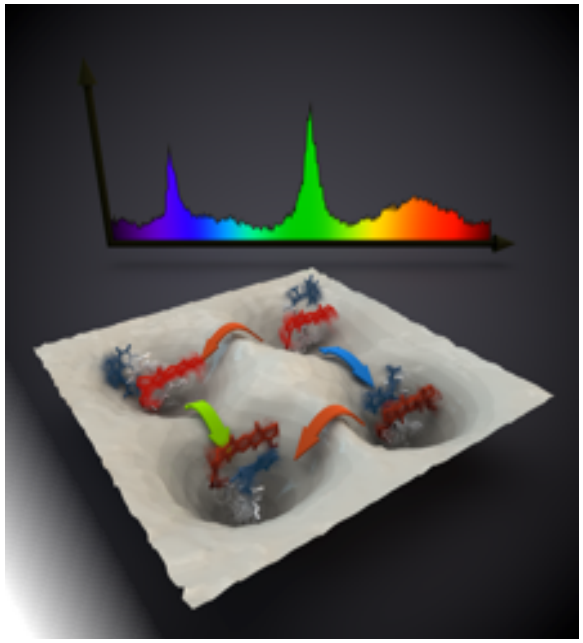
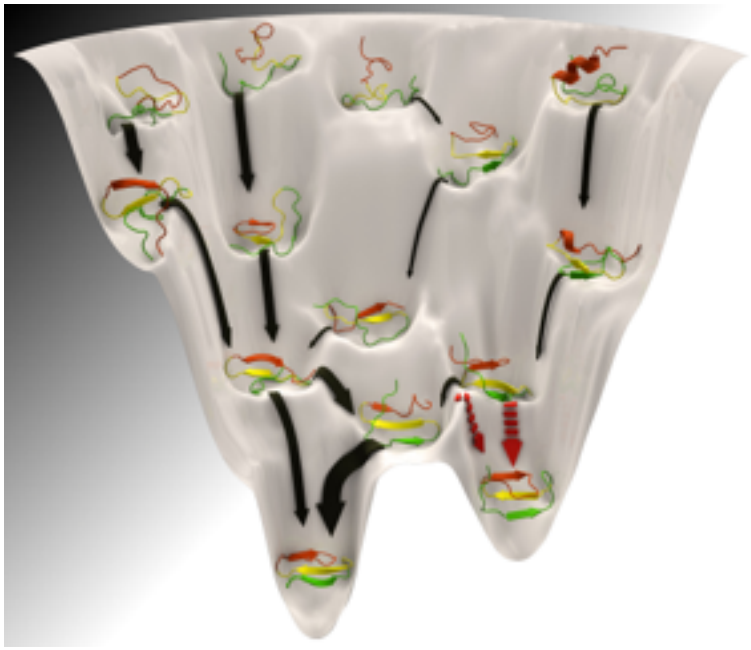
Finding slow modes and accessing very long timescales in molecular dynamics



