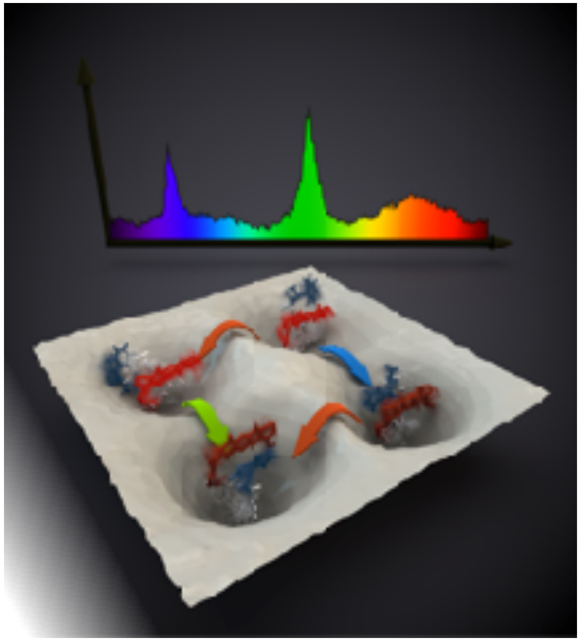
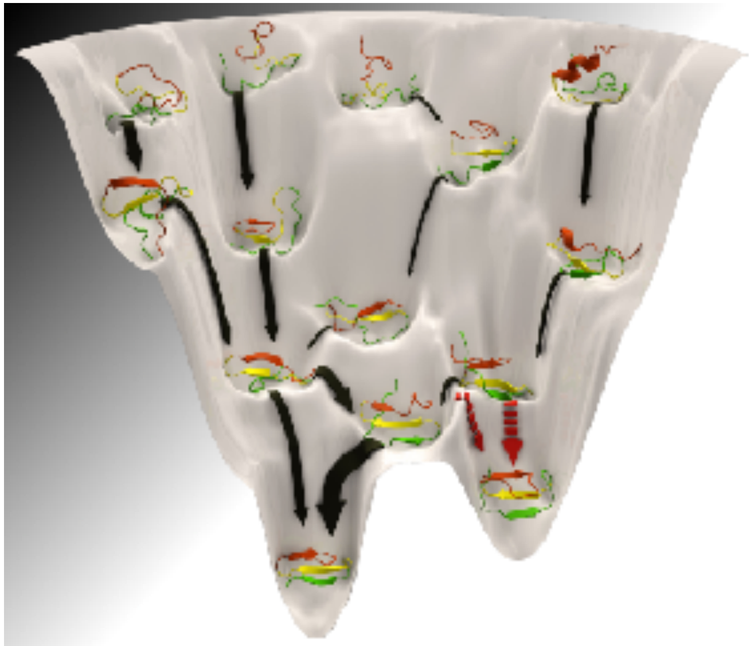
Protein dynamics and markov modeling





Frank Noé

Talk 01 - Introduction + Overview

Computational Molecular Biology

Freie Universität Berlin

Before we start…

installing for the first time?

```
conda config --add channels conda-forge
```

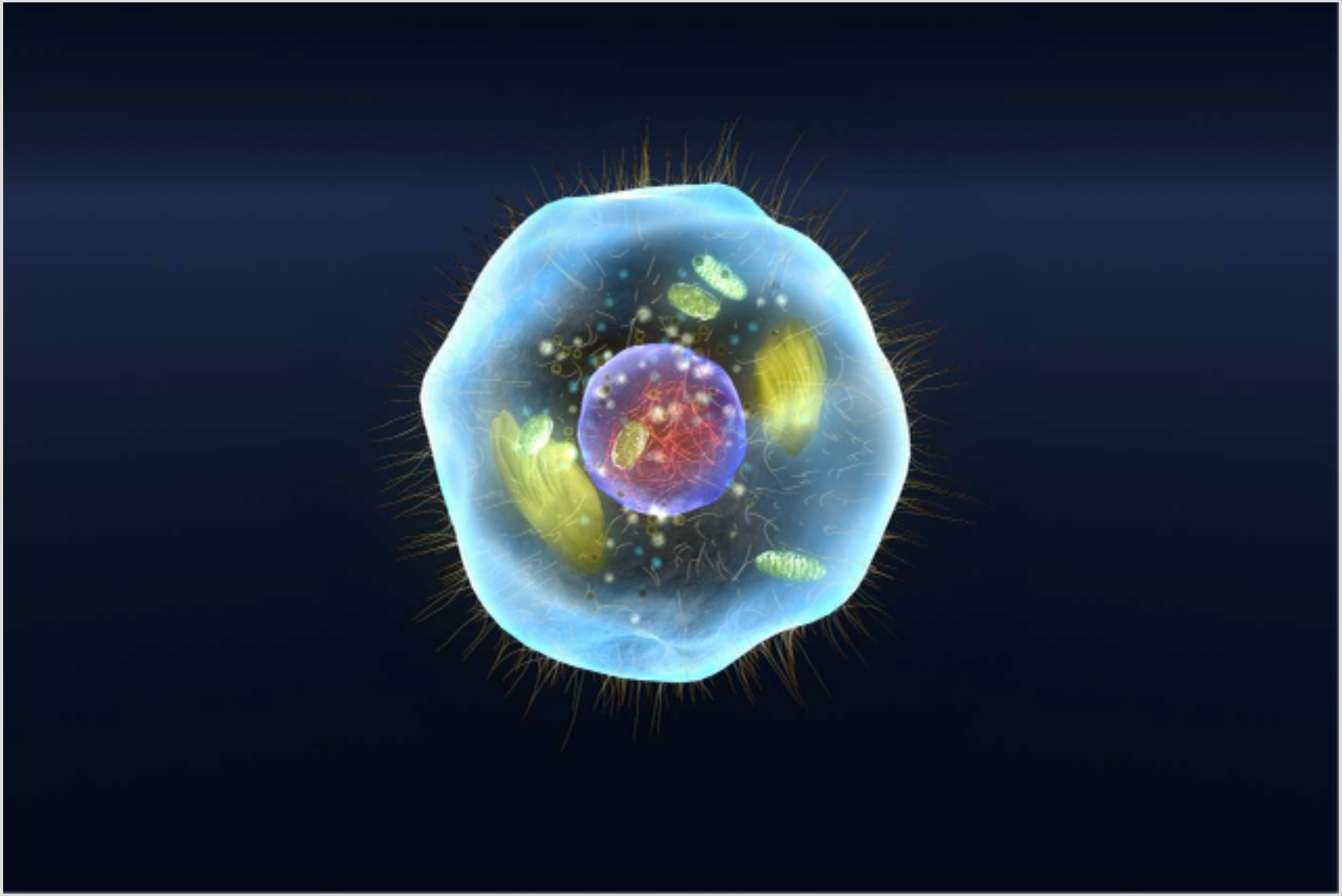
install / upgrade PyEMMA

```
conda install pyemma
```

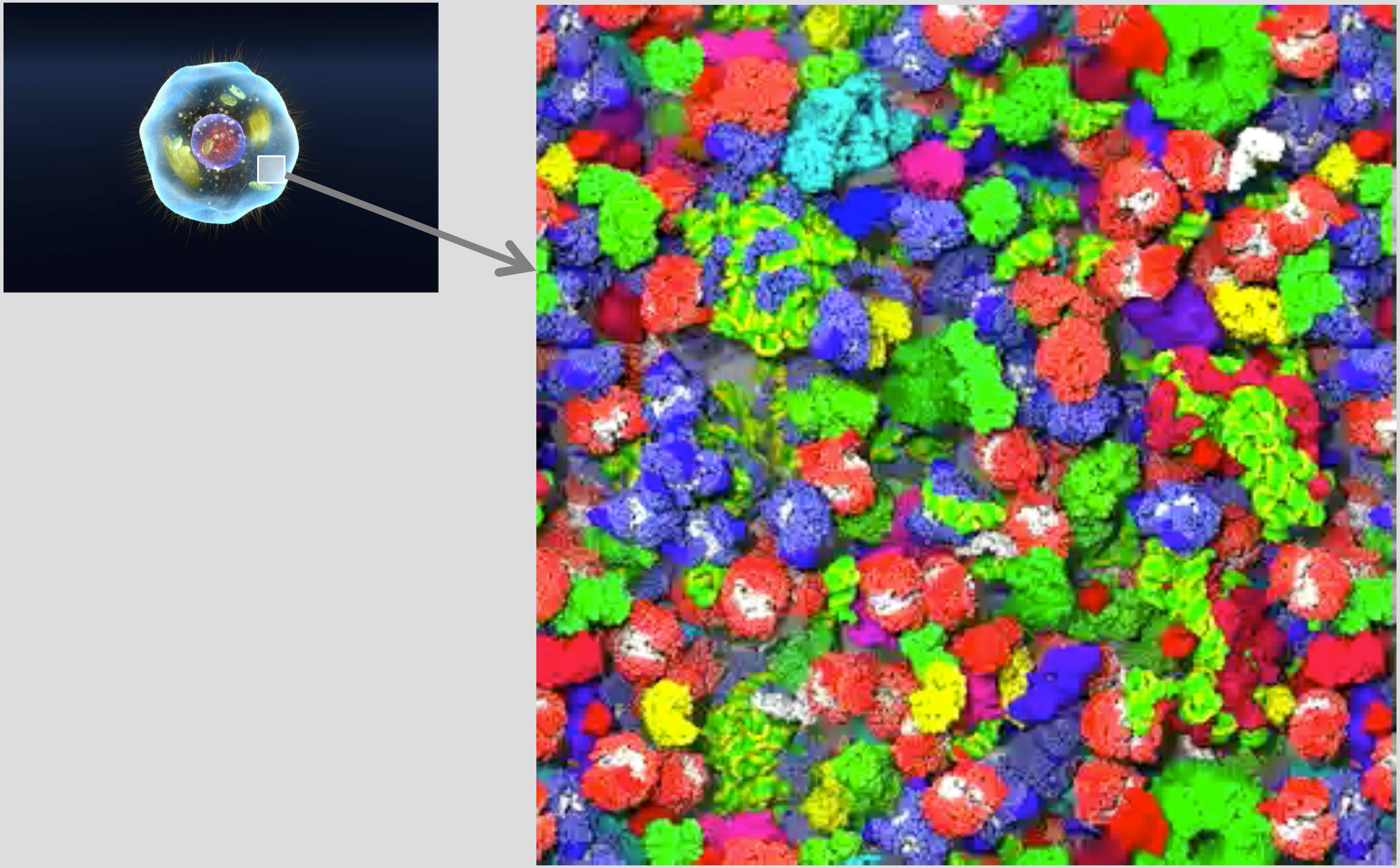
test your installation:

```
import pyemma
print pyemma.__version__
```

