Defining Conformational States of Proteins Using Dimensionality Reduction and Clustering Algorithms

Eugene Klyshko
Supervisor: Dr. Sarah Rauscher

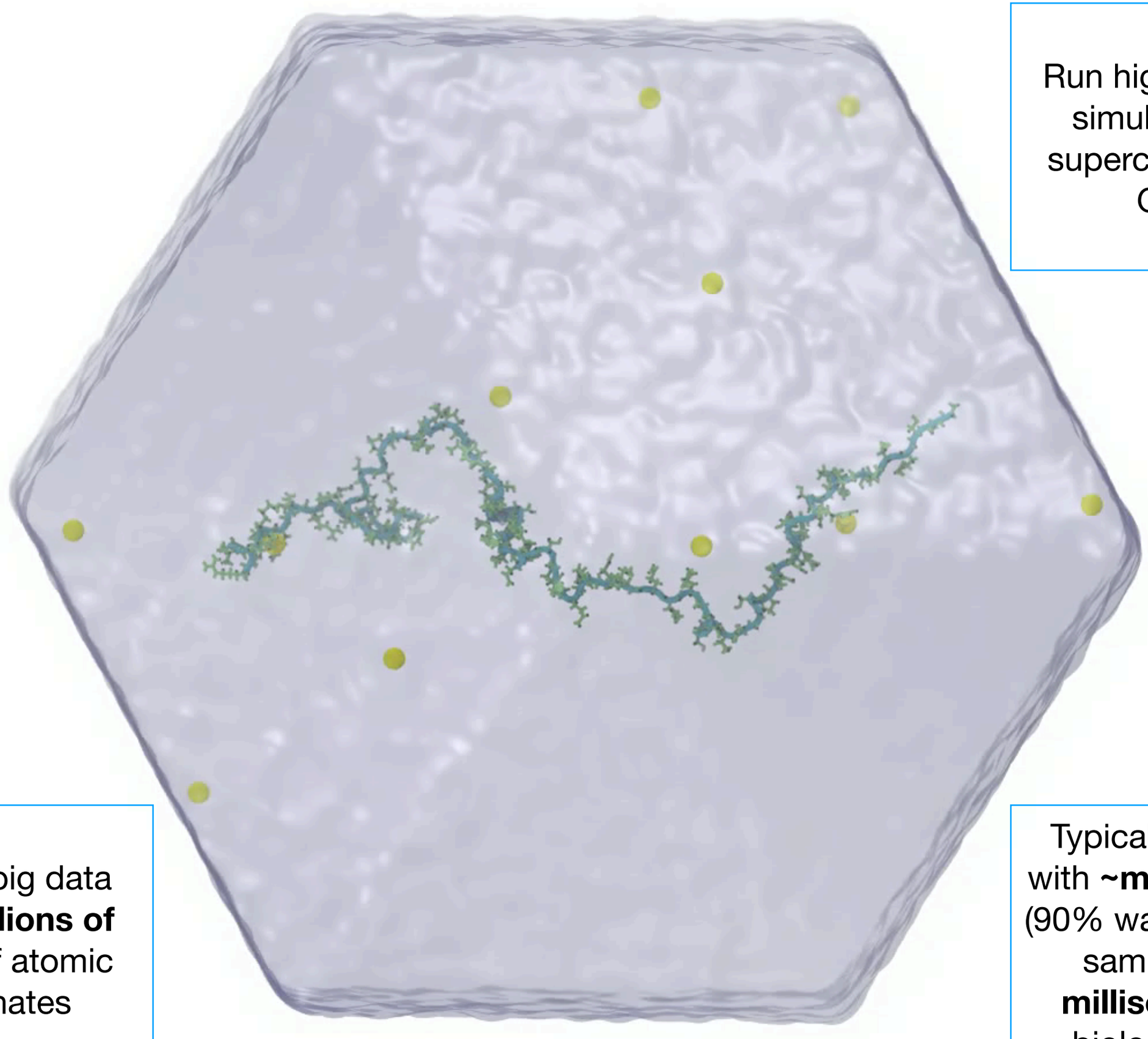
International HPC Summer School 2019,
RIKEN, Kobe City

