

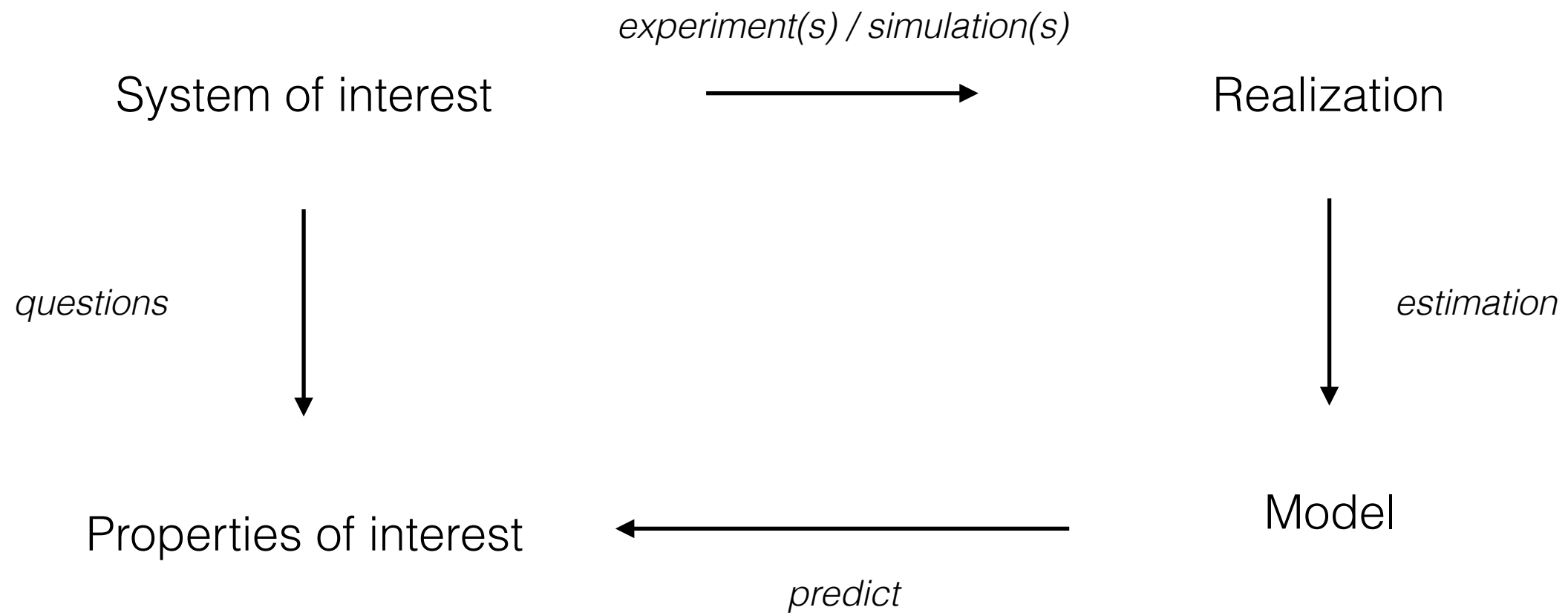
Markov state models

Theory, properties, estimation and validation

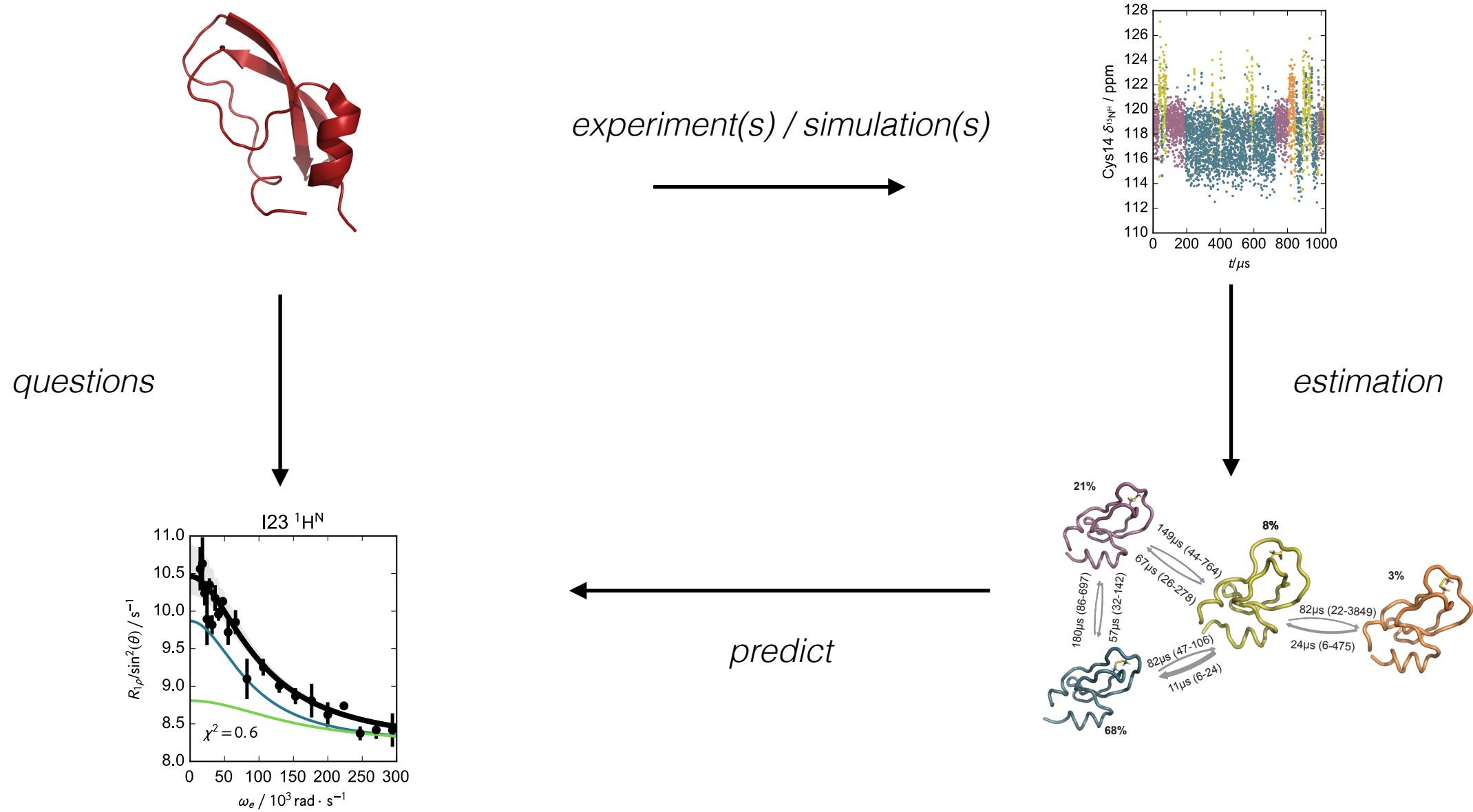
Simon Olsson

2019 PyEMMA Workshop
FU Berlin
Tuesday, Feb 19th

Motivation



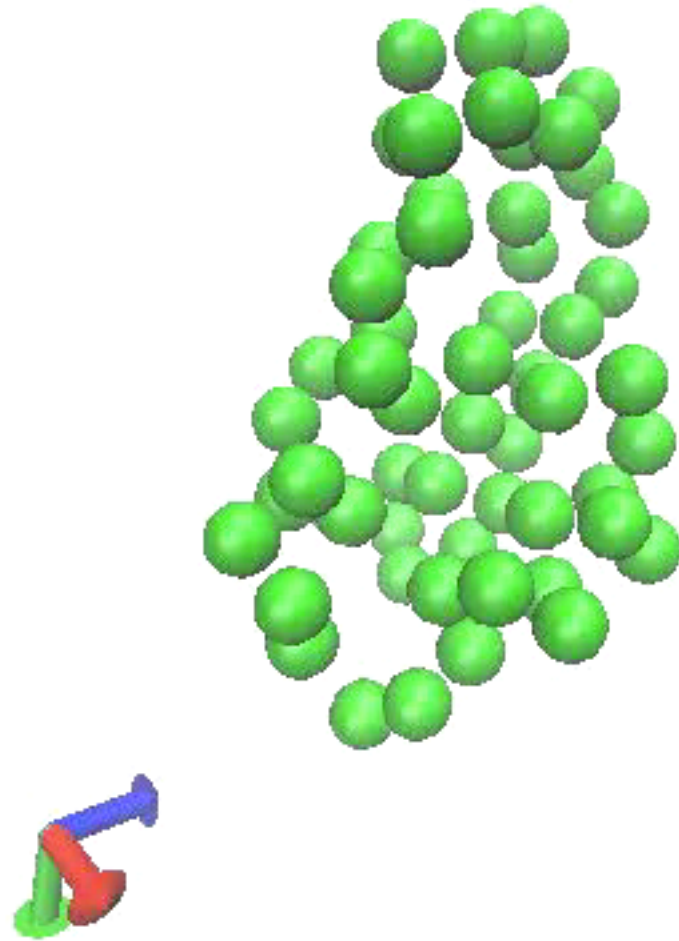
Motivation



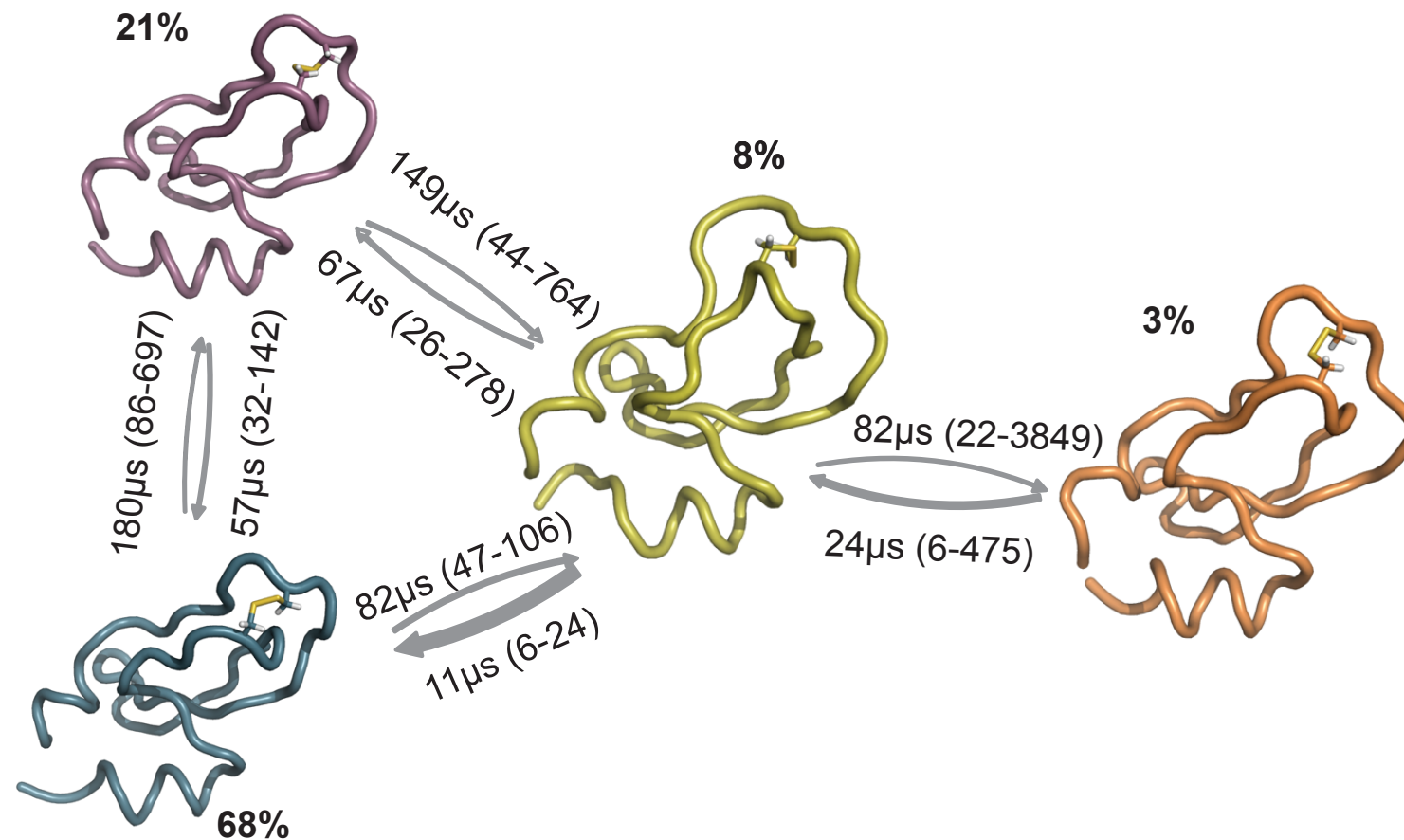
*„Find properties of a system of interest
using a simple model parametrized from observations“*

Example: BPTI

Simulation of BPTI



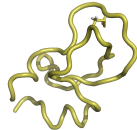


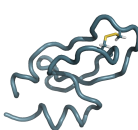
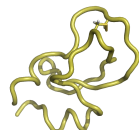


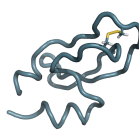
Markov state models



