Tim Hempel
Noé & Clementi Groups
Department of Mathematics and Computer Science
Department of Physics

Markov models for molecular dynamics

Markov School 2022

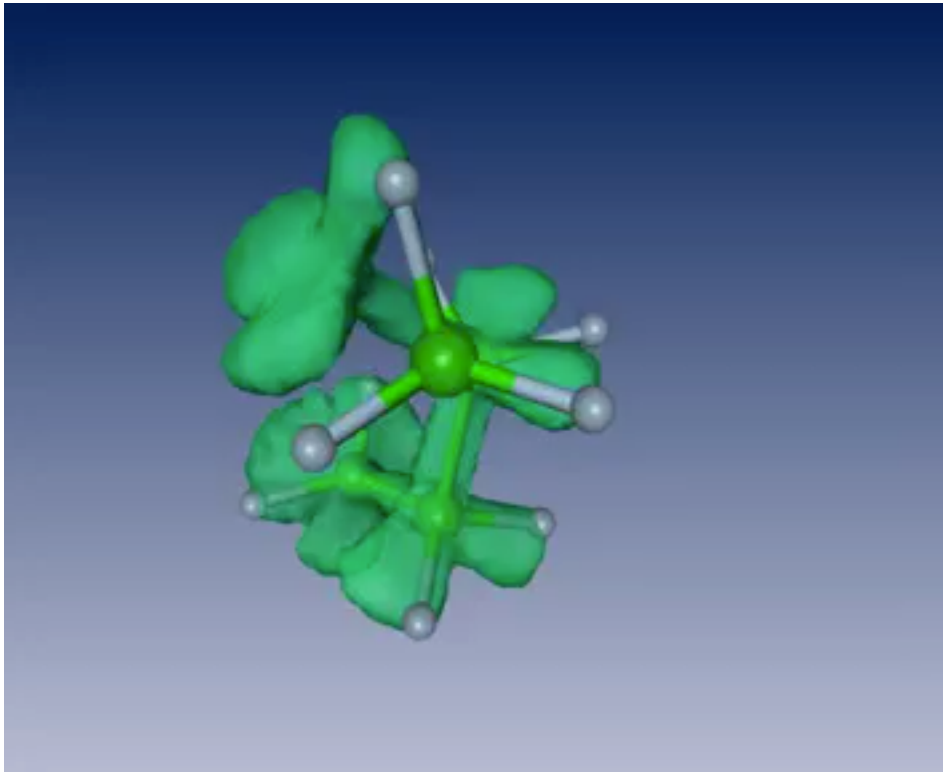
Slides by Jan-Hendrik Prinz, Simon Olsson & Tim Hempel

Conformational dynamics & Markov state models

Peptide dynamics

MD simulation of n-Butane

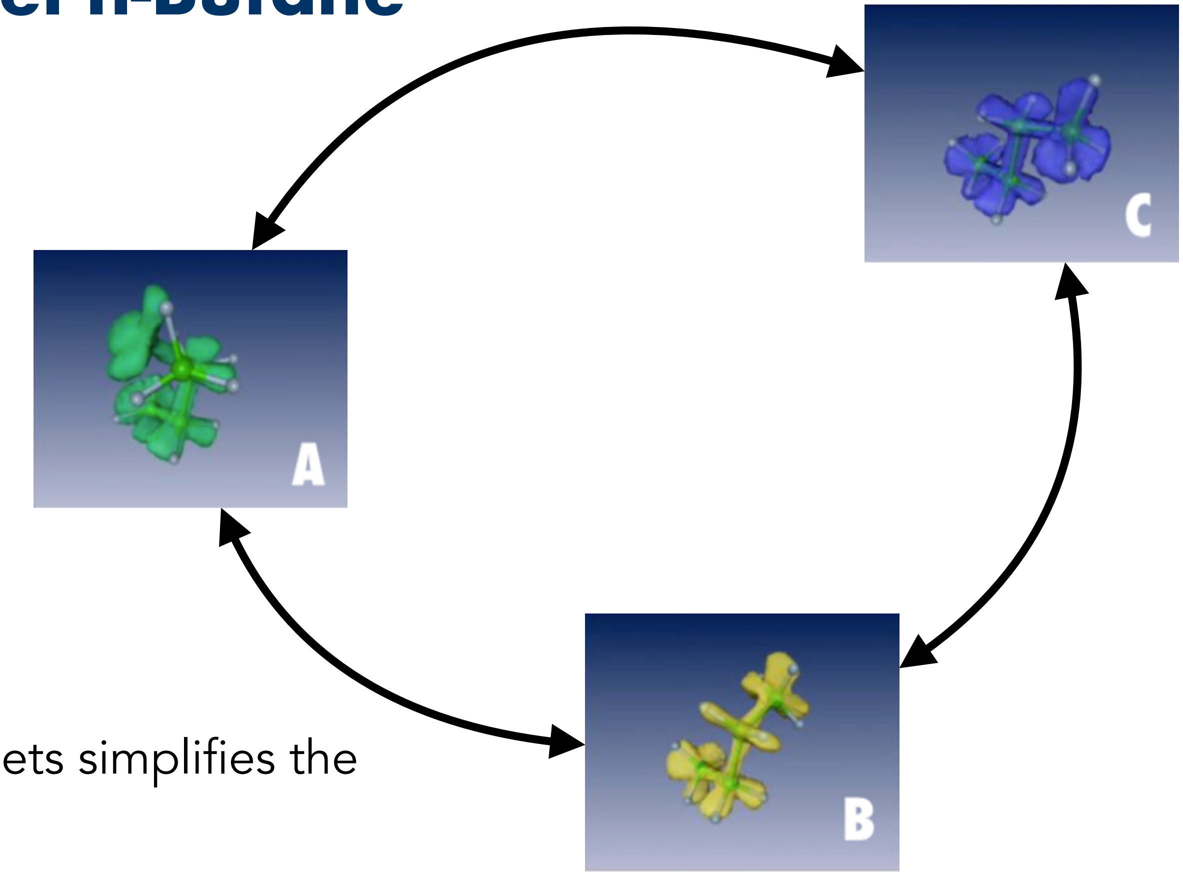
(14 atoms)



The peptide shows metastable dynamics (color-coded).

Example courtesy of JH Prinz

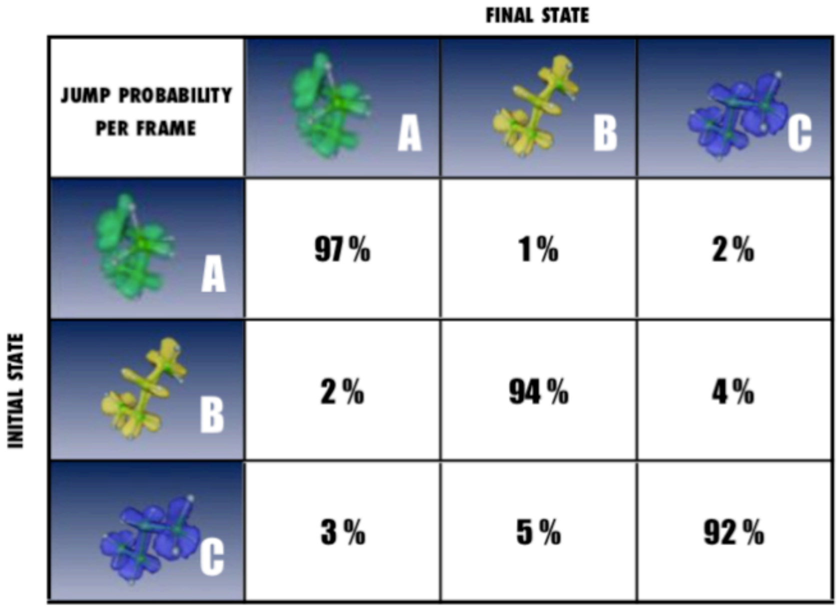
Simplified model n-Butane



Definition of metastable sets simplifies the dynamics substantially

-> transitions between „macro states“

Markov State Model of n-Butane



FINAL STATE

INITIAL STATE

JUMP PROBABILITY PER FRAME	A	B	C
A	97 %	1 %	2 %
B	2 %	94 %	4 %
C	3 %	5 %	92 %

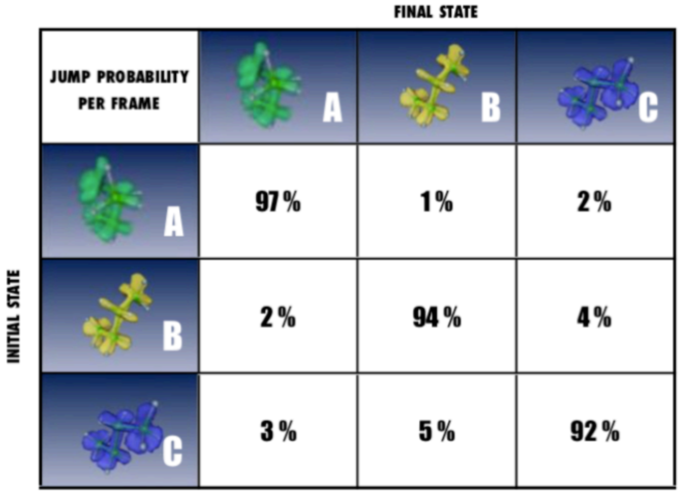
This is a matrix of conditional jump probabilities between macro states.

It is called an MSM transition matrix between metastable sets.

Markov State Model of n-Butane

How did we get there?

- Identify metastable states („assign colors")
- Estimate the transition probabilities.