July 8, 2019

Chemical & Physical Sciences
UNIVERSITY OF TORONTO
MISSISSAUGA

Physics
UNIVERSITY OF TORONTO

Motivation: reduce complexity of MD simulations



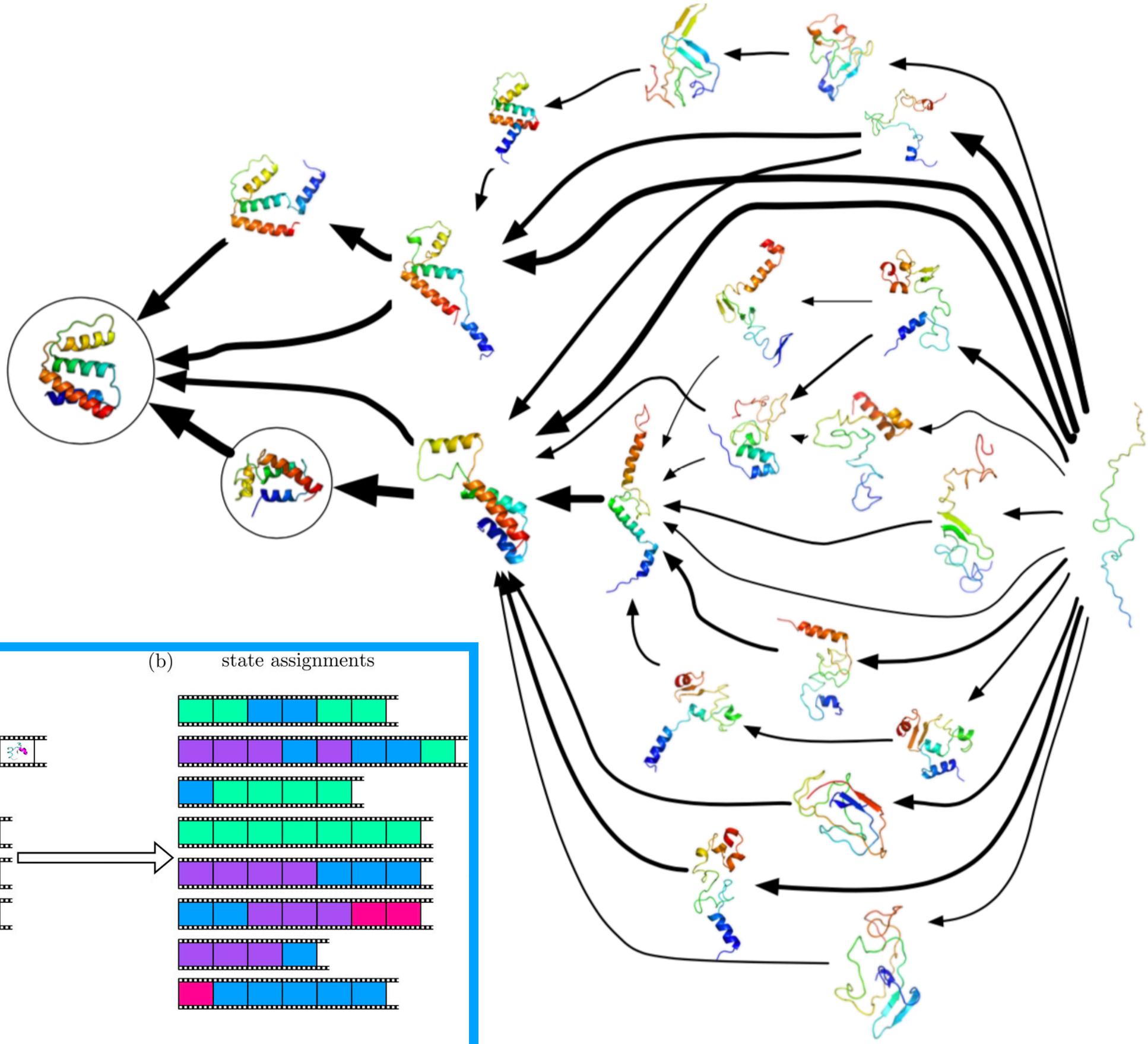
20 ns of the protein simulation

Run highly parallel simulations on supercomputers / GPUs

Produce big data sets – **millions of frames** of atomic coordinates

Typically systems with **~million atoms** (90% water) and can sample up to **milliseconds** of biological time

