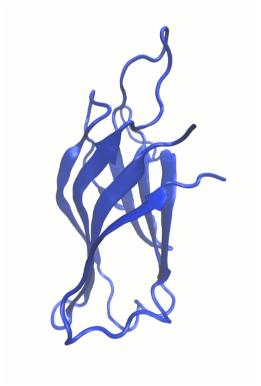
Tim Hempel
Noé Group
Department of Mathematics and Computer Science

Freie Universität Berlin

Molecular Dynamics Data Input and Featurization in PyEMMA

The classical MSM analysis pipeline

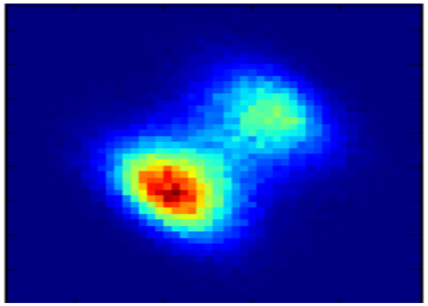
“MD data”



Featurization

“picking observables”, e.g. backbone torsions

high dimensional continuous trajectory

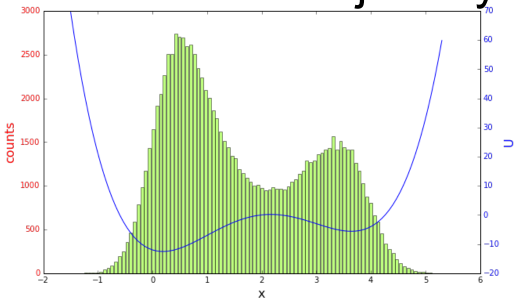


Coordinate transform

e.g. PCA, TICA

shortcut

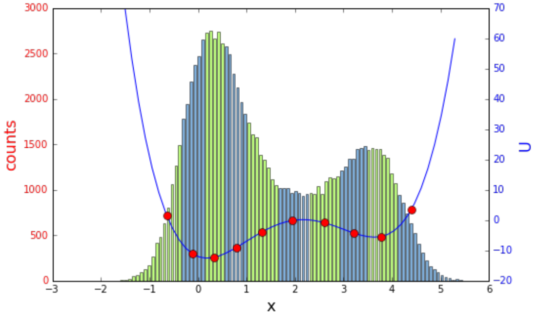
low dimensional continuous trajectory



clustering

e.g. k-means

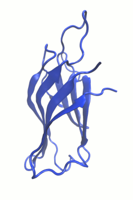
Discrete trajectory



Markov Model

The classical MSM analysis pipeline

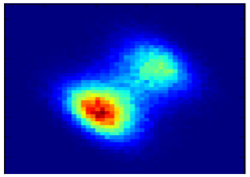
"MD data"



Featurization

"picking observables",
e.g. backbone
torsions

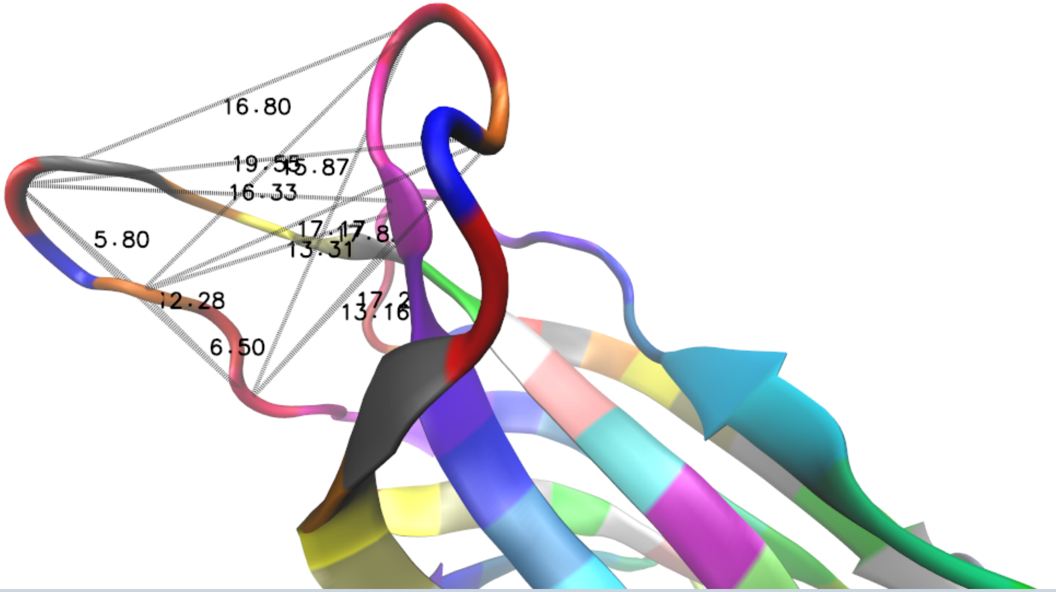
high dimensional
continuous trajectory



a) "what is the best description of my system?"
b) "what do I want to model?"