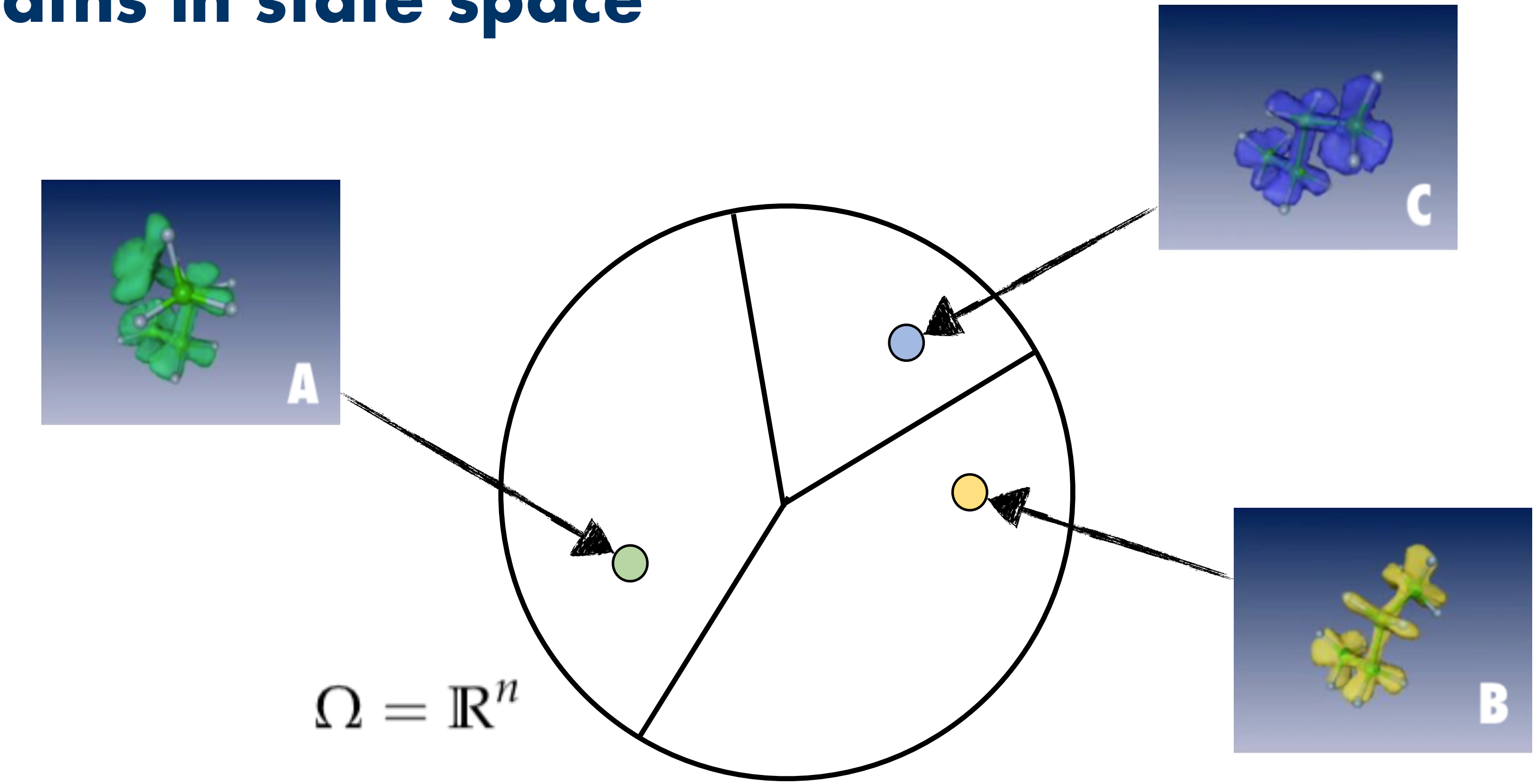
FINAL STATE

INITIAL STATE

JUMP PROBABILITY PER FRAME	A	B	C
A	97 %	1 %	2 %
B	2 %	94 %	4 %
C	3 %	5 %	92 %

Markov processes

Paths in state space

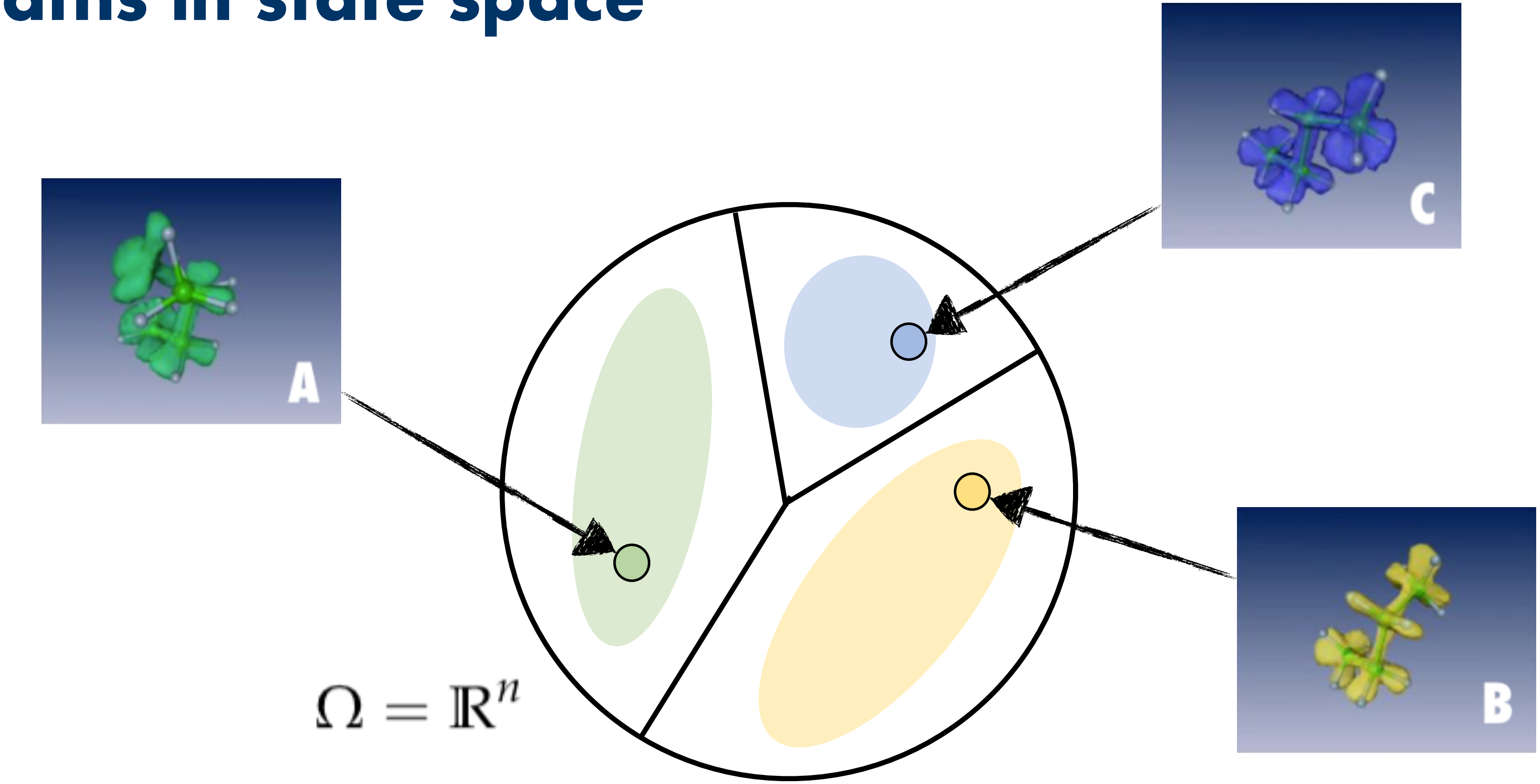


$$\Omega = \mathbb{R}^n$$

Points x in state space Ω correspond to conformations.

A trajectory is a path in state space.

Paths in state space



Only changes between long-living sets (color-coded) are interesting for us

-> metastability

Observations as stochastic process

View MD simulation as realization of a stochastic process

$$x : \underset{\substack{\uparrow \\ \text{time}}}{t} \in \mathbb{R}^{+} \mapsto x_t \in \underset{\substack{\nwarrow \\ \text{State space}}}{\Omega}$$

in a probability space.

Observations as stochastic process

View MD simulation as realization of a stochastic process

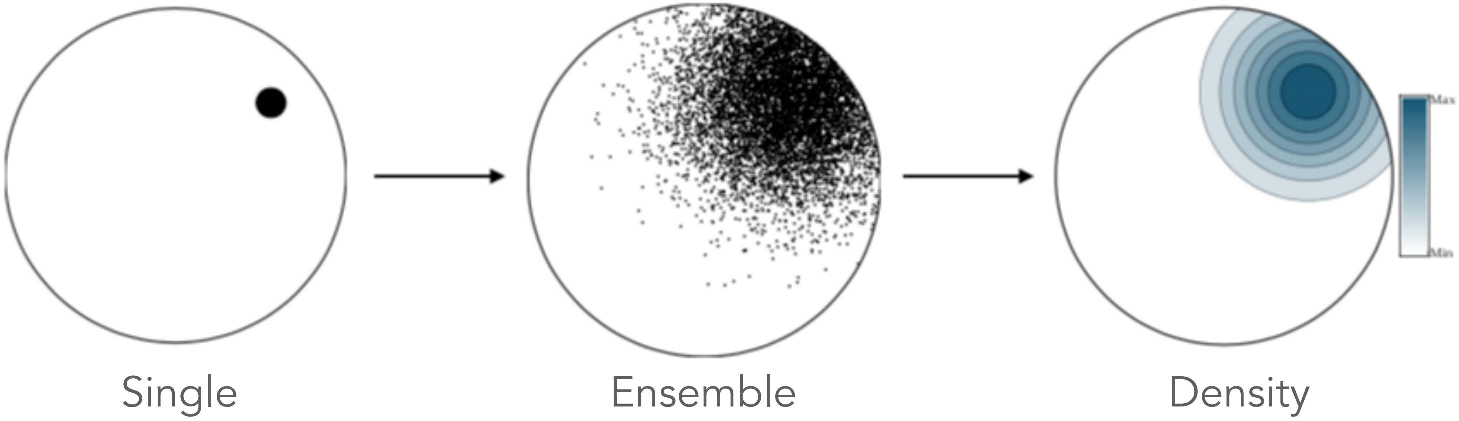
$$x \,:\, \underset{\substack{\uparrow\\ \text{time}}}{t} \in \mathbb{R}^{+} \mapsto x_t \in \underset{\substack{\nwarrow\\ \text{State space}}}{\Omega}$$

in a probability space.

We assume the probability space to be „nice“, such that continuous transition probability function can be defined.

$$\int_{\mathrm{A}} \mathrm{d}y\, \mathrm{p}_\tau(x,y) = \mathbb{P}[x_{t+\tau} \in \mathrm{A} | x_t = x], \quad \forall \mathrm{A} \in \mathcal{B}(\Omega), \forall t \geq 0$$

Modeling the density



Single

Ensemble

Density

Describe ensembles of configurations in Ω by a probability function

$$\mathrm{p} : x \in \Omega \mapsto \mathrm{p}(x) \in \mathbb{R}_0^+, \quad \int_\Omega \mathrm{d}x\, \mathrm{p}(x) = 1$$

Figure courtesy of JH Prinz

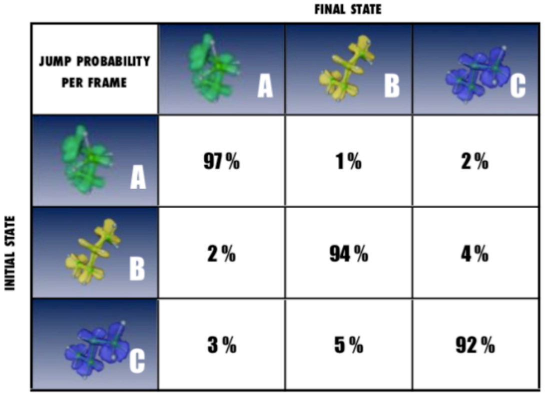
Modeling the dynamics

Assumption: The dynamics is Markovian

- there is no memory

$$\mathbb{P}[x_{t+\tau} \in \mathrm{A} | x_{t_1}, \ldots, x_t = x] = \mathbb{P}[x_{t+\tau} \in \mathrm{A} | x_t = x], \forall \mathrm{A} \subset \Omega, \forall t \geq 0$$

- We can write a transition matrix with conditional probabilities to model the system dynamics
- Chapman-Kolmogorov property connects jump probabilities for different lag times τ



		FINAL STATE		
INITIAL STATE	JUMP PROBABILITY PER FRAME	A	B	C
	A	97 %	1 %	2 %
	B	2 %	94 %	4 %
	C	3 %	5 %	92 %

$$\mathrm{p}_{\tau_1+\tau_2}(x, A) = \int_\Omega \mathrm{d}y\, \mathrm{p}_{\tau_1}(x, y)\, \mathrm{p}_{\tau_2}(y, A), \quad \forall A \subset \Omega.$$

$$P(k \cdot \tau) = P(\tau)^k$$

Assumptions I

Irreducibility

All states in state space can be reached from another in finite time.

$$\forall x \in \Omega, A \subset \Omega, \exists t < \infty \text{ s.t. } \mathrm{p}_t(x, \mathrm{A}) > 0$$