Metastability of states allow us to significantly simplify the dynamics of our system of interest

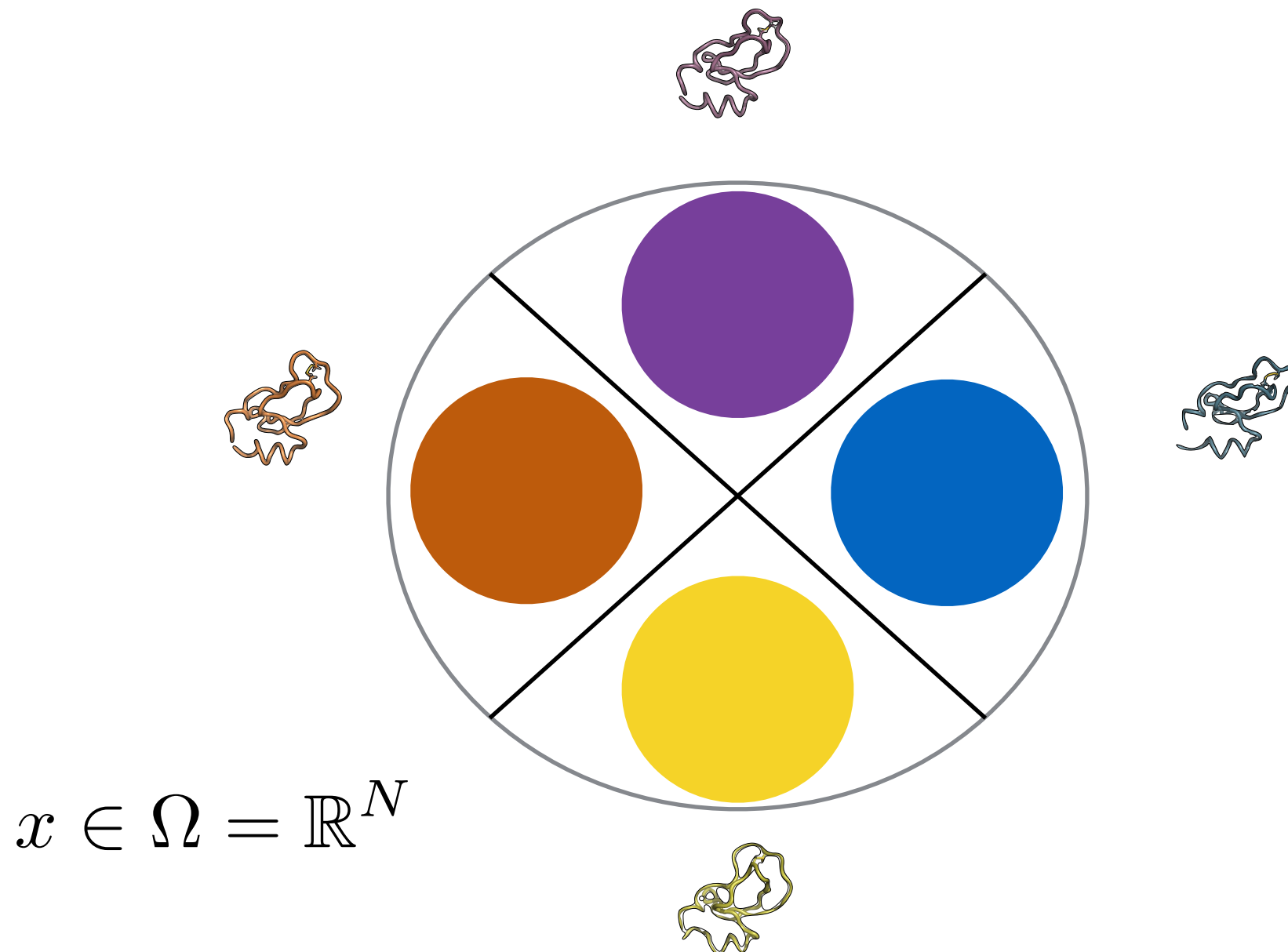
Markov state models

Final state

					
Initial state		96%	1%	2%	1%
		5%	95%	0%	0%
		1%	0%	97%	2%
		1%	0%	2%	97%

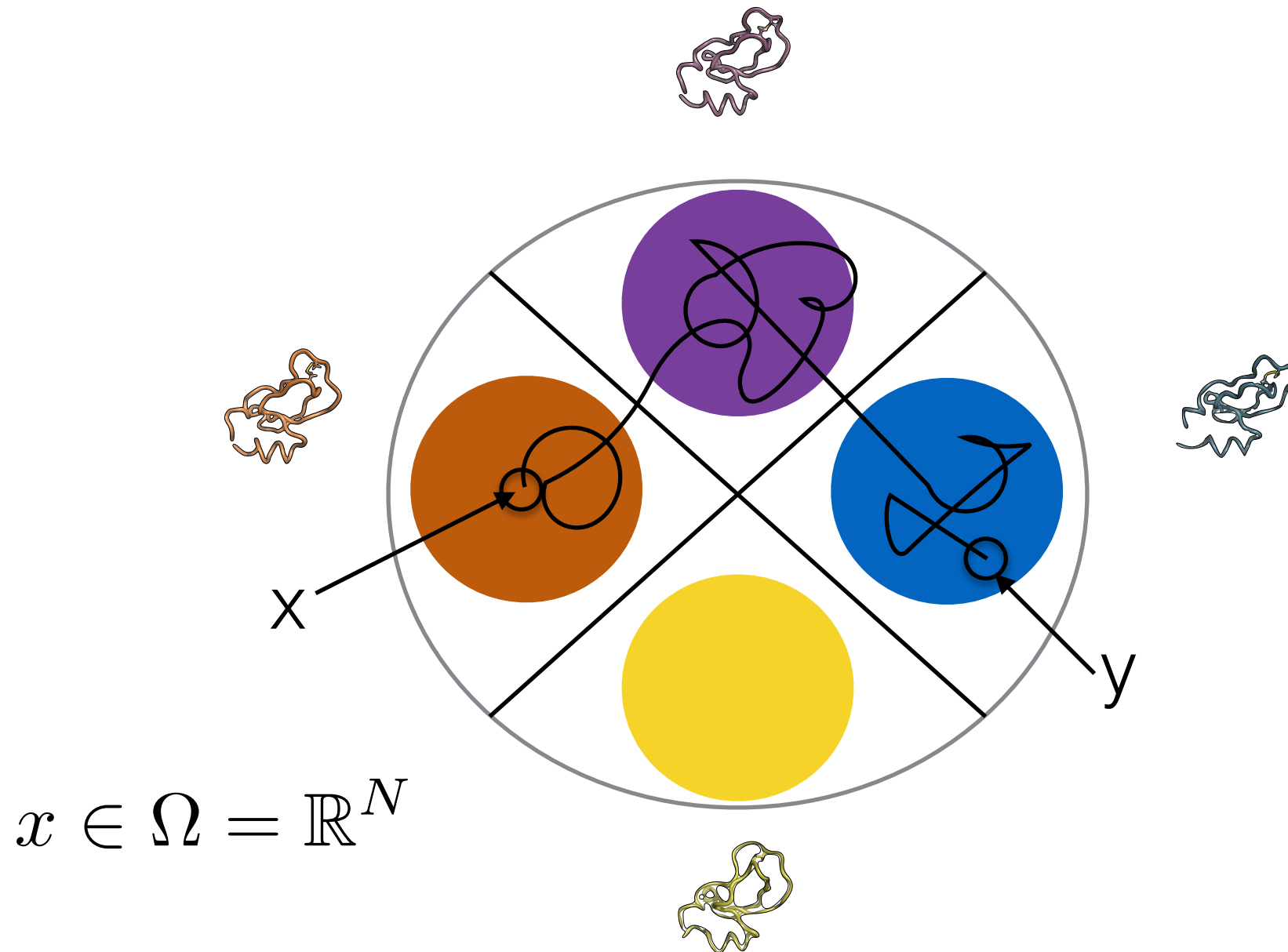
A Markov state model describes the dynamics of a system as conditional transition probabilities

What is meta-stability?



sets of configurations which are long-lived.
Markov state models assume these states, and exchange between them is important.

What is meta-stability?



$$x \in \Omega = \mathbb{R}^N$$

sets of configurations which are long-lived.
Markov state models assume these states, and exchange between them is important.