Network of states with transition probabilities

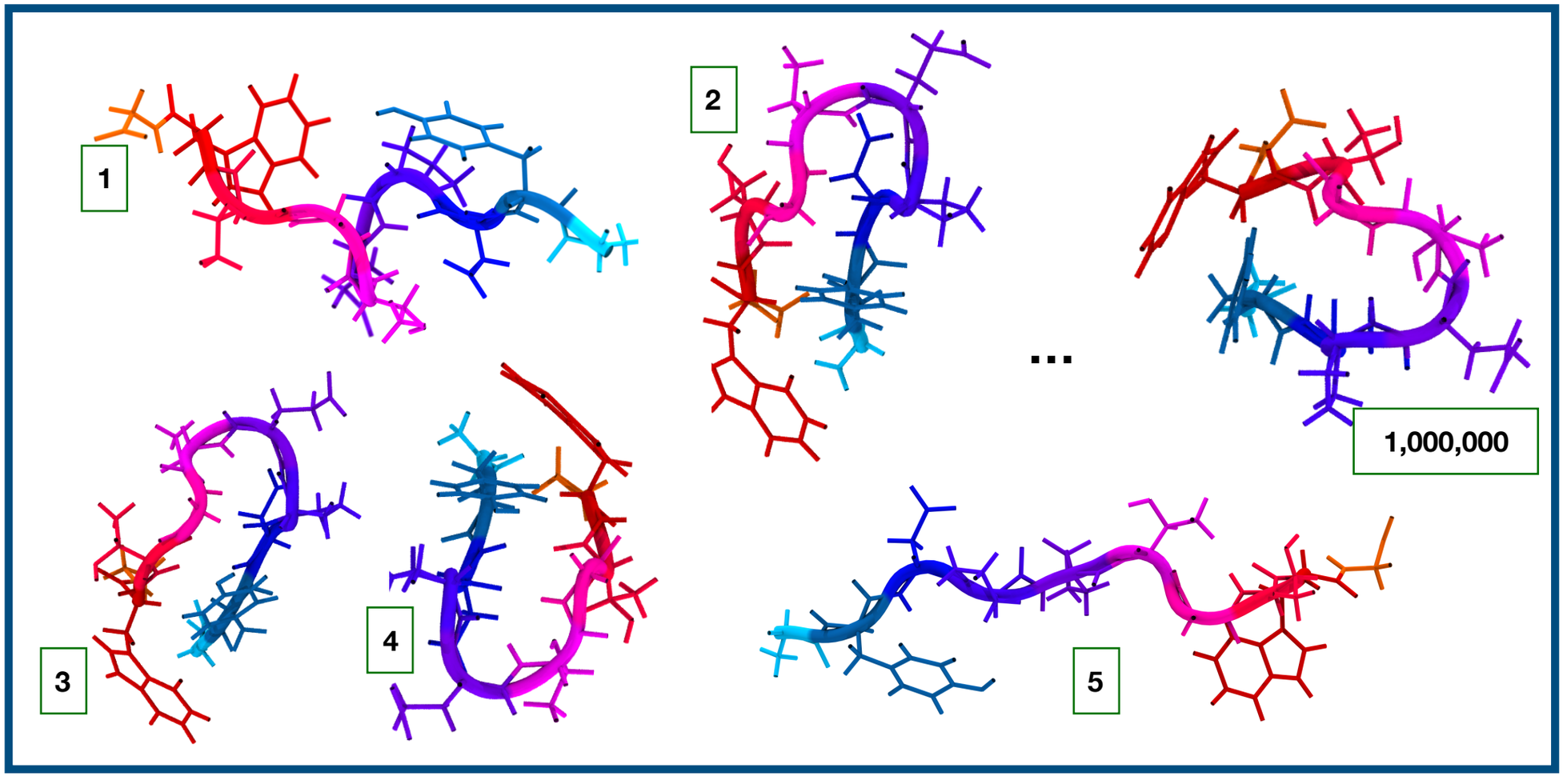


Voelz et al., *JACS* **2012** 134 (30), 12565-12577

The Problem:

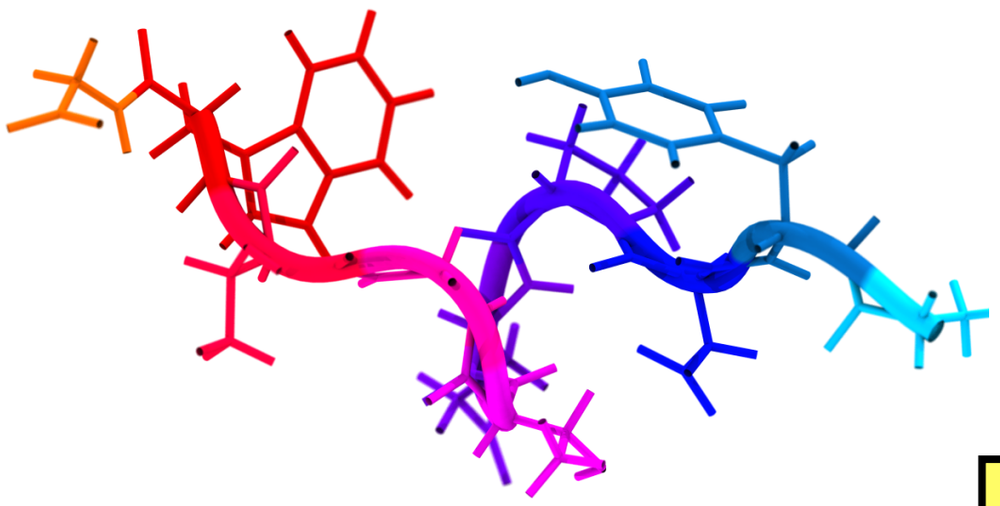
To find highly populated conformational states of the protein

