Multiscale characterization of macromolecular dynamics

Cecilia Clementi

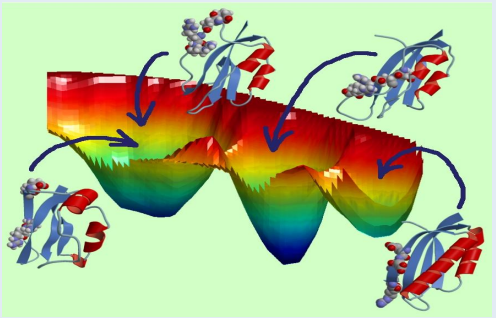
Center for Theoretical Biological Physics
& Department of Chemistry
Rice University, Houston, TX USA

Einstein Visiting Professor
Multiscale Modeling of Biophysical Systems
Department of Mathematics and Computer Science
Freie Universität, Berlin, Germany

Einstein Stiftung Berlin
Einstein Foundation Berlin

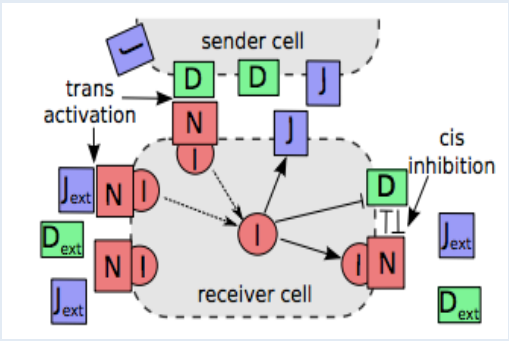
Clementi group: Multiscale characterization of macromolecular systems

Low dimensional representation of macromolecular dynamics

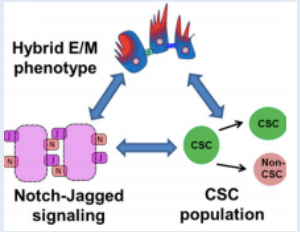


JCP in press (2017)
Curr. Op. Struct. Biol. 43, 141 (2017))
JCTC 12 5620 (2016)
JCTC 11, 5947 (2015)
JCTC 11, 5002 (2015)
JCP 142, 025103 (2015)
JCP 139, 145102 (2013)
Ann. Rev. Phys. Chem. 64, 295 (2013)
Chem. Sci. 5, 1401 (2014)

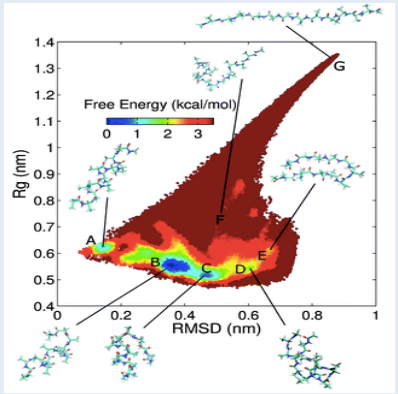
Theoretical framework for cell-cell signaling



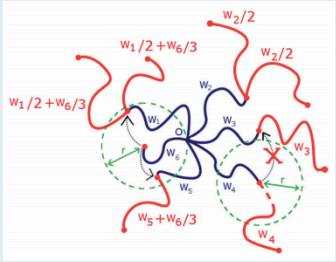
New J. Phys. 17, 055021 (2015)
PNAS 112, E402 (2015)



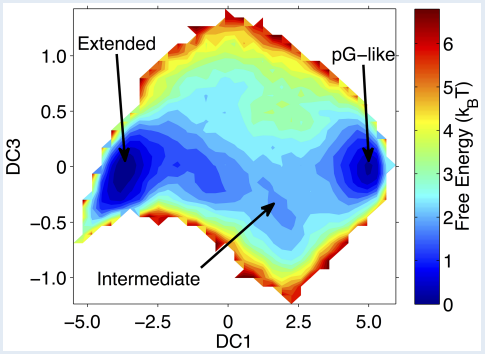
Adaptive sampling



eScience (2017) in press
PCCP 16, 19181 (2014)
JPCB 117, 12769 (2013)



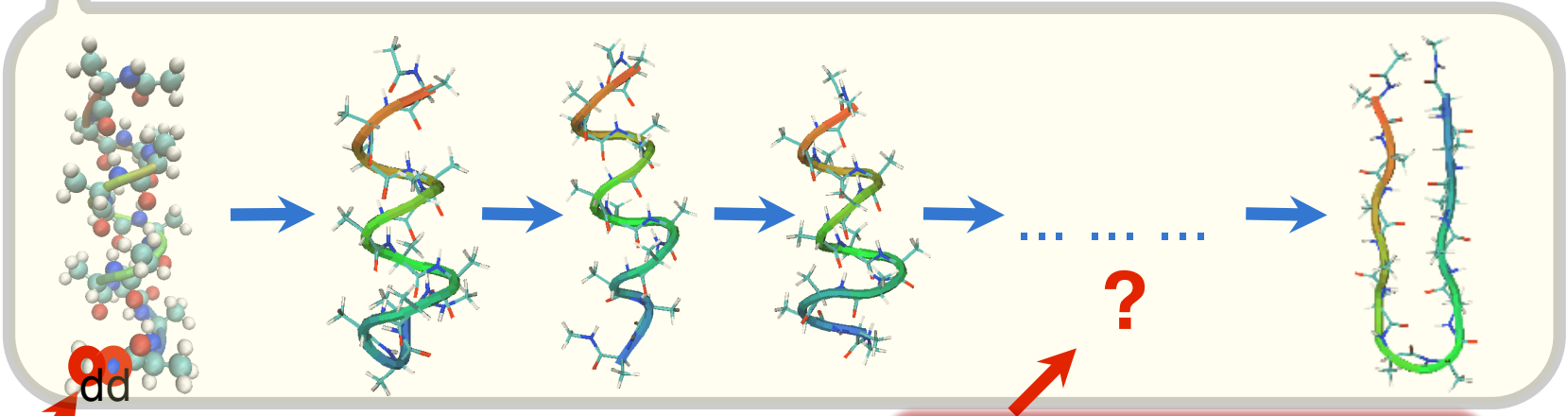
Systematic coarse-graining and multiscale approaches



PLOS Comp. Biol 10, e1003797 (2014)
JPCB 116, 8494 (2012)

Macromolecular Dynamics: Main goals and challenges

Trajectories in equilibrium distribution → Observables to compare with the experiment
Mechanism from the dynamics
New predictions



1. Force Field

A set of parameters and equations describing the interactions between atoms

2. Sampling

Can the simulations cover the rare events we are interested in?

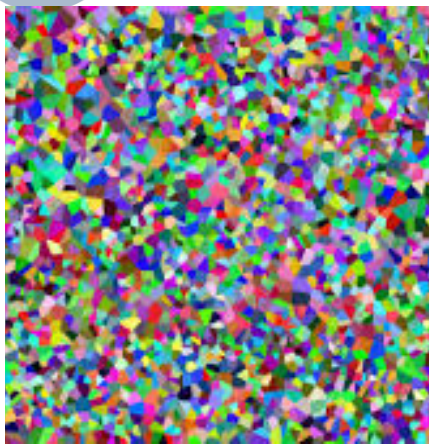
3. Data analysis

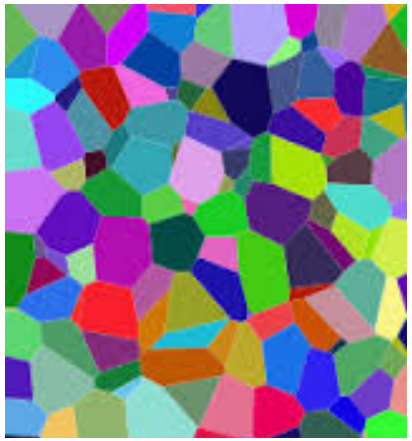
How do we understand the molecular mechanisms from the data?

