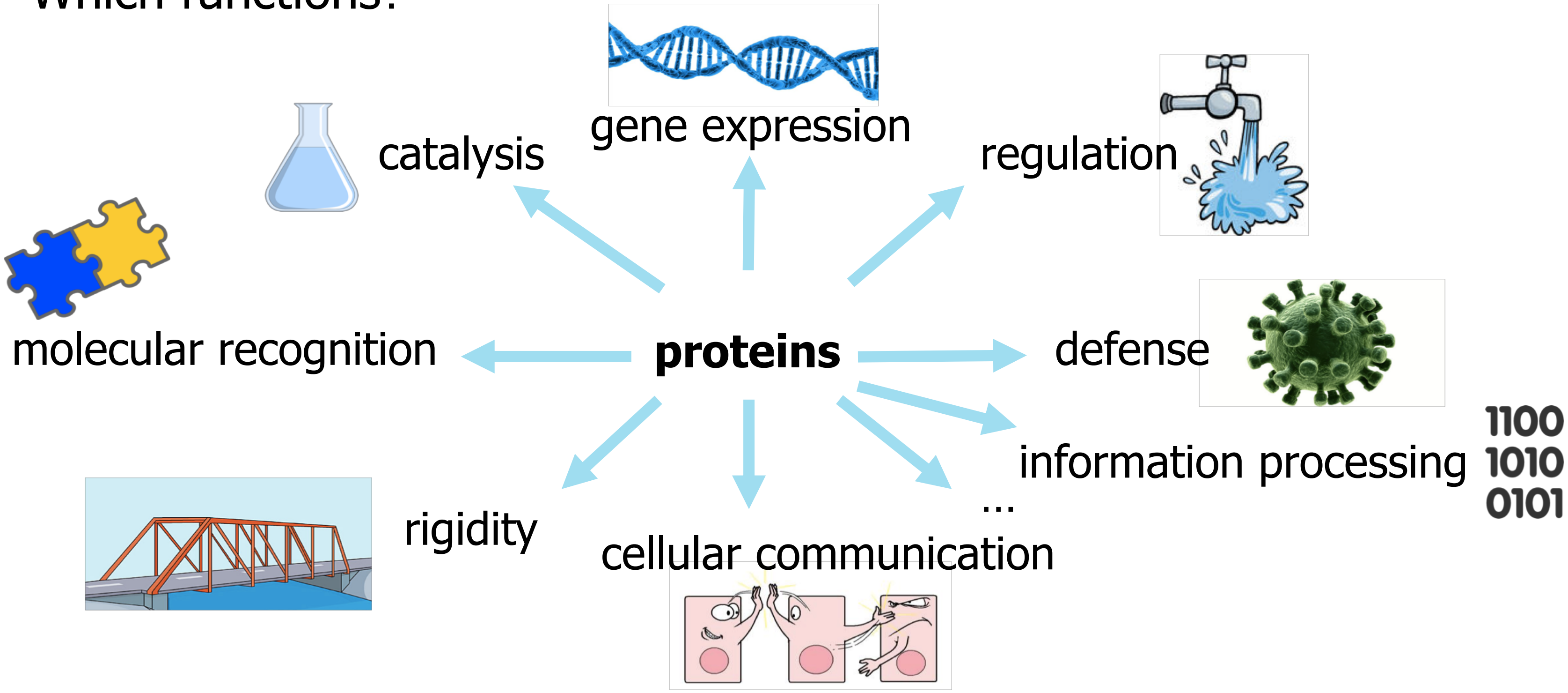
Protein dynamics and Markov modeling: Introduction + Overview

Fabian Paul

Computer Tutorial in Markov Modeling (PyEMMA) Talk 1

Protein 3-D structure and function

- Proteins are biomolecules that carry out their function *via* their 3-D structure, e. g. a receptor binding a molecule to detect a flavor or odor.
- Which functions?



proteins

catalysis

gene expression

regulation

defense

information processing

…

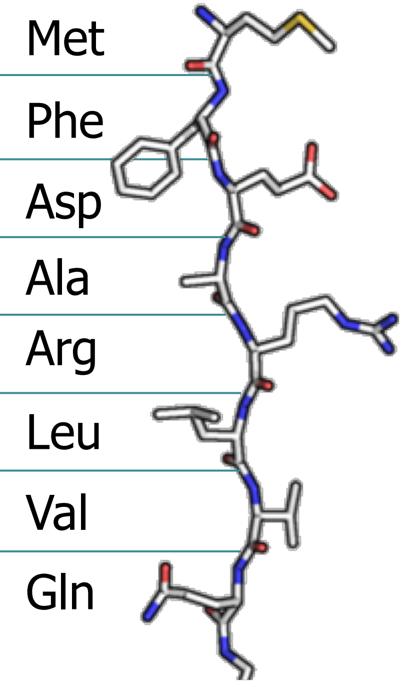
cellular communication

rigidity

molecular recognition

- To function, proteins have to change their 3-D structure with time, e. g.:
  - open ↔ closed (for regulation)
  - active ↔ inactive (for information processing, communication)
  - assembled ↔ disassembled (for rigidity+motion), …

Proteins in motion: time and timescales

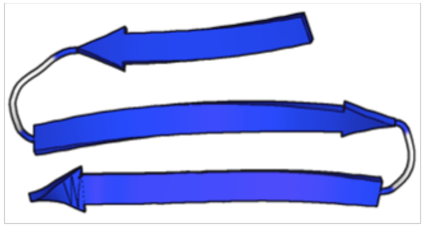


sequence

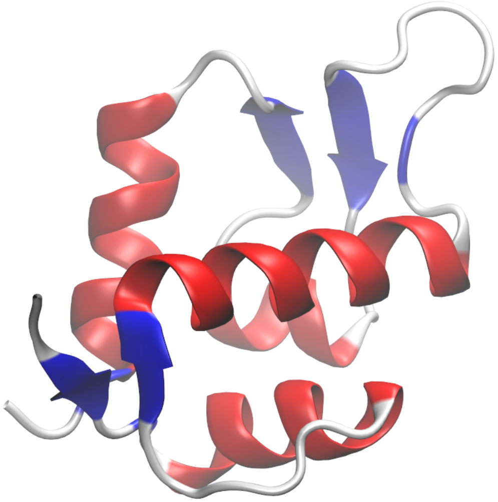
10^{-10}m

α-helix

β-sheet

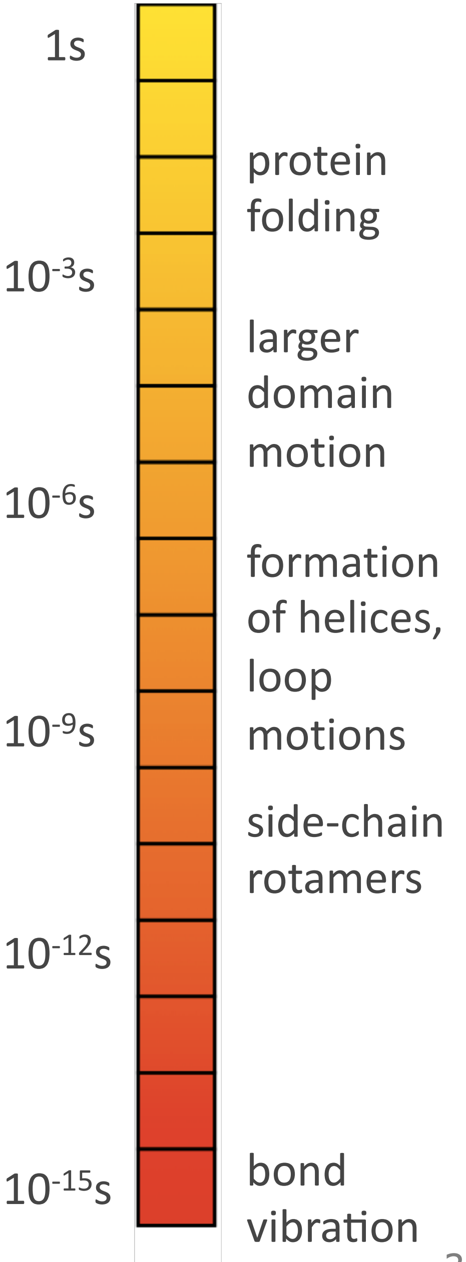


structural elements



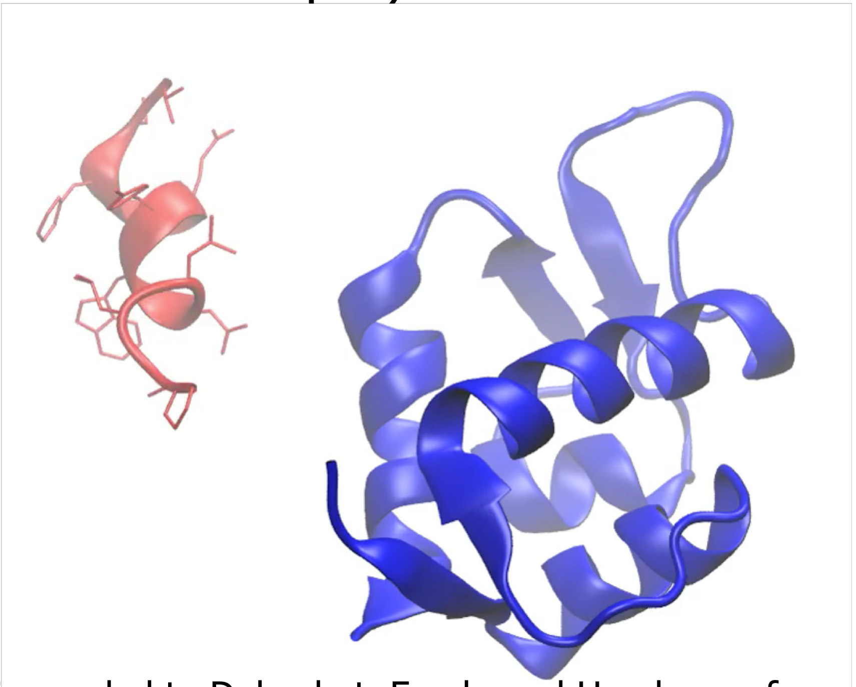
folded structure

10^{-8}m



Molecular dynamics (MD) simulation

- Experiment cannot resolve all temporal and spatial scales simultaneously. Experiments either have
  - high spatial resolution but low temporal resolution (e. g. cryo-electron microscopy*, X-ray diffraction)
  - high temporal resolution but limited spatial information. (e. g. single molecule fluorescence resonance energy transfer)
- Molecular dynamics simulation is an important tool that allows to observe molecules with simultaneously high temporal and high spatial resolution ("virtual microscope").



[*] Nobel prize in Chemistry 2017 awarded to Dubochet, Frank, and Henderson for cryo-electron microscopy

What is molecular dynamics (MD) simulation?

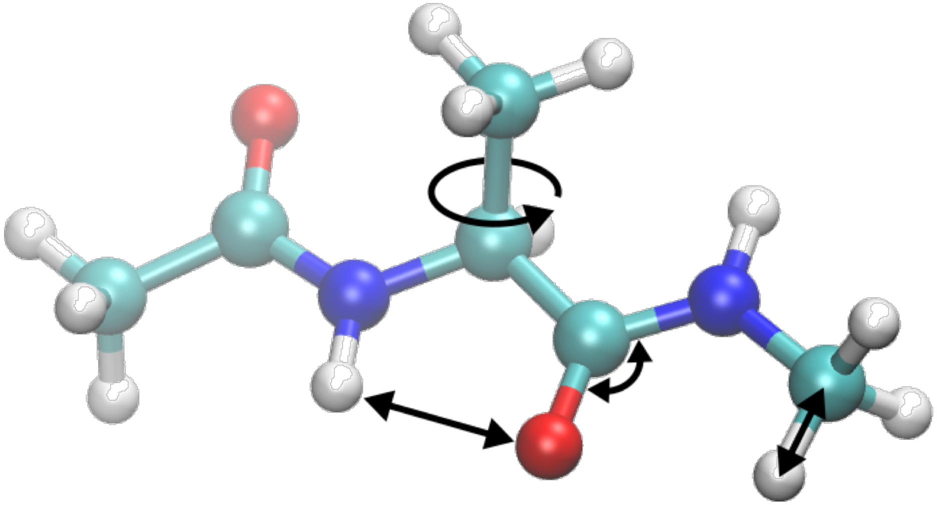
Molecular dynamics* uses classical mechanics to model molecular systems and consists of:

1. Equations of motion for the centers of masses $\boldsymbol{x}_i$ of the atoms, e. g. Langevin equations

$$m_i \ddot{\boldsymbol{x}}_i = -\gamma m_i \dot{\boldsymbol{x}}_i - \boldsymbol{\nabla}_i U(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) + \sqrt{2 k_B T \gamma}\, \boldsymbol{\eta}_i(t)$$

   with standard normally distributed random variates $(\eta_i)_j$

2. Molecular potential energy model $U(\boldsymbol{x})$ ”force field” that consists of energy terms for bonded and non-bonded interactions.