Molecular simulations

- Molecular simulations are realizations of stochastic process on Ω and are Markovian w.r.t. this space.

$$p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{y} = \mathbb{P}[\mathbf{x}(t + \tau) \in \mathbf{y} + d\mathbf{y} \mid \mathbf{x}(t) = \mathbf{x}]$$

$$\mathbf{x}, \mathbf{y} \in \Omega, \tau \in \mathbb{R}_{0+},$$

Transition probabilities are well defined

Molecular simulations

- Molecular simulations are realizations of stochastic process on Ω and are Markovian w.r.t. this space.

$$p(\mathbf{x}, \mathbf{y}; \tau) d\mathbf{y} = \mathbb{P}[\mathbf{x}(t + \tau) \in \mathbf{y} + d\mathbf{y} \mid \mathbf{x}(t) = \mathbf{x}]$$
$$\mathbf{x}, \mathbf{y} \in \Omega, \tau \in \mathbb{R}_{0+},$$

Transition probabilities are well defined

$$p(\mathbf{x}, A; \tau) = \mathbb{P}[\mathbf{x}(t + \tau) \in A \mid \mathbf{x}(t) = \mathbf{x}]$$
$$= \int_{\mathbf{y} \in A} d\mathbf{y} p(\mathbf{x}, \mathbf{y}; \tau).$$

Also applies for regions

Molecular simulations (2)

Ergodicity

No two or more segments of the space Ω are dynamically disconnected from each other.

and

For an infinitely long simulation we will have visited every state $\mathbf{x} \in \Omega$ infinitely many times.

Molecular simulations (3)

Reversibility

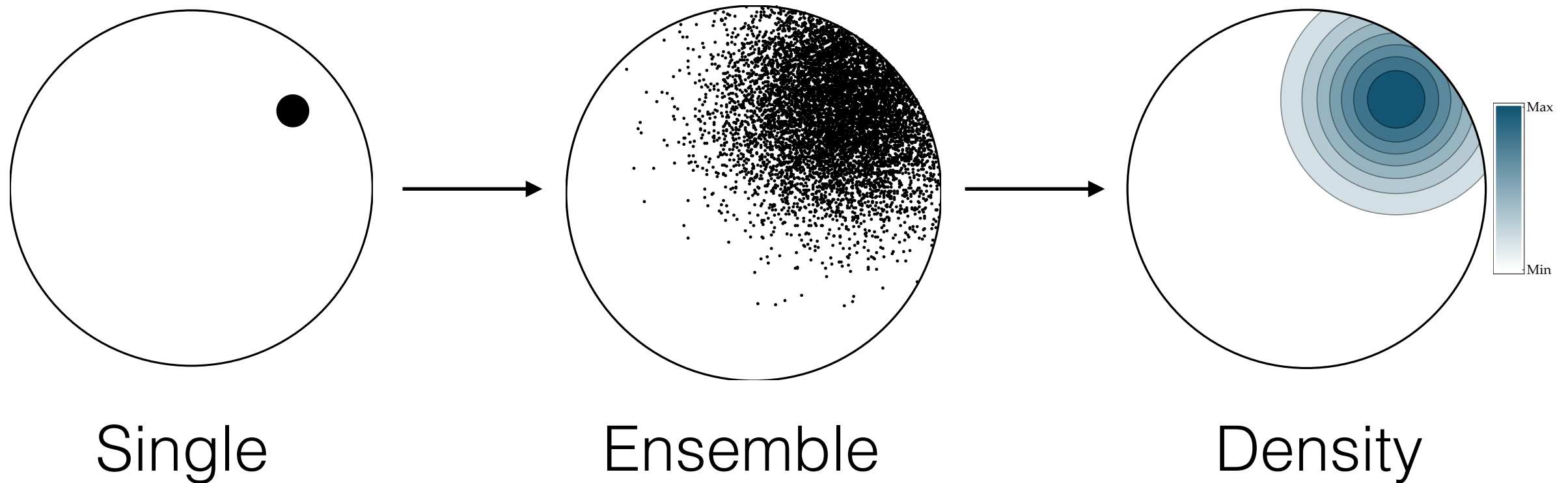
Simulations fulfill the detailed-balance condition:

$$\mu(\mathbf{x}) p(\mathbf{x}, \mathbf{y}; \tau) = \mu(\mathbf{y}) p(\mathbf{y}, \mathbf{x}; \tau)$$

$$\mu(\mathbf{x}) = Z(\beta)^{-1} \exp(-\beta H(\mathbf{x}))$$

At equilibrium the probability of jumping from any x to any y is the same as jumping from y to x .

An illustration of the transition density



Instead of single realizations we now focus on the evolution of an ensemble

Assumptions about the full dynamics

Markovian









$$\mathbb{P}(x_{t+\tau} \in A \mid x_{t_1}, \dots, x_t = x) = \mathbb{P}(x_{t+\tau} \in A \mid x_t = x)$$

*Factorization of the dynamics
into conditional probabilities*

Chapman-Kolmogorov property

$$p_{\tau_1+\tau_2}(x, A) = \int_{\Omega} p_{\tau_1}(x, y) p_{\tau_2}(y, A) dy$$

Direct combination of conditional probabilities with different lag-times

		Final state			
					
Initial state		96%	1%	2%	1%
		5%	95%	0%	0%
		1%	0%	97%	2%
		1%	0%	2%	97%

Assumptions about the full dynamics

Irreducibility

All states of the state space can be reached from any other state in a finite time.

Ensures unique stationary distribution.

Ergodicity

No states are disconnected

No cyclic dynamics.

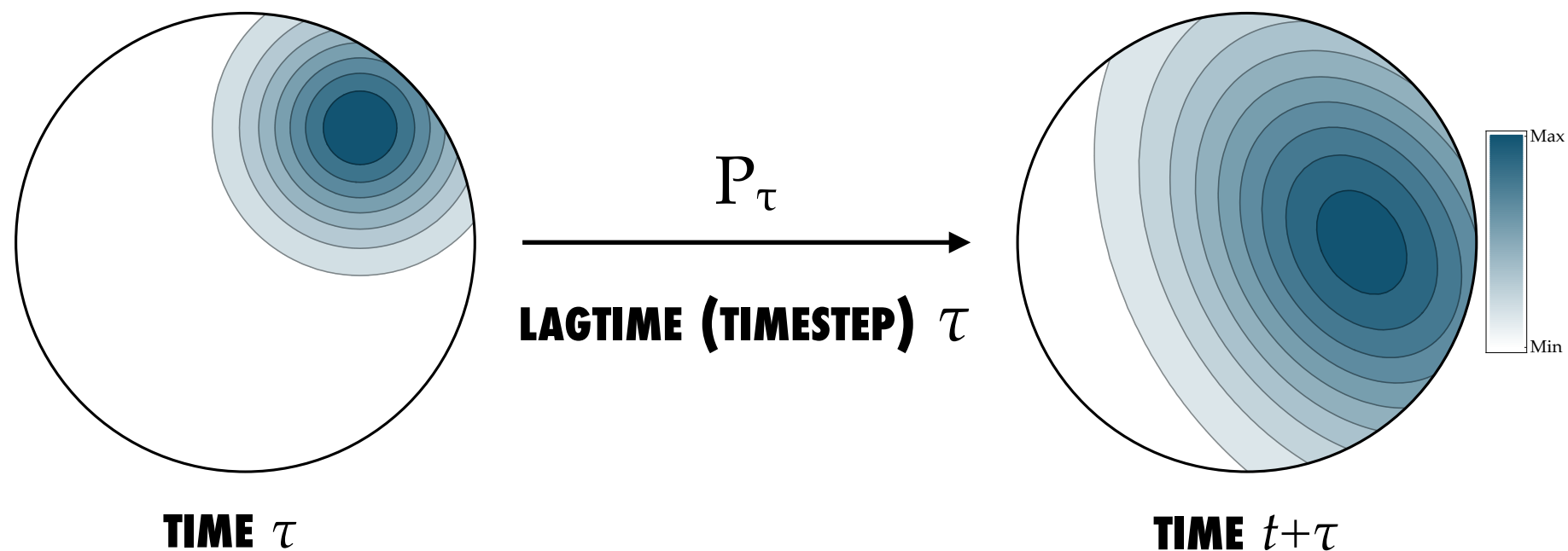
Ensures time and ensemble average properties are equal.

Reversibility

No net-probability flux at equilibrium. \Rightarrow no energy production/absorption \Rightarrow mass conservation.

Not strictly necessary for Markov models

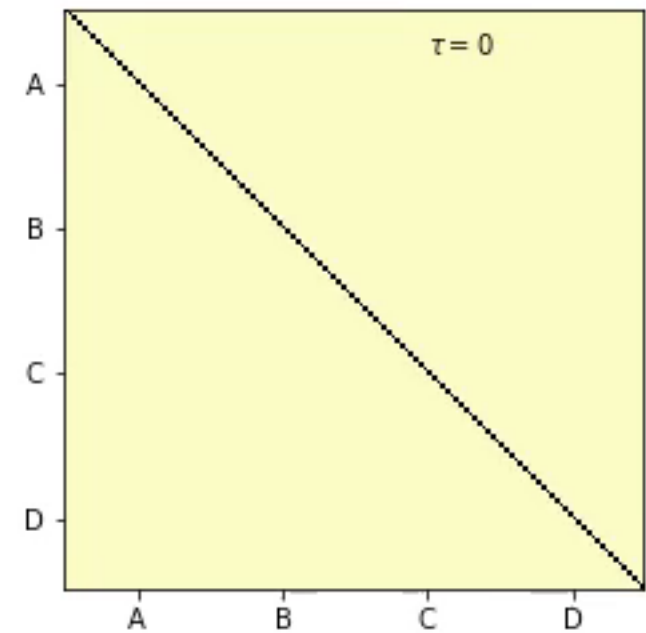
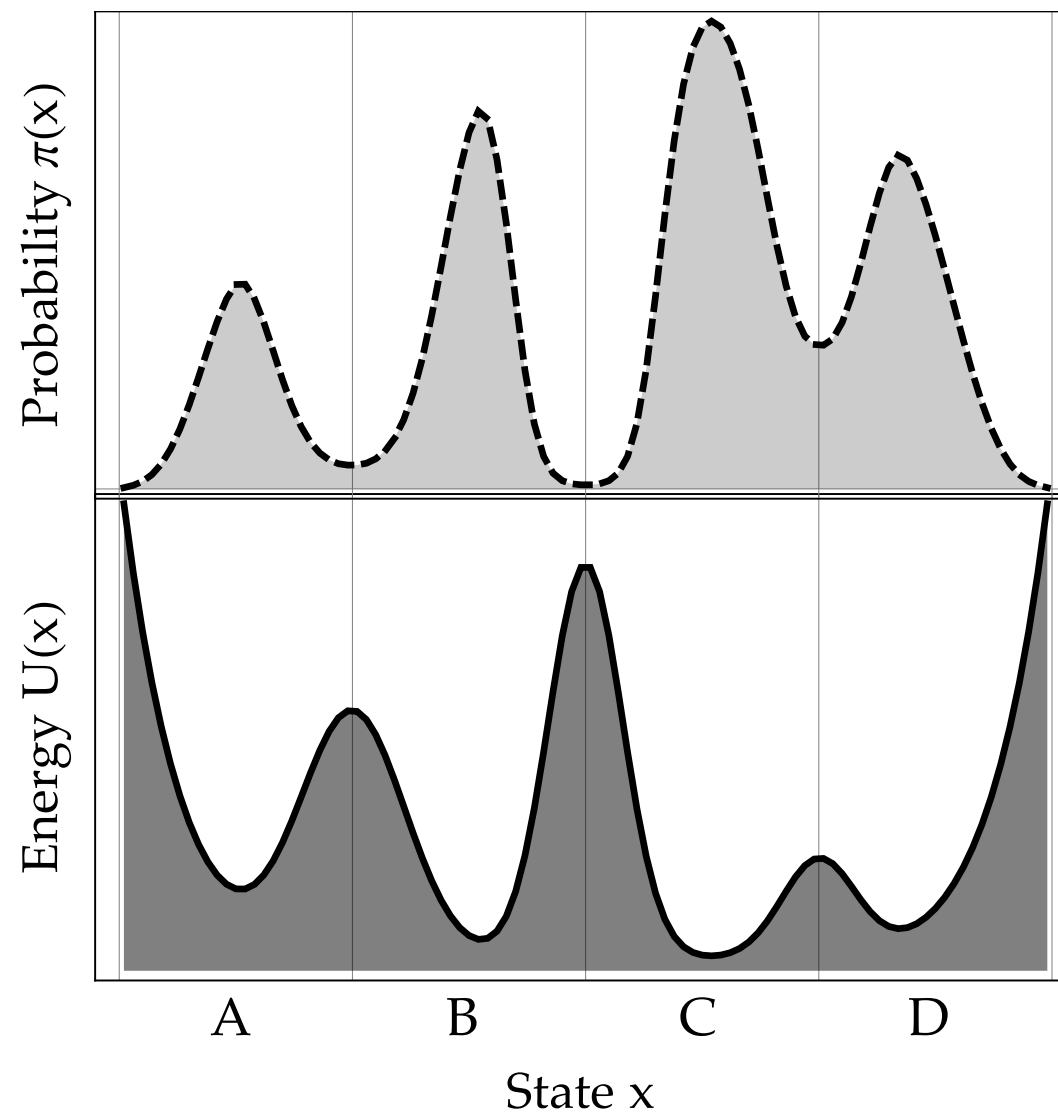
Ensemble view of dynamics