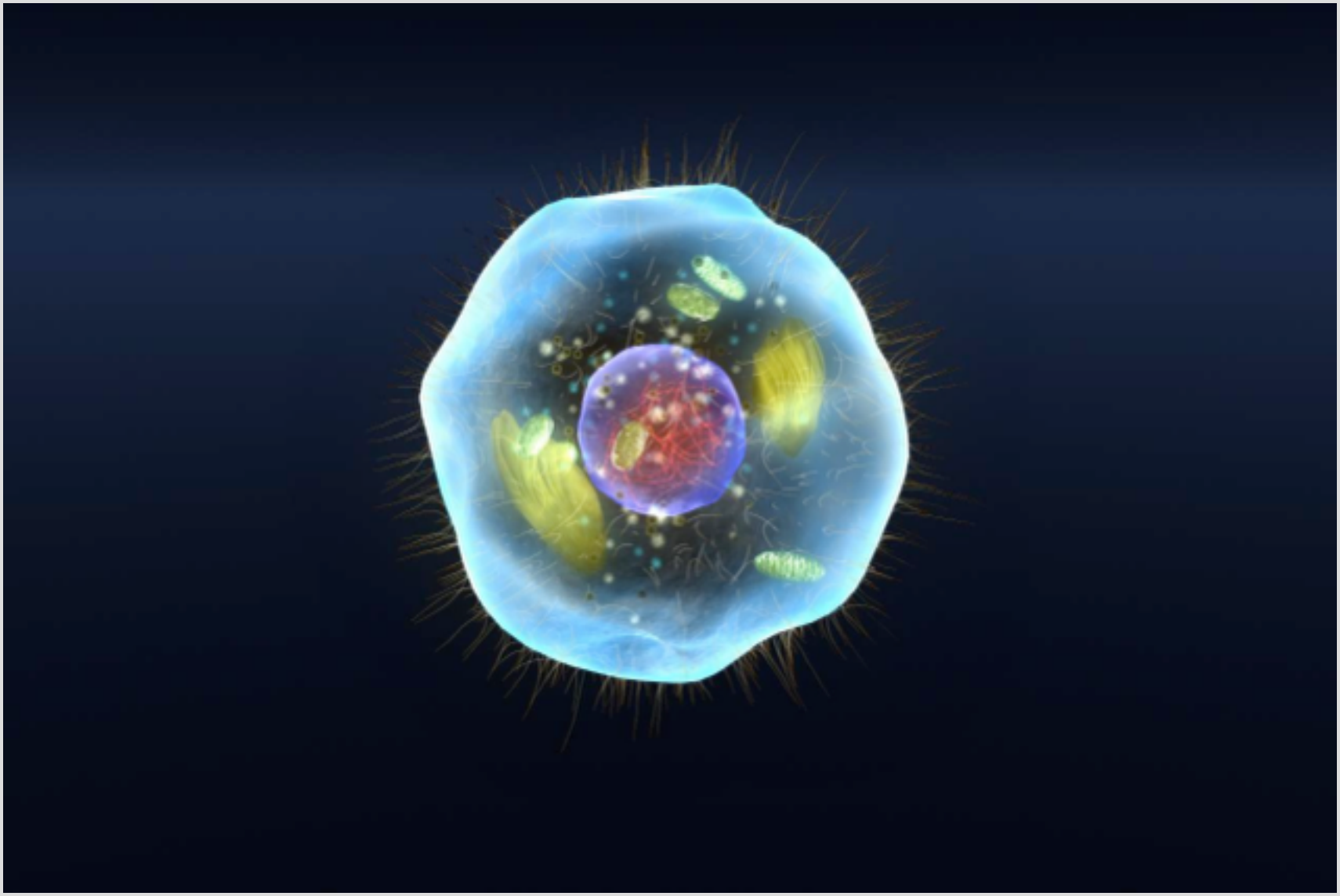


Frank Noé (FU Berlin)
frank.noe@fu-berlin.de

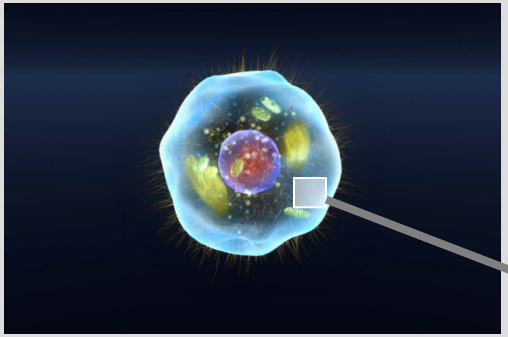
Computational Molecular Biology

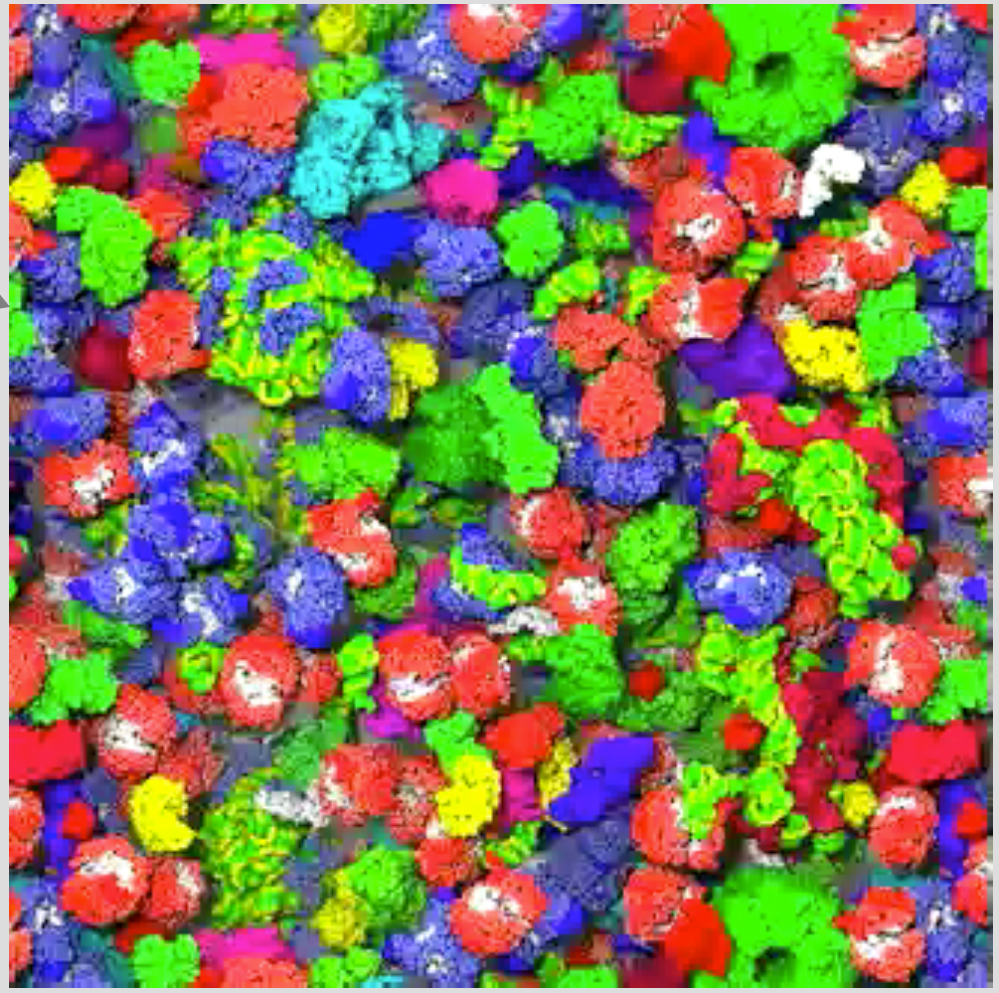
Freie Universität Berlin

Cell



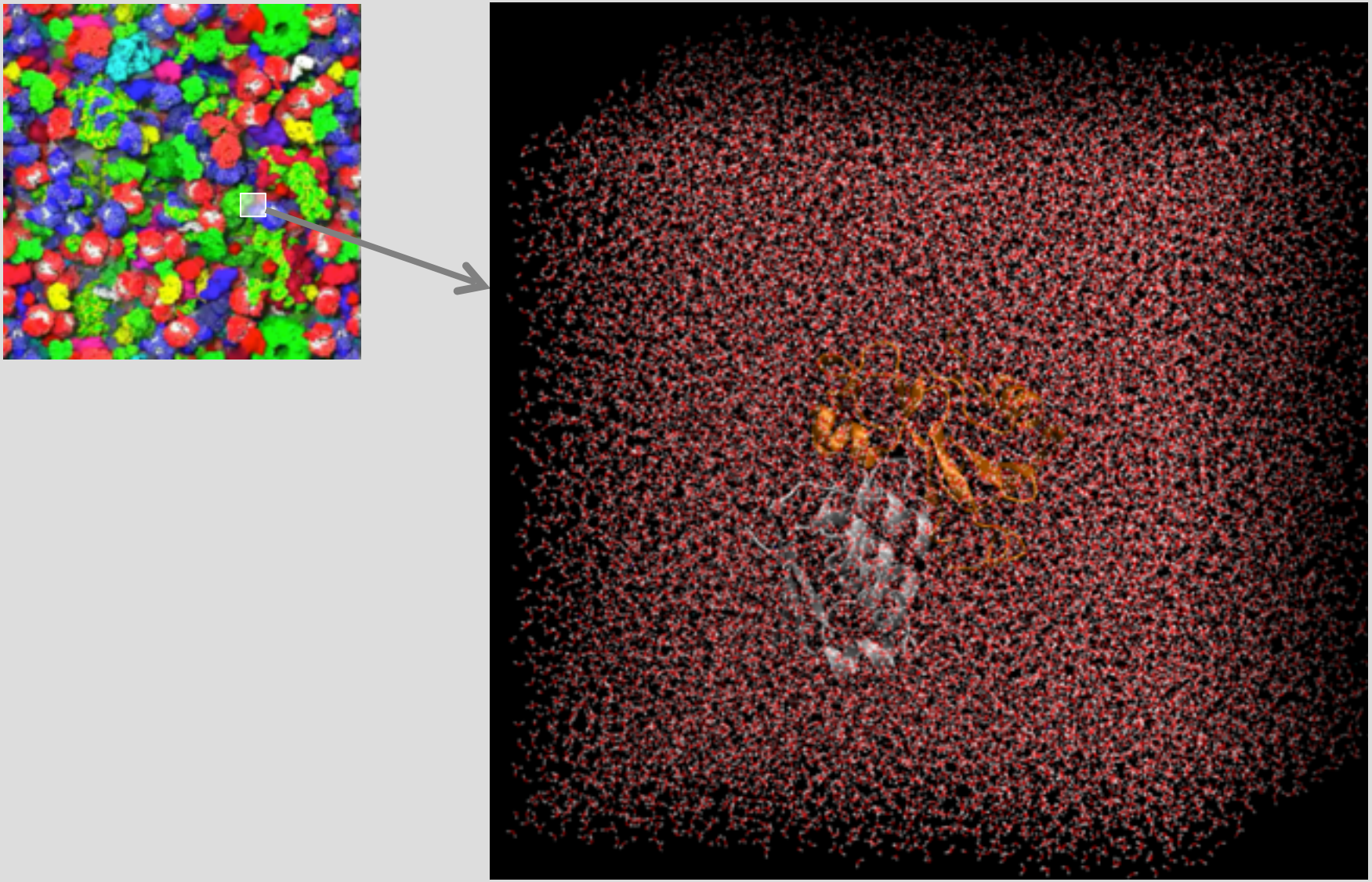
Proteins



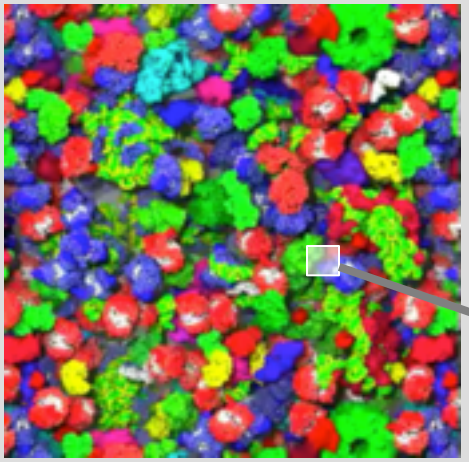


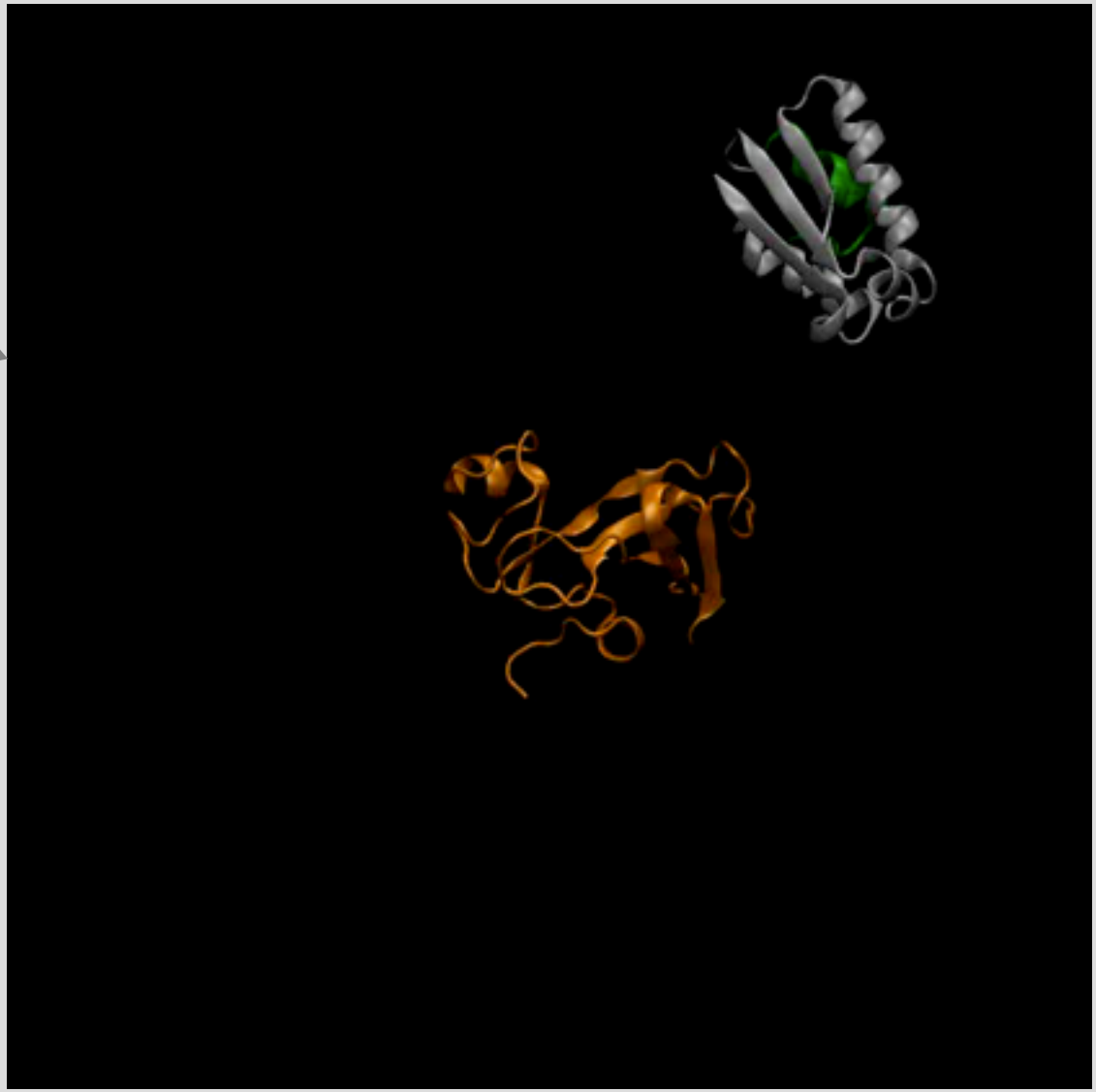
McGufee and Elcock, PloS Comput Biol 2010