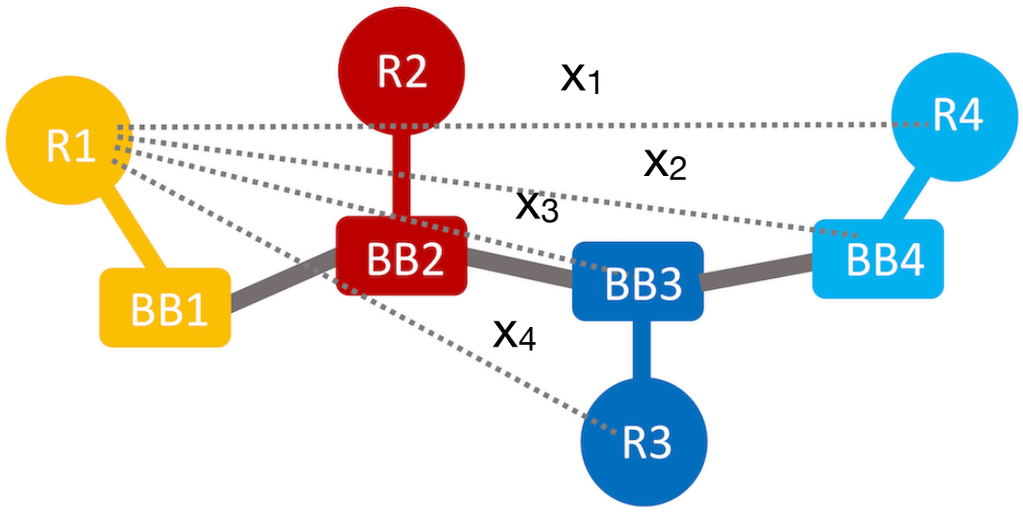
MD simulation with **~ 1 million** structures



The number of clusters (conformational states) is not known

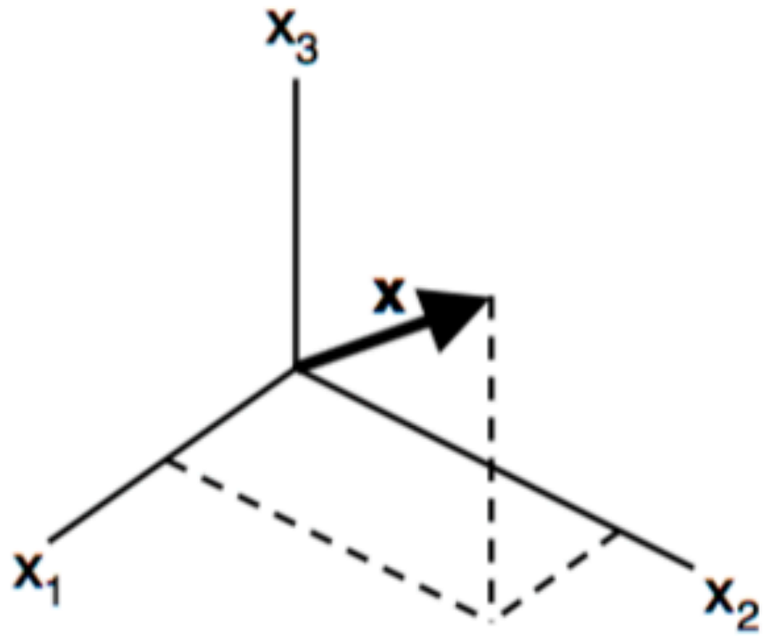
Create a feature vector for every structure