Cell

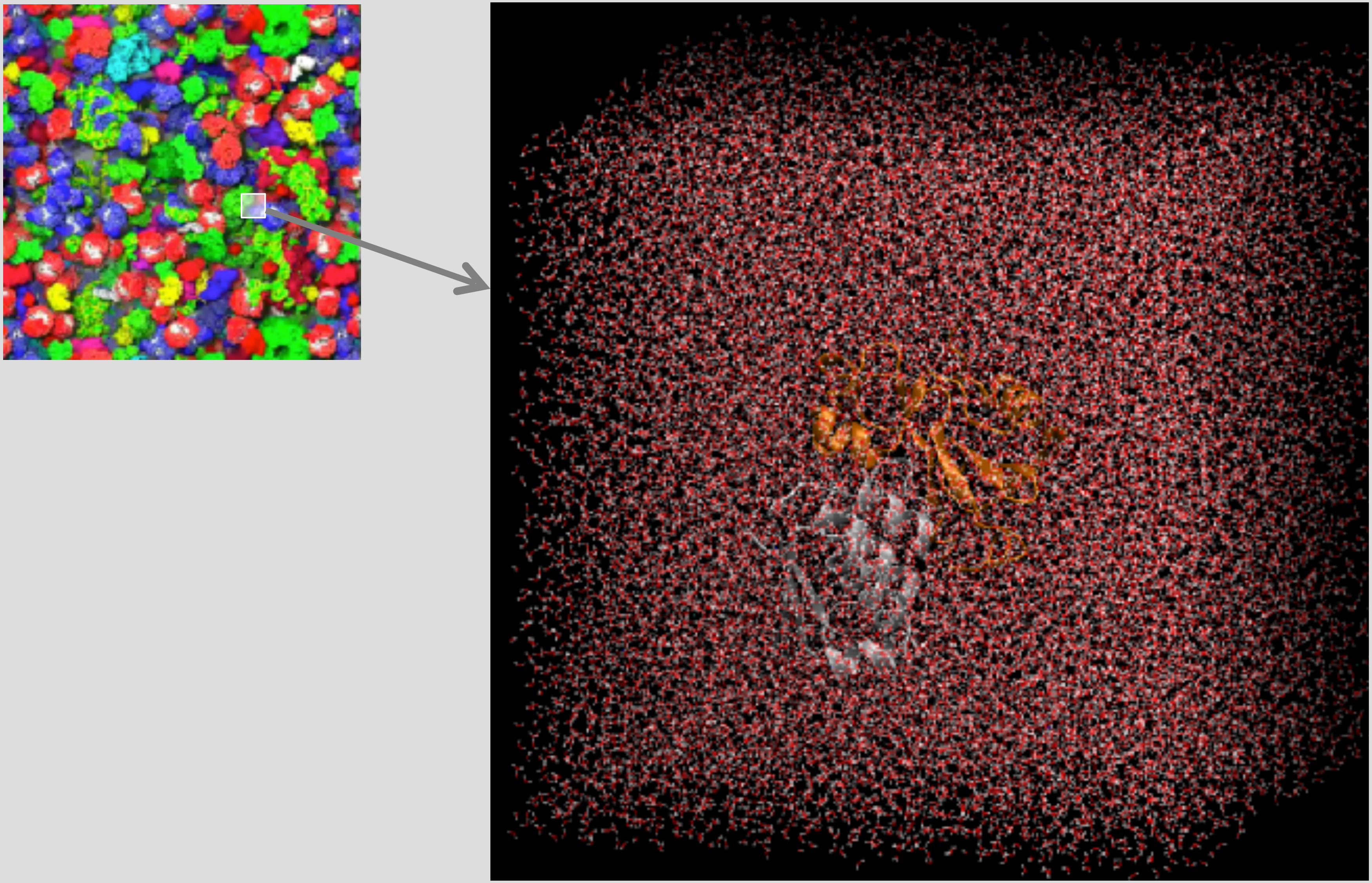


Proteins

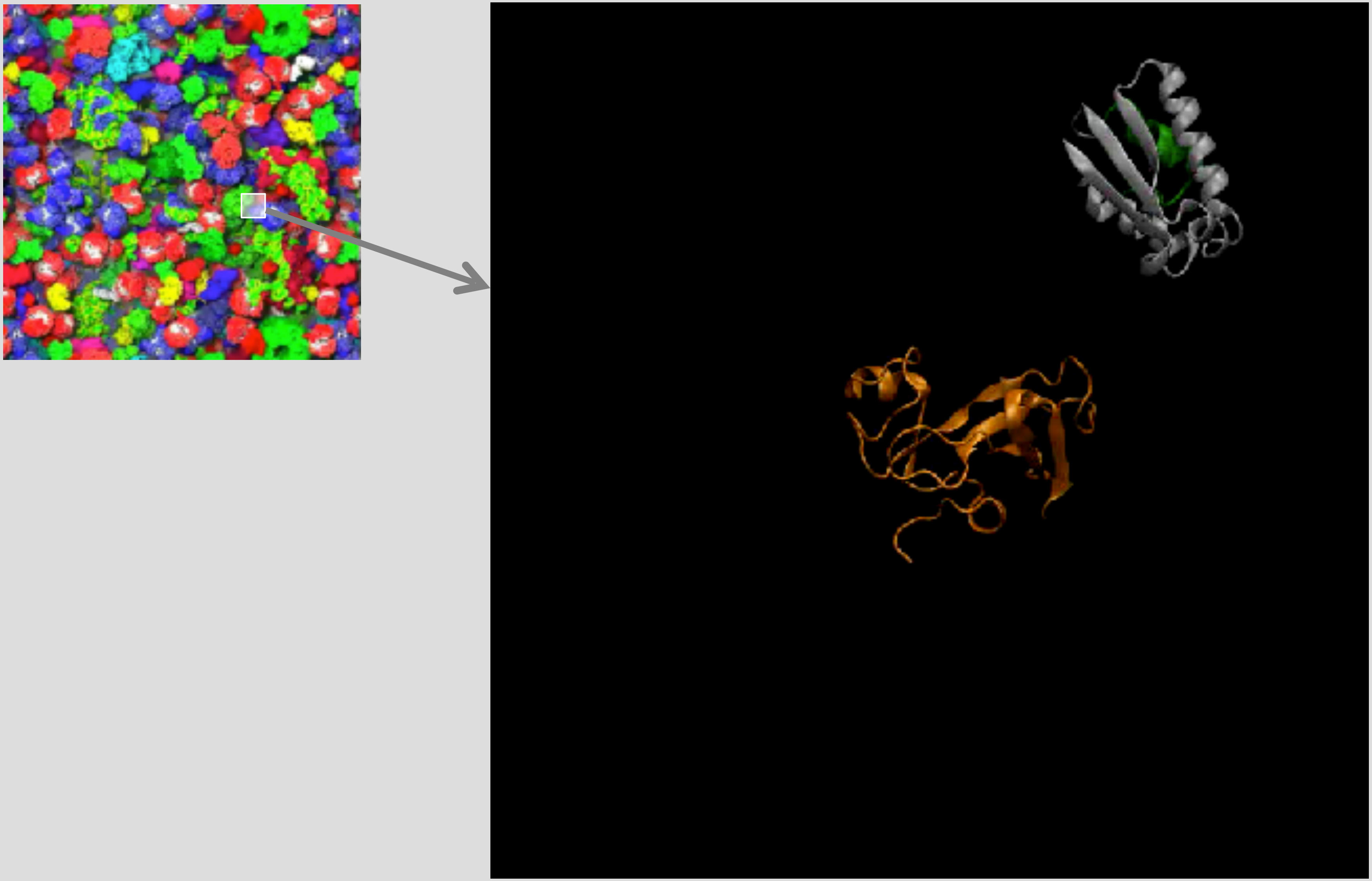


McGufee and Elcock, PloS Comput Biol 2010

Protein-Protein binding



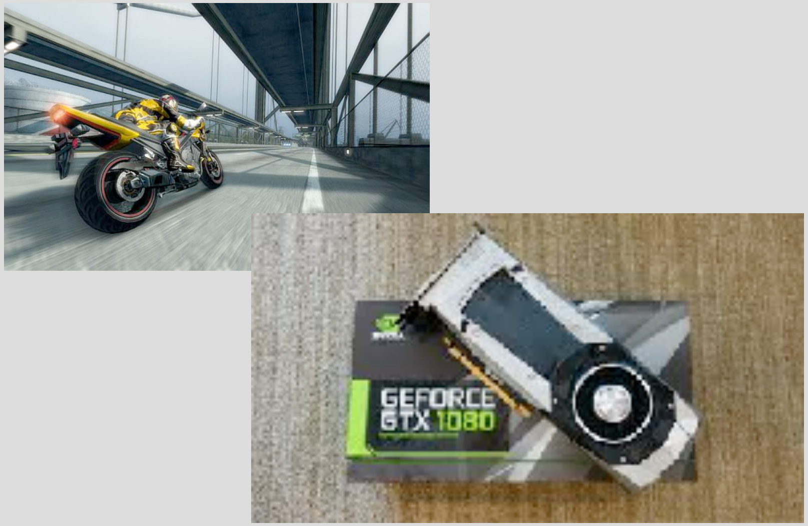
Protein-Protein binding



Plattner, Doerr, De Fabritiis, Noé

0.1 microseconds

50 K atom system (all atom, explicit solvent)

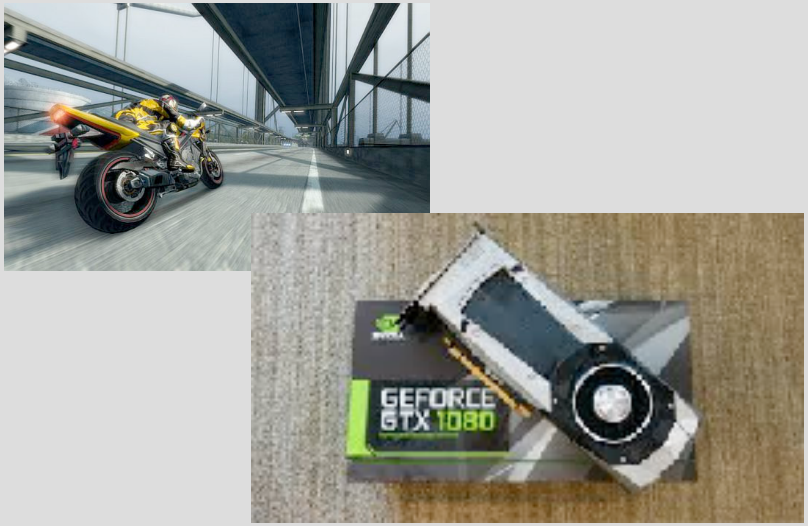


Rate

350 ns / day / GPU*
e.g. Amber, AceMD, OpenMM on Titan X

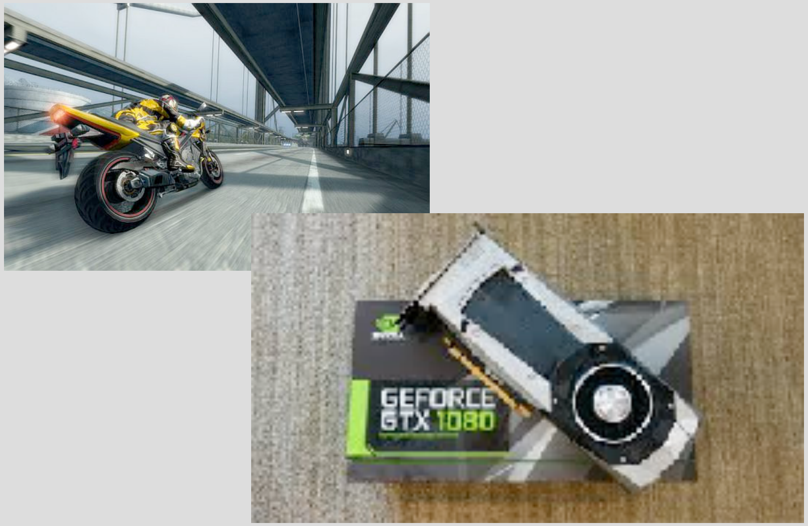
70 µs / day / Anton II

50 K atom system (all atom, explicit solvent)



Rate	350 ns / day / GPU* e.g. Amber, AceMD, OpenMM on Titan X	70 µs / day / Anton II
	200 GPUs	1 Anton II
	100 traj. of 350 ns / day	1 traj. of 10 µs / day
Throughput	70 µs / day	70 µs / day

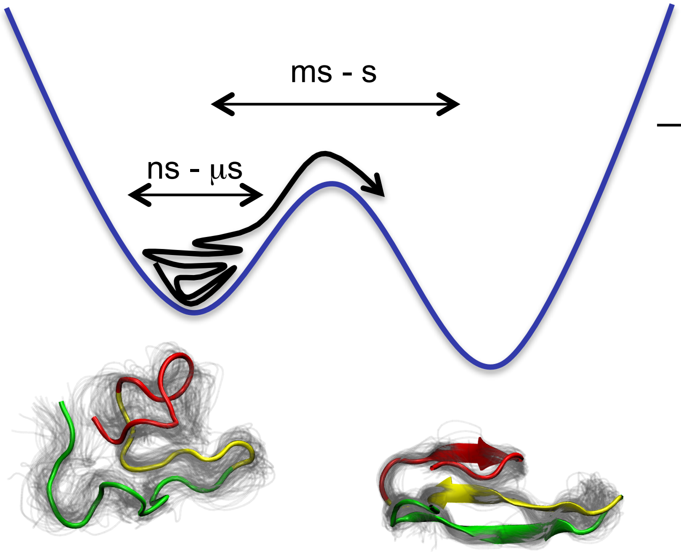
50 K atom system (all atom, explicit solvent)



Rate	350 ns / day / GPU* e.g. Amber, AceMD, OpenMM on Titan X	70 µs / day / Anton II
	200 GPUs	1 Anton II
	100 traj. of 350 ns / day	1 traj. of 10 µs / day
Throughput	70 µs / day	70 µs / day
Cost	200.000 USD	20.000.000 USD ???

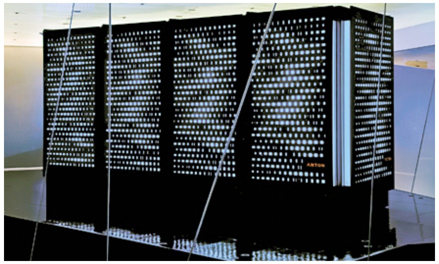
Conformation Dynamics / Markov models

Sampling Problem

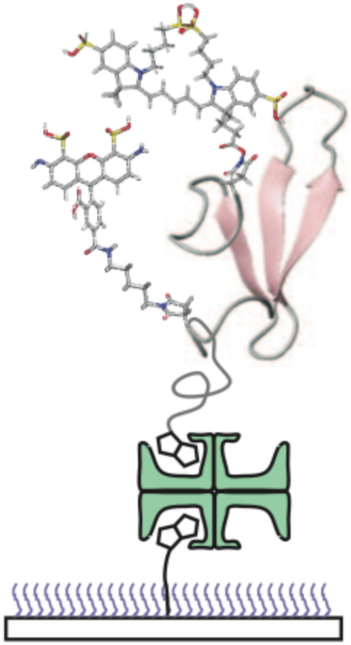


Analysis Problem

huge, complex datasets

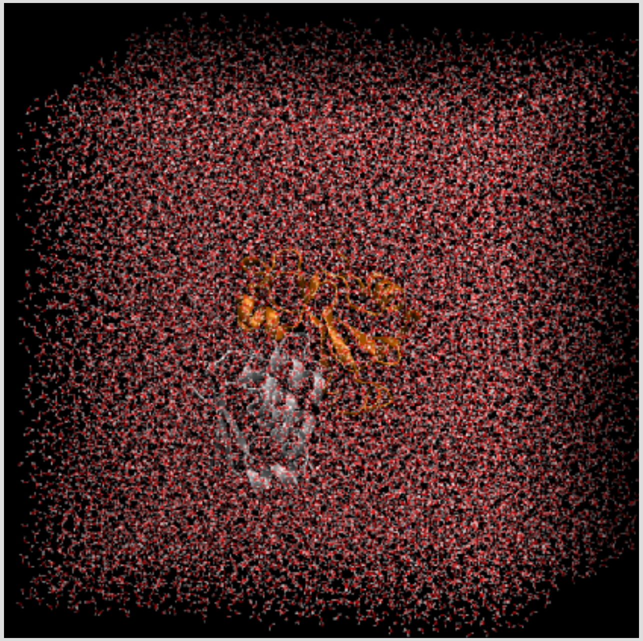


Reconciliation with Experiment



All atom MD

1000 x 1000 ns
in 1 month

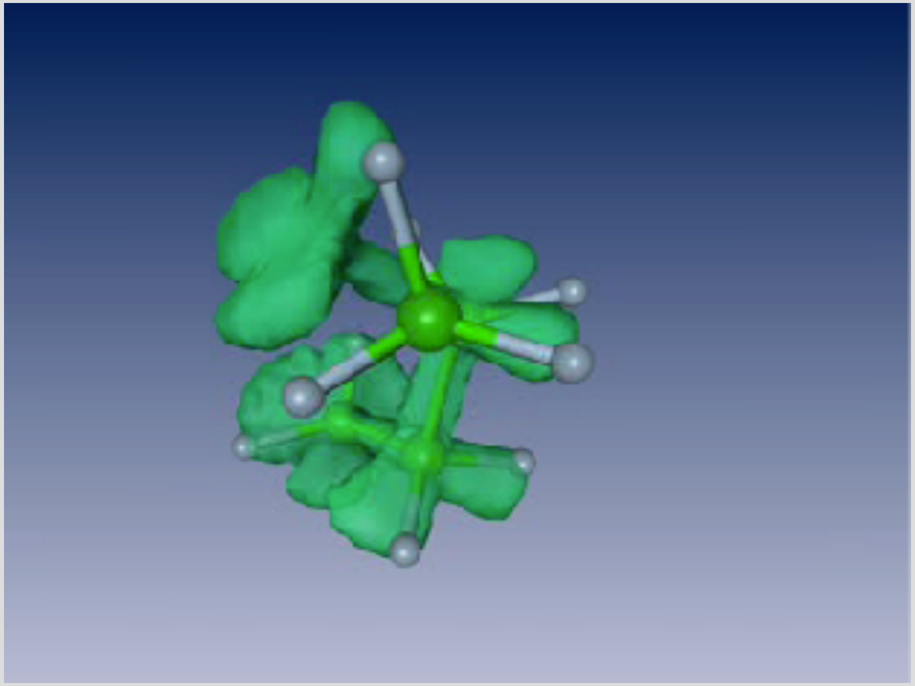


Analysis

Markov models

So what do we do?

Conformation dynamics



Boltzmann statistics

- Molecular motion is primarily driven by thermal fluctuations, and thus inherently stochastic
- A molecular system driven by thermal motion is also reversible and stationary, at least in between "*nonreversible checkpoints*".
- The stationary distribution is given by the Boltzmann distribution

$$\begin{aligned}\mu(\mathbf{x}) &= Z^{-1}\exp\left(-\frac{U(\mathbf{x})}{k_BT}\right)\\ Z &= \int_{\mathbf{x}\in\Omega}\exp\left(-\frac{U(\mathbf{x})}{k_BT}\right)\,\mathrm{d}\mathbf{x}\end{aligned}$$

Z or integrals over sets of $\mathbf{x}$ cannot be computed exactly for nontrival systems, and must therefore be sampled.

Expectation values / sampling problems

- Meaningful are expectation values:

$$\begin{aligned}\mathbb{E}[a] &= \int_{\mathbf{x}\in\Omega} \mu(\mathbf{x})\, a(\mathbf{x})\, \mathrm{d}\mathbf{x} \\ \mathbb{E}[(a,b);\tau] &= \int_{\mathbf{x}\in\Omega}\int_{\mathbf{y}\in\Omega} \mu(\mathbf{x})\, a(\mathbf{x})\, p(\mathbf{x}\to\mathbf{y};\tau)\, b(\mathbf{x})\, \mathrm{d}\mathbf{x}\, \mathrm{d}\mathbf{y}\end{aligned}$$

- Example 1: probability of being in the folded state F (rather than unfolded U):

$$p_F = \mathbb{E}[\mathbf{1}_F] = \int_{\mathbf{x}\in\mathrm{F}} \mu(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

  and the free energy difference of folding is then

$$\frac{\Delta G}{k_B T} = -\ln \frac{p_F}{1-p_F}$$

- Problem: In order to evaluate the above integrals, the parts of state space with significant weights $\mu(\mathbf{x})$ must be sampled. However, this is very hard because of free energy barriers / metastable states.

The Markov model trick

We rewrite the problem by introducing a state space partition $\{S_1, ..., S_n\}$ with $\Omega = \bigcup_i S_i$:

$$\begin{aligned} \mathbb{E}[a] &= \sum_i \pi_i \int_{\mathbf{x}\in S_i} \frac{\mu(\mathbf{x})}{\pi_i} a(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_i \pi_i \bar{a}_i \qquad (1) \\ \pi_i &= \int_{\mathbf{x}\in S_i} \mu(\mathbf{x})\,\mathrm{d}\mathbf{x} \end{aligned}$$

The first equation has become much easier - the local distribution $\mu(\mathbf{x})/\pi_i$ is easy to sample if the discrete states S_i do not contain internal barriers. However the second equation is still as hard. But we can rewrite it as follows:

$$\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{P}(\tau) \qquad (2)$$

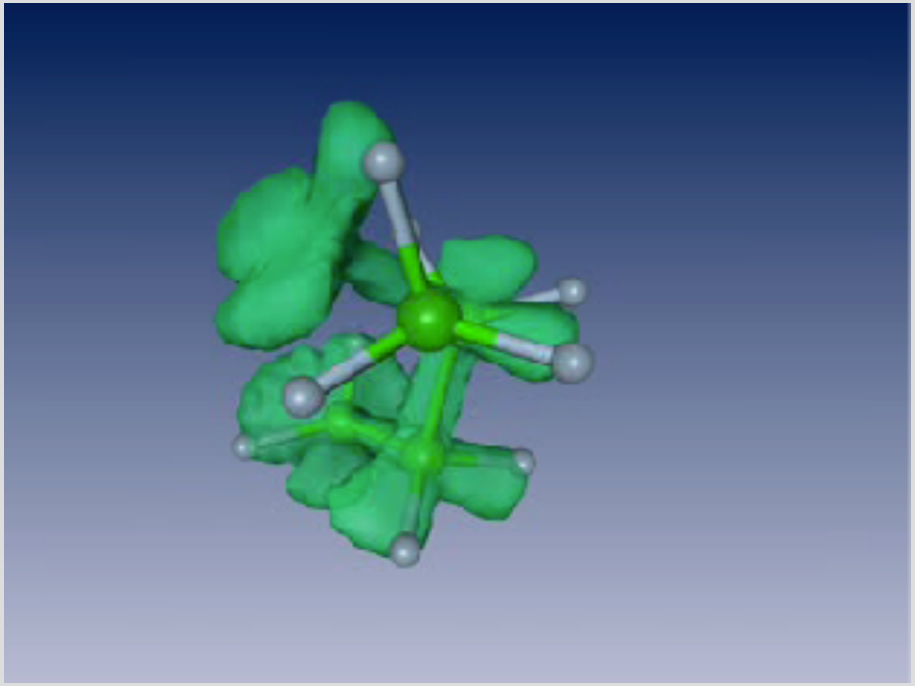
with the transition matrix

$$p_{ij}(\tau) = \int_{\mathbf{x}\in S_i} \int_{\mathbf{y}\in S_j} \frac{\mu(\mathbf{x})}{\pi_i} p(\mathbf{x} \to \mathbf{y}; \tau)\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

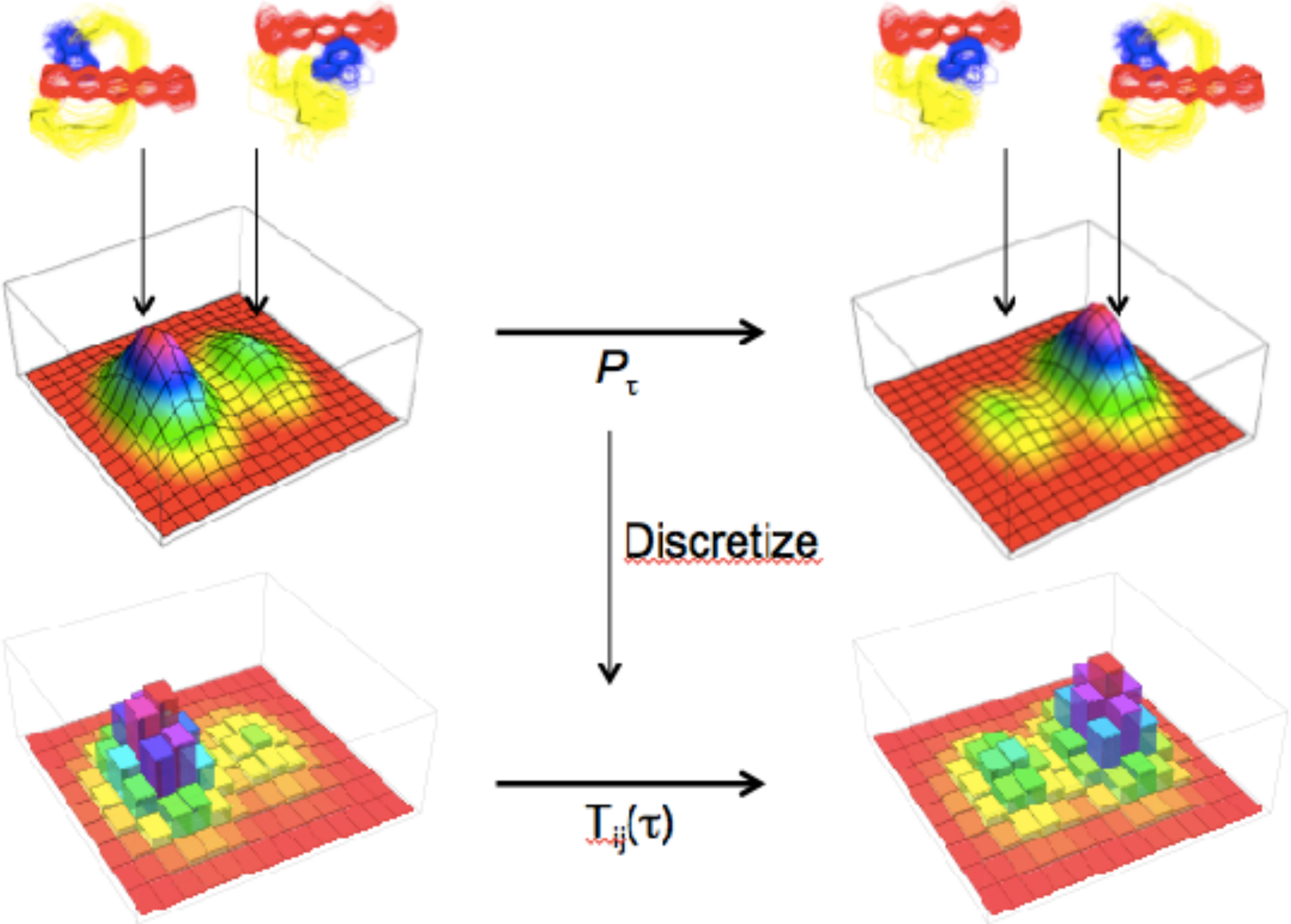
This is again relatively easy - we need to prepare starting points $\mathbf{x}$ according to the local distribution $\mu(\mathbf{x})/\pi_i$, then simulate for a (usually short) time τ and count the transition if it ends of in S_j. $p_{ij}(\tau)$ is just the fraction of transitions ending up in S_j after time τ given that we start from S_i. So we can estimate it without knowing π_i.

We can then reconstruct the unbiased $\boldsymbol{\pi} = [\pi_i]$ using Eq. (2), and use that in Eq. (1) to compute the expectation value. We have reduce the global sampling problem to a local sampling problem, which is much easier, and we have gained a perfect parallelization of our problem!