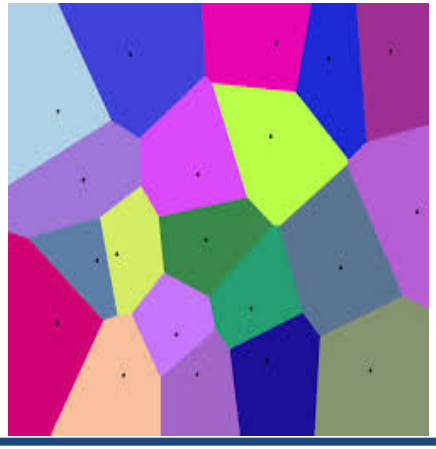
Coarse-Graining

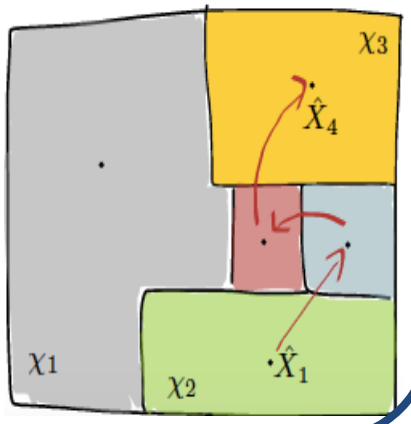
$\mathbb{R}^{3N}$

coarse-graining in conformation space



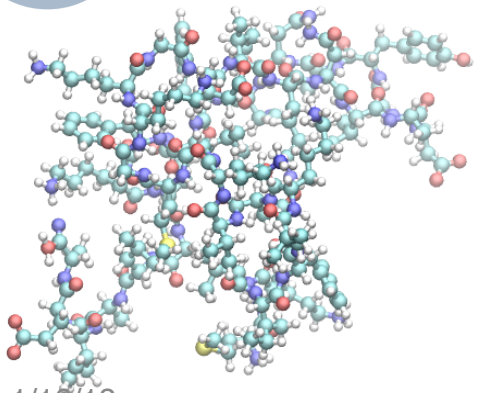


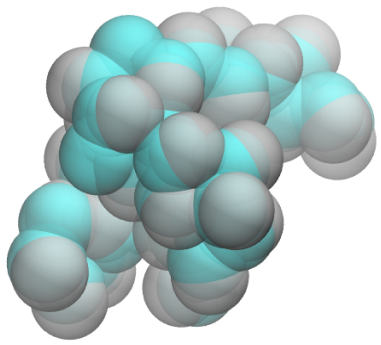


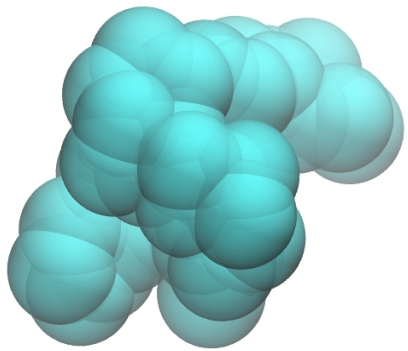


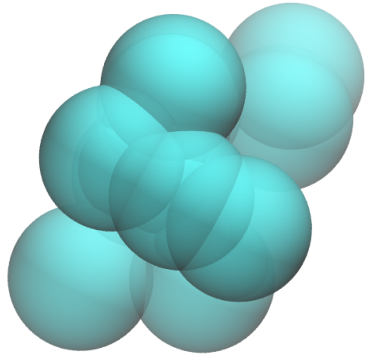
$\mathbb{R}^{3}$

coarse-graining in structural space





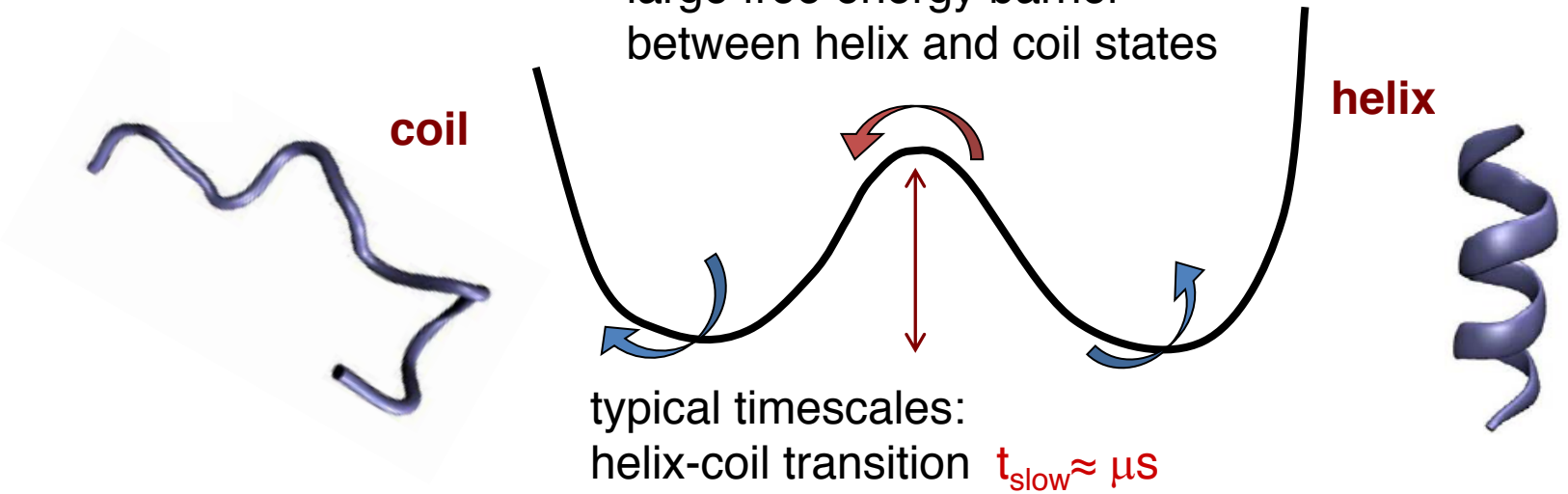




Macromolecular systems are often characterized by sets of different timescales, separated by large gaps

Example: helix-coil transition

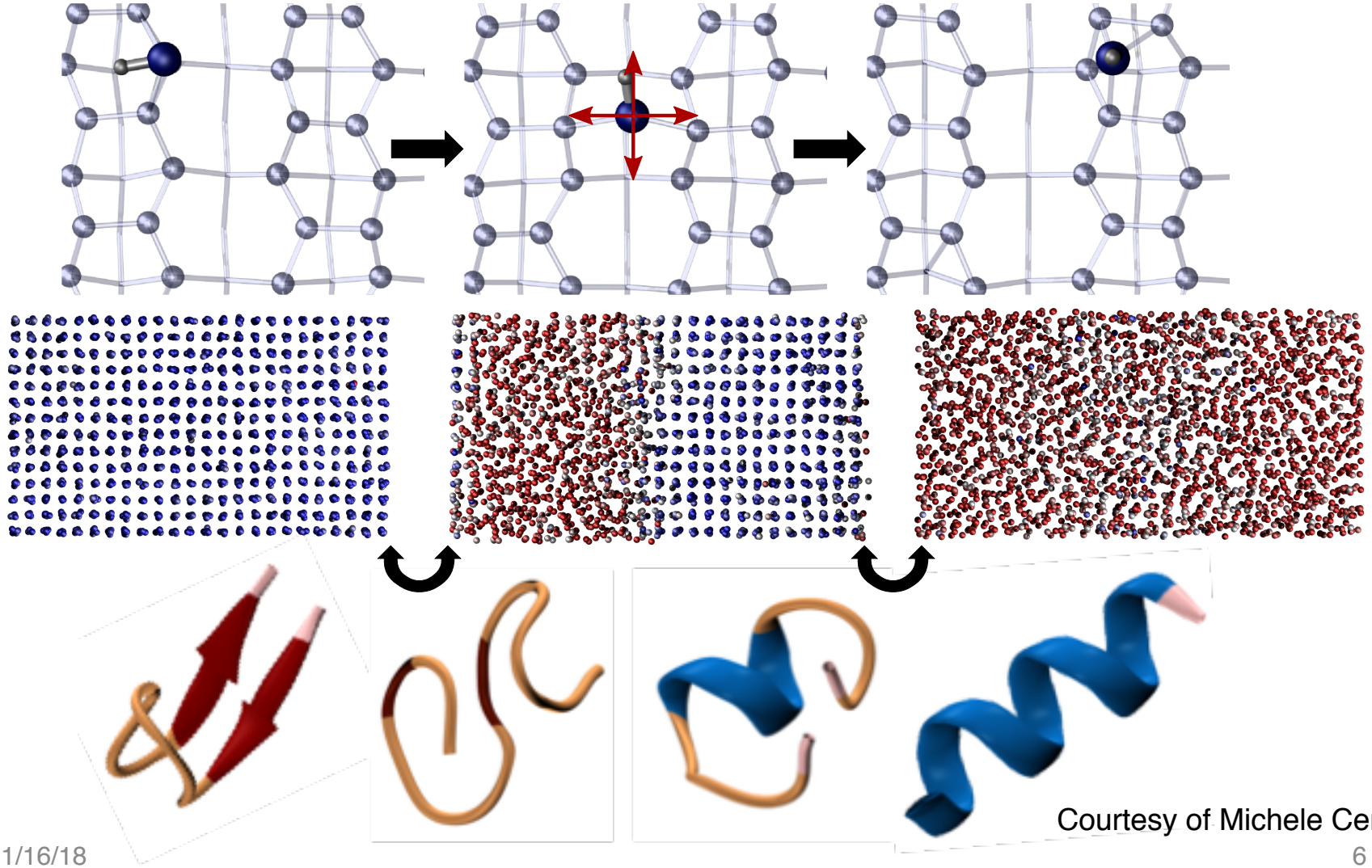
large free energy barrier
between helix and coil states



typical timescales:
helix-coil transition $t_{slow} \approx \mu s$
diffusion within a state $t_{fast} \approx$ ns

Can we extract the collective variables associated with the (rare) barrier-crossing events?