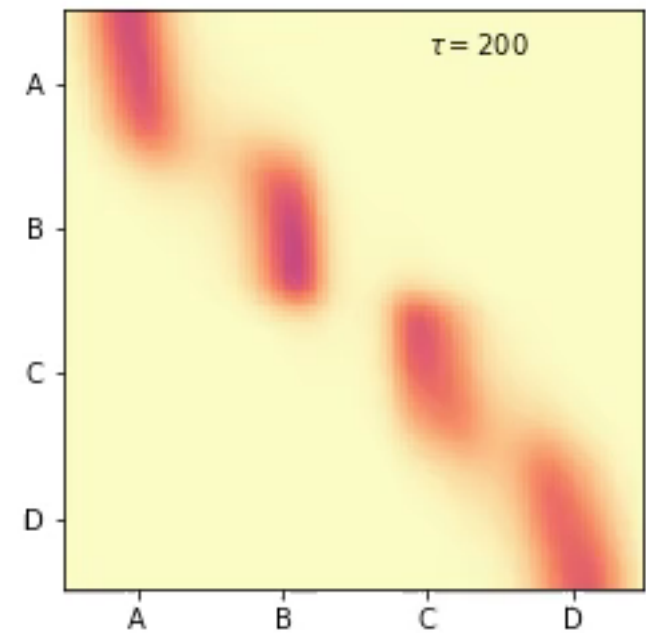
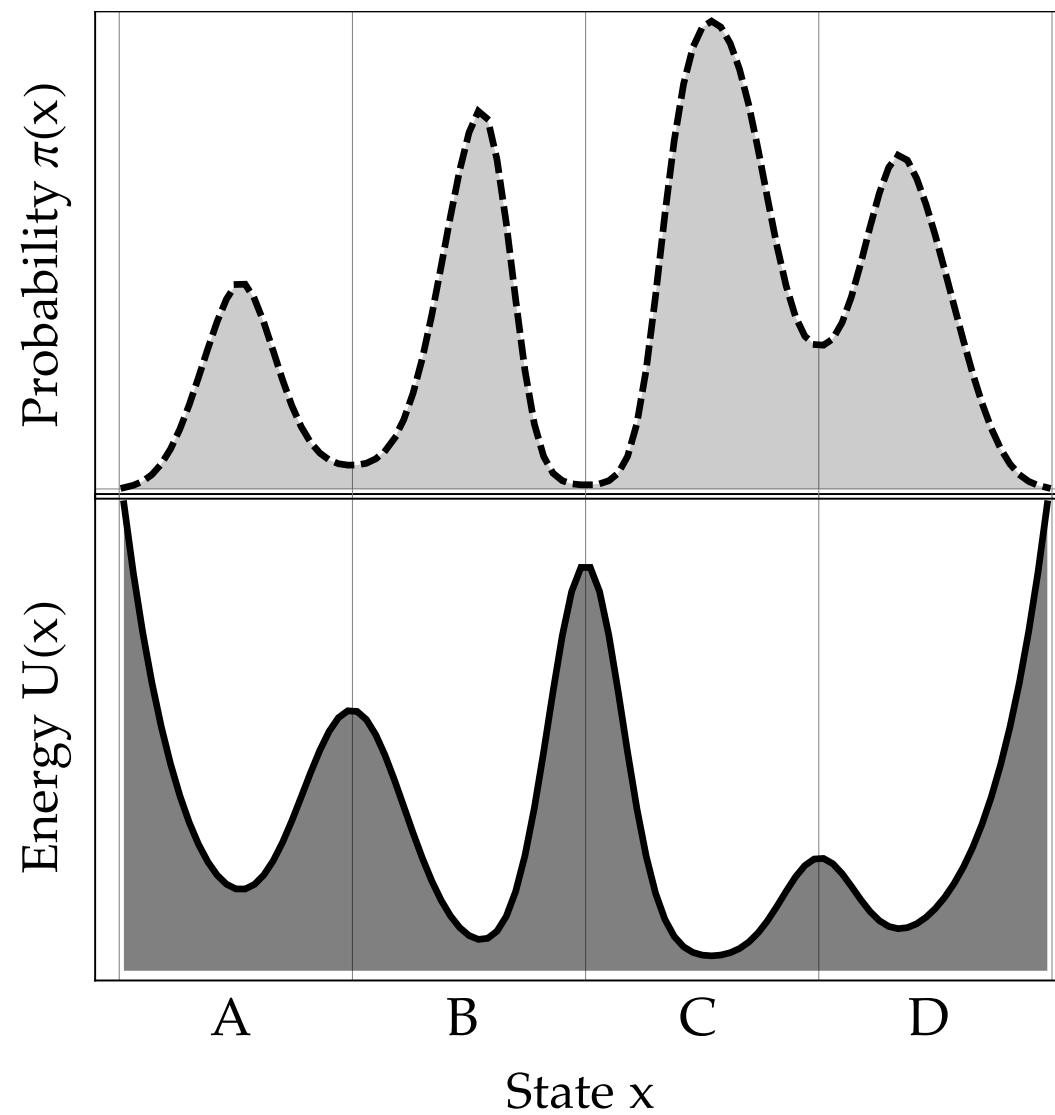
A propagator is an operator which transports probability densities in time

$$p_{t+\tau}(x) = [P_\tau p_t](x) = \int_{\Omega} dy p_\tau(y, x) p_t(y)$$

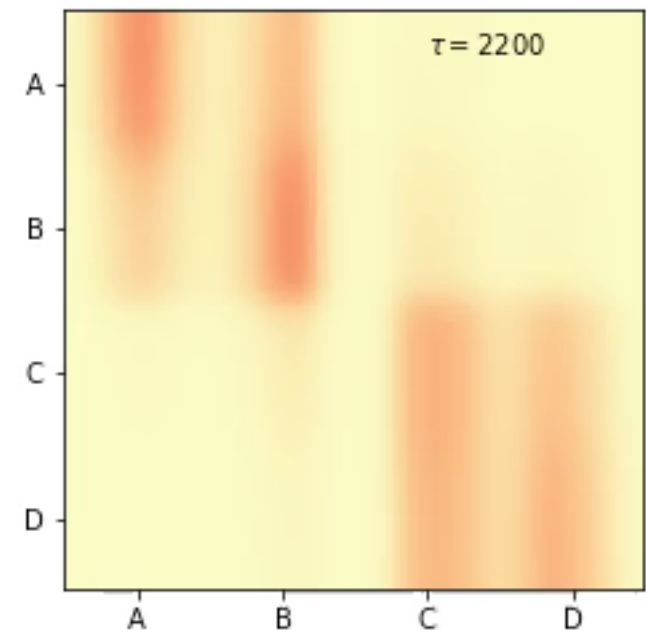
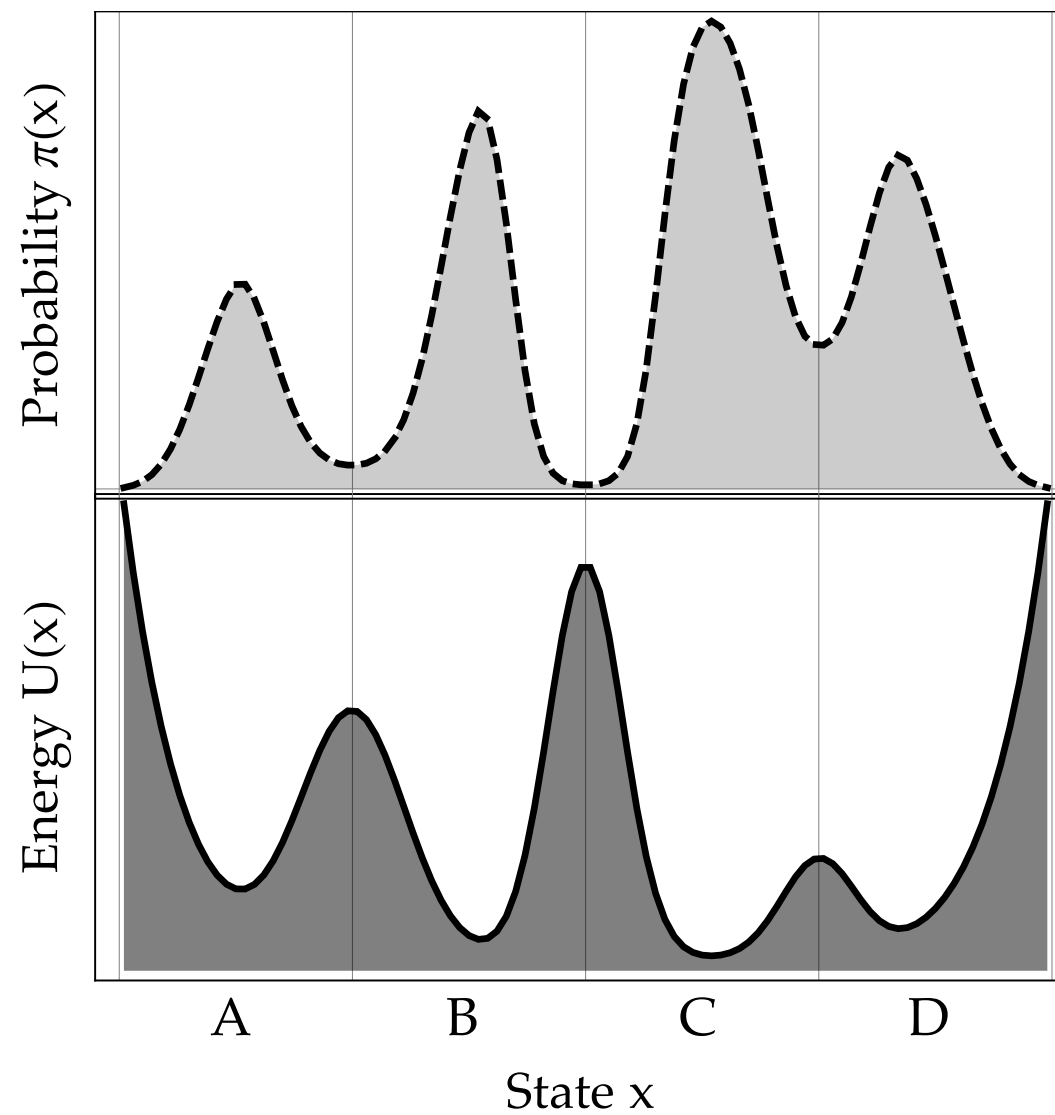
Propagator depends on lag time