Protein-Protein binding



Protein-Protein binding





Plattner, Doerr, De Fabritiis, Noé

0.1 microseconds

Part I

Direct approximation of slow modes / processes

Noé and Nüske, **Mult. Modeling Simul.** 11, 635-655 (2013) / also: **arXiv** (2012)

Nüske, Keller, Pérez-Hernandez, Mey, Noé, **JCTC** 10, 1739-1752 (2014)

Dynamical operators

Propagator

$$\rho_{t+\tau} = \mathcal{P}_\tau \rho_t = \int p_\tau(y \mid x)\, \rho_t(x)\, \mathrm{d}x$$

Transfer operator / Perron-Frobenius operator

= propagator for densities $u(x) = \frac{\rho(x)}{\pi(x)}$ with stationary density $\pi(x)$.

$$u_{t+\tau} = \mathcal{T}_\tau u_t = \int \frac{\pi(x)}{\pi(y)} p_\tau(y \mid x)\, \rho_t(x)\, \mathrm{d}x$$

Koopman operator

Adjoint to $\mathcal{P}$, adjoint to $\mathcal{T}$ with respect to π

$$f_{t+\tau} = \mathcal{K}_\tau f_t = \int p_\tau(y \mid x)\, f_t(y)\, \mathrm{d}y = \mathbb{E}[f_{t+\tau}(x)]$$

with detailed balance: $\pi(x)p_\tau(y \mid x) = \pi(y)p_\tau(x \mid y)$ we have $\mathcal{K}_\tau \equiv \mathcal{T}_\tau$

See: Mesic **Nonlinear Dyn.** 41, 309 (2005).

Slow processes

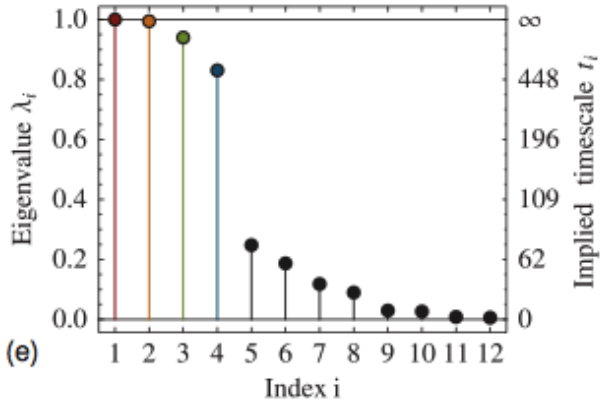
Backward propagator

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

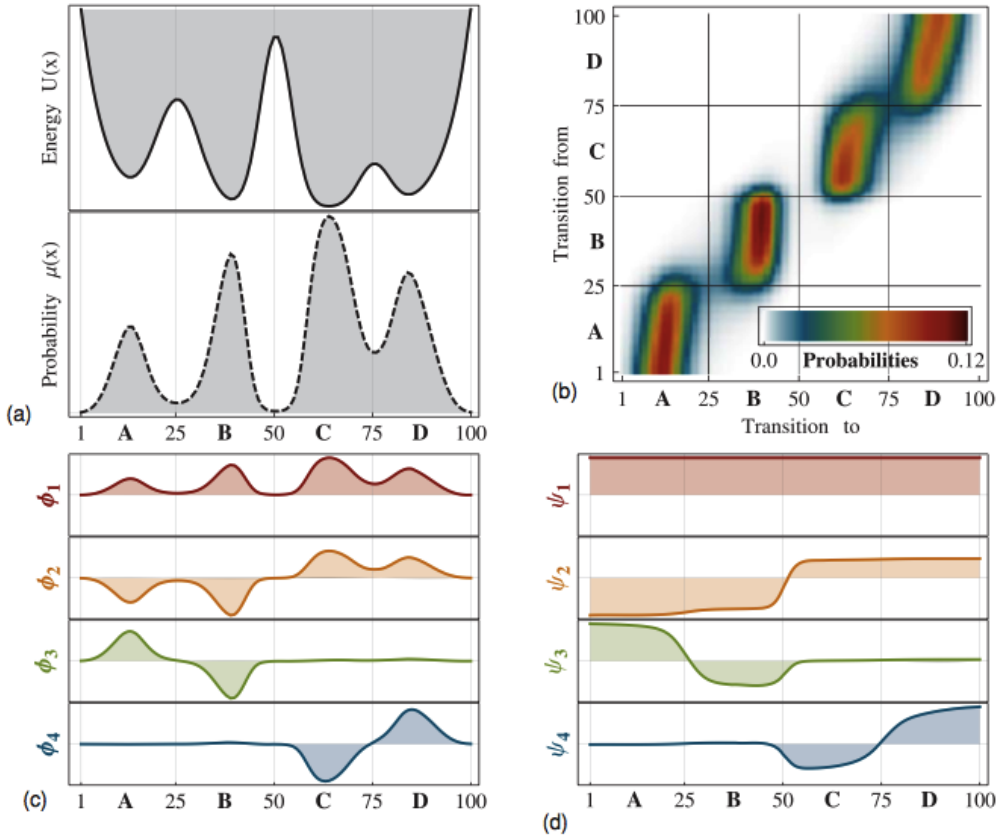
Spectral decomposition

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

Eigenvalues / timescales κ_i^{-1}



Processes:



Schütte et al: **J. Comput. Phys.** (1999), Prinz et al.: **J. Chem. Phys.** 134, p174105 (2011)

Collective variables: Slow modes versus reaction coordinates

Identify the slowest processes

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

Approximate Markov operator **eigenfunctions**

Identify the “best” / “slowest” coordinate connecting A and B

$$\begin{aligned} \tau_A, \tau_B & \quad \text{time to hit } A,\ B \text{ next} \\ q_{AB}(\mathbf{x}) & = \mathrm{Prob}(\tau_B < \tau_A) \end{aligned}$$

Approximate **committor**

Eigenfunctions and committor are closely related when considering reactions between metastable states

Schütte, Noé, Lu, Sarich, Vanden-Eijnden JCP 134, 204105

Molecular order parameters

Candidates x_1…x_n

- Cartesian coordinates
- Distances between heavy atoms or Ca's
- cos/sin of dihedrals
- Gaussian densities around sampled configurations (e.g. in RMSD space)
- Field variables or coordination numbers
- Your favorite order parameter
- All of them

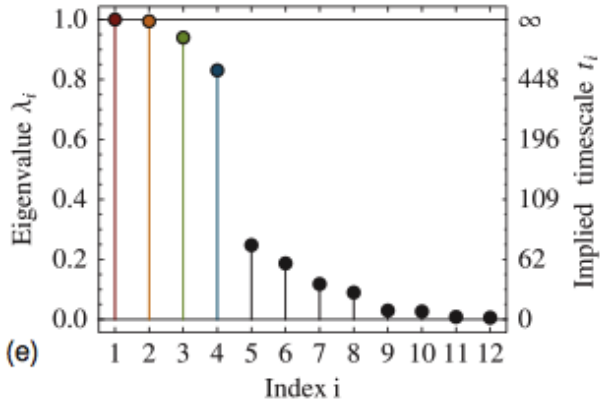
Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)

Slow processes

Backward propagator

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

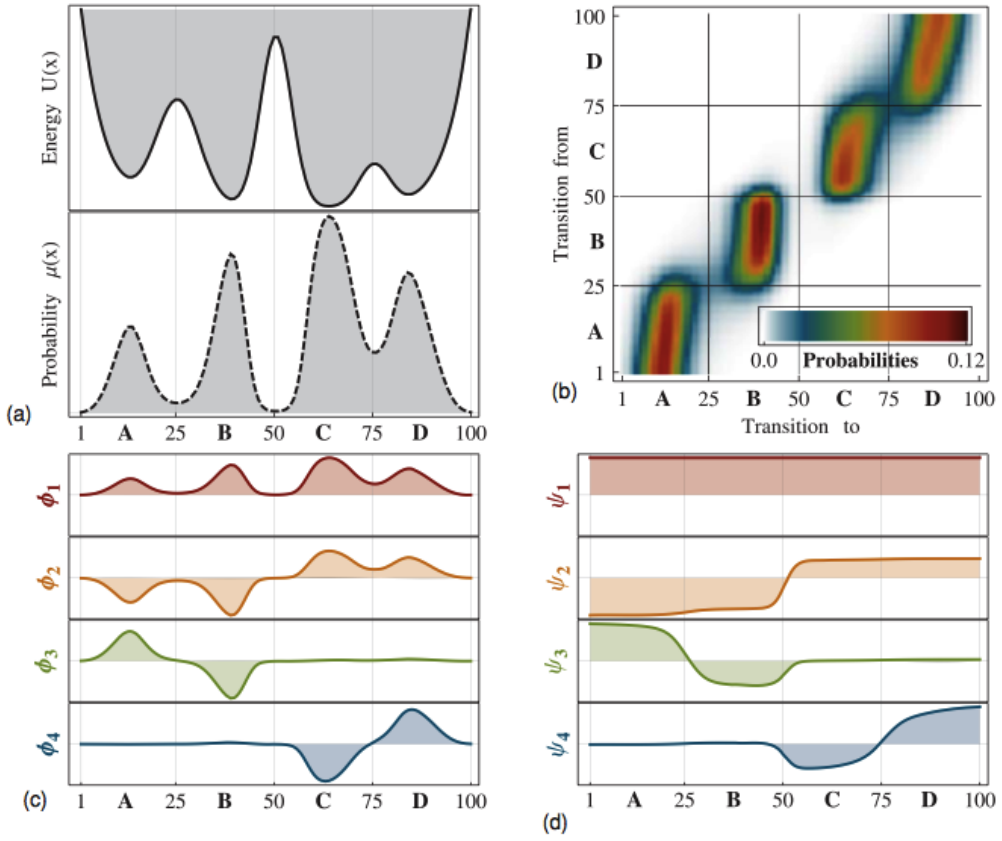
Eigenvalues / timescales κ_i^{-1}



Spectral decomposition

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

Processes:



Variational approach for Markov processes

Data-based version of: Fan, **PNAS** 35, 652-655 (1949)

The first m eigenfunctions $\psi_1, \ldots, \psi_m$ are the solution to the problem

$$\max_{f_1,\ldots,f_m} \sum_{i=1}^{m} \mathbb{E}\left[f_i\left(\mathbf{x}_t\right) f_i\left(\mathbf{x}_{t+\tau}\right)\right] \tag{1}$$

$$\text{s.t.} \quad \mathbb{E}\left[f_i\left(\mathbf{x}_t\right)^2\right] = 1$$

$$\mathbb{E}\left[f_i\left(\mathbf{x}_t\right) f_j\left(\mathbf{x}_{t+\tau}\right)\right] = 0, \ \textit{for } i \neq j$$

and the maximum value is the sum of $\lambda_1, \ldots, \lambda_m$

Properties:

- ψ_i and ψ_j are uncorrelated for $i \neq j$.
- ψ_i are the directions of slow kinetics with maximal autocorrelations $\mathbb{E}_\mu\left[\psi_i\left(\mathbf{x}_t\right) \psi_i\left(\mathbf{x}_{t+\tau}\right)\right] = \lambda_i\left(\tau\right)$.
- Population changes along ψ_i coordinates decay with $\lambda_i\left(\tau\right) = \mathrm{e}^{-\frac{\tau}{t_i}}$.
- For every other set of functions, the eigenvalues will be underestimated $\hat{\lambda}_i\left(\tau\right) \leq \lambda_i\left(\tau\right)$.

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

Variational approach for Markov processes: Linear variation

Ansatz: Define Basis set $\boldsymbol{\chi} = [\chi_1(\mathbf{x}), ..., \chi_n(\mathbf{x})]^\top$ and seek the linear expansions:

$$\hat{\psi}_i(\mathbf{x}) = \sum_j r_{ij}\chi_j(\mathbf{x})$$

Algorithm:

1. Estimate $\mathbf{C}^0$ and $\mathbf{C}^\tau$ with:

$$c^0_{ij} = \mathbb{E}_t[\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_t)]$$
$$c^\tau_{ij} = \mathbb{E}_t[\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_{t+\tau})]$$

2. Solve

$$\mathbf{C}^\tau \mathbf{r}_i = \mathbf{C}^0 \hat{\lambda}_i \mathbf{r}_i$$

3. Expand

$$\hat{\psi}_i(\mathbf{x}) = \sum_j r_{ij}\chi_j(\mathbf{x})$$

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

Variational approach for Markov processes: Empirical estimation

1. Define

$$\mathbf{X}_0 = \begin{bmatrix} \chi_1(\mathbf{x}_0) & \cdots & \chi_n(\mathbf{x}_0) \\ \vdots & & \vdots \\ \chi_1(\mathbf{x}_{T-\tau}) & \cdots & \chi_n(\mathbf{x}_{T-\tau}) \end{bmatrix} \qquad \mathbf{X}_\tau = \begin{bmatrix} \chi_1(\mathbf{x}_\tau) & \cdots & \chi_n(\mathbf{x}_\tau) \\ \vdots & & \vdots \\ \chi_1(\mathbf{x}_T) & \cdots & \chi_n(\mathbf{x}_T) \end{bmatrix}$$

2. Empirical covariance matrices: $\mathbf{C}^0$ and $\mathbf{C}^\tau$ with:

$$\mathbf{C}^0 = \mathbf{X}_0^\top \mathbf{X}_0$$
$$\mathbf{C}^\tau = \mathbf{X}_0^\top \mathbf{X}_\tau$$

3. Solve

$$\mathbf{C}^\tau \mathbf{r}_i = \mathbf{C}^0 \hat{\lambda}_i \mathbf{r}_i$$

4. The projections

$$\mathbf{\Psi} = \mathbf{X}\mathbf{R}$$

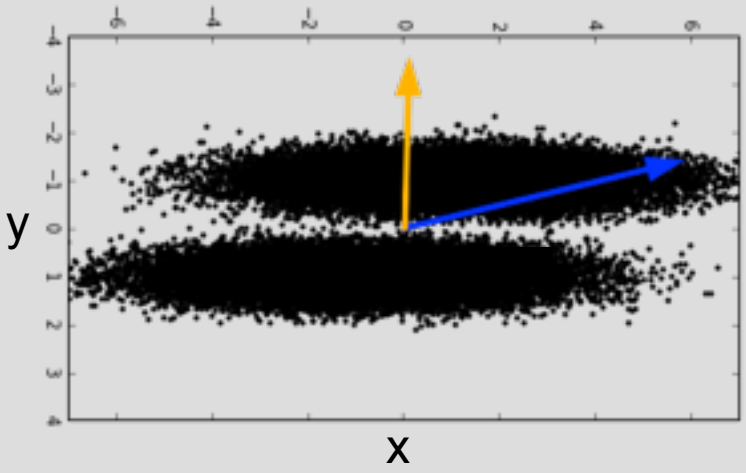
approximate the transfer operator eigenfunctions on the sampled configurations x_t.

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

Linear combination of basis functions

ψ_i are the directions of slow kinetics with maximal autocorrelations

$$\mathbb{E}_\mu \left[\psi_i \left(\mathbf{x}_t \right) \psi_i \left(\mathbf{x}_{t+\tau} \right) \right] = \lambda_i \left(\tau \right) = \mathrm{e}^{-\frac{\tau}{t_i}}.$$



Optimal for variational principle of conformation dynamics (maximum autocorrelation)

maximum variance

Simple Ansatz:

$$\psi_2(x, y) = ax + by$$

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

Comparison with other algorithms

Extended dynamic mode decomposition (EDMD)
Williams, Kevrikidis, Rowley **J. Nonlinear Sci.** 6, 1-40 (2015) / also: **arXiv** (2014)

Time-lagged independent component analysis TICA / Blind source separation
Molgedey and Schuster, **Phys Rev Lett** 72 3634-3637 (1994)

Dynamic mode decomposition (DMD)
Schmidt, Sesterhenn, **Ann. Meet. APS Div. Fluid Mech.** (2008)
Tu, Rowley, Luchtenburg, Brunton, Kutz, **J. Comput. Dyn.** (2014)

Markov state models (MSM)
Schütte, Fischer, Huisinga, Deuflhard: **J. Comput. Phys.** (1999)
see also papers by: Noé, Pande, Hummer, Weber, Swope, …

Extended dynamic mode decomposition

Relationship to Extended dynamic mode decomposition:

Williams, Kevrikidis, Rowley
J. Nonlinear Sci. 6, 1-40 (2015)

The VAMP eigenvalue problem can be written

$$\mathbf{K}\mathbf{r}_i = \mathbf{r}_i \lambda_i$$

with Koopman matrix

$$\begin{aligned}\mathbf{K} &= (\mathbf{C}^0)^{-1}\mathbf{C}^\tau \\ &= (\mathbf{X}_0^\top \mathbf{X}_0)^{-1}\mathbf{X}_0^\top \mathbf{X}_\tau = \mathbf{X}_0^+ \mathbf{X}_\tau^\top\end{aligned}$$

That is the optimal solution of the regression problem

$$\mathbf{Y} = \mathbf{X}\mathbf{K}$$

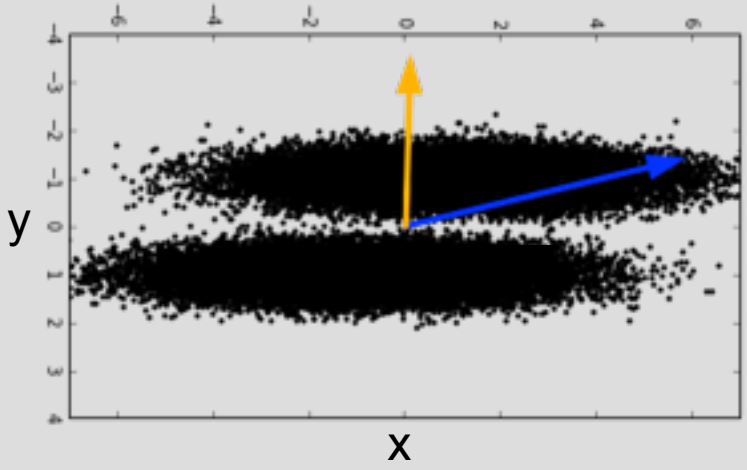
Remarks:

- The VAMP algorithm is equivalent to EDMD.
- VAMP/EDMD are also applicable to nonreversible and nonstationary dynamics.
- VAMP/EDMD results are variationally optimal and we have an eigenvalue bound in the reversible case.
- EDMD and DMD are usually implemented using singular value decomposition of the data matrix instead of computing a covariance matrix.