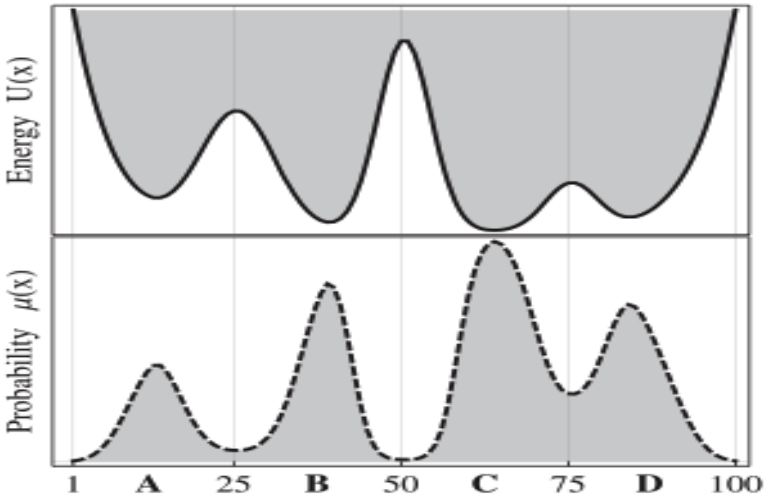
Can we extract the collective variables associated with the (rare) barrier-crossing events?



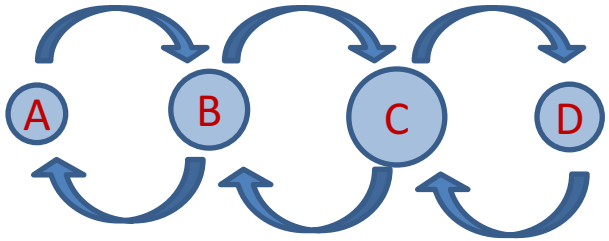
Courtesy of Michele Ceriotti

Can we extract the collective variables associated with the (rare) barrier-crossing events?

1D toy example

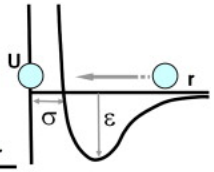


Prinz *et al.* J. Chem. Phys.134, 174105 (2011)

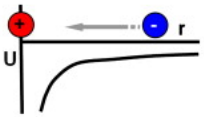


biomolecular system:
very high dimensionality

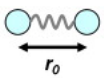
$$U = \sum_{i<j}\sum 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$



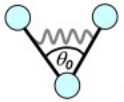
$$+ \sum_{i<j}\sum \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}$$



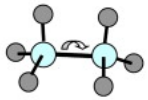
$$+ \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2$$



$$+ \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2$$



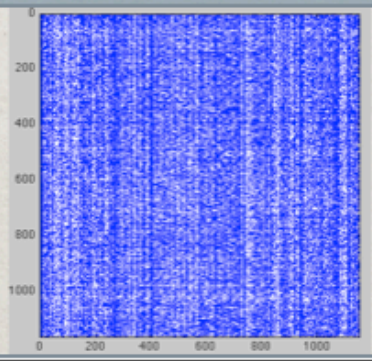
$$+ \sum_{torsions} k_\phi [1 + \cos(n\phi - \delta)]$$



Similar problems arise in different research fields

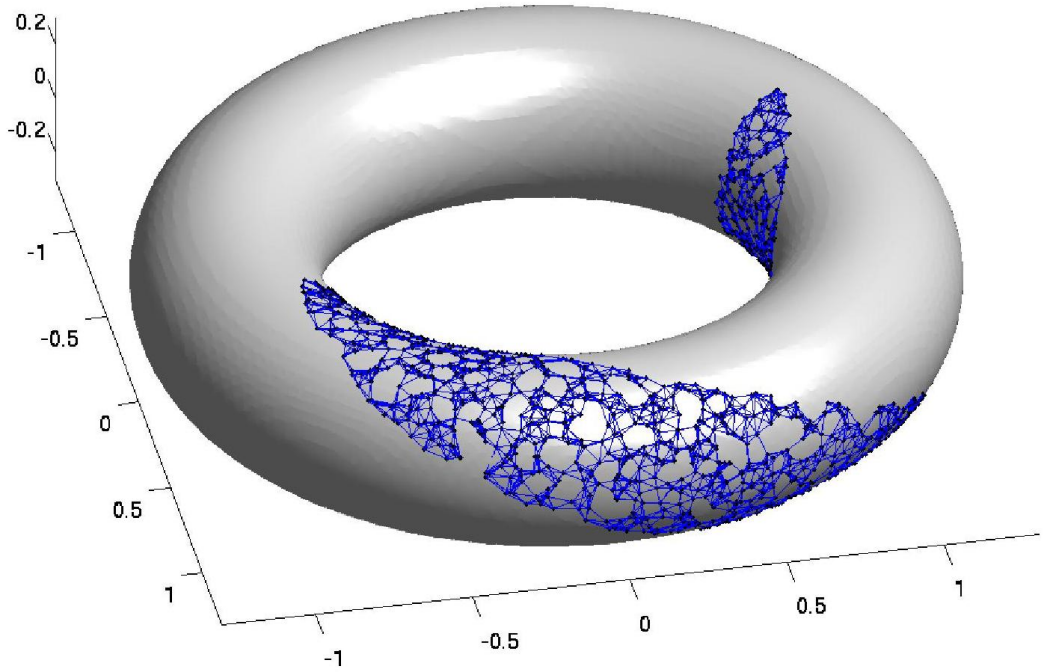
(computer science, engineering, applied math, statistics, biology, …)

Examples: classification of documents, image recognition

Data	Picture	Problems
Approx. 1000 articles from ScienceNews. Representation: a document-term matrix, with about 1000 terms.		Automatically sort articles into categories, given only a small set labeled by experts. Navigate the library.
A database of ~60000 grayscale 28x28 images of handwritten digits 0-9.		Automatically recognize digits (e.g. for ZIP codes, checks, etc...)

