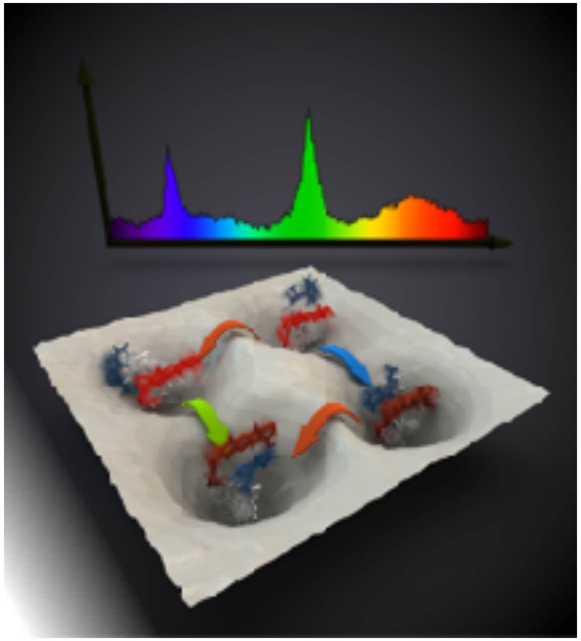
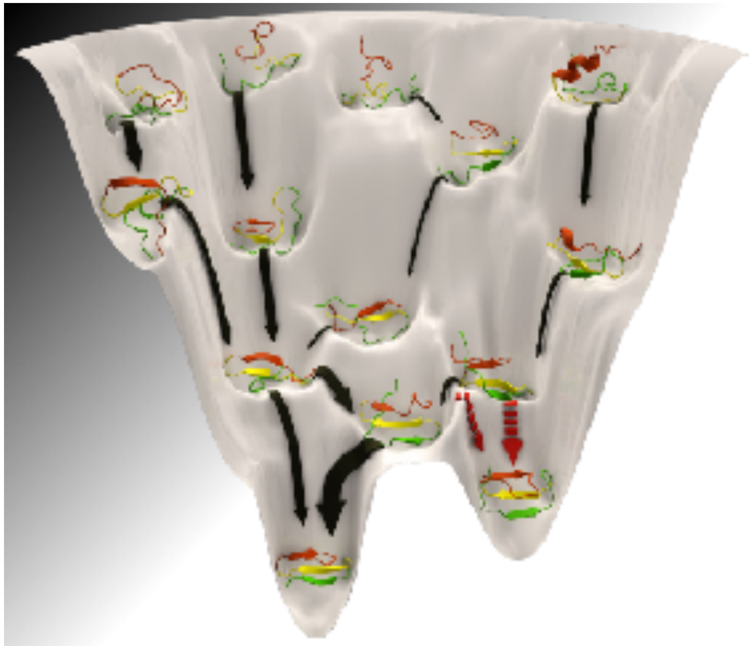
Accuracy of Markov State Models (MSMs) and Hidden Markov Models (HMMs)





Frank Noé (FU Berlin)
frank.noe@fu-berlin.de

Computational Molecular Biology

Freie Universität Berlin

Discretization

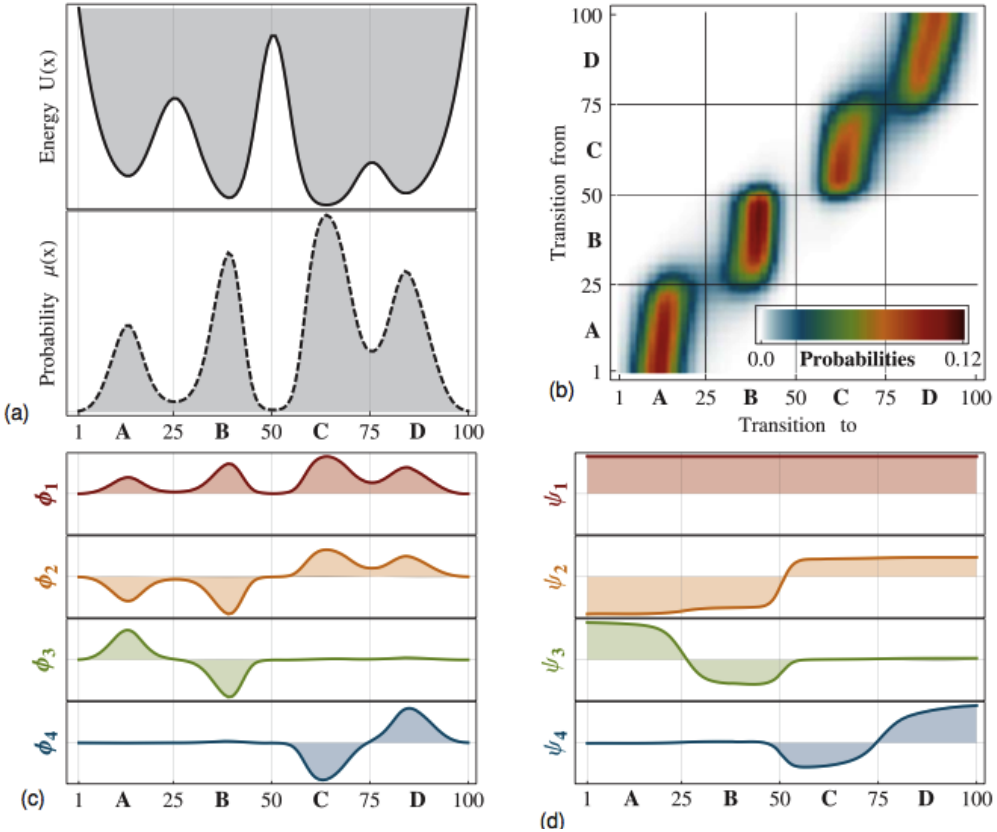
Backward propagator

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

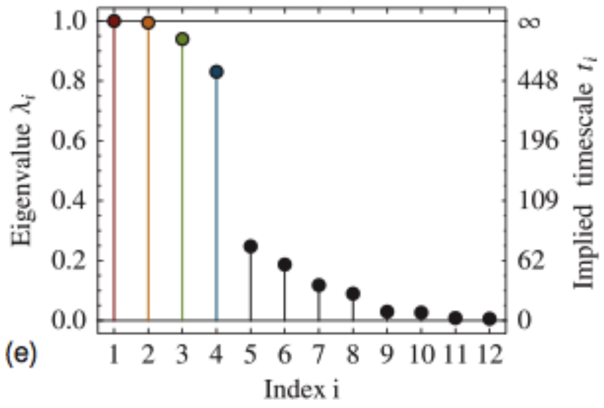
Spectral decomposition

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

Processes:

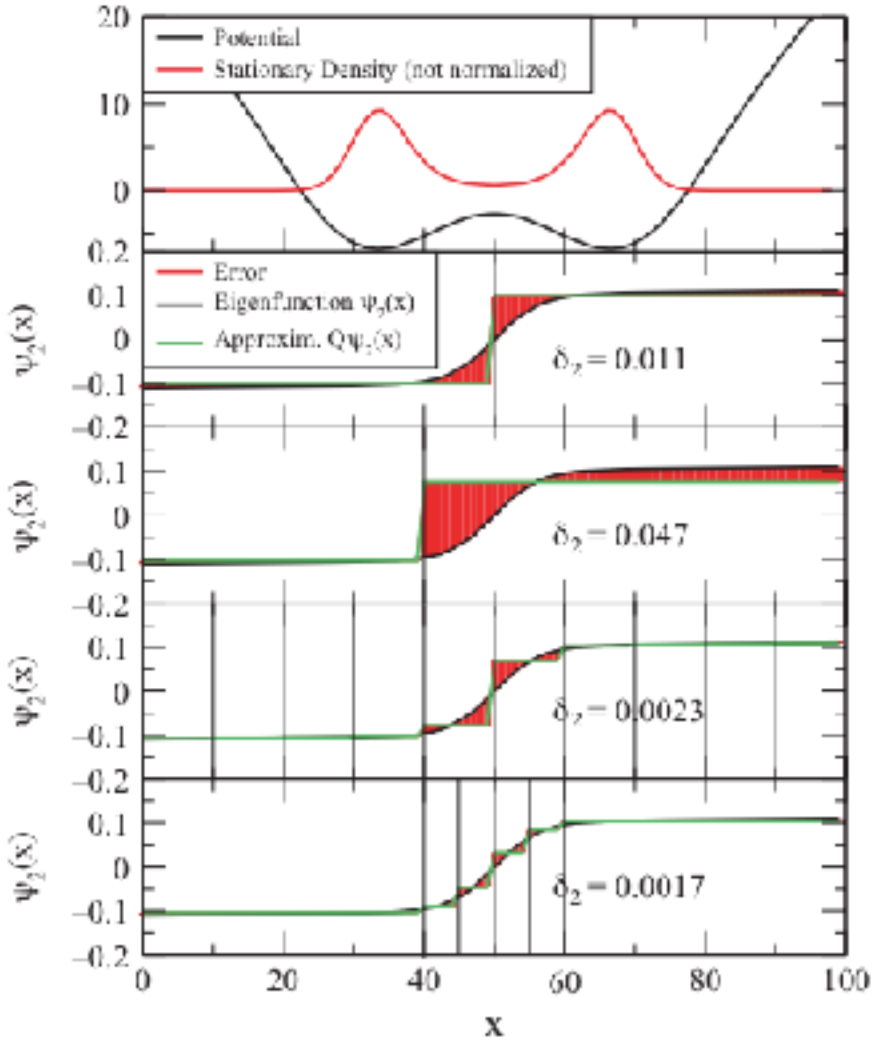


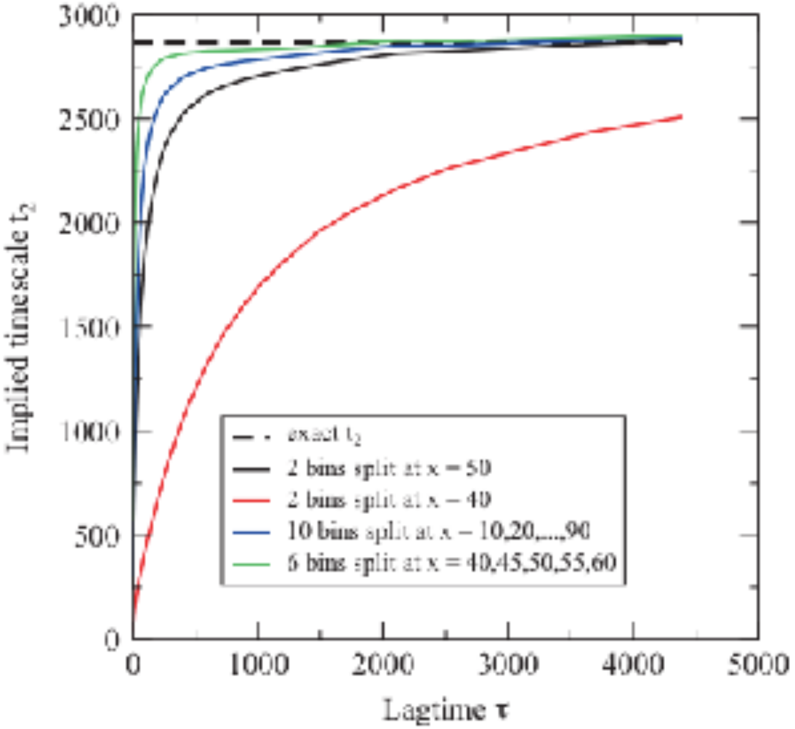
Eigenvalues / timescales κ_i^{-1}



Prinz et al, **JCP 134**, 174105 (2011)