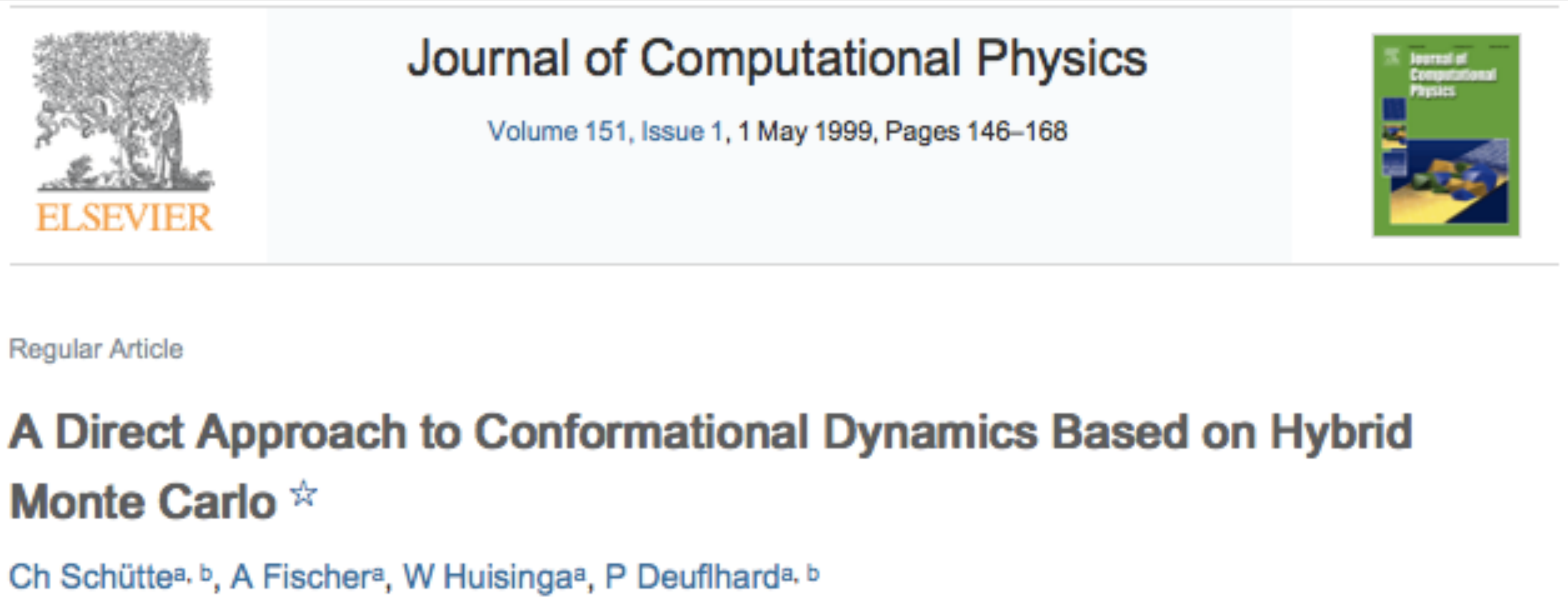
Conformation dynamics



Conformation Dynamics / Markov models



see also works by:
Andersen, Caflisch, Chodera, Deuflhard, Dill, Hummer, Pande, Schütte, Stock, Huisinga, Rao, Roux, Levy



ELSEVIER

Journal of Computational Physics

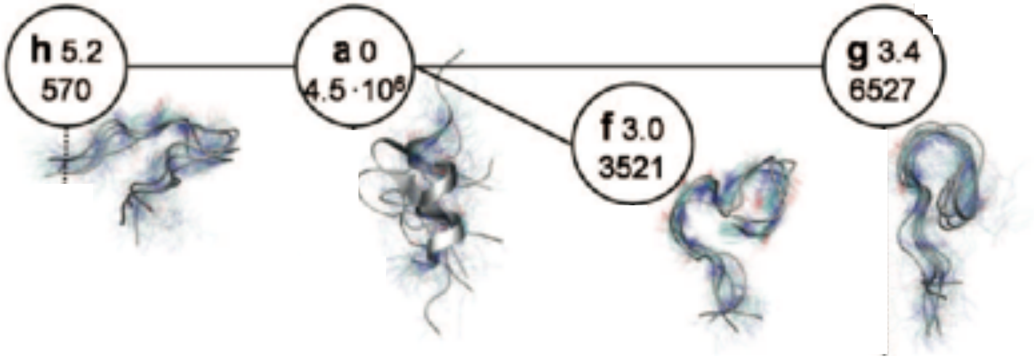
Volume 151, Issue 1, 1 May 1999, Pages 146–168

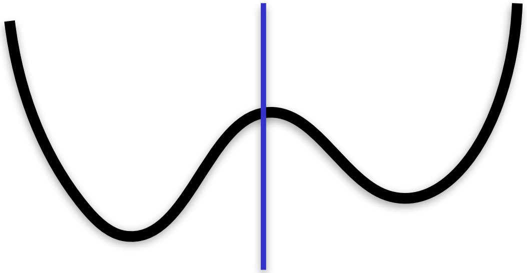
Regular Article

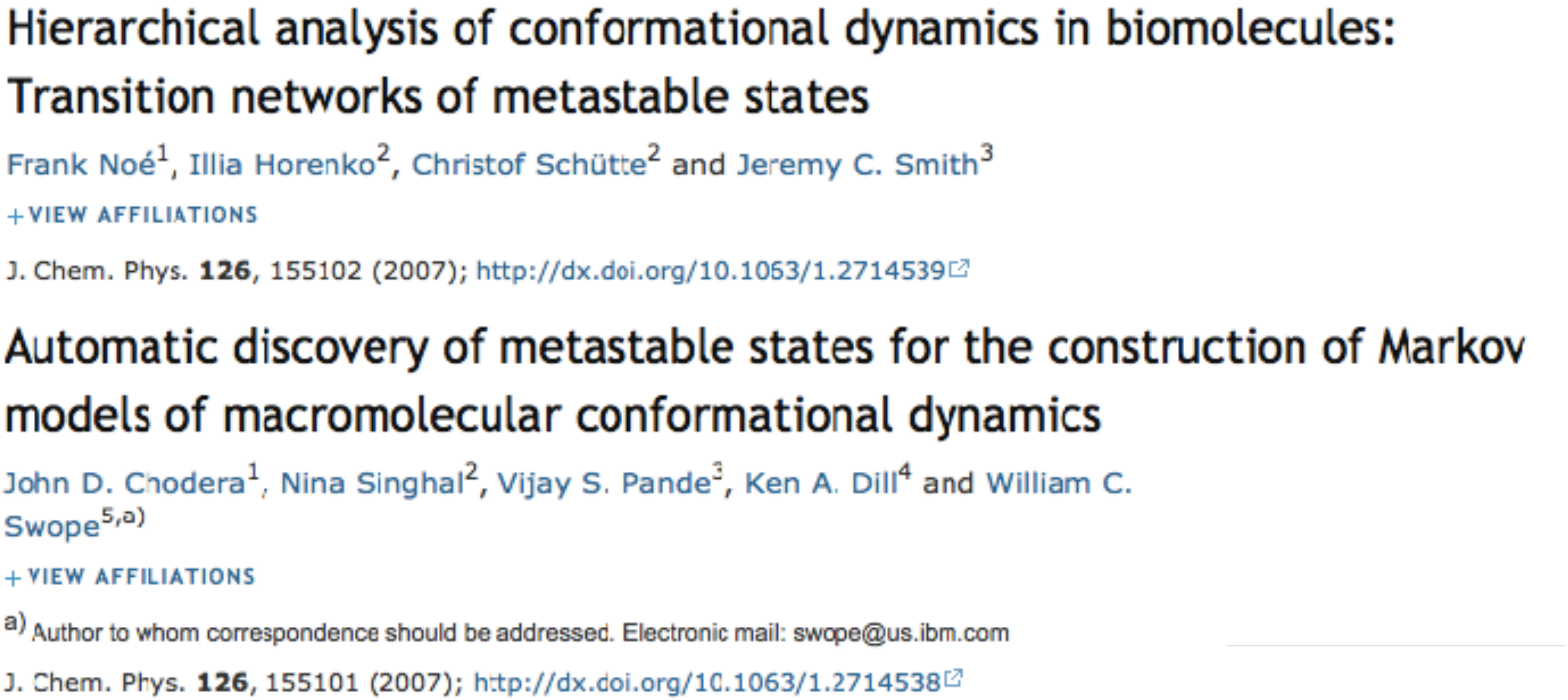
A Direct Approach to Conformational Dynamics Based on Hybrid Monte Carlo ☆

Ch Schütte[a, b], A Fischer[a], W Huisinga[a], P Deuflhard[a, b]

Generation 1 : focus on metastable states







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

Generation 2: understanding spectral properties of MSMs

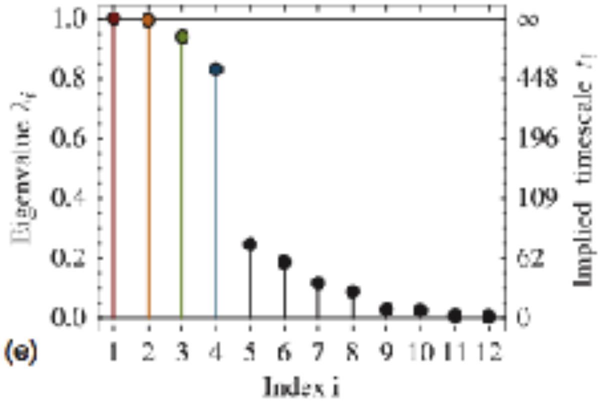
Propagator

$$p_\tau(\mathbf{z}_\tau) = \mathcal{P}(\tau)\, p_0(\mathbf{z}_0)$$

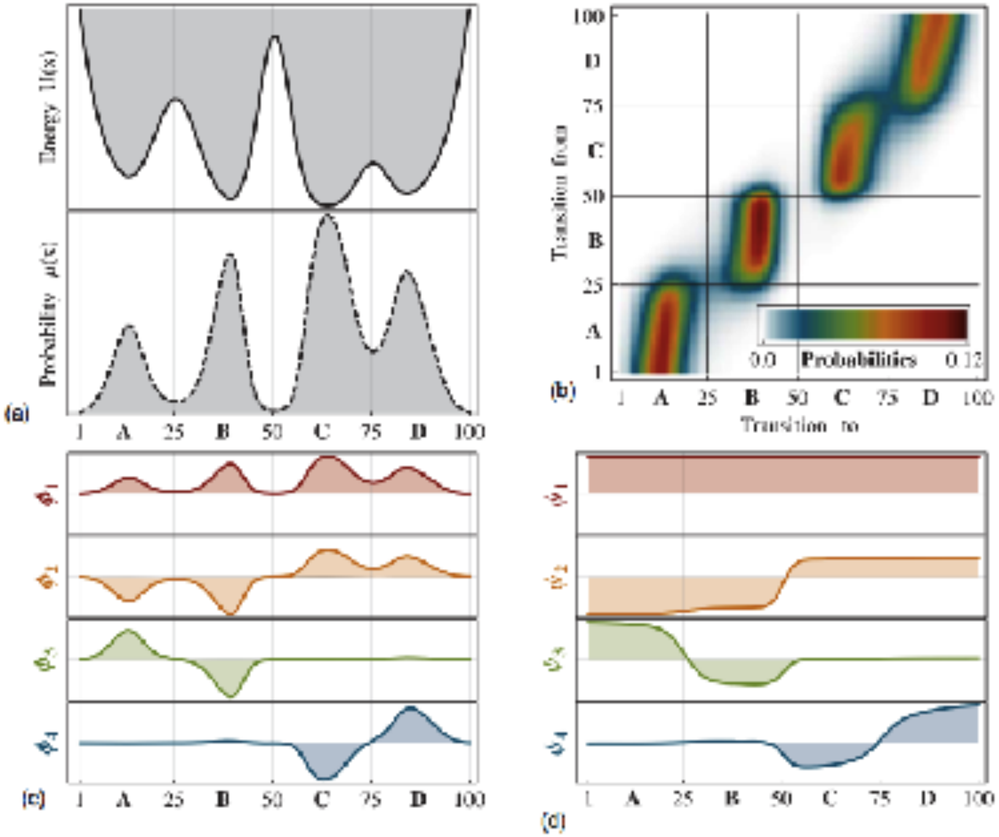
Spectral decomposition

$$p_\tau(\mathbf{z}_0, \mathbf{z}_\tau) = \mu(\mathbf{z}_\tau) + \sum_{i=2}^{\infty} \mathrm{e}^{\;\kappa_i \tau} \frac{\phi_i(\mathbf{z}_0)}{\mu(\mathbf{z}_0)} \phi_i(\mathbf{z}_\tau)$$

timescales

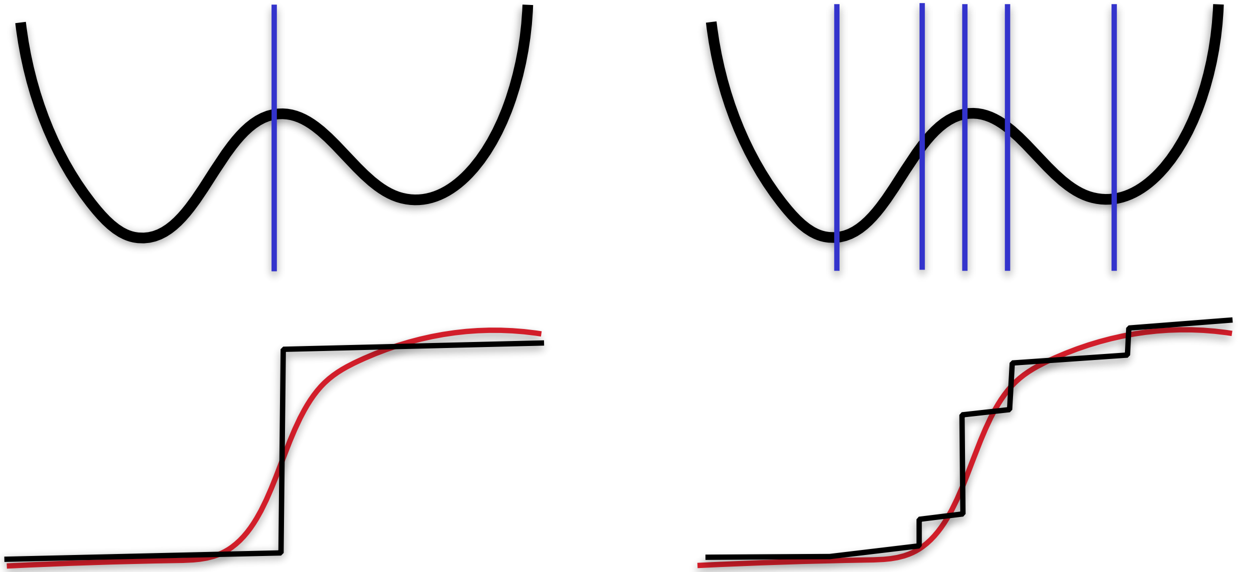


processes:



Prinz et al.: **J. Chem. Phys.** 134, p174105 (2011)

Generation 2: focus on discretizing transfer operator

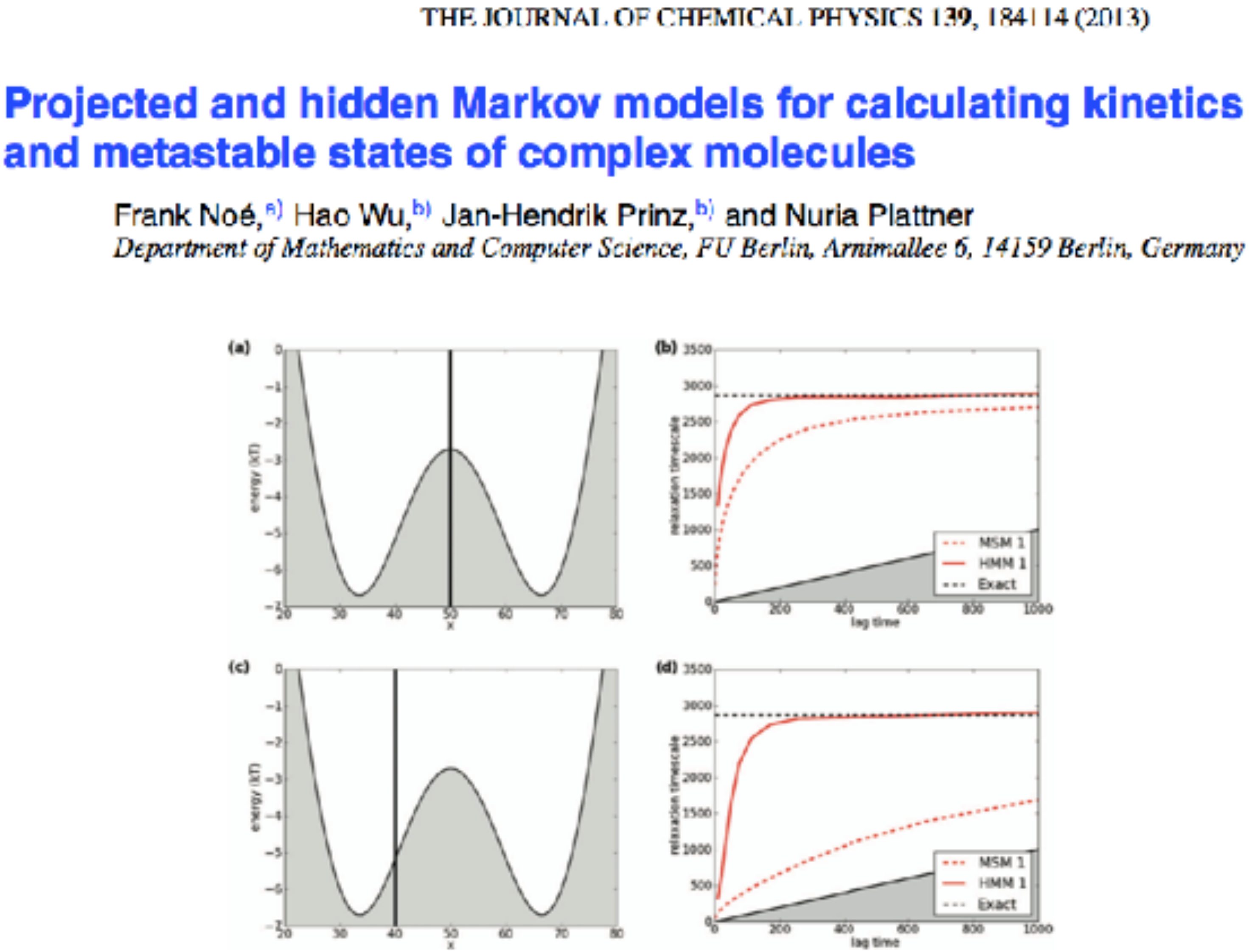


* No systematic error in the equilibrium distribution
* Systematic (discretization) error of MSM kinetics depends on eigenfunction approximation quality and lagtime.
* Timescales are always underestimated

Sarich, Noé, Schütte: On the approximation quality of Markov state models
Multiscale Model. Simul. (2010)

Prinz et al.: Markov models of molecular kinetics: generation and validation.
J. Chem. Phys. 134, p174105 (2011)

Generation 3: newer developments - HMMs

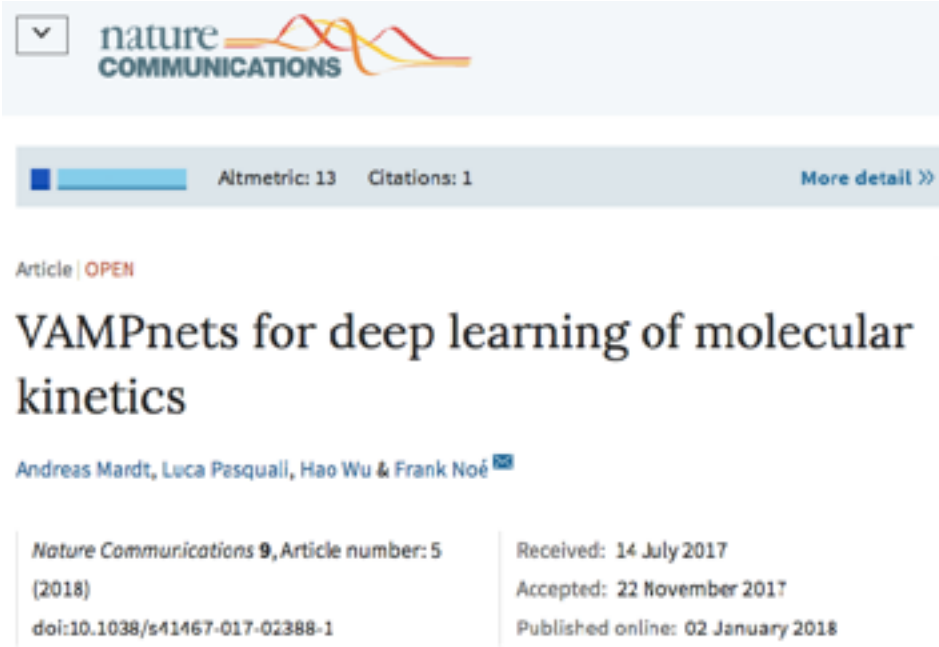


THE JOURNAL OF CHEMICAL PHYSICS 139, 184114 (2013)

Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules

Frank Noé,[a] Hao Wu,[b] Jan-Hendrik Prinz,[b] and Nuria Plattner
Department of Mathematics and Computer Science, FU Berlin, Arnimallee 6, 14159 Berlin, Germany

Generation 3: newer developments - VAMPnets



nature COMMUNICATIONS

Altmetric: 13 Citations: 1 More detail »

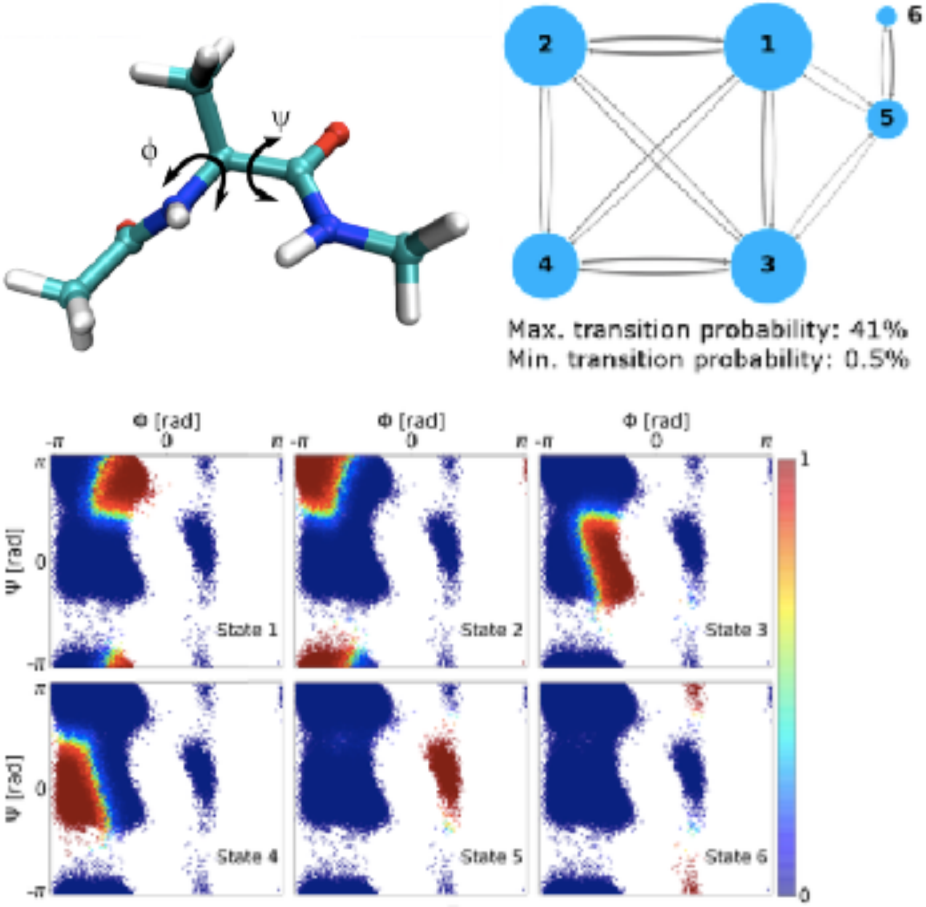
Article | OPEN

VAMPnets for deep learning of molecular kinetics

Andreas Mardt, Luca Pasquali, Hao Wu & Frank Noé

Nature Communications **9**, Article number: 5 (2018)
doi:10.1038/s41467-017-02388-1

Received: 14 July 2017
Accepted: 22 November 2017
Published online: 02 January 2018



Max. transition probability: 41%
Min. transition probability: 0.5%

Optimal reaction coordinates?

Backward propagator

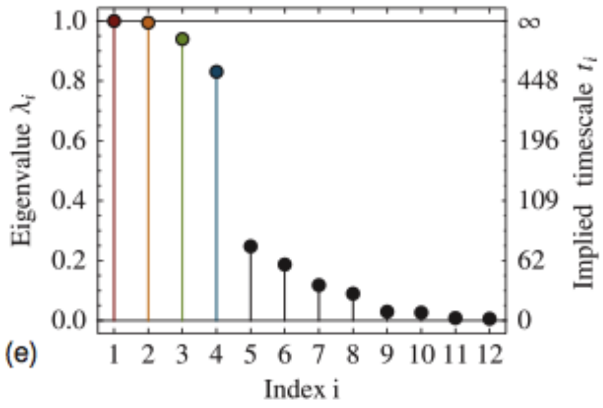
$$\rho_\tau = \mathcal{T}(\tau)\rho_0$$

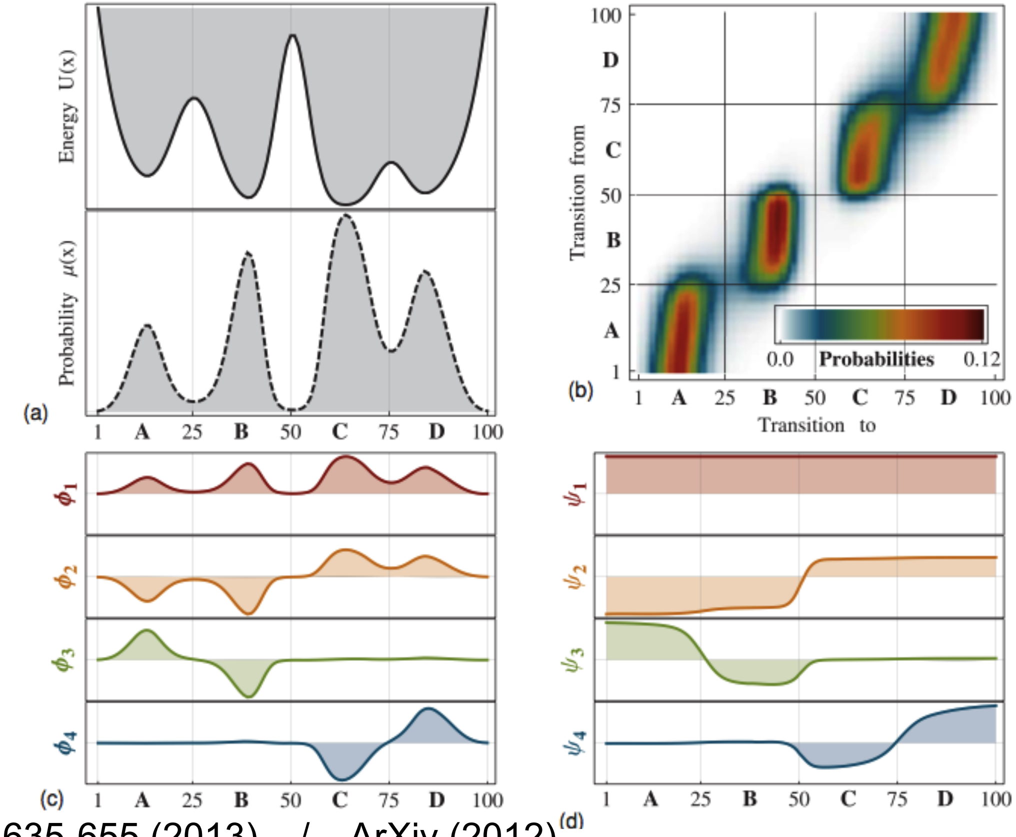
Spectral decomposition

$$\rho_\tau = \sum_{i=1}^{\infty} \mathrm{e}^{-\tau\kappa_i}\langle\psi_i \mid \rho_0\rangle\psi_i$$

Processes:

Eigenvalues / timescales κ_i^{-1}

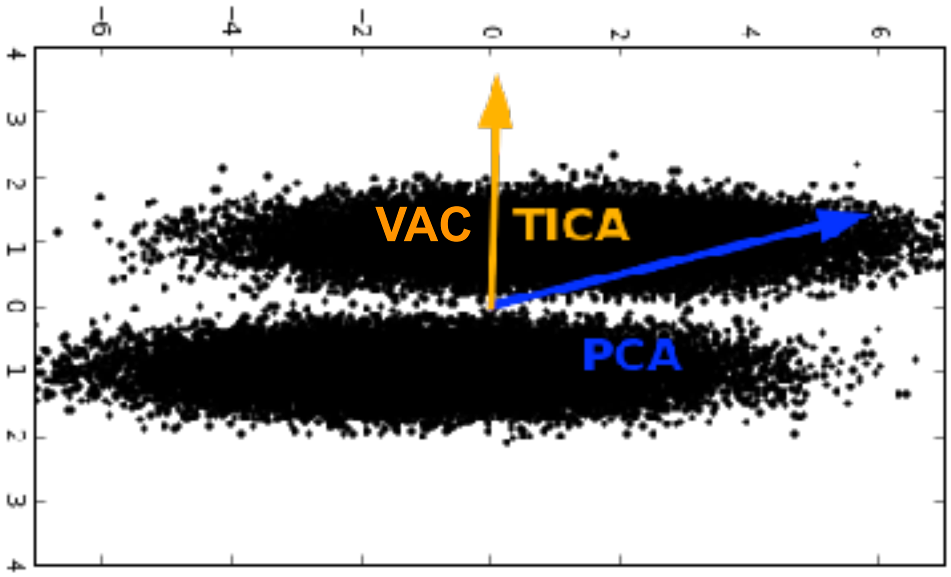




Noé and Nüske, **Multiscale Model. Simul.** 11, 635-655 (2013) / ArXiv (2012)

Nüske et al, **JCTC** 2014

How to find the slow coordinates?



Variational approach of conformation dynamics (VAC)

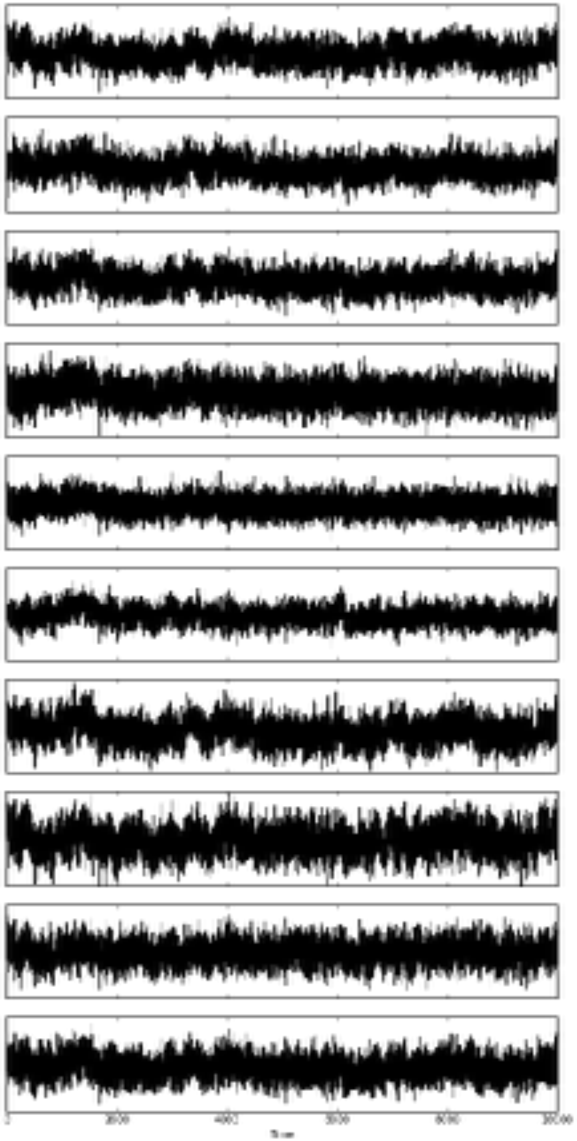
Noé and Nüske, **Multiscale Model. Simul.** 11, 635-655 (2013) / ArXiv (2012)
Nüske et al, **JCTC** 2014

Time-lagged independent component analysis (TICA)

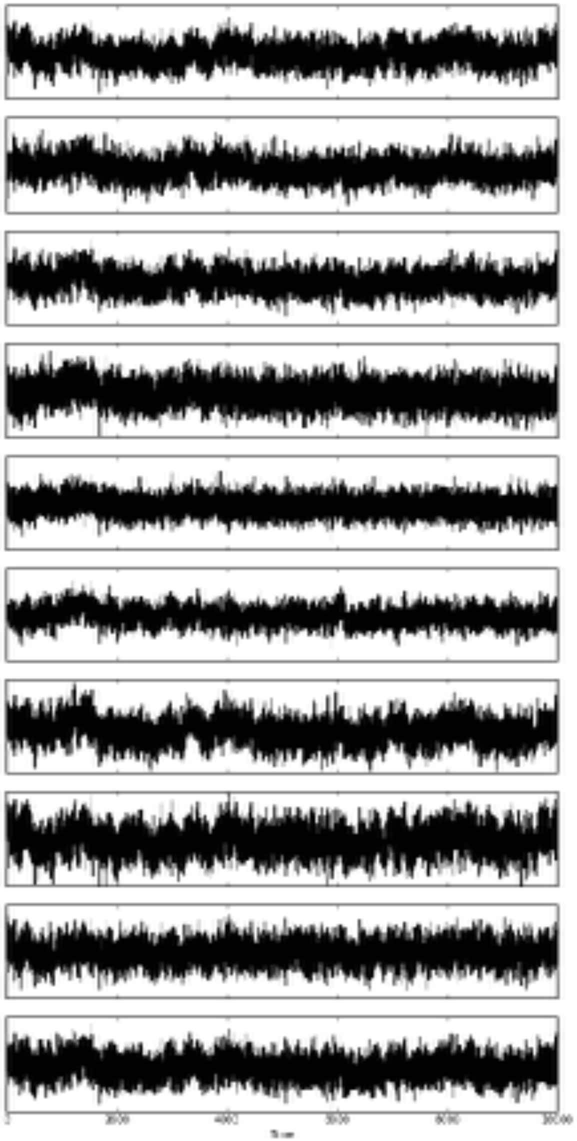
Molgedey and Schuster, **PRL** 1994
Perez-Hernandez et al, **JCP**, 139, 1502 (2013) Schwantes and Pande, **JCTC** 2013