Propagator depends on lag time



Propagator depends on lag time



So why is this?

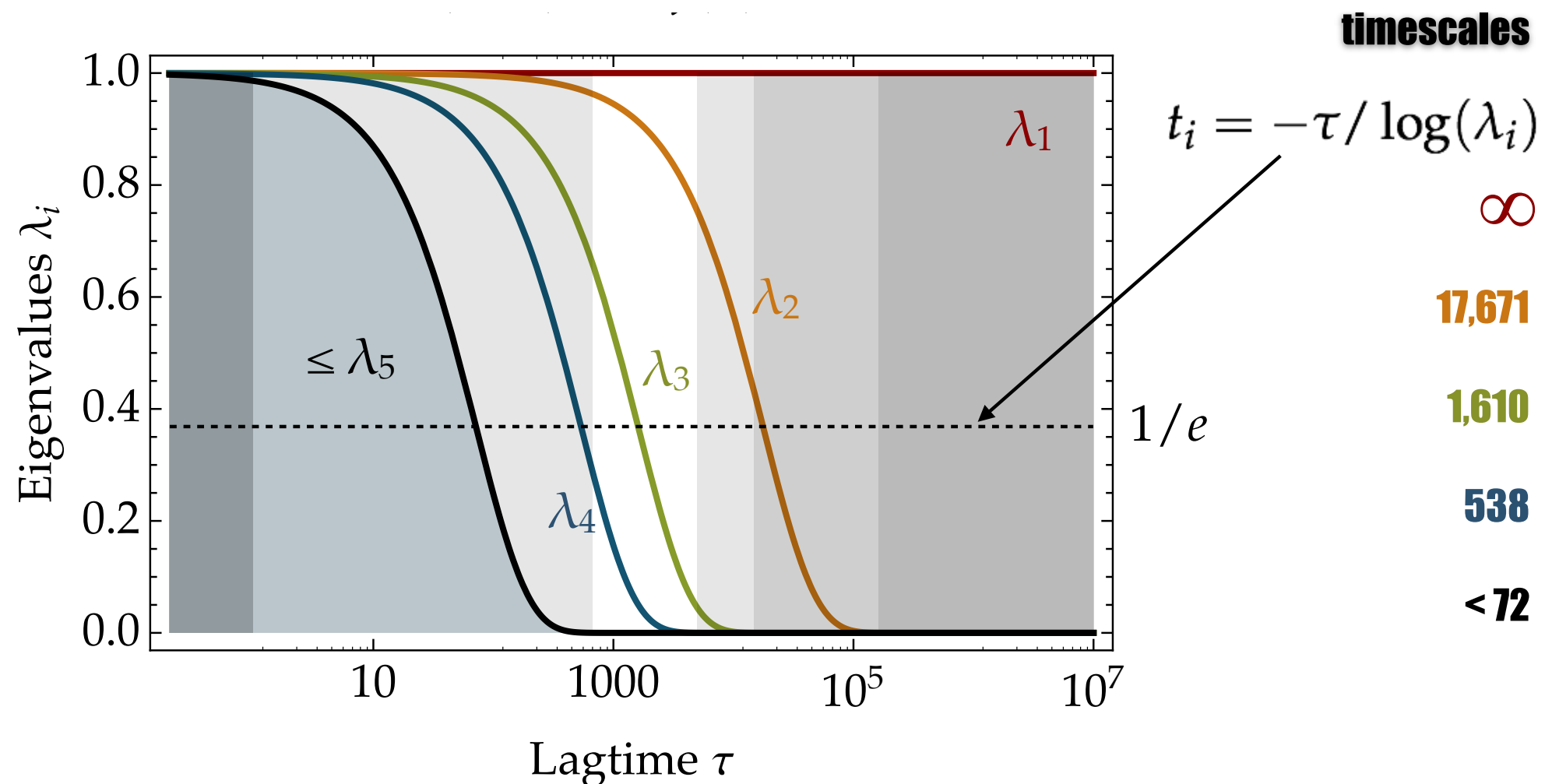
Implied time-scales

Eigenvalues of the propagator

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

Chapman-Kolmogorov Implies exponential lag-time dependence

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

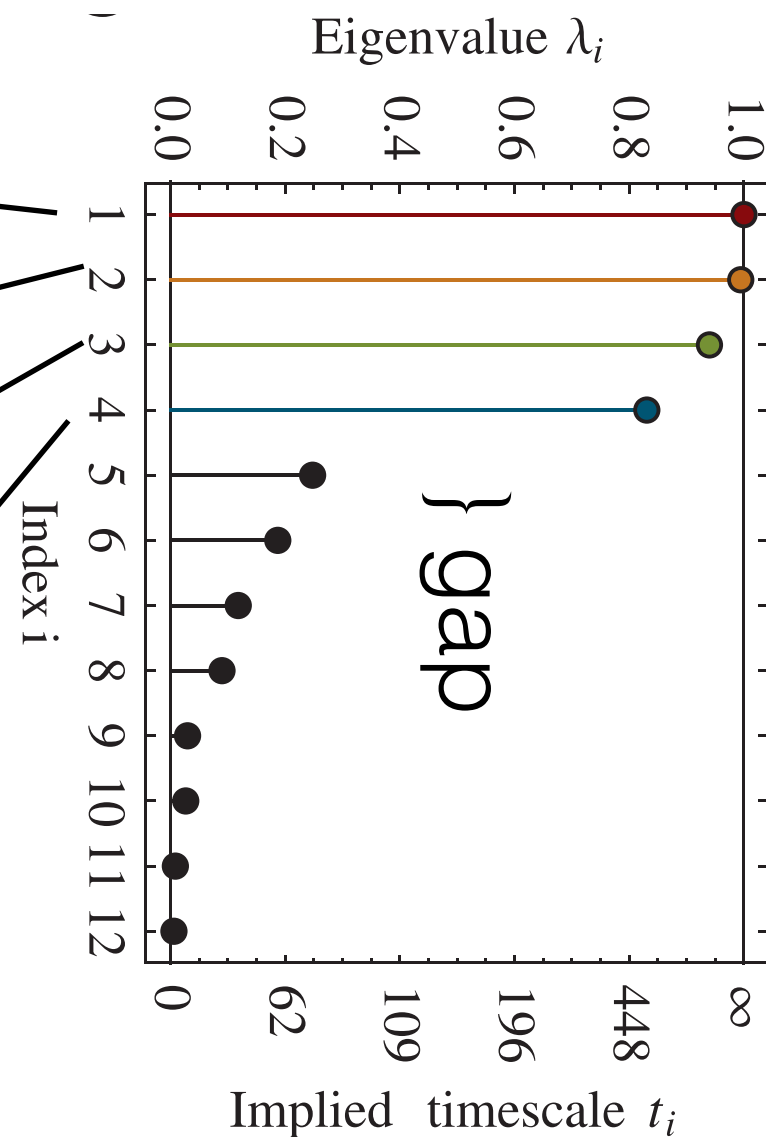
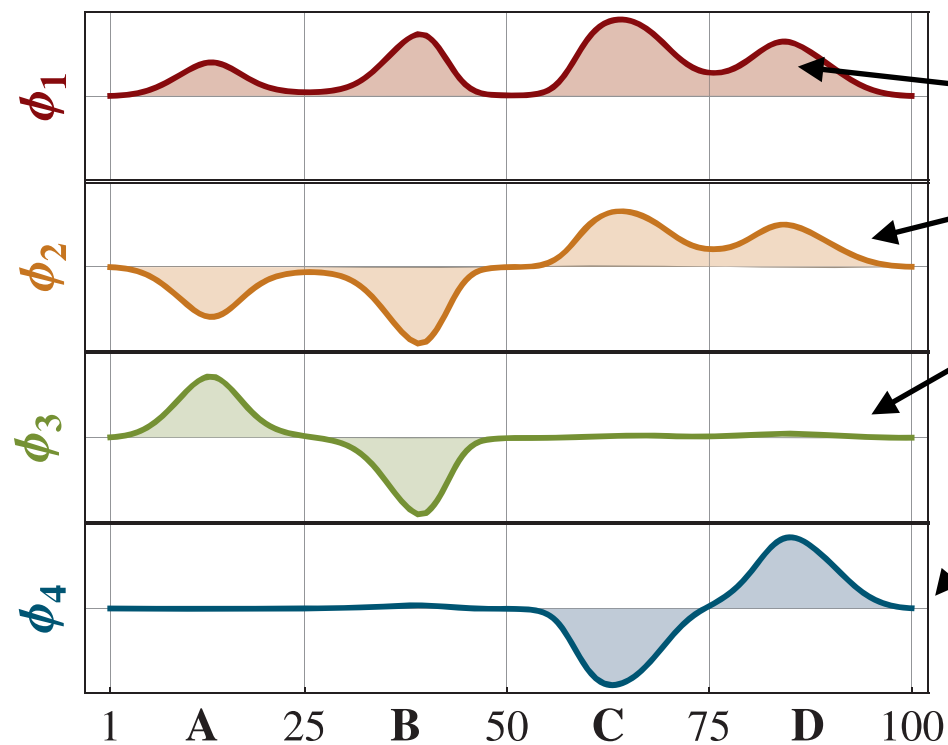


Meta-stability

- We can approximate the propagator by a finite number of processes with non-zero Eigenvalues
- If we have a gap in the Eigenvalue spectrum, we can choose the lag-time in a manner such that we fulfill this assumption
- When we do this, processes faster than the lag-time 'have decayed' or 'are not resolved'.