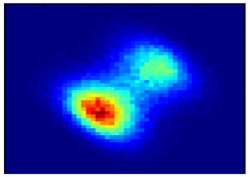
PyEMMA natively supported features:

- coordinates: all, heavy, Ca, selection
- angles:
  - backbone torsions
  - sidechain torsions
  - dihedrals
- distances or contacts between
  - all atom
  - Ca
  - heavy atom
- minimum distances
  - between residues or groups
- custom features



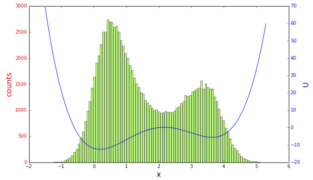
The classical MSM analysis pipeline

high dimensional continuous trajectory



Coordinate transform

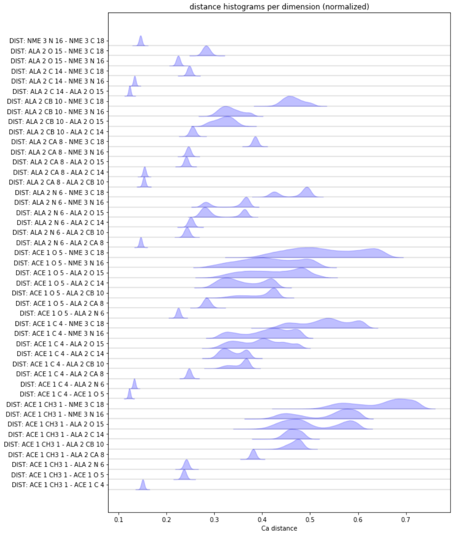
low dimensional continuous trajectory

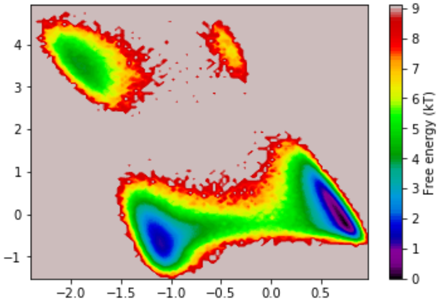


PyEMMA natively supported coordinate transforms:

- TICA (time-lagged independent component analysis)
    - strongly recommended
- PCA (principal component analysis)

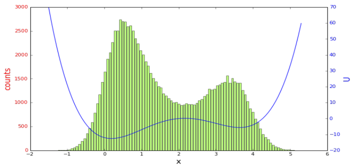
> *“What is the minimum dimensionality that still represents all of the important processes?”*





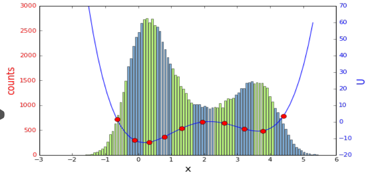
The classical MSM analysis pipeline

low dimensional continuous trajectory



clustering

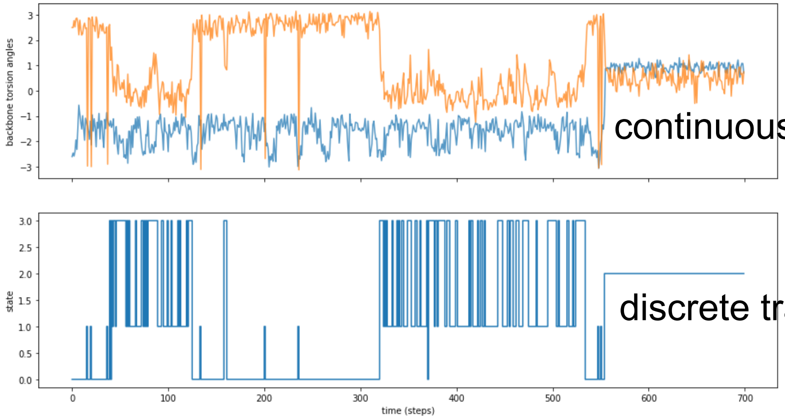
Discrete trajectory



PyEMMA natively supported clustering algorithms:

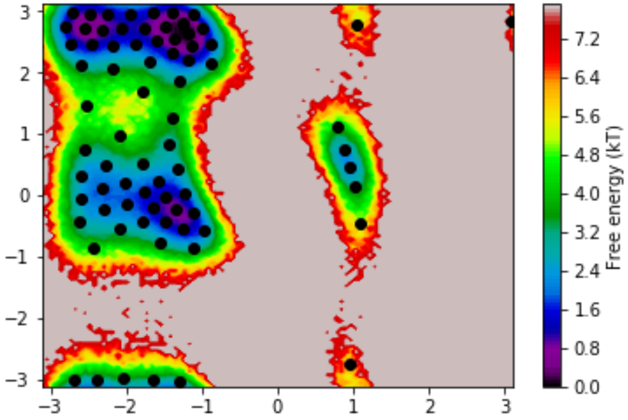
- k-means
- regular space
- uniform time

“What discretization resolves my processes best?”



continuous trajectory

discrete trajectory


Free energy (kT)