- ensures unique equilibrium probability π

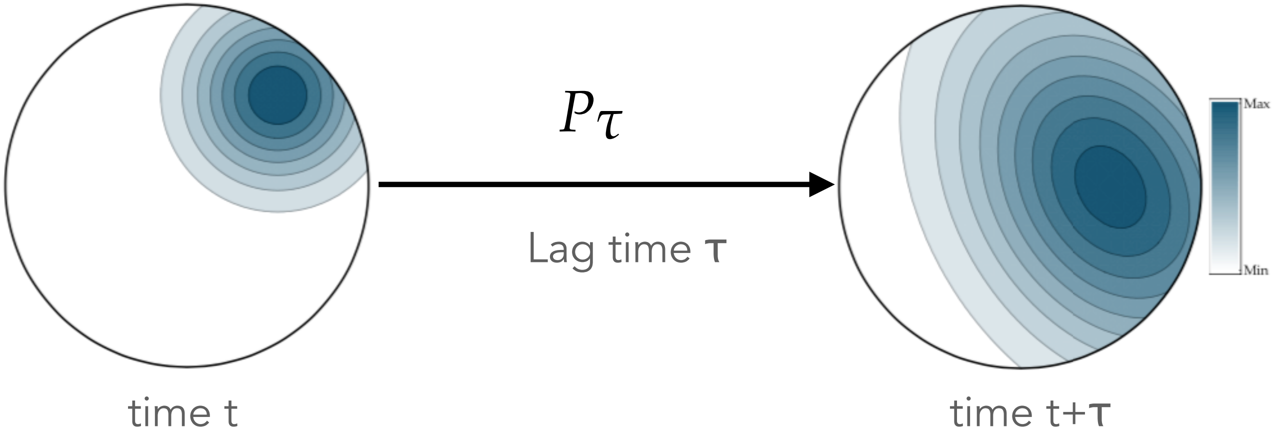
Ergodicity

- everything is connected („all states are accessible“)
- no cyclic dynamics („all states mix“)

$$\lim_{T\to\infty} T^{-1} \int_0^T \mathrm{d}t\, f(x_t) = \int_\Omega \mathrm{d}x\, \underset{\uparrow}{\pi(x)} f(x)$$

Stationary distribution of the Markov model

Propagator



Define the propagator as an operator that transports probability distributions in time

$$\mathrm{p}_{t+\tau}(x) = [\mathrm{P}_\tau \, \mathrm{p}_t](x) = \int_\Omega \mathrm{d}y \, \mathrm{p}_\tau(y,x) \mathrm{p}_t(y)$$

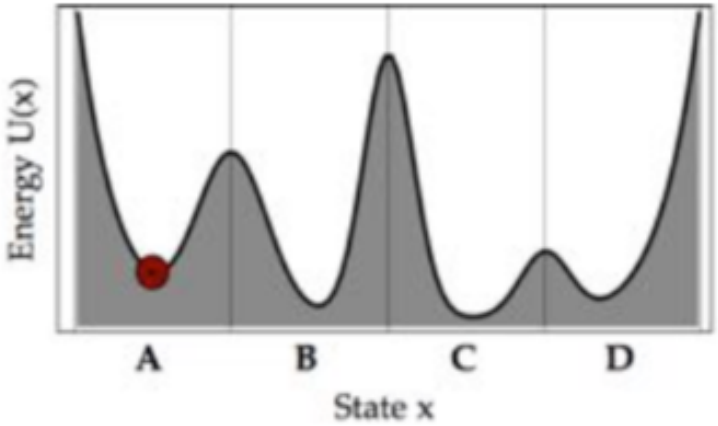
Example courtesy of JH Prinz

Example dynamics

Simple Brownian dynamics in a 1D potential

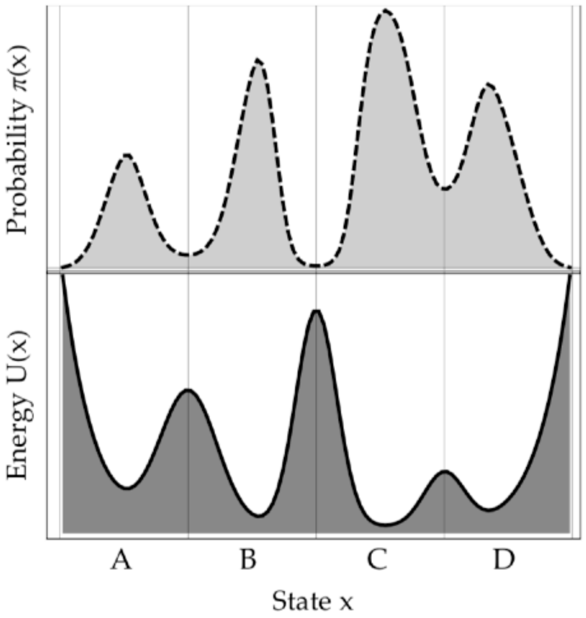
$$\gamma \, \mathrm{d}x_t = -\nabla U(x_t)\mathrm{d}t + \sigma \mathrm{d}W_t$$

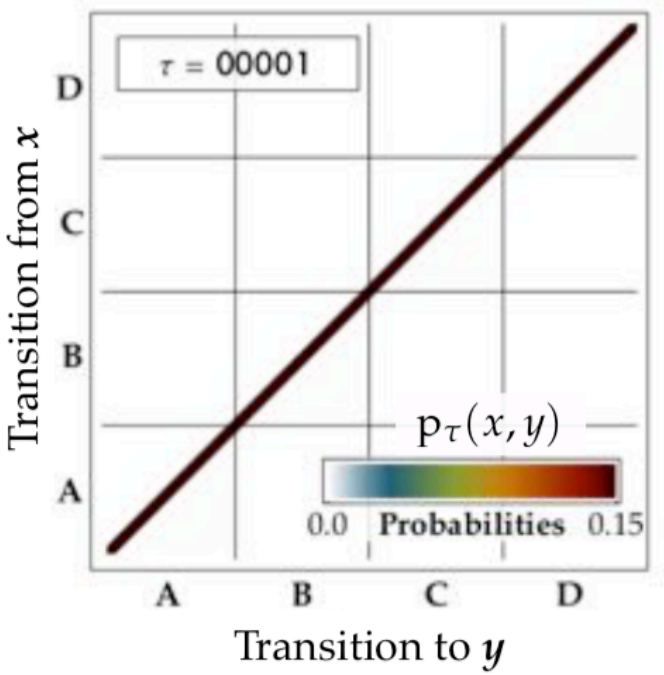
Potential landscape U(x) shows 4 distinct basins (metastable sets)



Lagtime dependence

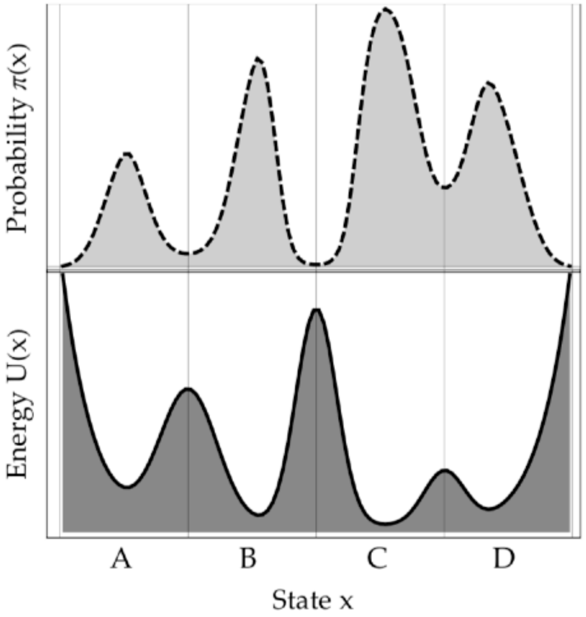
The propagator depends on the chosen lag time