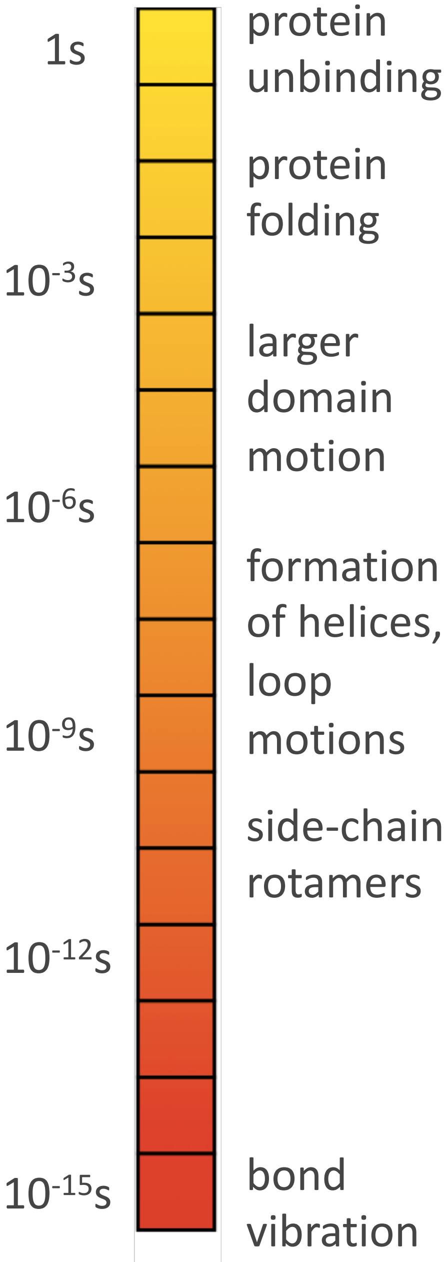
* Nobel prize in Chemistry 2013 awarded to Karplus, Levitt and Warshel for development of MD

Reachable time scales in MD simulation

Rate

100 ns / day / GPU*
e.g. Amber, AceMD, OpenMM

10 µs / day / Anton I



1s
10^{-3}s
10^{-6}s
10^{-9}s
10^{-12}s
10^{-15}s

protein unbinding

protein folding

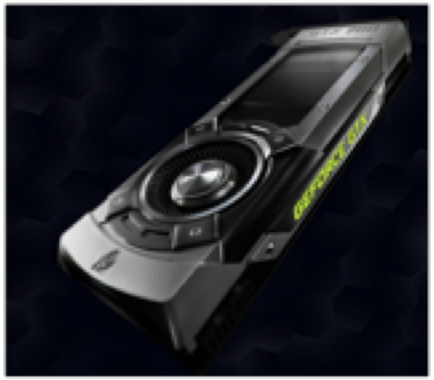
larger domain motion

formation of helices, loop motions

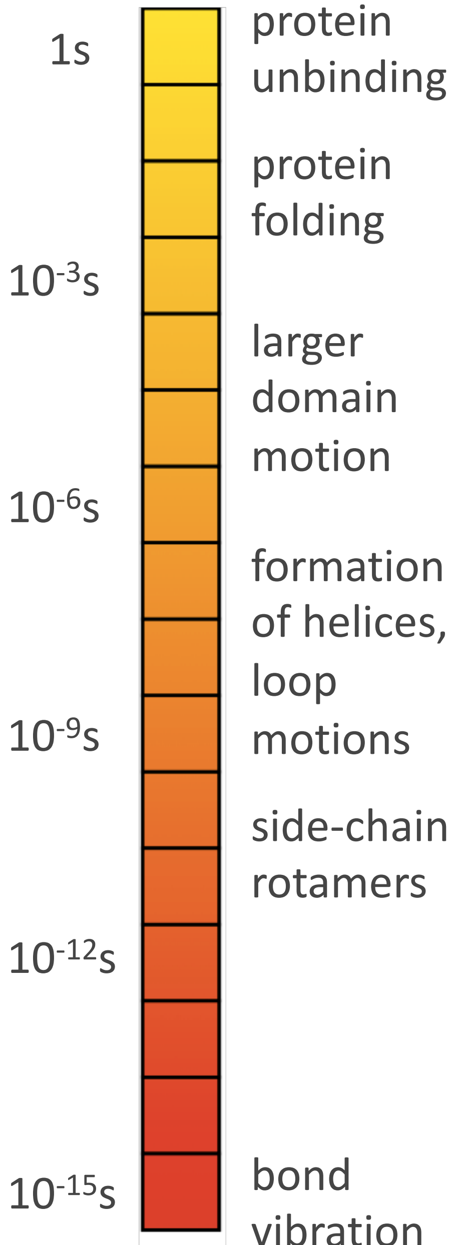
side-chain rotamers

bond vibration

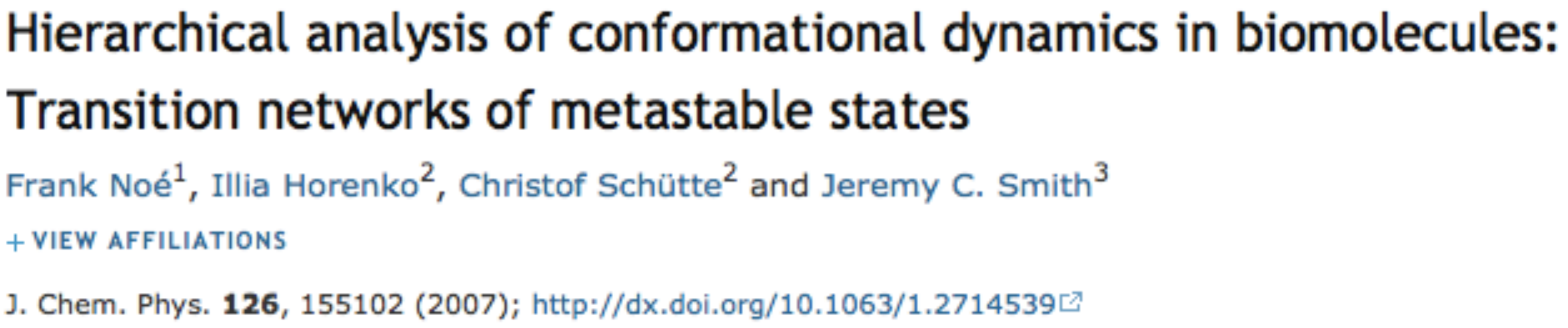
Reachable time scales in MD simulation



Rate	100 ns / day / GPU* e.g. Amber, AceMD, OpenMM	10 µs / day / Anton I
	100 GPUs	1 Anton I
	100 traj. of 100 ns / day	1 traj. of 10 µs / day
Throughput	10 µs / day	10 µs / day
Cost	100.000 USD	10.000.000 USD



First generation Markov state models (MSMs)

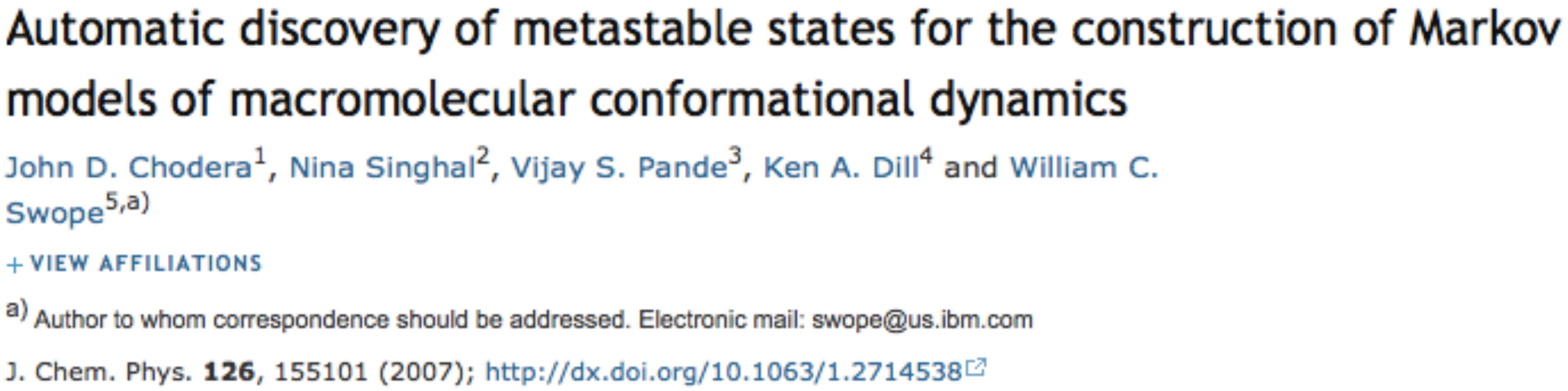


Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states

Frank Noé[1], Illia Horenko[2], Christof Schütte[2] and Jeremy C. Smith[3]

+ VIEW AFFILIATIONS

J. Chem. Phys. **126**, 155102 (2007); http://dx.doi.org/10.1063/1.2714539



Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics

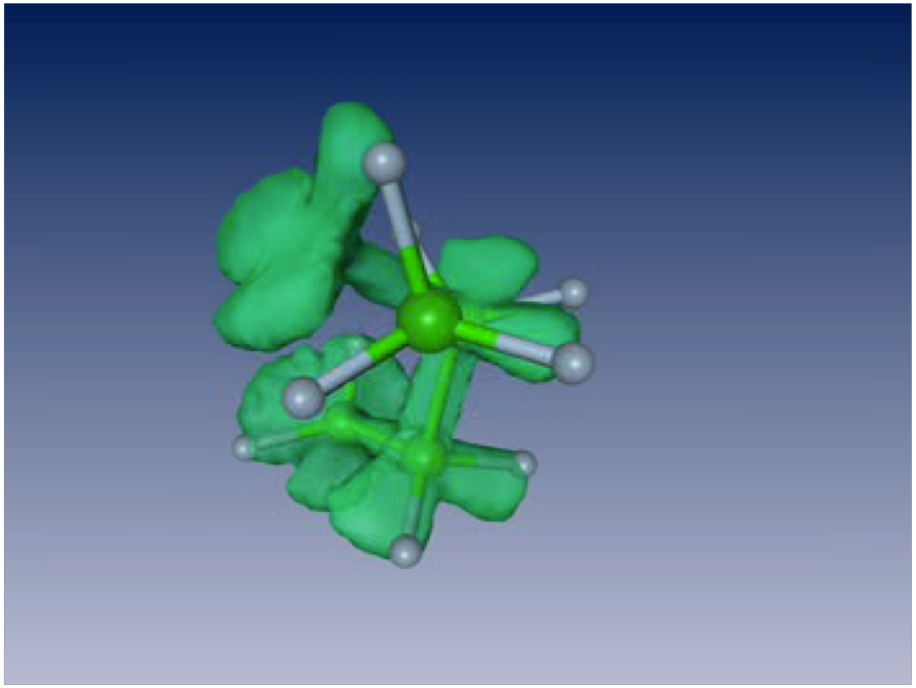
John D. Chodera[1], Nina Singhal[2], Vijay S. Pande[3], Ken A. Dill[4] and William C. Swope[5,a)]

+ VIEW AFFILIATIONS

a) Author to whom correspondence should be addressed. Electronic mail: swope@us.ibm.com

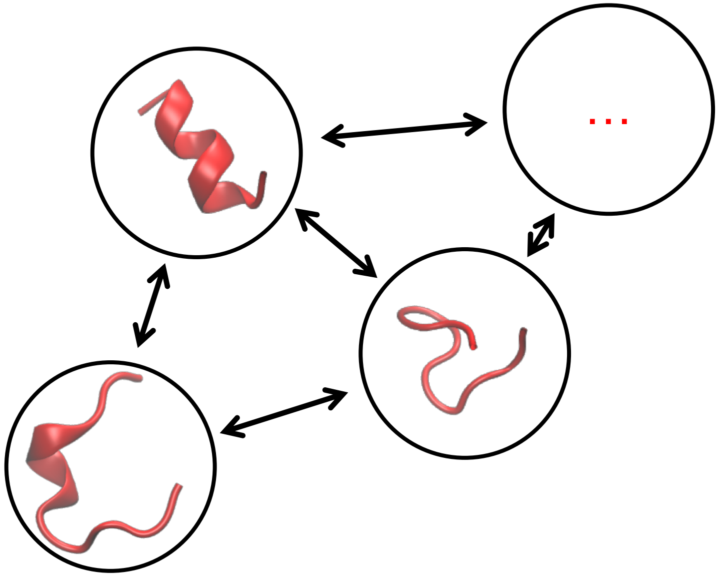
J. Chem. Phys. **126**, 155101 (2007); http://dx.doi.org/10.1063/1.2714538

Conformational dynamics



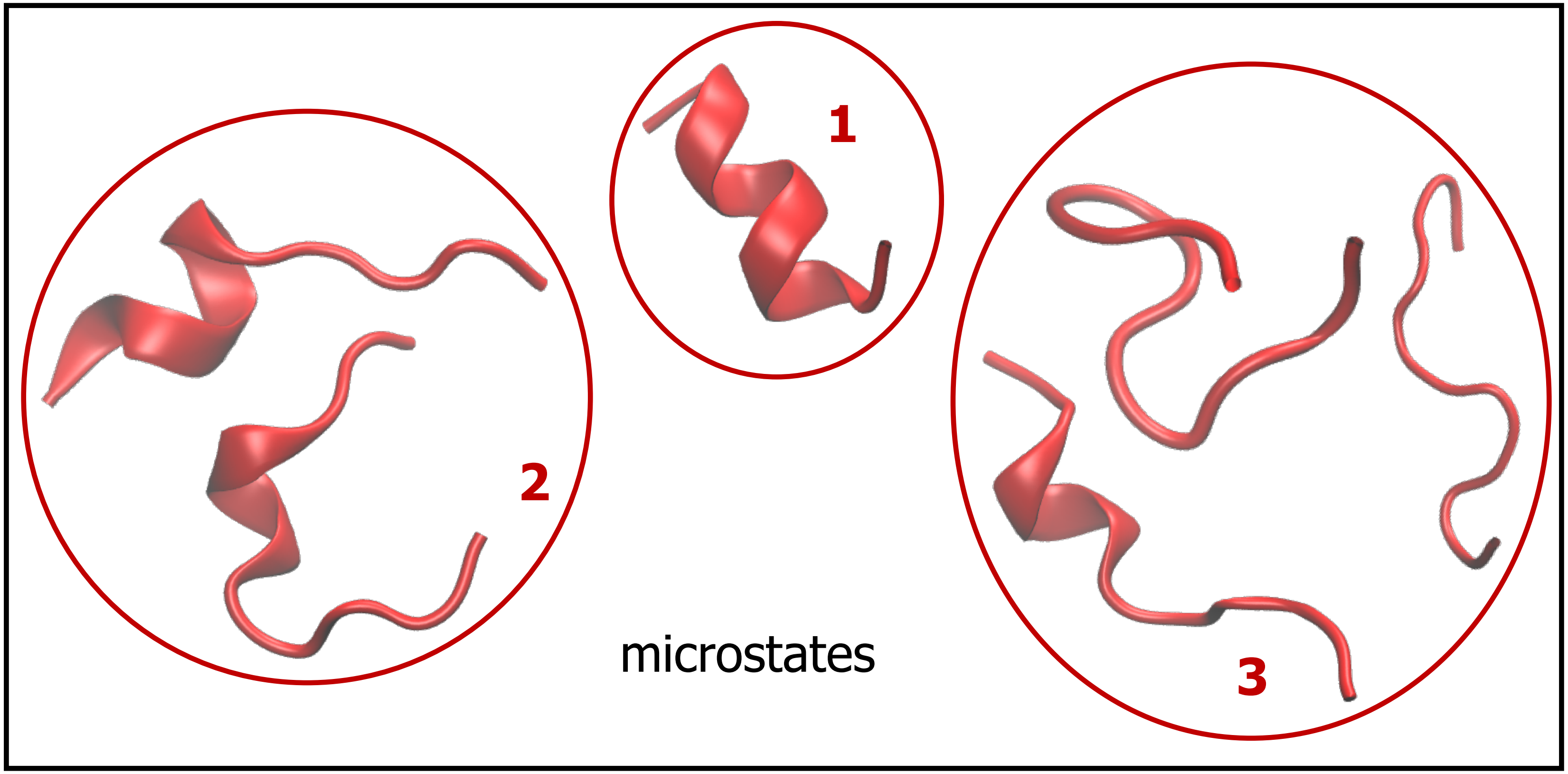
First generation Markov state models (MSMs)

- Markov state models (MSMs) can be used as a tool for the systematic analysis of multiple MD trajectories.
- A Markov state model consists of:
  1. a set of states $\{s_i\}_{i=1,\ldots N}$
  2. (conditional) transition probabilities between these state
$$T_{ij} = \mathbb{P}(s(t+\tau) = j \mid s(t) = i)$$
- Unlike MD trajectories, Markov state models are discrete in space and in time.



