

# Markov-based analysis of biomolecular systems pt. I



Toni Giorgino

National Research Council of Italy

[toni.giorgino@cnr.it](mailto:toni.giorgino@cnr.it)



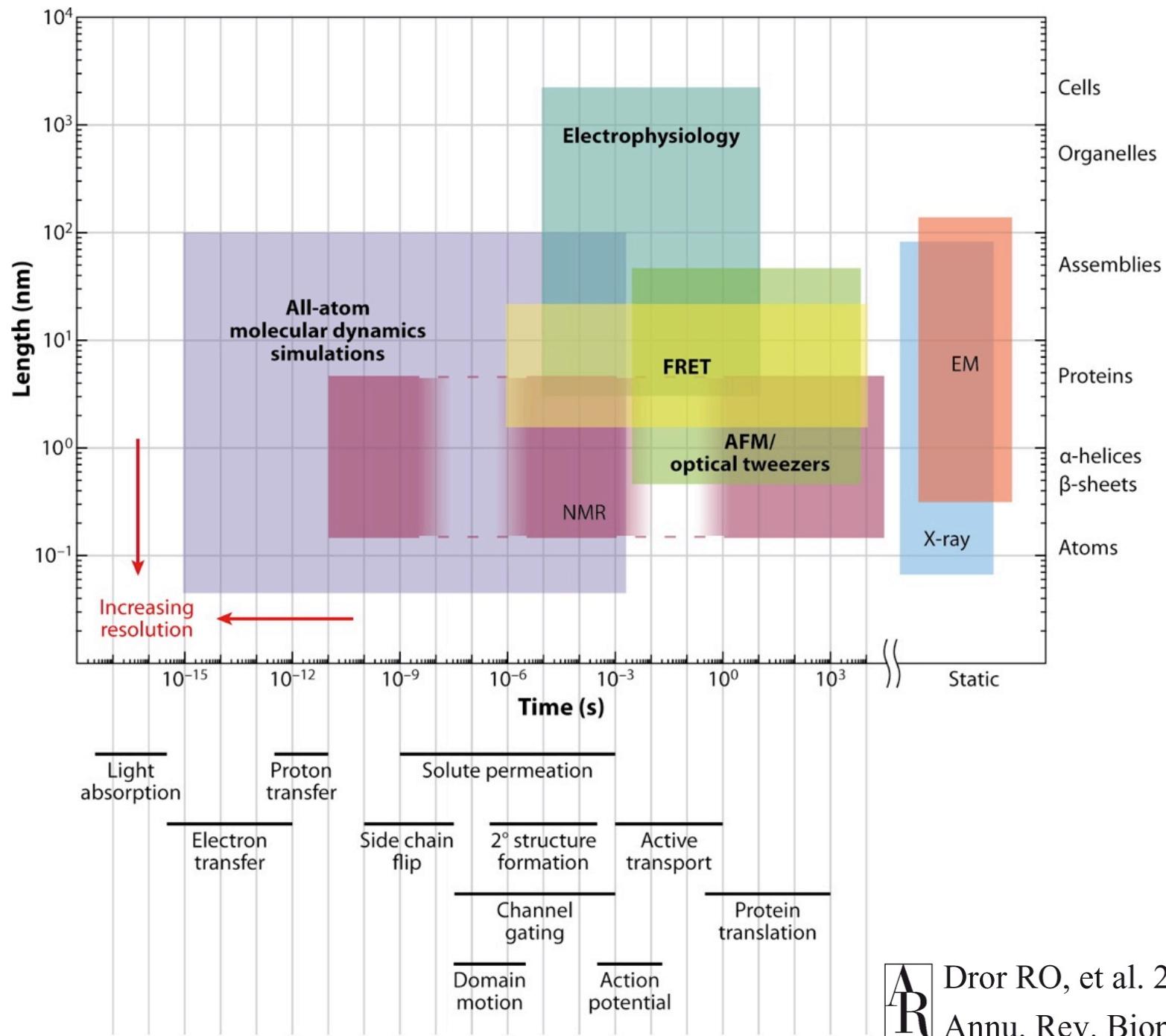
Master in Bioinformatics for Health Sciences  
Barcelona, 26 Apr 2019

# Introduction

- The aim of this class is to provide a practical overview of Markov state models in computational structural biology
- MSM emerging because:
  - reconstruct *kinetic information*\*<sup>†</sup>, including state transition networks, from simulated trajectories
  - start from **unbiased simulations** (no *a priori* reaction coordinate hypothesis necessary)
  - microsecond-scale (high-throughput) trajectory data are becoming accessible (e.g., with GPUs)
- Success cases: *ab initio* folding, drug binding, peptide binding, ...

\* as well as the corresponding structures

<sup>†</sup> see notes



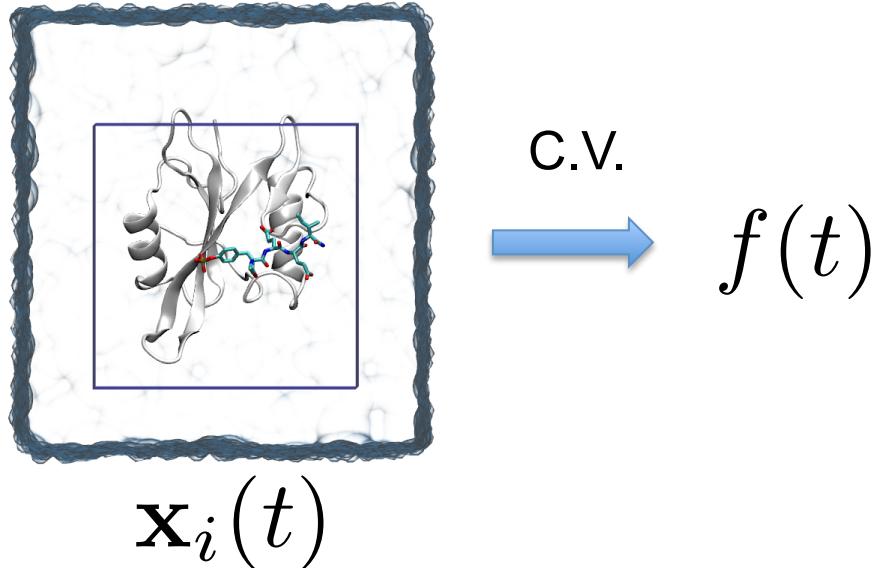
Dror RO, et al. 2012.

Annu. Rev. Biophys. 41:429–52

# **Trajectory analysis and collective variables**

et al., JCTC 2012

# Collective variables



- Distance
- Angle
- Energy
- Contacts
- Solvation
- ...

CVs are more *meaningful* than raw coordinates. To compute them:

- VMD (mouse)
- VMD (TCL scripting)
- PLUMED (scripting+driver)
- Other libraries (Python, R)

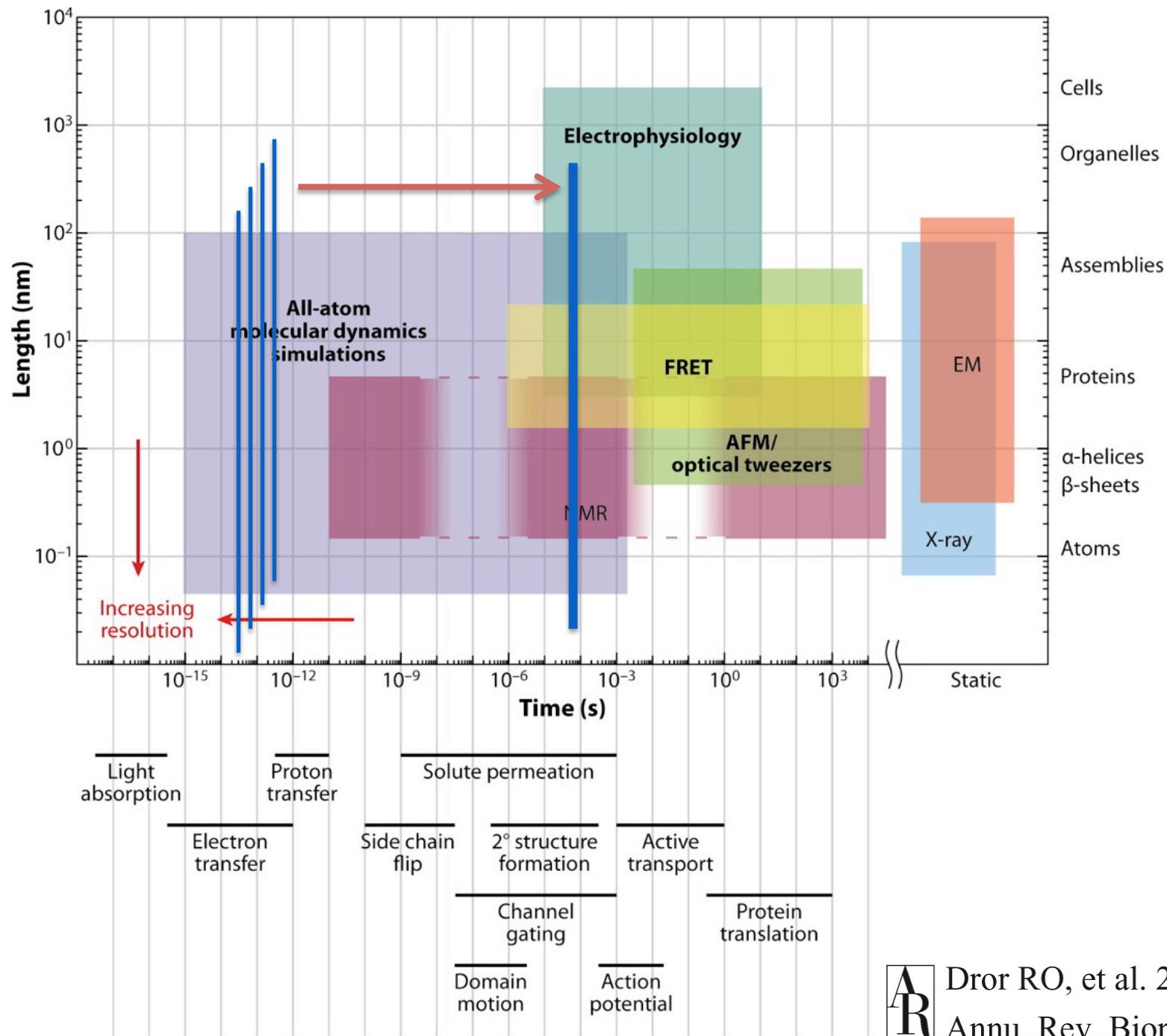
These tools “understand” a trajectory + structure files ( $C\alpha$ , helix...)

# Time-series analysis of CVs

- Useful when I have a system *at equilibrium*
- I.e., it is exploring states *around* the biologically relevant structure
- Problem: cannot make deductions beyond the observed time-scales.
- TS of the phenomena of interest are often very long (barriers).
- E.g.: drug binding, conf. changes, folding, permeation, ...

# Motivation

- Can we use an *ensemble* of simulations to estimate dynamic behaviours which happen on timescales longer than each of the observed trajectories?
- In other words, can we leverage several “short-sighted” (non-equilibrium) observations to extrapolate long-time behaviour?



Dror RO, et al. 2012.

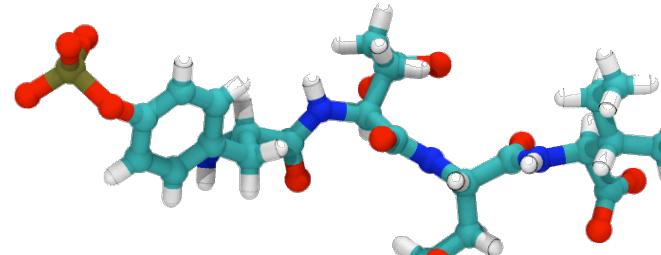
Annu. Rev. Biophys. 41:429–52

# **Basic example: two-state model (binding)**

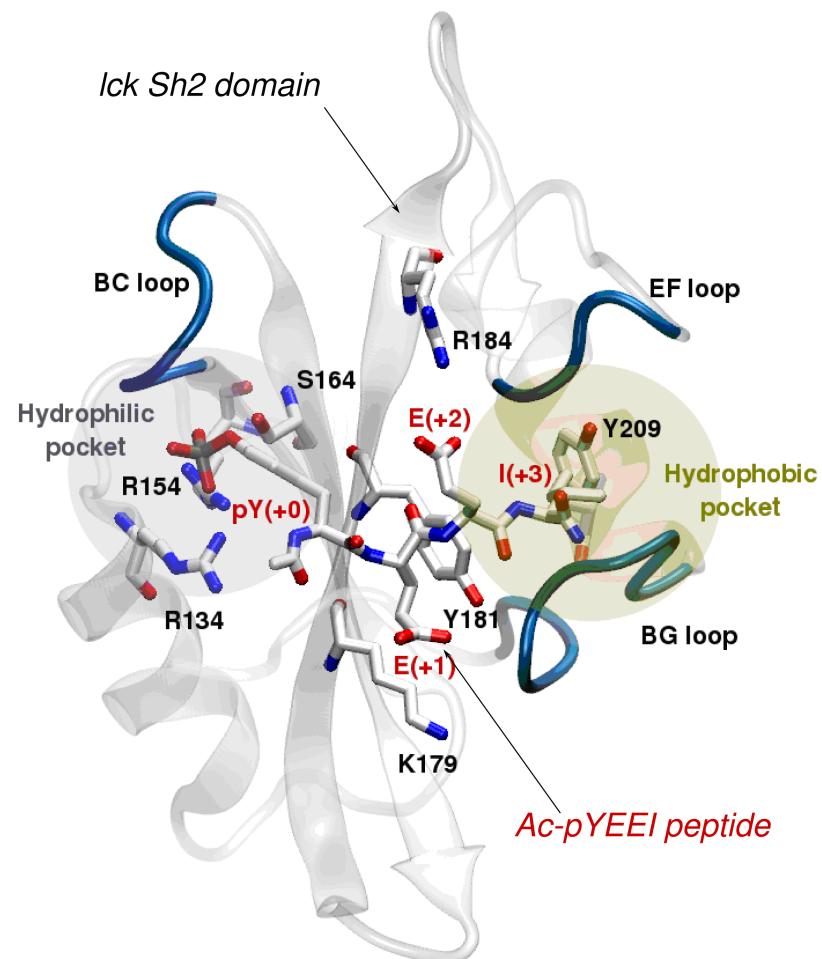
**Unbiased simulations:  
SH2 – pYEEI binding example**

# SH2 Domains

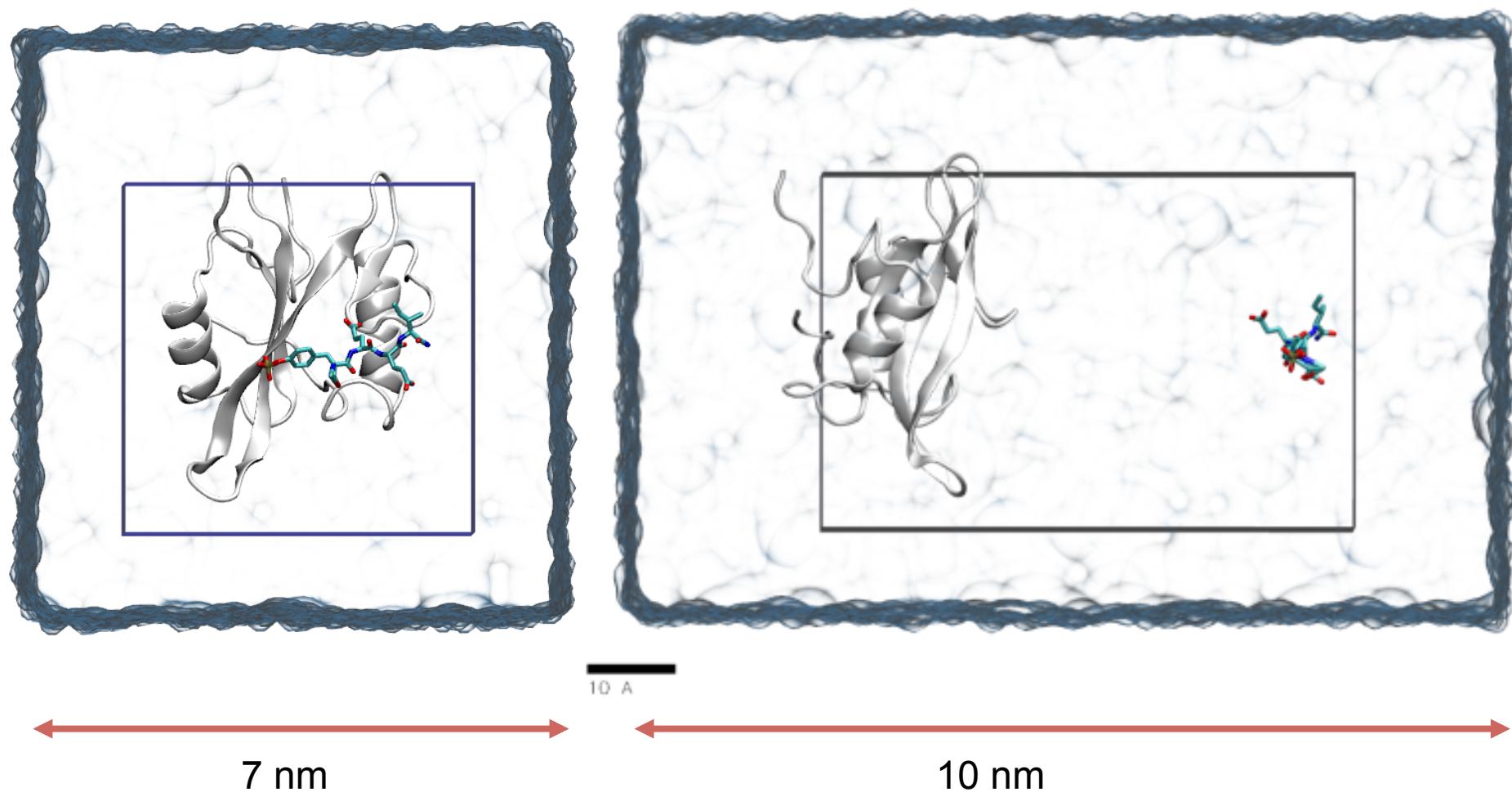
- Small domain (100 AA), well-conserved
- Recognize short pY-\* sequences with high specificity
- Prominent role in signaling
- Found in ~110 human proteins
- “Socket with two holes”