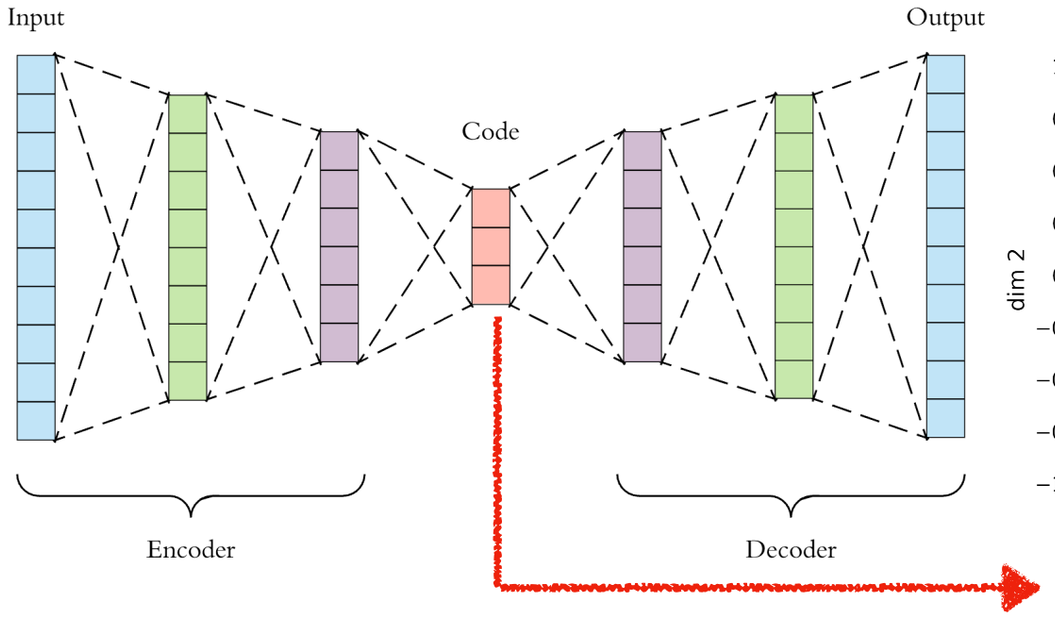


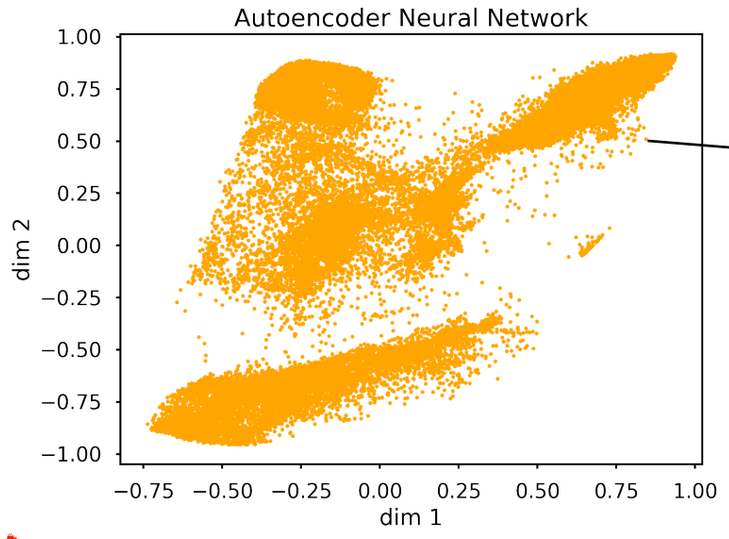
115-dimensional space

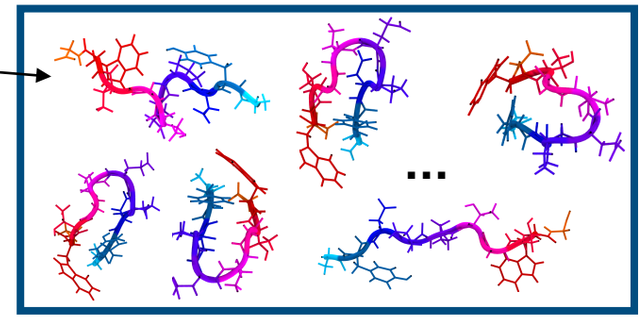
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_d \end{bmatrix}$$