First generation Markov state models: estimation

- Markov model estimation starts with: grouping of geometrically[1] or kinetically[2] related conformations into *clusters* or *microstates*

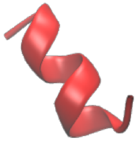
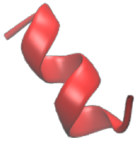
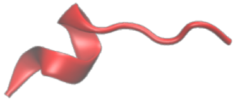
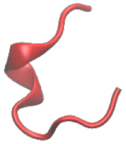


[1] Prinz *et al.*, *J. Chem. Phys.* **134**, 174105 (2011)

[2] Pérez-Hernández, **Paul**, *et al.*, *J. Chem. Phys.* **139**, 015102 (2013)

First generation Markov state models: estimation[1]

- We then assign every conformation in a MD trajectory to a microstate.

time t	τ	2τ	3τ	4τ	5τ	6τ	7τ
trajectory							
microstate s	1	1	2	3	3	2	3

- We count transitions between microstates and tabulate them in a count matrix $\mathbf{C}$

  e. g. $C_{11} = 1,\ C_{12} = 1,\ C_{23} = 2,\ \ldots$
- We estimate the transition probabilities T_{ij} from $\mathbf{C}$.
  - Naïve estimator: $\hat{T}_{ij} = C_{ij} / \sum_k C_{ik}$
  - Maximum-likelihood estimator [1]: $\hat{\mathbf{T}} = \underset{\mathbf{T}}{\operatorname{argmax}} \prod_{i,j} \left(T_{ij}\right)^{C_{ij}}$

[1] Prinz *et al.*, *J. Chem. Phys.* **134**, 174105 (2011)

[2] Pérez-Hernández, **Paul**, *et al.*, *J. Chem. Phys.* **139**, 015102 (2013)

First generation Markov state models: properties

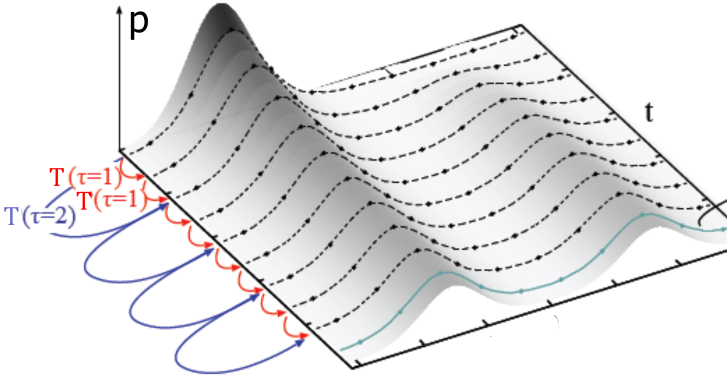
Markov state models:

- model the probability evolution of an ensemble

  let $p_i(t) = \mathbb{P}(s(t) = i)$, then

$$\mathbf{p}^T(n\tau) = \mathbf{p}^T(0)\mathbf{T}^n$$

  MSMs can extrapolate from the short-time estimate $\mathbf{T}(\tau)$ to long time scales.



- model the equilibrium distribution of an ensemble

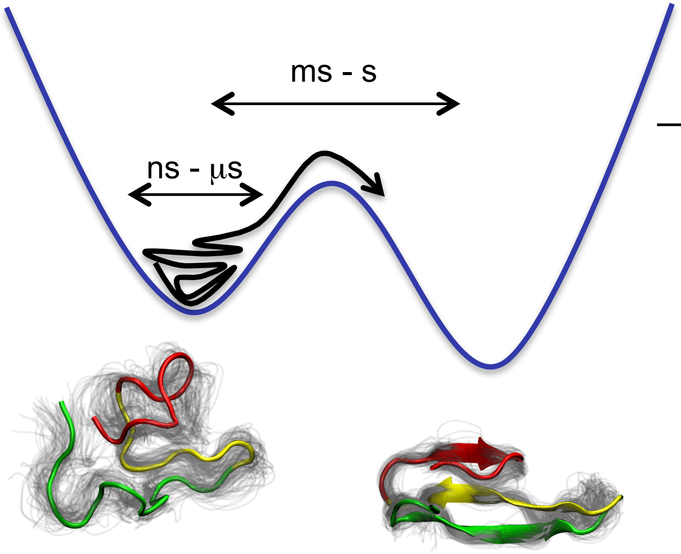
$$\boldsymbol{\pi}^T := \mathbf{p}^T(\infty) = \mathbf{p}^T(0) \lim_{n\to\infty} \mathbf{T}^n$$

  MSMs can even extrapolate to infinite time $\mathbf{p}^T(\infty)$. ($\tau \ll \infty$)

We can recover a coarse-grained version of the Boltzmann distribution ($\boldsymbol{\pi}$) without having to estimate $\mathbf{T}$ from data distributed according to the Boltzmann distribution.

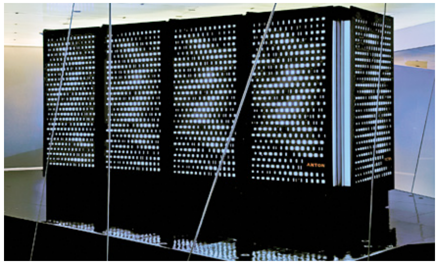
figure adapted from Nüske *et al., J. Chem. Theory Comput.* **10**, 1739 (2014)

Sampling Problem

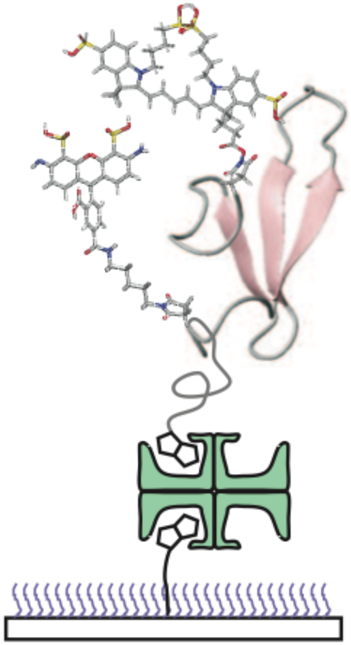


Analysis Problem

huge, complex datasets

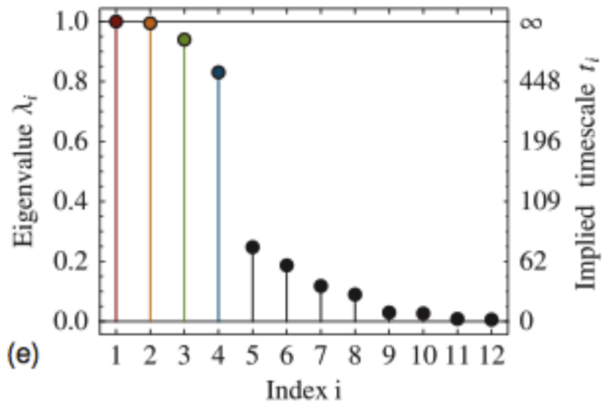


Reconciliation with Experiment

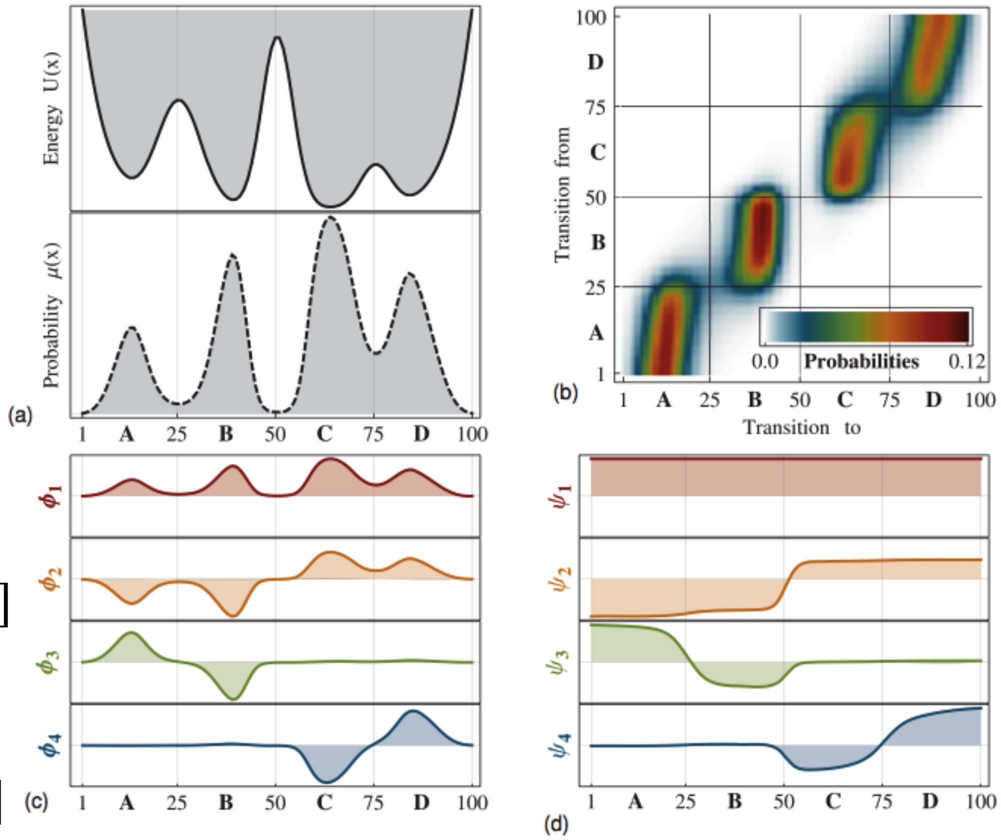


Second generation Markov models: spectral theory

time scales:



processes:



$$\boldsymbol{p}^T(\tau) = \boldsymbol{p}^T(0)\mathbf{T} = \sum_i \lambda_i\ \boldsymbol{\phi}_i\,[\boldsymbol{\psi}_i \cdot \boldsymbol{p}(0)]$$

$$\boldsymbol{p}^T(n\tau) = \sum_i \lambda_i^n \boldsymbol{\phi}_i\,[\boldsymbol{\psi}_i \cdot \boldsymbol{p}(0)]$$

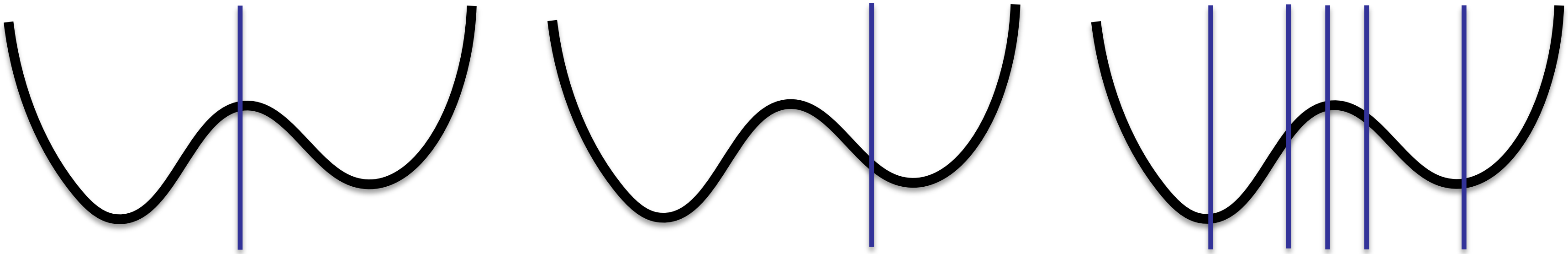
ϕ_i (left)

ψ_i (right)

- Eigenfunctions encode the slow relaxation processes.
- Eigenfunction point to the location of metastable states.

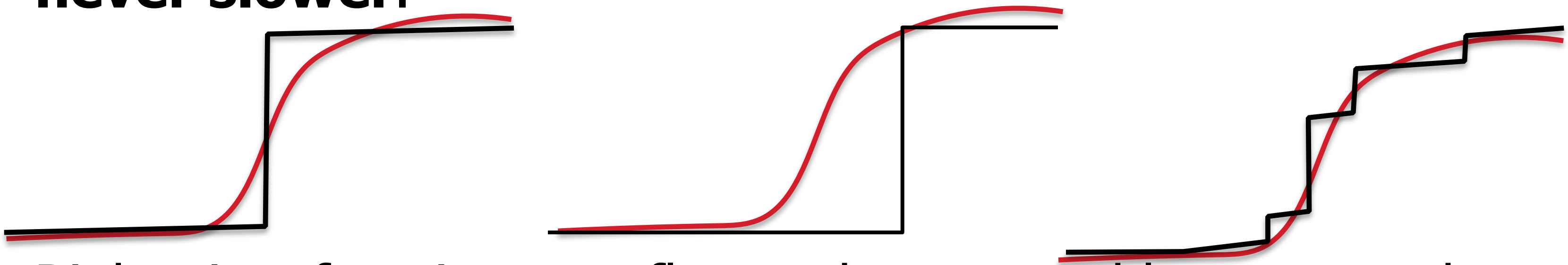
Prinz *et al.*, *J. Chem. Phys.* **134**, 174105 (2011) Sarich *et al.*, *SIAM Multiscale Model. Simul.* **8**, 1154 (2010).

Second generation Markov models: variational principle



good discretization bad discretization very good discretization

- If the cluster boundary is misplaced, transition across the boundary will be faster than transitions over the barrier **but never slower**.