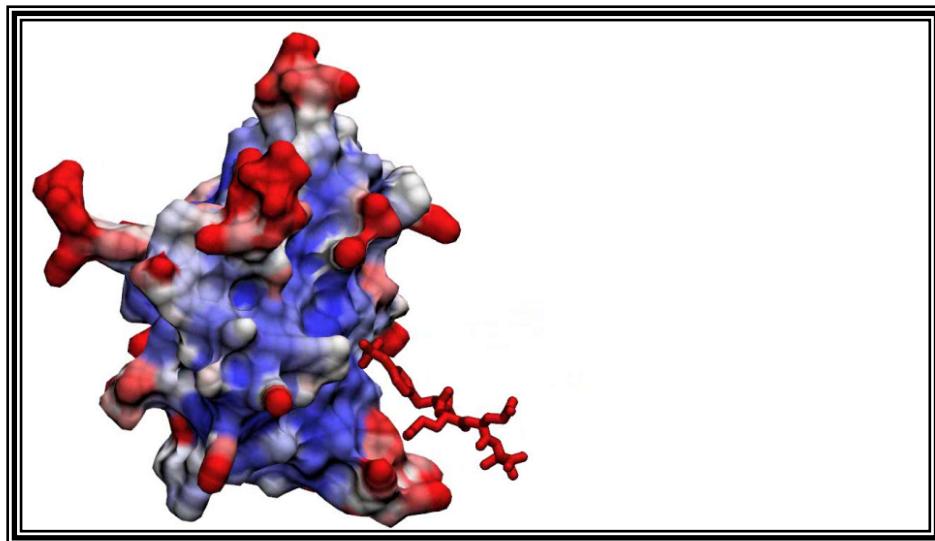
pTyr Glu Glu Ile



# In silico set-up



# Teaser (1:36)



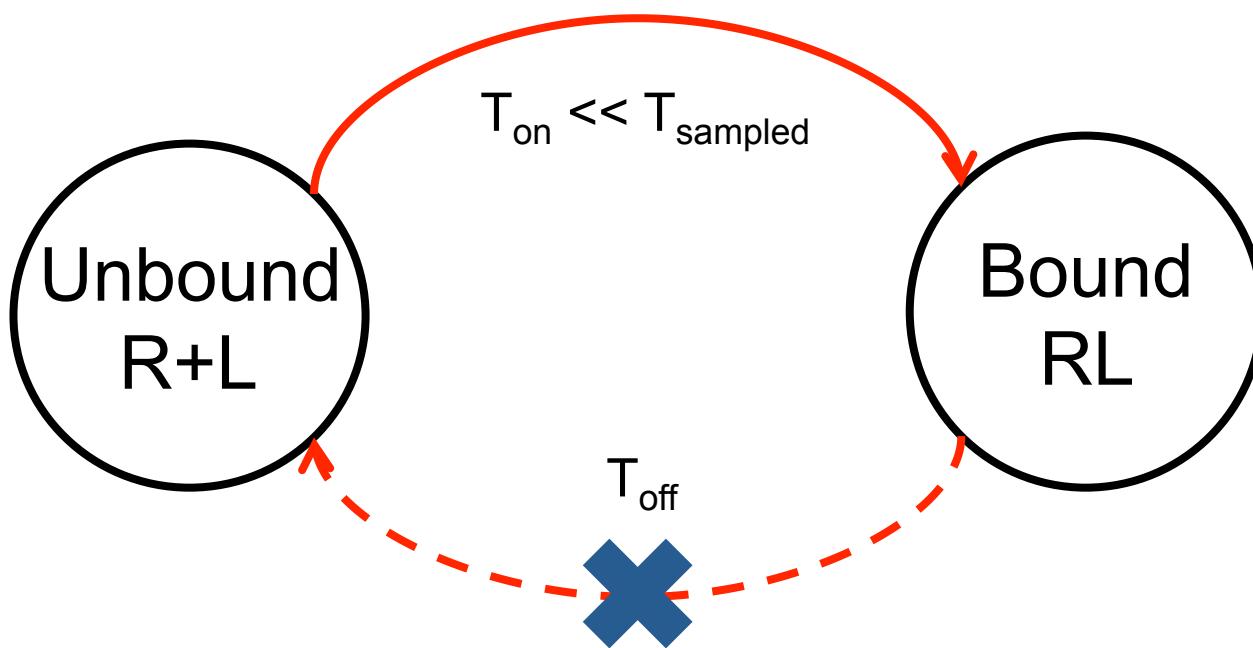
<http://goo.gl/ZLI42>

also [https://youtu.be/Xhof-pf\\_Eo8](https://youtu.be/Xhof-pf_Eo8)

- Video: *unbiased binding simulation of pYEEI to Lck SH2 domain.*
- *Protein surface and ligand's atoms color-coded by local flexibility, from blue to red (lowest to highest RMSF).*
- *As the peptide binds into its native conformation ( $RMSD < 2 \text{ \AA}$  from crystal structure), it loses flexibility; the same happens to the protein.*
- **By construction, we are not at equilibrium.**

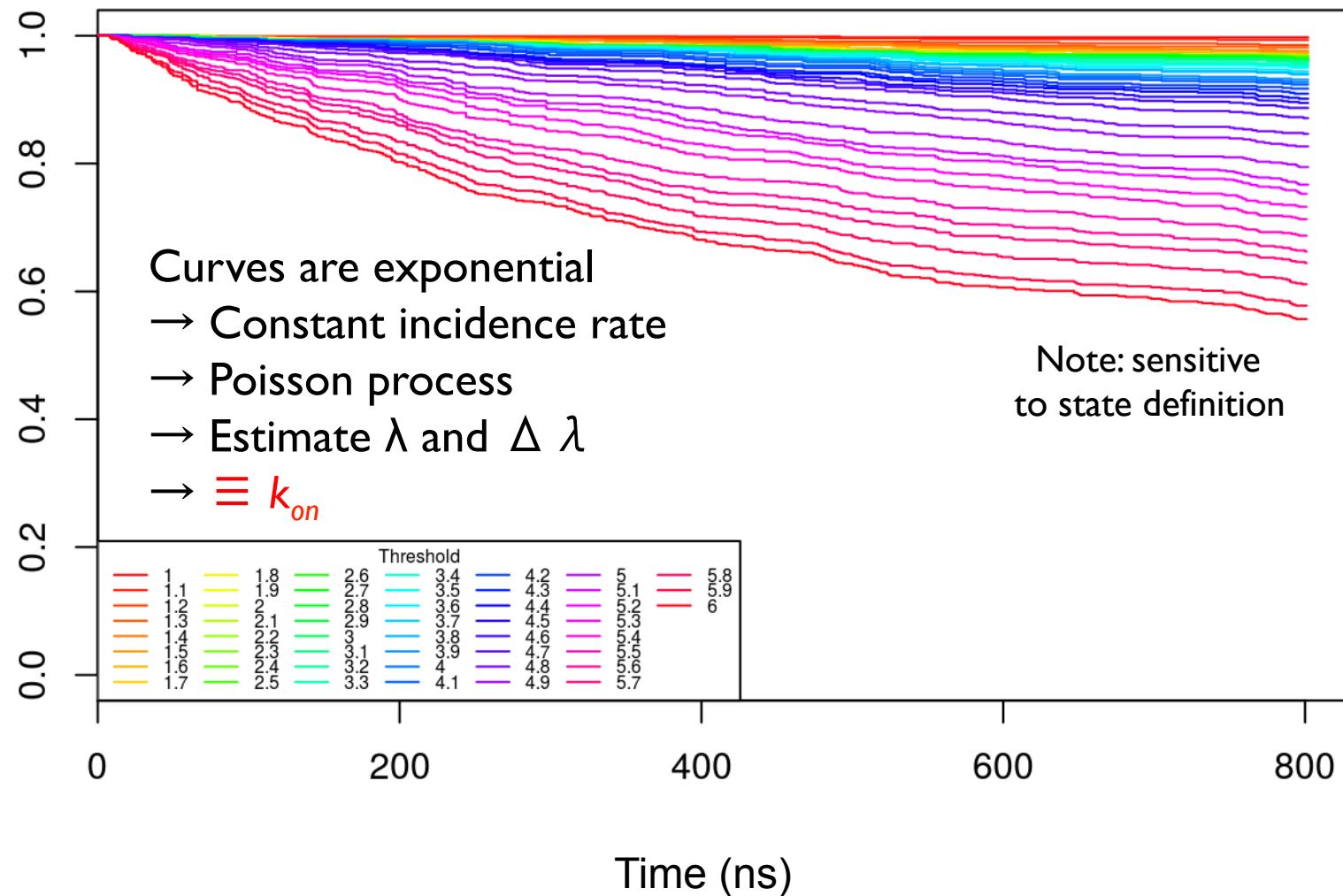
# Binding as an *irreversible* two-state model

On the sampled timescales...



$$T_{event} \ll T_{on} \ll T_{sampled} \ll T_{off}$$

*“Out of ~1000 equivalent trajectories, how many reached the bound state, and when?”*



# Let's look for a more general solution

This 2-state method is not sufficient:

1. One needs to define the *bound state*
  - which is not available if we lack a crystallographic or NMR structure.
2. We need to observe a sizeable number of *full* binding events.
  - Association processes are slow.  
They may be (far) outside our sampling power.
3. We get no information about *off* rates.
4. We get no information about *metastable states*.

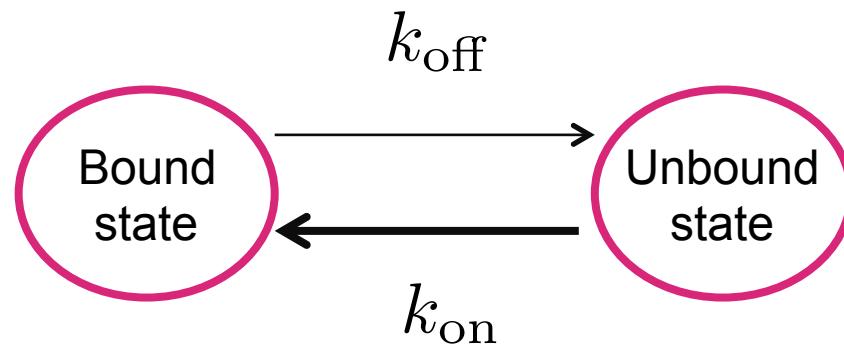
# **Going multi-state**

# Kinetics

We can use the *dynamic* information, computing the *kinetics* as well. This is why we are doing *unbiased dynamics* in the first place!

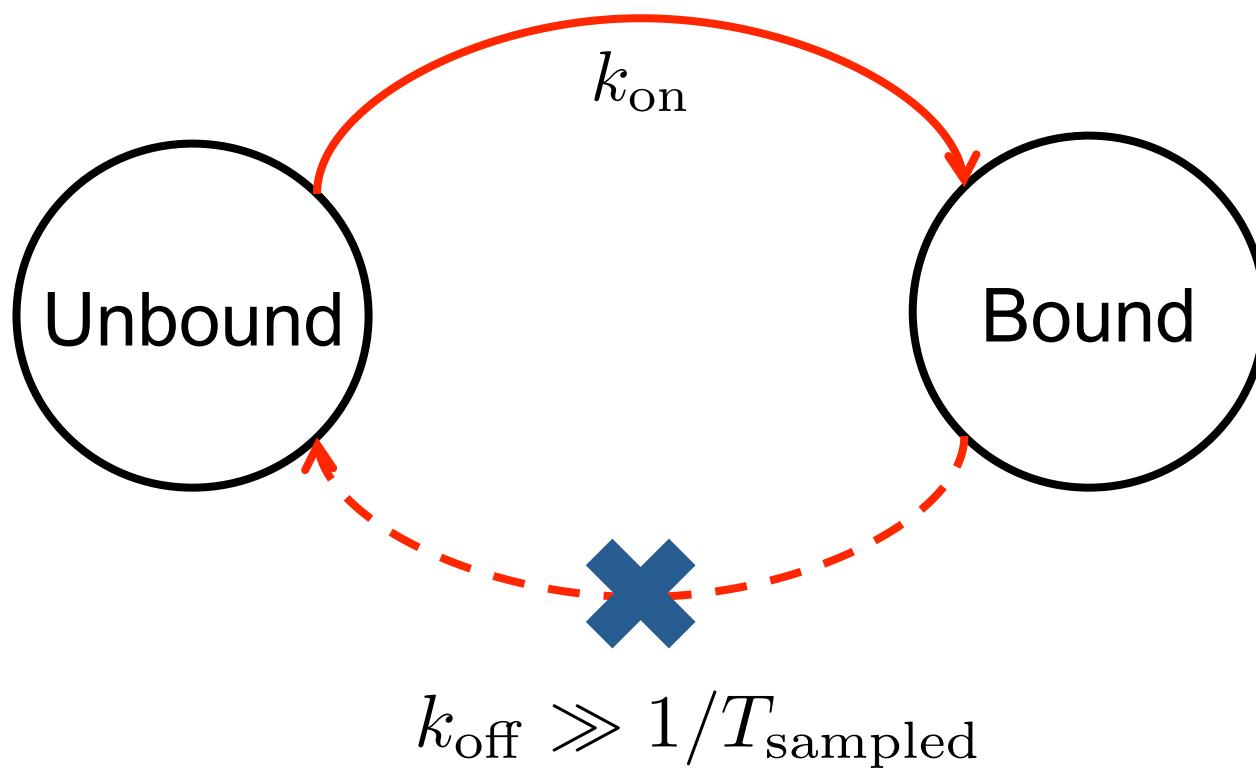
Binding kinetics is important for rational drug design as well (e.g. related to the time spent in complex with receptors).

$$\begin{aligned} K_D &= k_{\text{off}}/k_{\text{on}} \\ &\propto \tau_{\text{on}}/\tau_{\text{off}} \\ &\sim t_U^{(B)}/t_B^{(U)} \end{aligned}$$



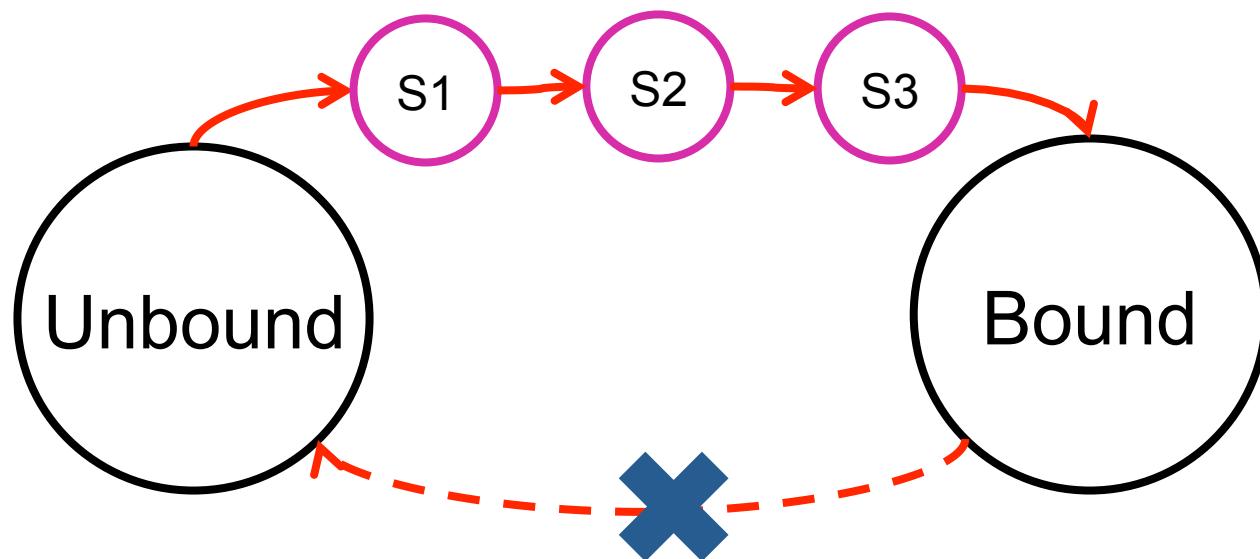
# Let's move beyond this model...

On the sampled timescales...



# Idea

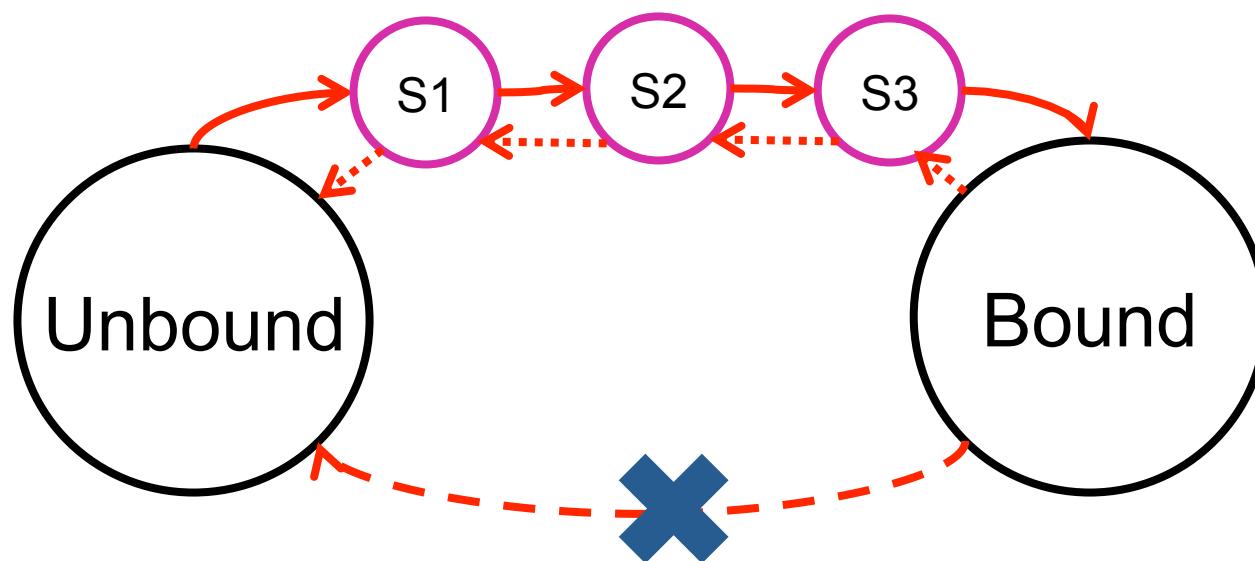
- We introduce several *intermediate steps*



$k_{\text{on}}$  reconstructed combining rates of different intermediate steps

# Idea

- We introduce several *intermediate steps*
- Compute MFPT



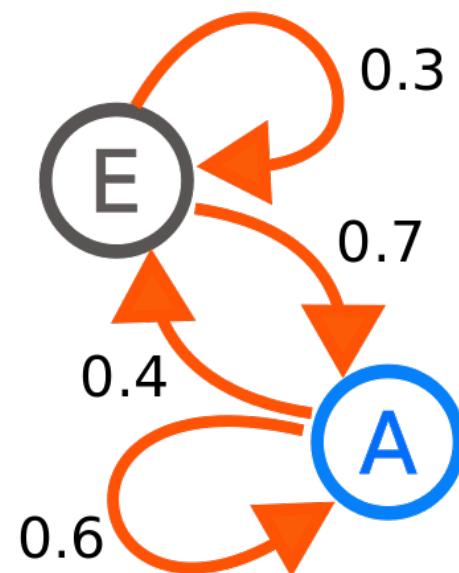
$k_{\text{on}}$  reconstructed combining rates of different intermediate steps

$k_{\text{off}}$  may also be reconstructed, if states are fine-grained!

# **Discrete-time Markov chains**



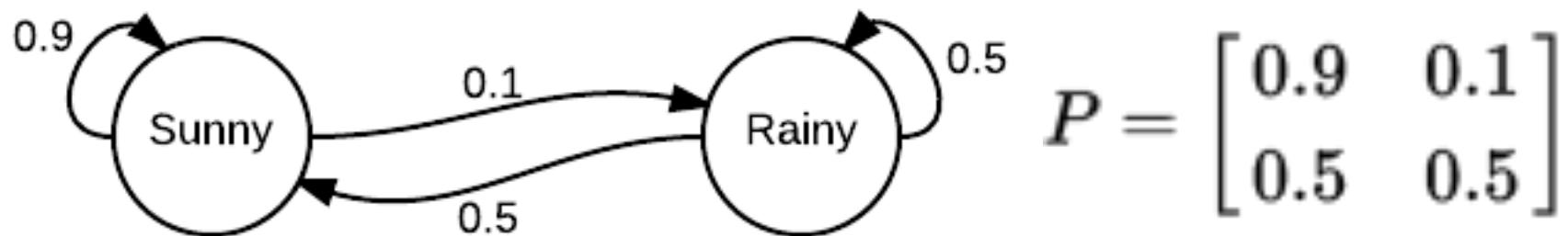
Andrei Markov  
1856-1922



# Discrete Time Markov Chains

- A **random** process.
- The system's state is a **discrete** variable.
- It undergoes transitions between states at uniformly-spaced (**discrete**) time points.
- Transition probabilities do not depend on the previous history of states (**memorylessness**).

# First example



Assume a deterministic initial condition:

$X_0 = \text{Sunny}$  with certainty; i.e.,

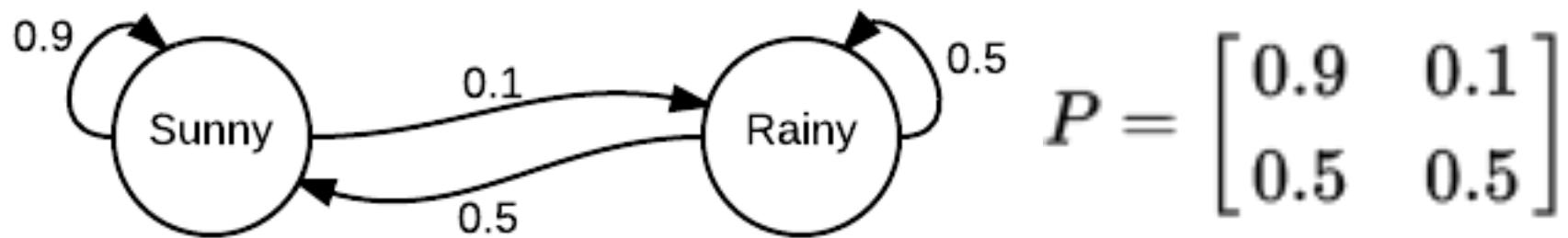
$$p(\text{Sunny} | t=0) = 1 \quad \text{and}$$

$$p(\text{Rainy} | t=0) = 0 \quad \text{i.e.}$$

$$s_0 = [1, 0]$$

...now, what is  $s_1$ ?