www.pyemma.org

Input

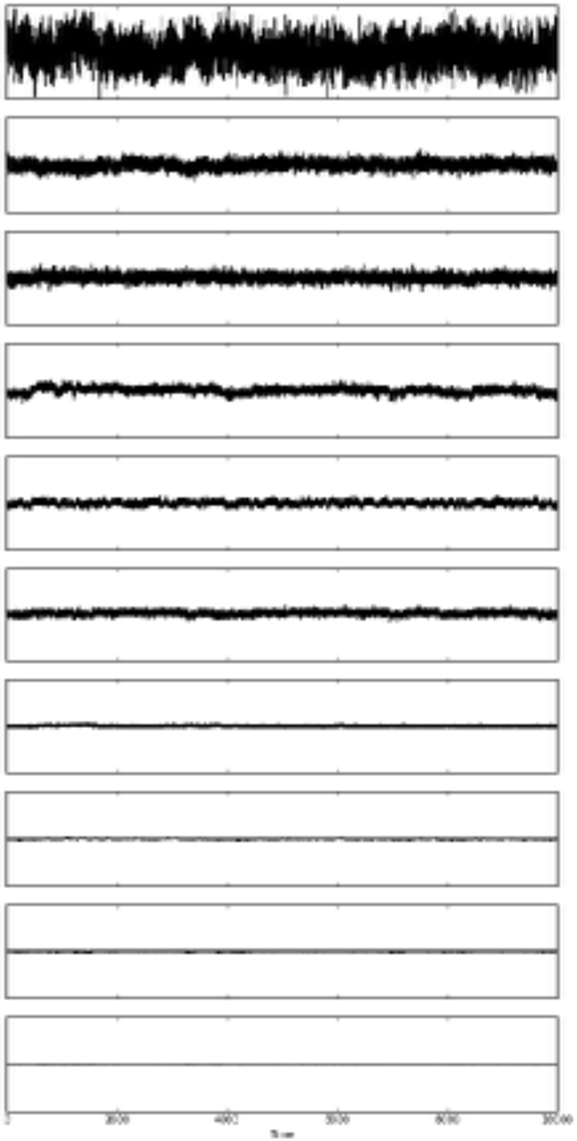


Input

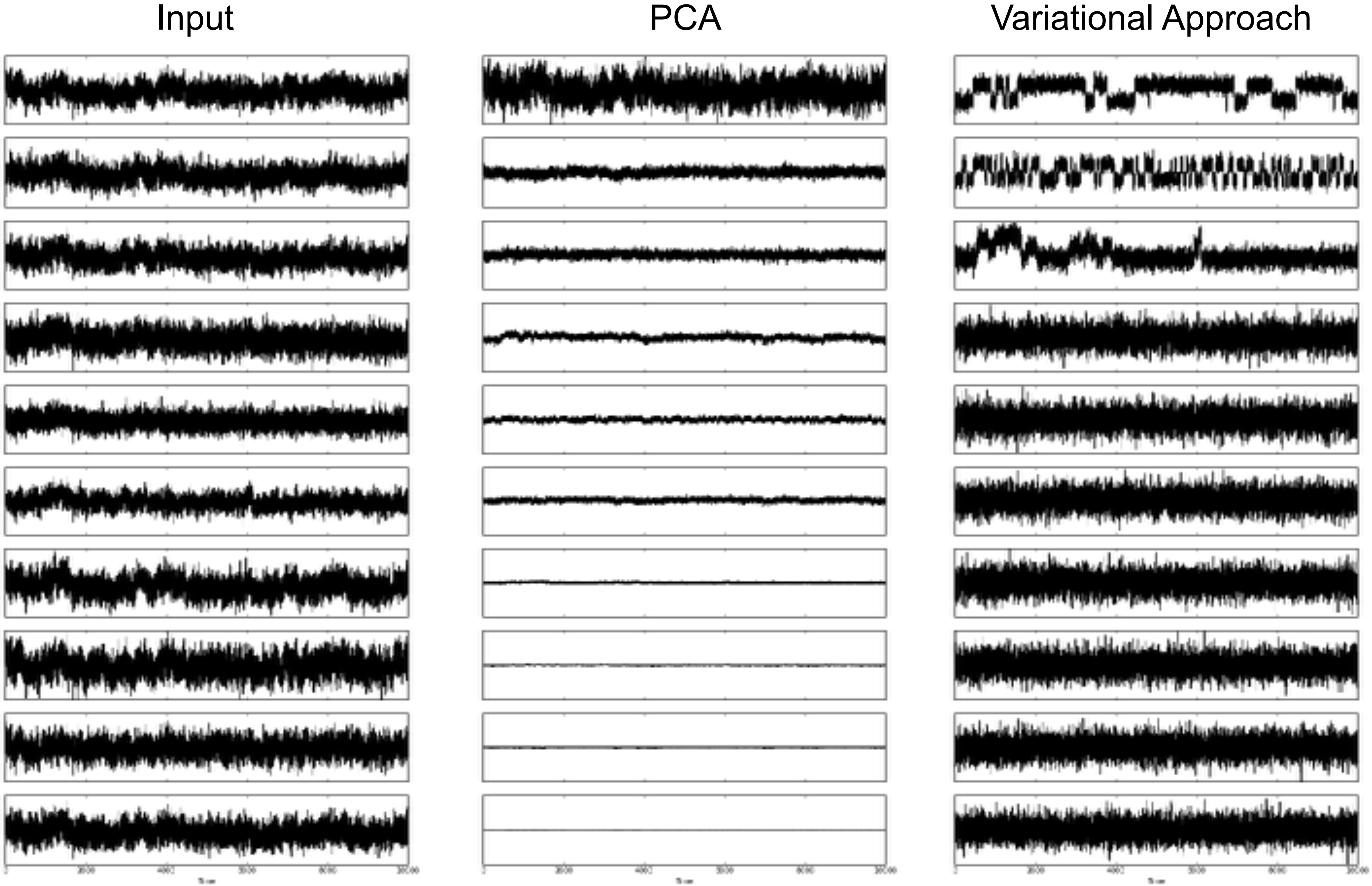


0
2000
4000
6000
8000
10000
Time

PCA



0
2000
4000
6000
8000
10000
Time



Input

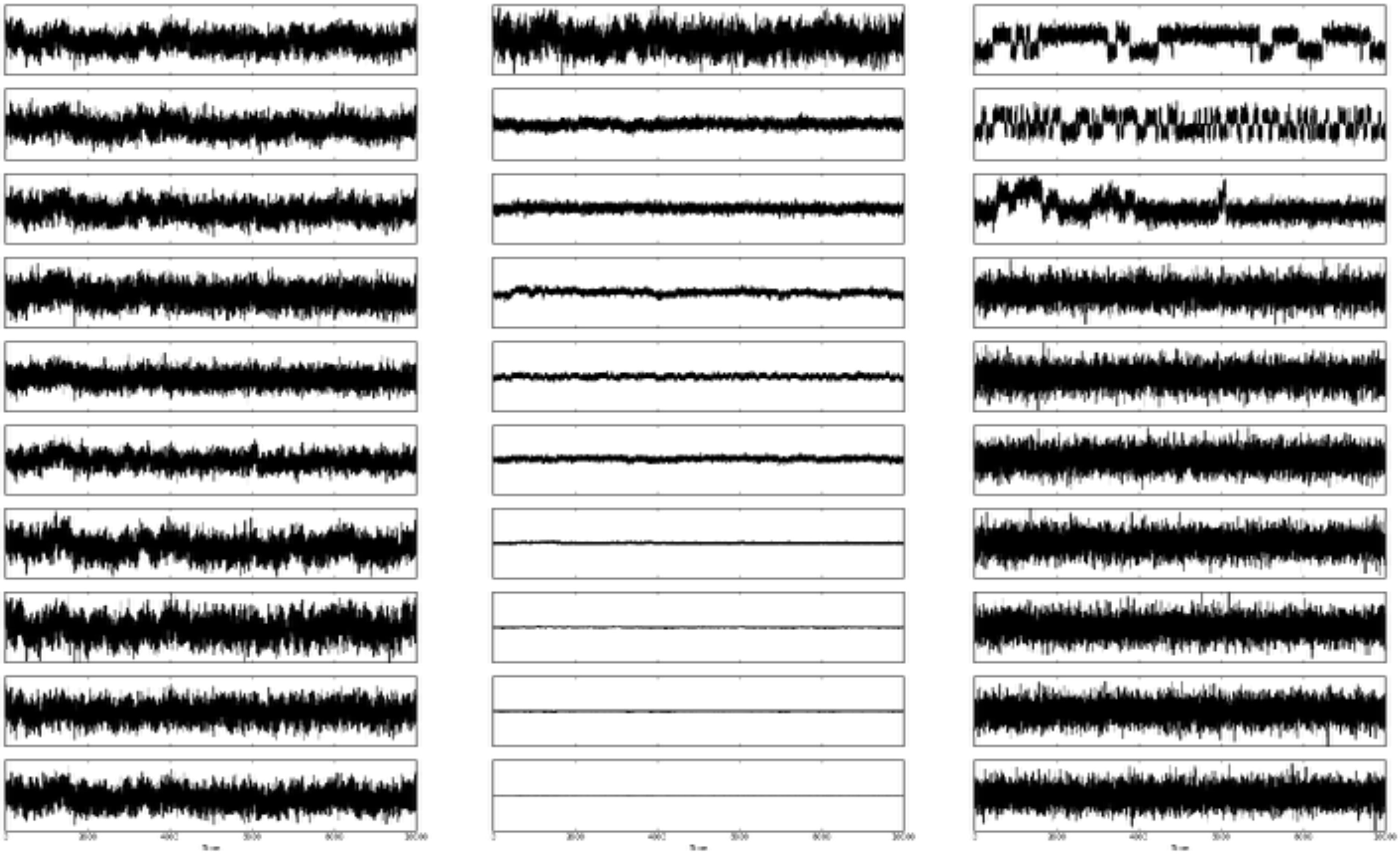
PCA

Variational Approach

Variational Approach

Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)

Perez-Hernandez et al, **JCP**, 139, 1502 (2013)
Identification of slow molecular order parameters for Markov model construction



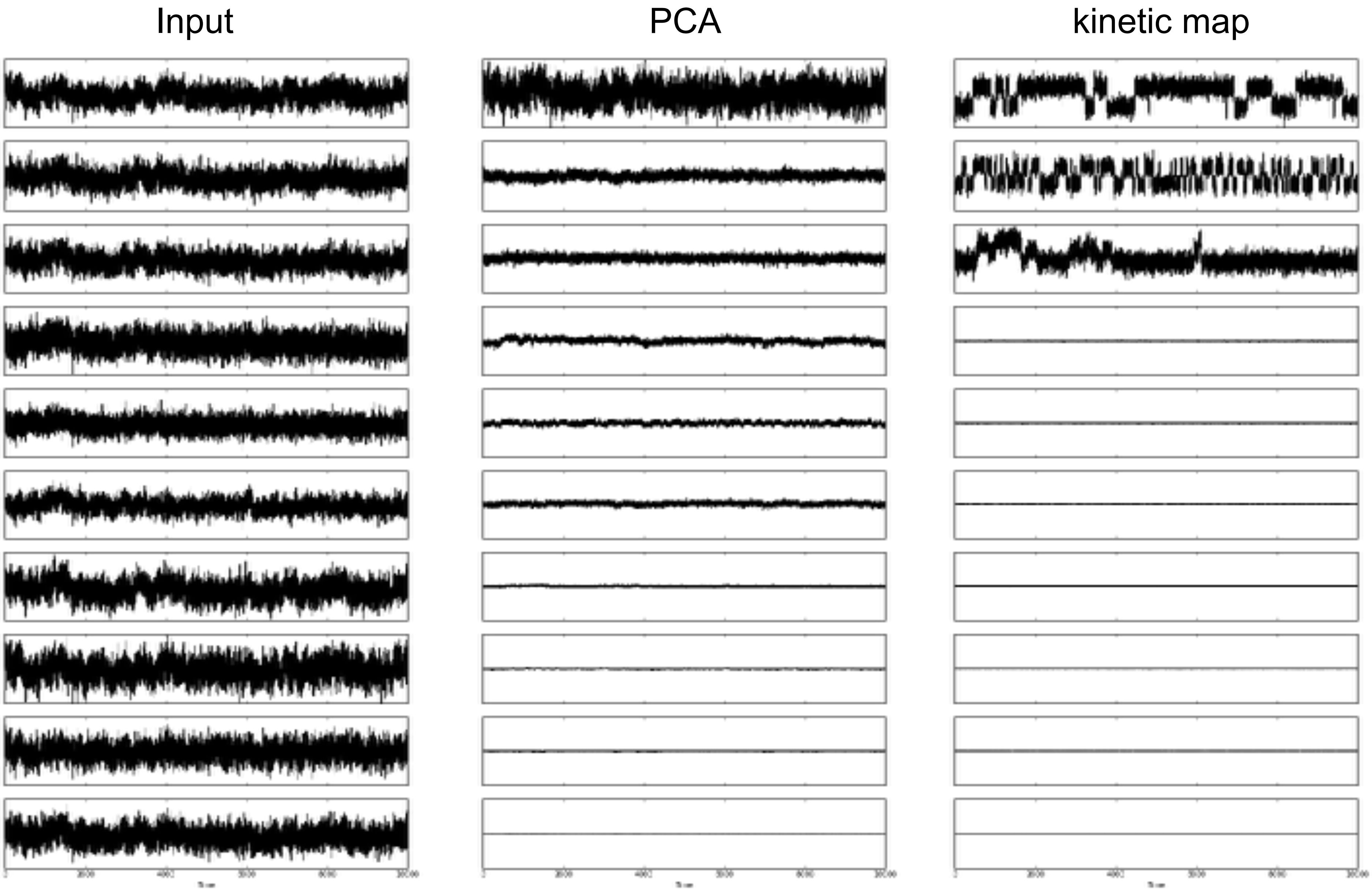
Input

PCA

Variational Approach

Variational Approach

Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)



Input

PCA

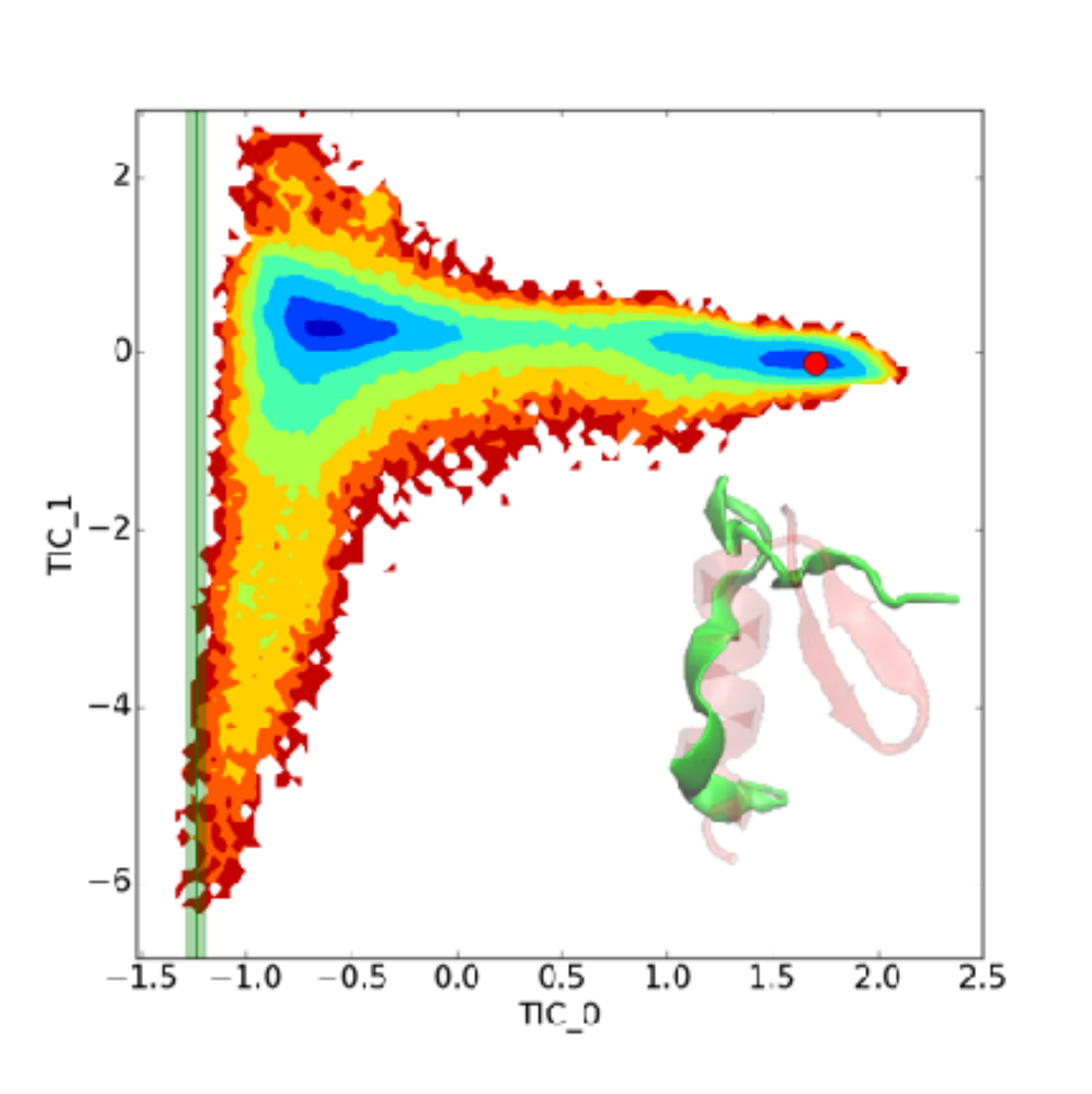
kinetic map

Variational Approach

Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)

Kinetic map:

Noé and Clementi, **JCTC** 11, 5002-5011 (2015)

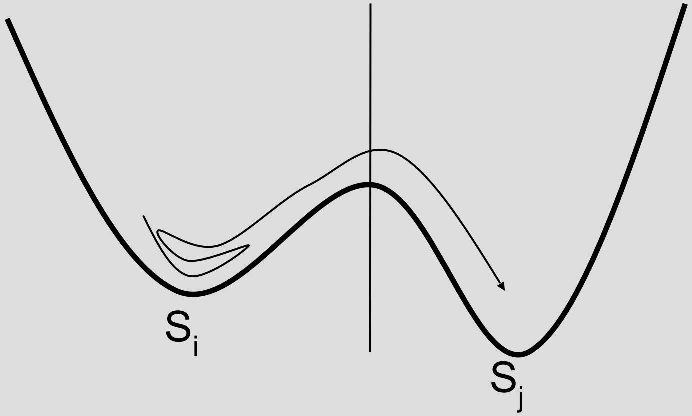


1FME peptide - Simulation data from DESRES, Lindorff-Larsen et al, Science 2011

Step 2: MSM estimation

Estimation of transition matrix

$$T_{ij}(\tau) = \frac{\mathbb{E}[\chi_i(\mathbf{x}(t))\,\chi_j(\mathbf{x}(t+\tau))]}{\mathbb{E}[\chi_i(\mathbf{x}(t))]} = \frac{c_{ij}^{\mathrm{corr}}(\tau)}{\pi_i},$$



Estimation:

Prinz et al.: **J. Chem Phys.** 134, 174105 (2011)
Bowman et al.: **J. Chem Phys.** 131, 124101 (2009)
Noé, **J Chem Phys** 128, 244103 (2008)

Statistical Error

$$p(Y|T) = \prod_{k=1}^{n-1} T_{y_k,y_{k+1}} = p(C|T) = \prod_{i,j=1}^{m} T_{ij}^{c_{ij}}$$

Linear Error Perturbation:

Sinhal, Pande, JCP 2006
Prinz, Smith, Noé, **Multiscale Model. Simul** 2011

Monte Carlo

Noé, **J Chem Phys** 128, 244103 (2008)
Chodera, Noé, **J Chem Phys** (2010)

Step 3: Analysis

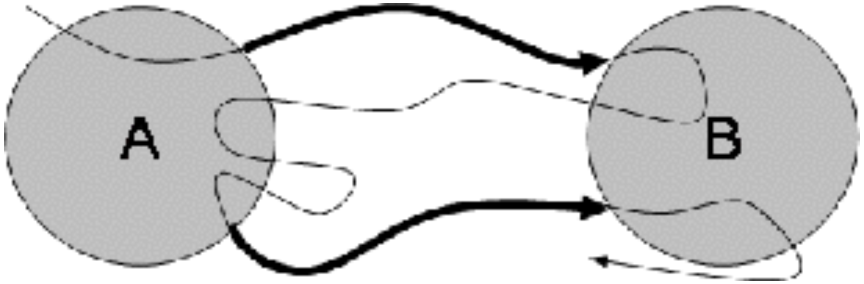
Transition path theory

Stationary probability

$$\pi^T = \pi^T \mathbf{T}(\tau).$$

Committor

$$-q_i^+ + \sum_{k\in I} T_{ik} q_k^+ = -\sum_{k\in B} T_{ik}.$$



Flux

$$f_{ij} = \pi_i q_i^- T_{ij} q_j^+.$$

$$f_{ij}^+ = \max\{0, f_{ij} - f_{ji}\}.$$

Metzner, Vanden-Eijnden, Schütte, **MMS** (2009)

Noé et al, **PNAS** (2009)

Bereszkovskii, Hummer, Szabo, **JCP** (2009)

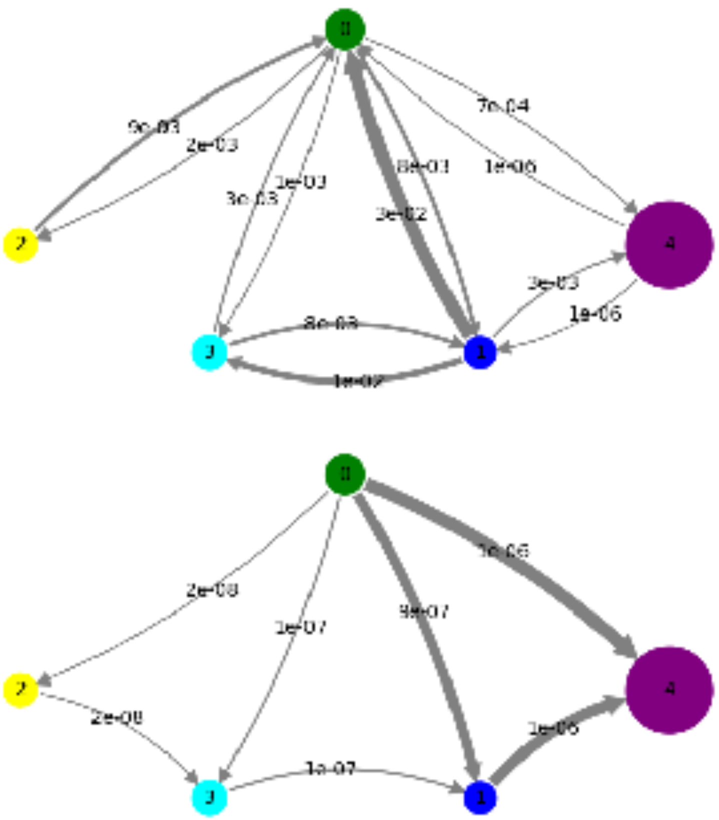
Metastable states (PCCA)