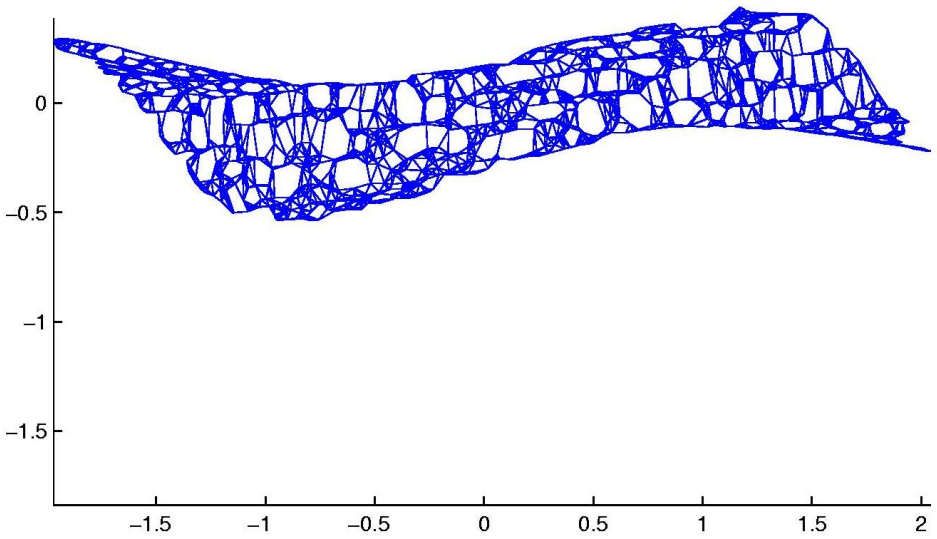
Courtesy of M.Maggioni

Mathematically this is a problem of non-linear dimensionality reduction



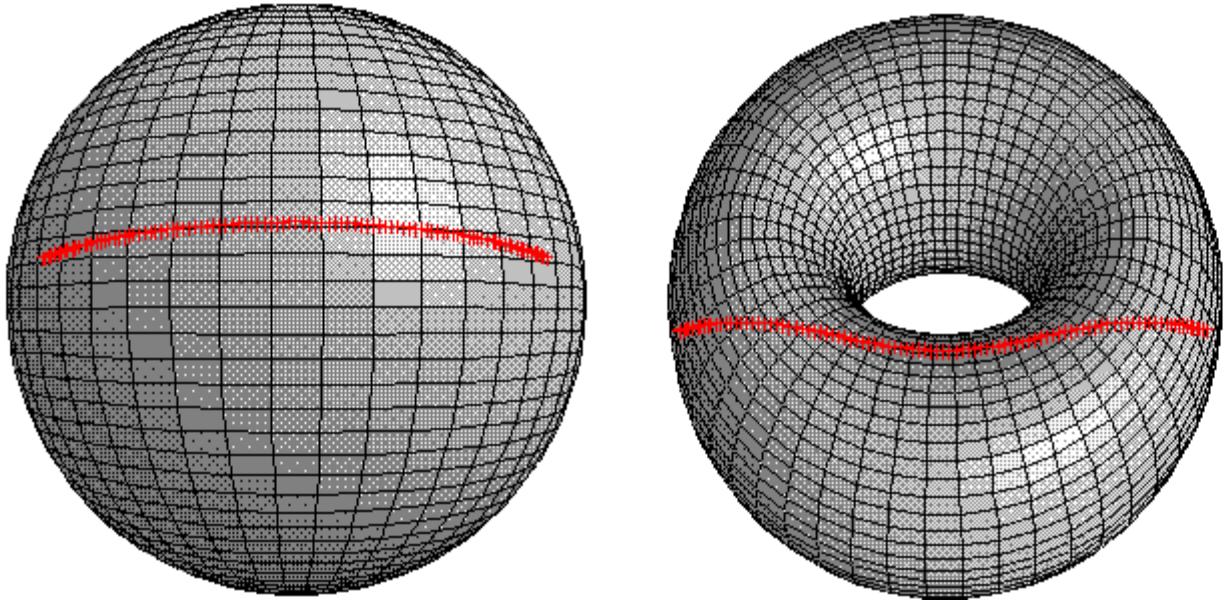
Example:

a set of points on a torus in 3d defines a 2d embedded “surface”



ISOMAP: Use geodesics to define the manifold

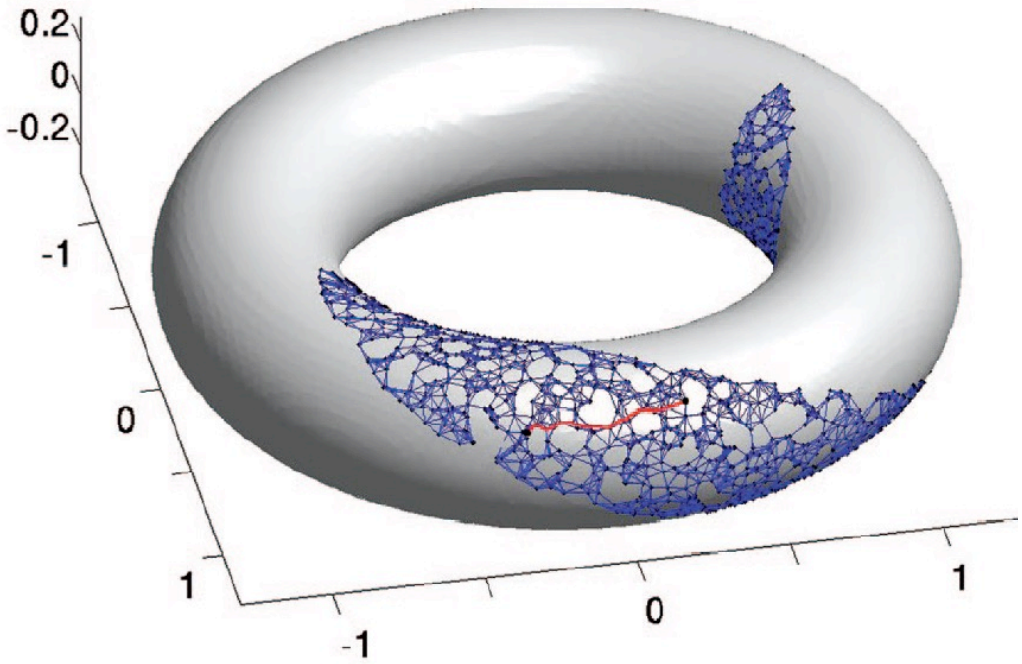
A geodesic is the shortest path between two points



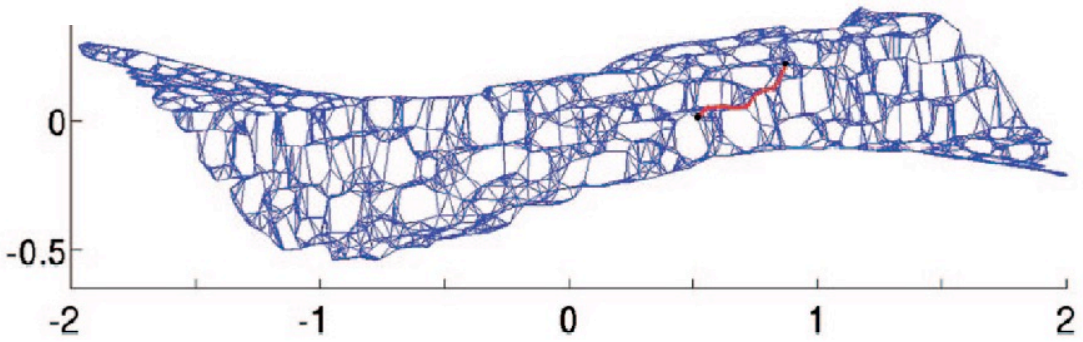
If we know the geodesics between any couple of points
then we know everything about the geometry of the system

ISOMAP idea from:
J. Tenenbaum, V. de Silva, & J. Langford
Science 290, 2319–2323 (2000)

Use geodesics to define the manifold

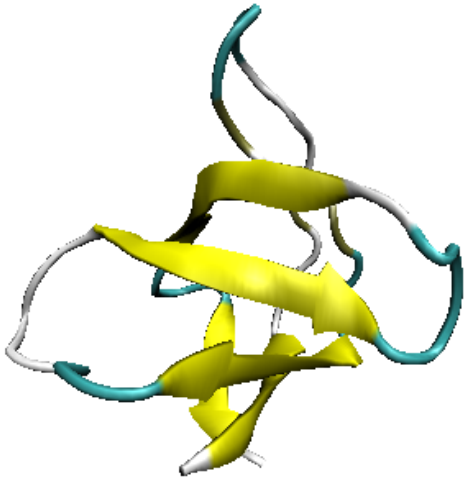


Network of nearest neighbors
can serve to approximate geodesics
between any pairs of points