What do you mean by processes?

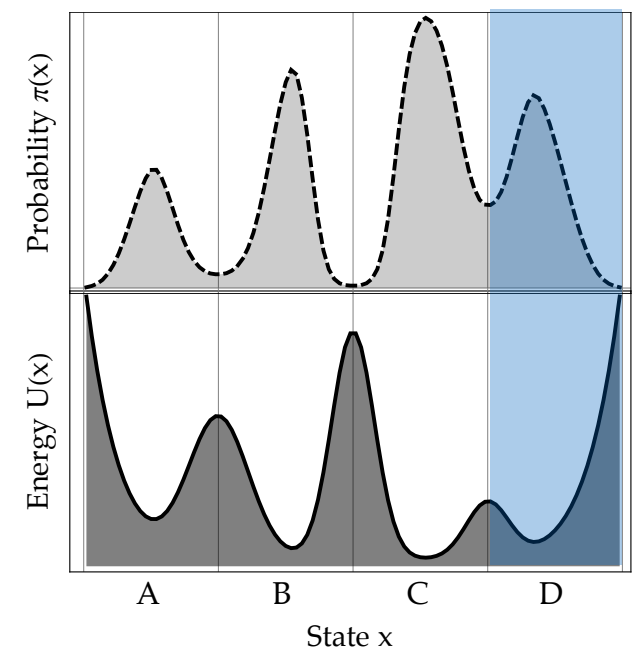
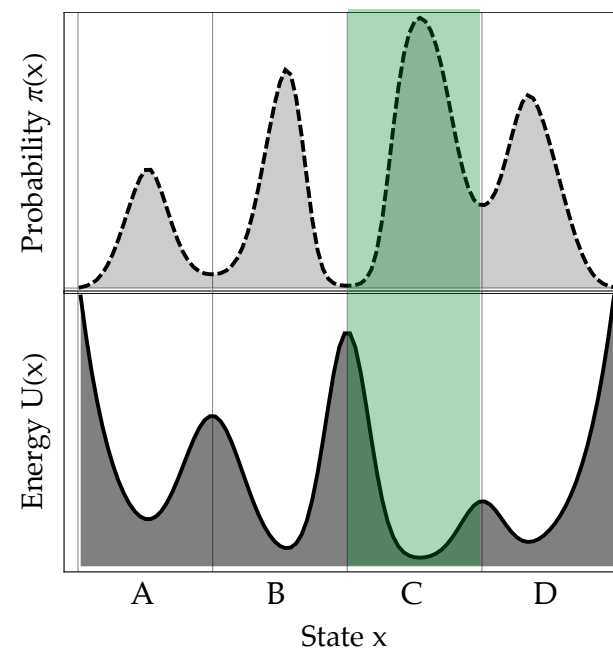
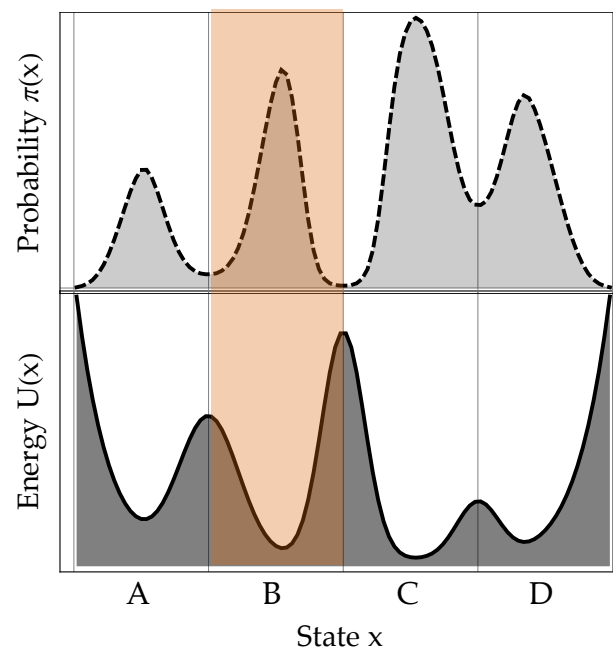
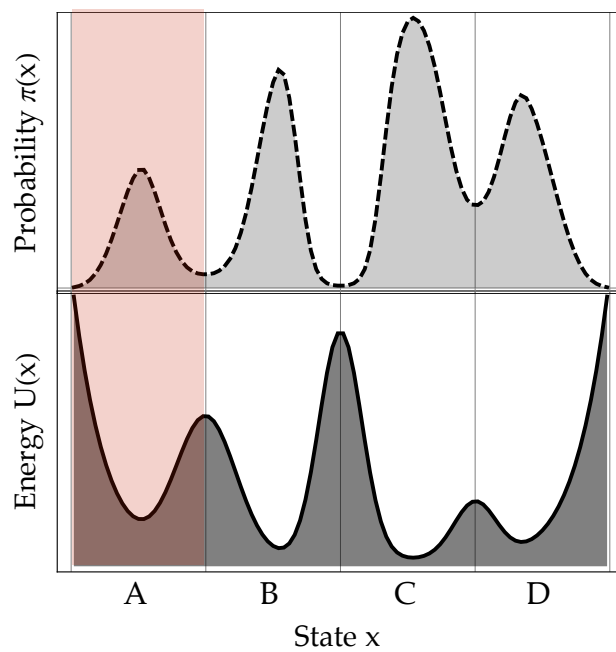
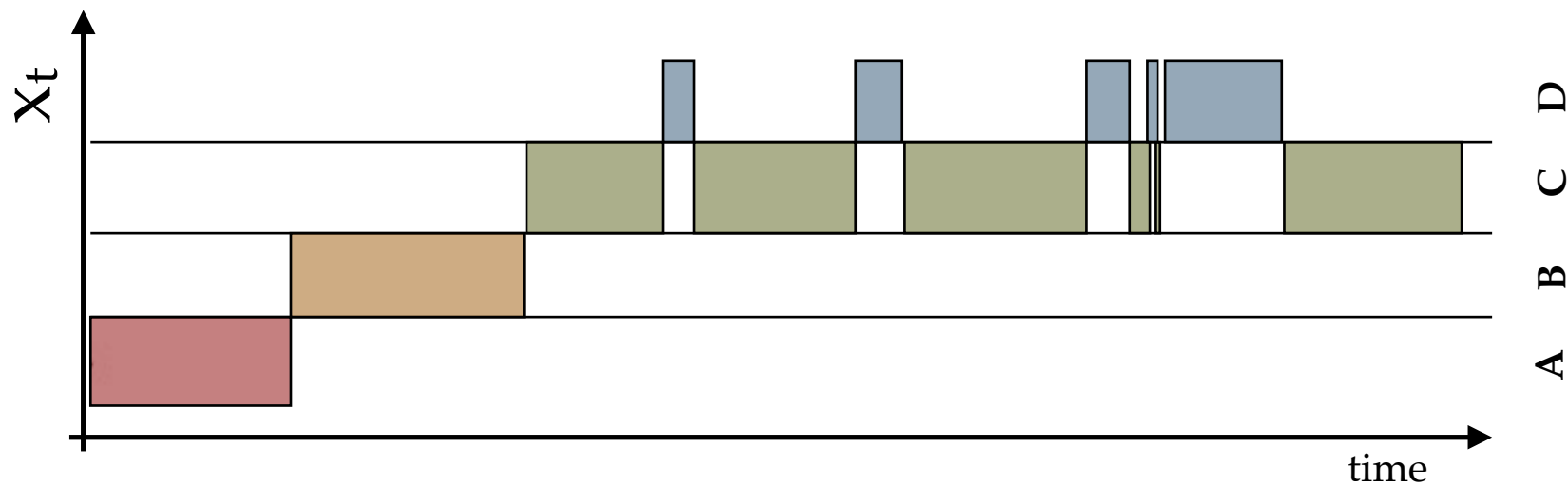
Eigenfunctions of P_τ



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

Estimation

Discretization of Ω



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

$$C_{ij}(\tau) = \delta(x_{n-\tau} = i, x_n = j)$$

Maximum likelihood estimator

We can express the probability of the observed data - discrete trajectory - given a transition probability matrix of an MSM

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

The aim is then to find the P which maximizes this expression - That is, the *Maximum likelihood estimator*.

Analytical solution for Non-reversible case

- We enforce the constraint that the transition probability matrix is row-stochastic:

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

- One can show the estimator is simply:

$$\hat{p}_{ij} = \frac{\hat{C}_{ij}}{\sum_j \hat{C}_{ij}}$$

Reversible estimator

- Enforces the detailed balance condition.
- No exact analytical solution:
 - Fixed-point iteration algorithm available.
 - Approximate solutions.
- Implemented in PyEMMA

Bayesian inference of MSMs

- The less simulation data we have, the more ambiguous the solution of the likelihood problem will be.
- Consequently, if we limit ourselves to the MLE, we are *ignorant* as to how **robust** our inferred MSM is.
- One way to quantify the uncertainty of MSMs is through **Bayesian inference**