Special case: Time-lagged independent component analysis (TICA)

Also known as: Blind source separation, Molgedey-Schuster transform

Original paper: Molgedey and Schuster, **PRL** 1994

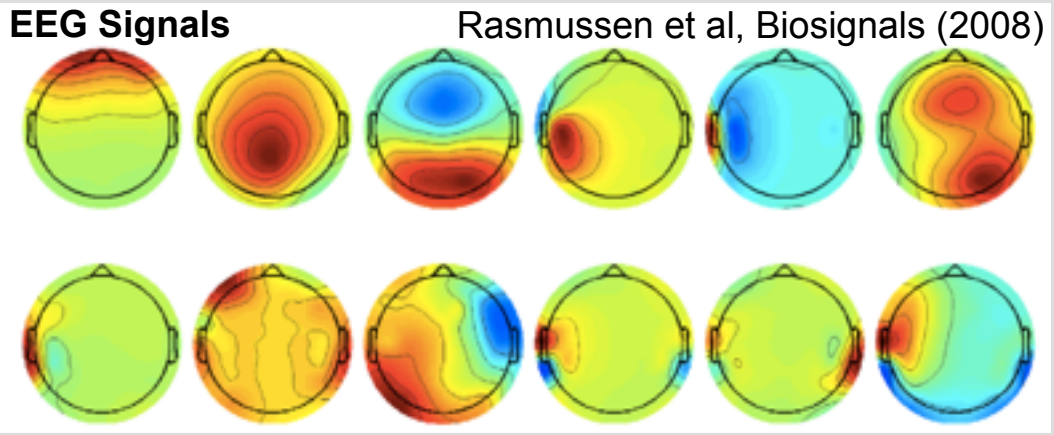


$$\psi_2(x, y) = ax + by$$

TICA is the variationally optimal approximation to the Markov operator eigenfunctions when using **linear combinations of mean-free molecular coordinates**
Perez-Hernandez et al, **JCP**, 139, 1502 (2013)

Applications of TICA

EEG Signals Rasmussen et al, Biosignals (2008)



Molecular Dynamics

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

Naritomi and Fuchigami, **JCP** (2011)
Slow dynamics in protein fluctuations …

Schwantes and Pande, **JCTC** (2013)
Improvements in Markov state modeling …

Special case: Markov state model

Choice of basis set: Characteristic functions on sets S_1, ..., S_n that partition phase space

$$\chi_i(\mathbf{x}_t) = \begin{cases} 1 & \mathbf{x}_t \in S_i \\ 0 & \text{else.} \end{cases}$$

where $\mathbf{q}$ are the position coordinates of the system.

Covariance matrices:

$$\begin{aligned} \mathbf{C}^0 &= \mathbf{X}_0^\top \mathbf{X}_0 = \mathrm{diag}(\sum_k N_{1k}(\tau), \ldots, \sum_k N_{nk}(\tau)) \\ \mathbf{C}^\tau &= \mathbf{X}_0^\top \mathbf{X}_\tau = [N_{ij}(\tau)]_{i,j=1,\ldots,n} \end{aligned}$$

where $N_{ij}(\tau)$ are the number of transitions from state i to j at lag time τ.

GEV problem:

$$\begin{aligned} \mathbf{C}^\tau \mathbf{R} &= \mathbf{C}^0 \mathbf{R} \hat{\mathbf{\Lambda}} \\ \mathbf{P}\mathbf{R} &= \mathbf{R} \hat{\mathbf{\Lambda}} \end{aligned}$$

with

$$\mathbf{P} = [p_{ij}(\tau)] \qquad p_{ij}(\tau) = \frac{N_{ij}(\tau)}{\sum_k N_{ik}(\tau)}$$

Related: Dynamic mode decomposition (DMD)

DMD: Schmidt, Sesterhenn, **Ann. Meet. APS Div. Fluid Mech.** (2008)
exact DMD: Tu, Rowley, Luchtenburg, Brunton, Kutz, **J. Comput. Dyn.** (2014)

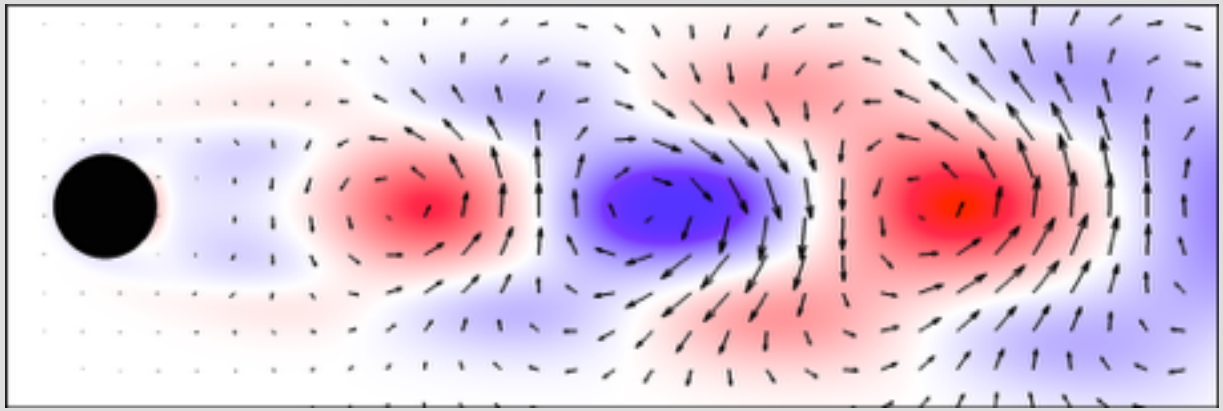
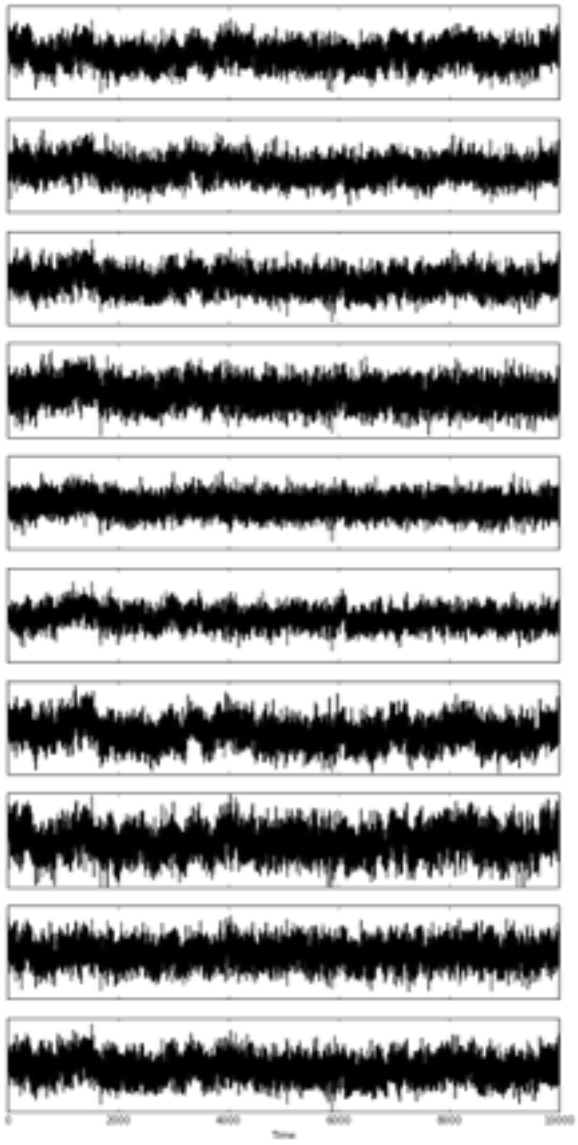


Image from: Chen, Tu, Rowley, J. Nonlinear Sci 22: 887-916 (2012)

DMD: computes left eigenvectors of K ==> dynamic modes
TICA/VAMP compute right eigenvectors of K ==> direct approximation of eigenfunctions