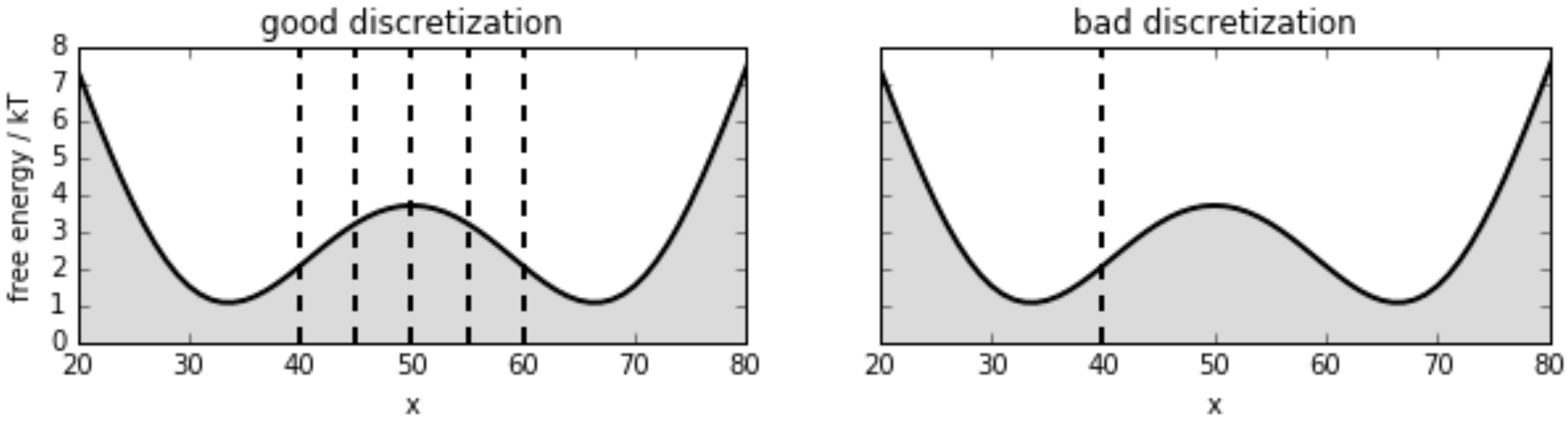
Discretization



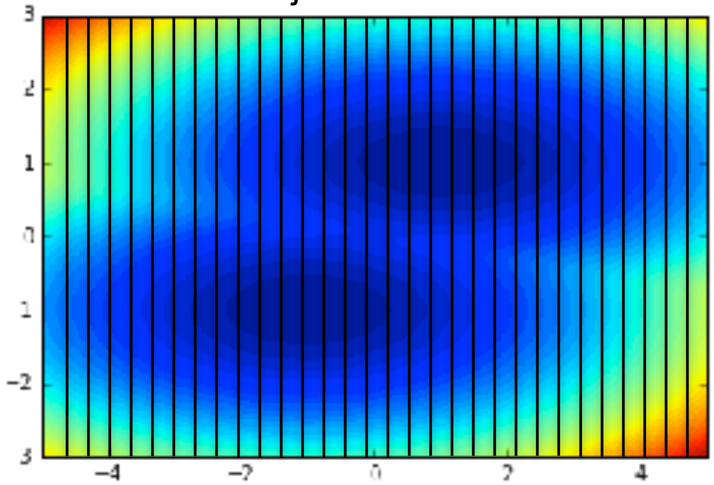


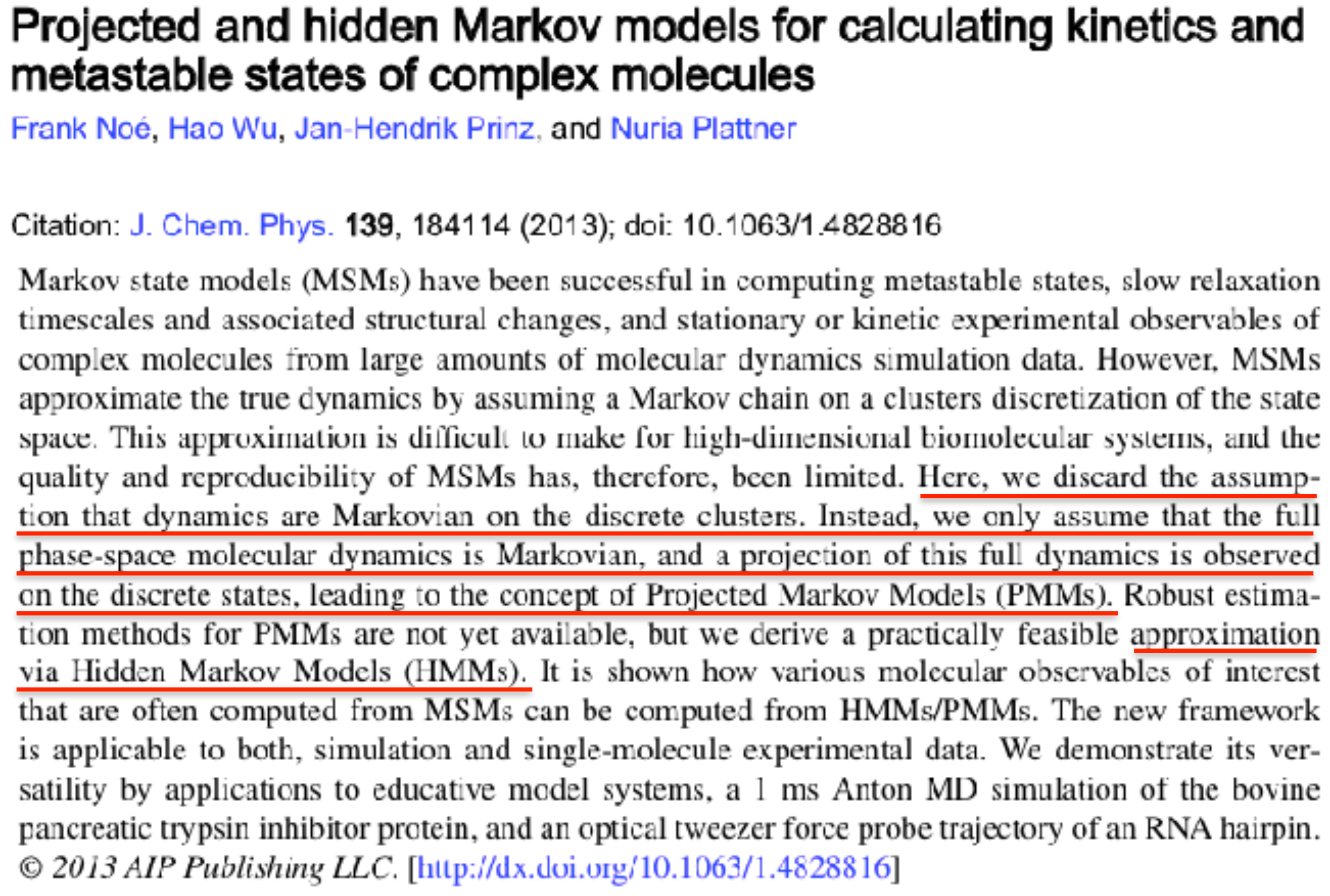
Prinz et al, **JCP 134**, 174105 (2011)

Two flavors of discretization errors



Projection error





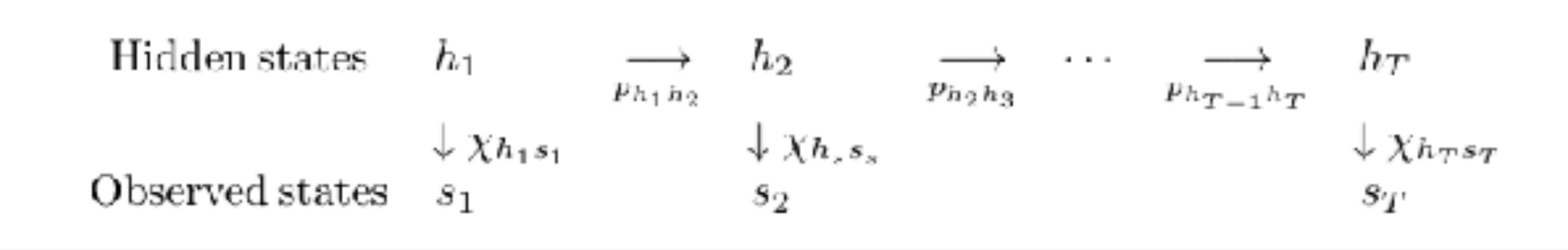
Projected and hidden Markov models for calculating kinetics and metastable states of complex molecules

Frank Noé, Hao Wu, Jan-Hendrik Prinz, and Nuria Plattner

Citation: J. Chem. Phys. **139**, 184114 (2013); doi: 10.1063/1.4828816

Markov state models (MSMs) have been successful in computing metastable states, slow relaxation timescales and associated structural changes, and stationary or kinetic experimental observables of complex molecules from large amounts of molecular dynamics simulation data. However, MSMs approximate the true dynamics by assuming a Markov chain on a clusters discretization of the state space. This approximation is difficult to make for high-dimensional biomolecular systems, and the quality and reproducibility of MSMs has, therefore, been limited. Here, we discard the assumption that dynamics are Markovian on the discrete clusters. Instead, we only assume that the full phase-space molecular dynamics is Markovian, and a projection of this full dynamics is observed on the discrete states, leading to the concept of Projected Markov Models (PMMs). Robust estimation methods for PMMs are not yet available, but we derive a practically feasible approximation via Hidden Markov Models (HMMs). It is shown how various molecular observables of interest that are often computed from MSMs can be computed from HMMs/PMMs. The new framework is applicable to both, simulation and single-molecule experimental data. We demonstrate its versatility by applications to educative model systems, a 1 ms Anton MD simulation of the bovine pancreatic trypsin inhibitor protein, and an optical tweezer force probe trajectory of an RNA hairpin.

Hidden Markov Model



$$\begin{tikzcd}
\text{Hidden states} & h_1 \arrow[r, "p_{h_1 h_2}"'] \arrow[d, "\chi_{h_1 s_1}"] & h_2 \arrow[r, "p_{h_2 h_3}"'] \arrow[d, "\chi_{h_2 s_2}"] & \cdots \arrow[r, "p_{h_{T-1} h_T}"'] & h_T \arrow[d, "\chi_{h_T s_T}"] \\
\text{Observed states} & s_1 & s_2 & & s_T
\end{tikzcd}$$

Hidden Markov Model

$$\begin{tikzcd}
\text{Hidden states} & h_1 \arrow[r, "p_{h_1 h_2}"'] \arrow[d, "\chi_{h_1 s_1}"] & h_2 \arrow[r, "p_{h_2 h_3}"'] \arrow[d, "\chi_{h_2 s_2}"] & \cdots \arrow[r, "p_{h_{T-1} h_T}"'] & h_T \arrow[d, "\chi_{h_T s_T}"] \\
\text{Observed states} & s_1 & s_2 & & s_T
\end{tikzcd}$$

Maximum-likelihood Estimation: We define the forward and backward variables

$$\begin{aligned}
\alpha_{t,i} &= \mathbb{P}(s_1, \ldots s_t, h_t = i \mid \bar{\pi}) \\
\beta_{t,i} &= \mathbb{P}(s_{t+1}, \ldots s_T, \mid h_t = i)
\end{aligned}$$

and the observation matrices

$$\mathbf{O}_t = \operatorname{diag}(\chi_{1,s_t}, \ldots, \chi_{m,s_t})$$

The maximum-likelihood estimator is obtained by iterating:

1. **Expectation** step: Estimate the forward-backward variables

$$\boldsymbol{\alpha}_{1:T}, \boldsymbol{\beta}_{1:T} = \arg\max \mathbb{P}\left(s_{1:T} \mid \boldsymbol{\alpha}_{1:T}, \boldsymbol{\beta}_{1:T}, \tilde{\mathbf{P}}, \chi\right)$$