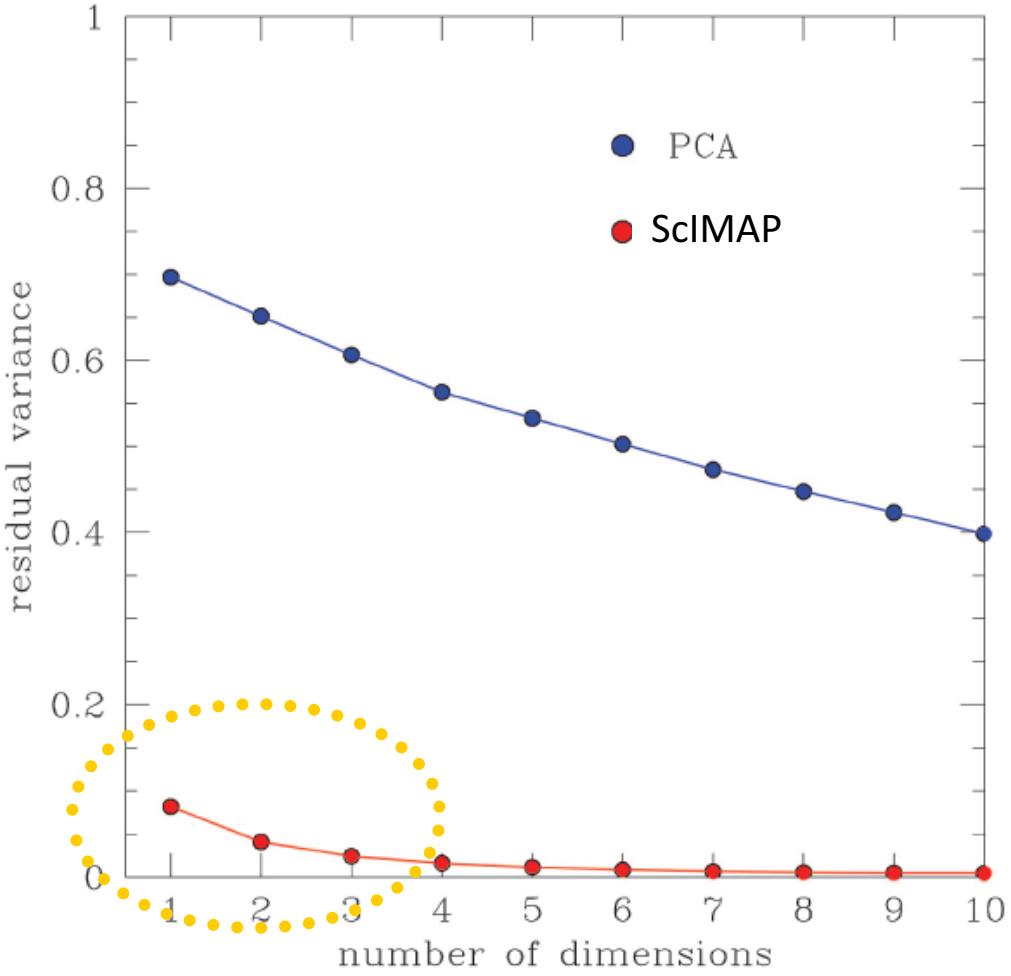


P. Das, M. Moll, H. Stamati, L.E. Kavraki, & C.Clementi
Proc. Natl. Acad. Sci. USA 103, 9885-9890 (2006)

Application to SH3 folding dynamics



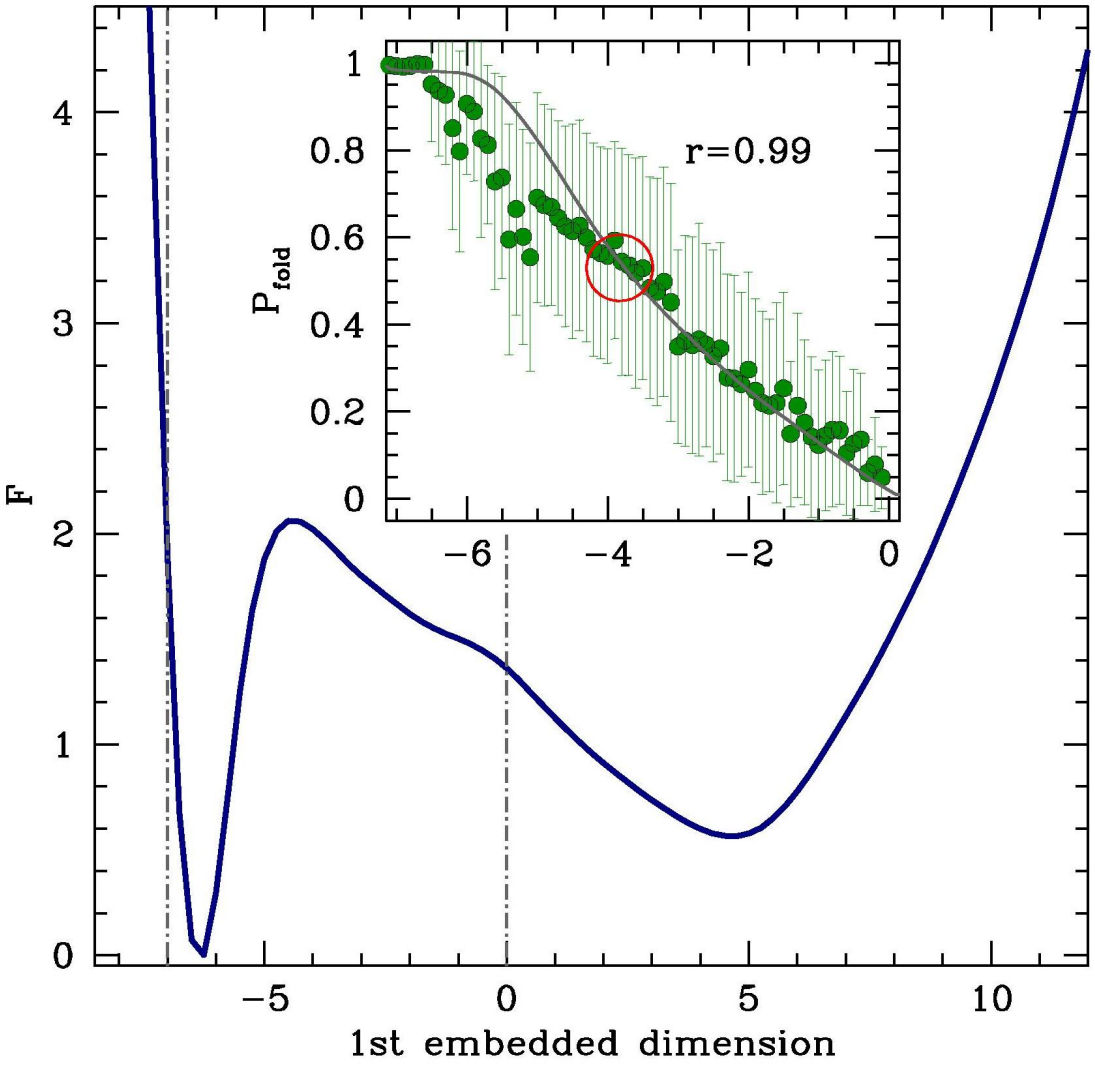
Native structure
of SH3 protein



P. Das, M. Moll, H. Stamati, L.E. Kavraki, & C.Clementi
Proc. Natl. Acad. Sci. USA 103, 9885-9890 (2006)

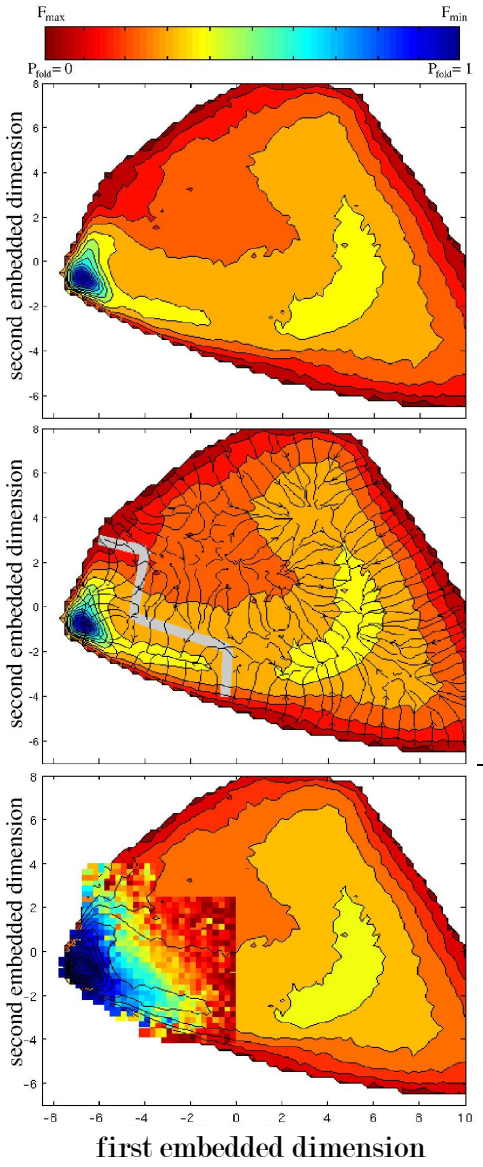
Application of nonlinear dimensionality reduction to SH3 folding