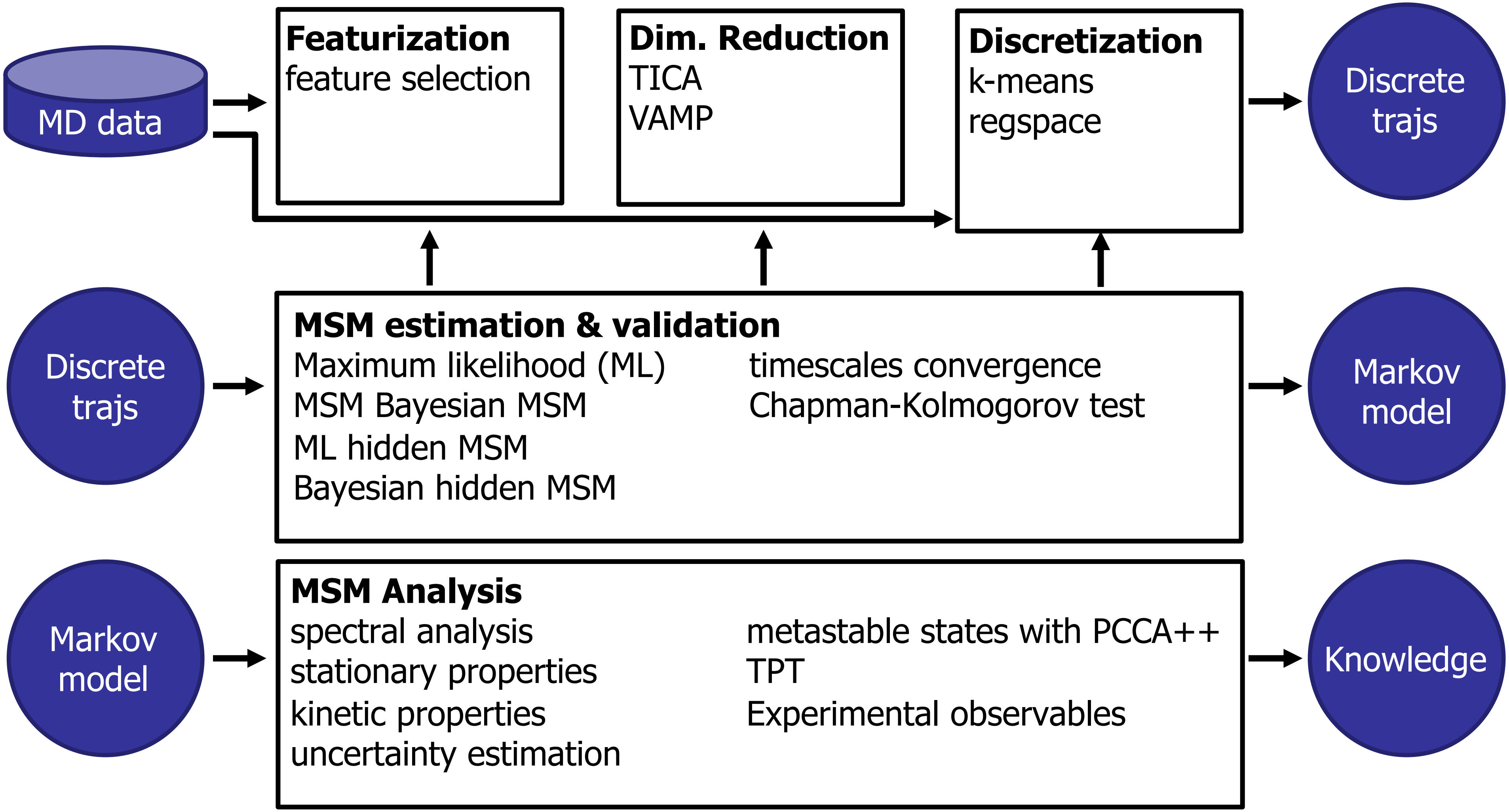


- Right eigenfunctions are flat on the metastable states and change only at/near the barrier. A good discretization allows to represent the eigenfunction well.
- Equivalence between eigendecomposition and maximizing “slowness”!

Second generation Markov models: variational principle

- Equivalence between eigendecomposition and maximizing "slowness".
$$R = \sum_{i=1}^{m} \mathrm{cov}\big(f_i(\boldsymbol{x}_t), f_i(\boldsymbol{x}_{t+\tau})\big) \leq R_m^{\mathrm{opt}}$$
where $f_1(x), \ldots f_m(x)$ are uncorrelated functions with variance 1. Can maximize the score for multiple functions simultaneously.
- Variational principle: generate a guess (for the functions) and rank it with the variational score. The higher, the better.
- Any algorithm that generates functions which maximize the score is suitable. Not limited to eigendecompositions / linear algebra.
- Works in very high-dimensional space.
- Result will be close to the true eigenfunctions.
→ Approximations will retain properties of the eigenfunctions: encode the slow dynamics, point towards the metastable states.

Markov modeling workflow



MD data

Featurization
feature selection

Dim. Reduction
TICA
VAMP

Discretization
k-means
regspace

Discrete trajs

Discrete trajs

MSM estimation & validation
Maximum likelihood (ML)
MSM Bayesian MSM
ML hidden MSM
Bayesian hidden MSM

timescales convergence
Chapman-Kolmogorov test

Markov model

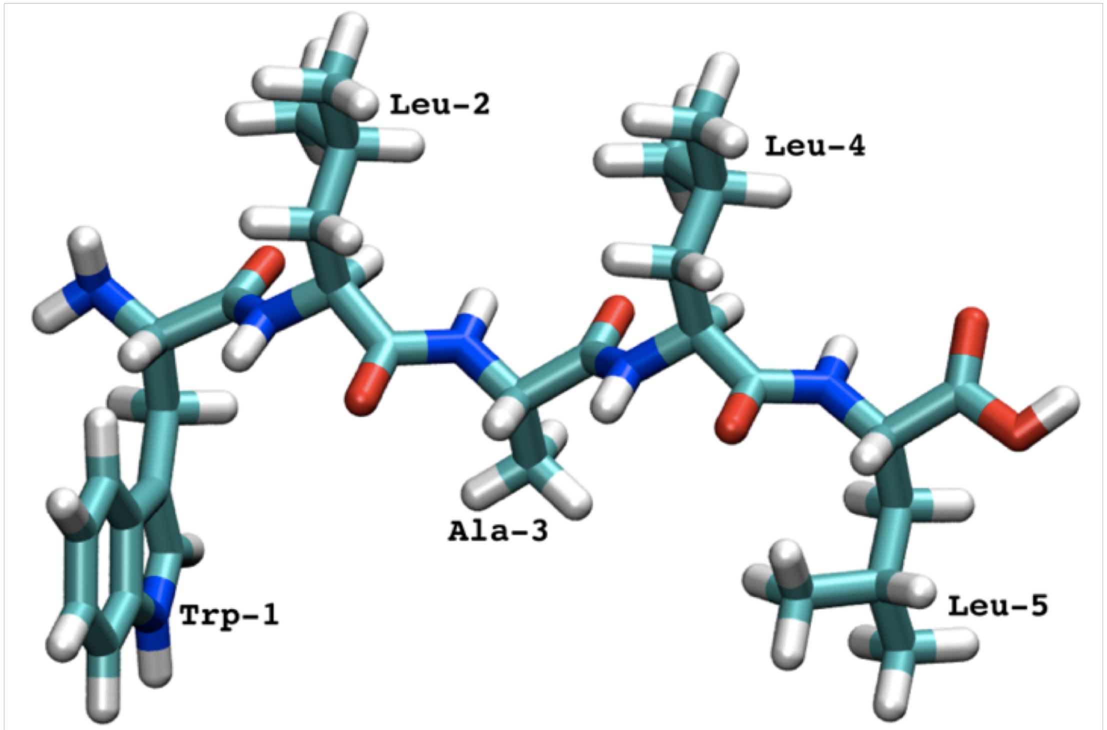
Markov model

MSM Analysis
spectral analysis
stationary properties
kinetic properties
uncertainty estimation

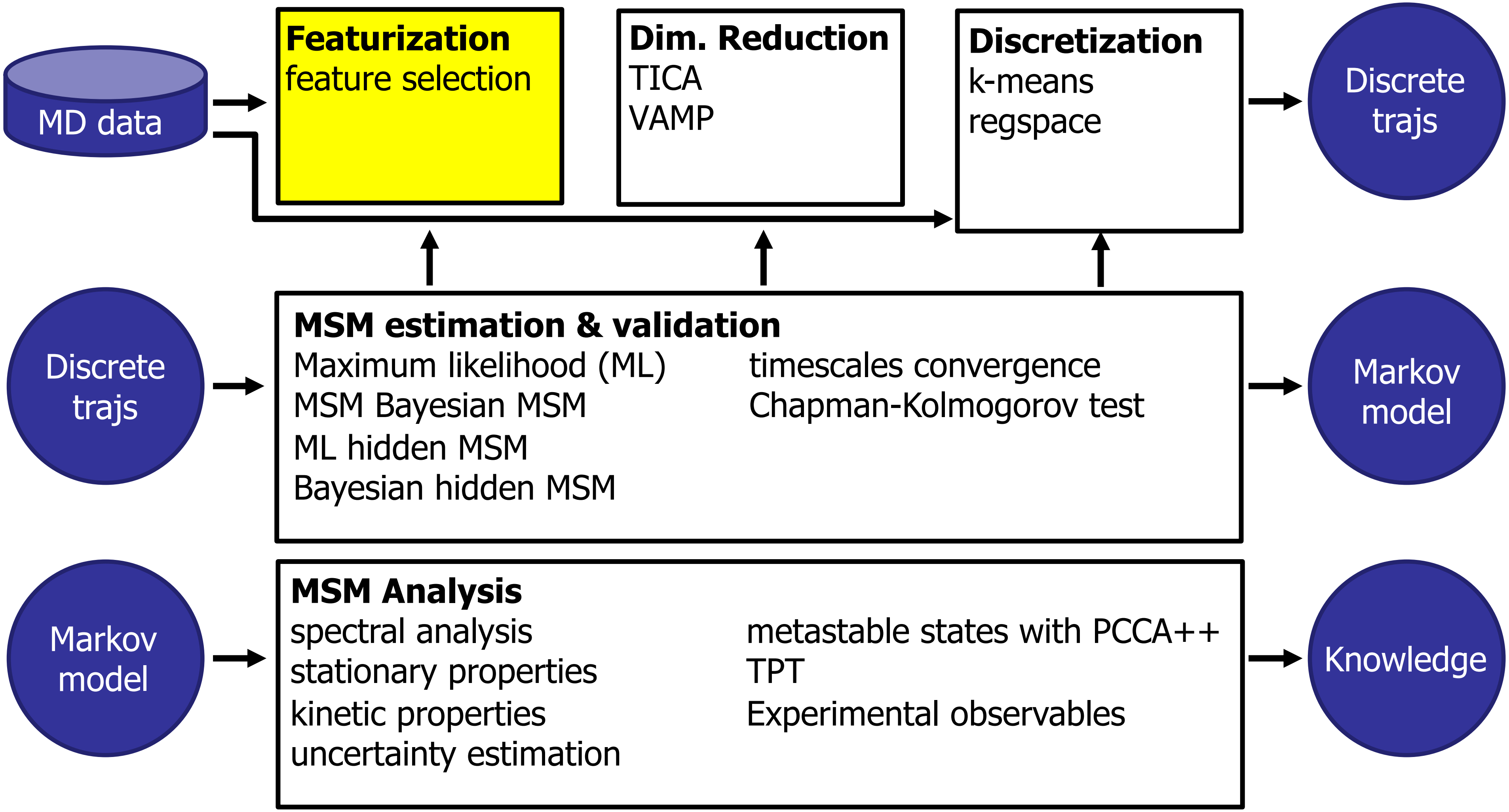
metastable states with PCCA++
TPT
Experimental observables

Knowledge

Markov modeling workflow: pentapeptide demo



Markov modeling workflow



MD data

Featurization
feature selection

Dim. Reduction
TICA
VAMP

Discretization
k-means
regspace

Discrete trajs

Discrete trajs

MSM estimation & validation
Maximum likelihood (ML)
MSM Bayesian MSM
ML hidden MSM
Bayesian hidden MSM

timescales convergence
Chapman-Kolmogorov test

Markov model

Markov model

MSM Analysis
spectral analysis
stationary properties
kinetic properties
uncertainty estimation

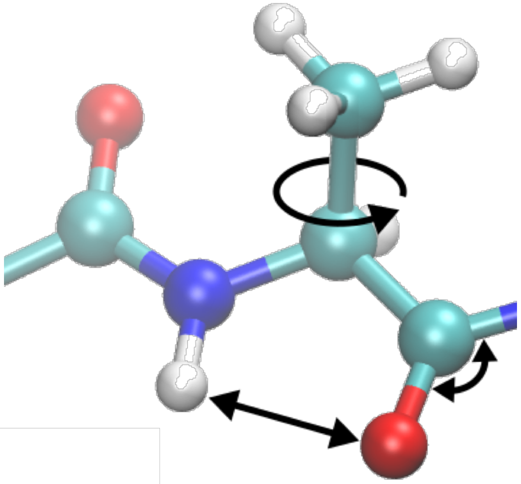
metastable states with PCCA++
TPT
Experimental observables

