Tim Hempel
Noé group
FB Mathematik und Informatik

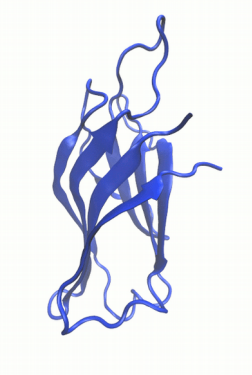
Freie Universität Berlin

Hands on PyEMMA
Data Input, Featurization, Discretization

The classical MSM pipeline

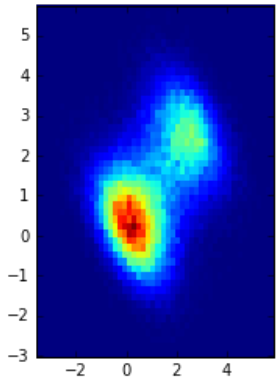
Freie Universität Berlin

"MD data"



Featurization
"picking observables", e.g. backbone torsions

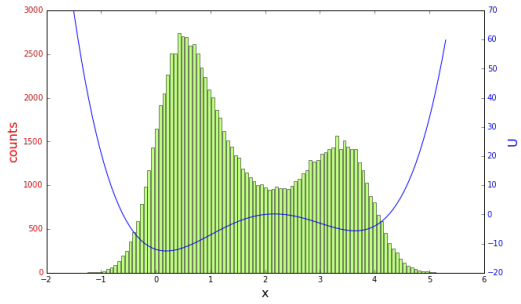
high dimensional continuous trajectory



Coordinate transform

e.g. PCA, TICA

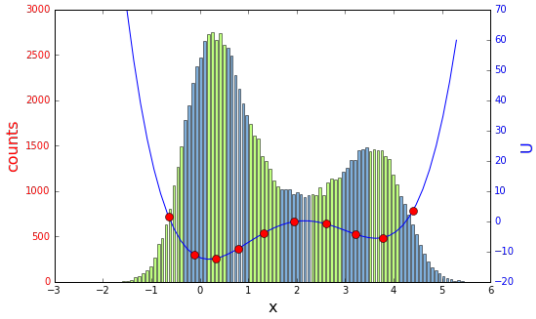
low dimensional continuous trajectory



shortcut

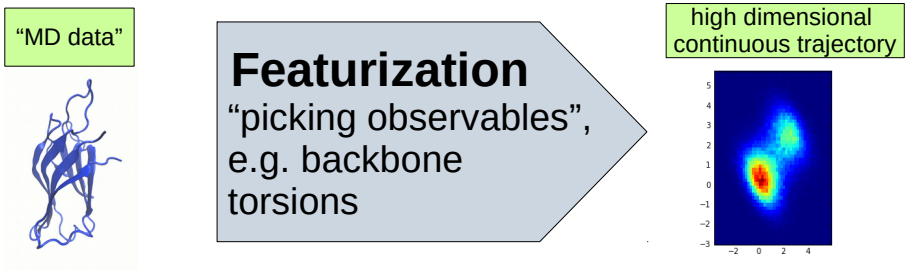
clustering
e.g. k-means

Discrete trajectory



Markov Model

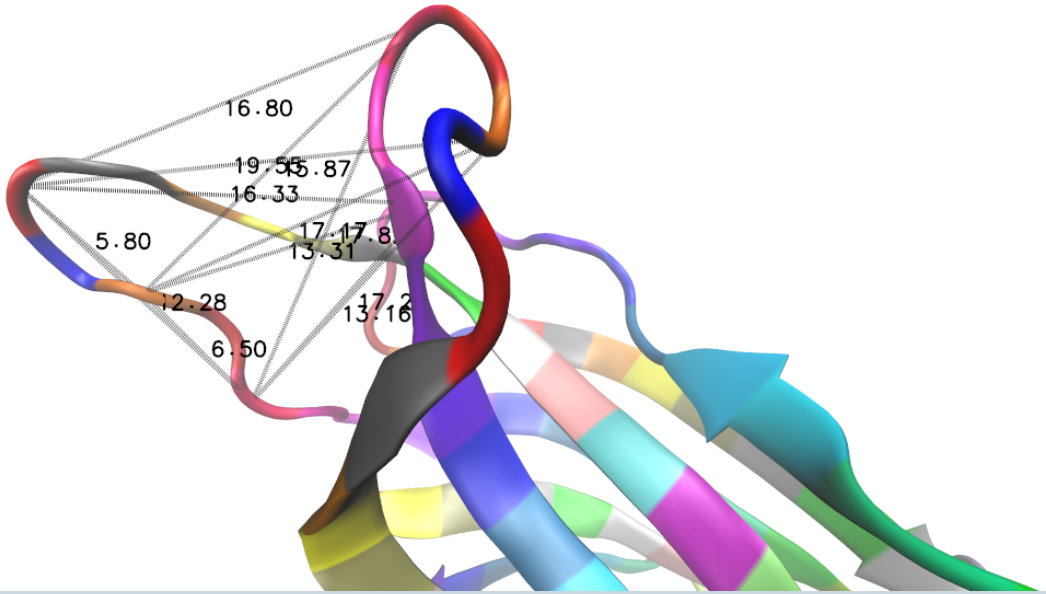