Bayesian inference of MSMs

Likelihood from before

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

Bayesian inference of MSMs

Likelihood from before

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

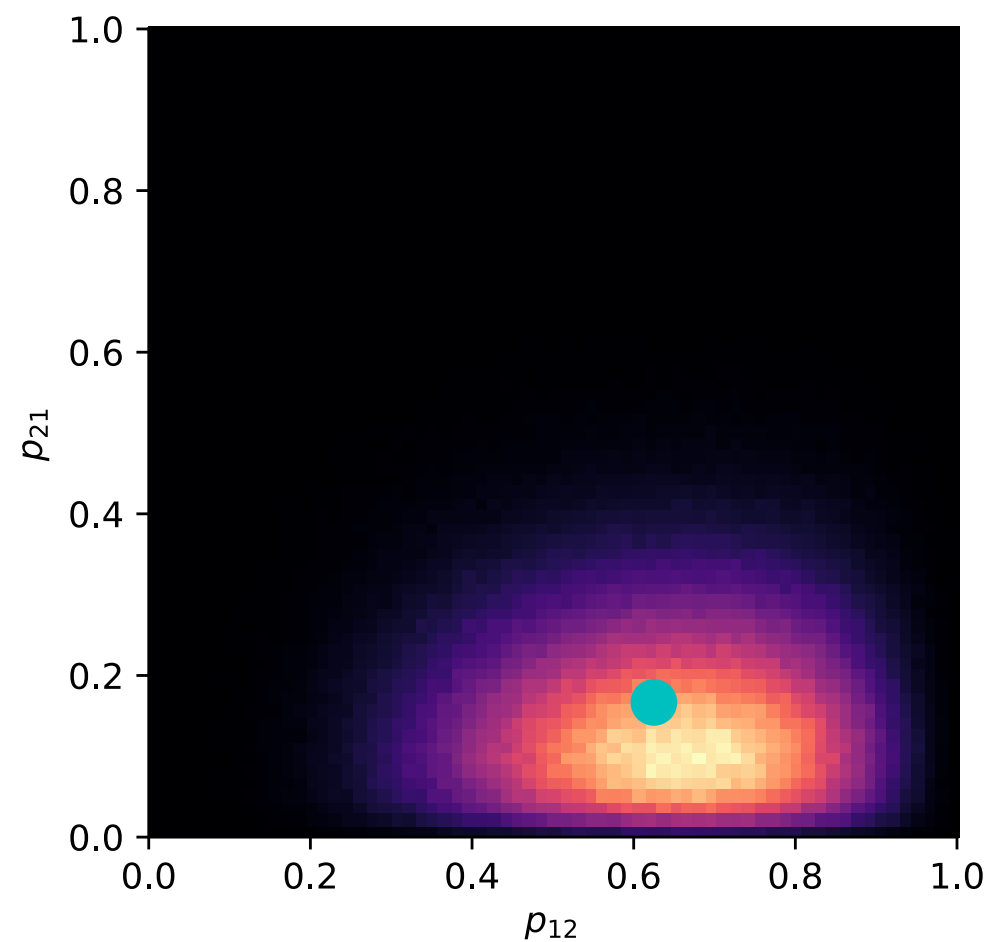
Introduction of prior information

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

The prior can encode useful constraints: row-stochasticity, reversibility, fixed stationary distribution, sparsity etc

Bayesian inference of MSMs

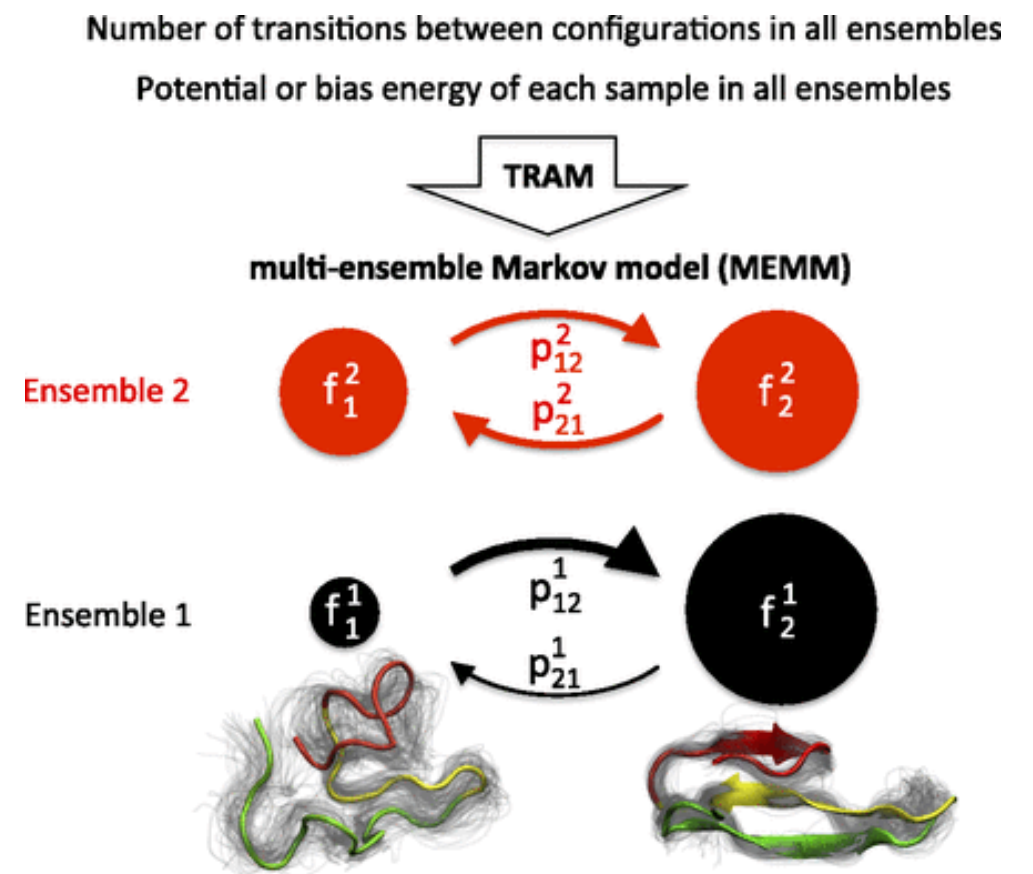
Inference is done by MCMC sampling



Alternative estimators

Transition(-based) Reweighting Analysis Method

- Allows taking into account simulation data from multiple thermodynamic ensembles.
- That means, we can **use data from enhanced sampling simulations together with unbiased simulation data to generate models more efficiently.**
- More about this tomorrow.

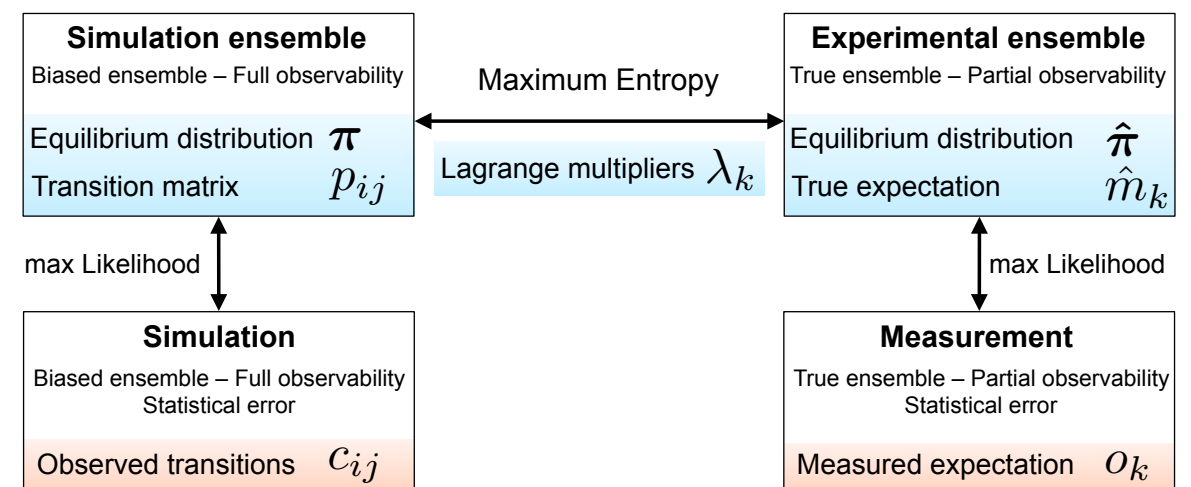


Wu et al. *PNAS* 2016, 113(23), E3221–E3230

Implemented in PyEMMA

Augmented Markov models

- Enables integration of external information into the estimation of Markov state models.
- Fx use of experimental constraints from biophysical experiments such as NMR.
- A notebook tutorial distributed with PyEMMA 2.5 and up.



Olsson et al. *PNAS* 2017, 114(31), pp. 8265-8270. doi: 10.1073/pnas.1704803114

Implemented in PyEMMA

Analysis of our estimate

$P_{ij}(1)$	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

Time-scales are always under-estimated

Increasing the lag-time

**COUNT
MATRIX**

$C_{ij}(100)$	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

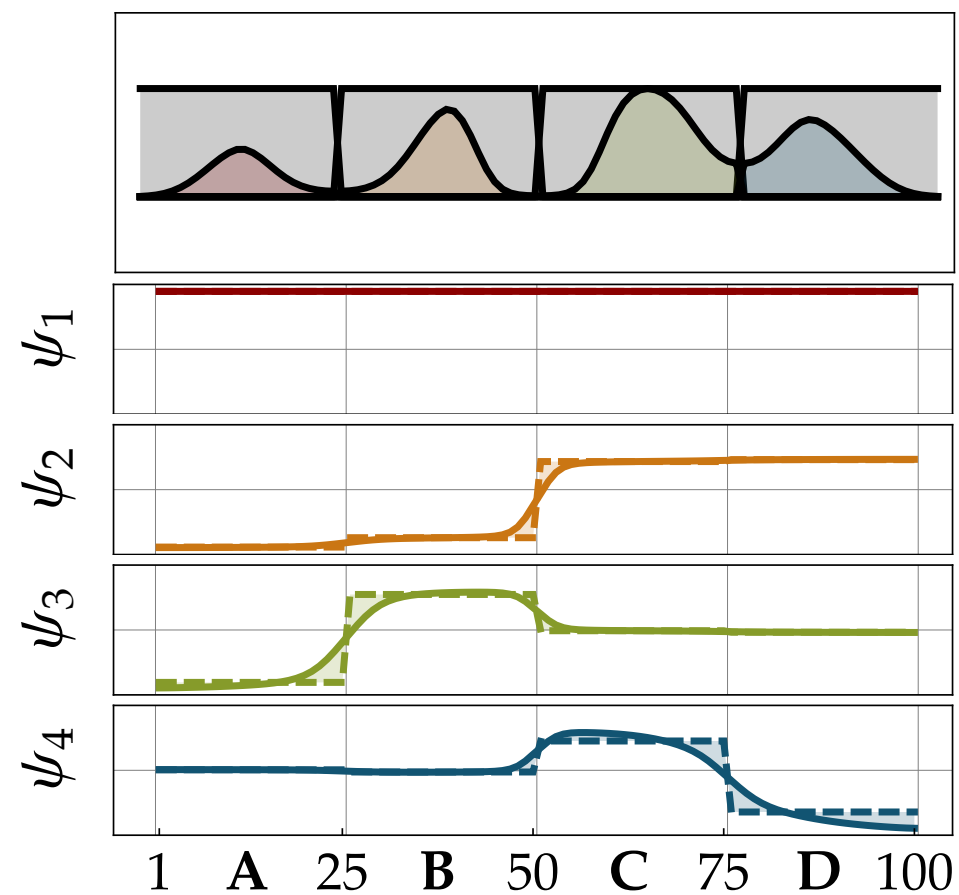
May improve estimates of predicted time-scales