Free energy as a function of the first "collective" coordinate



P. Das, M. Moll, H. Stamati, L.E. Kavraki, & C.Clementi
Proc. Natl. Acad. Sci. USA 103, 9885-9890 (2006)

Application of nonlinear dimensionality reduction to SH3 folding



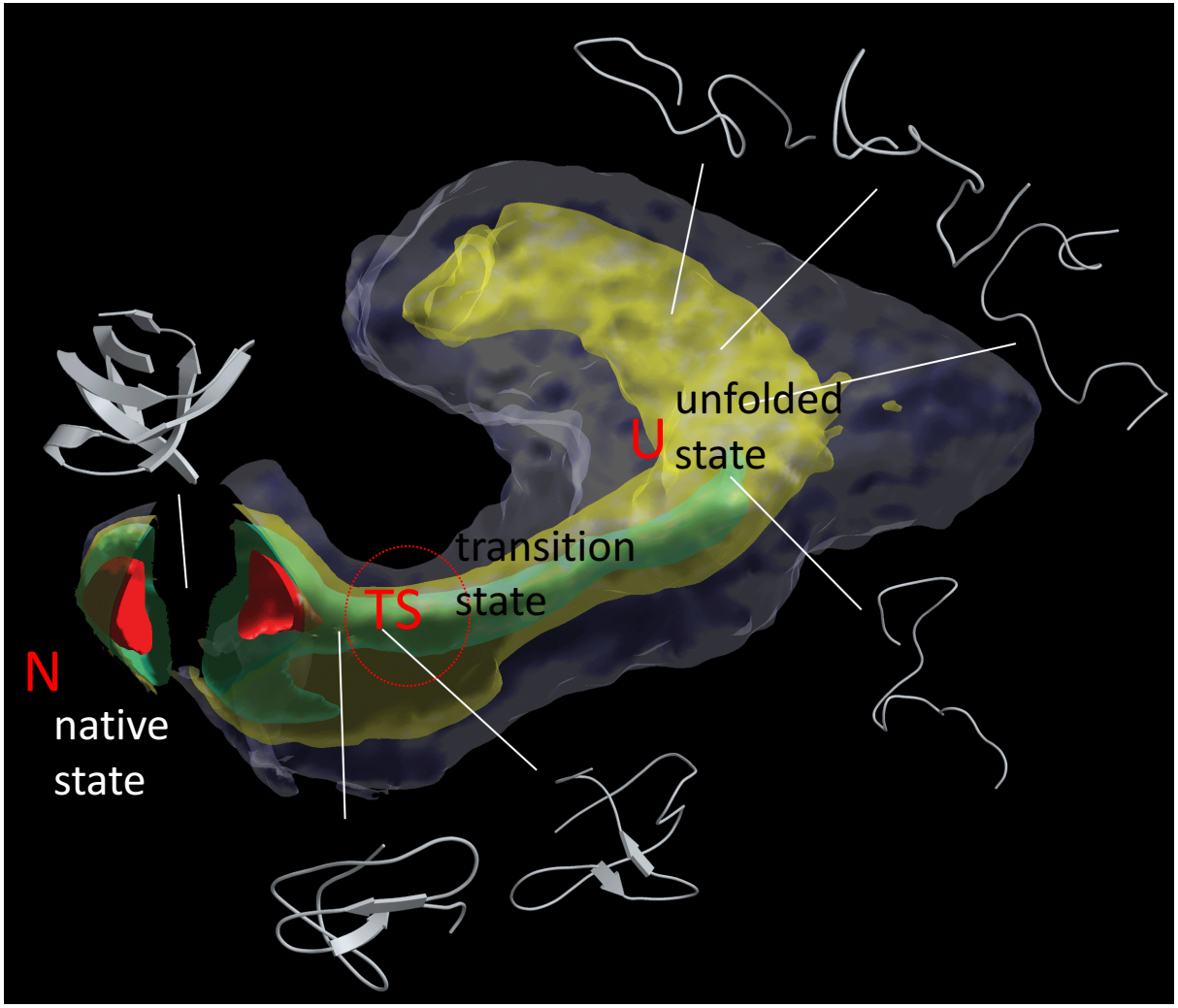
Contour plot of the 2D embedded free energy with two distinct minima – folded and unfolded state. The main folding route closely follows the 1st embedded dimension.

A free energy gradient field can be associated to the free energy contour plot.

The region with $P_{fold} \sim 0.5$ matches the separatrix identified by the gradient flux.

P. Das, M. Moll, H. Stamati, L.E. Kavraki, & C.Clementi
Proc. Natl. Acad. Sci. USA 103, 9885-9890 (2006)

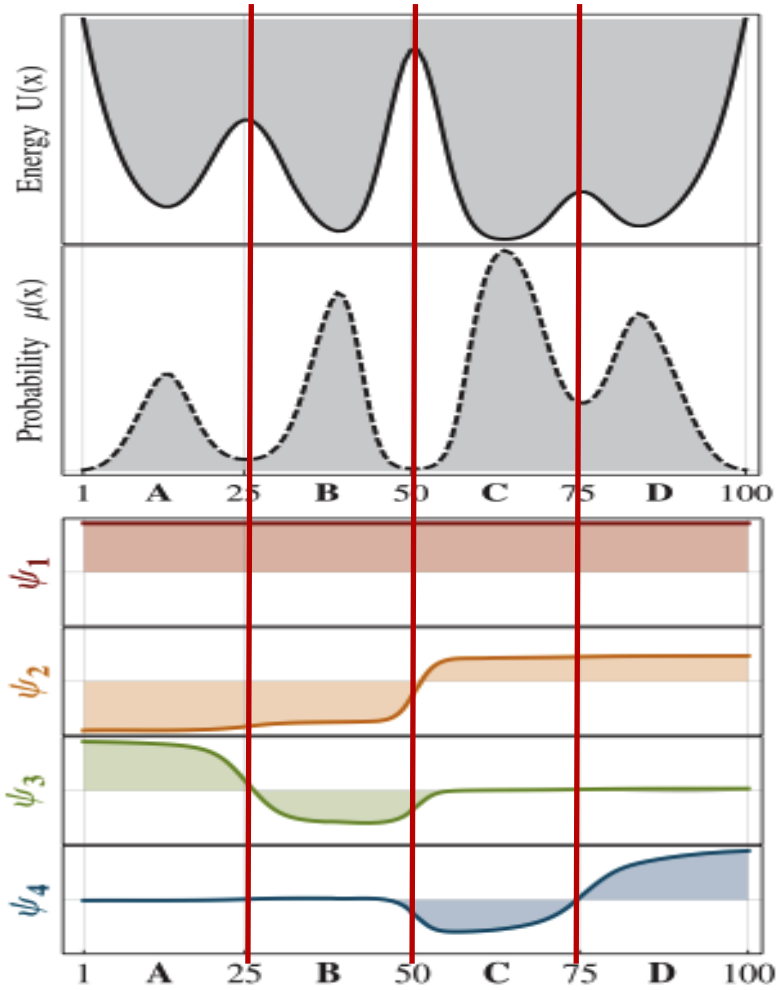
Application of nonlinear dimensionality reduction to SH3 folding



P. Das, M. Moll, H. Stamati, L.E. Kavraki, & C.Clementi
Proc. Natl. Acad. Sci. USA 103, 9885-9890 (2006)

Can we extract the collective variables associated with the (rare) barrier-crossing events?

1D toy example

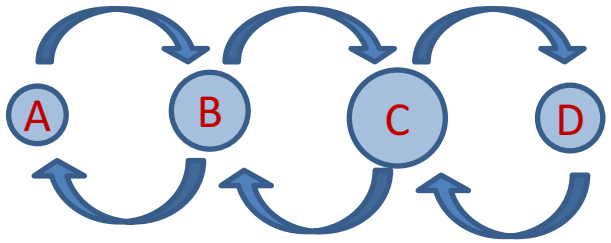


Prinz *et al.* J. Chem. Phys.134, 174105 (2011)

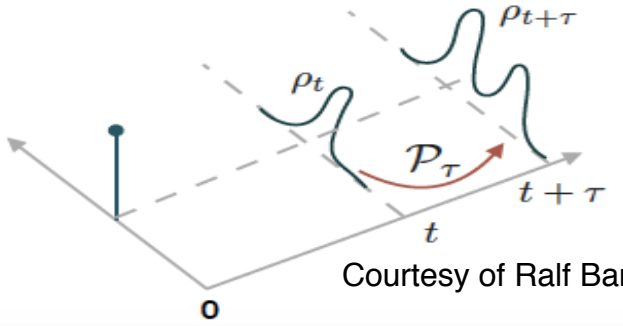
Mathematical answer:

Estimate the dominant **eigenvalues** and **eigenfunctions** of the backward Markov propagator underlying the MD

Schütte, C. *et al.* J. Comput. Phys. 151, 146 (1999)



Dominant **eigenvalues** and **eigenfunctions** of the backward Markov propagator underlying the MD provide information of both the equilibrium and the slow kinetic properties of a molecular system



Courtesy of Ralf Banisch

Propagator of probability density of states, $\rho_t(\mathbf{x})$

$$\begin{aligned}\rho_{t+\tau}(\mathbf{y}) &= \int_{\mathbf{x}\in\Omega} \rho_t(\mathbf{x})\, p_\tau(\mathbf{y}\mid\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \mathcal{P}\circ\rho_t(\mathbf{x})\end{aligned}$$

$$\begin{aligned}&\mathcal{P}(\tau)\circ\rho(\mathbf{x}) \\ &\approx \textstyle\sum_{k=1}^{m} e^{-\kappa_k\tau}\langle\phi_k,\rho\rangle_{\pi^{-1}}\phi_k\end{aligned}$$

Backward propagator of weighted densities, $v_t(\mathbf{x}) = \rho_t(\mathbf{x})/\pi(\mathbf{x})$

$$\begin{aligned}v_{t+\tau}(\mathbf{y}) &= \frac{1}{\pi(\mathbf{y})}\int_{\mathbf{y}} p_\tau(\mathbf{y}\mid\mathbf{x})\,\pi(\mathbf{x})\,v_t(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \mathcal{T}\circ v_t(\mathbf{x}).\end{aligned}$$

$$\begin{aligned}&\mathcal{T}(\tau)\circ u(\mathbf{y}) \\ &\approx \textstyle\sum_{k=0}^{m} e^{-\kappa_k\tau}\langle\psi_k,u\rangle_\pi\psi_k\end{aligned}$$