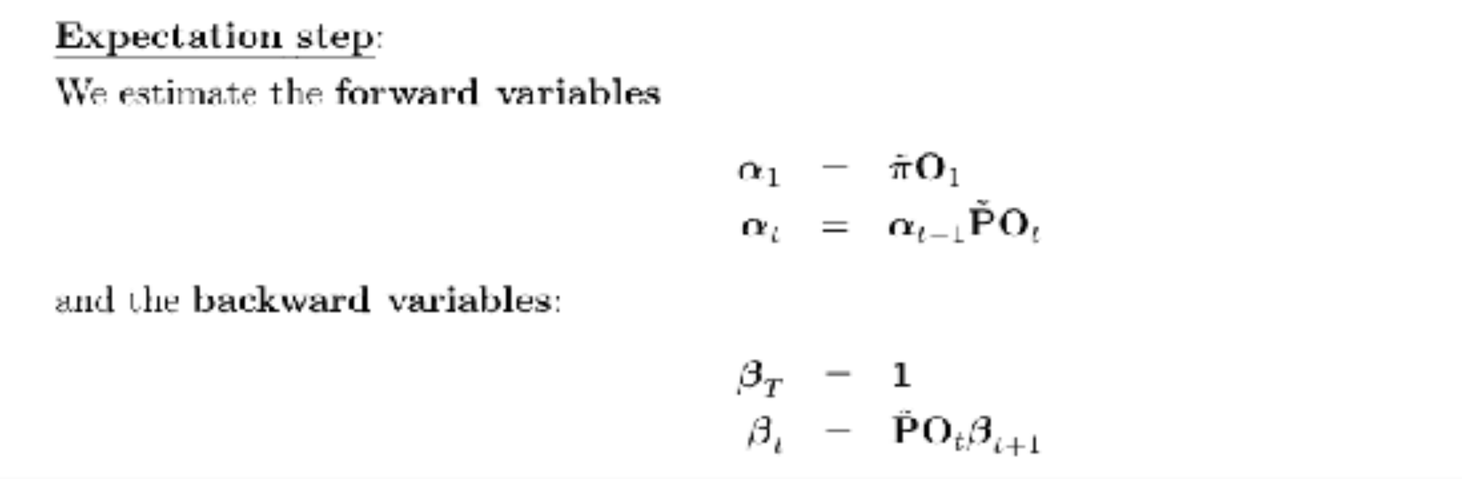
   and compute likelihood

$$L = \log \mathbb{P}\left(s_{1:T} \mid \tilde{\mathbf{P}}, \chi\right)$$

2. **Maximization** step: Estimate the parameters

$$\tilde{\mathbf{P}}, \chi = \arg\max \mathbb{P}\left(s_{1:T} \mid \tilde{\mathbf{P}}, \chi, \boldsymbol{\alpha}_{1:T}, \boldsymbol{\beta}_{1:T}\right)$$

Hidden Markov Model



<ins>**Expectation step**</ins>:

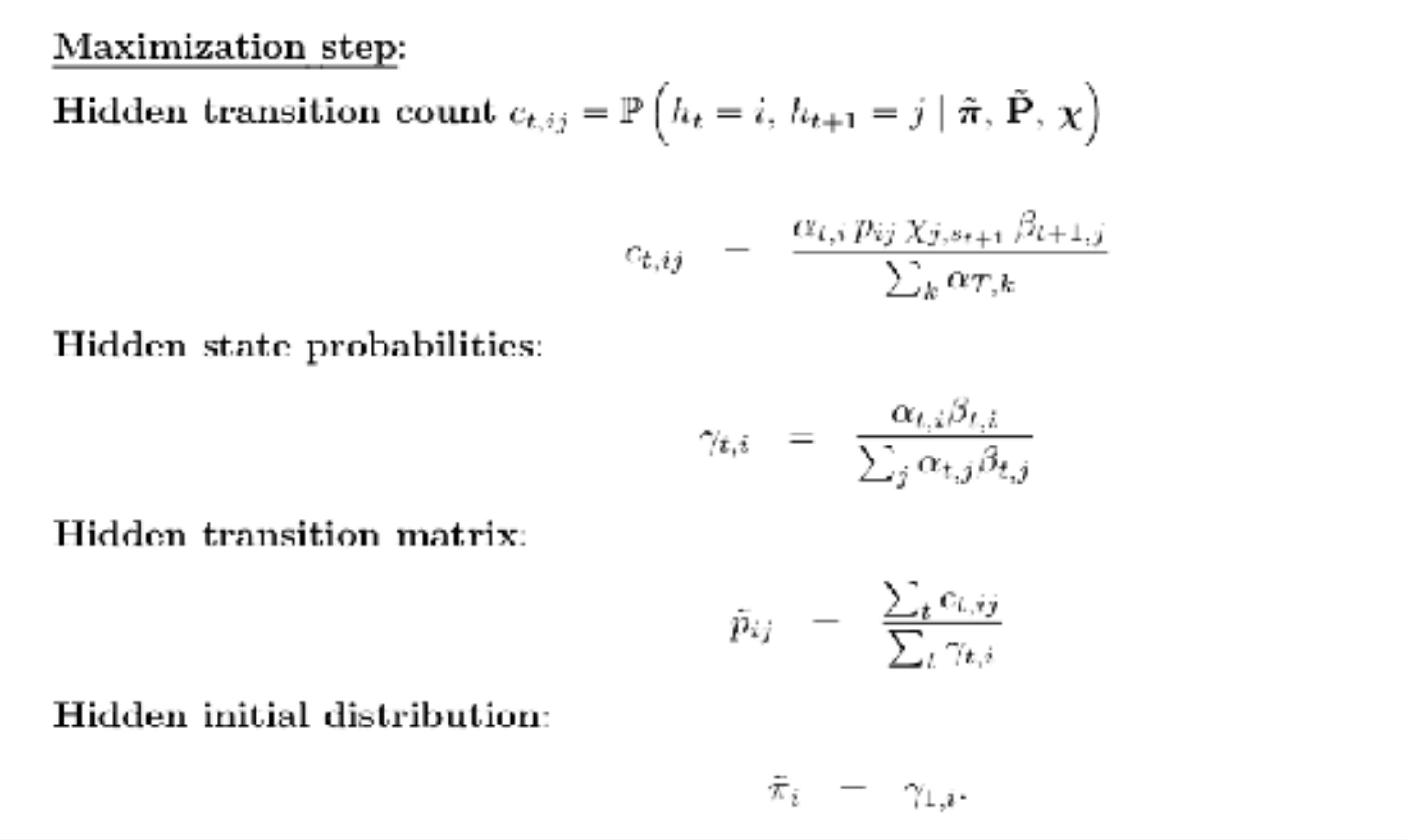
We estimate the **forward variables**

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \check{\boldsymbol{\pi}} \mathbf{O}_1 \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1} \check{\mathbf{P}} \mathbf{O}_t \end{aligned}$$

and the **backward variables**:

$$\begin{aligned} \boldsymbol{\beta}_T &= \mathbf{1} \\ \boldsymbol{\beta}_t &= \ddot{\mathbf{P}} \mathbf{O}_t \boldsymbol{\beta}_{t+1} \end{aligned}$$

Hidden Markov Model



Maximization step:

Hidden transition count $c_{t,ij} = \mathbb{P}\left(h_t = i,\, h_{t+1} = j \mid \tilde{\boldsymbol{\pi}},\, \tilde{\mathbf{P}},\, \boldsymbol{\chi}\right)$

$$c_{t,ij} = \frac{\alpha_{t,i}\, p_{ij}\, \chi_{j,s_{t+1}}\, \beta_{t+1,j}}{\sum_k \alpha_{T,k}}$$

Hidden state probabilities:

$$\gamma_{t,i} = \frac{\alpha_{t,i}\beta_{t,i}}{\sum_j \alpha_{t,j}\beta_{t,j}}$$