Machine learning to reduce dimensions: from 115 to 2



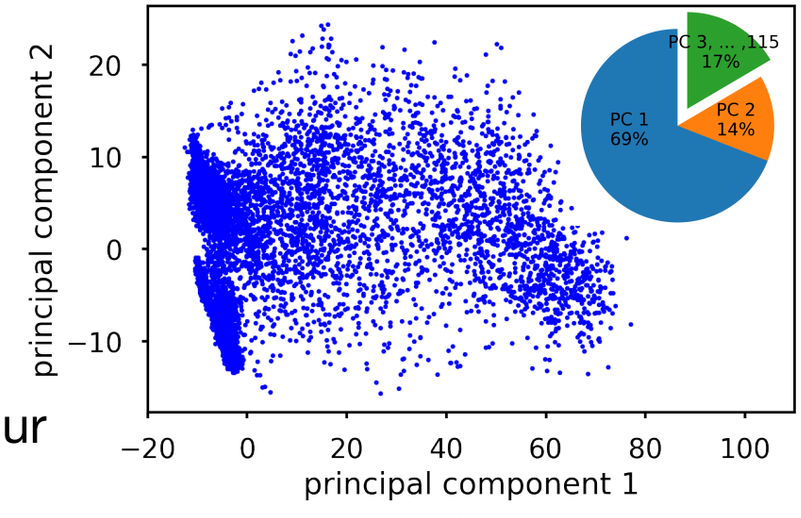




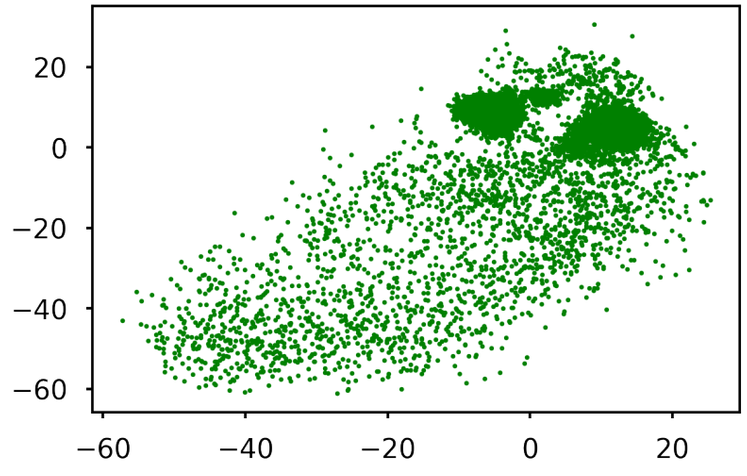
Several techniques used:

1. Autoencoder neural network
2. **P**rincipal **C**omponent **A**nalysis
3. **M**ulti**d**imensional **S**caling
4. **t**-distributed **S**tochastic **N**eighbour **E**mbedding
5. **U**niform **M**anifold **A**pproximation and **P**rojection