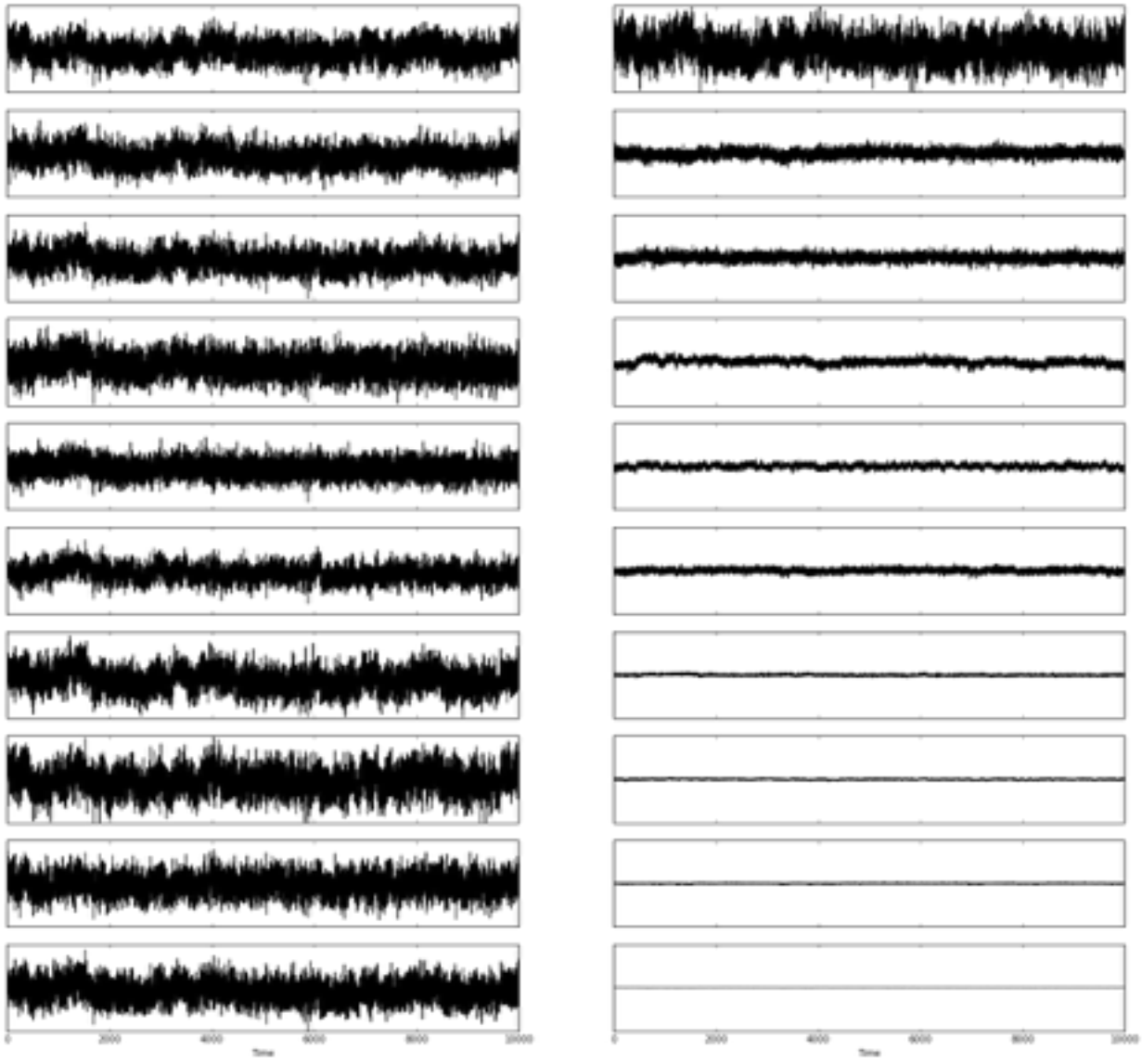
Examples

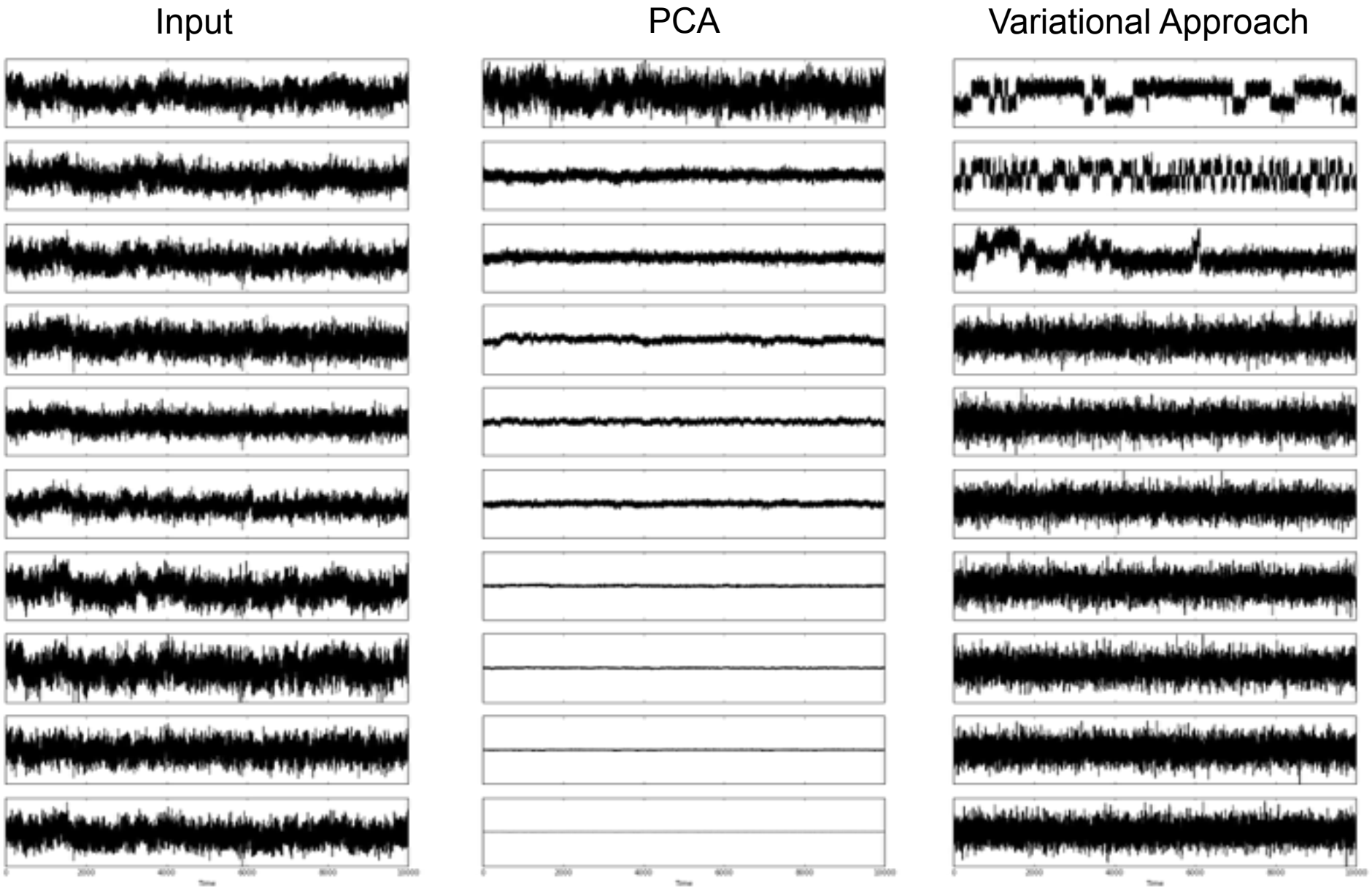
Input



Input

PCA

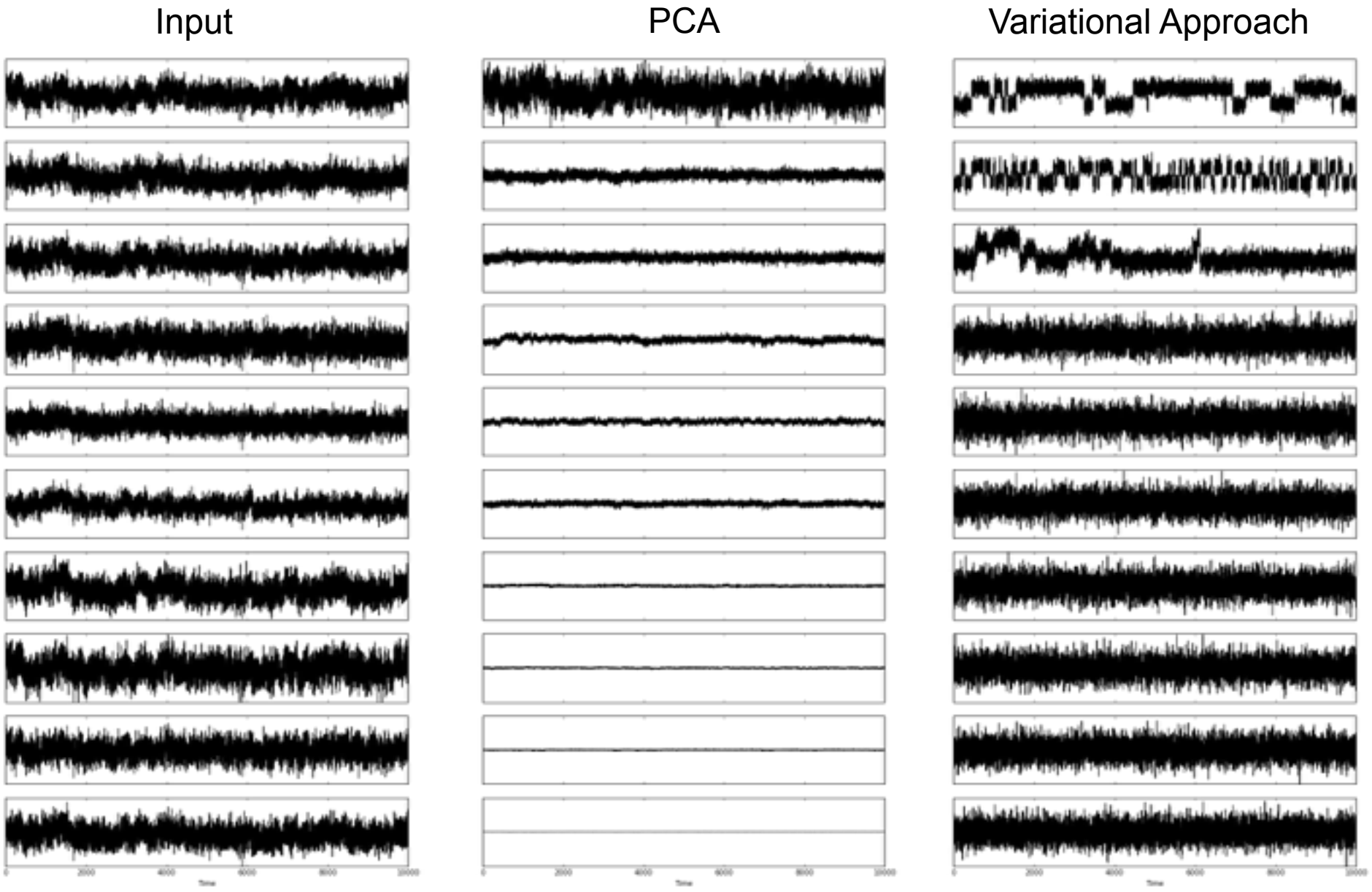




Variational Approach

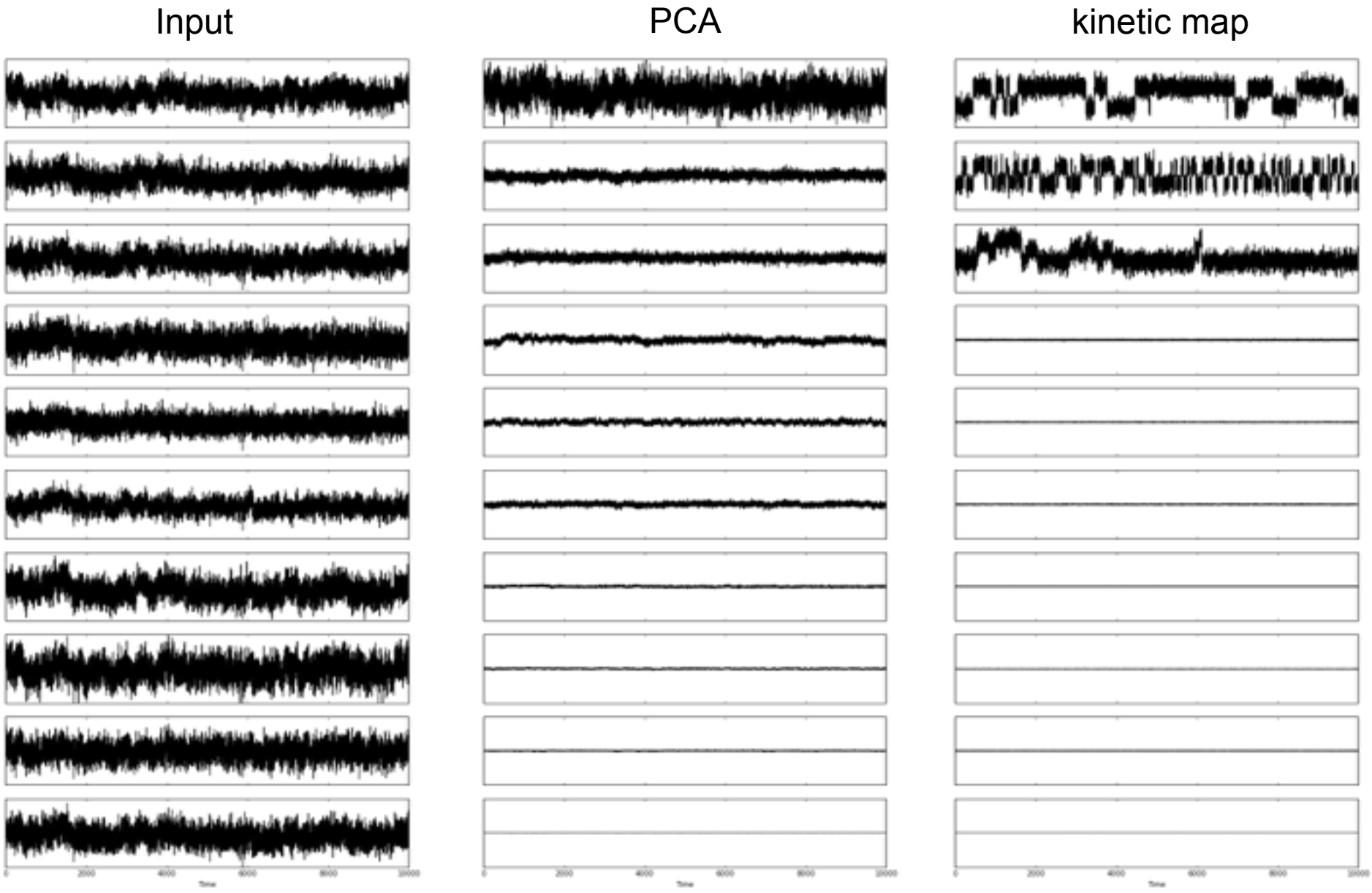
Noé and Nüske, **MMS** 11, 635-655 (2013)
Nüske et al, **JCTC** 10, 1739-1752 (2014)

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



Variational Approach

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



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)

Expansion in molecular order parameters

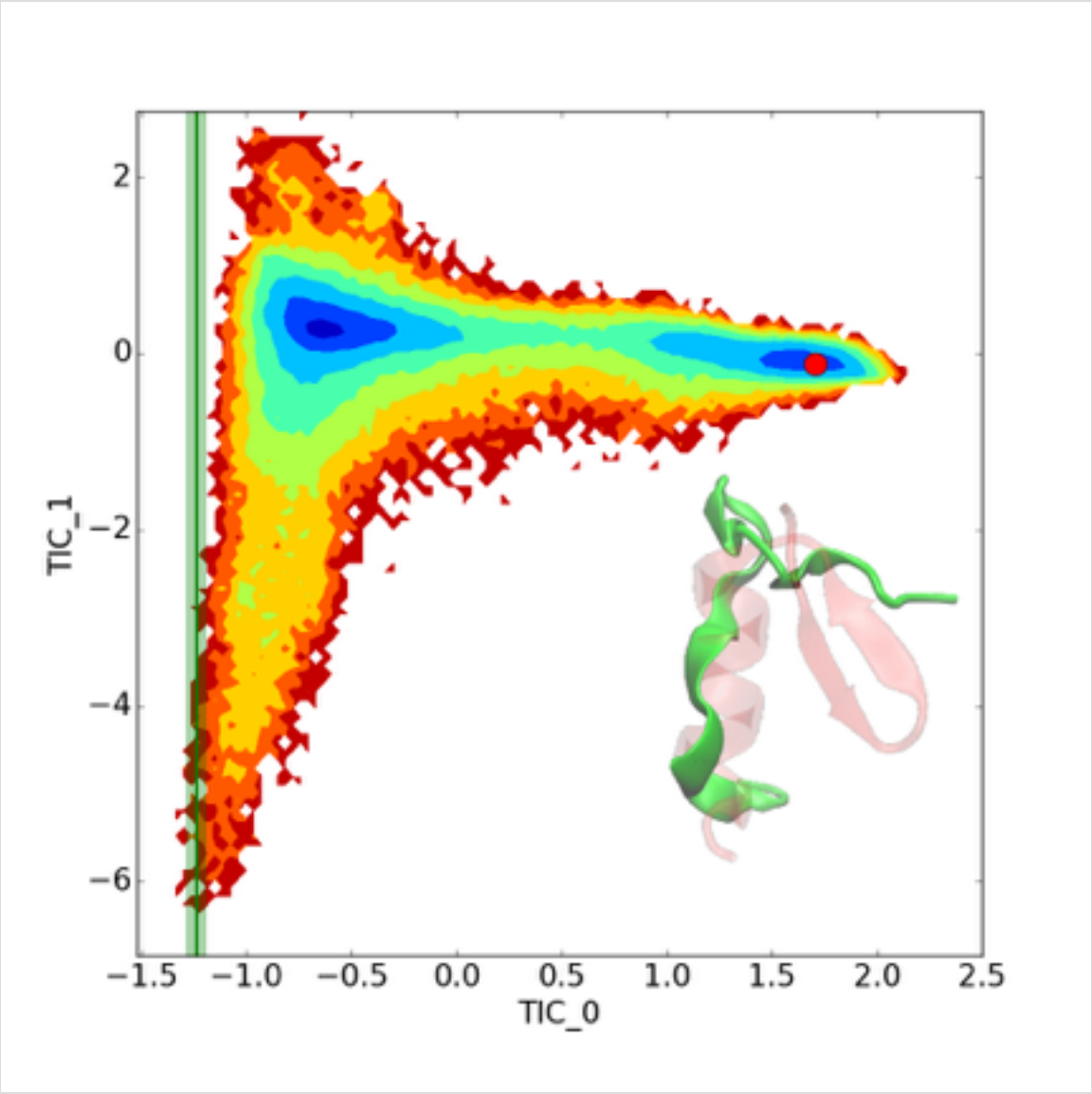
Candidates x_1…x_n

- Cartesian coordinates
- Distances between heavy atoms or Ca's
- cos/sin of dihedrals
- Gaussian densities around sampled configurations (e.g. in RMSD space)
- Field variables or coordination numbers
- Your favorite order parameter
- All of them

Simple Ansatz:

$$\psi_i(\mathbf{x}) = b_{i,1}x_1 + \ldots + b_{i,n}x_n$$

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