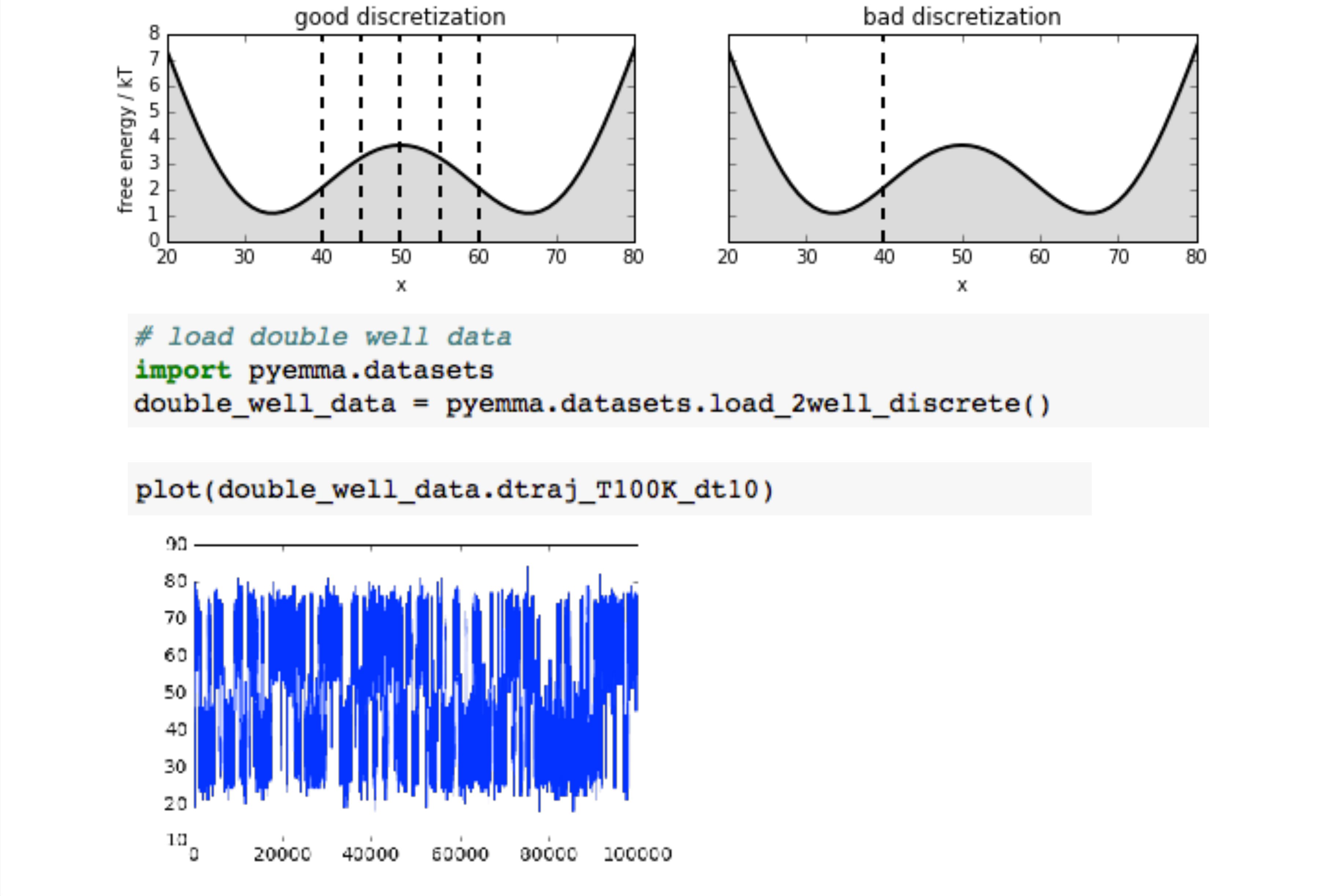
Hidden transition matrix:

$$\tilde{p}_{ij} = \frac{\sum_t c_{t,ij}}{\sum_t \gamma_{t,i}}$$

Hidden initial distribution:

$$\tilde{\pi}_i = \gamma_{1,i}.$$

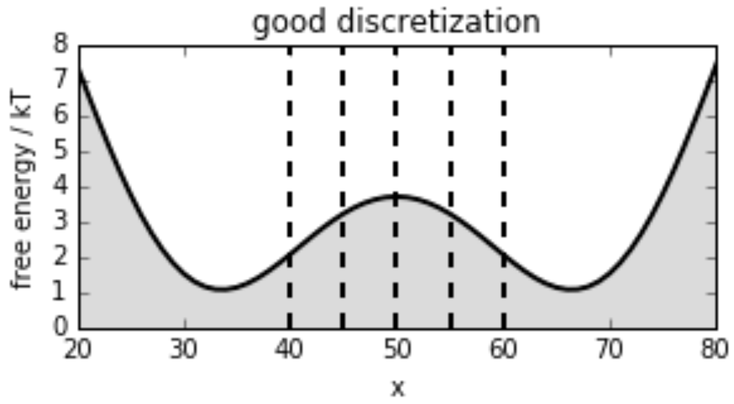
Case 1: coarse discretization



```
# load double well data
import pyemma.datasets
double_well_data = pyemma.datasets.load_2well_discrete()
```

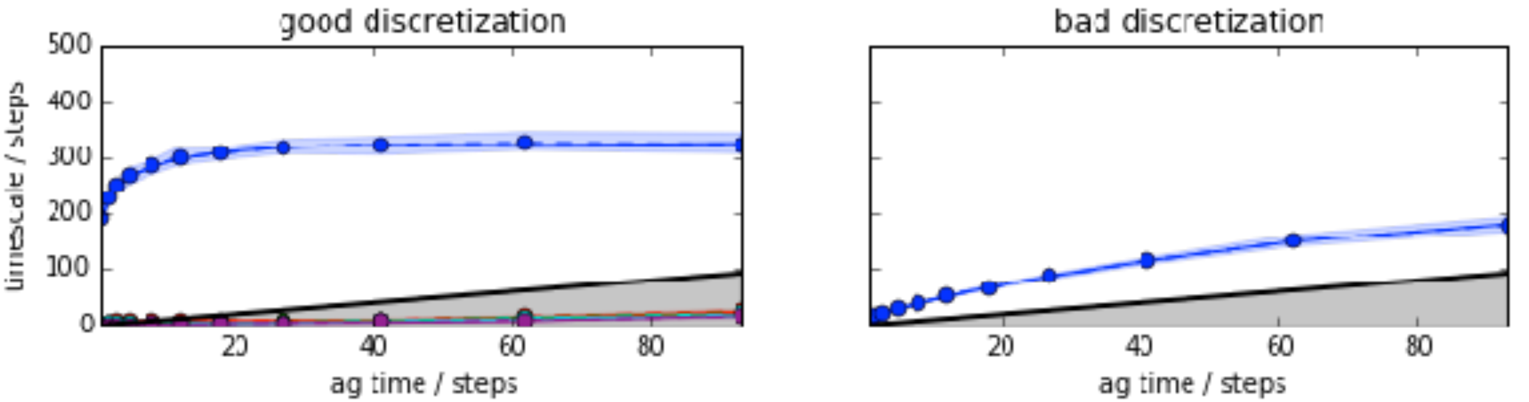
```
plot(double_well_data.dtraj_T100K_dt10)
```

Case 1: coarse discretization

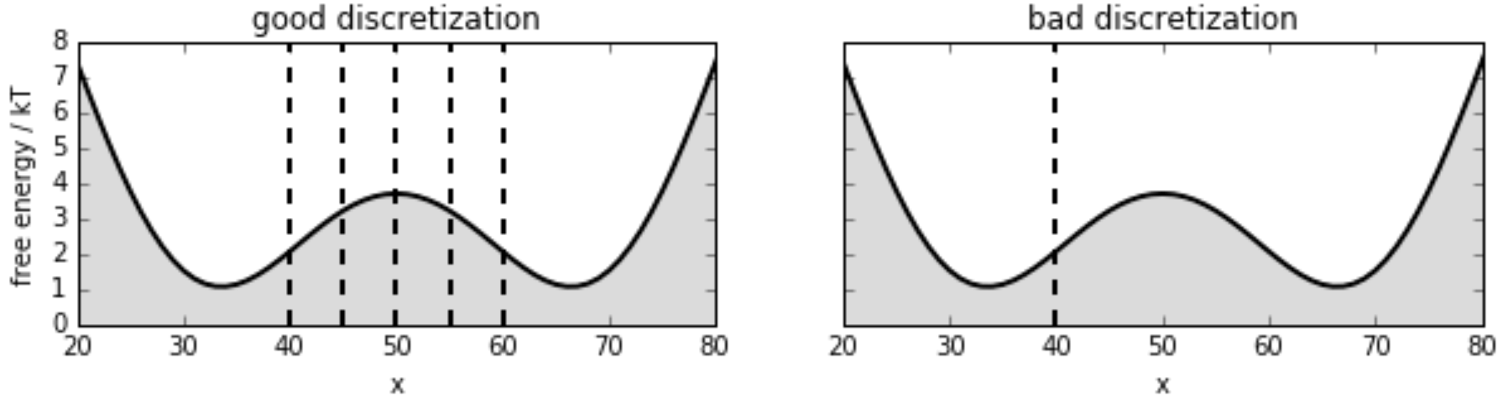


MSM: Implied timescales

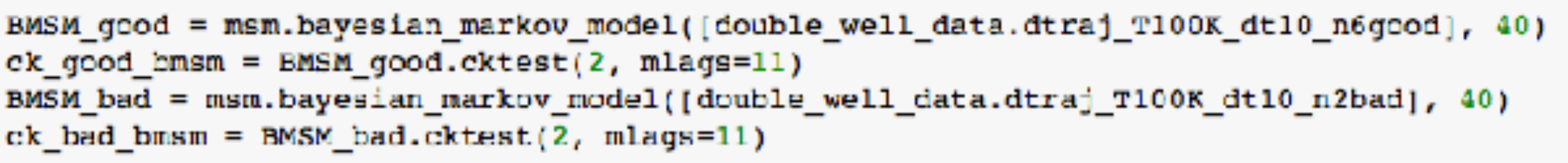
```
its_good_bmsm = msm.timescales_msm([double_well_data.dtraj_T100K_dt10_n6good], lags = 100, errors='bayes')
its_bad_bmsm = msm.timescales_msm([double_well_data.dtraj_T100K_dt10_n2bad], lags = 100, errors='bayes')
```