# First example



*What is  $s_1$ ?*

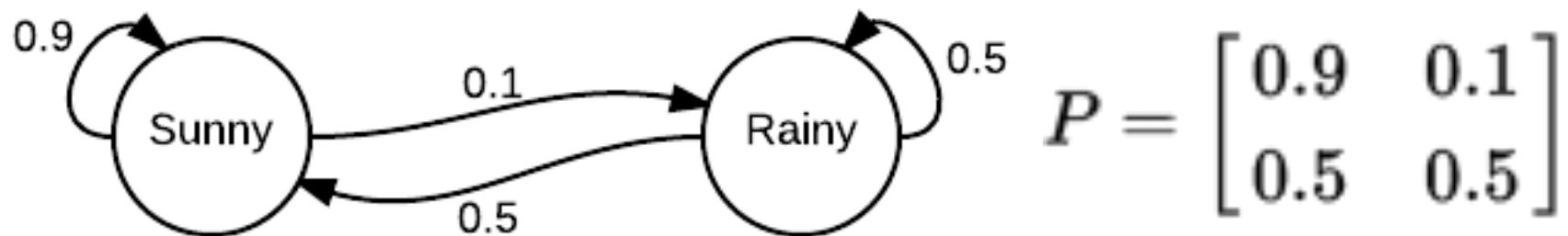
$$s_1 \quad \left\{ \begin{array}{l} p(\text{Sunny} \mid t=1) = 0.9 \\ p(\text{Rain} \mid t=1) = 0.1 \end{array} \right.$$

In matrix form...

$$s_1 = s_0 P$$

*...now, what is  $s_2$ ?*

# First example



*What is  $s_2$ ?*

$$s_2 \left[ \begin{array}{l} p(\text{Sunny} \mid t=2) = 0.9 p(\text{Sunny} \mid t=1) + 0.5 p(\text{Rainy} \mid t=1) = 0.86 \\ p(\text{Rainy} \mid t=2) = 0.1 p(\text{Sunny} \mid t=1) + 0.5 p(\text{Rainy} \mid t=1) = 0.14 \end{array} \right]$$

In matrix form...

$$s_2 = s_1 P$$

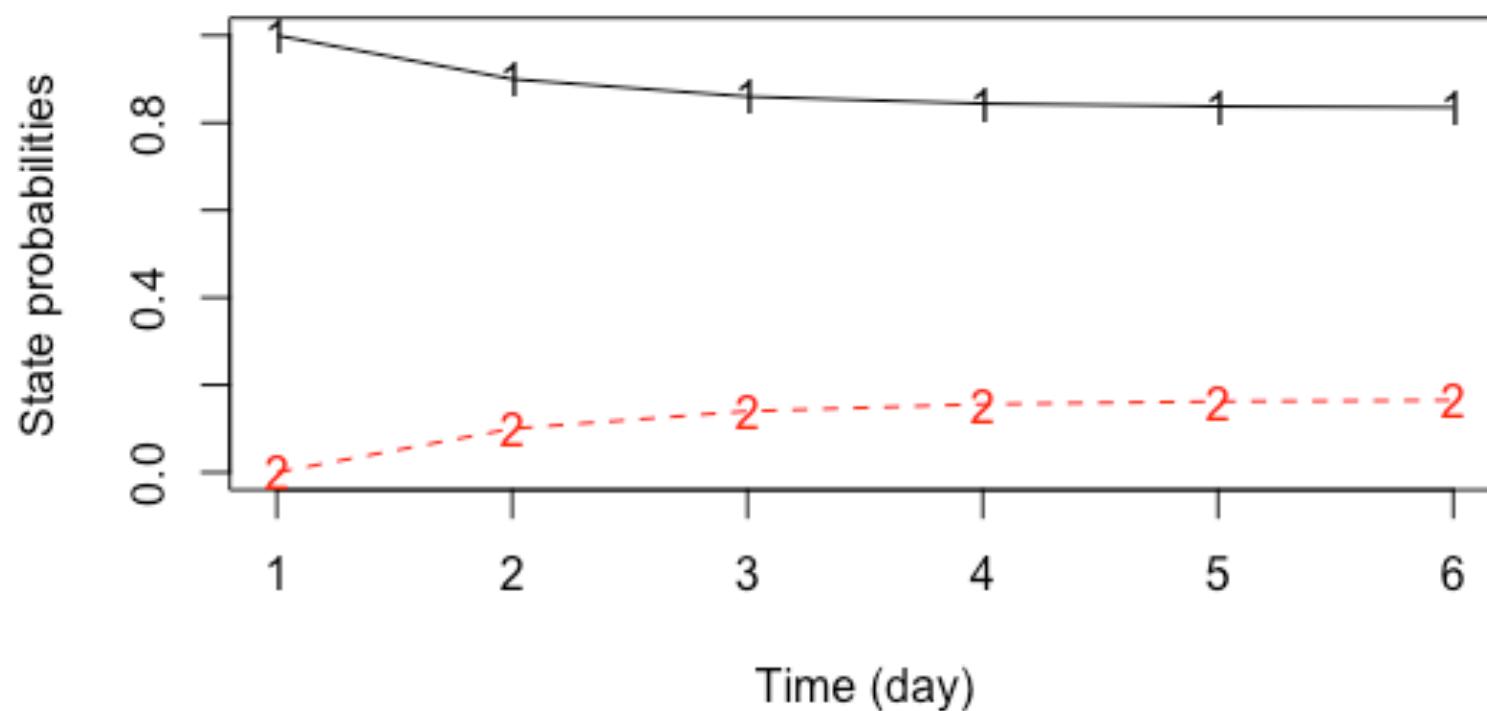
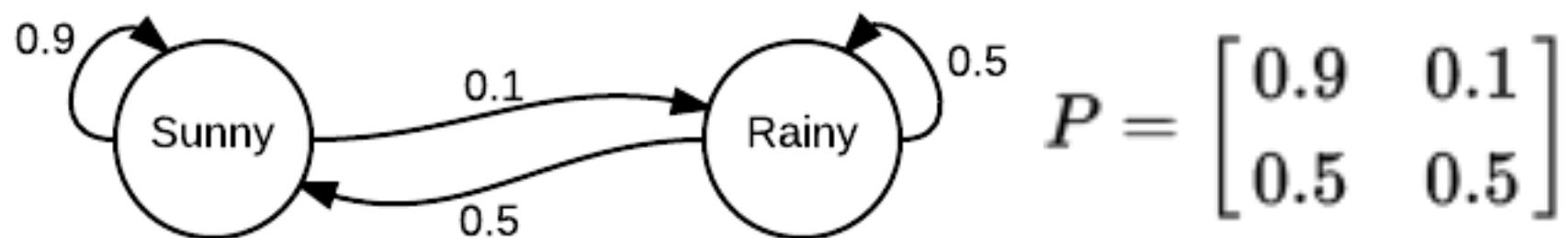
And in general...

$$s_{t+1} = s_t P$$

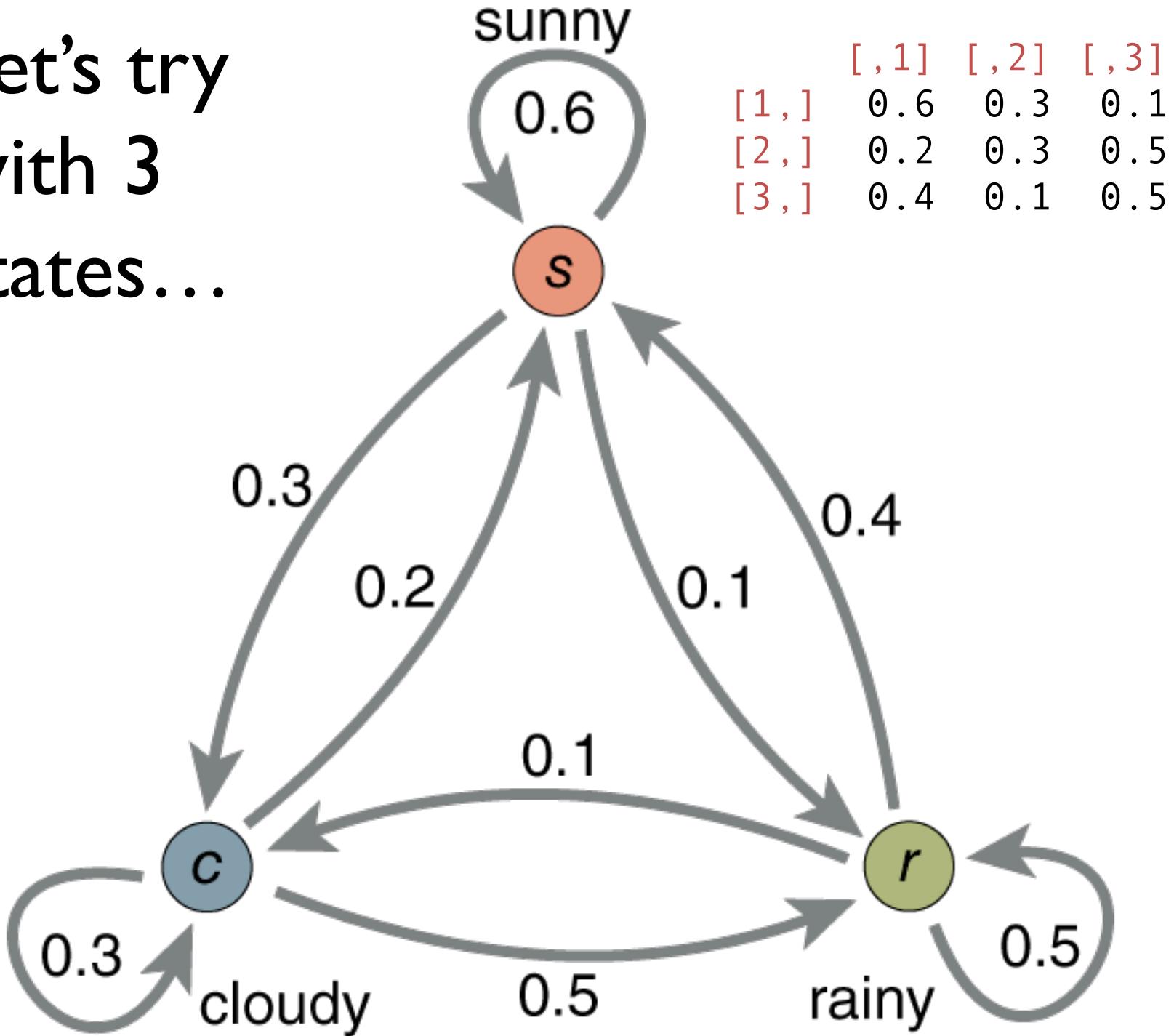
Meaning...

$$s_t = s_0 P^t$$

# Let's do a numerical test...



Let's try  
with 3  
states...

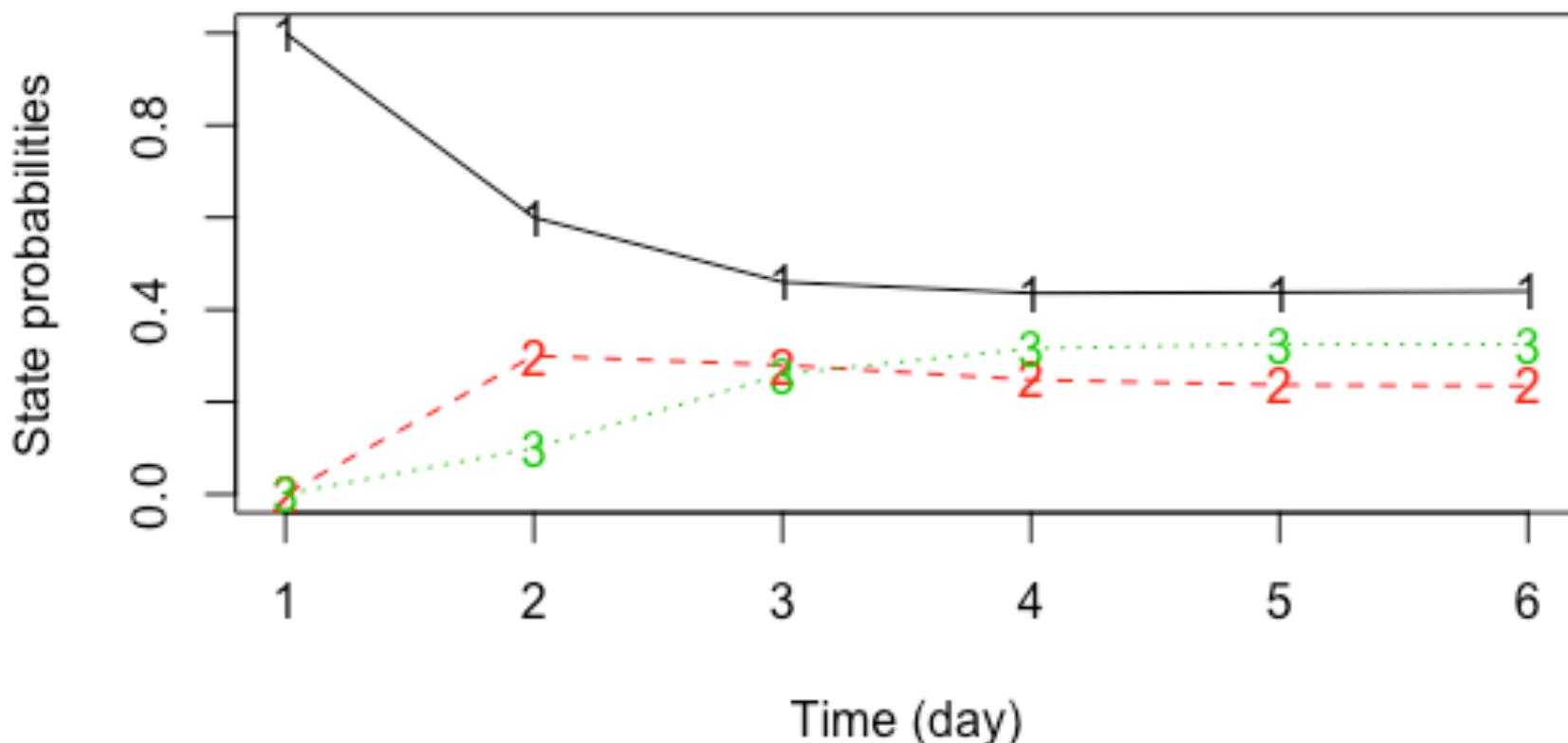


Starting from  
*Sunny*...



	[ , 1 ]	[ , 2 ]	[ , 3 ]
[ 1 , ]	0 . 6	0 . 3	0 . 1
[ 2 , ]	0 . 2	0 . 3	0 . 5
[ 3 , ]	0 . 4	0 . 1	0 . 5

Initial state:  $s_0 = [1,0,0]$

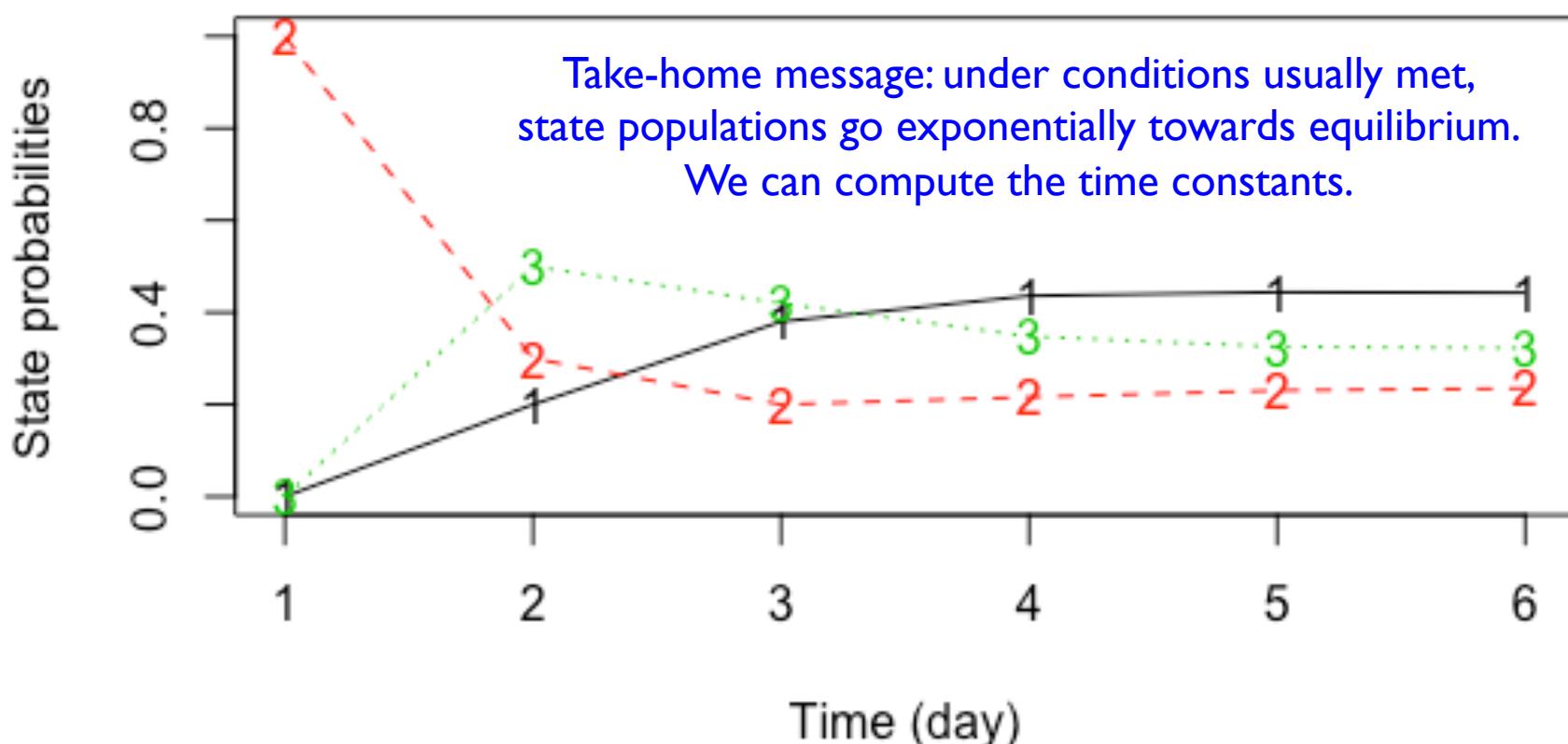


Starting from  
Cloudy...



	[ , 1 ]	[ , 2 ]	[ , 3 ]
[ 1 , ]	0 . 6	0 . 3	0 . 1
[ 2 , ]	0 . 2	0 . 3	0 . 5
[ 3 , ]	0 . 4	0 . 1	0 . 5

Initial state:  $s_0 = [0,1,0]$



# Important quantities we can compute

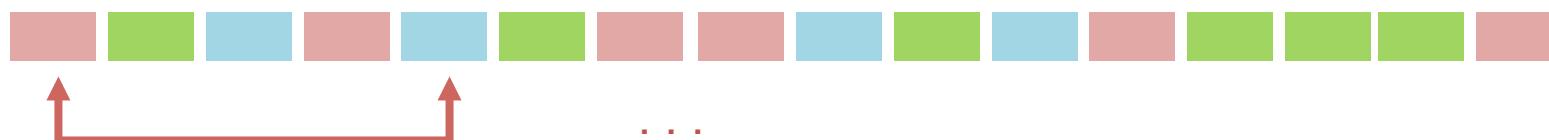
- Stationary distribution ( $\rightarrow$  eq. probabilities)
- Relaxation times
- Mean-first passage times ( $\rightarrow$  kinetic rates)
- And others we won't discuss:
  - Commitor probabilities
  - Fluxes
  - ...