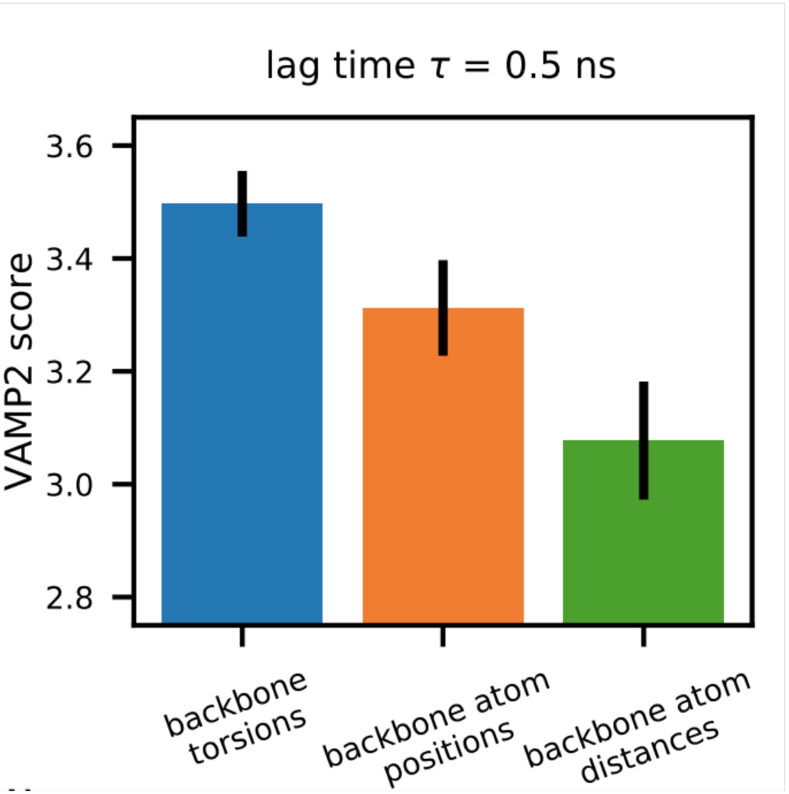
Knowledge

Feature selection

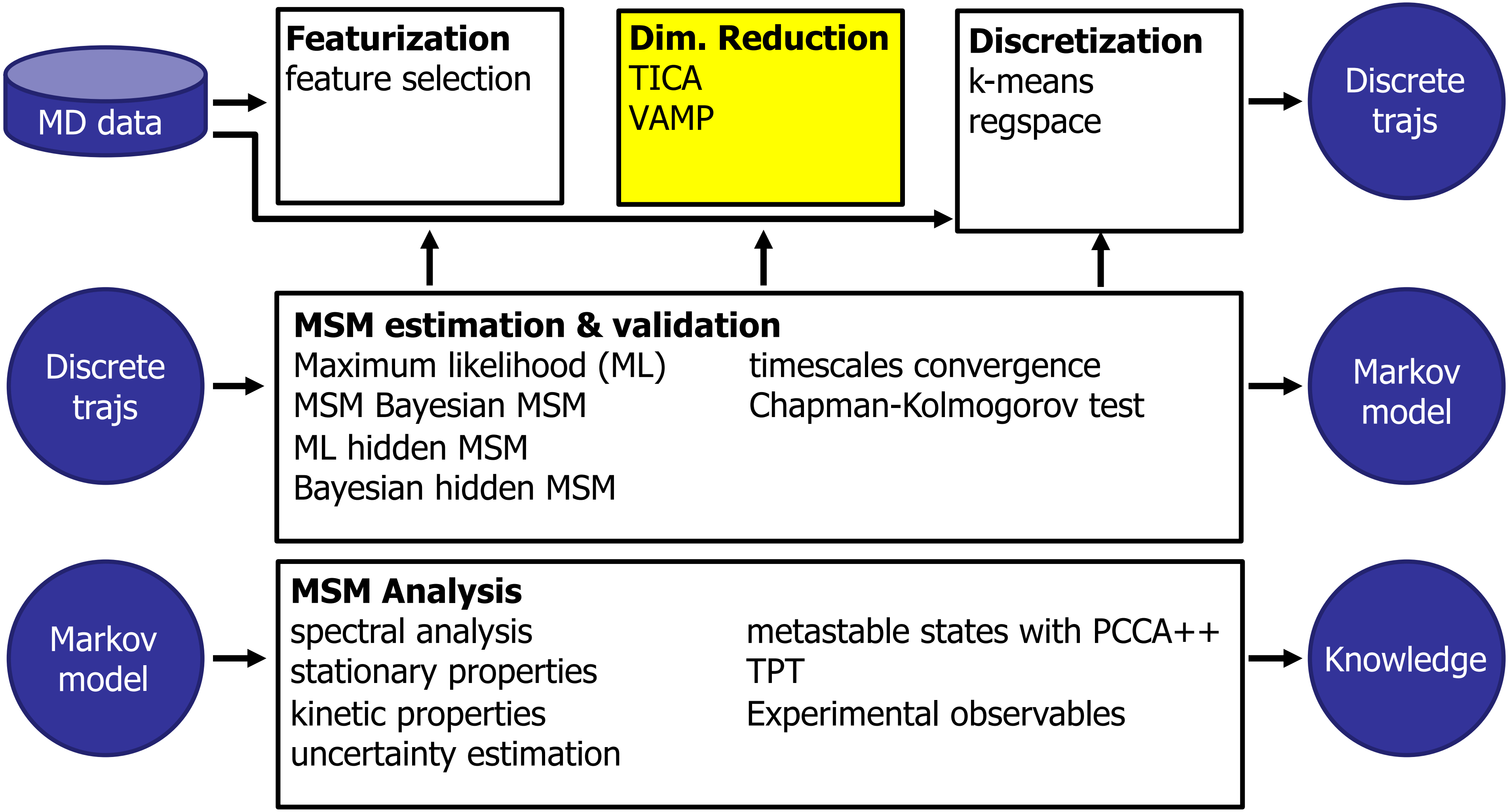
Select the set of molecular features that gives the most metastable kinetic model (the higher VAMP score, the better).

Use cross-validation prevent interpreting noise as a rare event.





Markov modeling workflow



MD data

Featurization
feature selection

Dim. Reduction
TICA
VAMP

Discretization
k-means
regspace

Discrete trajs

Discrete trajs

MSM estimation & validation
Maximum likelihood (ML)
MSM Bayesian MSM
ML hidden MSM
Bayesian hidden MSM

timescales convergence
Chapman-Kolmogorov test

Markov model

Markov model

MSM Analysis
spectral analysis
stationary properties
kinetic properties
uncertainty estimation

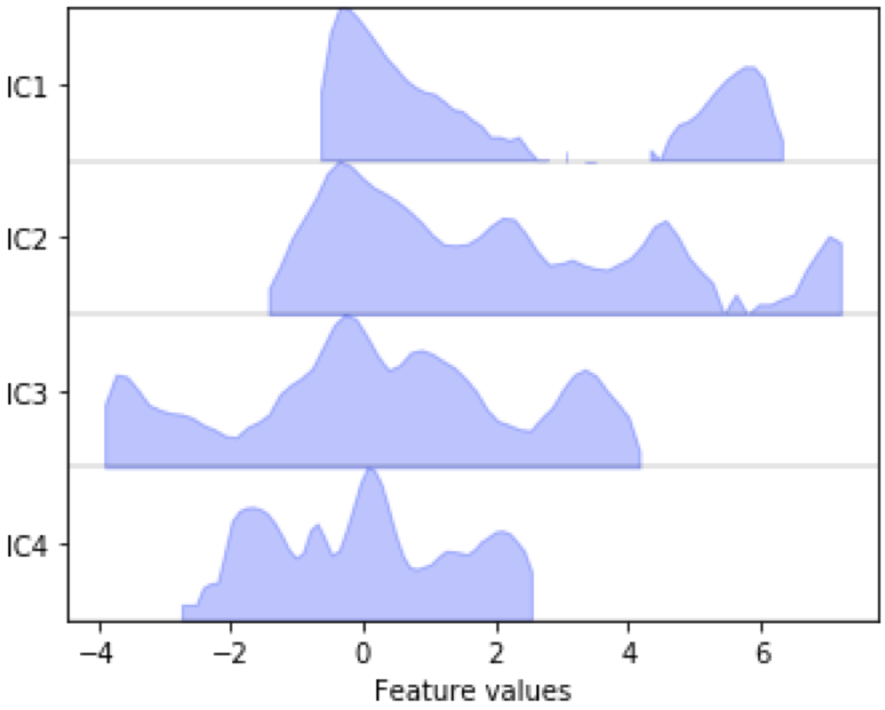
metastable states with PCCA++
TPT
Experimental observables

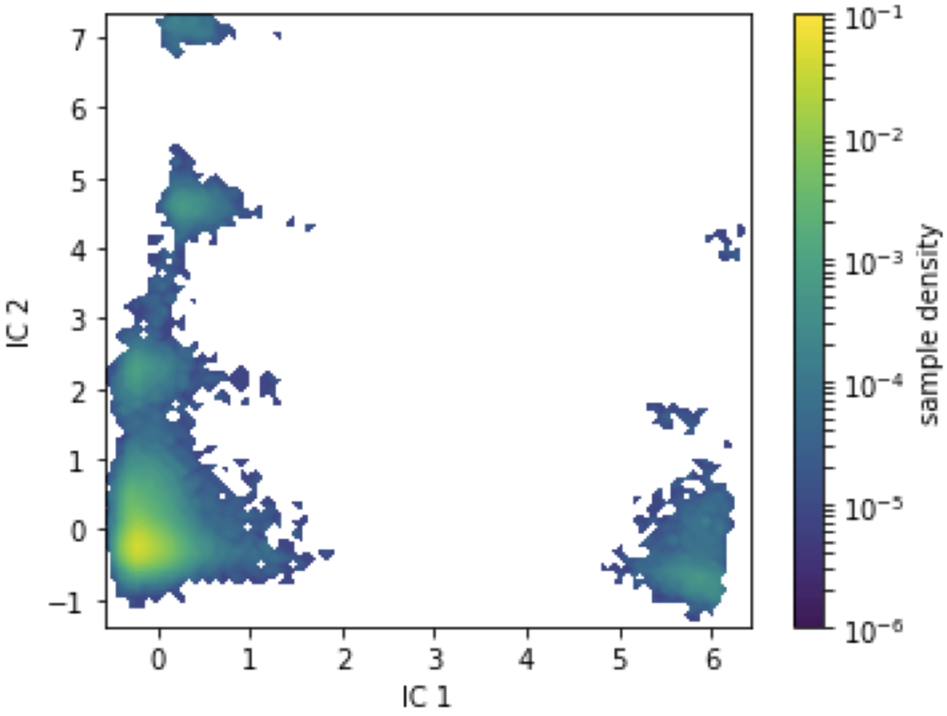
Knowledge

Dimension reduction

Find order parameters (“independent components”) that describe the slowest transitions in the MD data.

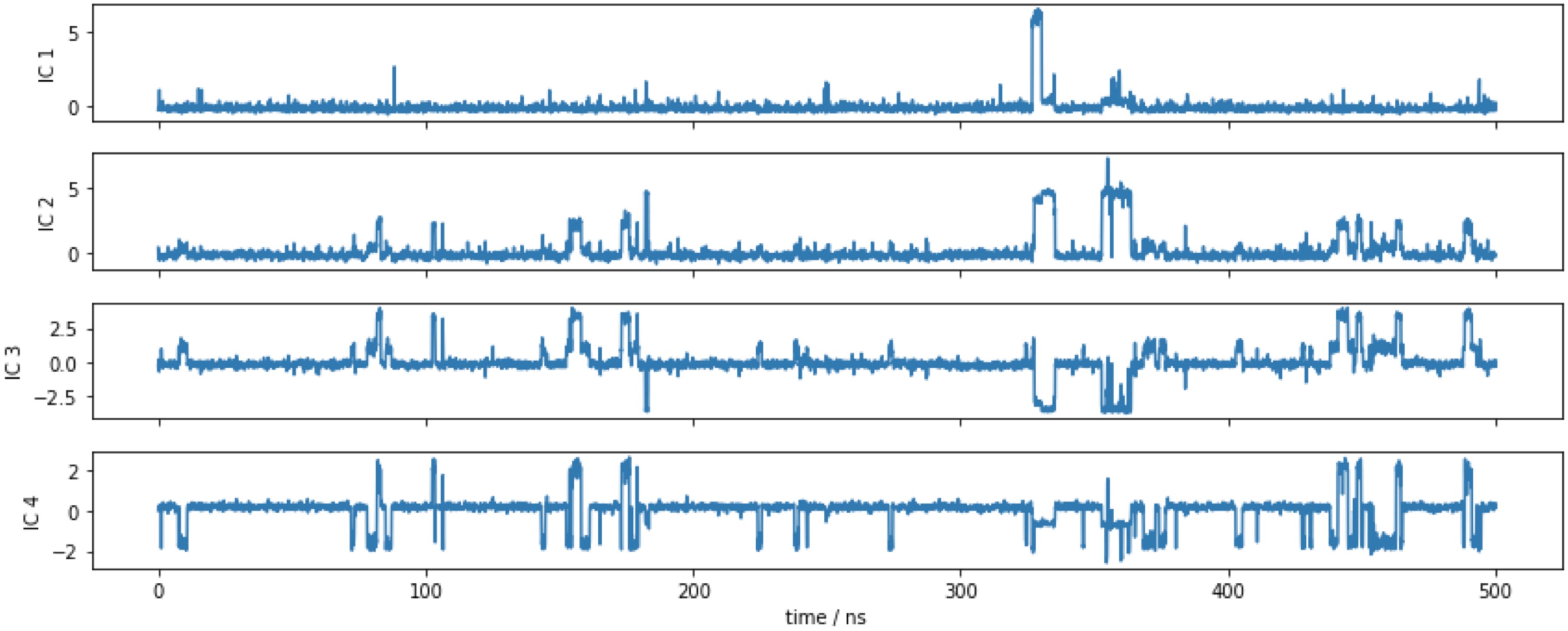
Reduction to two dimensions allows to visualize various functions of the conformational state as 2-D plots, e.g. a histogram samples



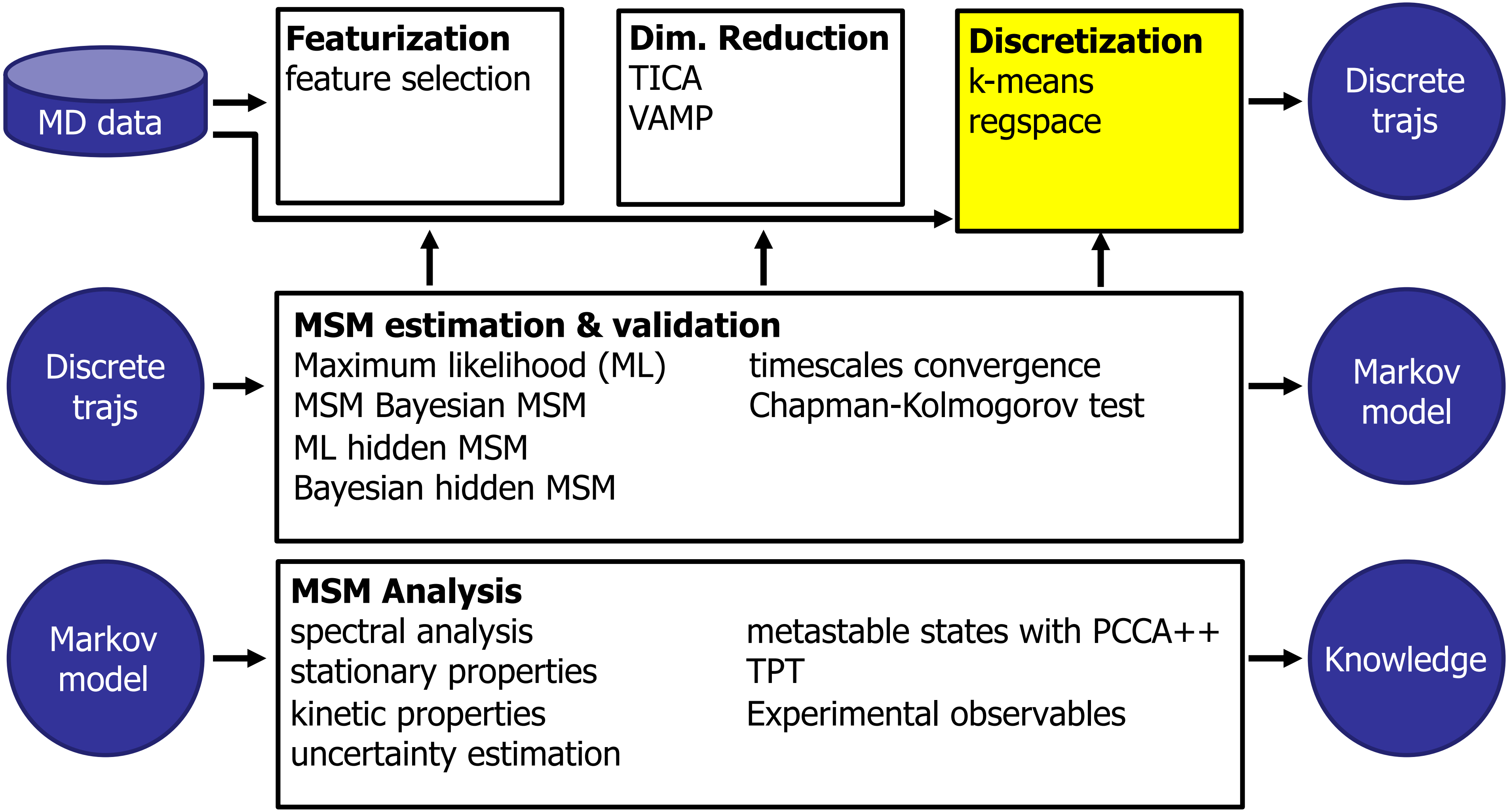


Dimension reduction

Rare events appear clearly in the time series representations of the independent components.