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

Doing this in PyEMMA…


```
from pyemma import coordinates, plots

# define input
feat = coordinates.featurizer("protein.pdb")
feat.add_residue_mindist()
data = coordinates.source("protein.xtc", feat)

# TICA
tica = coordinates.tica(data, lag=100, dim=2)
Y = tica.get_output()[0]

# visualize
plots.plot_free_energy(Y[:, 0], Y[:, 1])
```

Estimation

Estimation problems

1) Nonreversible estimate

$$\hat{\mathbf{C}}(0) = \frac{1}{N}\mathbf{X}^\top\mathbf{X}$$
$$\hat{\mathbf{C}}(\tau) = \frac{1}{N}\mathbf{X}^\top\mathbf{Y}$$

2) Standard reversible estimate

$$\hat{\mathbf{C}}_{\mathrm{sym}}(0) \approx \frac{1}{2N}\left(\mathbf{X}^\top\mathbf{X} + \mathbf{Y}^\top\mathbf{Y}\right)$$
$$\hat{\mathbf{C}}_{\mathrm{sym}}(\tau) \approx \frac{1}{2N}\left(\mathbf{X}^\top\mathbf{Y} + \mathbf{Y}^\top\mathbf{X}\right)$$

This leads to a strong bias for out-of equilibrium data!