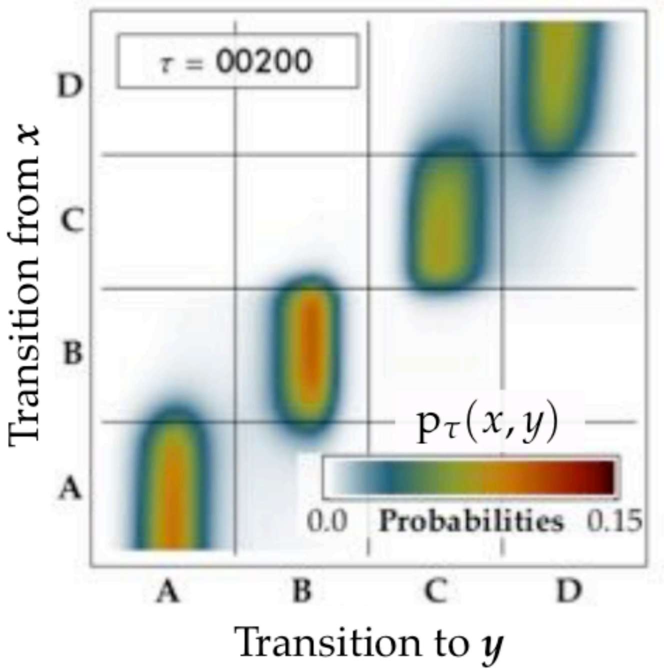




Lagtime dependence

The propagator depends on the chosen lag time

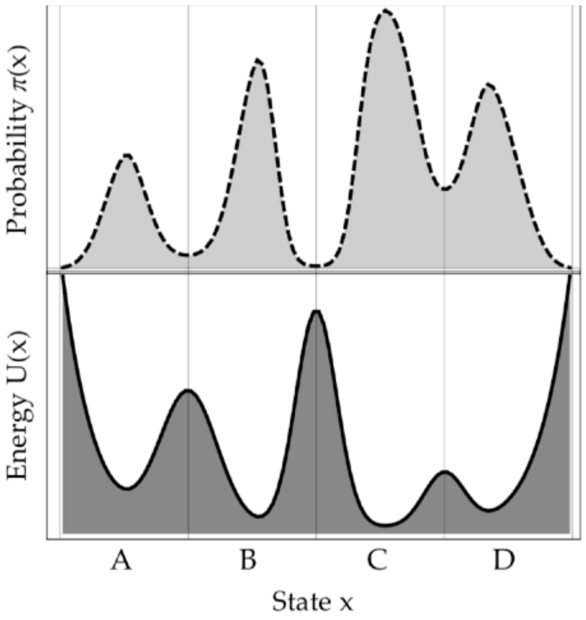


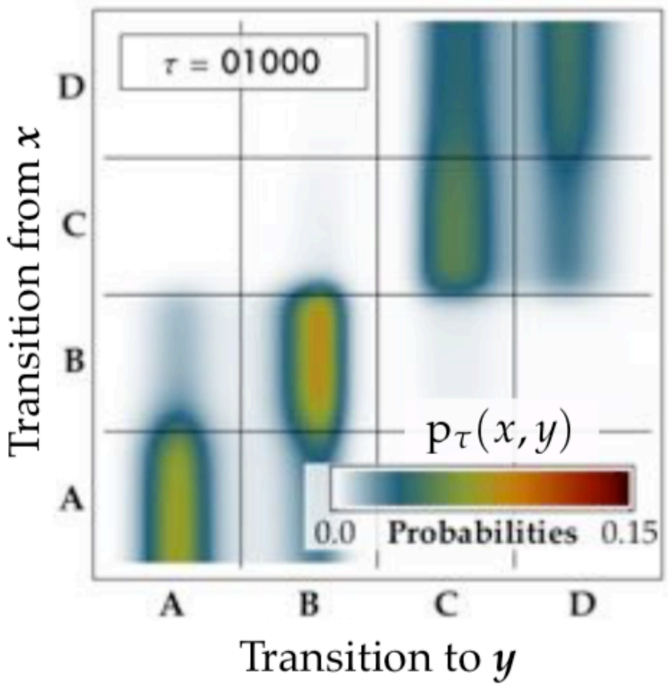


4 metastable sets {A}, {B}, {C}, {D}

Lagtime dependence

The propagator depends on the chosen lag time

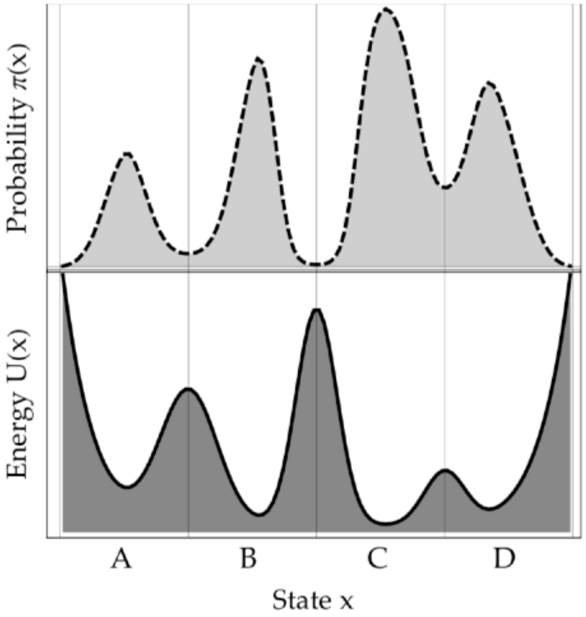


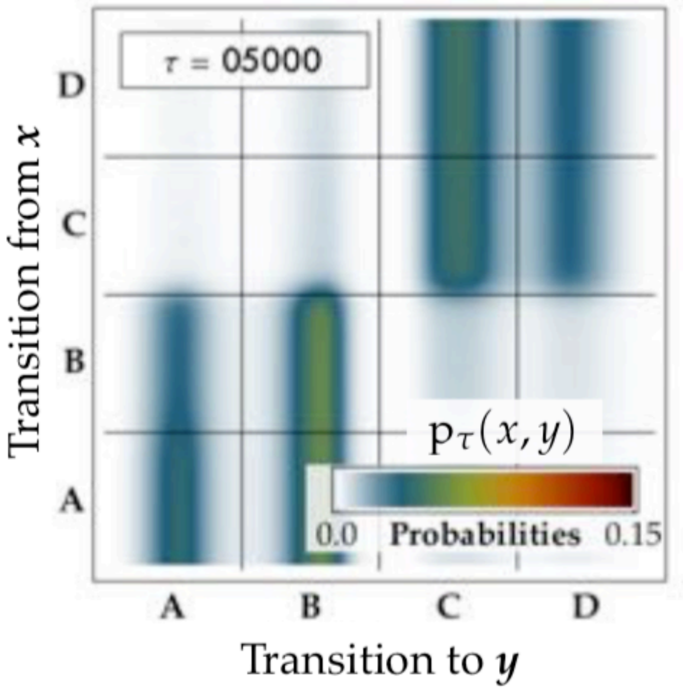


3 metastable sets {A}, {B}, {C, D}

Lagtime dependence

The propagator depends on the chosen lag time

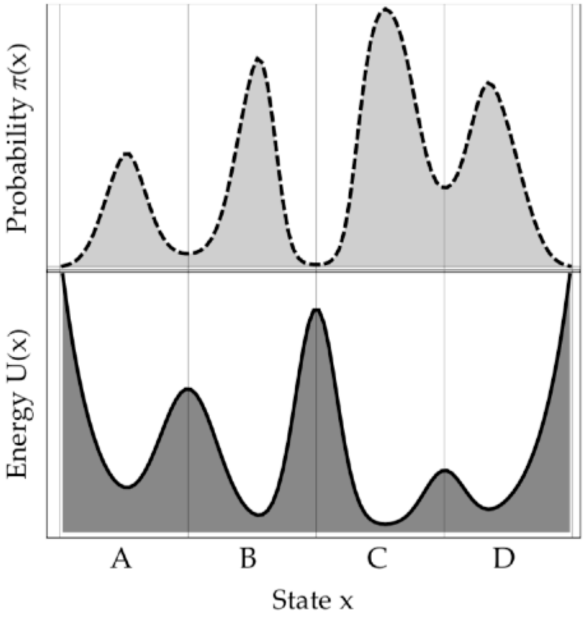


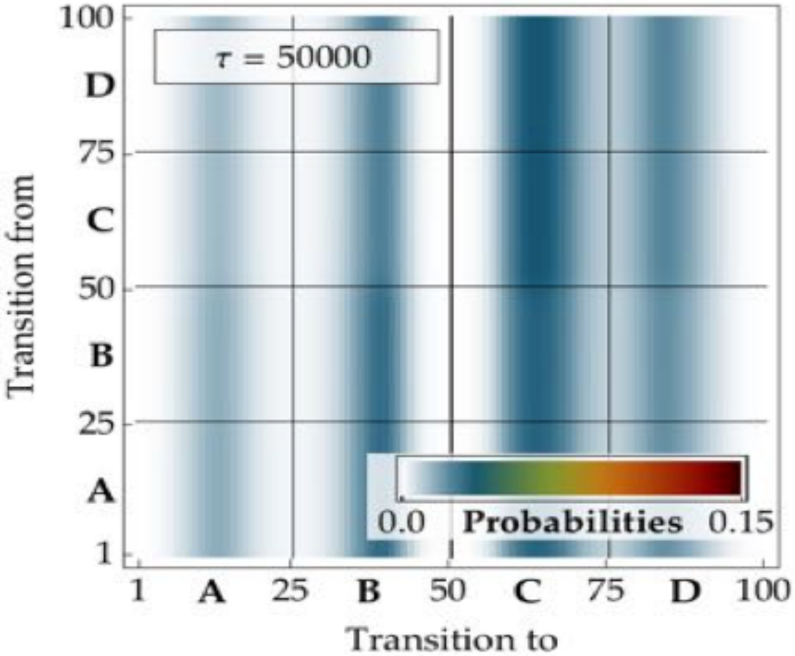


2 metastable sets {A, B}, {C, D}

Lagtime dependence

The propagator depends on the chosen lag time





1 metastable set {A, B, C, D}

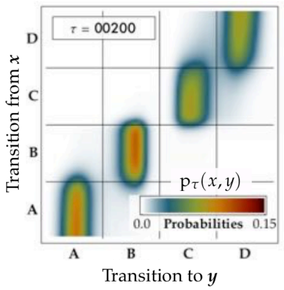
Eigenspectrum of the propagator

Eigenvalues

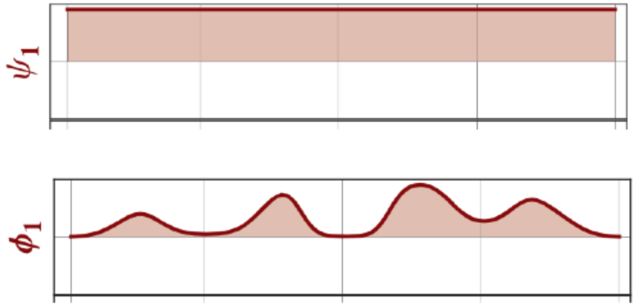
$$\mathrm{P}_\tau \, \phi_i = \lambda_i \, \phi_i$$

- The first eigenvalue is always 1
- All other eigenvalues are < 1

The first eigenvector (with eigenvalue 1) corresponds to the *stationary distribution* that is often denoted by a π.



→



First right eigenvector

First left eigenvector

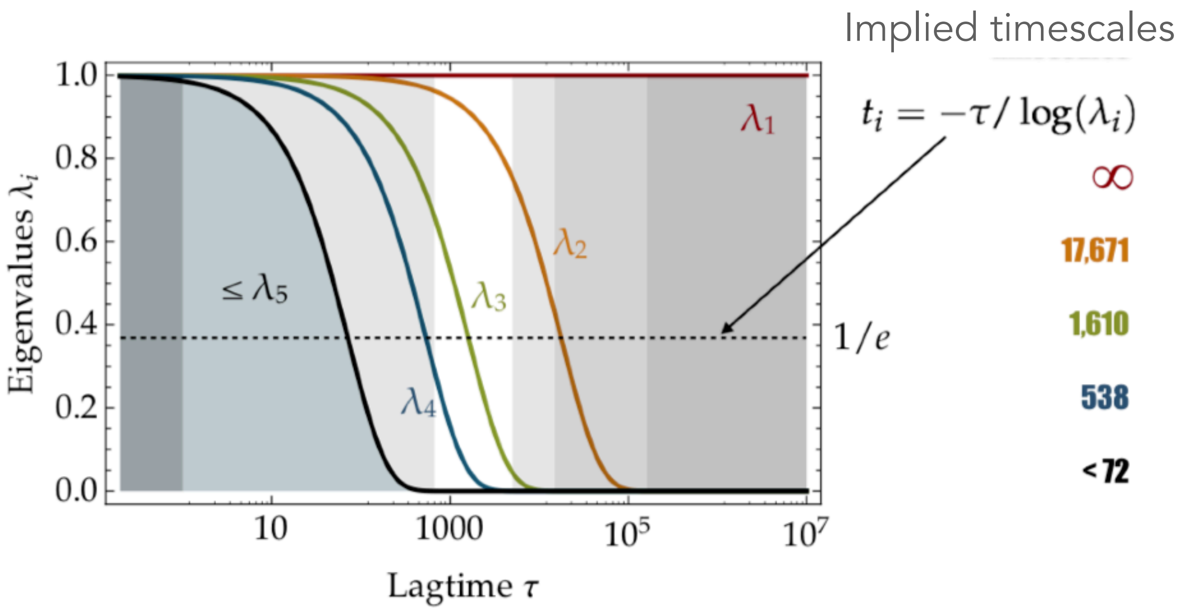
Eigenspectrum of the propagator

Eigenvalues

$$P_\tau \, \phi_i = \lambda_i \, \phi_i$$

Chapman-Kolmogorov implies exponential decay of eigenvalues with lag time

$$\lambda_i(k \cdot \tau) = \lambda_i^k(\tau)$$



Implied timescales

$$t_i = -\tau / \log(\lambda_i)$$

∞

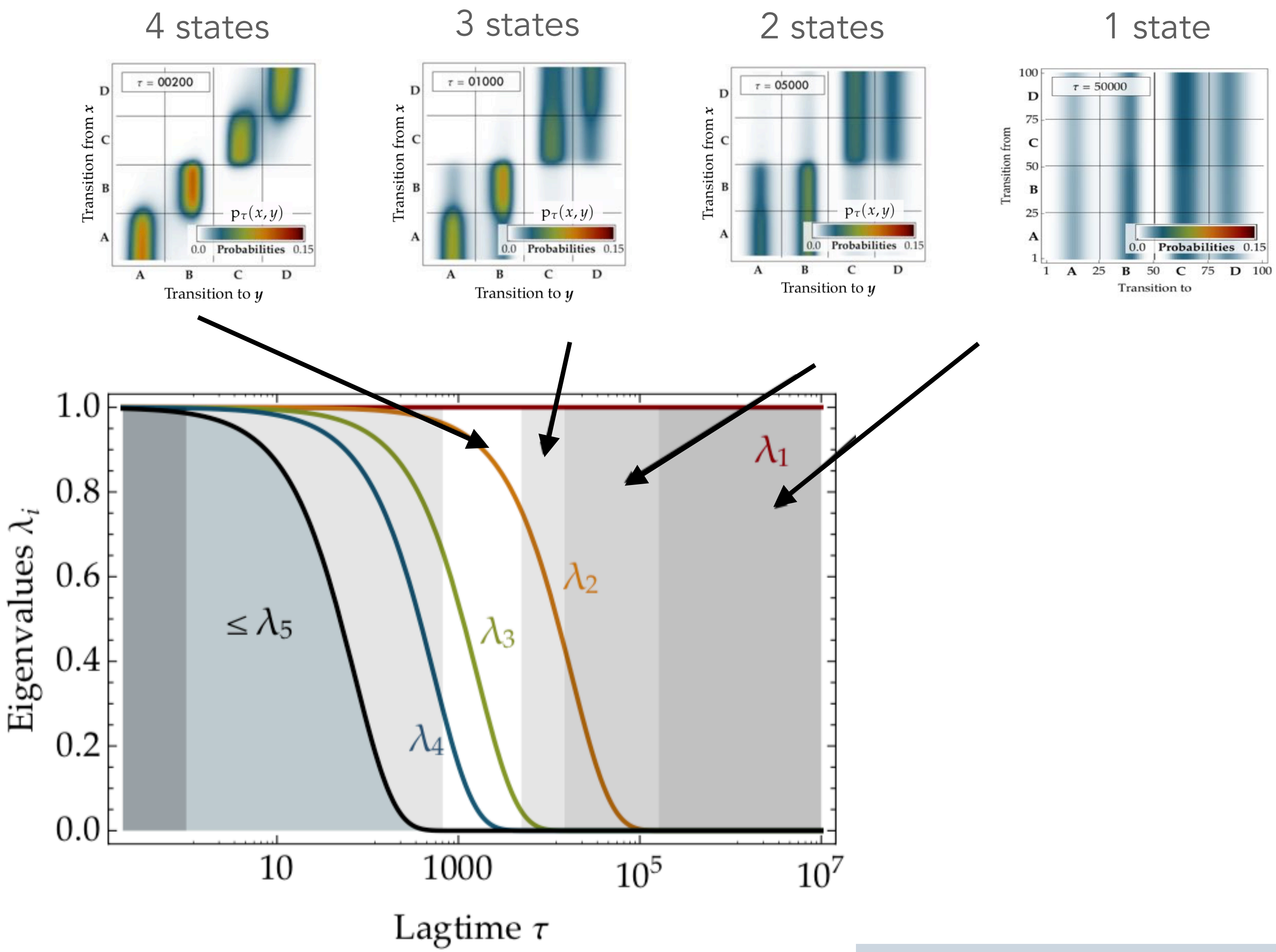
17,671

1,610

538

< 72

Eigenspectrum of the propagator


4 states

3 states

2 states

1 state

Assumptions III

The propagator can be approximated using only a finite number m < M of processes with non-zero eigenvalues, i.e. $\forall \tau > \tau_{\min},\ i > m : \lambda_i(\tau) \approx 0$

such that the dynamics can be written as

$$u_{t+k\tau}(\mathbf{x}) = \mathcal{T}_{\text{slow}}(k\tau) \circ u_t(\mathbf{x}) + \cancel{\mathcal{T}_{\text{fast}}(k\tau) \circ u_t(\mathbf{x})},$$