Markov modeling workflow



MD data

Featurization
feature selection

Dim. Reduction
TICA
VAMP

Discretization
k-means
regspace

Discrete trajs

Discrete trajs

MSM estimation & validation
Maximum likelihood (ML)
MSM Bayesian MSM
ML hidden MSM
Bayesian hidden MSM

timescales convergence
Chapman-Kolmogorov test

Markov model

Markov model

MSM Analysis
spectral analysis
stationary properties
kinetic properties
uncertainty estimation

metastable states with PCCA++
TPT
Experimental observables

Knowledge

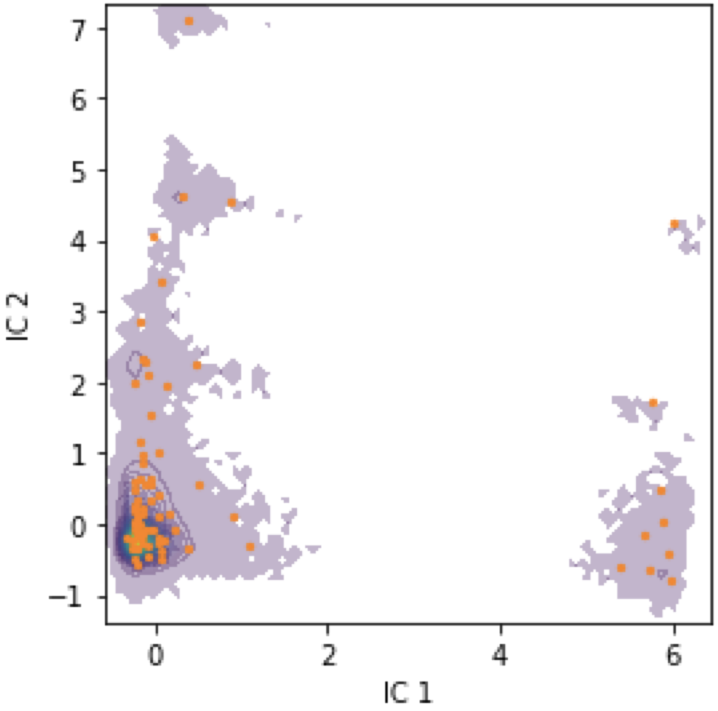
State space discretization / clustering

MSM require discretization of state space.

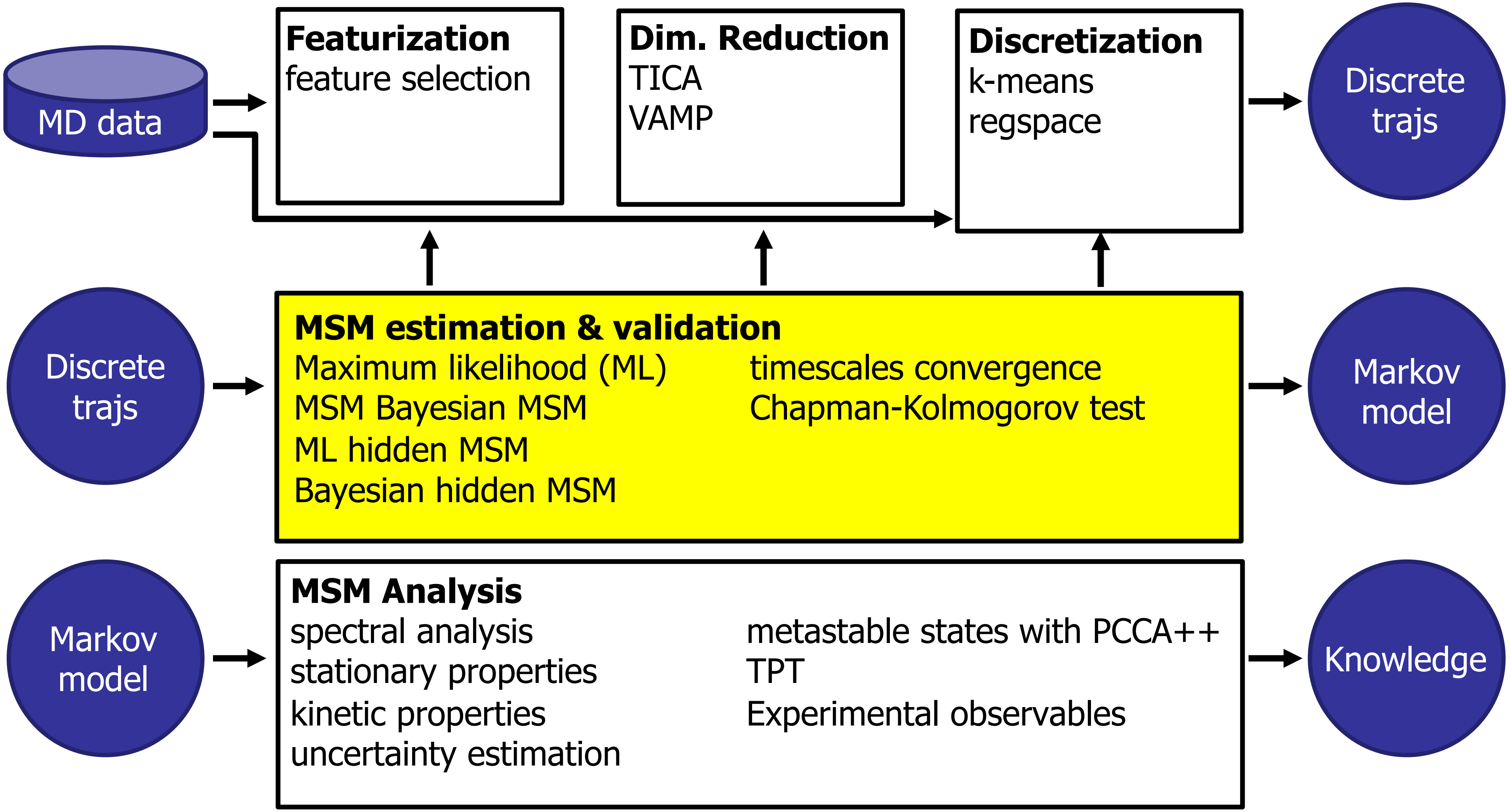
Use off-the-shelf clustering methods (k-means, …) to dissect the space into a number of non-overlapping (Voronoi) cells.

The space of independent components is already the ideal space in which to cluster.

The in the next step count transition between cells and estimate MSM.



Markov modeling workflow



MD data

Featurization
feature selection

Dim. Reduction
TICA
VAMP

Discretization
k-means
regspace

Discrete trajs

Discrete trajs

MSM estimation & validation

Maximum likelihood (ML)
MSM Bayesian MSM
ML hidden MSM
Bayesian hidden MSM

timescales convergence
Chapman-Kolmogorov test

Markov model

Markov model

MSM Analysis

spectral analysis
stationary properties
kinetic properties
uncertainty estimation

metastable states with PCCA++
TPT
Experimental observables

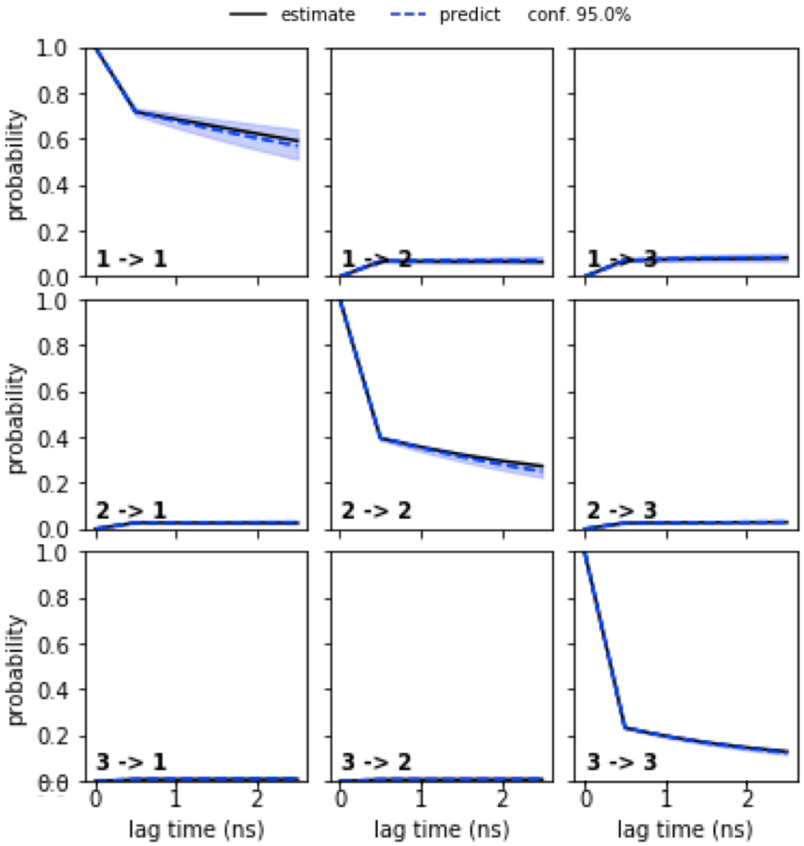
Knowledge

MSMS validation: Chapman-Kolmogorov test

The previous steps (feature selection, dimension reduction, clustering) can‘t be done with error. Already the operation of reducing the dimension introduced an error.

Errors affect the ability of the MSM to predict the future evolution of ensembles probabilities.

$$T(n\tau) \underset{?}{=} \left(T(\tau)\right)^n$$

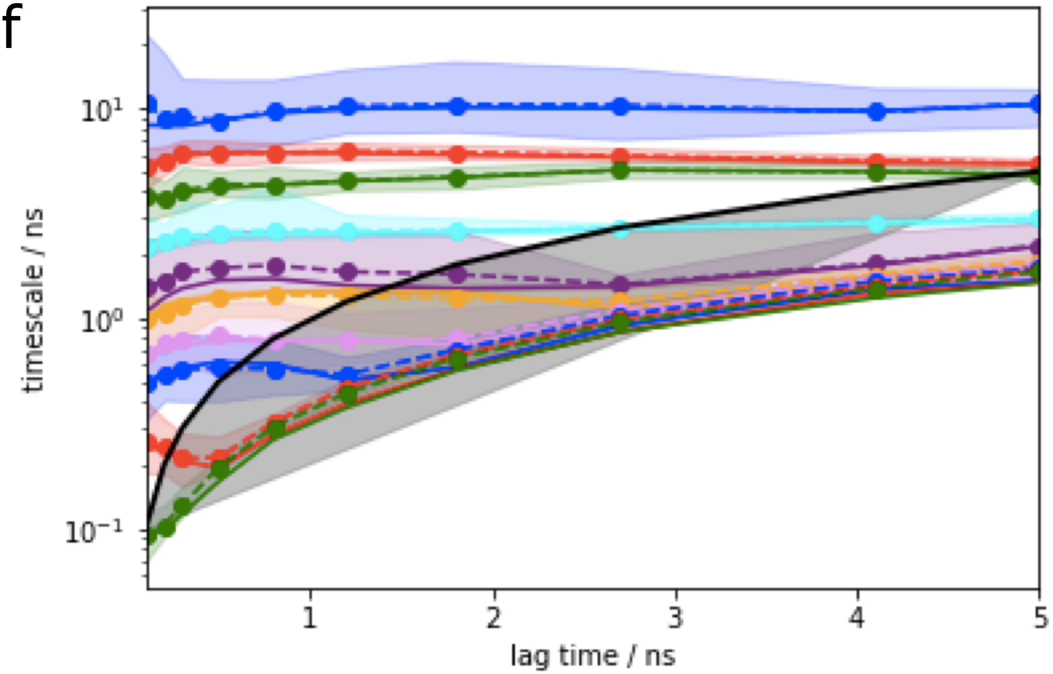


MSM validation: implied time scale test

The previous steps (feature selection, dimension reduction, clustering) can't be done with error. Already the operation of reducing the dimension introduced an error.

Errors affect the ability of the MSM to predict the future evolution of ensembles probabilities.