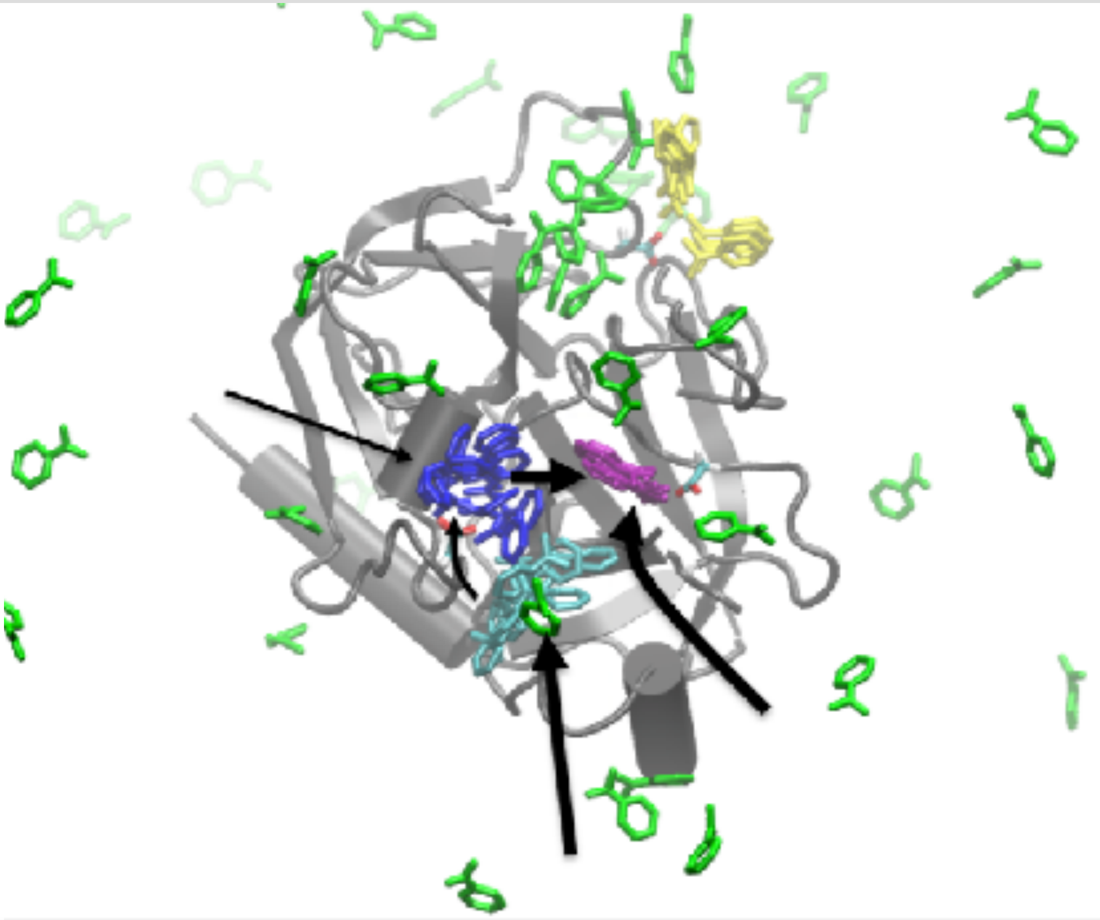
Deuflhard, Weber.: **Linear Alg. Appl.** 398C, 161 (2005)

Experimental observables

Noé et al, **PNAS** 108, p 4822 (2011)
Lindner et al, **JCP** 139, 175102 (2013)

Step 4: Coarse-graining

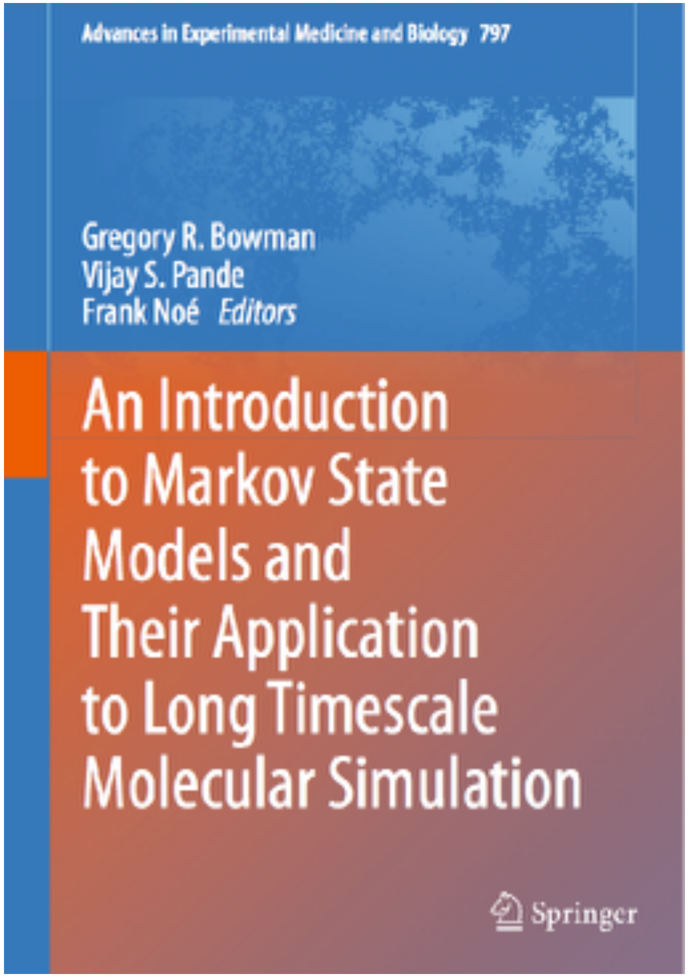




Scherer et al. **JCTC** 11, 5525–5542 (2015).

Markov State Models

Review book



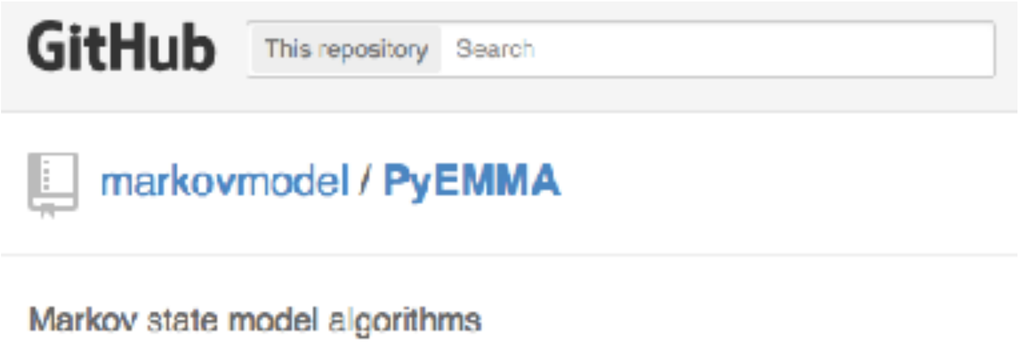
Advances in Experimental Medicine and Biology 797

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

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

Springer

PyEMMA software



markovmodel / PyEMMA

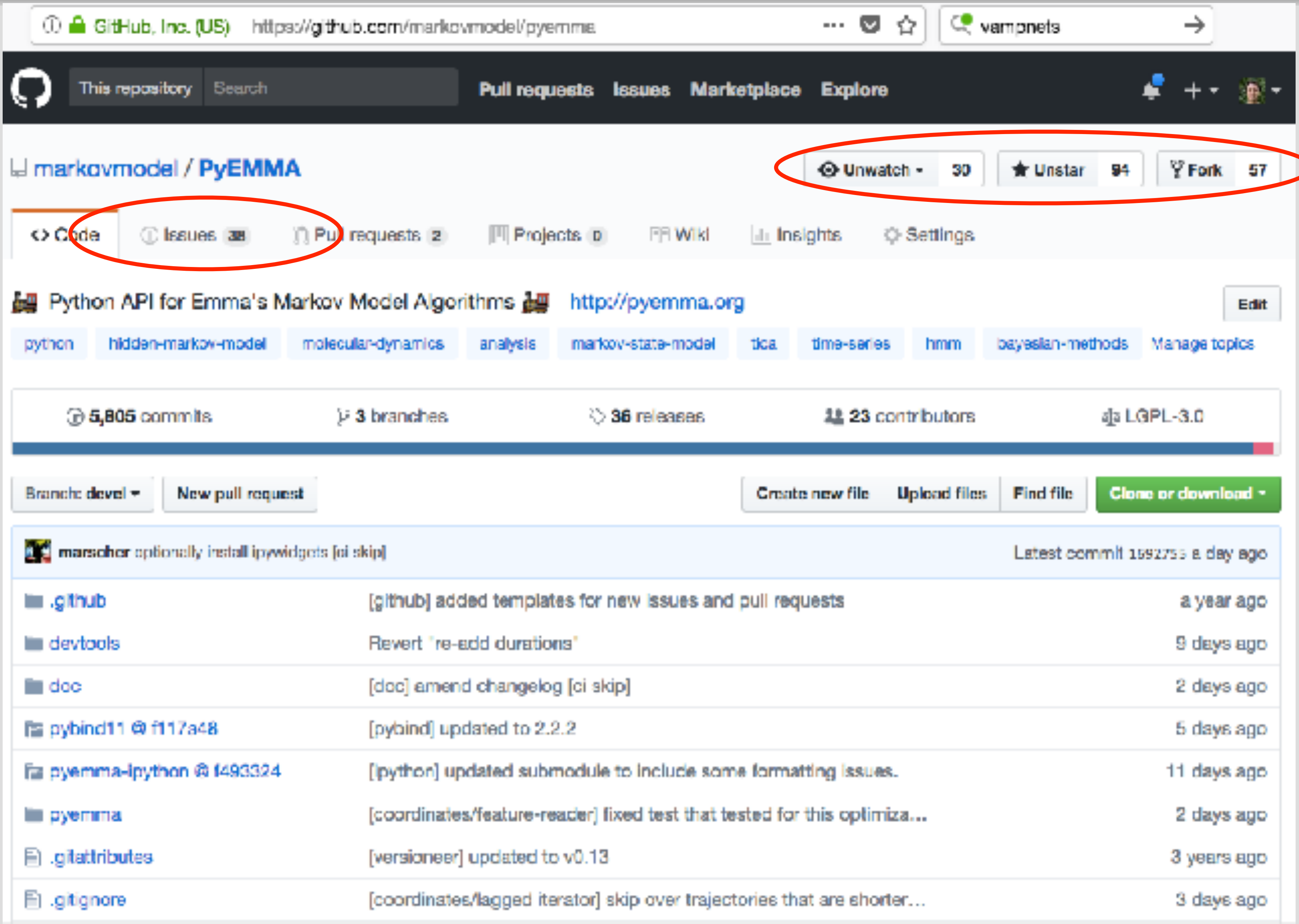
Markov state model algorithms

code: www.github.com/markovmodel

docs: www.pyemma.org

M. K. Scherer, B. Trendelkamp-Schroer, F. Paul, G. Pérez-Hernández, M. Hoffmann, N. Plattner, C. Wehmeyer, J.-H. Prinz, and F. Noé, “PyEMMA 2: A software package for estimation, validation, and analysis of Markov models,” **J. Chem. Theory Comput.** 11, 5525–5542 (2015)

PyEMMA github site



GitHub, Inc. (US) https://github.com/markovmodel/pyemma

vampnets

markovmodel / **PyEMMA**

Python API for Emma's Markov Model Algorithms http://pyemma.org

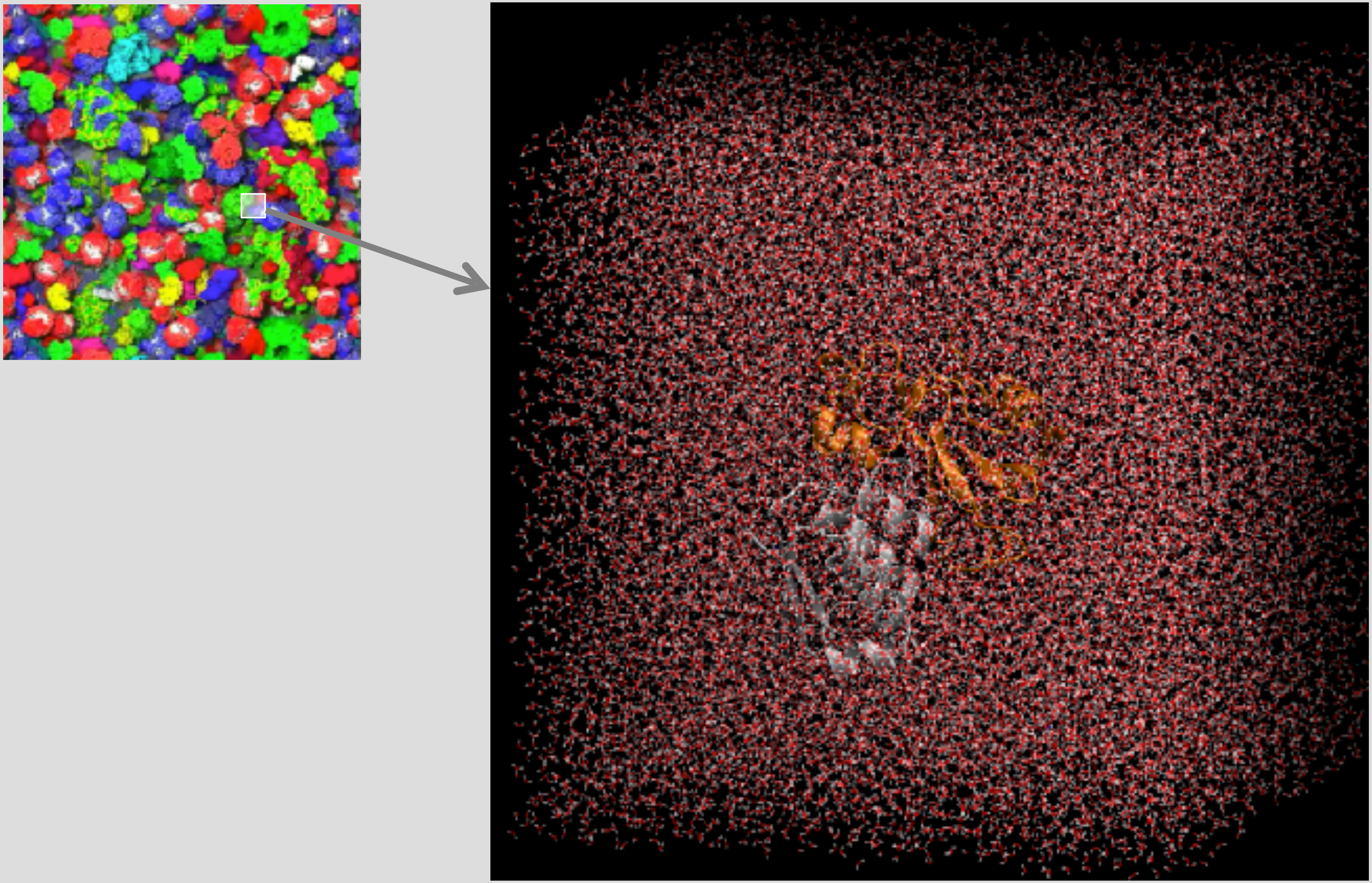
5,805 commits · 3 branches · 36 releases · 23 contributors · LGPL-3.0

marscher optionally install ipywidgets [ci skip] — Latest commit 1592755 a day ago

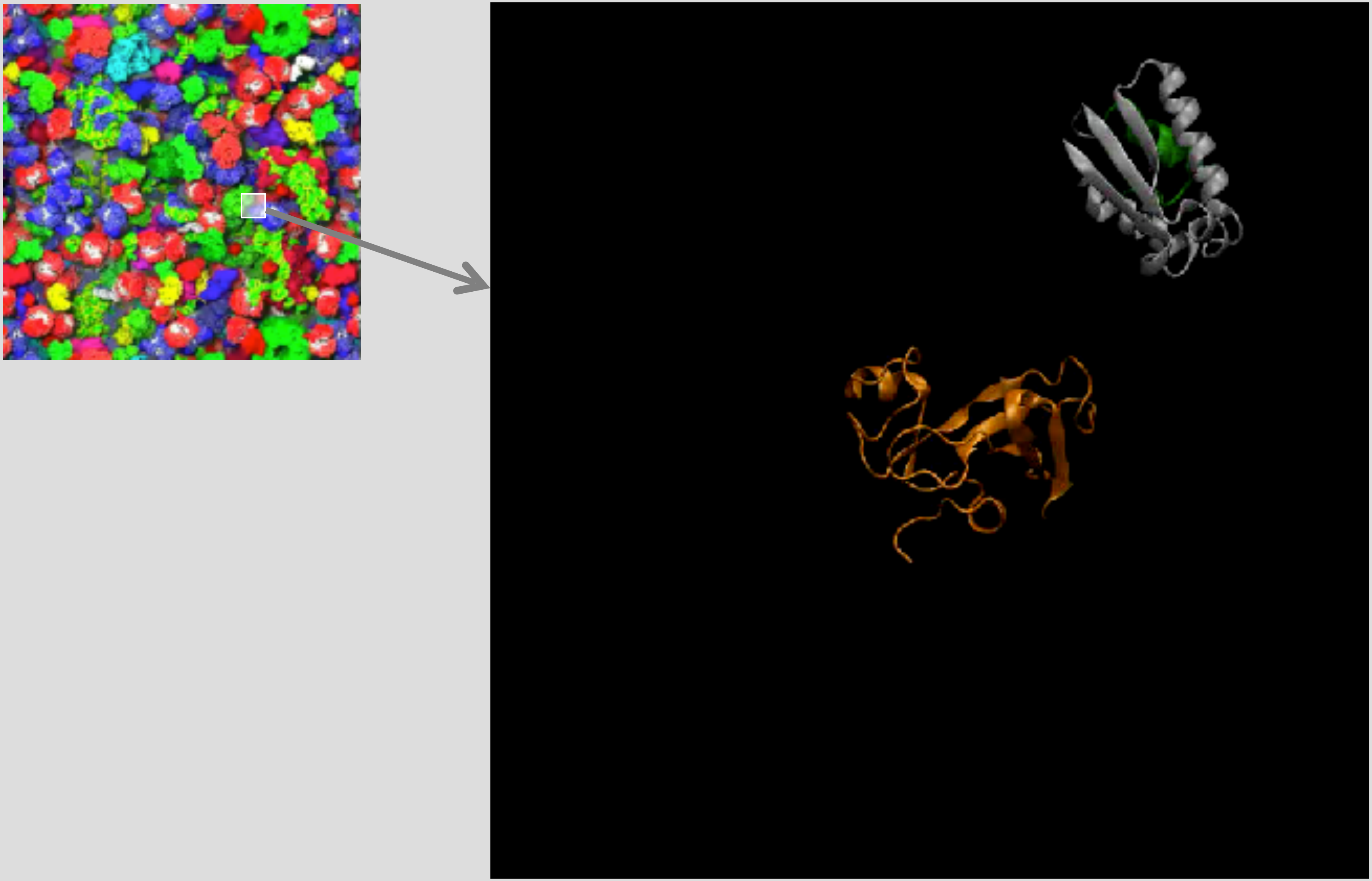
.github	[github] added templates for new issues and pull requests	a year ago
devtools	Revert "re-add durations"	9 days ago
doc	[doc] amend changelog [ci skip]	2 days ago
pybind11 @ f117a48	[pybind] updated to 2.2.2	5 days ago
pyemma-ipython @ f493324	[ipython] updated submodule to include some formatting issues.	11 days ago
pyemma	[coordinates/feature-reader] fixed test that tested for this optimiza…	2 days ago
.gitattributes	[versioneer] updated to v0.13	3 years ago
.gitignore	[coordinates/lagged iterator] skip over trajectories that are shorter…	3 days ago