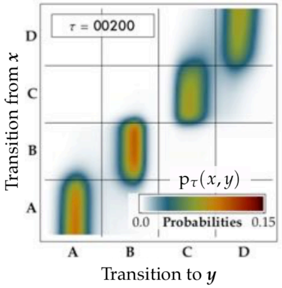
$$= \sum_{i=1}^{m} \lambda_i^k \, \langle u_t, \phi_i \rangle \, \psi_i(\mathbf{x}) + \cancel{\mathcal{T}_{\text{fast}}(k\tau) \circ u_t(\mathbf{x})},$$

If the eigenvalue spectrum has a gap, a lag time τ can be chosen to fulfill this assumption.

„The fast processes have decayed“

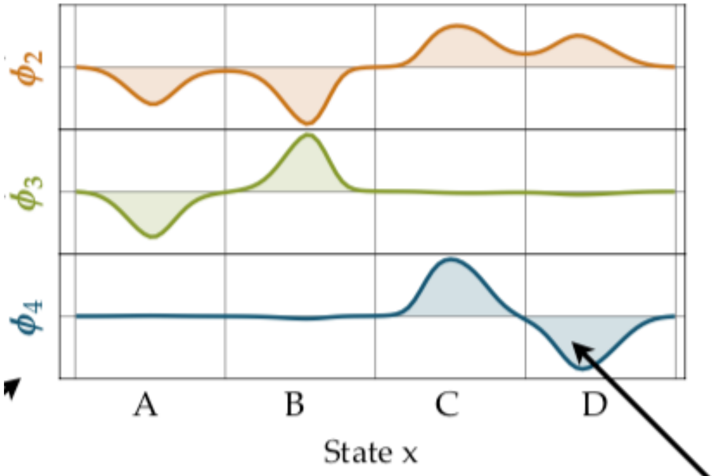
Eigenspectrum of the propagator

Separation into eigenvector/eigenvalue pairs $\mathrm{P}_\tau\,\phi_i = \lambda_i\,\phi_i$



→

Eigenfunctions



timescales

∞

17,671

1,610

538

Sign structure indicates
Metastable states

Markov State Models

From continuous state space to a finite set of states

- everything we learnt for continuous models is also true for discrete MSMs

How to construct a simple MSM from data in a full continuous state space?

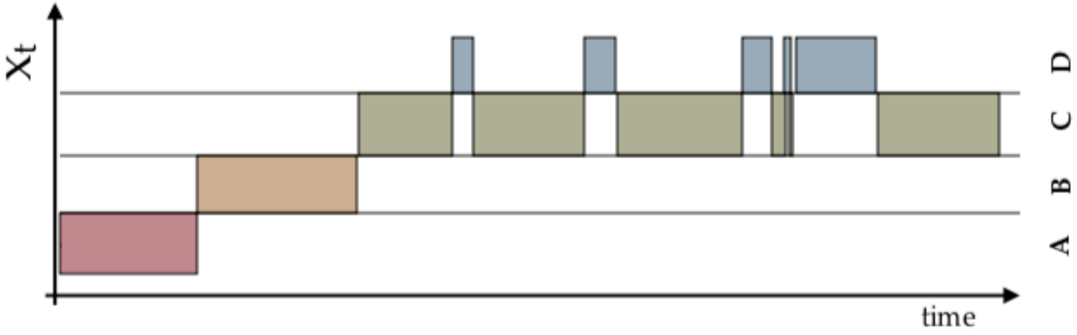
Prinz, JH., Wu, H., Sarich, M., Keller, BG., Senne, M., Held, M., Chodera, JD., Schütte, C. and Noé, F. Markov models of molecular kinetics: generation and validation. J. Chem. Phys. 134, 174105 (2011).

Estimation

How to construct an MSM from simulation data?

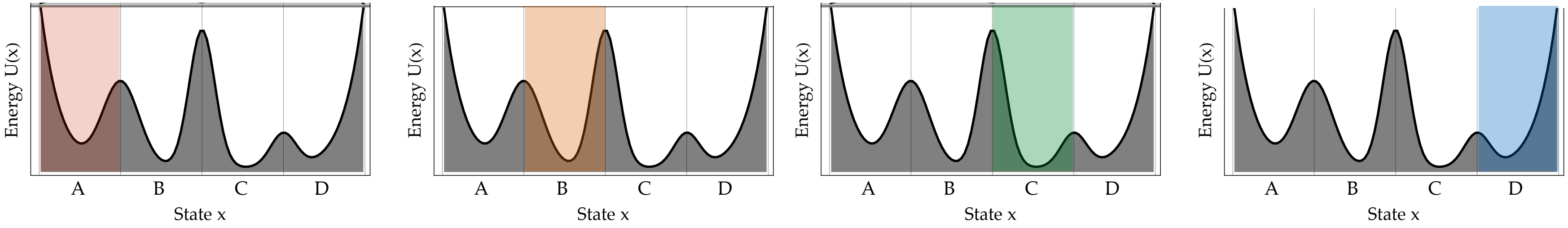
Discretization

Example of realization of a Markov process



Counting transitions

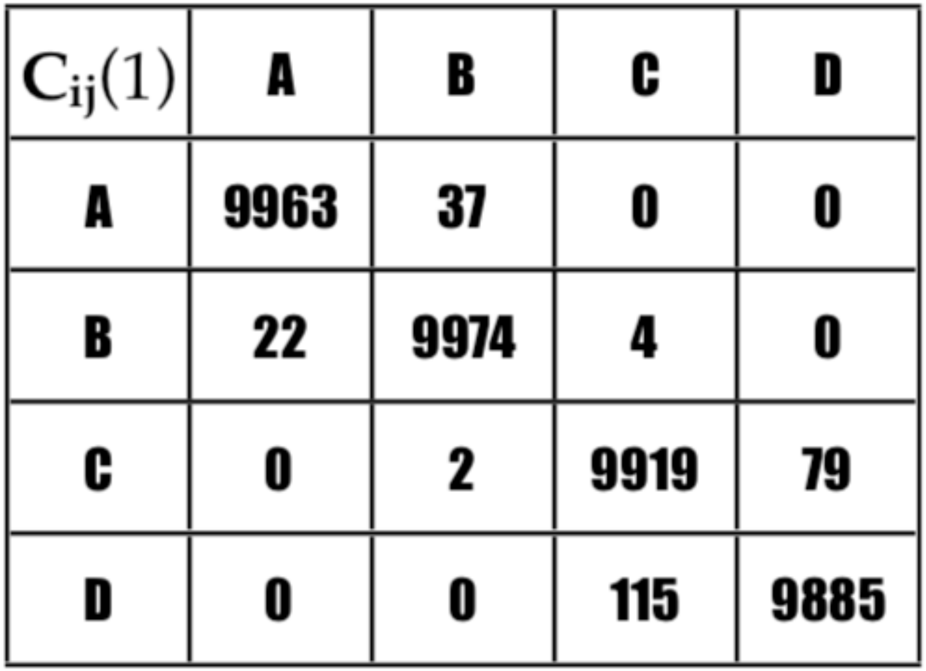
	A	B	C	D
A	10	1	0	0
B	0	13	1	0
C	0	0	24	5
D	0	0	5	14



Count matrix

Generate a Markov model from discretized time series by counting transitions. In this example:

Count matrix:



$C_{ij}(1)$	A	B	C	D
A	9963	37	0	0
B	22	9974	4	0
C	0	2	9919	79
D	0	0	115	9885

Likelihood

Given the transition probabilities of an MSM, we can compute the observation probability for a full (discrete) trajectory:

$$\begin{aligned}\mathbb{P}(x_1, \ldots, x_t \mid P) &= \prod_{k=1}^{L} p_{x_{k-1}, x_k} \\ &= p_{x_0, x_1} \cdot \ldots \cdot p_{x_L - 1, L} \\ &= \prod_{ij} p_{ij}^{c_{ij}} \\ &= p_{11}^{c_{11}} \cdot \ldots\end{aligned}$$

Naive approach: Find the MSM that has the highest likelihood given the observed data

-> Maximum Likelihood Estimator (MLE)

Analytic solution

Given the constraints of the MSM transition matrix

$$\sum_j p_{ij} = 1, \quad \forall i$$

Find an analytic expression for the MLE

$$P^{\mathrm{MLE}} = \underset{P}{\operatorname{argmax}} \prod_{k=1}^{L} p_{x_{k-1},x_k}$$

Using Lagrange multipliers

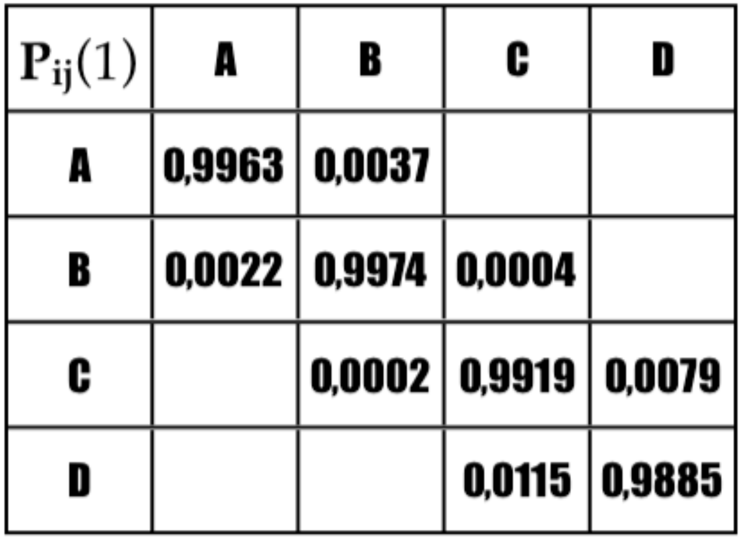
$$\begin{aligned}\hat{p}_{ij} &= \frac{\sum_{n=\tau}^{L} \delta(x_{n-\tau} = i, x_n = j)}{\sum_{n=\tau}^{L} \delta(x_{n-\tau} = i)} \\ &= \frac{\hat{C}_{ij}}{\sum_j \hat{C}_{ij}}\end{aligned}$$

row-normalized transition counts

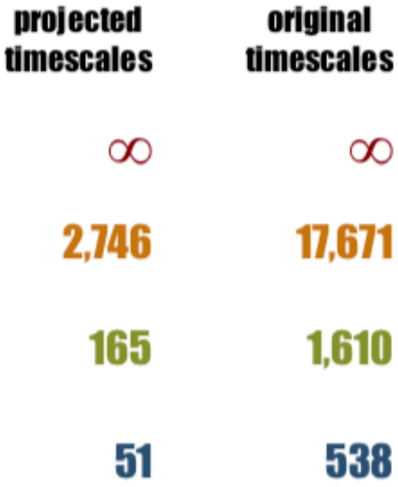
MLE transition matrix

Compute transition matrix from the count matrix to parametrize the simple 4 state MSM.