Schütte, C. *et al.* J. Comput. Phys. 151, 146 (1999)

Different strategies and algorithms have been proposed to estimate these eigenfunctions and eigenvalues from MD trajectory data

Examples include:

- Markov State Model (MSM)

  Swope, B.C., Pitera J.D., Suits, F. J., Phys. Chem. B 108, 6571 (2004)
  Singhal, N., Pande, V.S., J. Chem. Phys. 123, 204909 (2005)
  Chodera, J.D. *et. al.,* J. Chem. Phys. 126, 155101 (2007)
  Noé et al. J. Chem. Phys. 126, 155102 (2007)
  Buchete, N.-V., Hummer, G., J. Phys. Chem. B 112, 6057 (2008)
  Sarich, M., Noé, F., Schütte, C., Multiscale Model. Simul. 8, 1154 (2010)
  Prinz, J.-H. *et al.,* J. Chem. Phys., 134, 174105 (2011)

- Diffusion Map (and Locally Scaled version, LSDMap)

  Coifman, R.R. *et al.,* PNAS 102, 7426 (2005)
  Nadler, B. *et al.,* NIPS 21, 113 (2006)
  Rohrdanz, M.A. *et al.,* J. Chem. Phys. 134, 124116 (2011)

- Variational Principle for Conformation Dynamics (including time-lagged Independent Component Analysis, tICA)

  Noé, F. *et al.*, Multiscale Model. Simul., 11, 635 (2013).
  Perez Hernandez, G. *et al.*, J. Chem. Phys., 139, 015102 (2013).
  Schwantes, C.R., Pande, V.S., J. Chem. Theory Comput. 9, 2000 (2013)

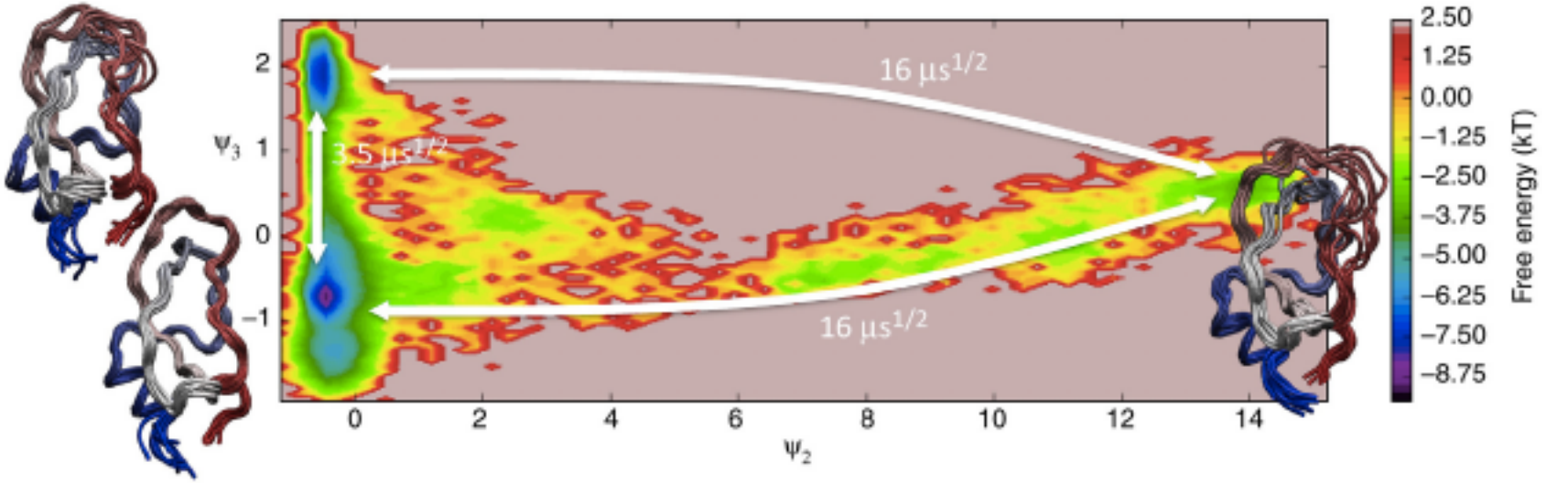
Available online at www.sciencedirect.com

ScienceDirect

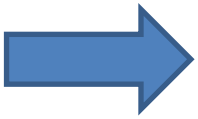
Current Opinion in Structural Biology

Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods

Frank Noé[1] and Cecilia Clementi[2]



Coarse-graining in conformation space: common challenges



Conceptual:

definition of suitable metric space

Computational:

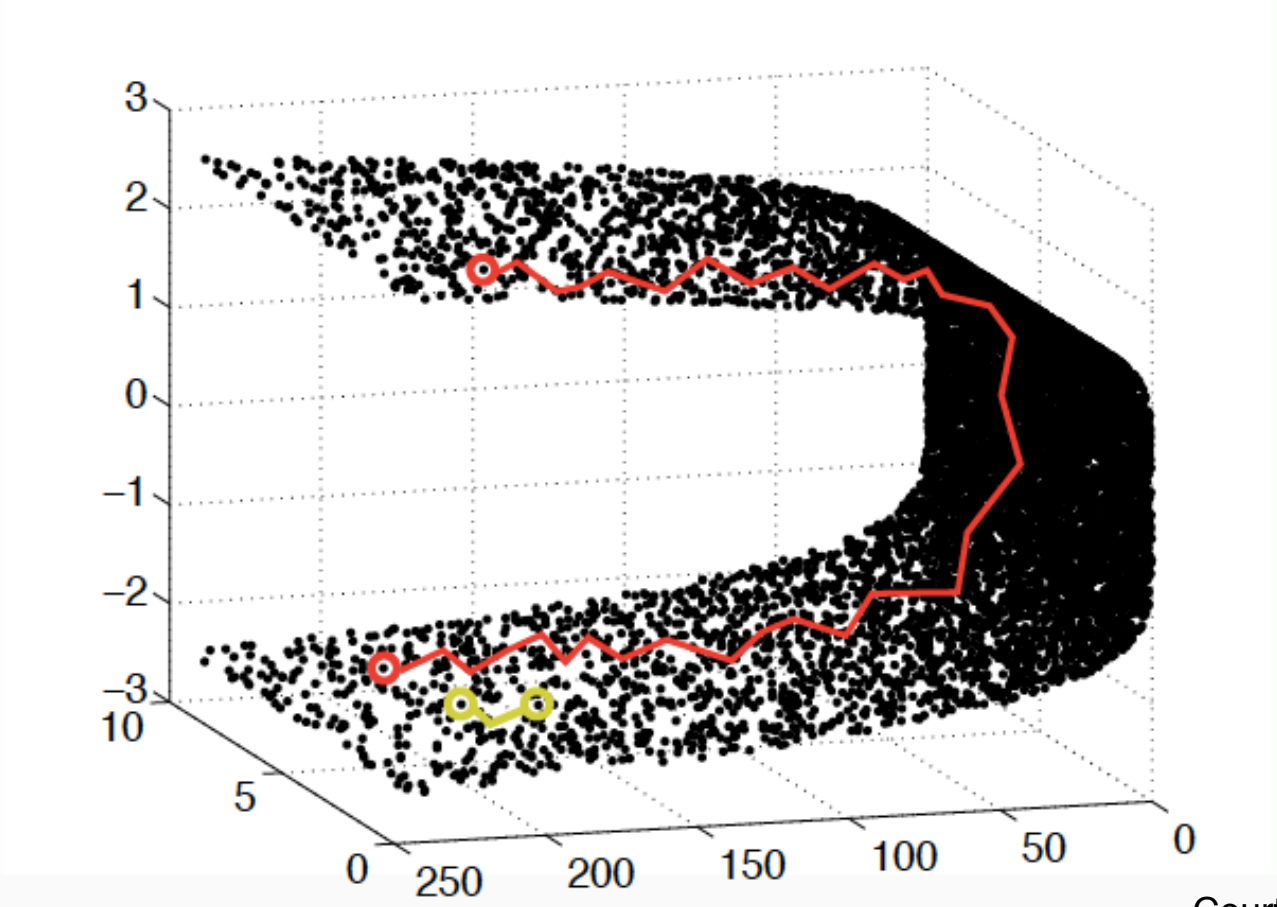
big data (diagonalization of large matrices),
model selection,
uncertainty quantification

Interpretation:

Connection to experimental observables

Diffusion Map

Diffusion maps idea: Build a Markov chain on the data points. Base jump probabilities on Euclidean distance, but only allow local jumps. Then compute distances based on how the Markov chain traverses the dataset.