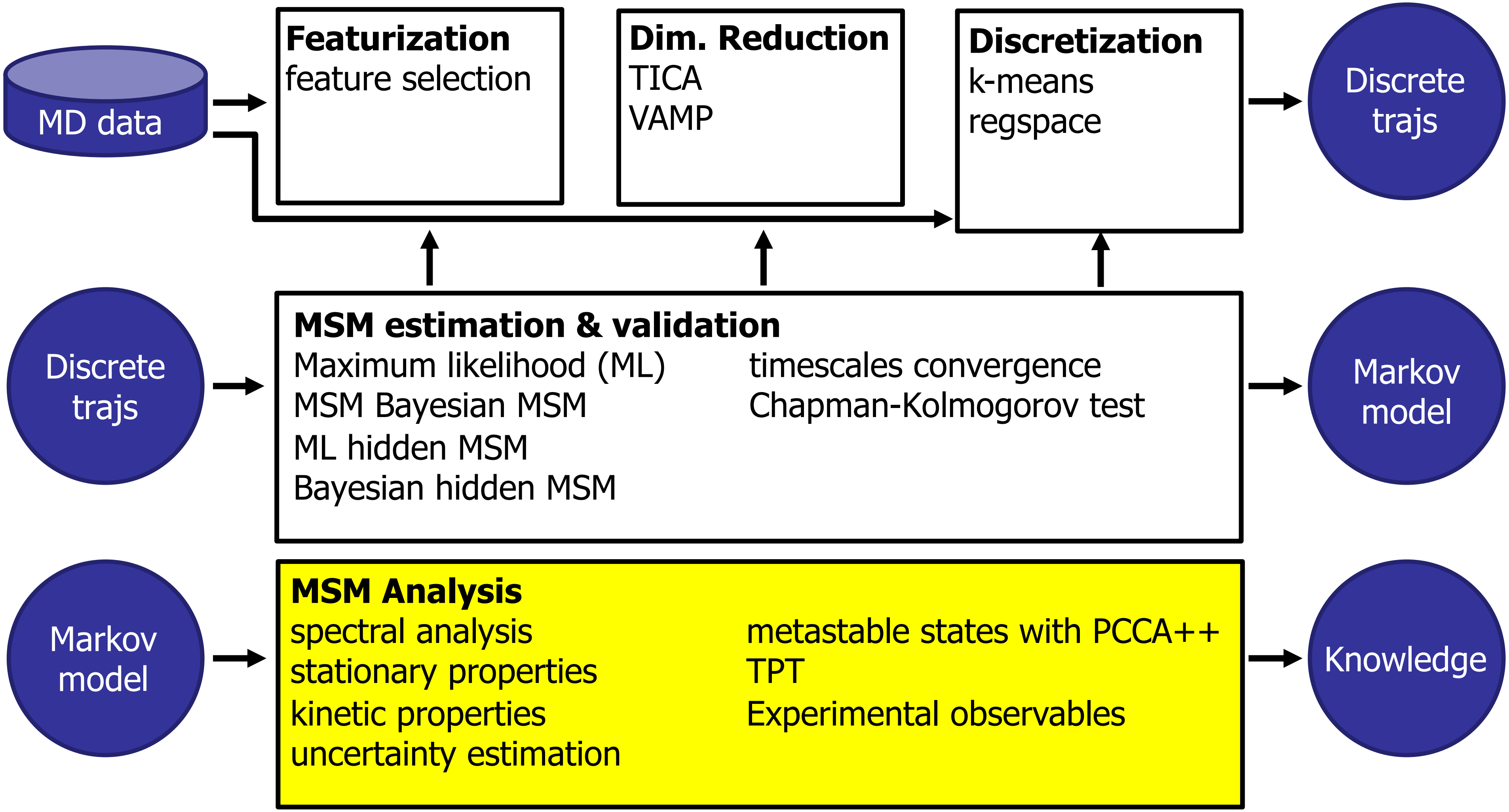
$$T(n\tau) \underset{?}{=} \left(T(\tau)\right)^n$$



Inserting the eigen-decomposition of the transition matrix, this equation can be transformed to $\lambda(n\tau) = (\lambda(\tau))^n$ or

$$\mathrm{ITS}(n\tau) = -\frac{n\tau}{\ln\lambda(n\tau)} = -\frac{n\tau}{\ln(\lambda(\tau))^n} = -\frac{\tau}{\ln\lambda(\tau)} = \mathrm{ITS}(\tau)$$

Markov modeling workflow



MD data

Featurization
feature selection

Dim. Reduction
TICA
VAMP

Discretization
k-means
regspace

Discrete trajs

Discrete trajs

MSM estimation & validation
Maximum likelihood (ML)
MSM Bayesian MSM
ML hidden MSM
Bayesian hidden MSM

timescales convergence
Chapman-Kolmogorov test

Markov model

Markov model

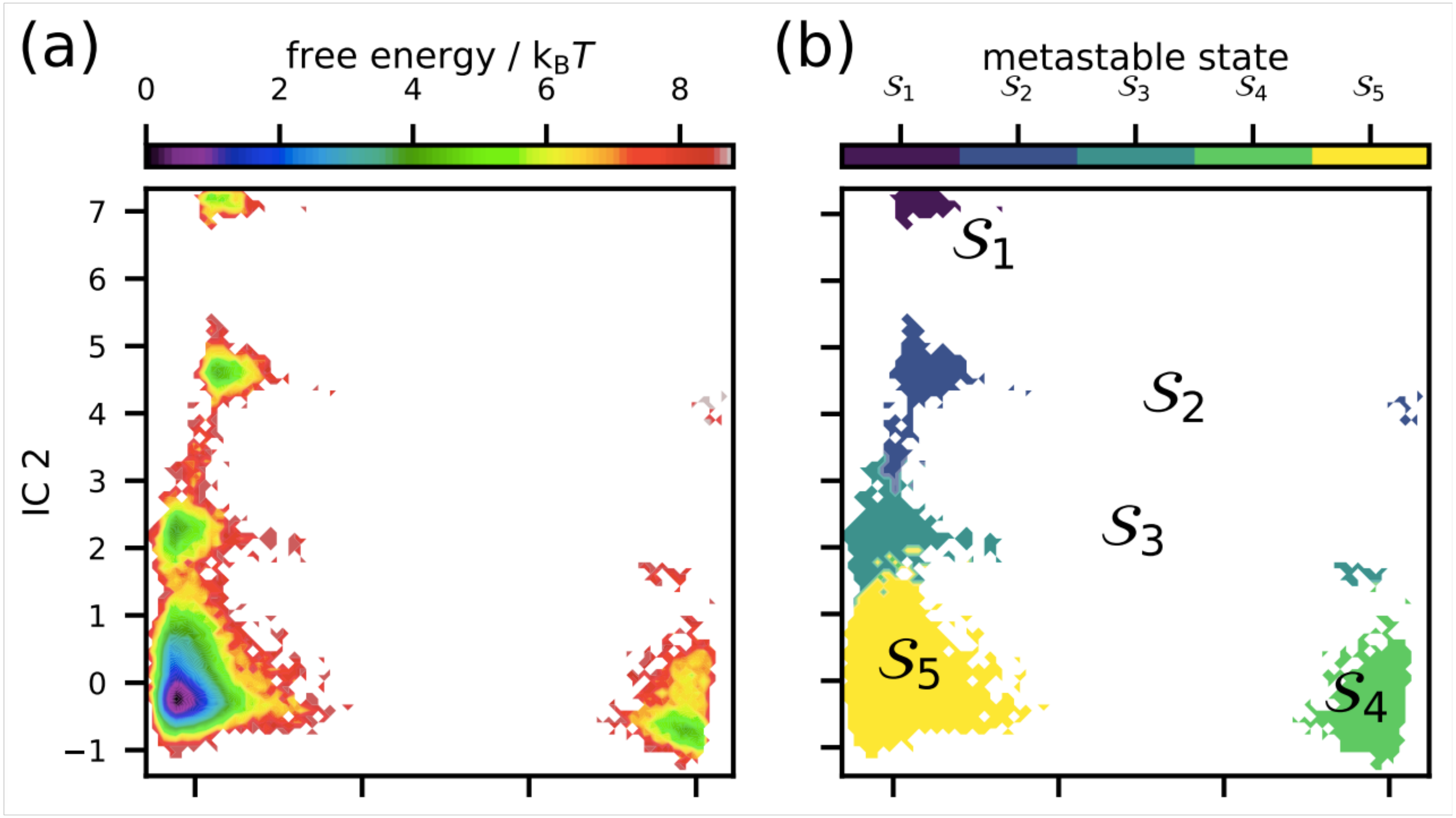
MSM Analysis
spectral analysis
stationary properties
kinetic properties
uncertainty estimation

metastable states with PCCA++
TPT
Experimental observables

Knowledge

MSM analysis: free energy landscapes

- (a) Reweighted free energy surface projected onto the first two independent components exhibits five minima which
- (b) PCCA++ identifies the five minima as metastable states.

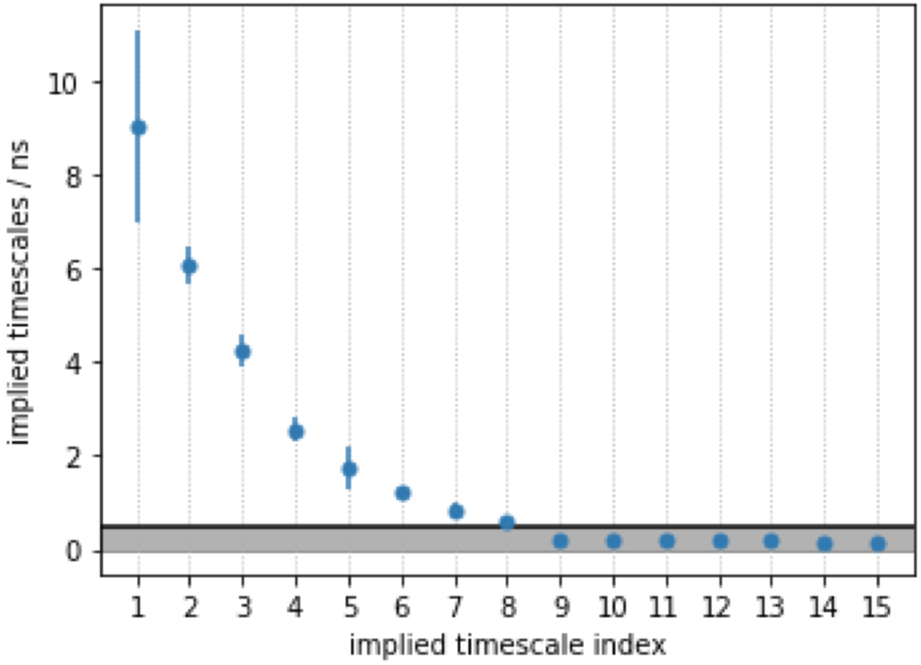


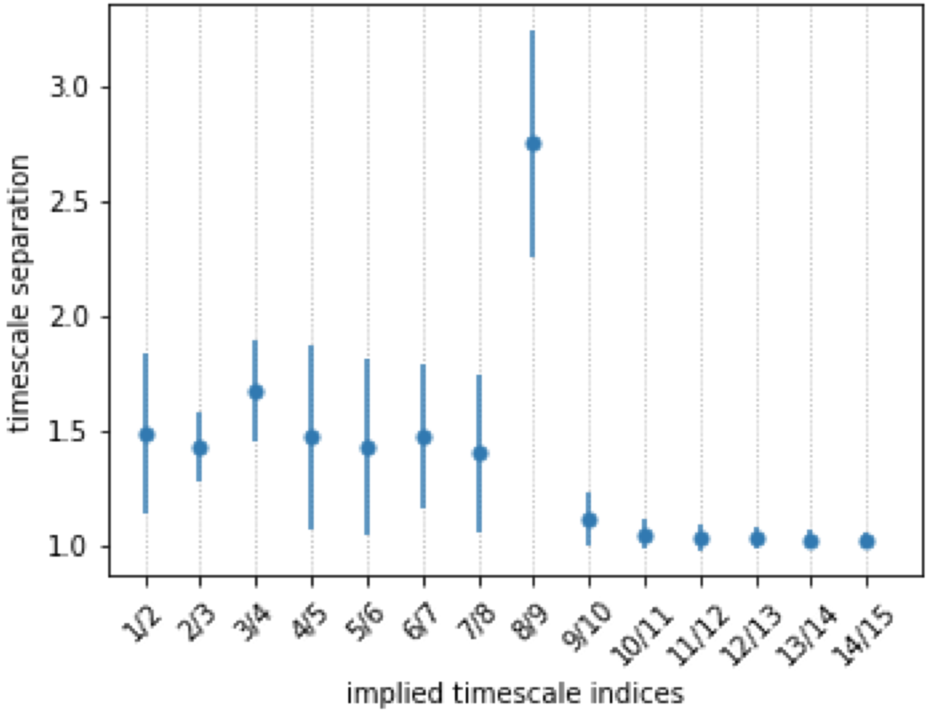
MSM analysis: relaxation processes

The eigendecomposition of the transition matrix yields:

- Eigenvalues that encode the relaxation timescales (time, the system takes to return to equilibrium "implied timescales") and
- Eigenvectors that encode the conformations between which probability is moved as the system relaxes to equilibrium.

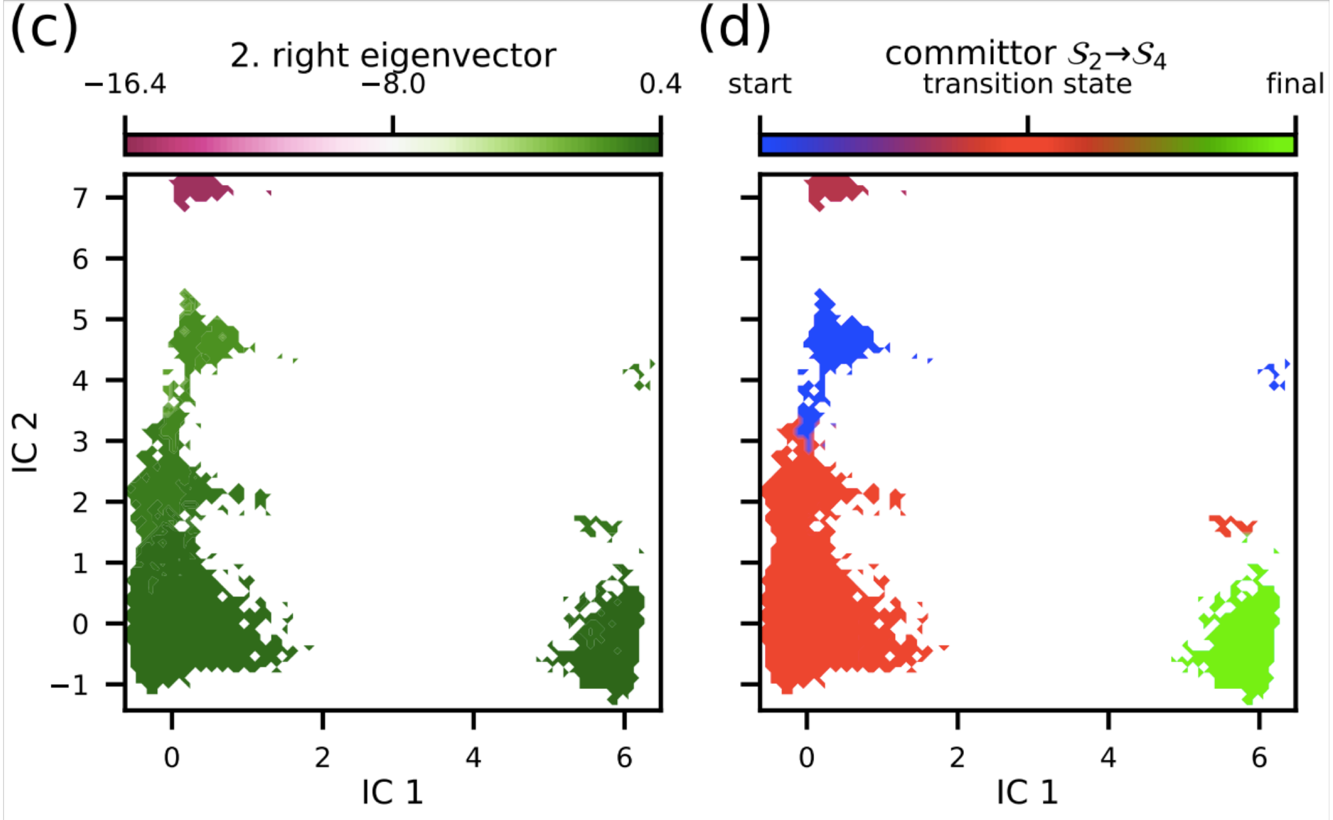
If there is a gap in between ITS, one can truncate the spectrum.