Transition matrix:



$P_{ij}(1)$	A	B	C	D
A	0,9963	0,0037		
B	0,0022	0,9974	0,0004	
C		0,0002	0,9919	0,0079
D			0,0115	0,9885



projected timescales	original timescales
∞	∞
2,746	17,671
165	1,610
51	538

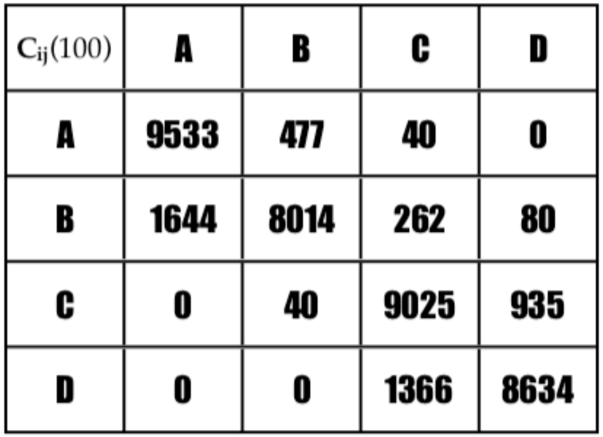
The timescales of projected models are always underestimated!

Djurdjevac, N., Sarich, M. & Schütte, C. *Estimating the eigenvalue error of Markov State Models*. Multiscale Model. Sim.

Lagtime dependence

Increasing the lagtime (use every n-th step) when counting will improve the estimation of the timescales

Count matrix at lagtime 100:



$C_{ij}(100)$	A	B	C	D
A	9533	477	40	0
B	1644	8014	262	80
C	0	40	9025	935
D	0	0	1366	8634

projected timescales	original timescales
∞	∞
15,397	17,671
1211	1,610
379	538

We have to choose the lagtime such that the MSM implied timescales are converged.

Assumptions II

Detailed balance („microscopic reversibility")

$$\pi(x)\mathrm{p}_\tau(x,y) = \pi(y)\mathrm{p}_\tau(y,x)$$

- allows to define a meaningful scalar product

$$\langle f,g\rangle_\pi = \int \mathrm{d}x\, f(x)g(x)\pi(x)$$

- Propagator is symmetric w.r.t. stationary distribution-weighted scalar product

In equilibrium, there is no net flux of particles,
i.e. we cannot draw energy from the system

Reversible dynamics

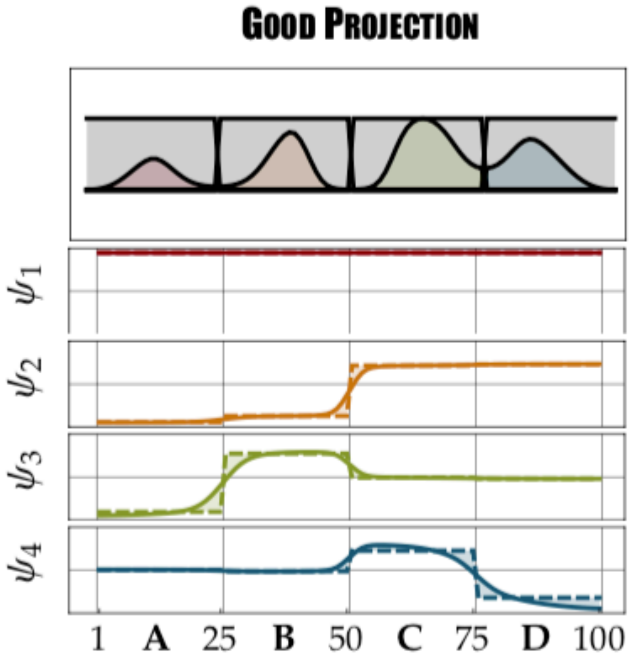
MLE estimate does not necessarily obey detailed balance. We can add a detailed balance constraint

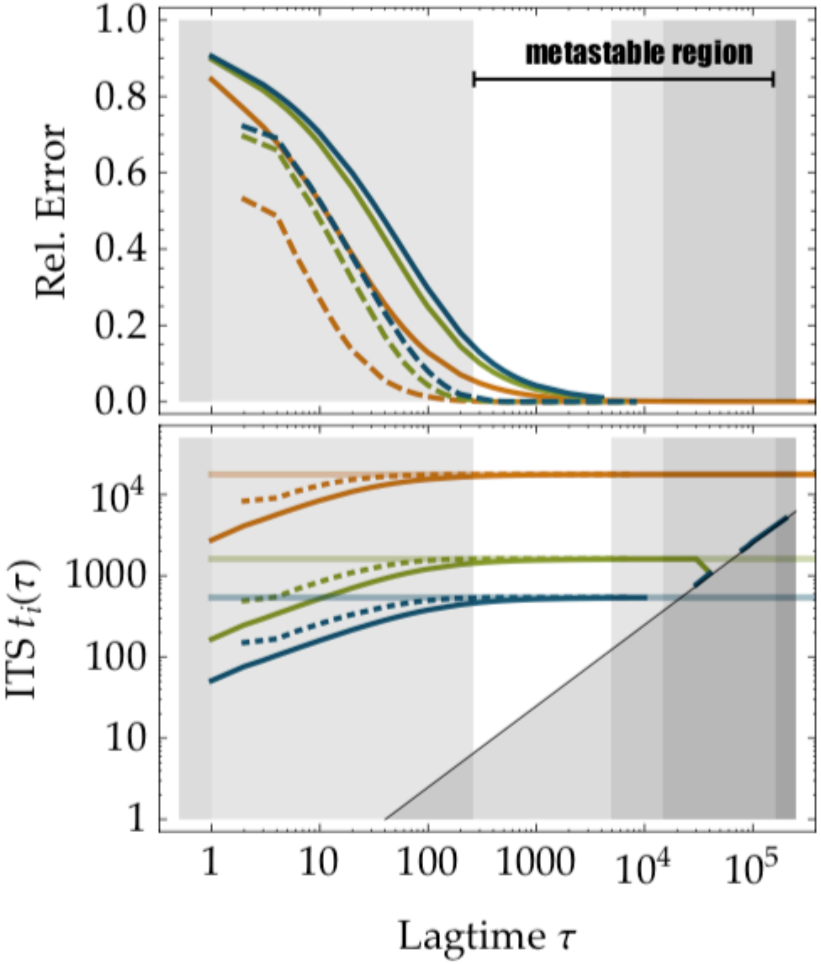
$$\pi_i p_{ij} = \pi_j p_{ji}, \quad \forall i, j$$

There is no analytic solution for this estimator but it can be solved iteratively. The final solution is a model that obeys detailed balance and maximizes the likelihood under that constraint.

Example dynamics

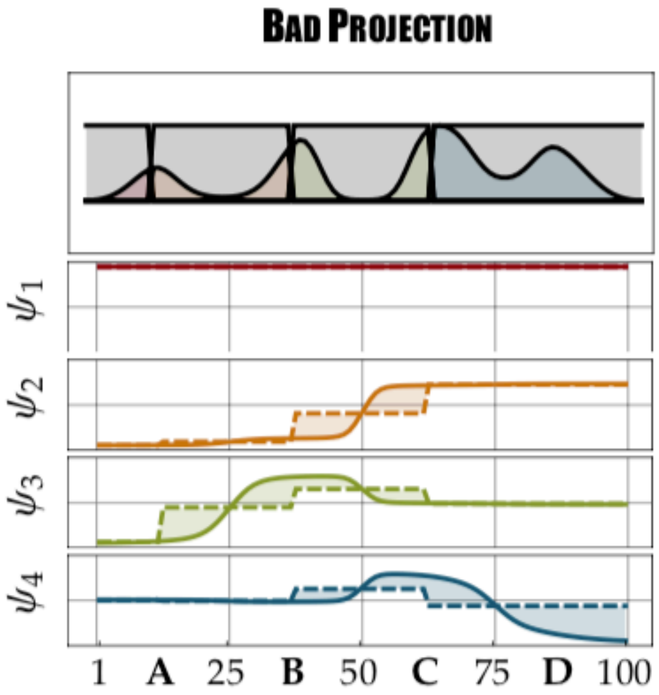
$$t_i = -\tau / \log(\lambda_i)$$

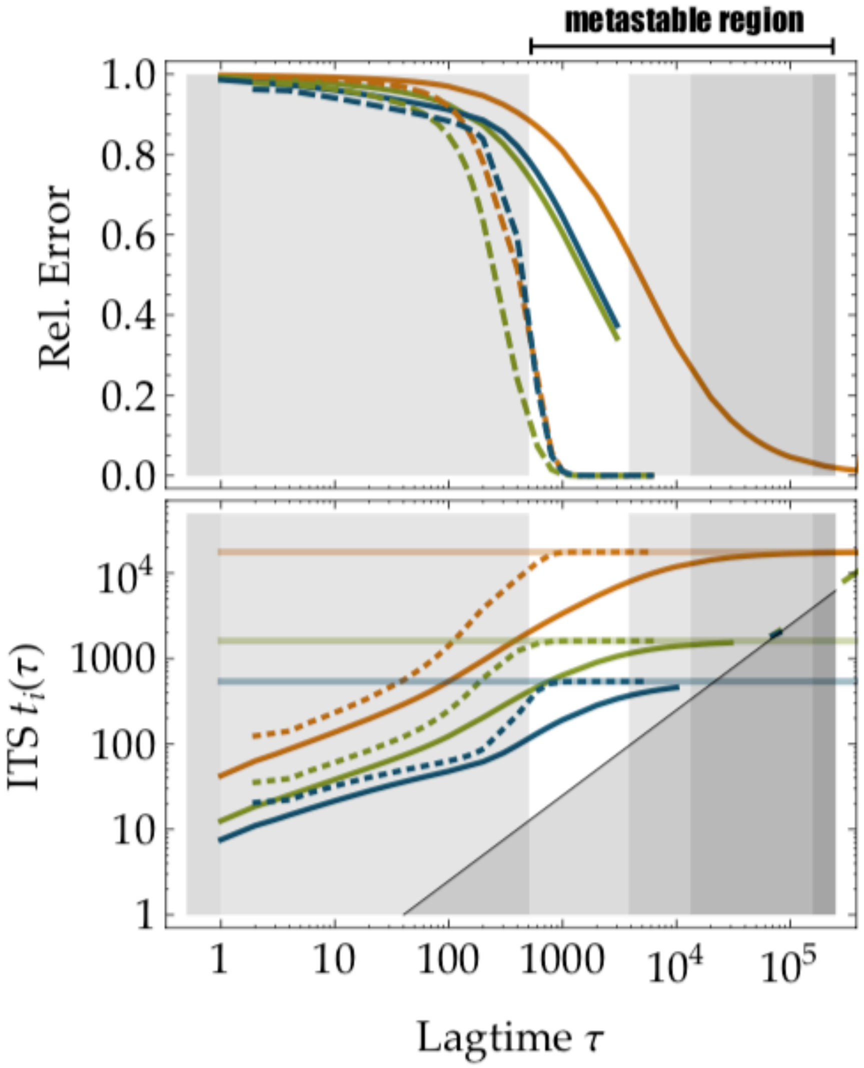




Example dynamics

$$t_i = -\tau / \log(\lambda_i)$$





Problems

Observations in the projected (discretized) space are often non-Markovian

- MSM not the most appropriate choice to express the dynamics of a non-Markovian time series