**Learning matrices  
from trajectories  
(of discrete states)**

# Markov models

Total sampled time (e.g. 100  $\mu\text{s}$ )

Define the discrete state of the system in discrete time (e.g. via reaction coordinates)



Compute the transition probability matrix  
sliding a window of lag time  $\tau$

		To		
		Red	Green	Blue
From	Red	0	0	0
	Green	0	0	0
	Blue	0	0	0

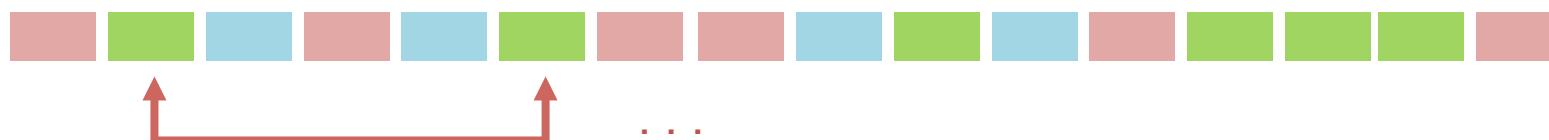
From

		To		
		Red	Green	Blue
From	Red	0	0	1
	Green	0	0	0
	Blue	0	0	0

# Markov models

Total sampled time (e.g. 100  $\mu\text{s}$ )

Define the discrete state of the system in discrete time (e.g. via reaction coordinates)



Compute the transition probability matrix  
sliding a window of lag time  $\tau$

		To		
		Red	Green	Blue
From	Red	0	0	1
	Green	0	0	0
	Blue	0	0	0

		To		
		Red	Green	Blue
From	Red	0	0	1
	Green	0	1	0
	Blue	0	0	0

# Markov models

Total sampled time (e.g. 100  $\mu\text{s}$ )



Define the discrete state of the system in discrete time (e.g. via reaction coordinates)



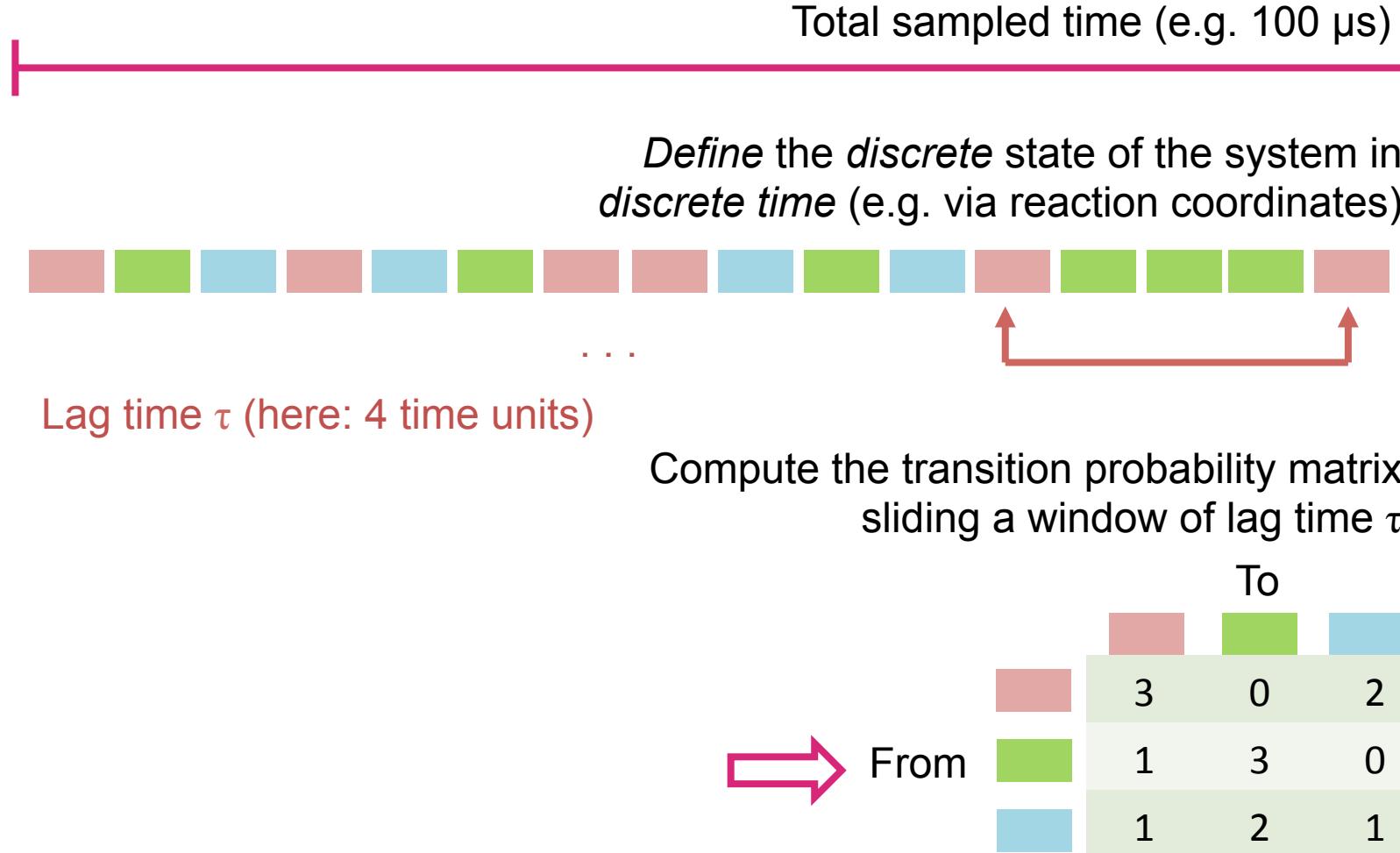
Compute the transition probability matrix  
sliding a window of lag time  $\tau$

		To		
		Red	Green	Blue
		0	0	1
From	Red	0	1	0
	Green	0	0	0
	Blue	0	0	0



		To		
		Red	Green	Blue
		0	0	1
From	Red	0	1	0
	Green	0	0	0
	Blue	1	0	0

# Markov models



Repeat until the end of the trajectory.

Note the Markovian assumption:  
transition probabilities **do not** depend  
on history (neither, for us, on time).

(Note how “Early” and “late” events  
are squashed in the same matrix.)

*Important:* you can accumulate counts  
from **different** trajectories in the **same** matrix!

It's all valid, independent sampling  
of the same underlying system.

Example: multiple weather stations in the same city.  
They may be active at different times!

**Transition counts**

		To	
		From	
		From	
		3	0
		1	3
		1	2
		2	

**Transition probabilities**

To			$\Sigma_j$
From			
3/5	0	2/5	1
$\frac{1}{4}$	$\frac{3}{4}$	0	1
$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$	1

Normalize  
by rows

$$P_{ij}$$

**Probability vector**

1	0	0
---	---	---

$$\times \begin{bmatrix} 3/5 & 0 & 2/5 \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & \frac{2}{4} & \frac{1}{4} \end{bmatrix}$$

$$s_i$$

$$P_{ij}$$

**Evolved (after  $\tau$ ) state**

3/5	0	2/5
-----	---	-----

$$s'_j$$

**Probability vector**

1	0	0
---	---	---

$s_i$

$$\times \begin{bmatrix} \frac{3}{5} & 0 & \frac{2}{5} \\ \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{4} & \frac{2}{4} & \frac{1}{4} \end{bmatrix} = P_{ij}$$

**Evolved state**

$\frac{3}{5}$	0	$\frac{2}{5}$
---------------	---	---------------

$s'_j$

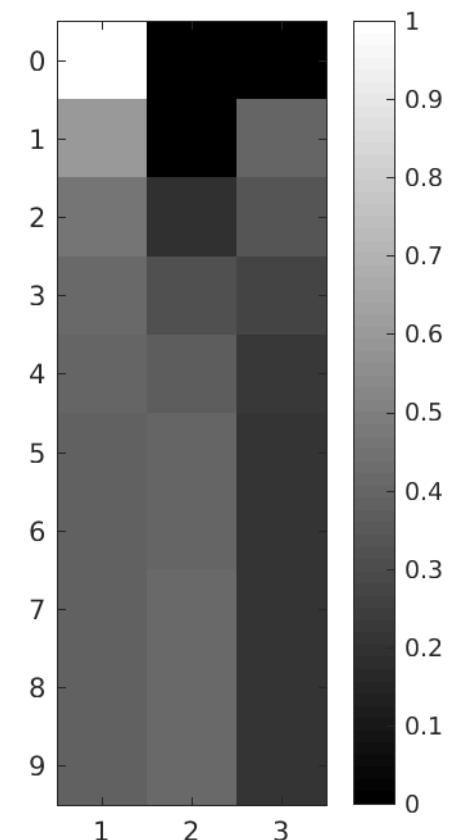
$$s' = sP$$

$$s'' = (sP)P = sP^2$$

$$s^{(n)} = sP^n$$

1	0	0
.60	0	.40
.46	.20	.34
.41	.32	.27
.39	.37	.23
.39	.40	.22
...	...	...

time  
↓



### Probability vector

$$\begin{matrix} 1 & 0 & 0 \end{matrix}$$

$s_i$

x

3/5	0	2/5
1/4	3/4	0
1/4	2/4	1/4

$P_{ij}$

### Evolved state (after tau)

$$\begin{matrix} 3/5 & 0 & 2/5 \end{matrix}$$

$s'_j$

$$s' = sP$$

$$s'' = (sP)P = sP^2$$

$$s^{(n)} = sP^n$$

...

$$s^\infty = ?$$

$$s^\infty P = s^\infty$$

Left eigenvector of P (eigenvalue 1)

Is the stationary state  
= equilibrium probabilities  
= the free energy surface

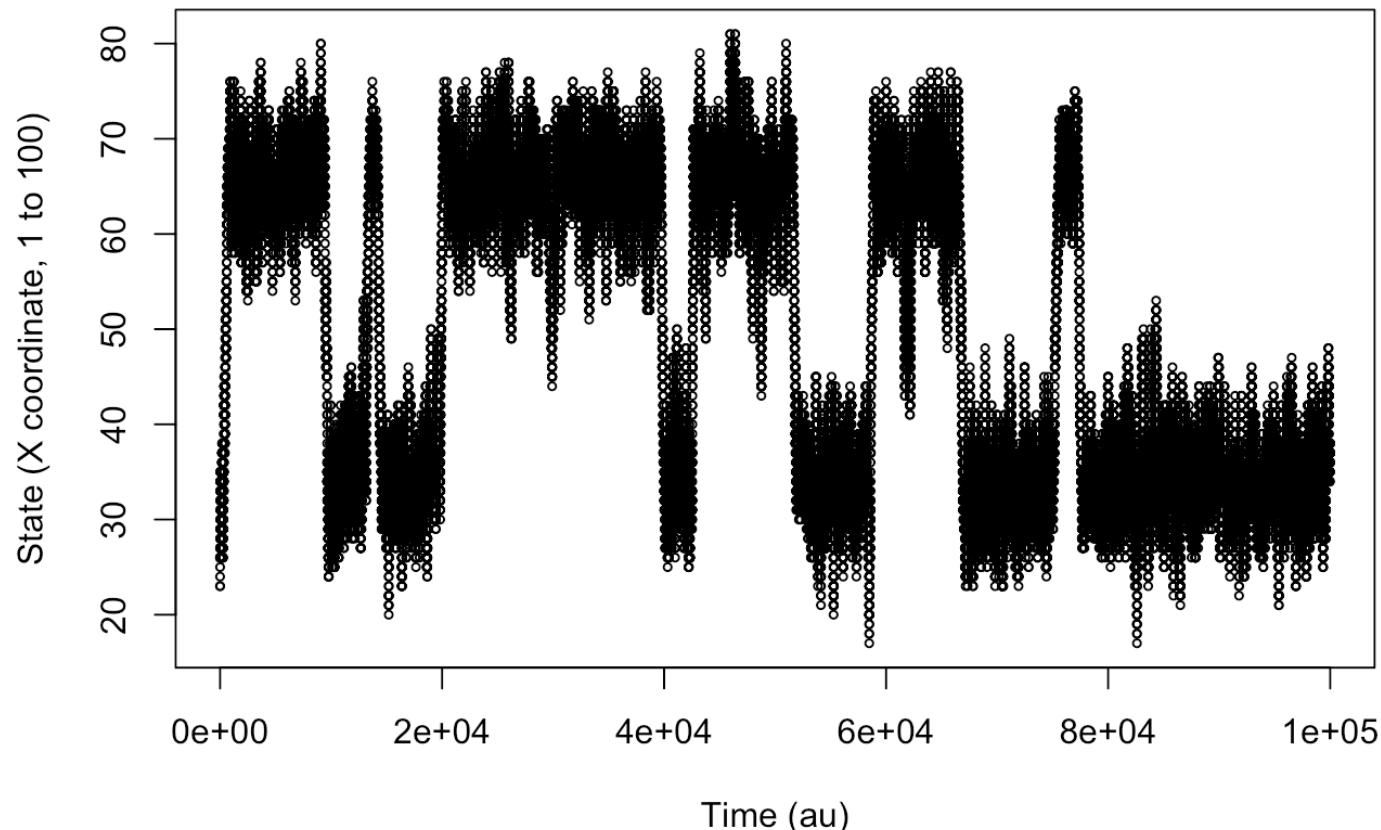
```
[a,b]=eig(P')
a(:,3)/sum(a(:,3))
= 0.385 0.410 0.205 = [5/13 16/39 8/39]
```

# **Markov modeling a 1D trajectory**

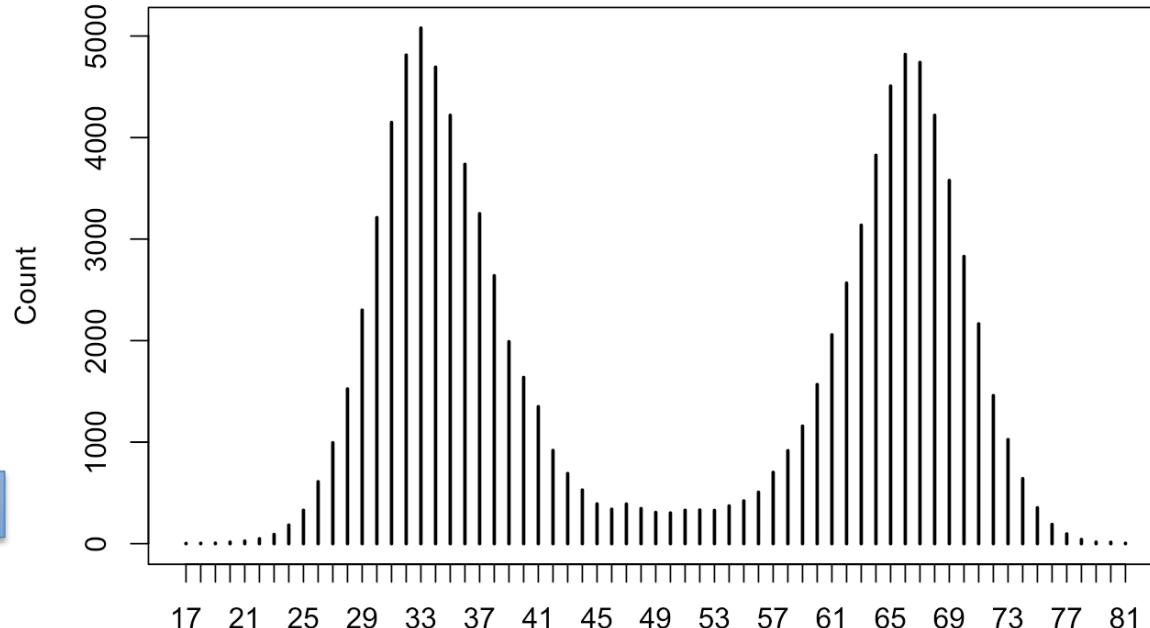
(Please find the extended version online,  
“Markov state models of a 1D trajectory”,  
with R code)

# Start with a 1-D trajectory

- Already discretized in 100 bins

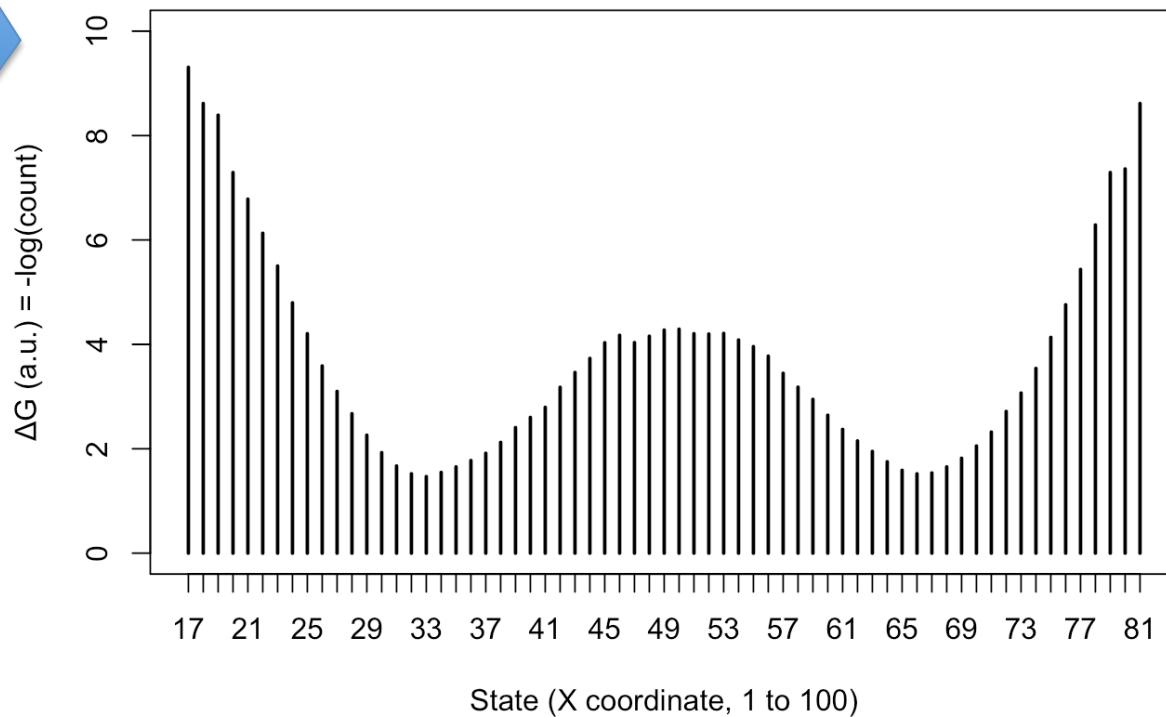


## Histogram



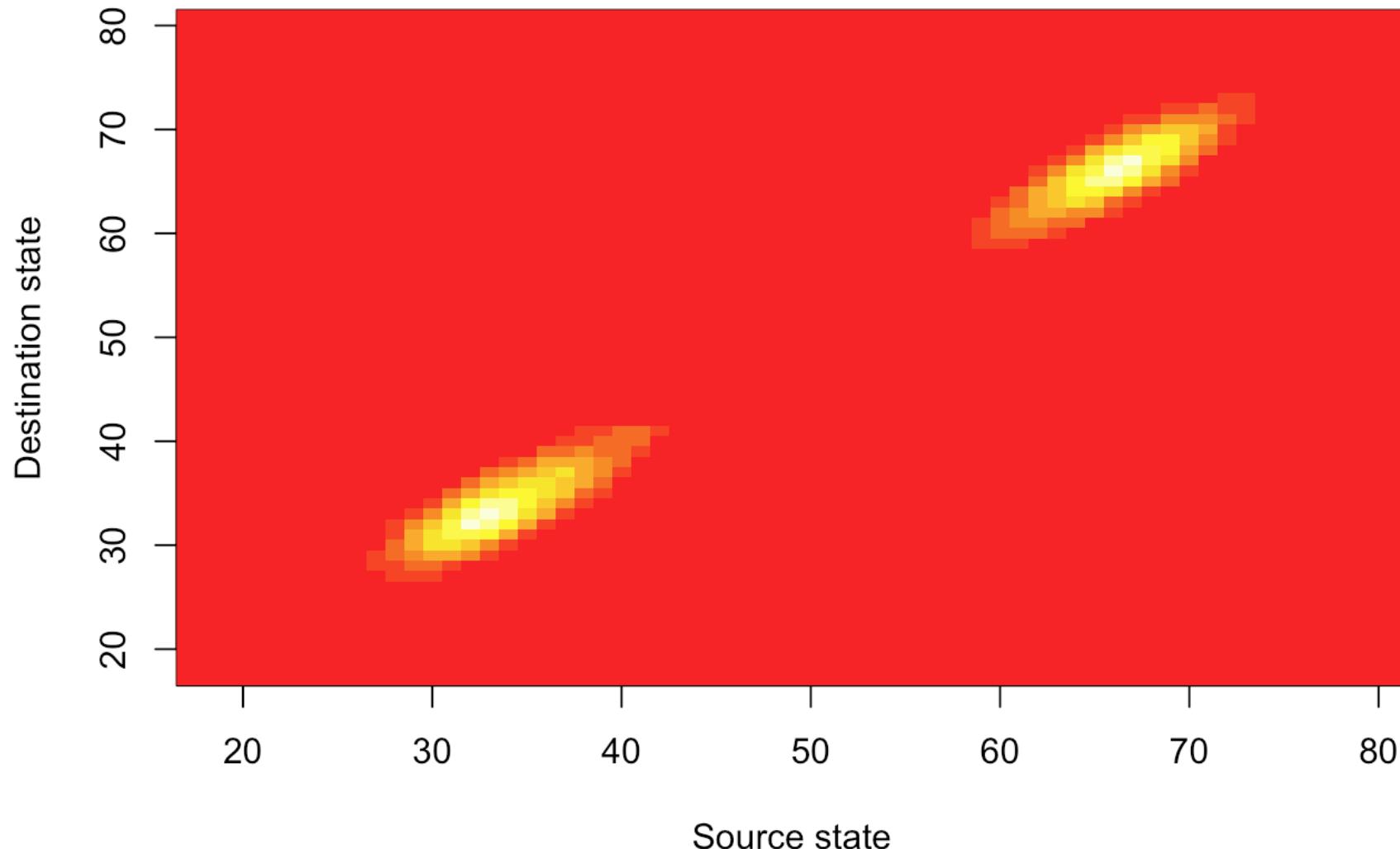
## Boltzmann inversion

-log(count)