Application to protein-protein association

Protein-Protein binding



Protein-Protein binding

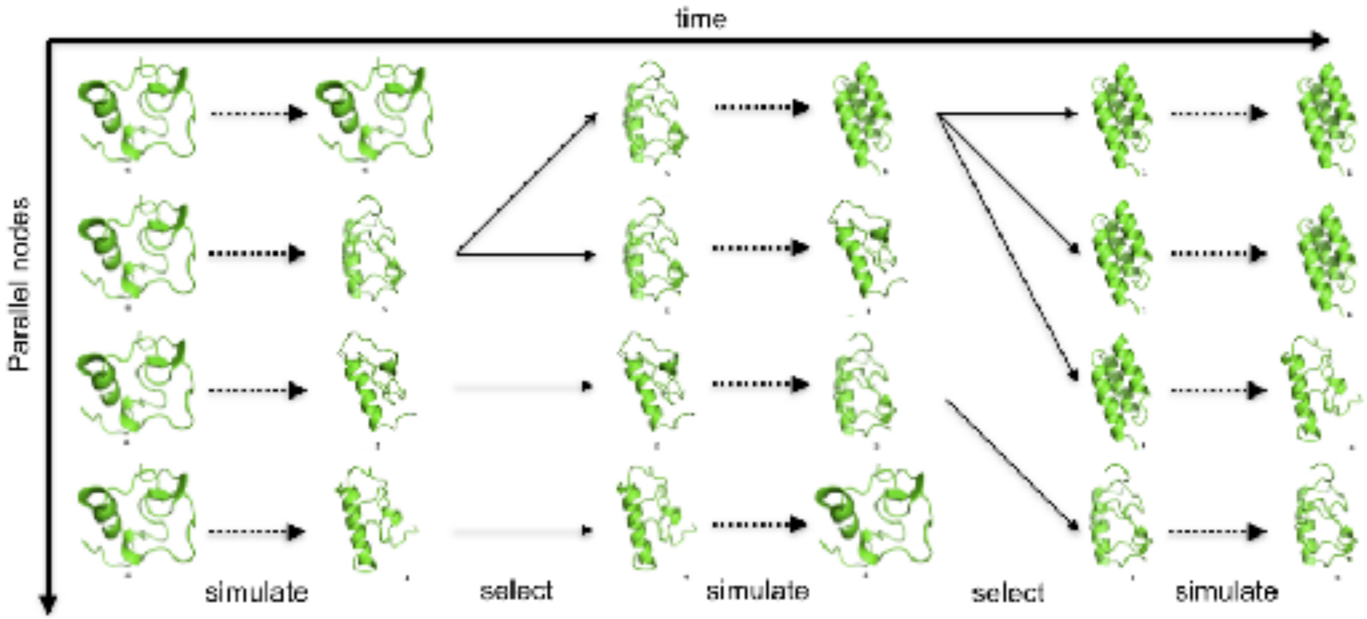


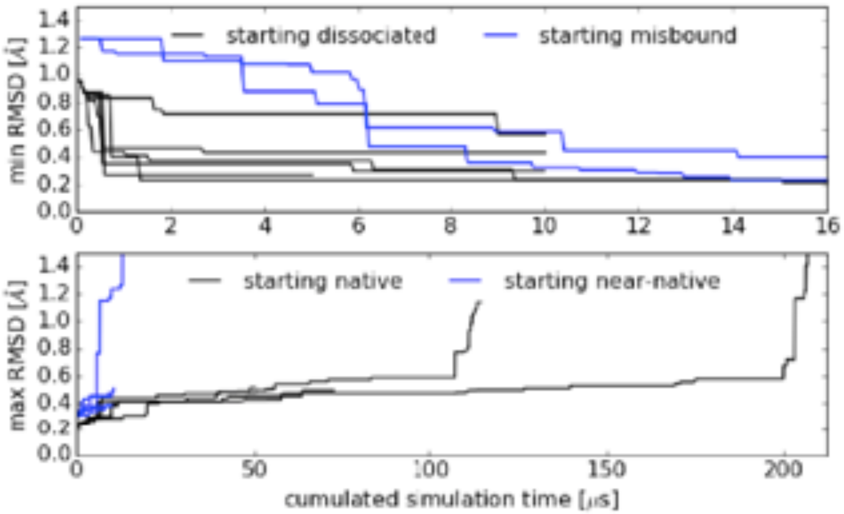
Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry (2017)

0.1 microseconds

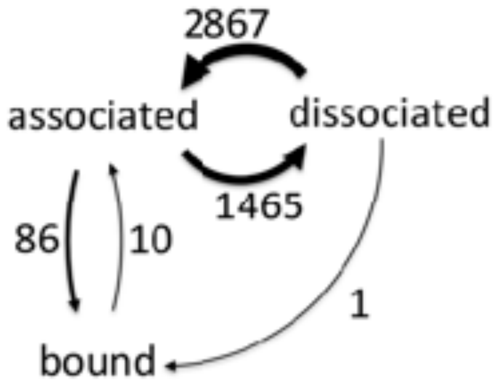
1) Adaptive molecular dynamics

Prototype: **github.com/markovmodel/adaptivemd**





2 ms simulation time total



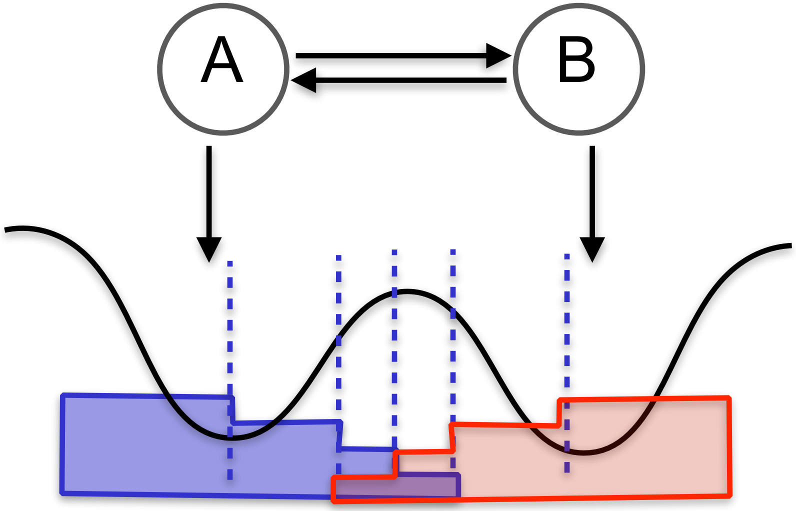
Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry (2017)

2) Dimension reduction (10000 => 10) using variational approach

3) Discretization using k-means

4) Hidden Markov model based on microstates

Noé et al, JCP 139, 184114 (2013)

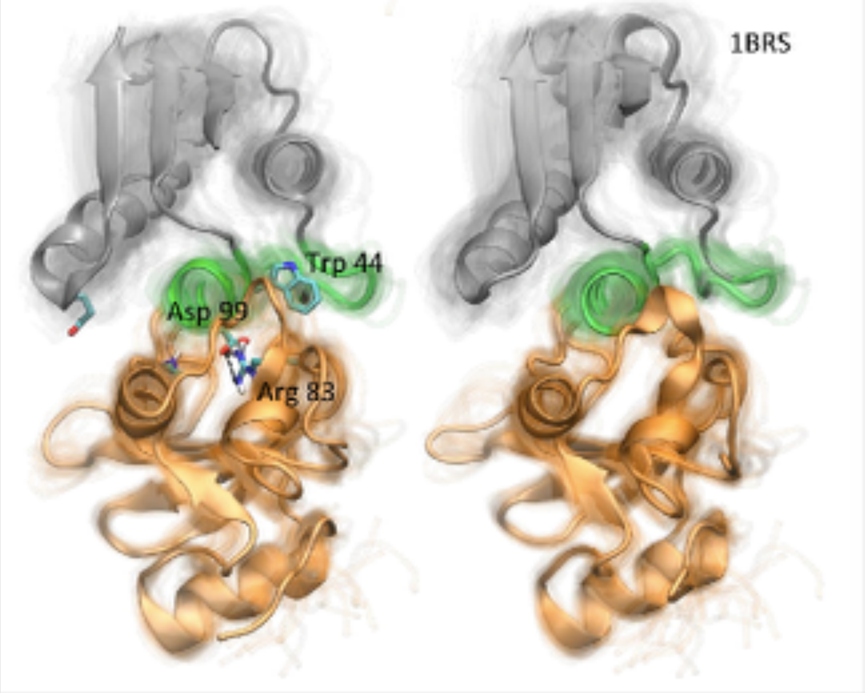


Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry (2017)

Validation of the model

- crystal structure 1BRS predicted by the most stable HMM state (95% population)

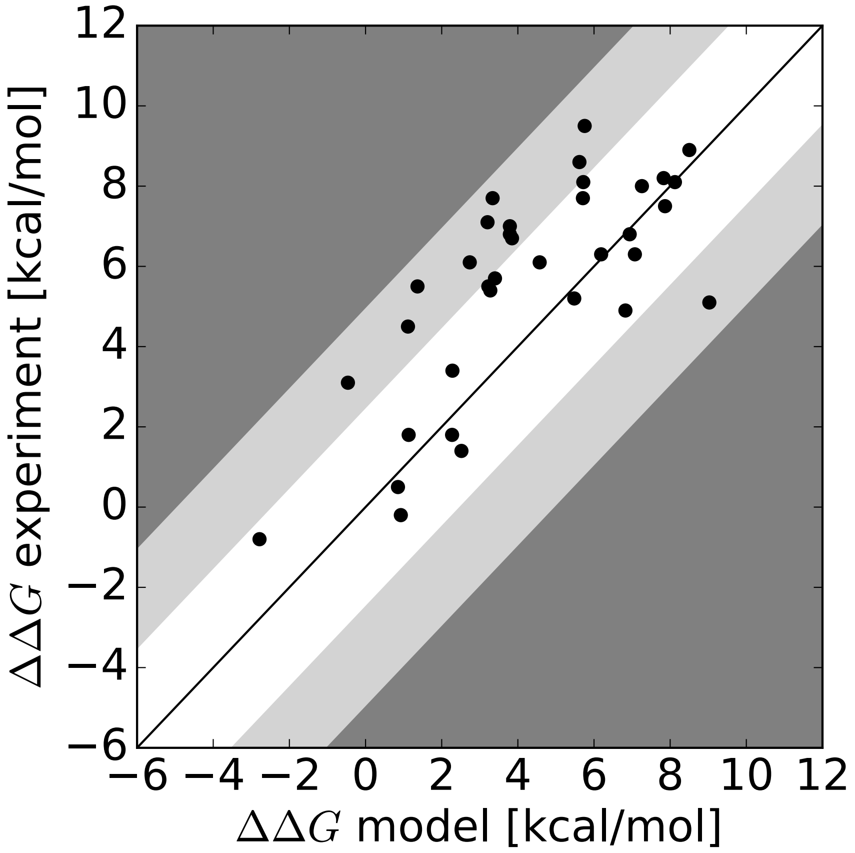
  average heavy-atom RMSD 2.1 A

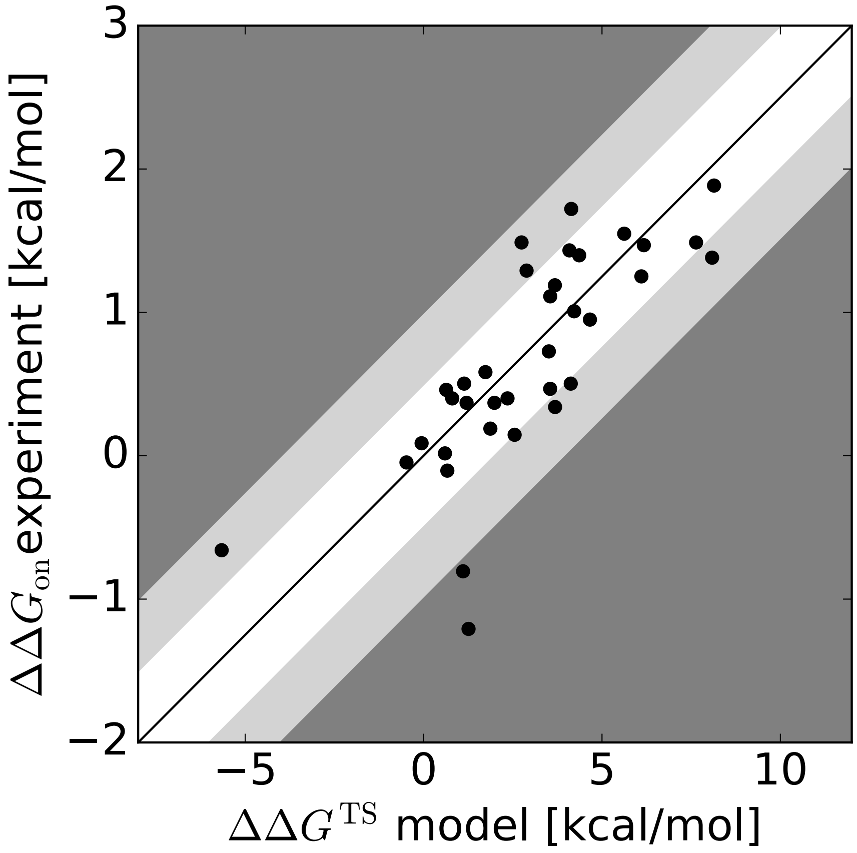


	Model	95% confidence interval	Experiment
• **Binding free energy**	**14.8** kcal / mol	(12.3 … 19.3)	**16.8** kcal/mol
• **Association rate**	**0.74 10^8** $s^{-1}M^{-1}$	(0.72 … 0.75)	**$1\cdot10^8$** $s^{-1}M^{-1}$
• **Dissociation rate**	**2.7 10^{-3}** s^{-1}	($2.8\cdot10^{-6}$ … $1.8\cdot10^{-1}s^{-1}$)	(**$4.8\cdot10^{-5}$** s^{-1} … **$5.0\cdot10^{-4}$** s^{-1})

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

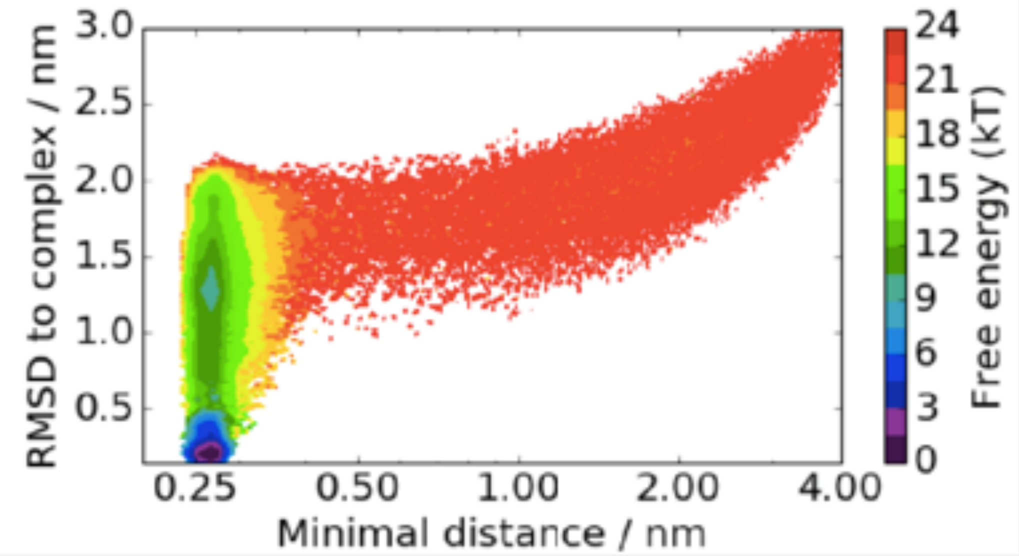
Mutants by first-order perturbation theory

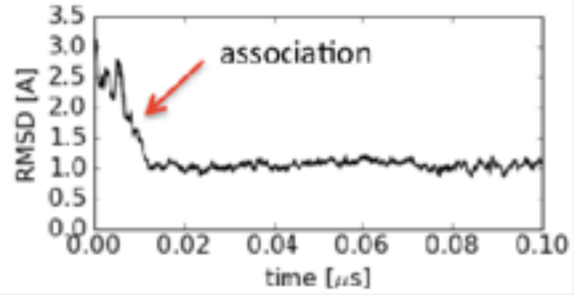


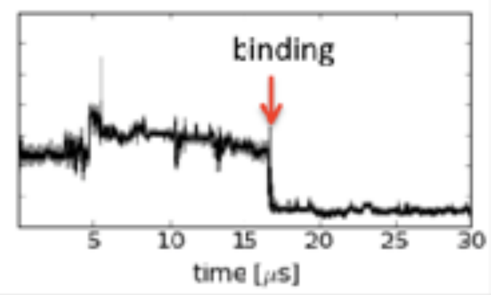


Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry (2017)

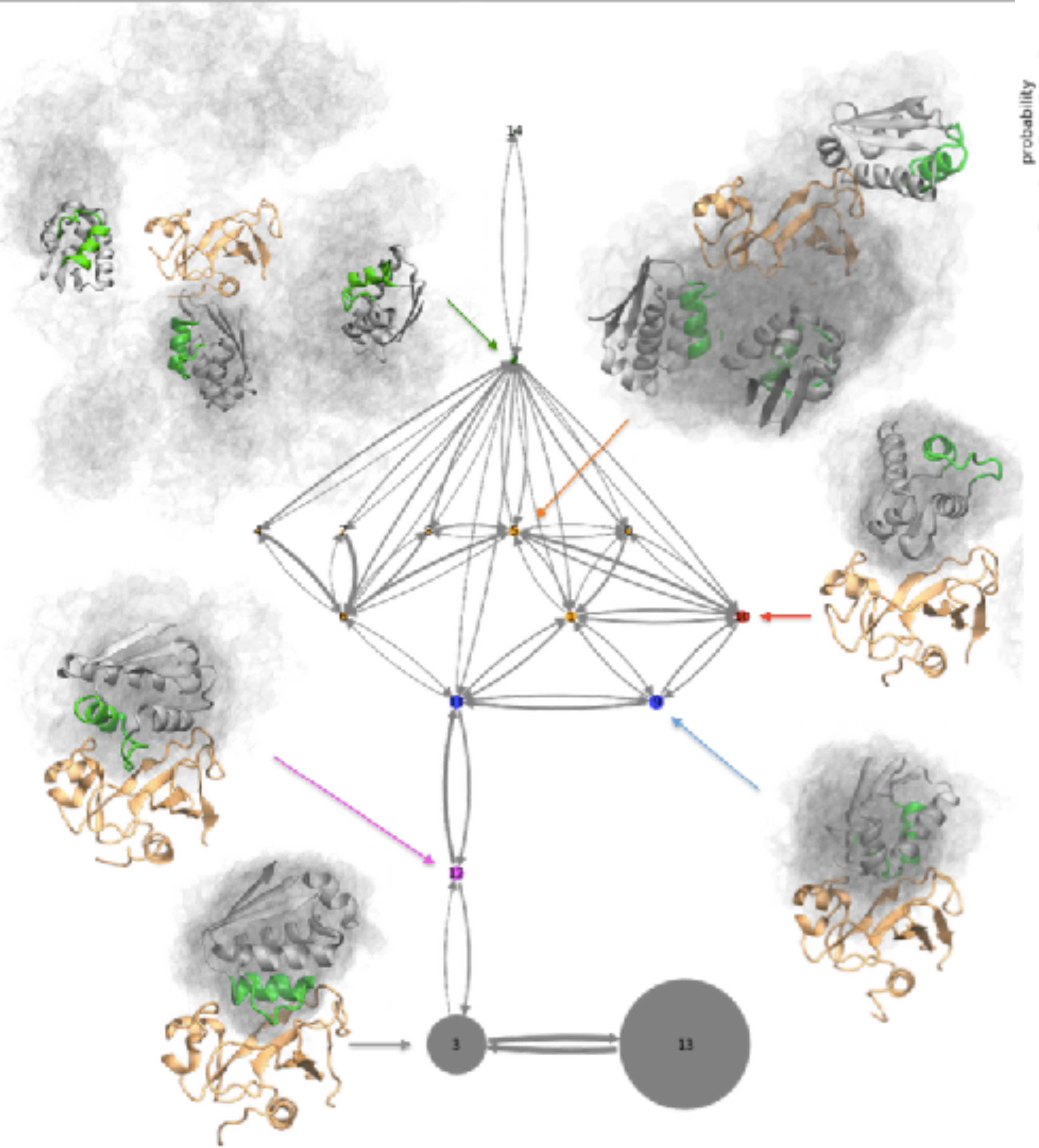
Estimation (Reversible Markov state model)