But

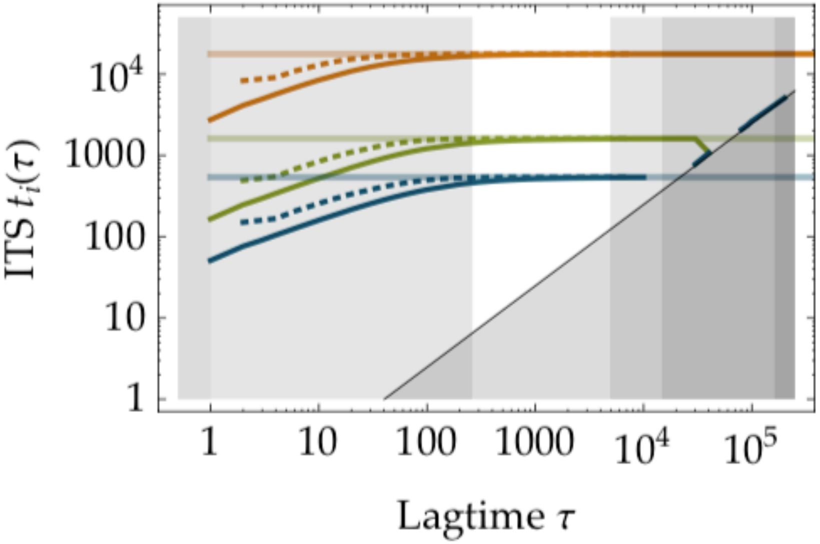
- We don’t want to compress the dynamics into a transition matrix, we want to model a system

So

- Ensure that the observations are „as Markovian as possible“

Validation measures

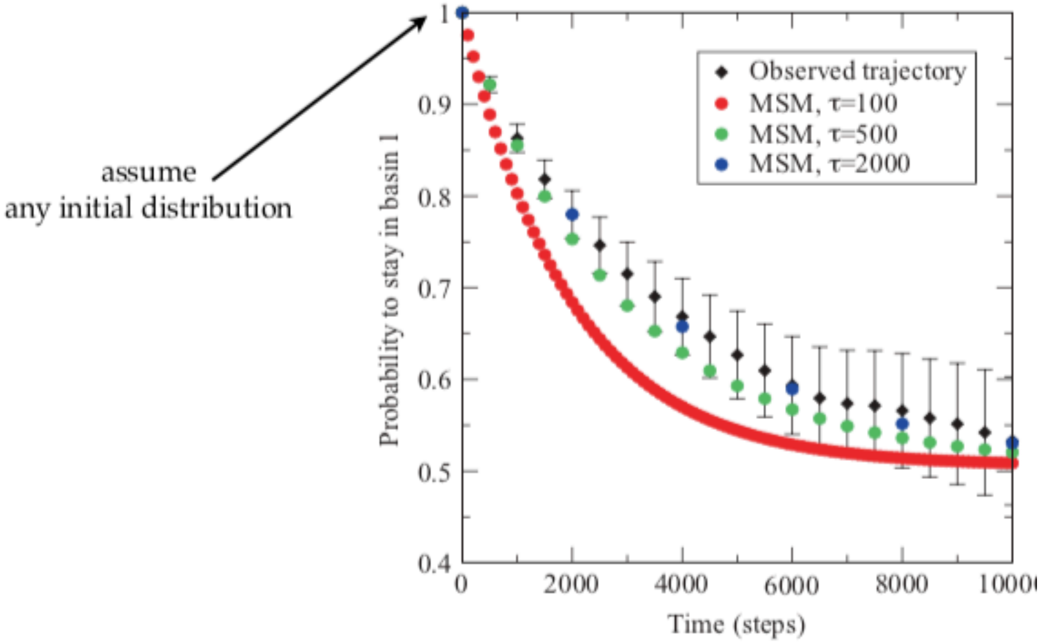
Implied timescales test



Validation measures

Chapman-Kolmogorov equation $P(\tau)^k = P(k \cdot \tau)$

Compare the evolution in the model with the data



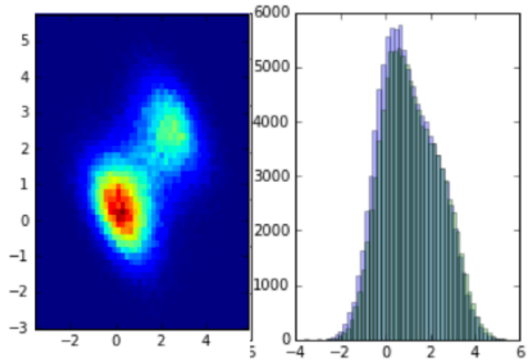
assume any initial distribution

Scheme for generation

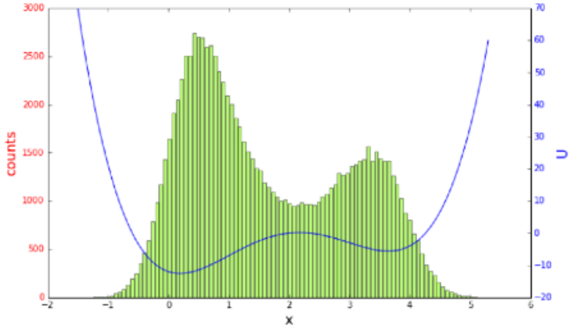
- use a fine enough **discretization** and construct a large transition matrix
- check **implied timescales convergence** and select a lag time
- use dominant eigenvectors to estimate the **metastable subsets**
- use metastable sets as discretization and construct a small **metastable transition matrix**
- **validate** the model using Chapman-Kolmogorov test

Example workflow

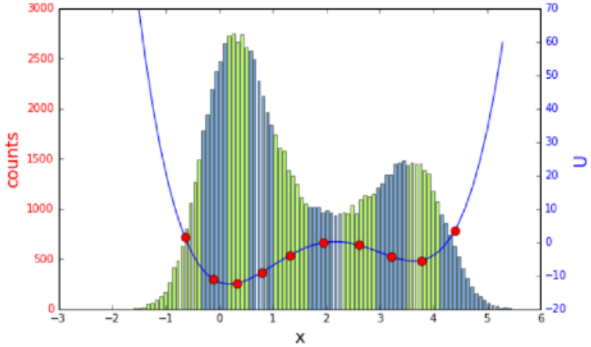
Toy example: 2D asymmetric double well potential



Coordinate transform
e.g. PCA, TICA



clustering



Discrete trajectory

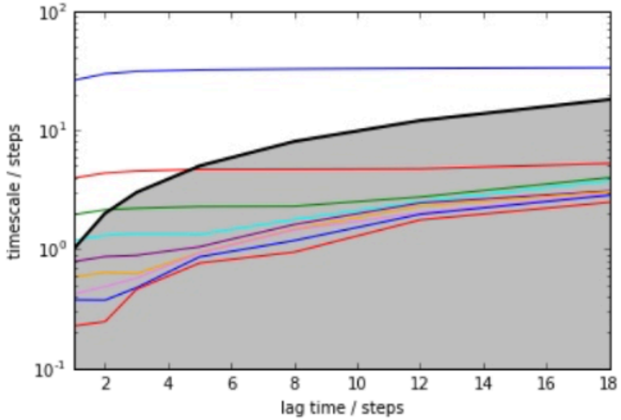
estimation

Transition matrix

Example workflow

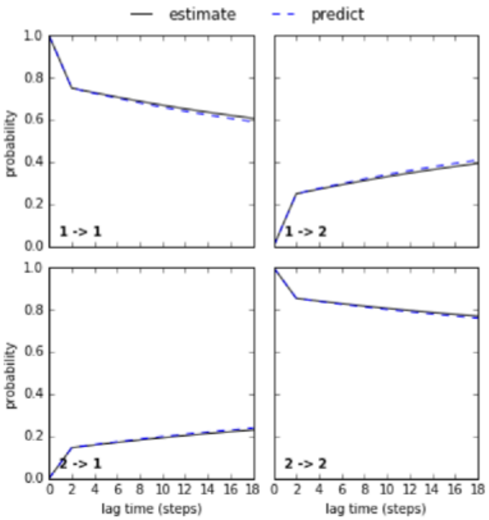
validation:

implied time scales convergence



$$\tau_k = -\frac{\tau}{\ln(\lambda_k)}$$

Chapman-Kolmogorow test

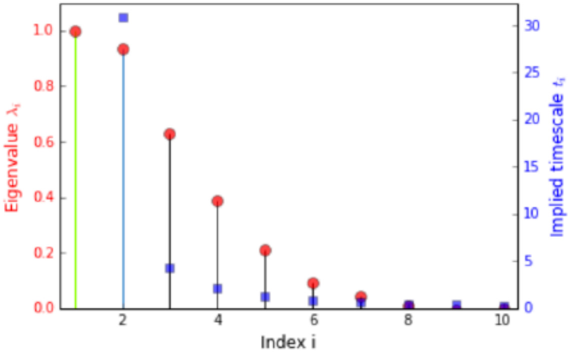


$$P(k \cdot \tau) = P(\tau)^k$$

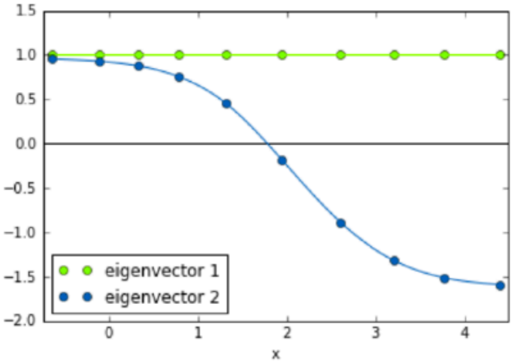
Example workflow

Properties

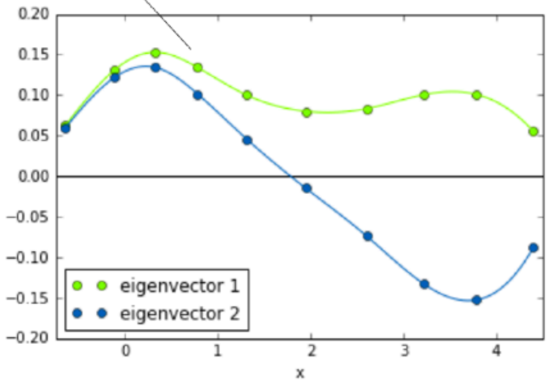
Eigenvalues



Right eigenvectors



Left eigenvectors



Equilibrium distribution

Relaxation time scales

slow processes

Error estimation

- besides the MLE estimate, other MSMs can lead to the same observation
- Bayes' rule allows to find the probability of a model given the observations
- Likelihood from before (MLE):

$$\mathbb{P}(x_i, \ldots, x_t \mid P) = p(C \mid P) \propto \prod_{i,j=1}^{n} p_{ij}^{c_{ij}}$$

- introduce prior information

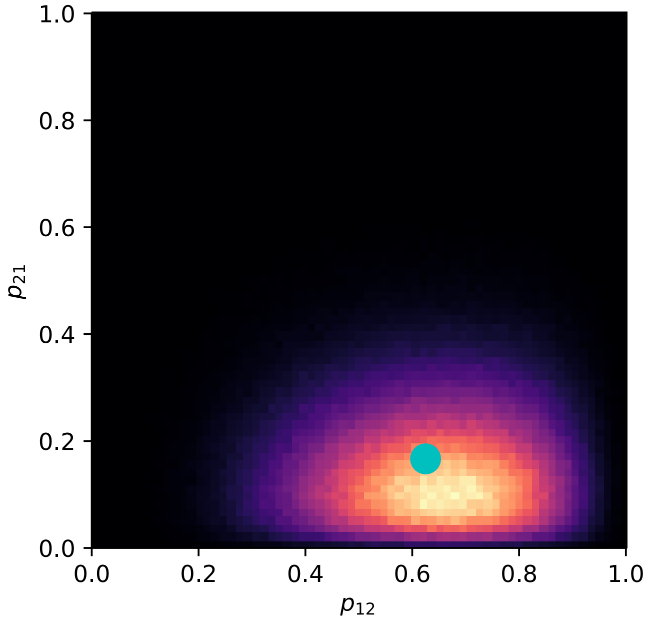
$$p(P \mid C) \propto p(C \mid P) p(P)$$

The prior can encode useful constraints, e.g. reversibility, fixed stationary distributions, sparsity etc.