## The transition count matrix\*

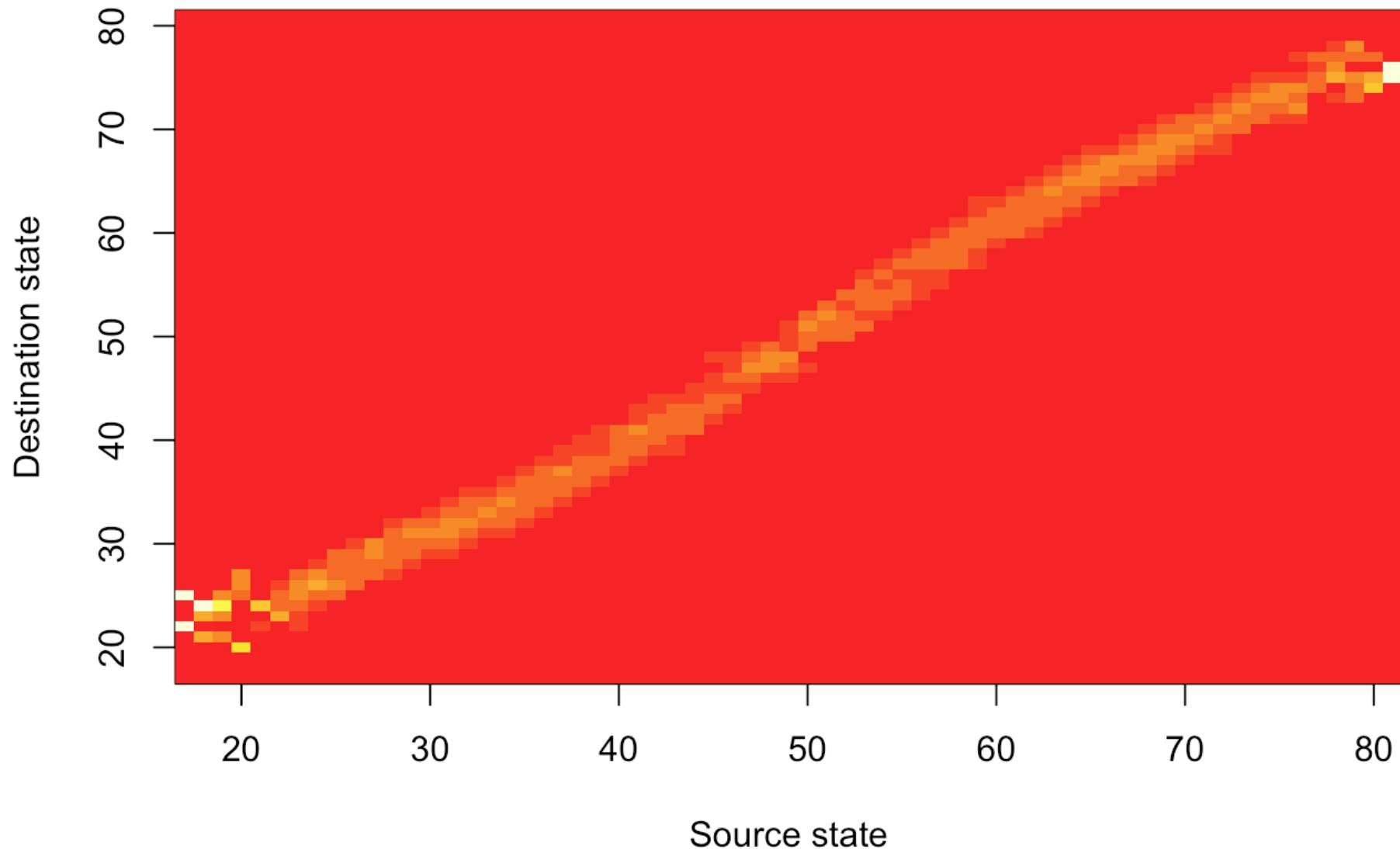
*How many times we have seen state  $i$  going to  $j$  after  $\tau = 10$  time units*



\* shown as an image for compactness

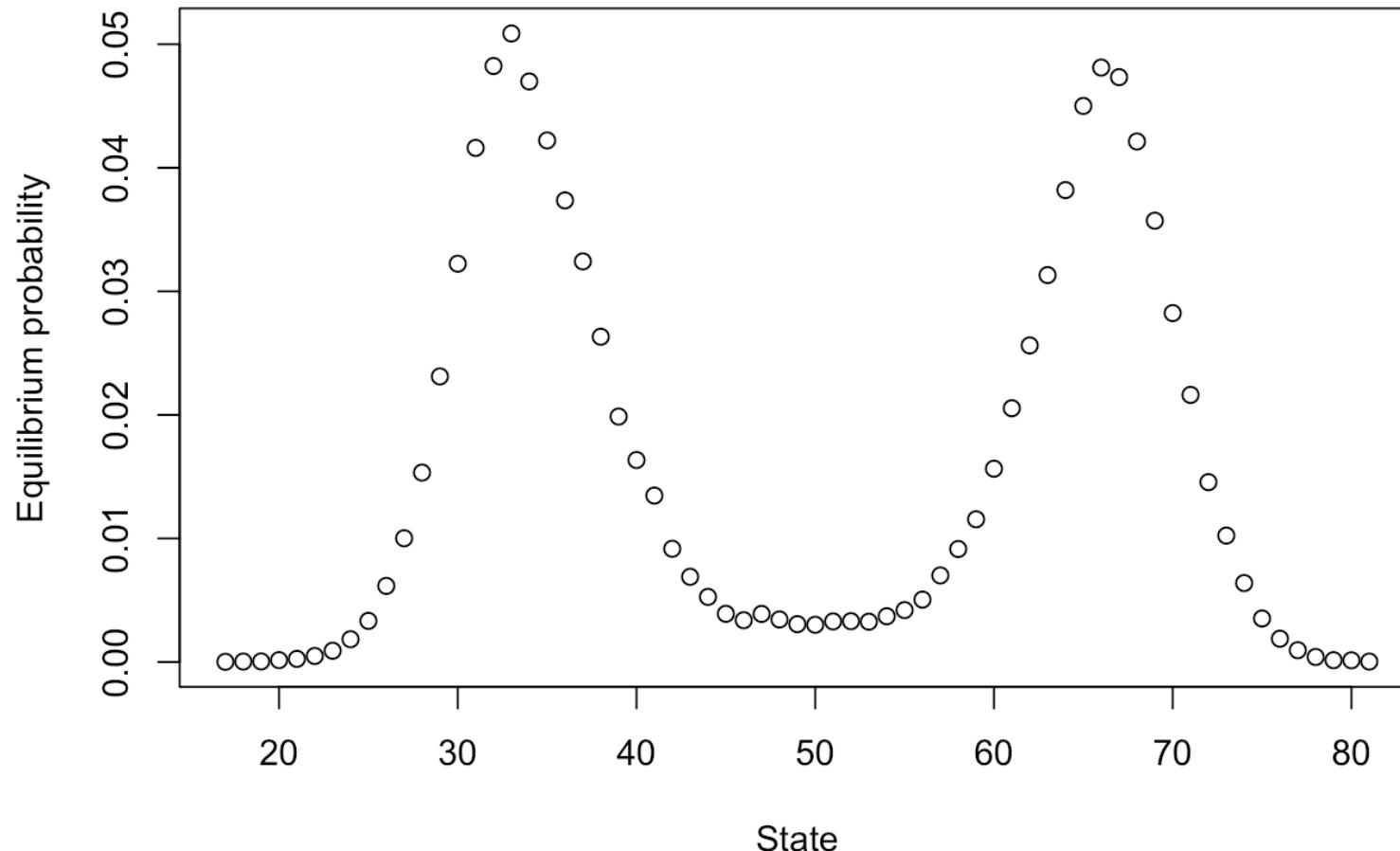
# The transition *probability* matrix

Rows normalized to sum to 1



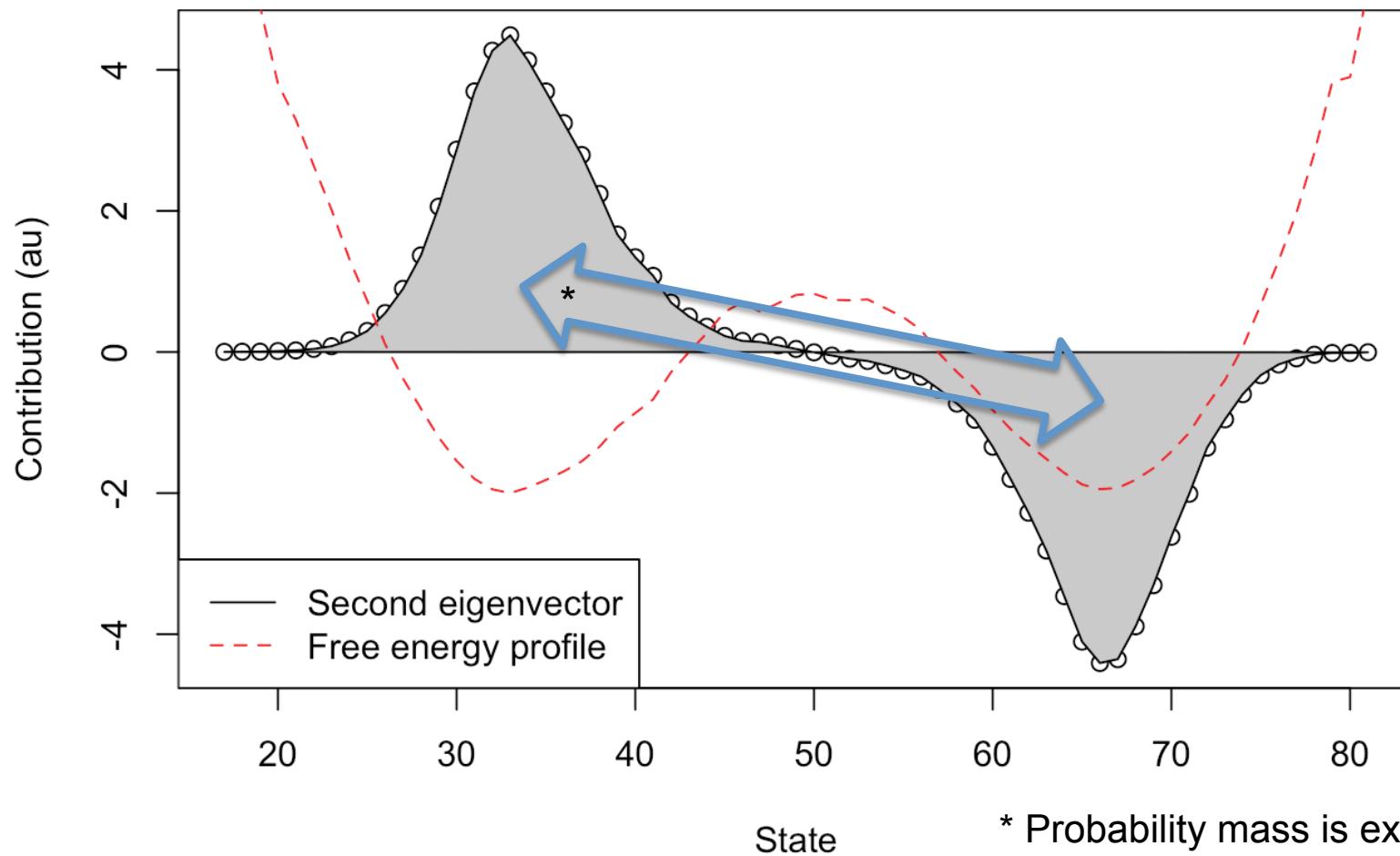
# First eigenvector ( $\mu_1 = 1$ )

This is the stationary state (normalize so it sums to 1)



# Second eigenvector ( $\mu_2=0.997$ )

This is the slowest relaxation mode: ITS  $\tau_2 = 3610$  time units

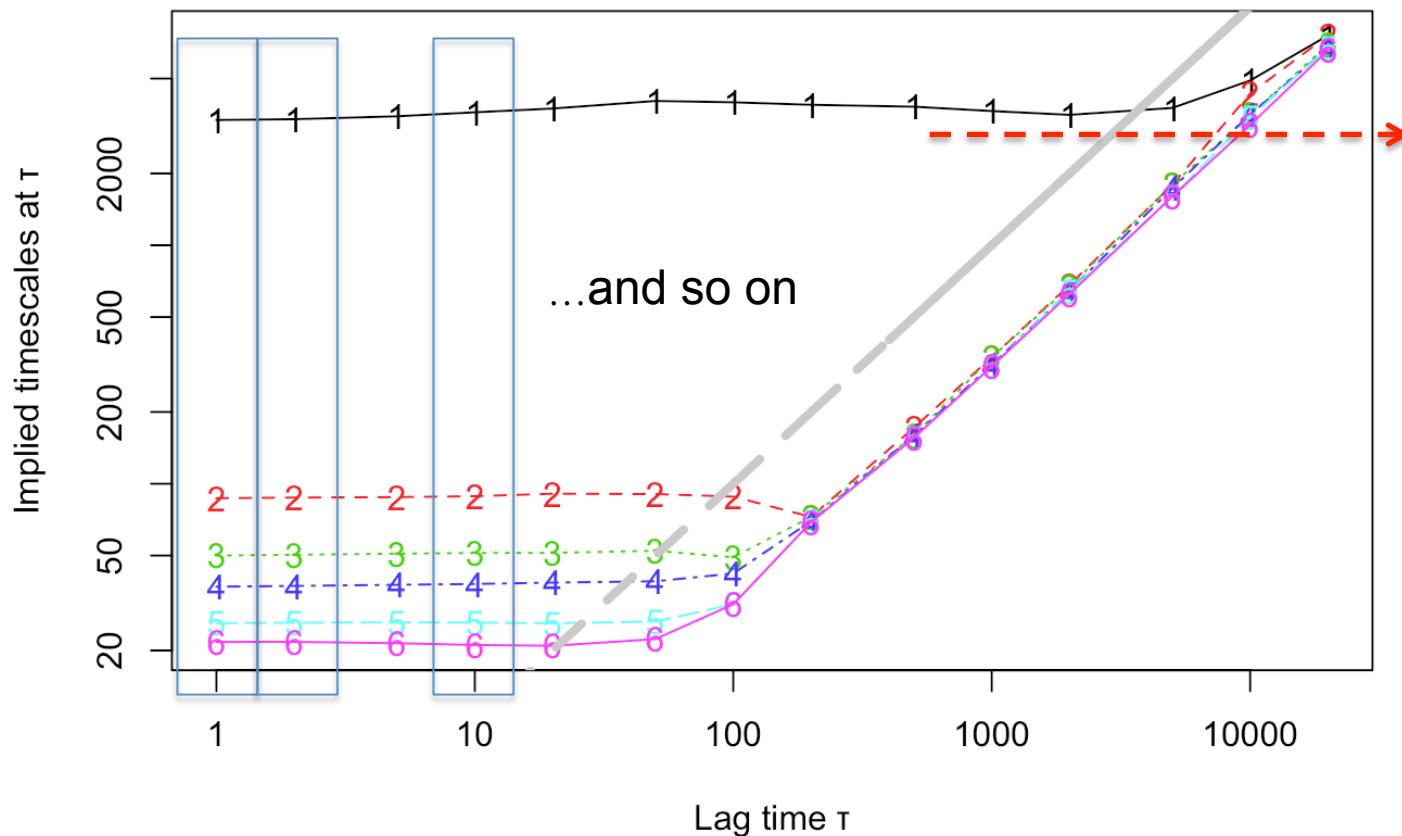


# Markovianity

- The state transition probabilities only depend on the current state.
- Examples
  - Today's weather, not yesterday's
  - Where the ligand is, not how did it got there
- The property may be false at short timescales but true at longer ones (system's memory)
- It does depend on the chosen states

# Implied timescales plot

Repeat the eigenvalues determination for several lags. Check convergence.

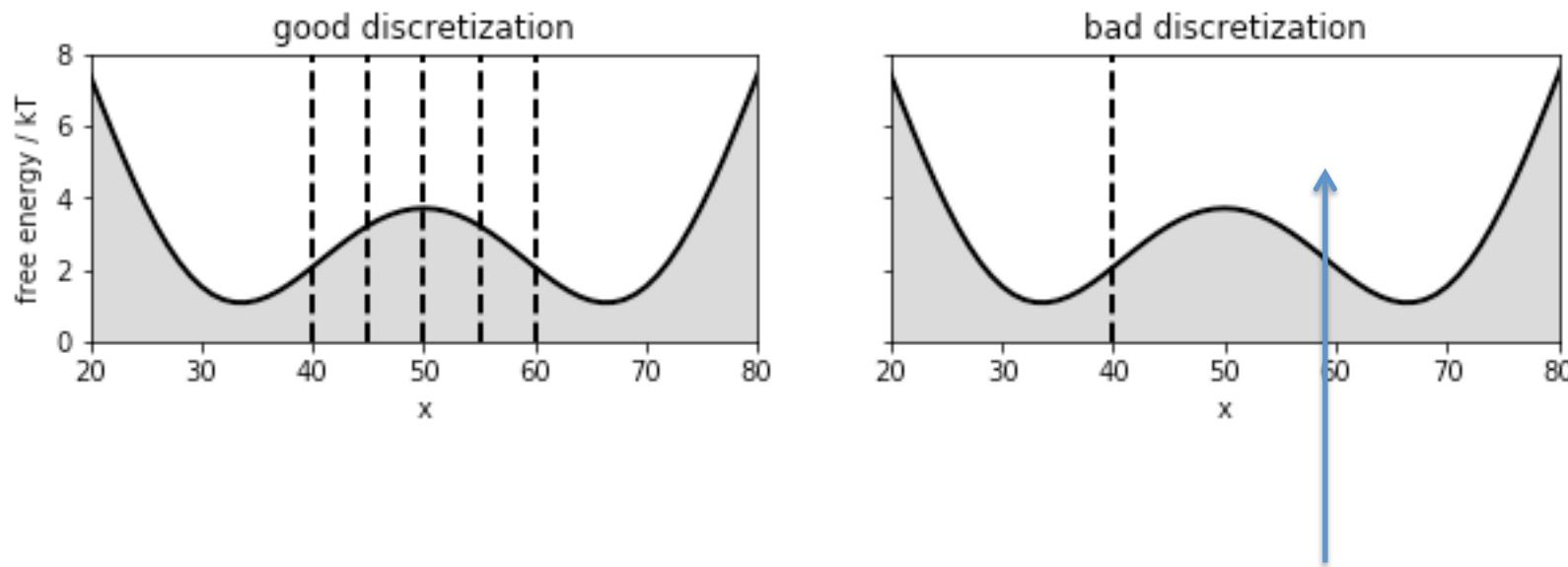


Here,  
convergence is  
achieved very  
early.

