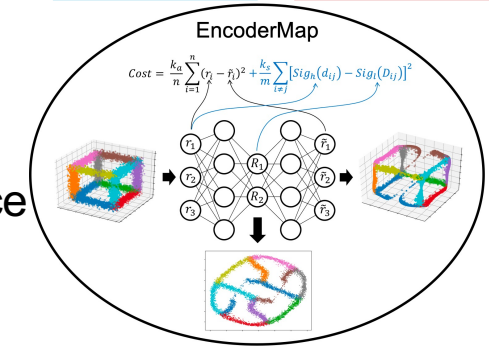
~60% of structures assigned;
average cluster RMSD:
1.34 Å

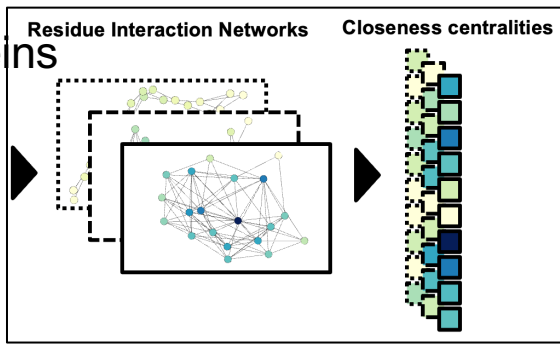
TC5b (40 temperature RE trajectories)

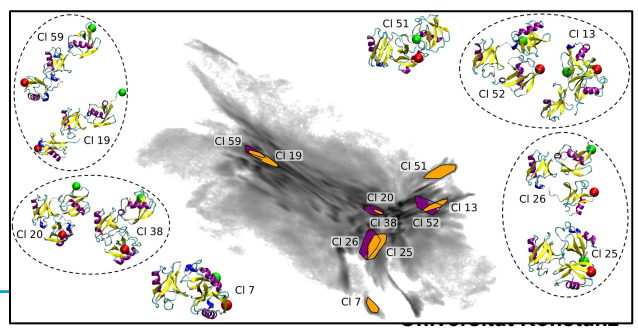
Outline

- Setting the stage
- EncoderMap
  - Learning meaningful representations of conformational phase space
  - Generating protein conformations and visualizing molecular motion
- Extracting meaningful feature sets from graph representations of proteins
  - Generating residue interaction landscapes
- Utilizing low-dimensional embeddings for clustering
  - Identifying conformational states
- Backmapping based sampling
  - Linking scales through low-dimensional representations









Acknowlegements