The classical MSM analysis pipeline



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

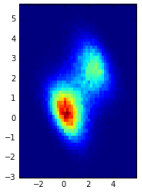


The classical MSM analysis pipeline

"MD data"

Featurization
"picking observables", e.g. backbone torsions

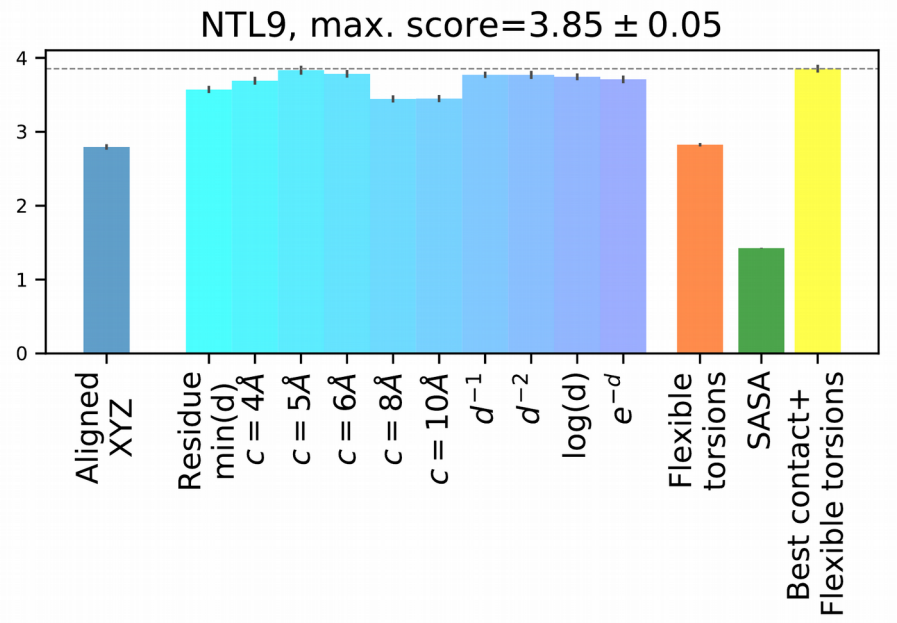
high dimensional continuous trajectory



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

How to choose features: VAMP score

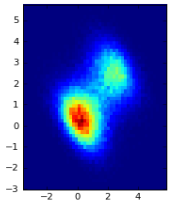
- It is possible to choose features based upon VAMP-2 score
- Score mirrors the number of processes captured by a fixed number of dimensions
- in practice, need cross-validation to avoid overfitting



Martin Scherer et al, arXiv:1811.11714 [physics.bio-ph]