- Reasons:
- (a) true two-state dynamics;
  - (b) absence of orthogonal degrees of freedom;
  - (c) fine space discretization

# Macrostates

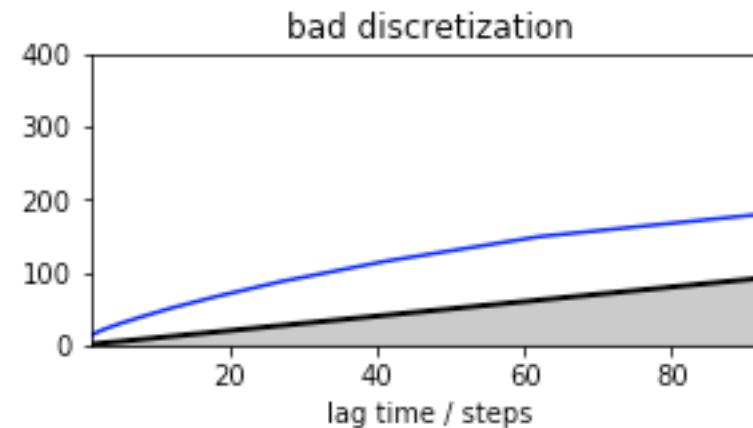
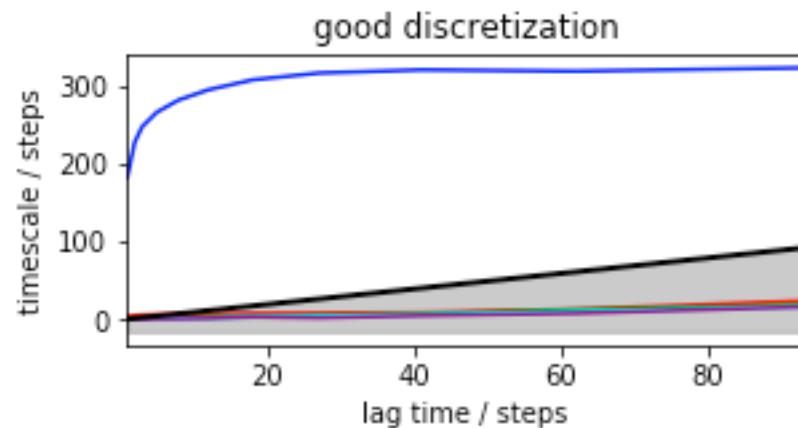
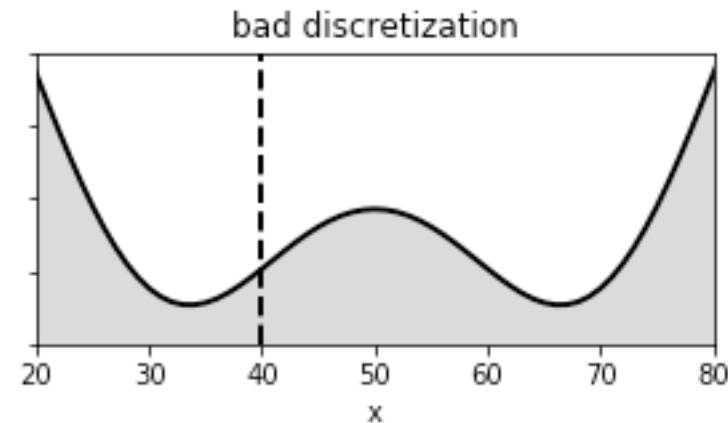
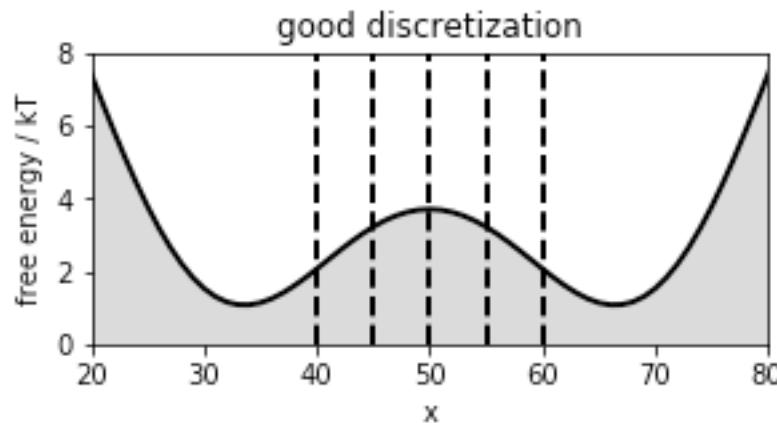
A bad choice of the discretization breaks the Markovianity assumption



In the “bad discretization” case, the barrier is embedded in one of the states. This generates a “long term memory” effect: the rightmost state could actually be short-lived (if we are on the left of the barrier) or long-lived (if we are on its right). These two cases are convoluted into the same, so that the present state information itself is not sufficient to predict the “future” of the system any more.

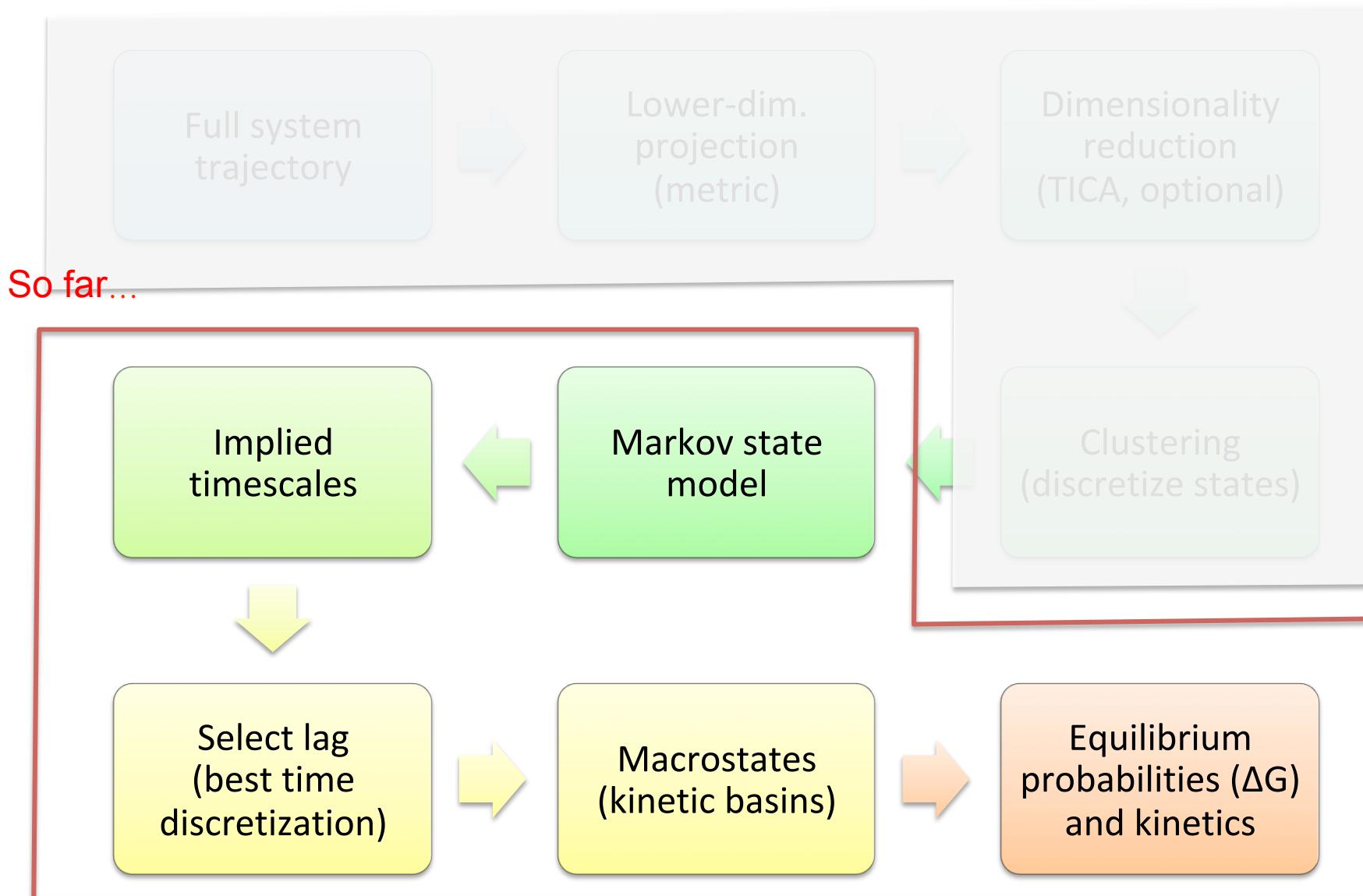
# Macrostates

A bad choice of the discretization breaks the  
Markovianity assumption



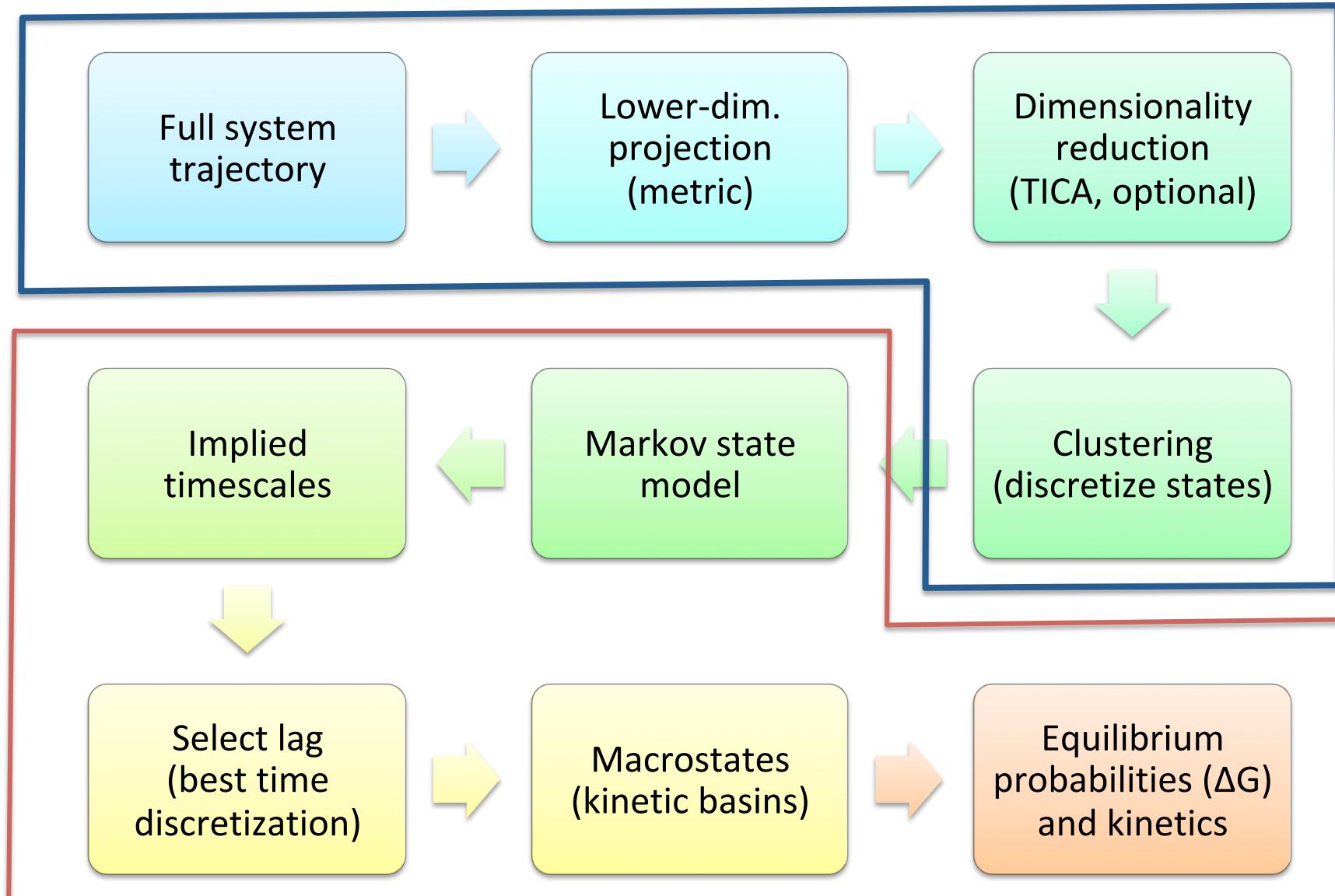
**Back to molecules:  
>> 1 dimension, continuous:  
clustering**

# Dealing with trajectories...



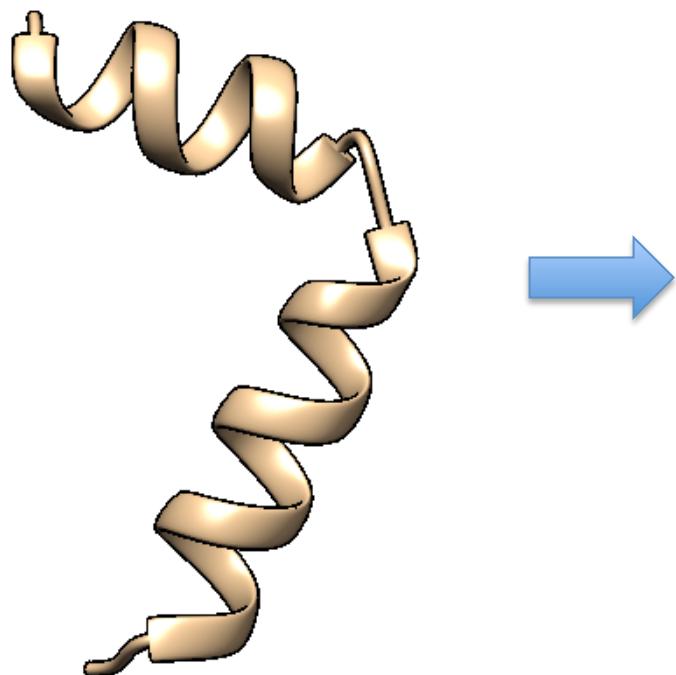
# Dealing with trajectories...

This is what we need to add



# Metric projection

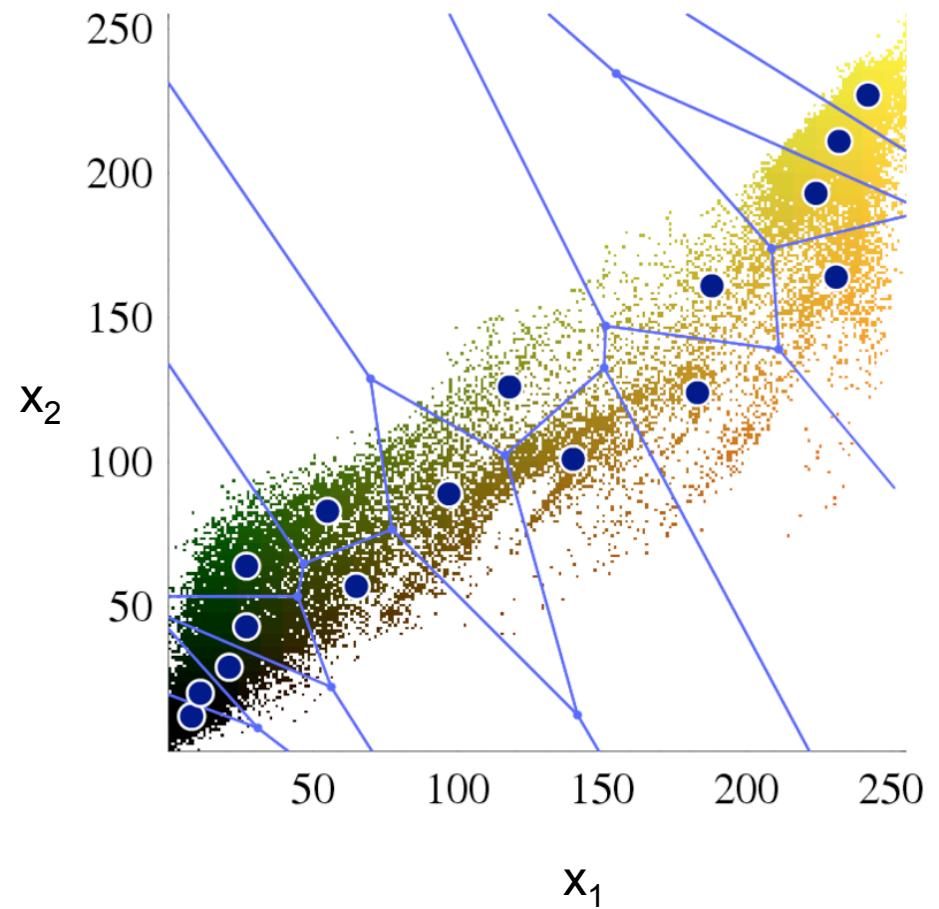
The first step is to project the system state in a lower-dimensional space (“metric”). Many choices are available, e.g.



- Manually chosen distances
- Atom coordinates
- $N$  phi/psi Ramachandran angles
- Distance matrix
- Contact matrix
- ...

# Clustering (1<sup>st</sup> level)

- Now move to a discrete state space
- Reduction from the low-dimensional space is done by *clustering*
  - Usually called “microstates”
- Several algorithms are implemented in MSM packages (e.g. grid;  $k$ -means; etc. We won’t discuss them)



# **Conclusions and Resources**

# Conclusions

- MSM methods make efficient use of *unbiased* sampling
- Cluster →  $P_{ij}$  → eigenvectors → free energies ( $K_D$ ) and kinetics ( $k_{on}/k_{off}$ )
- Still require *huge* (but not unreachable) amounts of sampling/simulation for biologically interesting systems
- Strong mathematical foundation, with good software available

# Software + Tutorials

- All are Python-based
  - They include clear walkthroughs (highly recommended) with datasets
- **HTMD** – [www.html.org](http://www.html.org)
  - Also analysis + system build + adaptive ...
  - Can aggregate large-scale datasets
- PyEMMA – [www.emma-project.org](http://www.emma-project.org)
- MSMBuilder - [msmbuilder.org](http://msmbuilder.org)

# Related topics

If you liked this, you may also like...