Equilibrium estimation from out-of equilibrium data

1. Extend data

$$\mathbf{X}^{\ddagger} = [\mathbf{X}\,\mathbf{1}]$$

2. Compute (nonreversible) Koopman matrix $\mathbf{K}$ from data

3. Compute reweighting vector $\mathbf{u}$ using the eigenvalue problem

$$\hat{\mathbf{u}} = \hat{\mathbf{u}}\mathbf{K}$$

and normalization $\mathbf{u} = \frac{1}{\hat{u}_{n+1}}\hat{\mathbf{u}}$

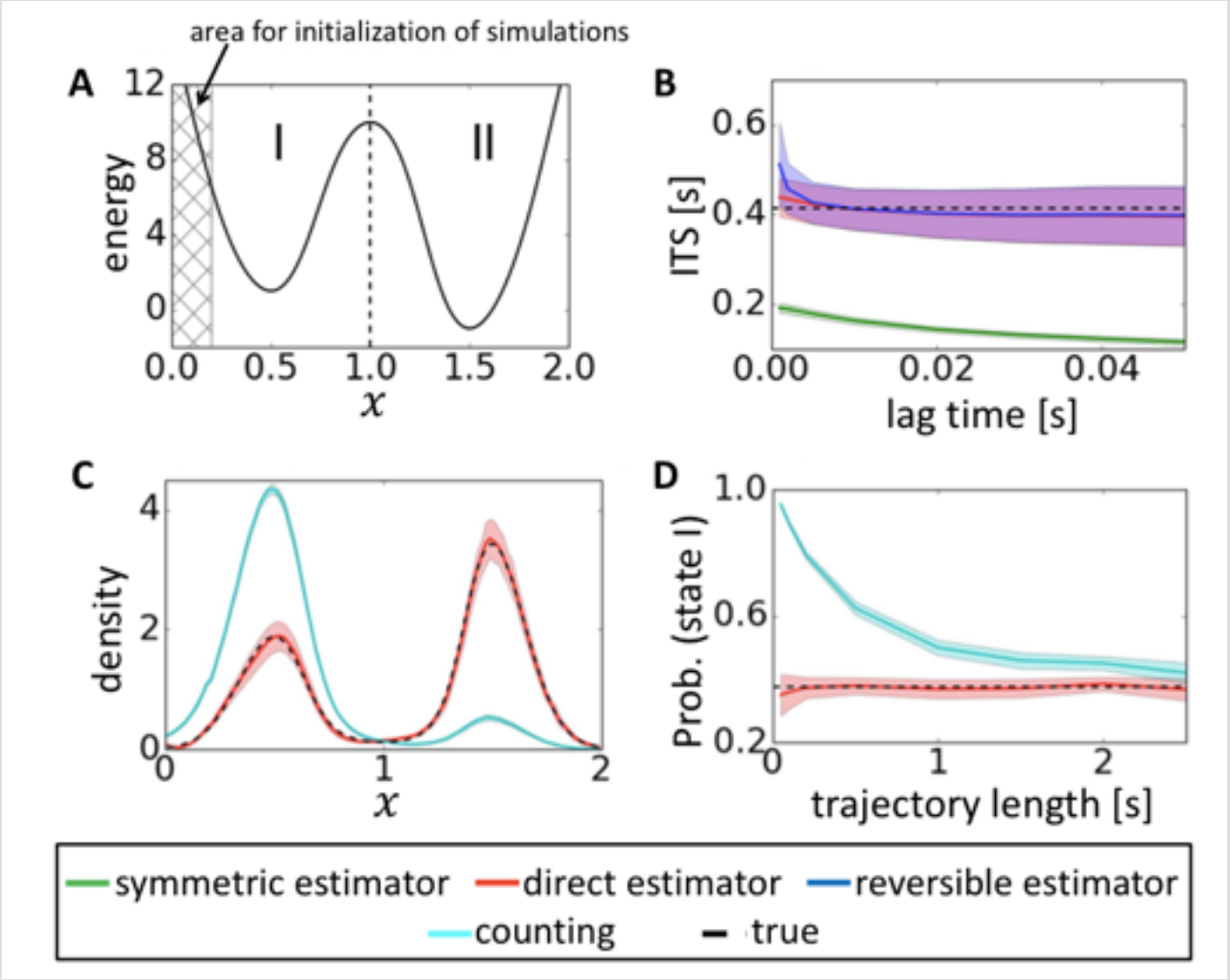
4. Equilibrium expectations:

$$\mathbb{E}_{\pi}[f(\mathbf{x}_t)] = \frac{1}{N}\sum_{t=1}^{T-\tau} \mathbf{u}\,\boldsymbol{\chi}(\mathbf{x}_t)\,f(\mathbf{x}_t)$$

$$\mathbb{E}_{\pi}[\chi_i(\mathbf{x}_t)\chi_j(\mathbf{x}_{t+k\tau})] = \frac{1}{N}\sum_{t=1}^{T-\tau} \mathbf{u}\,\boldsymbol{\chi}(\mathbf{x}_t)\,\boldsymbol{\chi}(\mathbf{x}_t)\boldsymbol{\chi}^{\top}(\mathbf{x}_t)\mathbf{K}^k$$

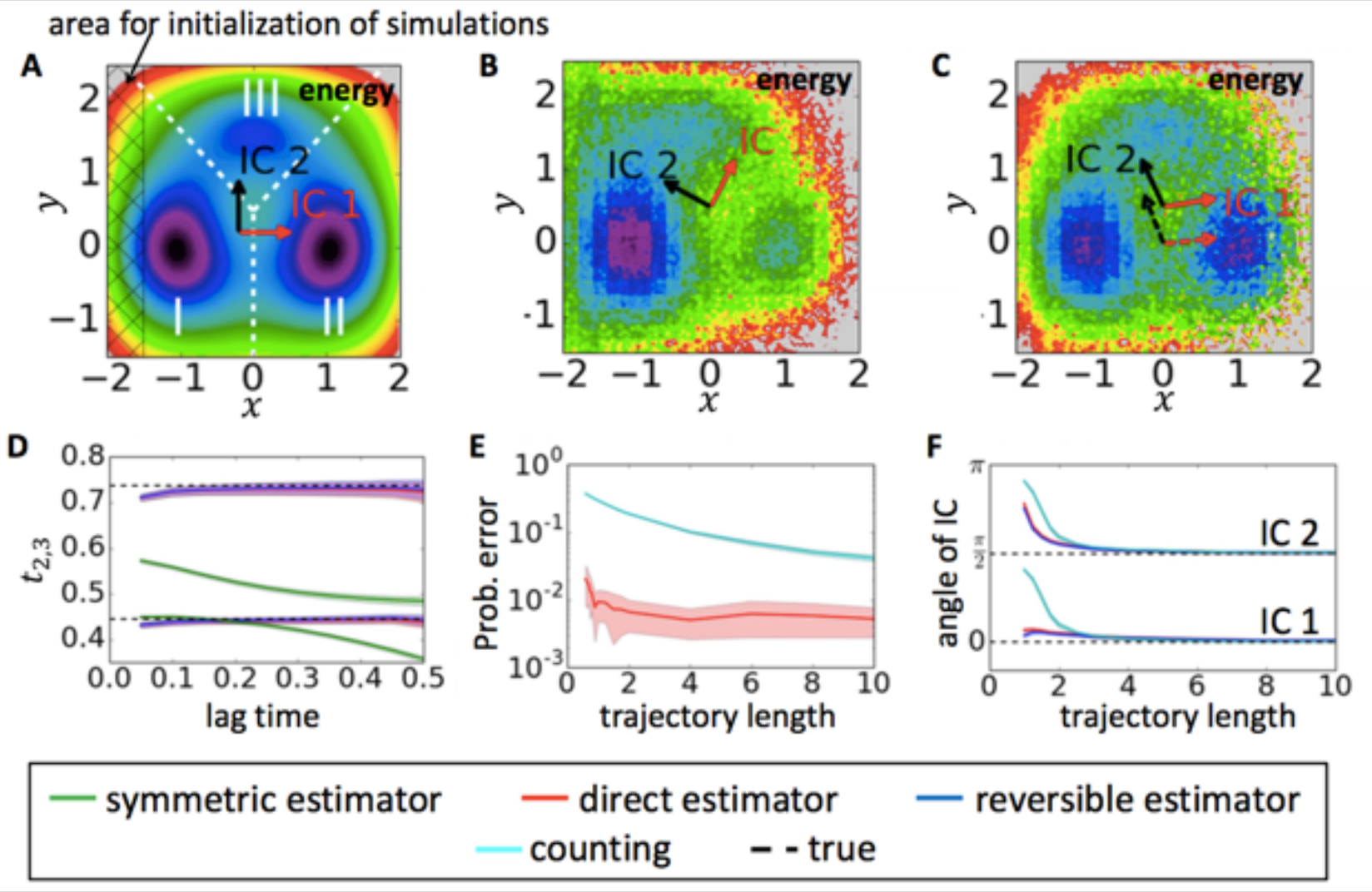
Wu et al, **arXiv** 1610.06773 (2016)

Equilibrium estimation from out-of equilibrium data



Wu et al, **arXiv** 1610.06773 (2016)

Equilibrium estimation from out-of equilibrium data



Wu et al, **arXiv** 1610.06773 (2016)

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

European Research Council
Established by the European Commission

SFB 958

SFB 740

EINSTEIN Foundation.de