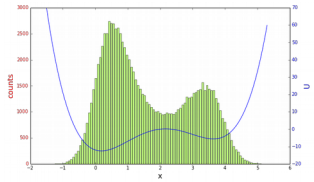
The classical MSM analysis pipeline

high dimensional continuous trajectory



Coordinate transform

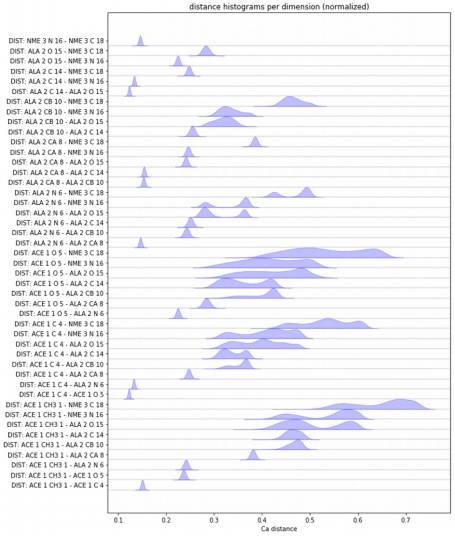
low dimensional continuous trajectory

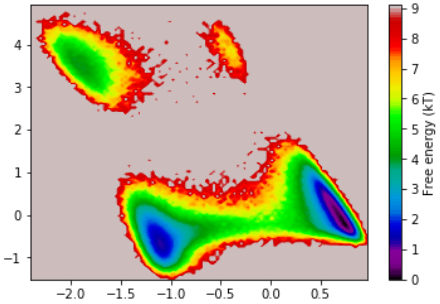


PyEMMA natively supported coordinate transforms:

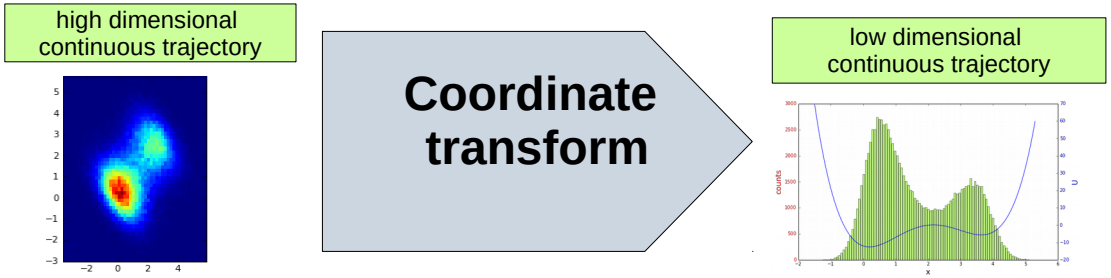
- TICA (time-lagged independent component analysis)
- VAMP (variational approach for Markov processes)
- PCA (principal component analysis, not recommended)

"What is the minimum dimensionality that still represents all of the important processes?"





The classical MSM analysis pipeline



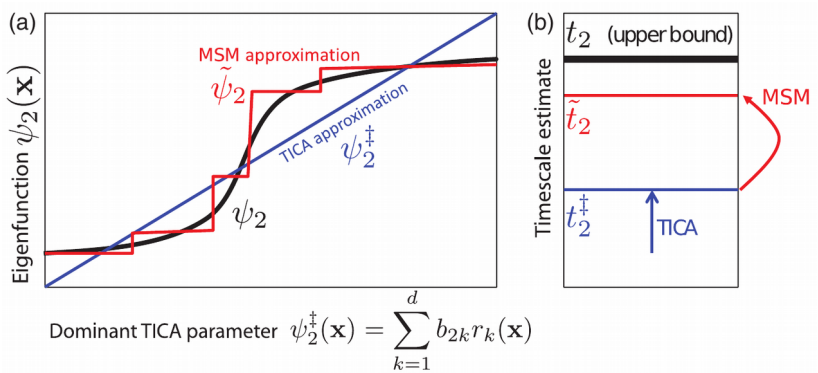
What is TICA?

- Linear approximation to the slow eigenfunctions of the system
- Maximizes auto-correlation for a given lag time (hyper-parameter)
- Estimation from data using

$$C_{ij}^{\mathrm{TICA}}(\tau) = \langle x_i(t) x_j(t+\tau) \rangle_t$$

- Generalized eigenproblem yields basis functions

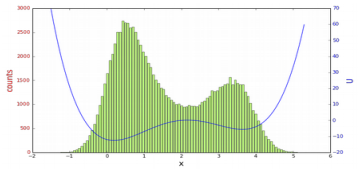
$$C(\tau)\vec{u}_i = C(0)\lambda_i(\tau)\vec{u}_i$$



Guillermo Pérez-Hernández et al., 2013, JCP

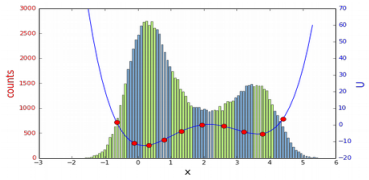
The classical MSM analysis pipeline

low dimensional continuous trajectory



clustering

Discrete trajectory



"What discretization resolves my processes best?"

PyEMMA natively supported clustering algorithms:

- k-means
- regular space
- uniform time