- Estimation of reversible transition matrices
- Hidden Markov Models doi:[10.1063/1.4828816](https://doi.org/10.1063/1.4828816)
- TICA
- Conformational fluxes
- Adaptive sampling
- Role of drug residence time in drug discovery

# Literature

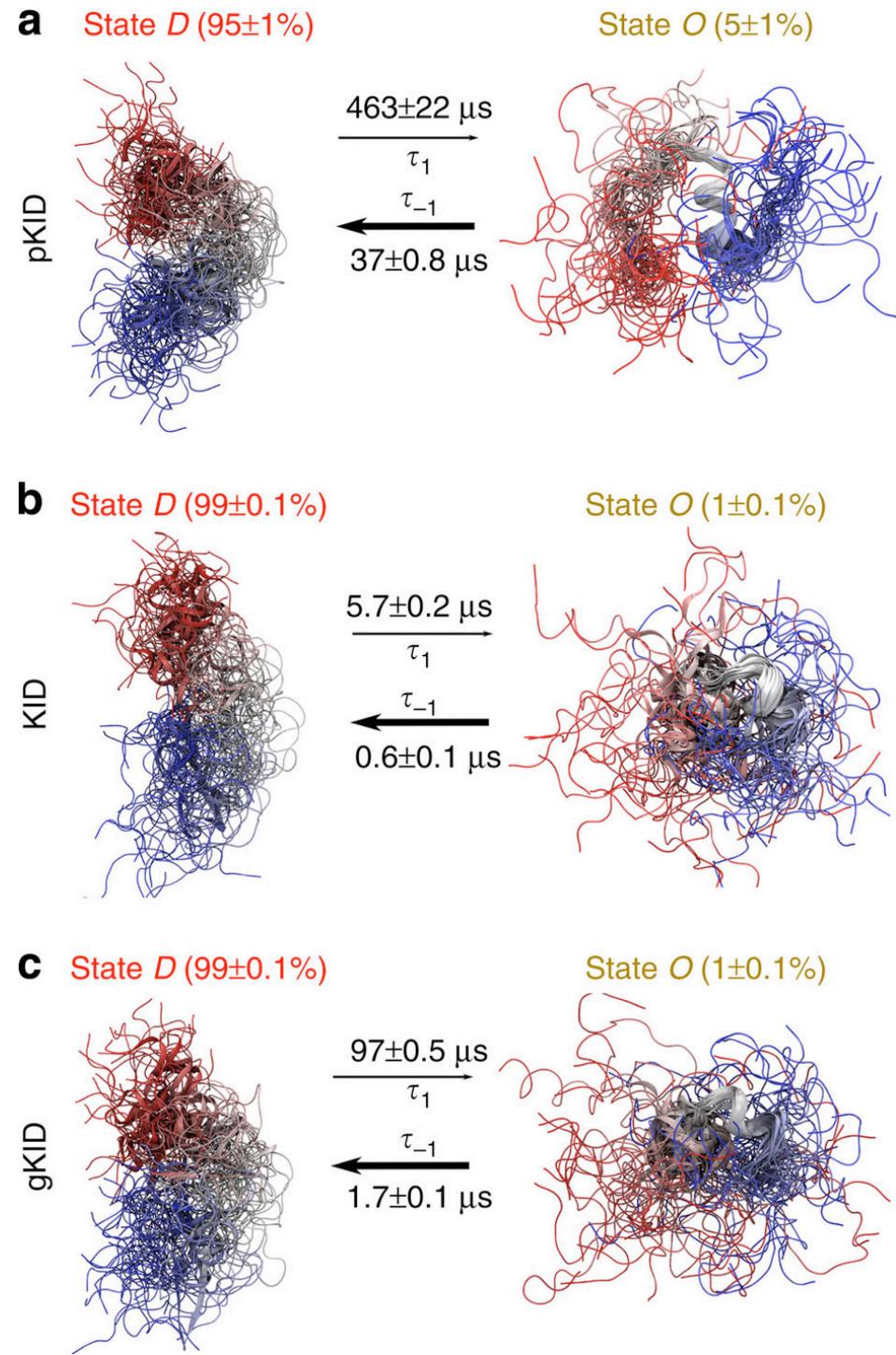
- Buch et al., 10.1073/pnas.1103547108
  - Rigid ligand + protein association, simplest case
- Swinney, PMID:19152211
  - Importance of kinetics in drug design
- Chodera and Noe, 10.1016/j.sbi.2014.04.002
  - Excellent overview (1)
- Voelz et al., 10.1021/ja9090353
  - Full reconstruction of a millisecond folding
- Pande et al., 10.1016/j.ymeth.2010.06.002
  - Excellent overview (2)
- Paul et al., 10.1038/s41467-017-01163-6
  - Peptide-peptide association
- Doerr et al., 10.1021/ct400919u
  - Adaptive sampling
- Bryan et al., “The \$25 B eigenvector”
  - The original PageRank algorithm; e.g. [jdc.math.uwo.ca/M1600b/I/pagerank-1600.pdf](http://jdc.math.uwo.ca/M1600b/I/pagerank-1600.pdf)

# **Examples from the literature**

# *Kinetic modulation of a disordered protein domain by phosphorylation*

N. Stanley, S. Esteban and G.  
De Fabritiis, Nat. Commun. 5,  
5272 (2014)

doi:10.1038/ncomms6272



# Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations

Ignasi Buch, Toni Giorgino, and Gianni De Fabritiis<sup>1</sup>

Computational Biochemistry and Biophysics Laboratory, Universitat Pompeu Fabra, Barcelona Biomedical Research Park, C/Doctor Aiguader 88, 08003 Barcelona, Spain

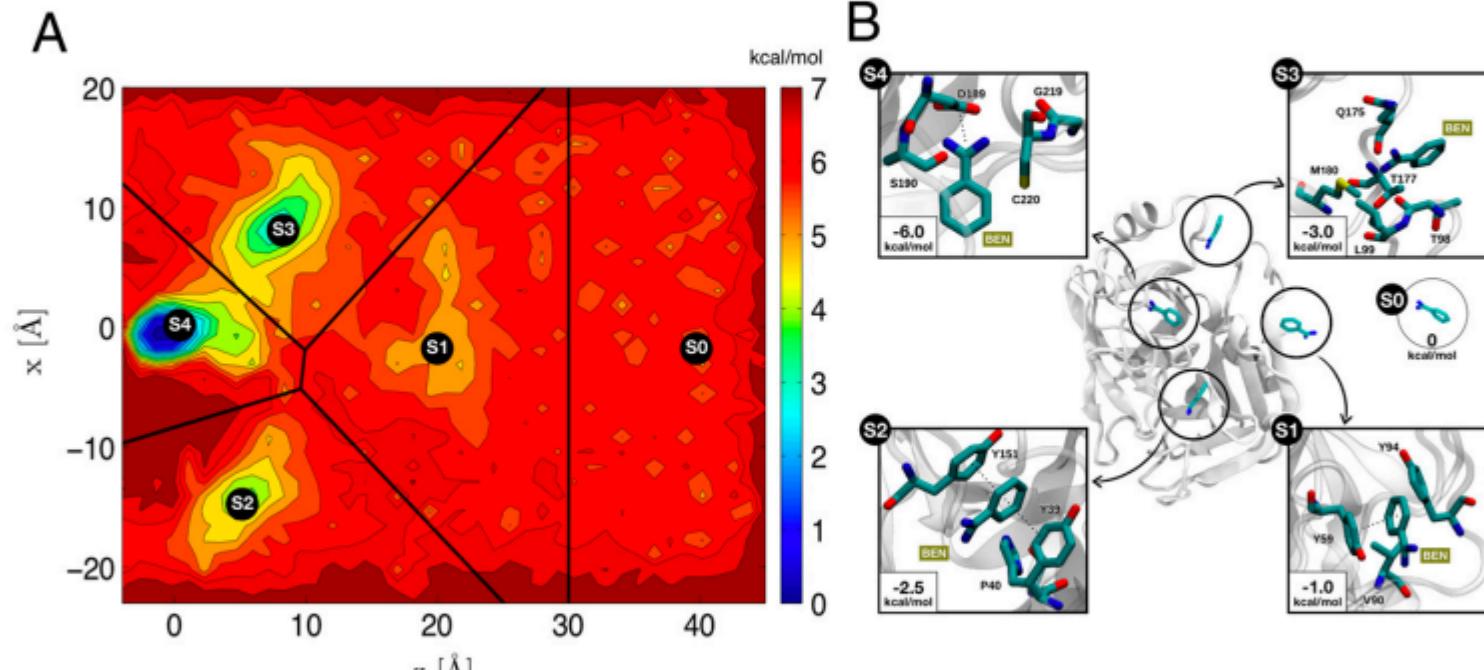
Edited by Arieh Warshel, University of Southern California, Los Angeles, CA, and approved May 11, 2011 (received for review March 4, 2011)

The understanding of protein-ligand binding is of critical importance for biomedical research, yet the process itself has been very difficult to study because of its intrinsically dynamic character. Here, we have been able to quantitatively reconstruct the complete binding process of the enzyme-inhibitor complex trypsin-benzamidine by performing 495 molecular dynamics simulations of the

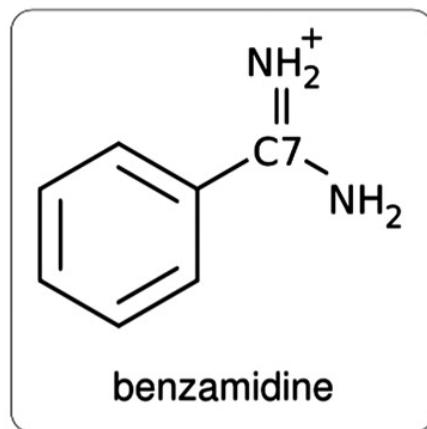
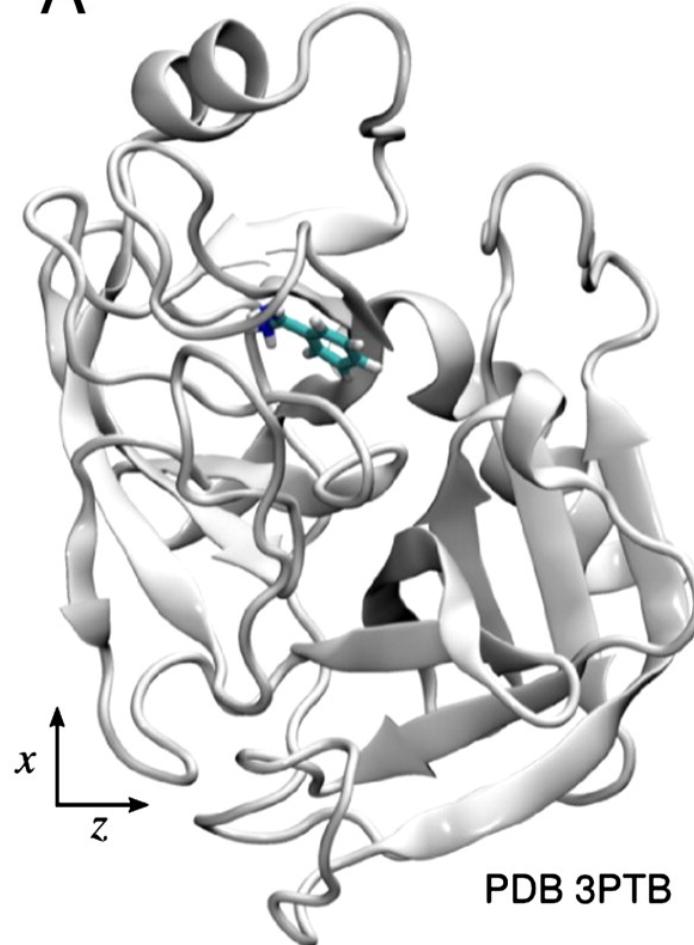
reproduce with atomic resolution the crystallographic mode of binding, but we also provide the kinetically and energetically meaningful transition states of the process.

Free ligand binding has been used in the past to describe computational experiments in which, typically, a ligand is placed at a certain distance from the target protein and first by diffusion and

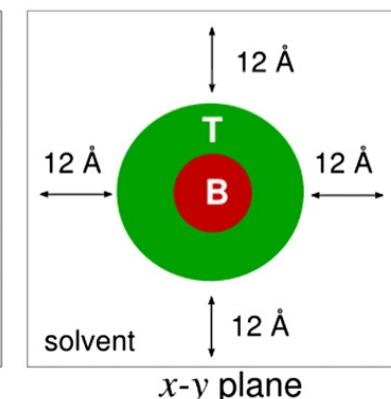
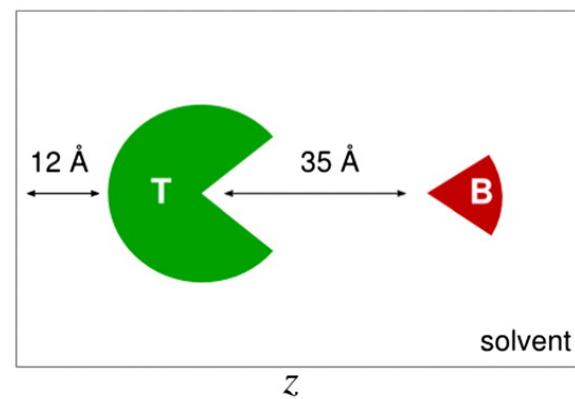
in the pro-

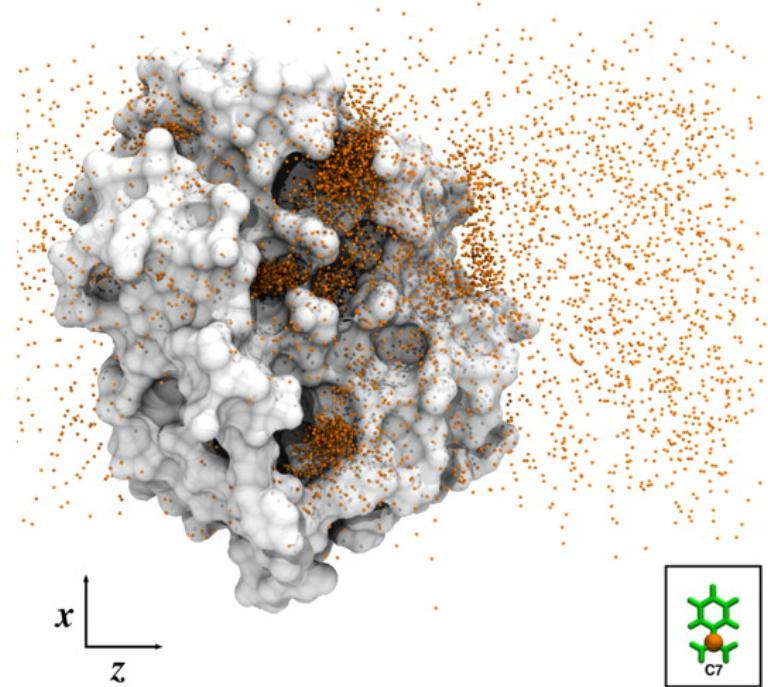
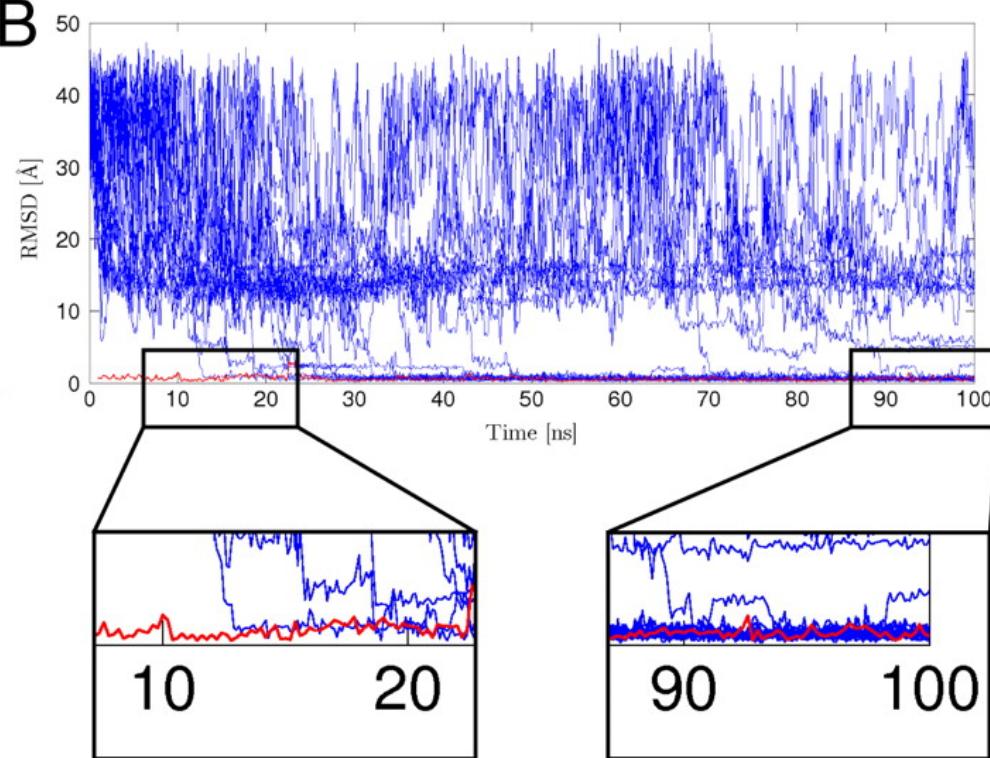