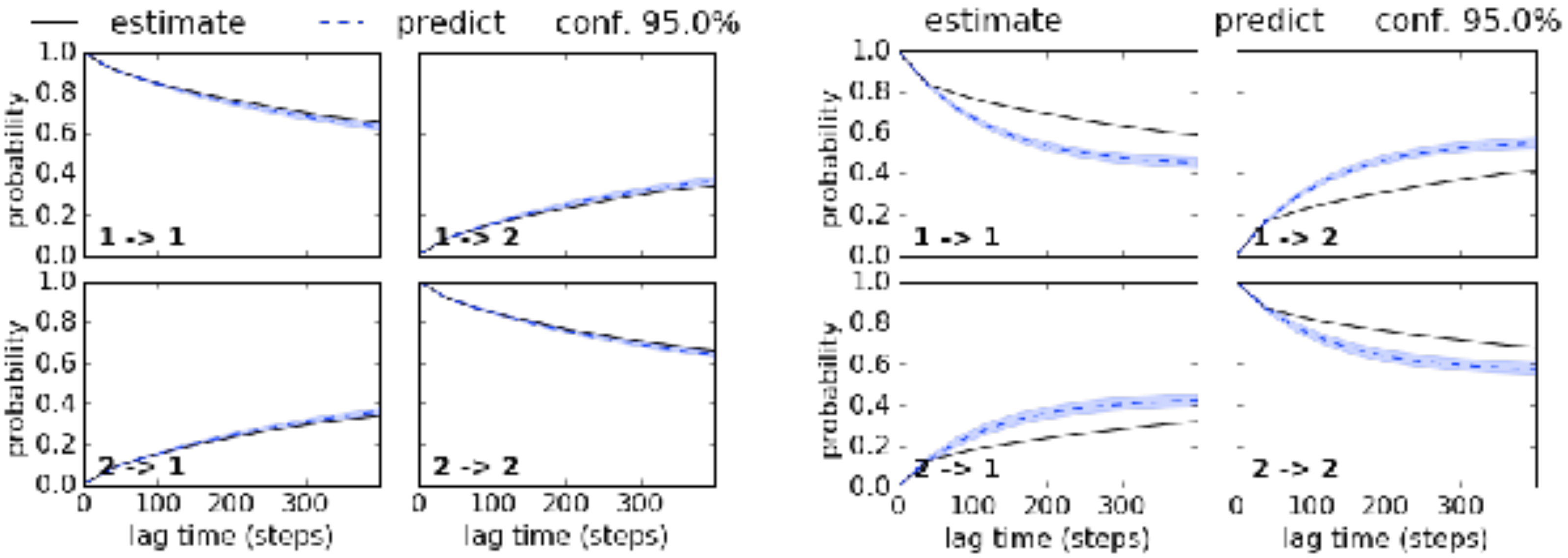
Case 1: coarse discretization



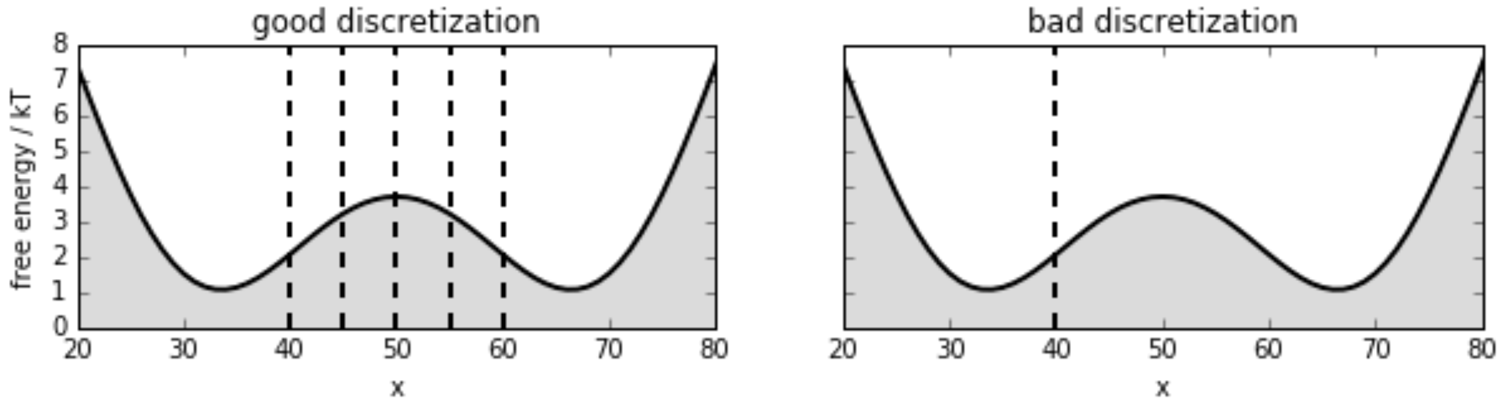
MSM: Chapman-Kolmogorov Test



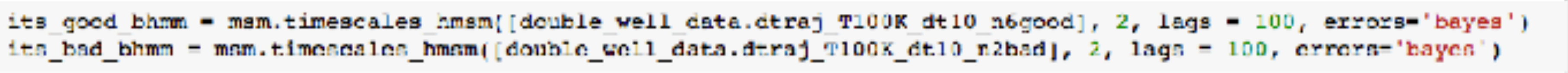
```
BMSM_good = msm.bayesian_markov_model([double_well_data.dtraj_T100K_dt10_n6good], 40)
ck_good_bmsm = BMSM_good.cktest(2, mlags=11)
BMSM_bad = msm.bayesian_markov_model([double_well_data.dtraj_T100K_dt10_n2bad], 40)
ck_bad_bmsm = BMSM_bad.cktest(2, mlags=11)
```