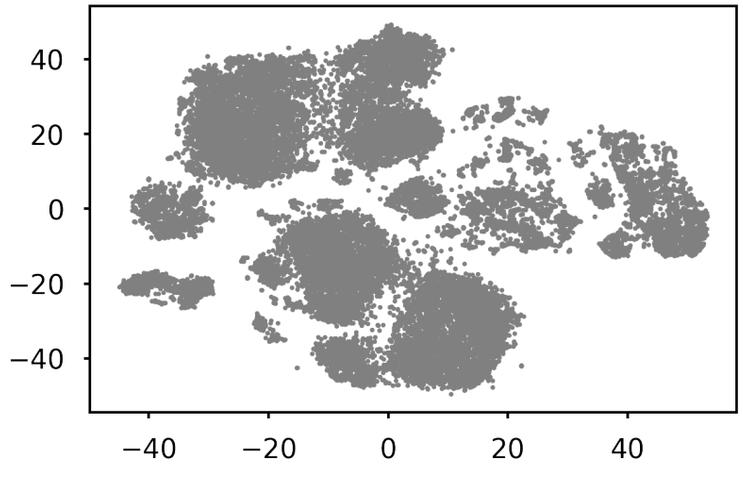
PCA



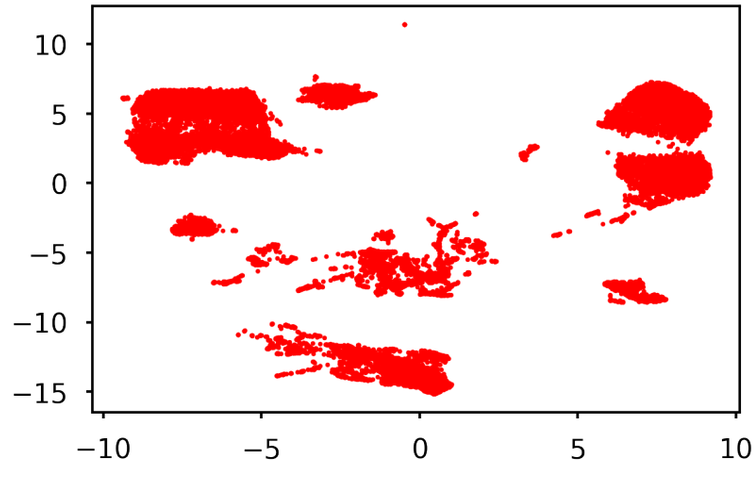
MDS



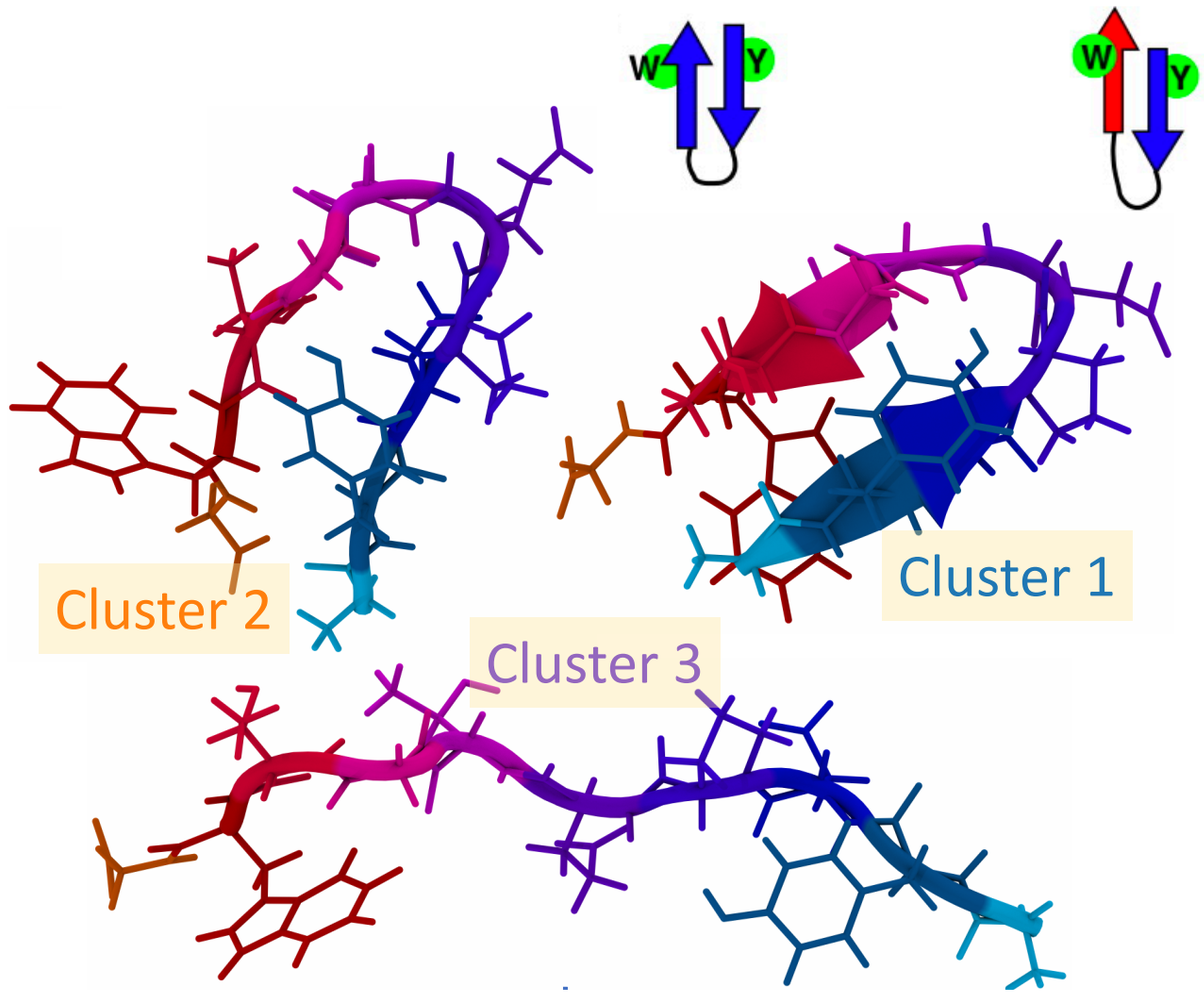
t-SNE

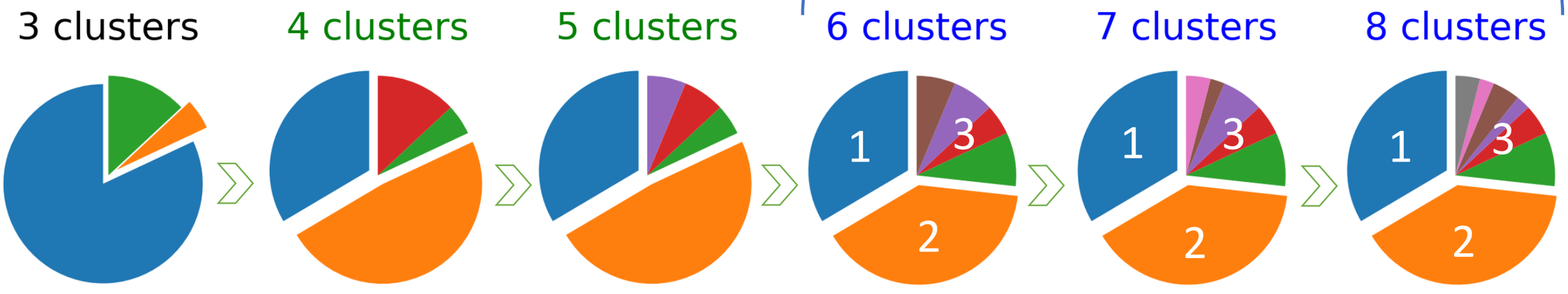


UMAP

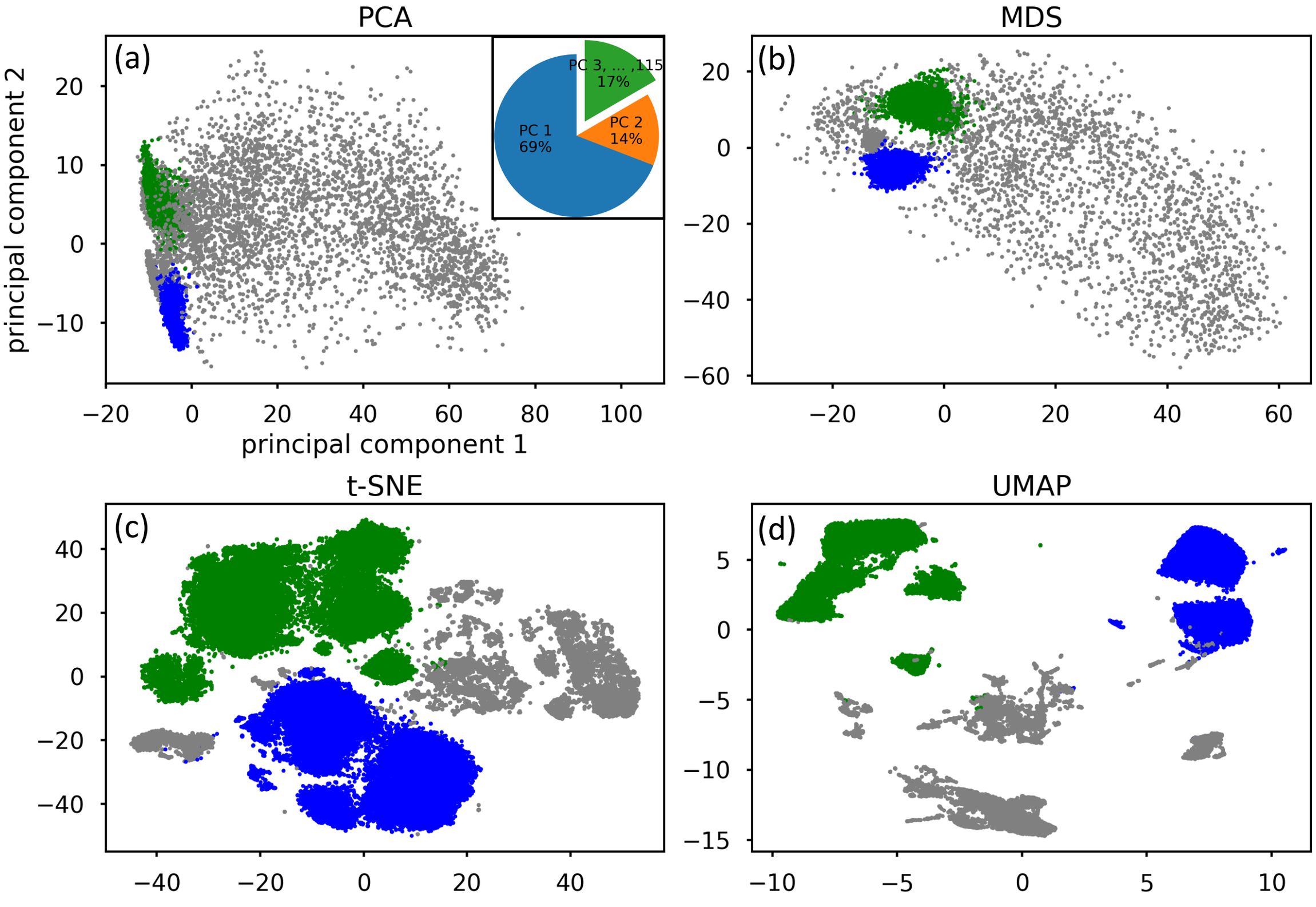


Results: clustering





Clustering in reduced dimensions



Conclusion & future directions

1. This approach (dimensionality reduction + clustering) helps to characterize conformational states of proteins

2. Large data-sets require a lot memory usage and parallel algorithms - HPC clusters will help

3. Next step is to find transition probabilities between states

4. Plan to develop methodology for a special type of proteins - Intrinsically Disordered proteins