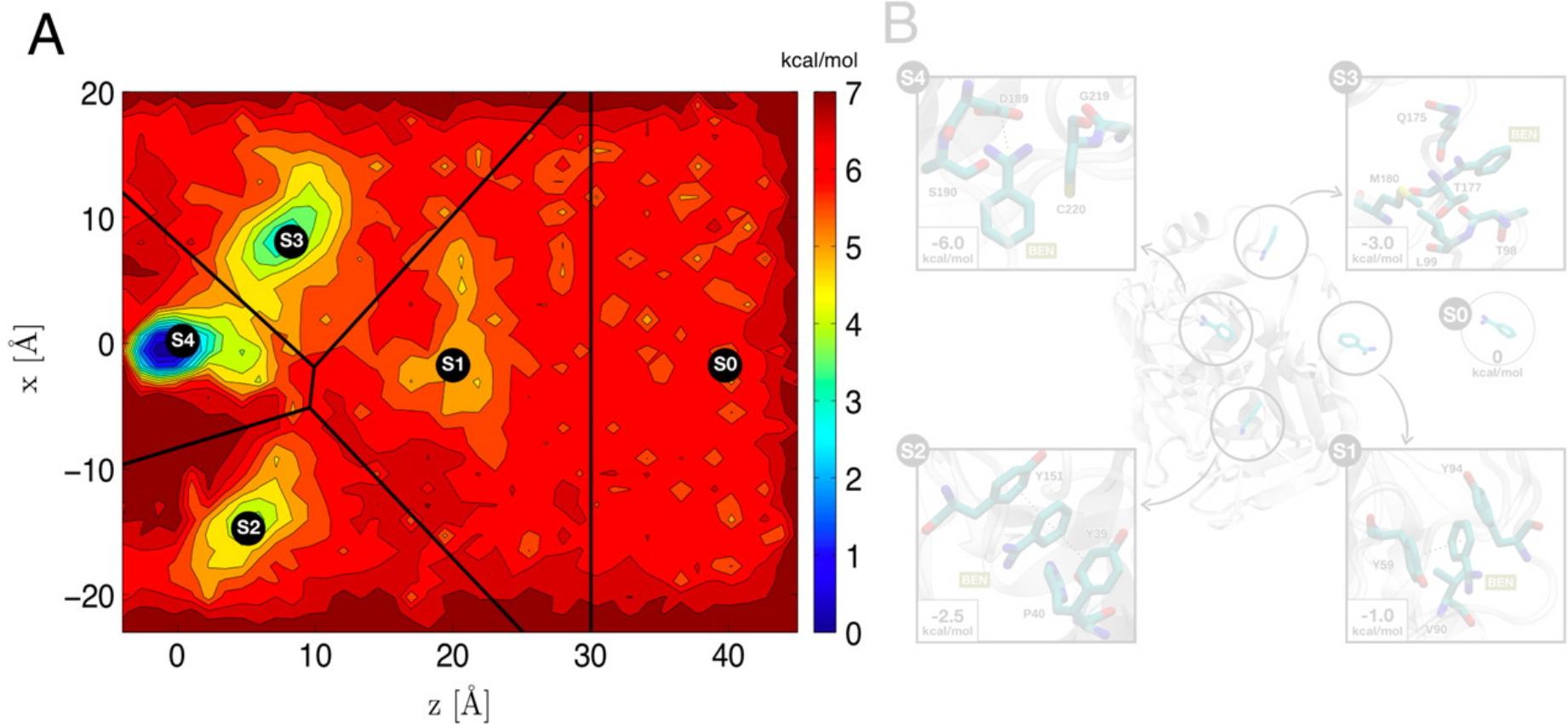
A



B

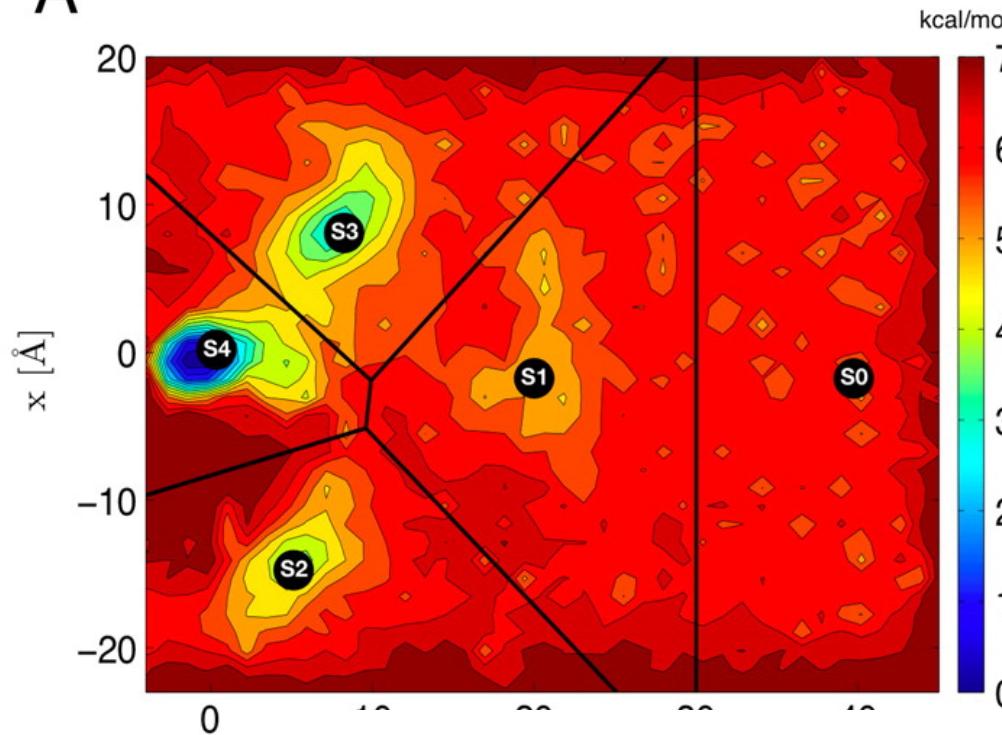


**A****B**

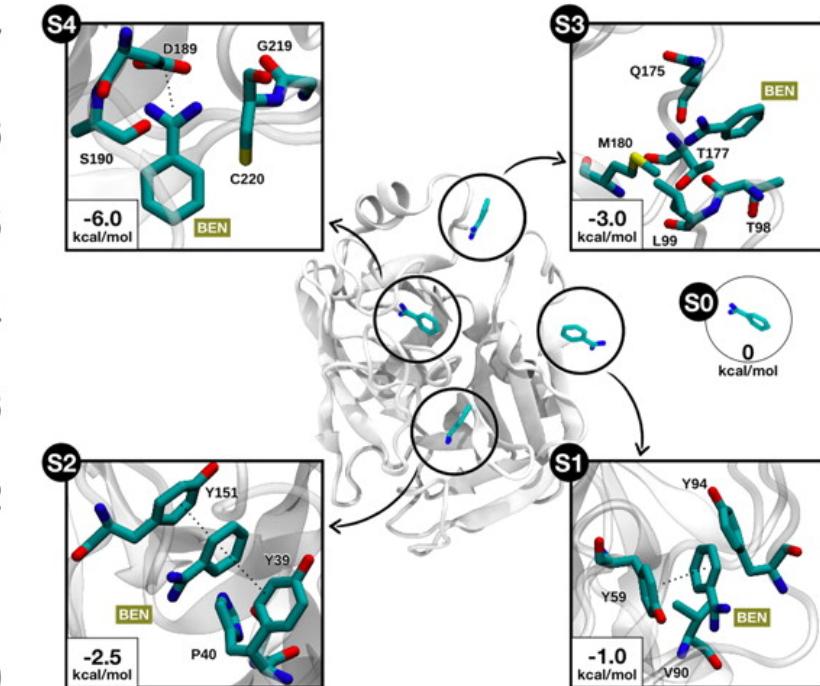


Identification of metastable states. (A) PMF in the  $xz$  plane. Five different metastable states can be identified from the different free-energy minima (S0 to S4). The relative free energy between the unbound state S0 and the bound state S4 is -6 kcal/mol. The most probable transition to the bound state S4 may be from S3 from the fact that the barrier between the two states is just 1.5 kcal/mol. (B) Structural characterization of metastable states. In states S1 and S2, benzamidine is stabilized by  $\pi$ - $\pi$  stacking interactions with Y151 and Y39 side chains. In S3, a hydrogen bond may be formed between NH<sub>2</sub> groups of benzamidine (only heavy atoms shown for clarity) and Q175 side chain, or by a cation- $\pi$  interaction between the Q175 side chain again, and benzamidine's benzene ring.

A



B



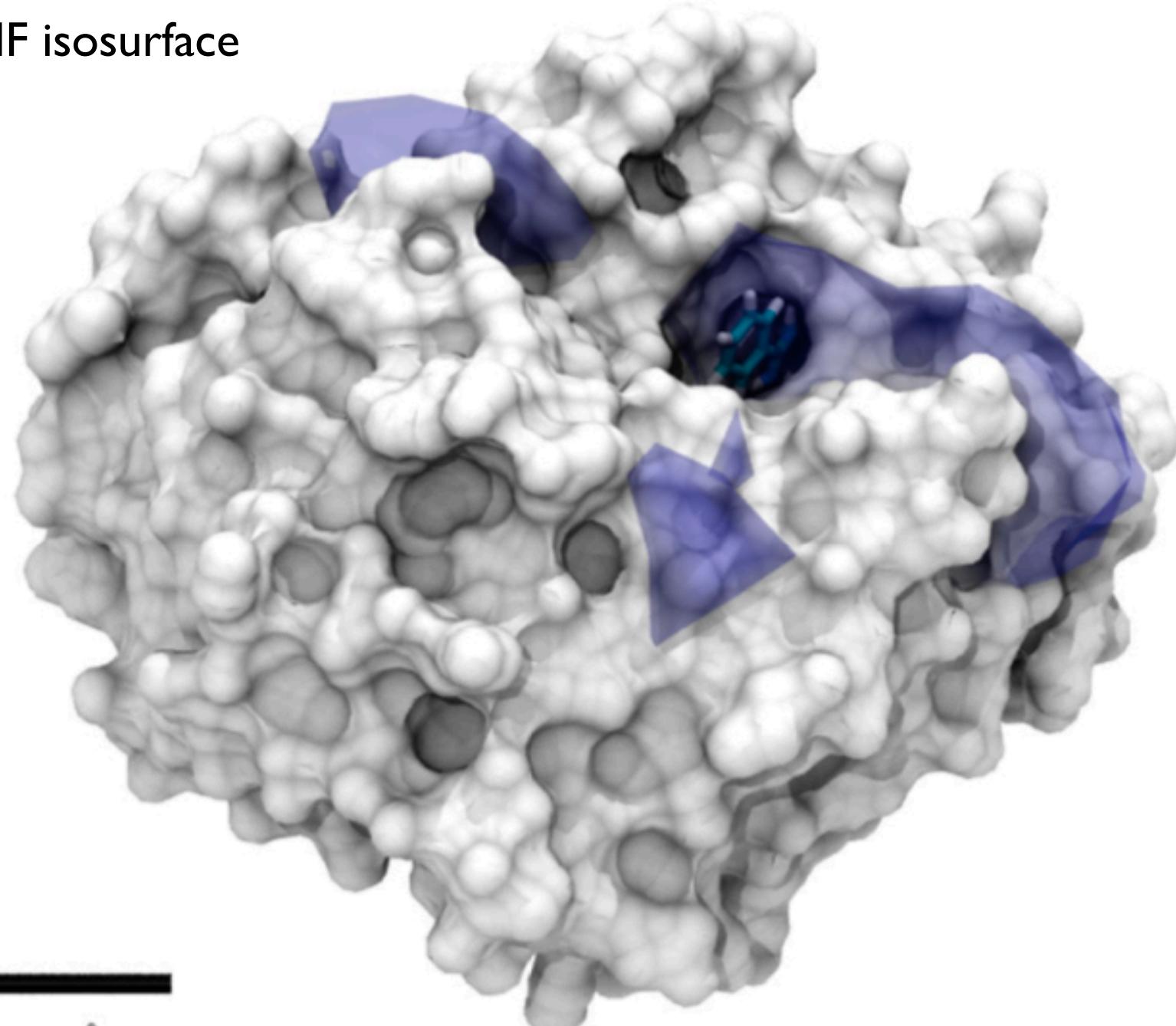
**Table S1. Transition probabilities for the metastable states in the five-state coarse-grained model at a lag time of 50 ns**

Identification of the metastable states was identified from the energy landscape. State S0 and the state S1 were identified from S2 and S3 from the energy landscape. Characterization of the interactions with the ligand benzimidazole (BEN) and of benzimidazole (BEN) interactions with the side chains between the Q175 and the T177 residues.

	S0	S1	S2	S3	S4
S0	0.069	0.090	0.130	0.305	0.406
S1	0.066	0.094	0.124	0.300	0.416
S2	0.059	0.091	0.186	0.313	0.352
S3	0.073	0.096	0.103	0.361	0.366
S4	0.021	0.032	0.043	0.107	0.797

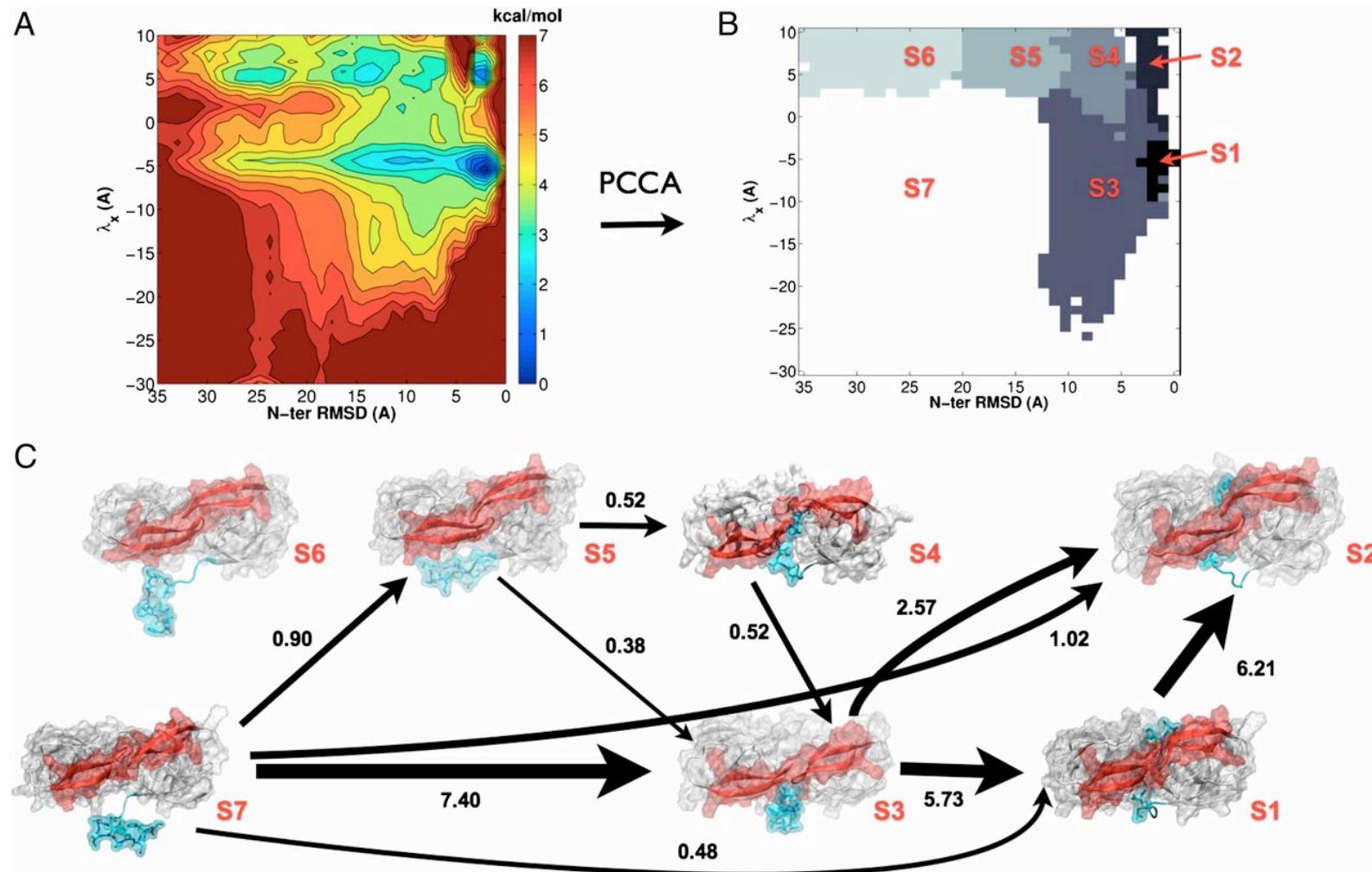
states can be identified in the unbound state. State S4 may be characterized by π-π stacking between NH<sub>2</sub> groups and π interaction between the Q175 and the T177 residues.

3D PMF isosurface



10 Å

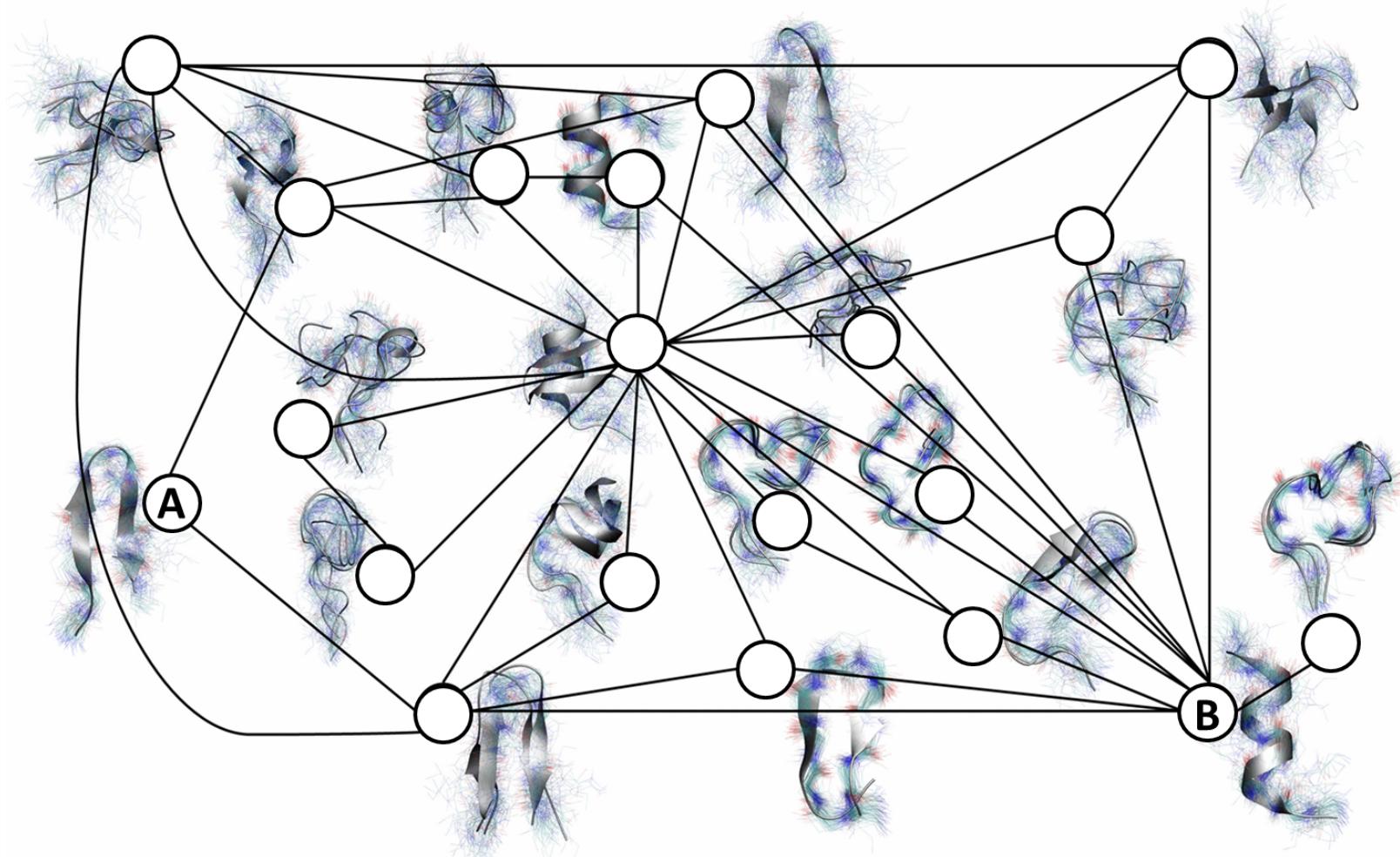
# Kinetic characterization of the critical step in HIV-1 protease maturation

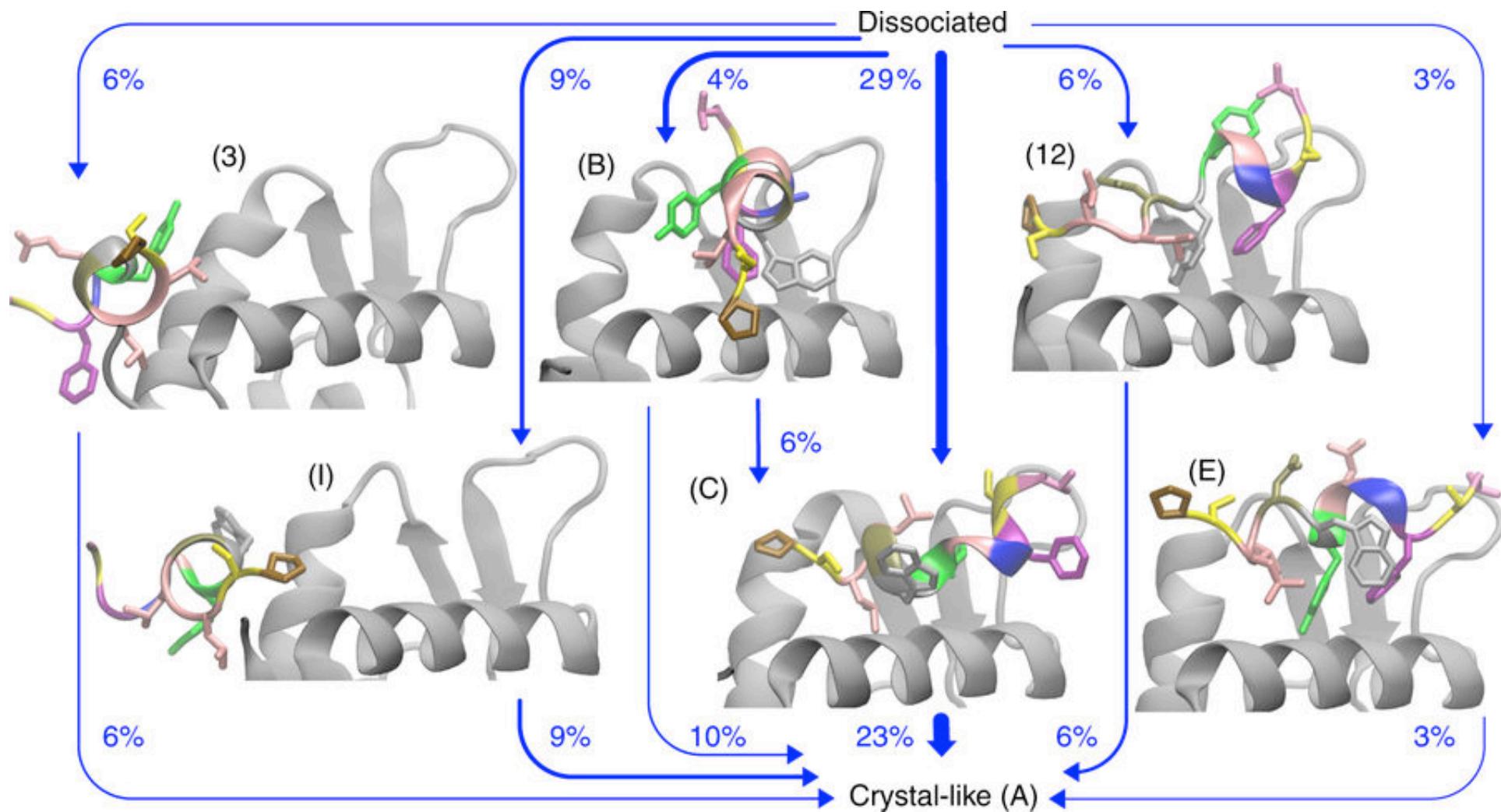


[www.pnas.org/content/109/50/20449.full](http://www.pnas.org/content/109/50/20449.full)

# Folding kinetics

It is possible to model (small) protein folding processes at all-atom resolution.  
However, the folding landscape is quite convoluted!





Binding mechanism comprised by the 60% most probable pathways. Structures of metastable (on-pathway) intermediates are shown, labels are as in Fig. 1. Arrows indicate the direction and relative magnitude of the reactive flux from the dissociated state to the crystal-like bound state. PMI residues that form PMI–Mdm2 contacts with at least a probability of 0.5 in a given macro-state are shown as sticks

**End**