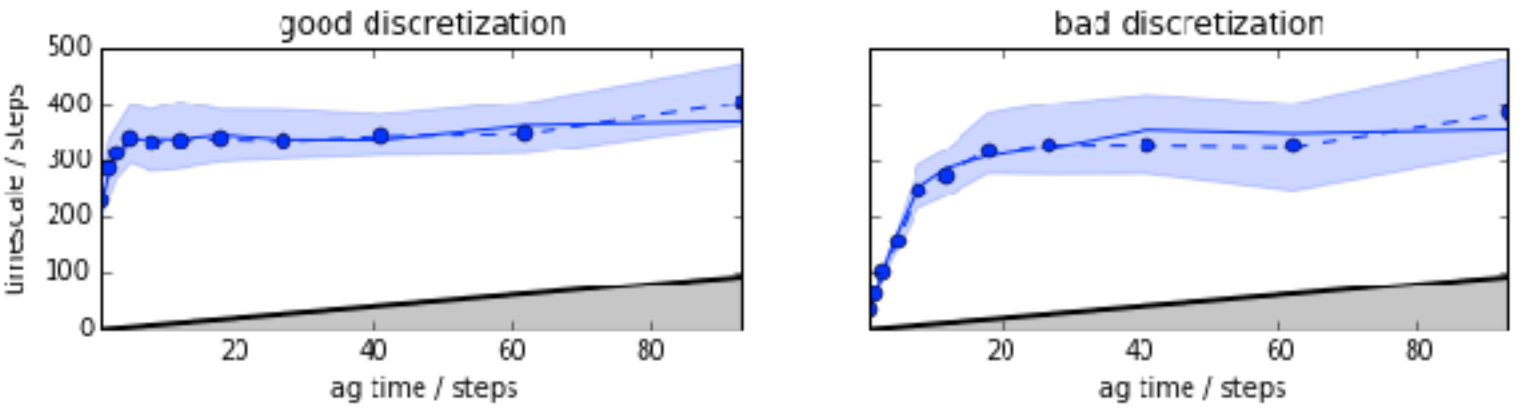
Case 1: coarse discretization



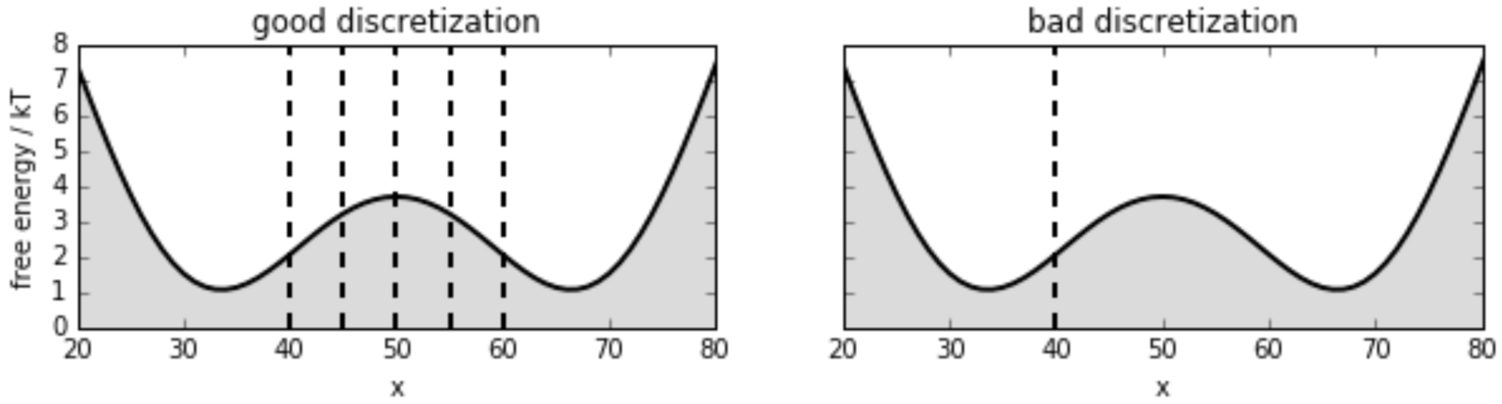
HMM: Implied timescales



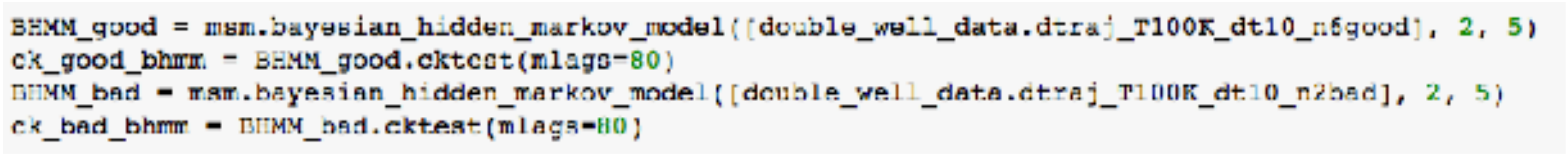
```
its_good_bhmm = msm.timescales_hmsm([double_well_data.dtraj_T100K_dt10_n6good], 2, lags = 100, errors='bayes')
its_bad_bhmm = msm.timescales_hmsm([double_well_data.dtraj_T100K_dt10_n2bad], 2, lags = 100, errors='bayes')
```



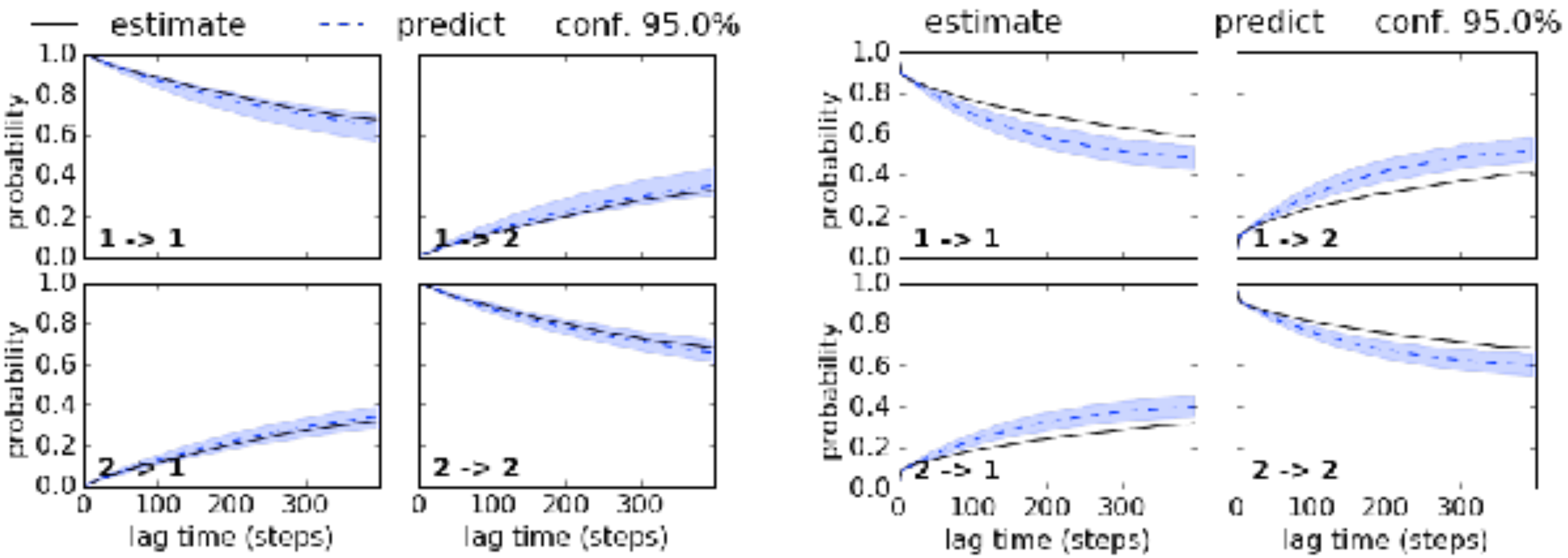
Case 1: coarse discretization



HMM: Chapman-Kolmogorov Test

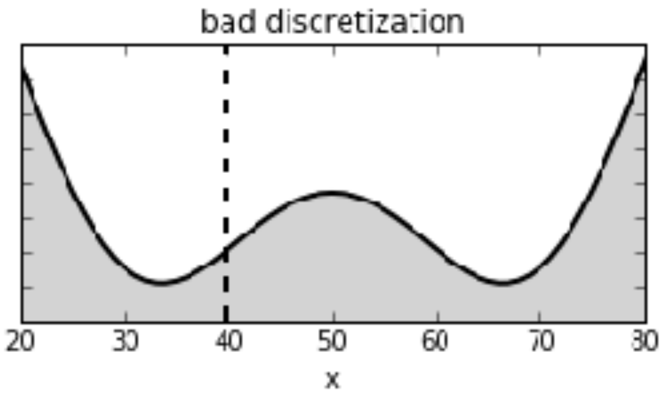


```
BHMM_good = msm.bayesian_hidden_markov_model([double_well_data.dtraj_T100K_dt10_n6good], 2, 5)
ck_good_bhmm = BHMM_good.cktest(mlags=80)
BHMM_bad = msm.bayesian_hidden_markov_model([double_well_data.dtraj_T100K_dt10_n2bad], 2, 5)
ck_bad_bhmm = BHMM_bad.cktest(mlags=80)
```

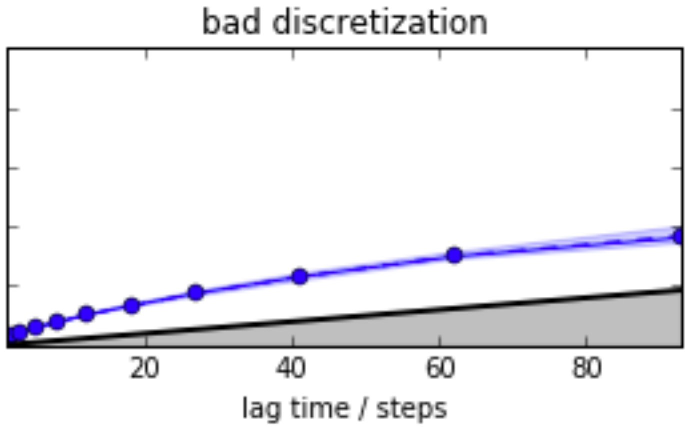


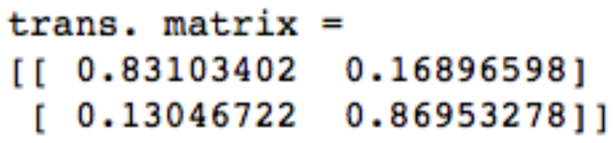
Case 1: coarse discretization

Why does the HMM work better than the MSM?