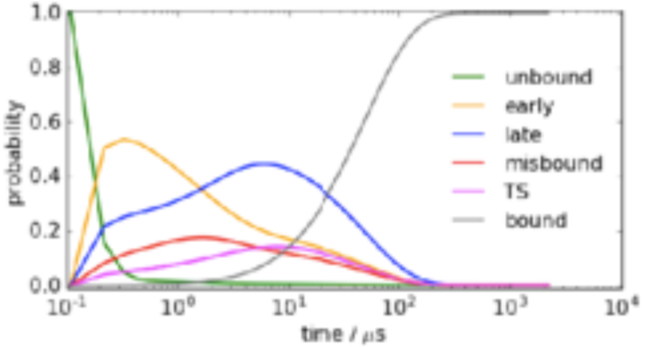






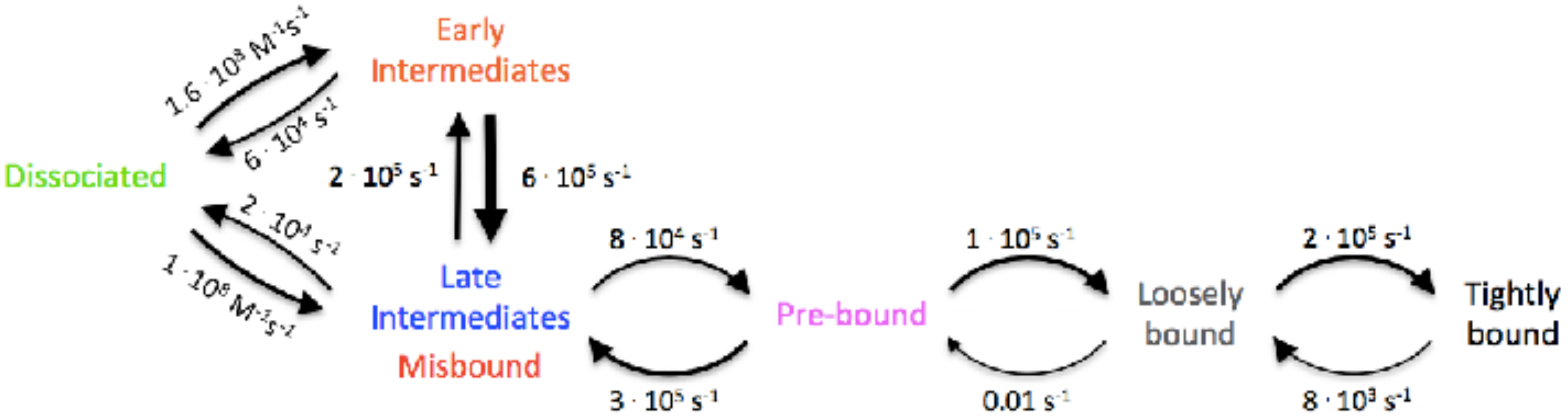
Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry (2017)





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

Coarse-grained model



Dissociated

$1.6 \cdot 10^8\ M^{-1}s^{-1}$

$6 \cdot 10^4\ s^{-1}$

Early Intermediates

$2 \cdot 10^5\ s^{-1}$

$6 \cdot 10^5\ s^{-1}$

$2 \cdot 10^3\ s^{-1}$

$1 \cdot 10^8\ M^{-1}s^{-1}$

Late Intermediates

Misbound

$8 \cdot 10^4\ s^{-1}$

$3 \cdot 10^5\ s^{-1}$

Pre-bound

$1 \cdot 10^5\ s^{-1}$

$0.01\ s^{-1}$

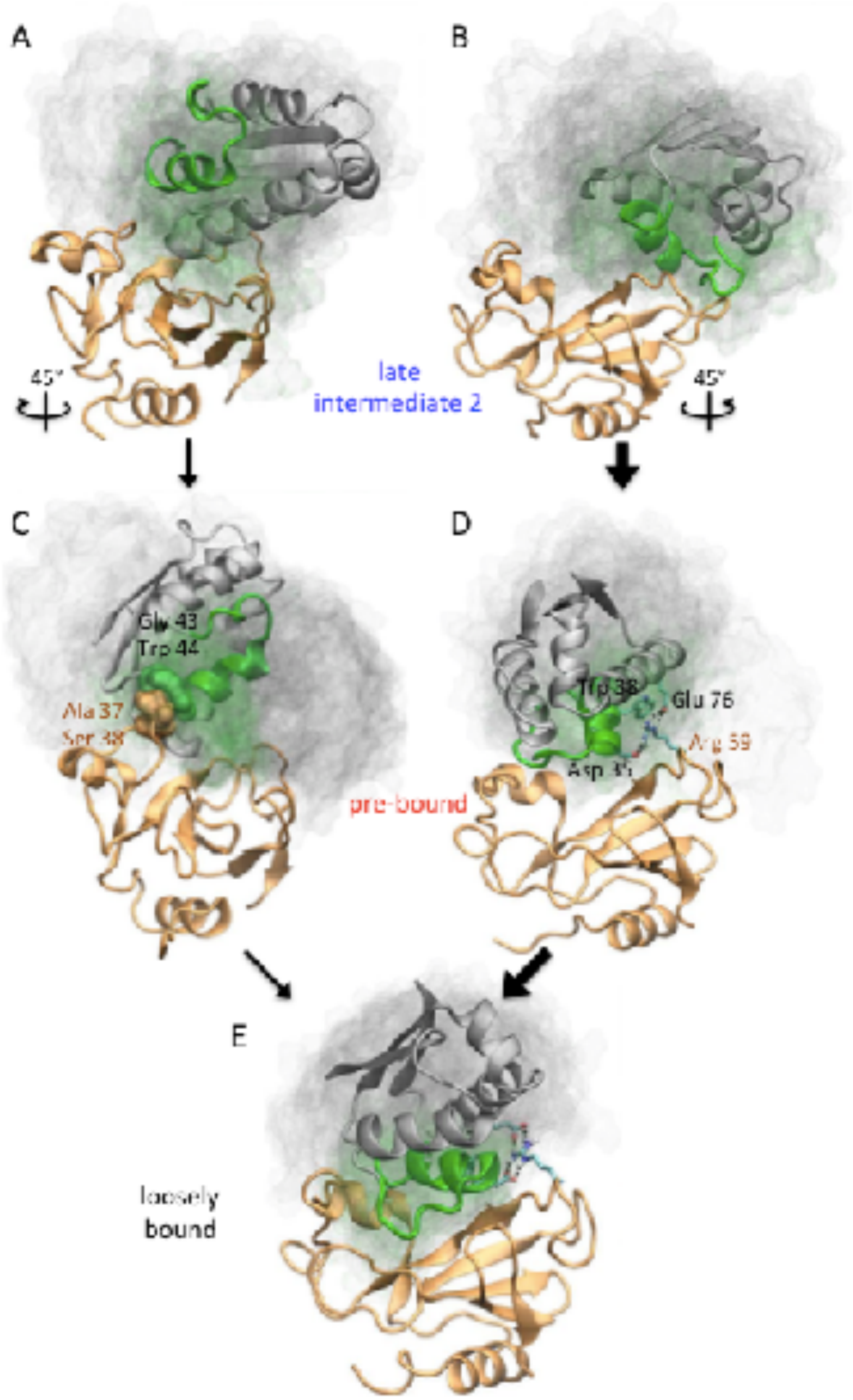
Loosely bound

$2 \cdot 10^5\ s^{-1}$

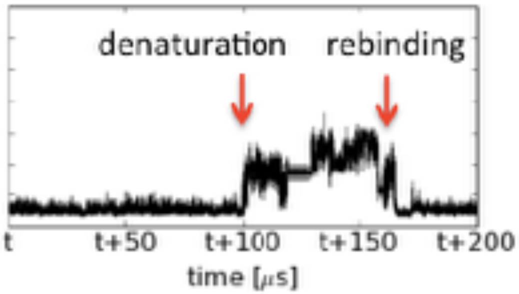
$8 \cdot 10^3\ s^{-1}$

Tightly bound

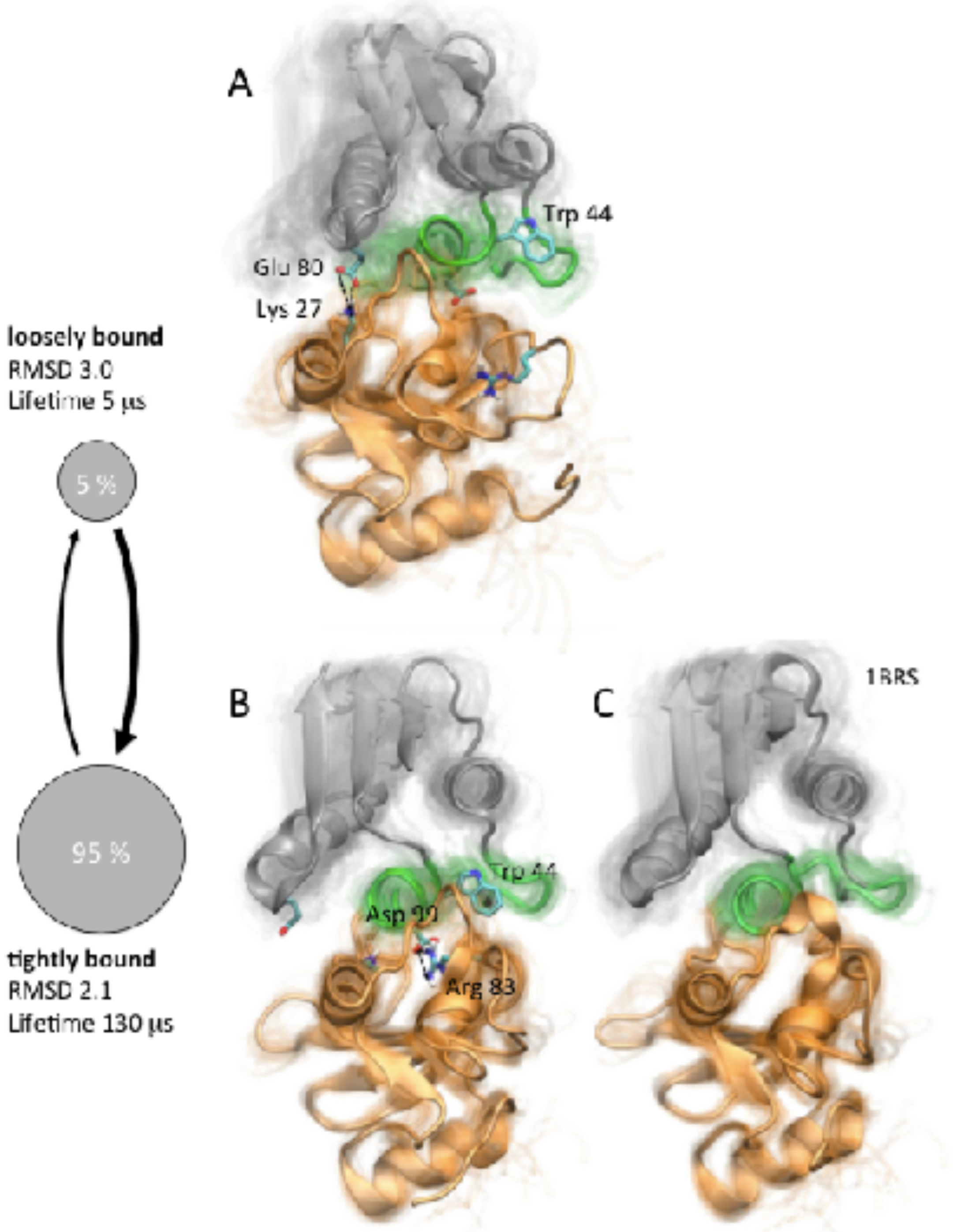
Plattner, Doerr, De Fabritiis, Noé
Nature Chemistry (2017)



Geminate rebinding



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



loosely bound
RMSD 3.0
Lifetime 5 µs

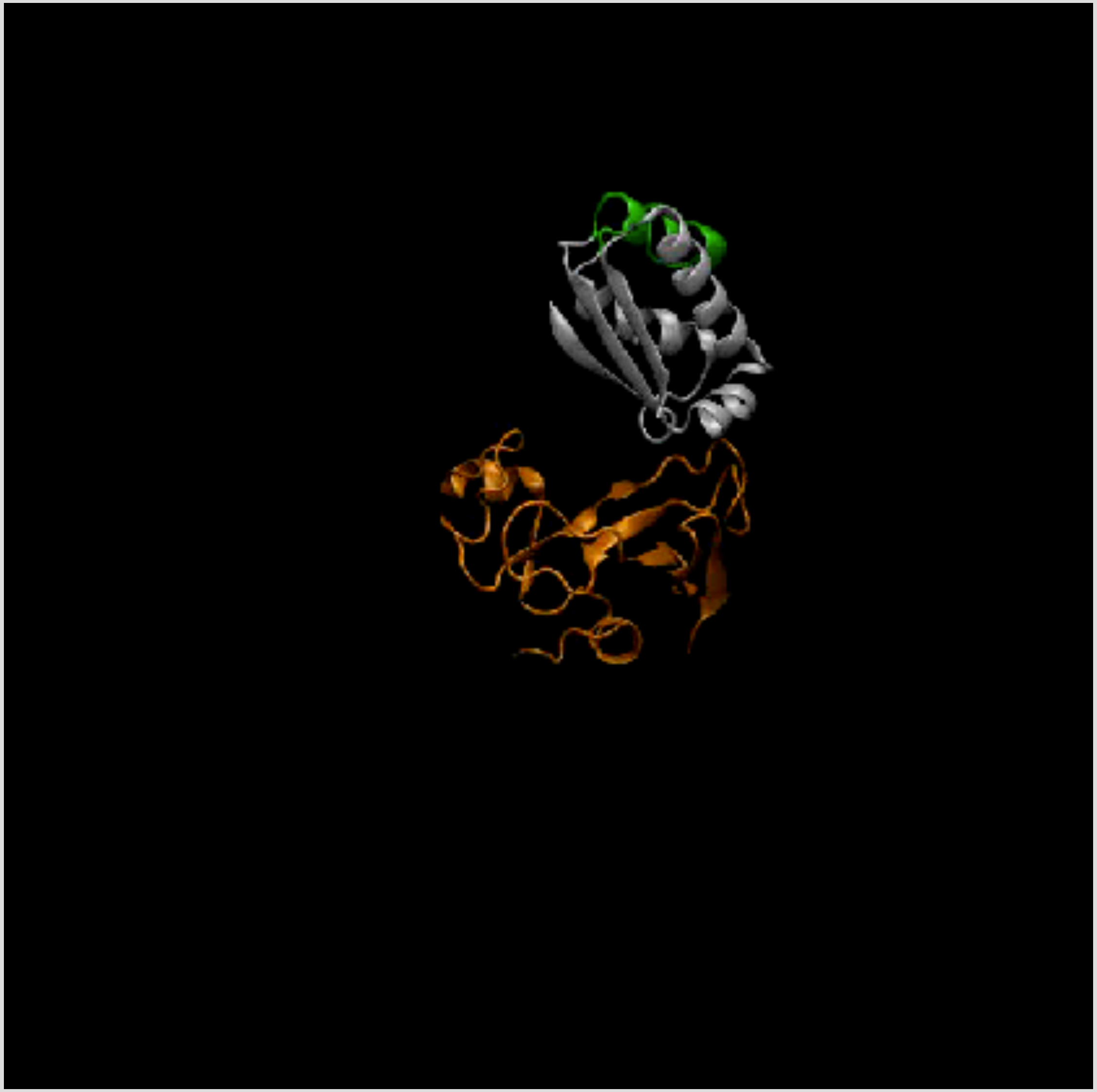
5 %

95 %

tightly bound
RMSD 2.1
Lifetime 130 µs

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

Protein-Protein binding

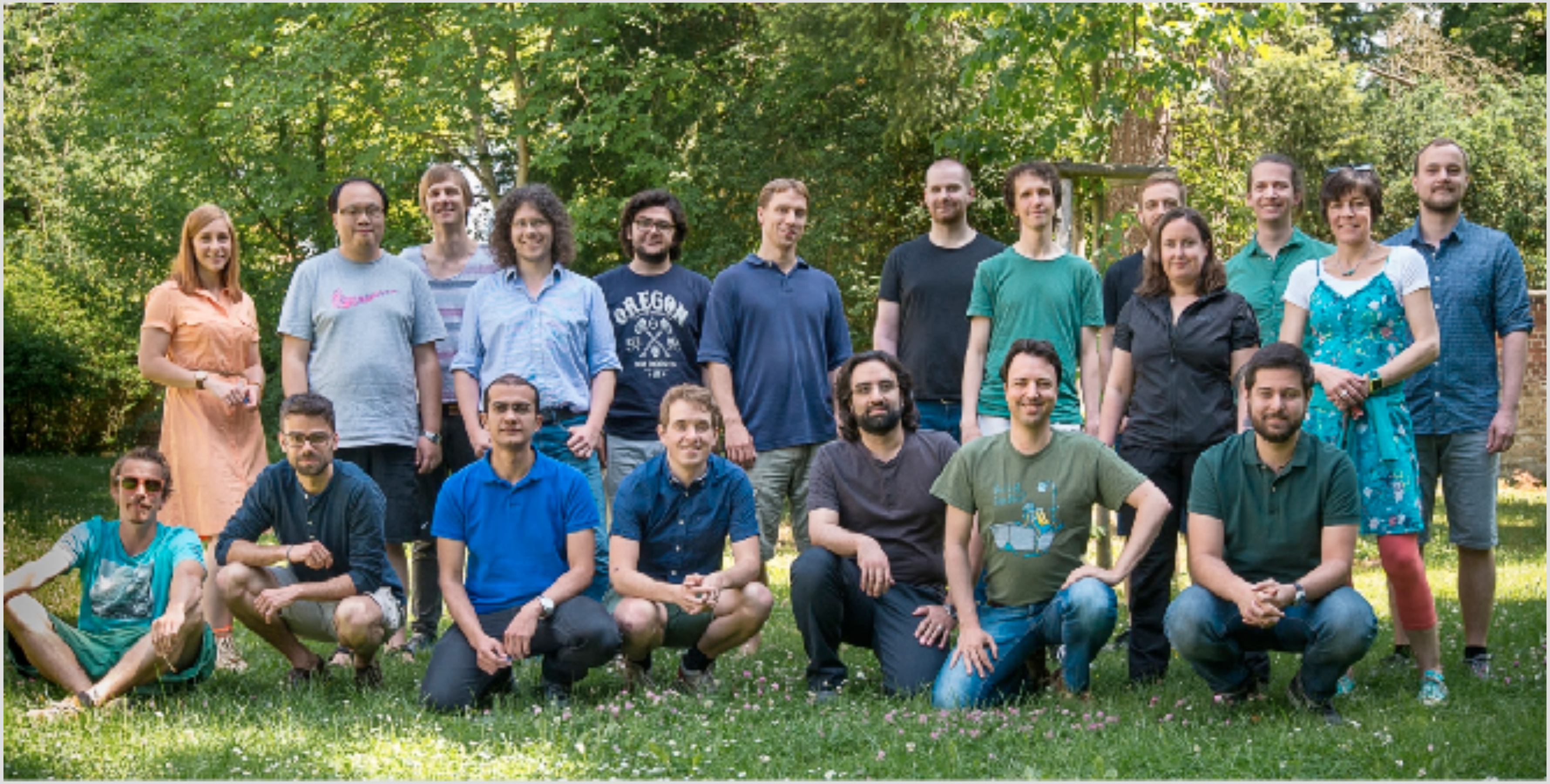


Plattner, Doerr, De Fabritiis, Noé

Nature Chemistry (in press)

0.1 milliseconds

Acknowledgements



Collaborations

Cecilia Clementi (Rice University)
Christof Schütte (FU Berlin)
Eric Vanden-Eijnden (Courant NY)
Thomas Weikl (MPI Potsdam)
Edina Rosta (King's College London)
Bettina Keller (FU Berlin)

Vijay Pande (Stanford)
Volker Haucke (FMP Berlin)
Stephan Sigrist (FU Berlin)
Oliver Daumke (MDC)
John Chodera (MSKCC NY)
Gianni de Fabritiis (Barcelona)

Funding

erc
European Research Council
Established by the European Commission

DFG
SFB 958
SFB 740

EINSTEIN Foundation.de