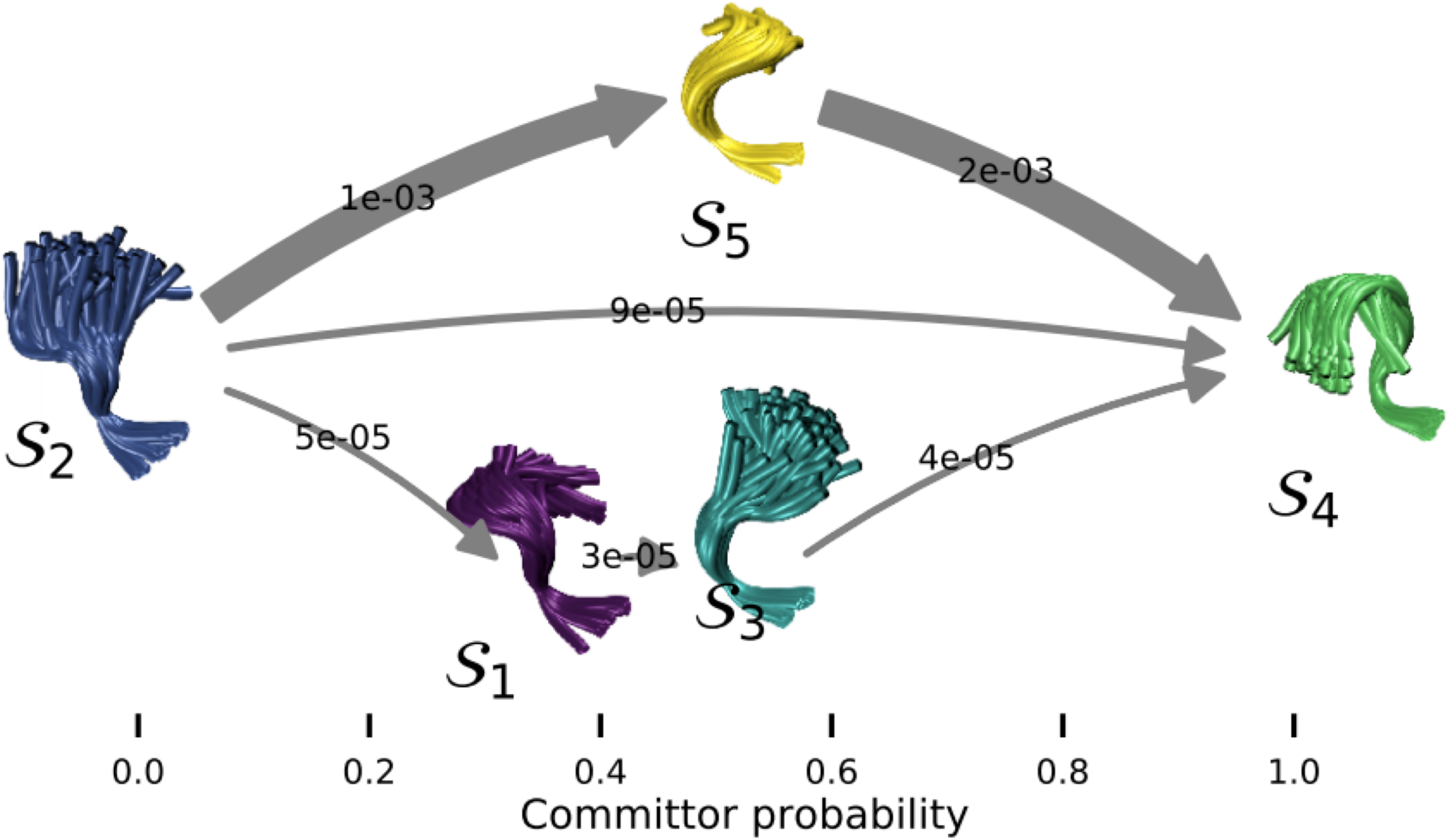


MSM analysis: relaxation processes

- (c) The second right eigenvector shows that the slowest process shifts probability between the least probable state (S1) and the other states, in particular states (S4, S5), whereas
- (d) the committor S2 → S4 indicates that states S(1,3,5) act as a transition region between states S2 and S4.



Transition path theory / analysis

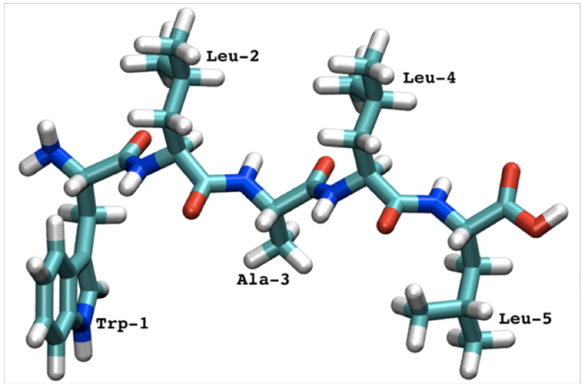


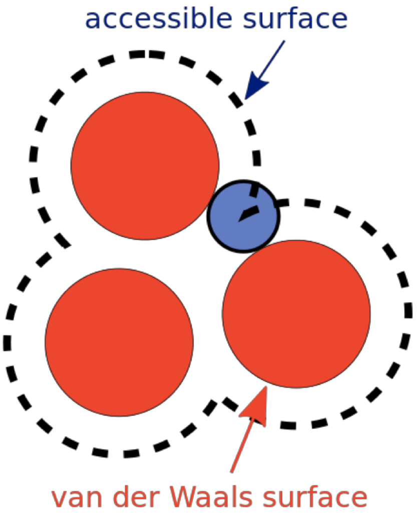
MSM analysis: Experimental observables

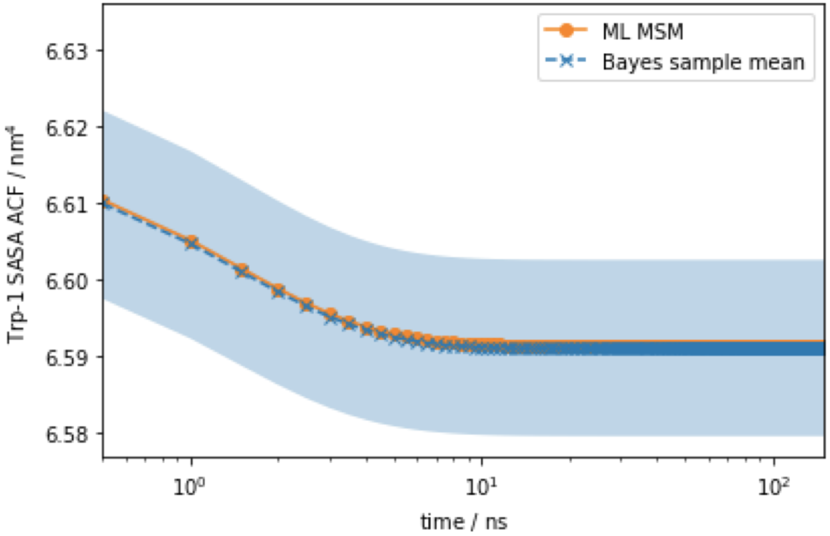
$$\mathrm{afc}(x;\tau) = \frac{\mathbf{x}^{\top}\mathrm{diag}(\mathbf{p})\mathbf{T}\mathbf{x}}{\mathbf{x}^{\top}\mathrm{diag}(\mathbf{p})\ \mathbf{x}}$$

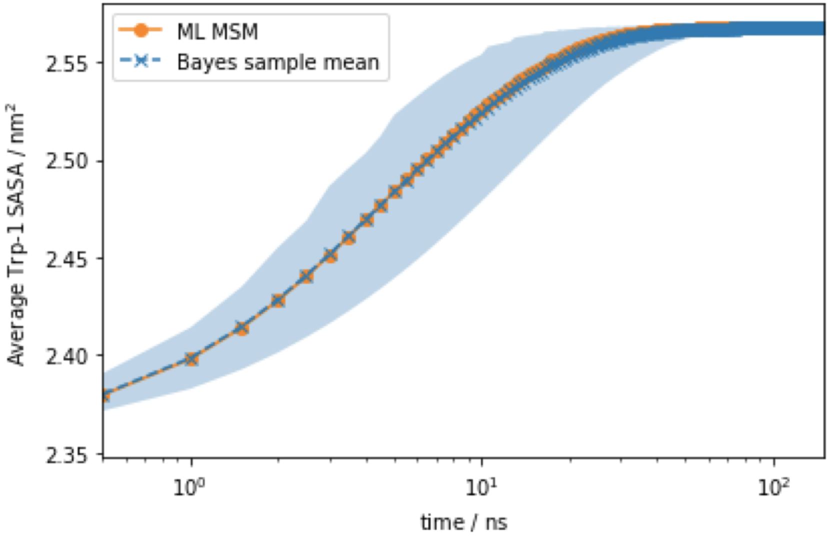
Example analysis of the conformational dynamics of a pentapeptide backbone:

- (a) the Trp-1 SASA autocorrelation function yields a weak signal which, however,
- (b) can be enhanced if the system is prepared in the nonequilibrium condition S1.

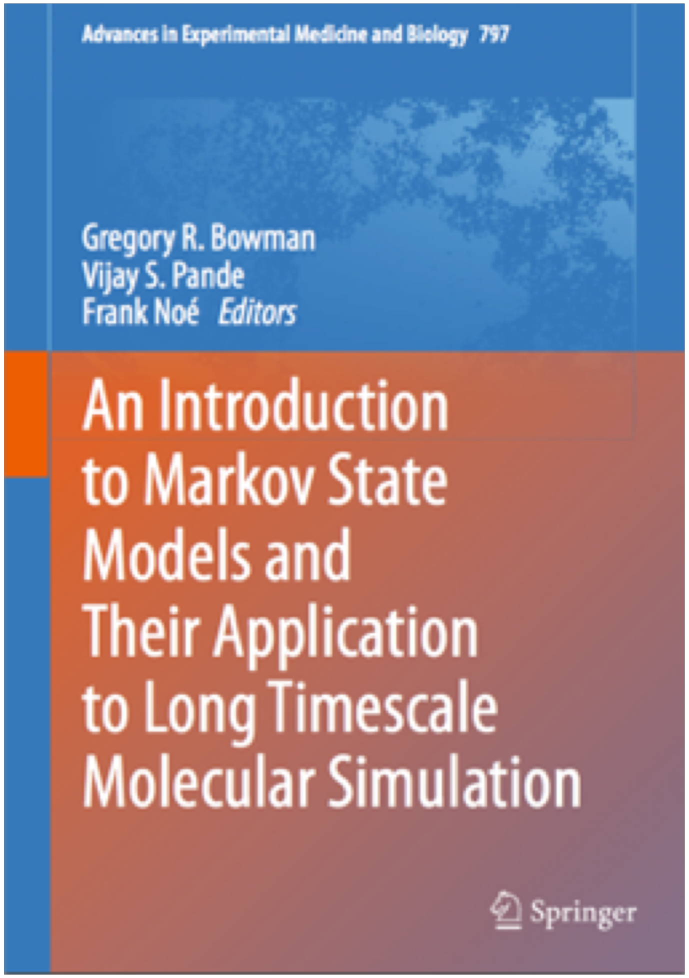








Review book



Advances in Experimental Medicine and Biology 797

Gregory R. Bowman
Vijay S. Pande
Frank Noé *Editors*

An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation

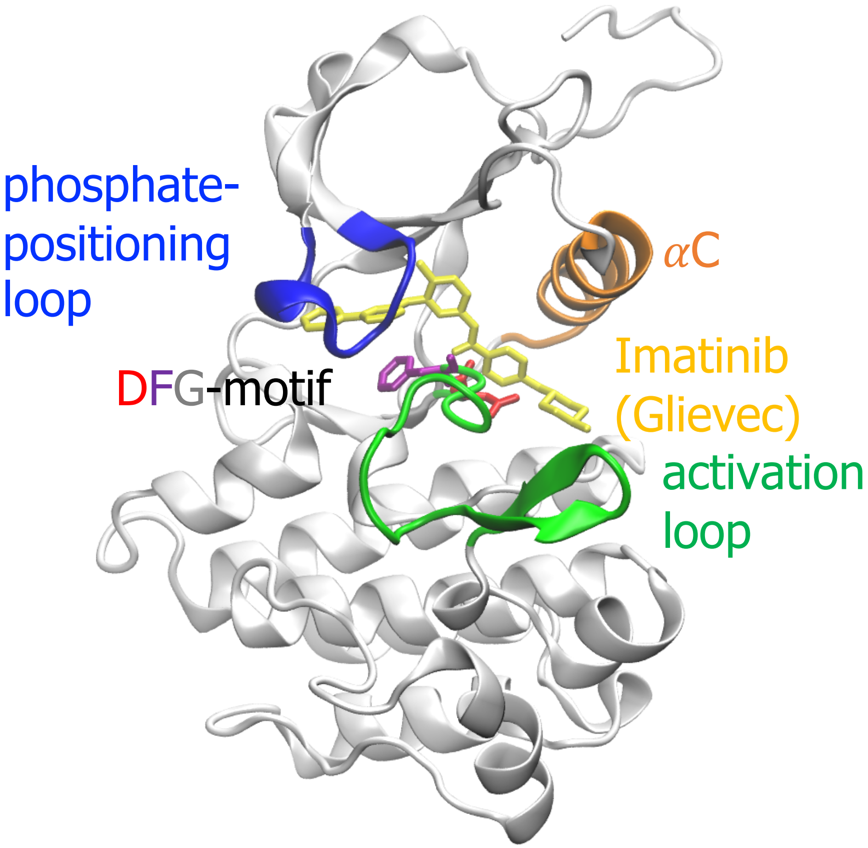
Springer

Thanks!

Prof Frank Noé, Martin Scherer, Simon Olsson, Christoph Wehmeyer, Tim Hempel, Brooke Husic, Moritz Hoffmann, Sebastian Stolzenberg and the whole Pyemma team.

Thank you for your attention!

Job advertisement



Open postdoc position in the lab of Prof Benoît Roux, University of Chicago, USA starting March 2020.

- Process of Imatinib-Abl kinase binding/ conformational change.
- Covalent kinase inhibitors.