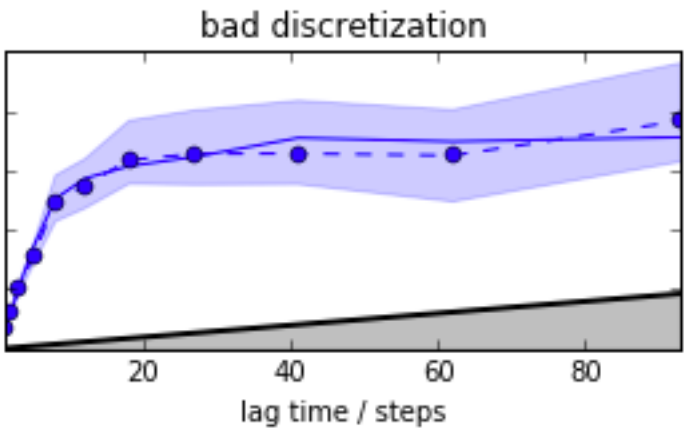
MSM at lag 40:

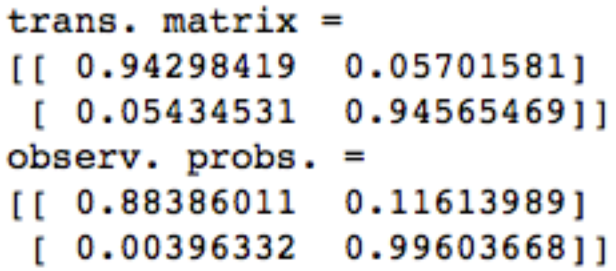




```
trans. matrix =
[[ 0.83103402  0.16896598]
 [ 0.13046722  0.86953278]]
```

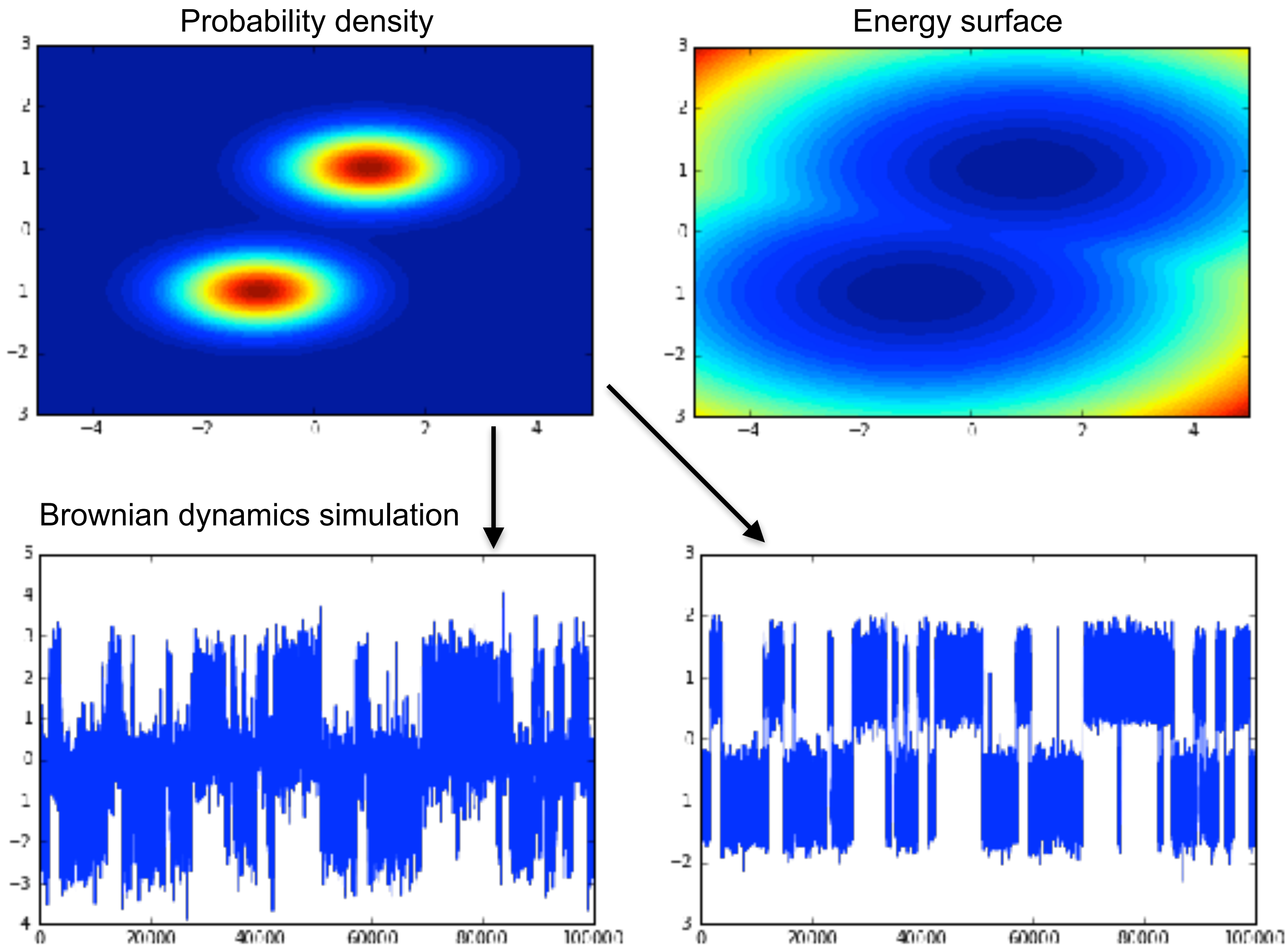
HMM at lag 40:





```
trans. matrix =
[[ 0.94298419  0.05701581]
 [ 0.05434531  0.94565469]]
observ. probs. =
[[ 0.88386011  0.11613989]
 [ 0.00396332  0.99603668]]
```

Case 2: bad projection



References

[1] Prinz, J.-H., H. Wu, M. Sarich, B. G. Keller, M. Senne, M. Held, J. D. Chodera, Ch. Schütte and F. Noé: Markov models of molecular kinetics: Generation and Validation. J. Chem. Phys. 134, 174105 (2011)

[2] Sarich, M., F. Noé, Ch. Schütte: On the Approximation Quality of Markov State Models. Multiscale Model. Simul. 8, 1154-1177 (2010)

[3] Swope, W. C., J. W. Pitera and F. Suits: Describing protein folding kinetics by molecular dynamics simulations: 1. Theory, J. Phys. Chem. B. 108, 6571-6581 (2004)

[4] Beauchamp, K. A., R. McGibbon, Y. S. Lin and V. S. Pande: Simple few-state models reveal hidden complexity in protein folding. Proc. Natl. Acad. Sci. USA 109, 17807-17813 (2012)

[5] Noé, F. and F. Nüske: A variational approach to modeling slow processes in stochastic dynamical systems. SIAM Multiscale Model. Simul. 11. 635-655 (2013).

[6] Trendelkamp-Schroer, B., H. Wu, F. Paul and F. Noé: Estimation and uncertainty of reversible Markov models. arxiv.org/pdf/1507.05990 (2015)

[7] Rabiner, L. R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. IEEE 77, 257--286 (1989)

[8] Noé, F., H. Wu, J.-H. Prinz and N. Plattner, N.: Projected and Hidden Markov Models for calculating kinetics and metastable states of complex molecules. J. Chem. Phys. 139, 184114 (2013)

[9] Chodera, J. D., P. Elms, F. Noé, B. Keller, C. M. Kaiser, A. Ewall-Wice, S. Marqusee, C. Bustamante, N. Singhal Hinrichs: Bayesian hidden Markov model analysis of single-molecule force spectroscopy: Characterizing kinetics under measurement uncertainty. http://arxiv.org/abs/1108.1430.

Acknowledgements

Positions available
frank.noe@fu-berlin.de

Collaborations

Cecilia Clementi (Rice)
Christof Schütte (FU Berlin)
Eric Vanden-Eijnden (Courant Institut NY)
Thomas Weikl (MPI Potsdam)
Marcus Sauer, Sören Doose (Uni Würzburg)

Vijay Pande (Stanford)
Katja, Faelber, Oliver Daumke (MDC)
John Chodera (MSKCC NY)
Gianni de Fabritiis (Barcelona)

Funding

erc

DFG
Deutsche Forschungsgemeinschaft