Bayesian inference of MSMs

- MCMC sampling on transition matrix
- yields a set of transition matrices
- we can estimate model confidence by evaluating properties on all sampled transition matrices

$$\mathbb{E}(f(P)) \approx \frac{1}{N} \sum_{P \sim \mathbb{P}(P|x_1,\ldots x_t)} f(P)$$



(1) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and Uncertainty of Reversible Markov Models. *The Journal of Chemical Physics* **2015**, *143* (17), 174101. https://doi.org/10.1063/1.4934536.

Figure courtesy of S Olsson

Analysis

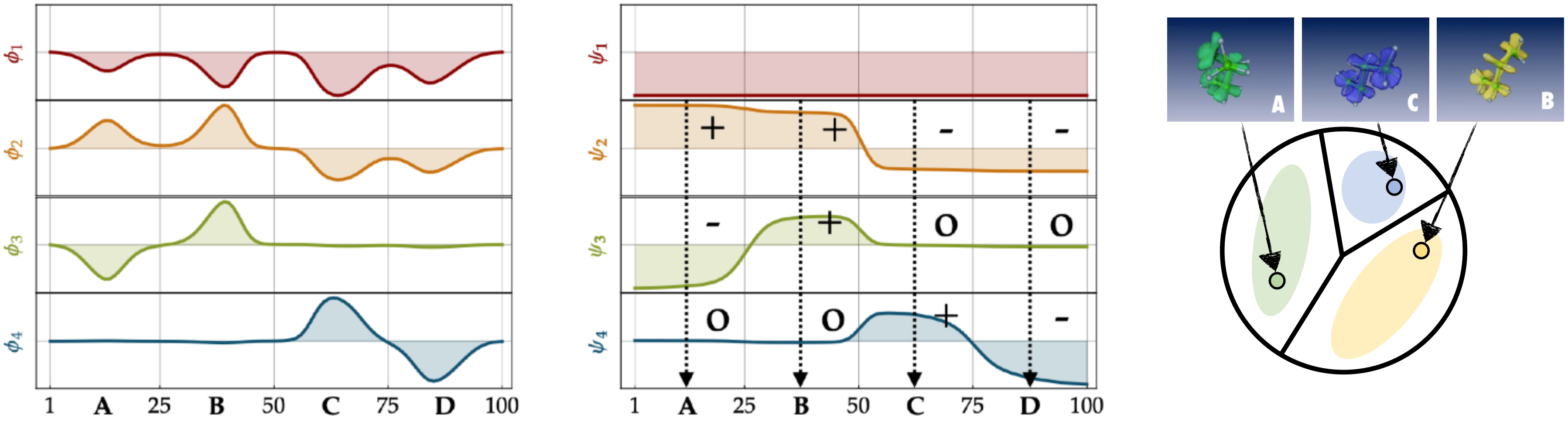
Target properties

We can compute

- equilibrium properties (observable averages)
- relaxation timescales (eigenvalues)
- dominant processes (eigenvectors)
- stationary distribution / equilibrium distribution (first normalized eigenvector)
- metastable sets (Eigenvectors / PCCA)
- correlation functions
- mean first passage times
- transition path probabilities

PCCA++

Idea: find metastable sets from the eigenvectors.

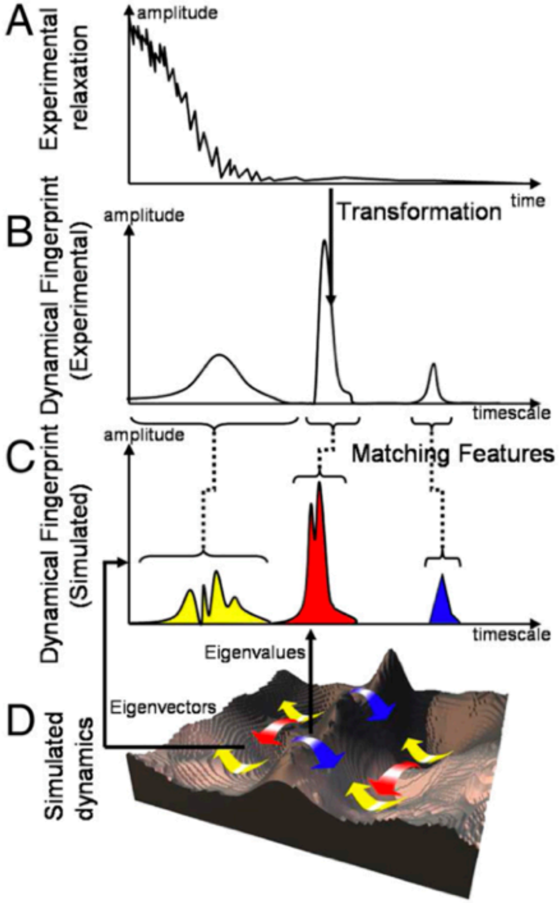


- sign structure of the right eigenvectors are used for a „spectral clustering"

Röblitz, S. & Weber, M. *Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification*. Adv Data Anal Classif 7, 147–179 (2013)

Dynamical fingerprints

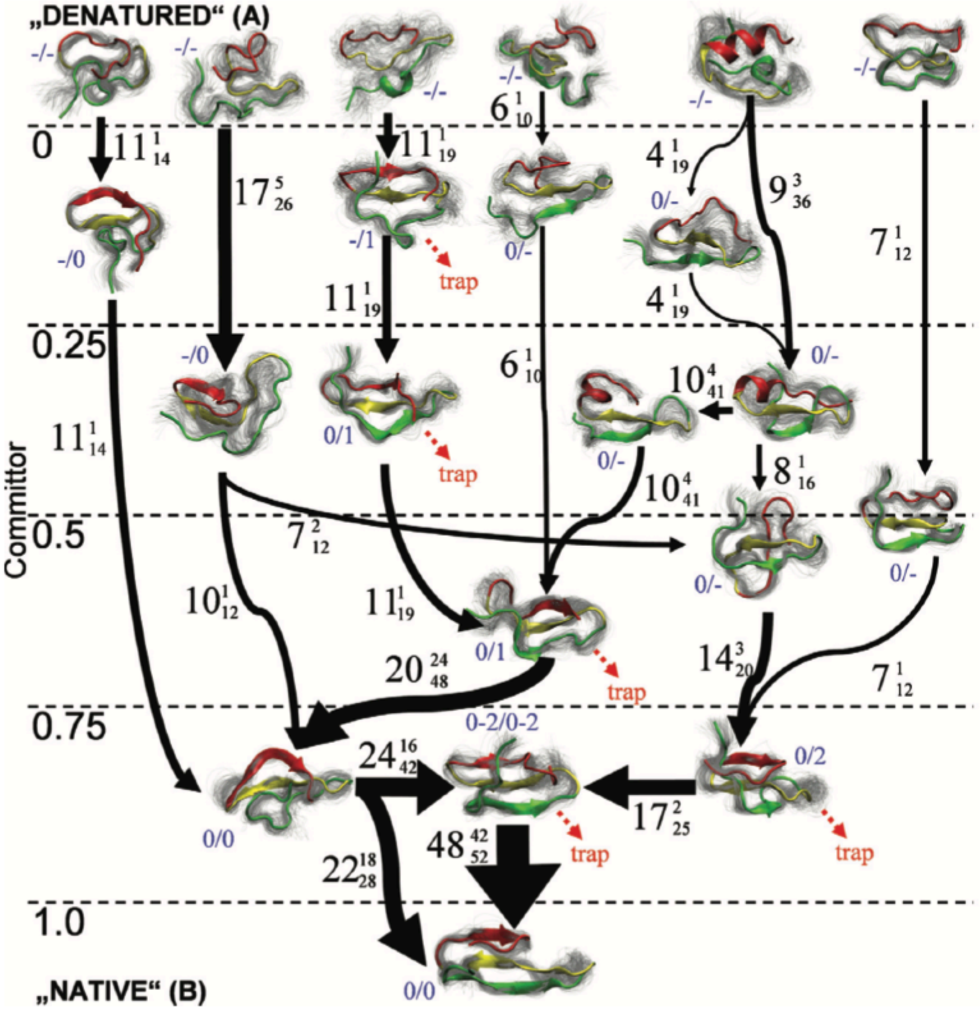
Idea: Relate relaxation experiments to computational ones by computing a dynamical spectrum



Noé, F. et al. *Dynamical fingerprints for probing individual relaxation processes in biomolecular dynamics with simulations and kinetic experiments*. Proc. Nat. Acad. Sci. USA 108, 4822–4827 (2011).

Path properties

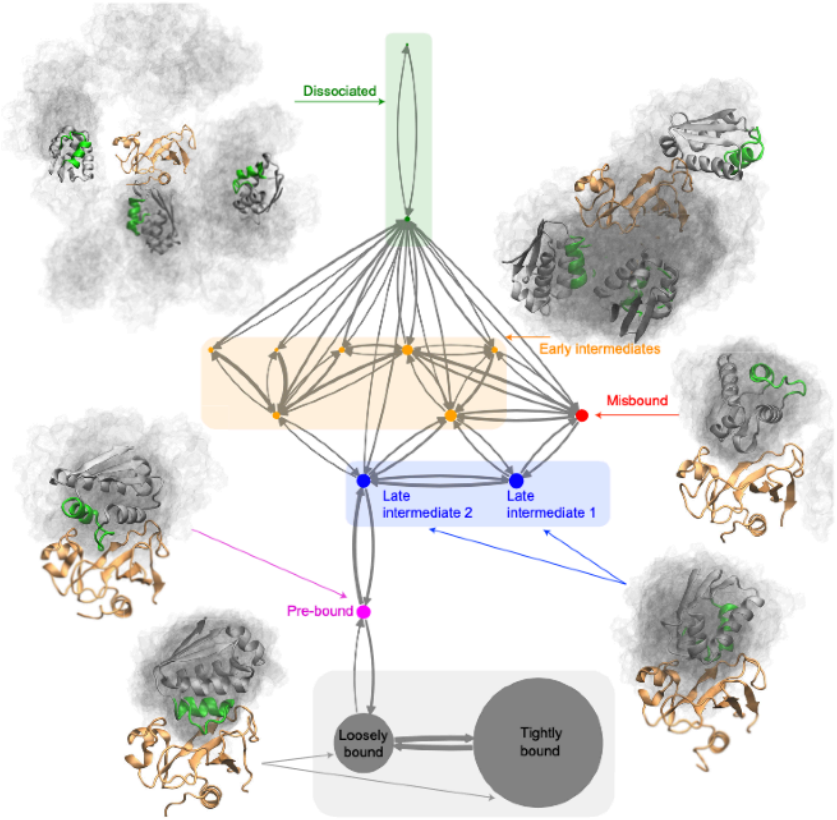
Compute path probabilities, e.g., for folding of a protein



Noé, F., Schütte, C., Vanden-Eijnden, E. & Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. Proc. Nat. Acad. Sci. USA 106, 19011–19016 (2009).

Binding/unbinding kinetics

Model binding kinetics of e.g. protein-protein dissociation, determination of dissociation constant



Plattner, N.; Doerr, S.; Fabritiis, G. D.; Noé, F. Complete Protein–Protein Association Kinetics in Atomic Detail Revealed by Molecular Dynamics Simulations and Markov Modelling. *Nature Chemistry* **2017**, *9* (10), 1005. https://doi.org/10.1038/nchem.2785.

Binding process - 100 microseconds

Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry 9, 1005 (2017)

Thanks for your attention