Projection/discretization error

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

metastable region

GOOD PROJECTION

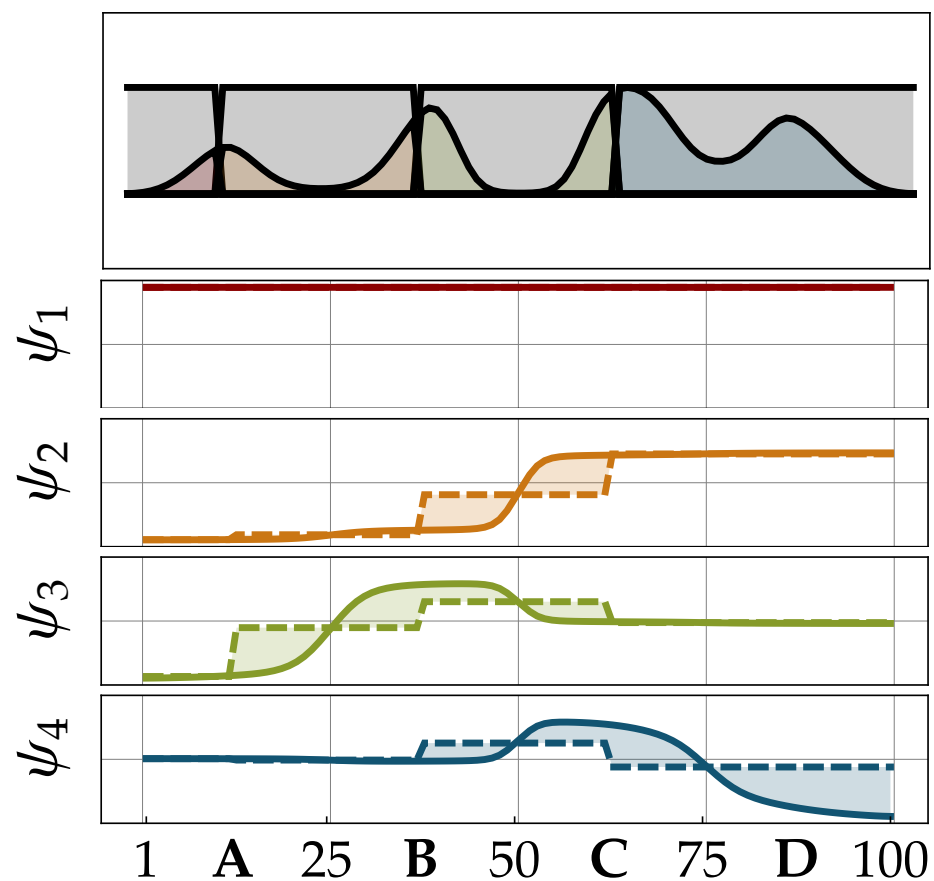


Projection/discretization error

metastable region

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

BAD PROJECTION



Known problems

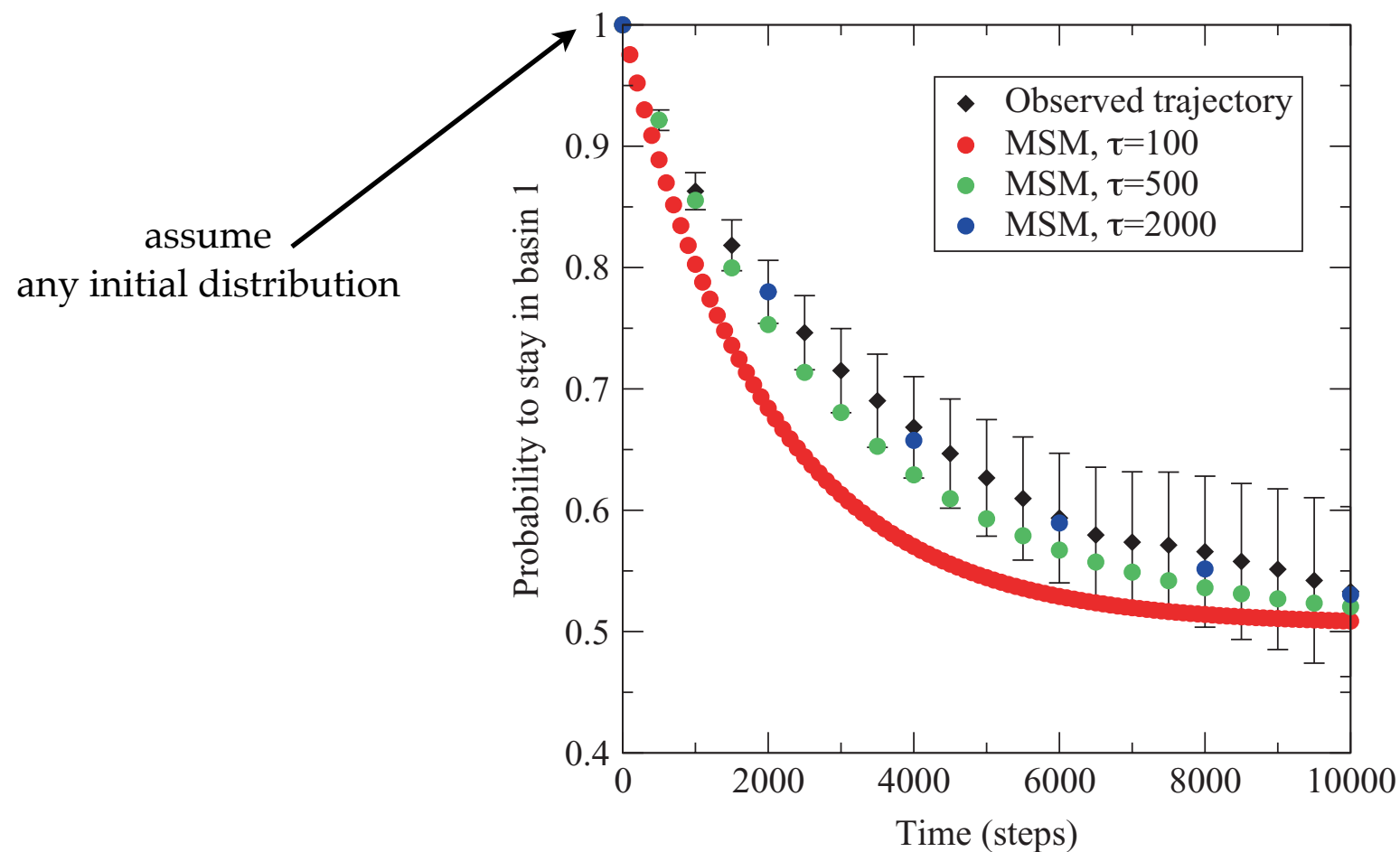
- Observations (projections, discretizations) are in many cases not Markovian
- However, we are often interested in understanding the full system not just the observation.
- Since we often have a lot of freedom to choose the projections and discretization, it is important to choose one which is as Markovian as possible.

Validation

Chapman-Kolmogorov test

Compare the evolution of the data with the model

$$\underbrace{T^k(\tau)}_{\text{Markov model prediction}} \approx \underbrace{T(k\tau)}_{\text{estimation from data}}$$



General scheme for Markov state model generation

- Discretize a suitable projection of your data.
- Construct a transition matrix.
- Estimate the number of meta-stable states (time-scale gap)
- Perform Chapman-Kolmogorov test.

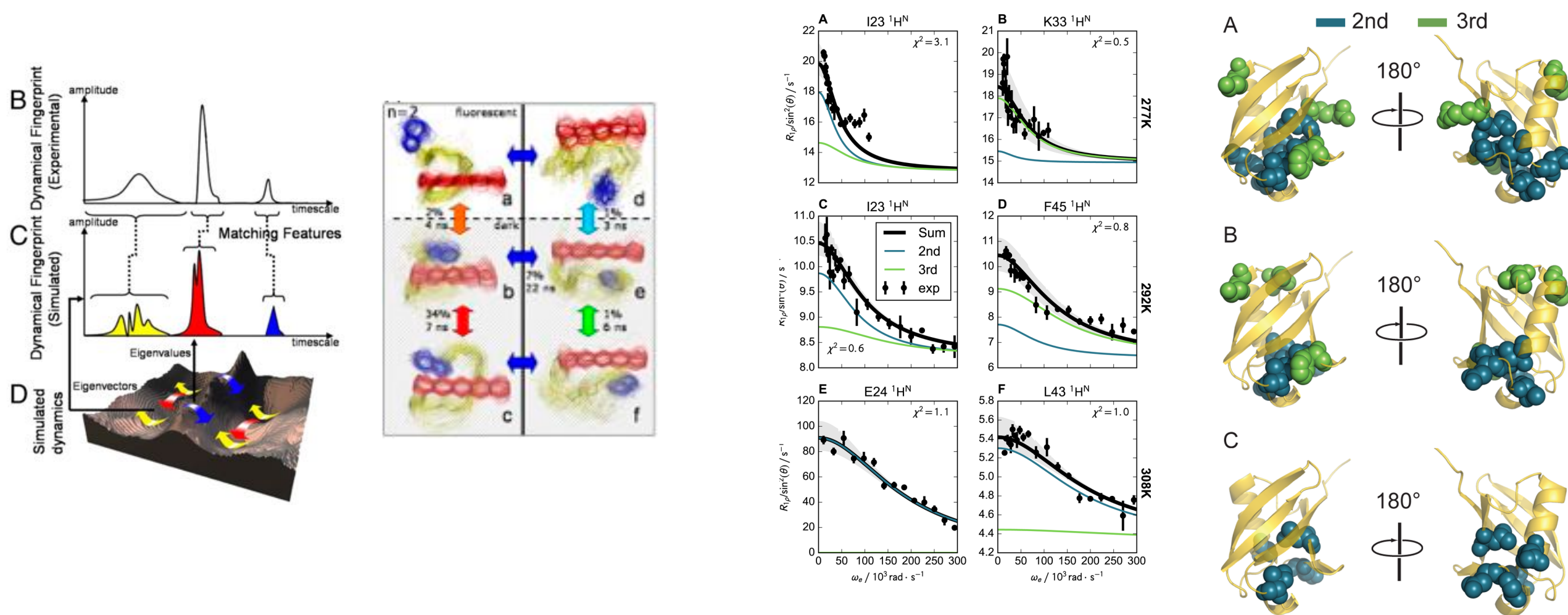
Analysis

Useful predictions from a MSM

Common properties

- Relaxation time-scales
- Dominant processes
- Stationary distribution (thermodynamics)
- Meta-stable sets (more about this later)
- Correlation functions (spectroscopic observables)
- Mean first passage times
- Path probabilities

Spectroscopic observables



Noé 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).

Olsson & Noé *Mechanistic Models of Chemical Exchange Induced Relaxation in Protein NMR*. 139, 200–210 JACS (2017)

Summary

- Markov state models are derived coarse-grained models of the full original (Markovian) dynamics .
- MSMs may be parameterized (estimated/learned) from simulation data to compute properties of interest.
- MSMs are particularly useful if the projection/discretization error can be minimized: then the predicted quantities match the original.

Questions?