Courtesy of Ralf Banisch

Diffusion Map

Define transition probability kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\epsilon(\mathbf{x}_i)\epsilon(\mathbf{x}_j)}\right)$$

Normalize the kernel as: $\tilde{K}_{i,j} = \dfrac{K(\mathbf{x}_i, \mathbf{x}_j)}{\sqrt{\sum_k K(\mathbf{x}_i, \mathbf{x}_k)\sum_k K(\mathbf{x}_j, \mathbf{x}_k)}}$

Define $D_i = \sum_j \tilde{K}_{i,j}$,

and construct the diffusion map transition matrix:

$$M_{i,j} = \frac{\tilde{K}_{i,j}}{D_i}$$

Calculate the first m eigenvalues and eigenvalues of M, $\{\tilde{\psi}_i\}$.

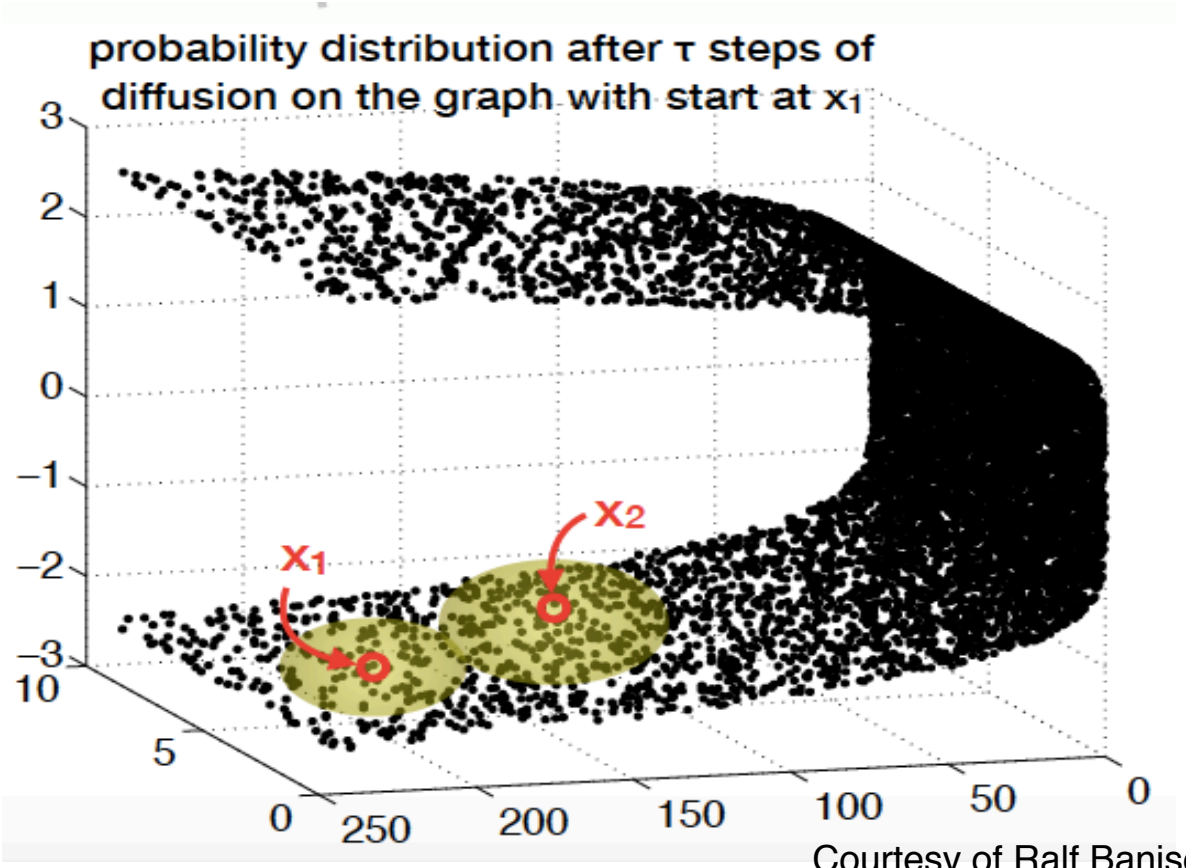
Coifman, R.R. *et al.* PNAS 102, 7426 (2005)
Nadler, B., *et al.* NIPS 21, 113 (2006)
Rohrdanz, M.A. *et al.* J. Chem. Phys. 134 (2011)

Diffusion Distance

$$D_\tau^2(\mathbf{x}_1,\mathbf{x}_2) = \|p_\tau(\mathbf{y}\mid\mathbf{x}_1) - p_\tau(\mathbf{y}\mid\mathbf{x}_2)\|_{\pi^{-1}}^2 = \int_{\mathbf{y}\in\Omega} \frac{|p_\tau(\mathbf{y}\mid\mathbf{x}_1) - p_\tau(\mathbf{y}\mid\mathbf{x}_2)|^2}{\pi(\mathbf{y})}\,\mathrm{d}\mathbf{y} =$$

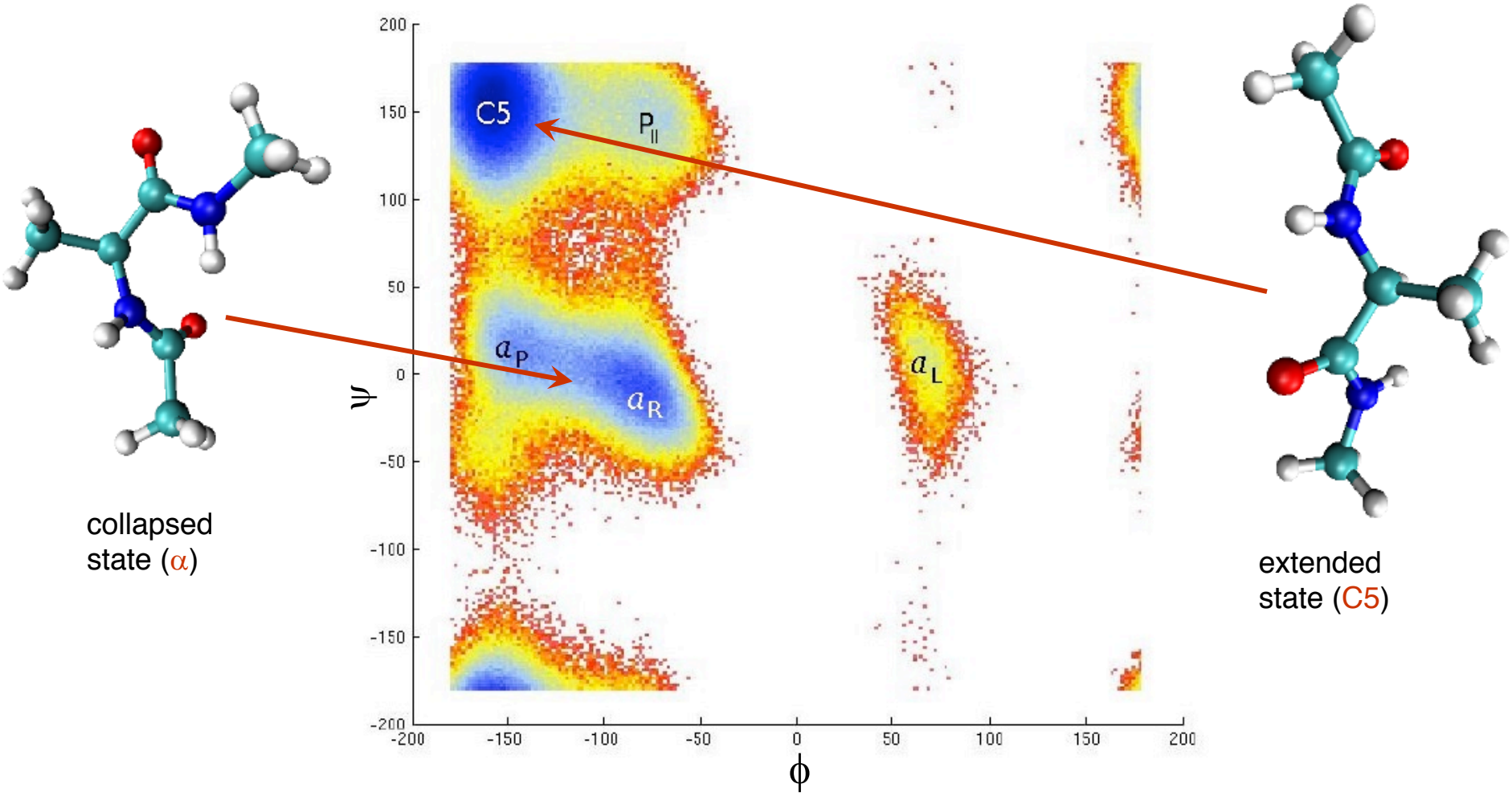
$$= \sum e^{-2\kappa_i\tau}(\psi_i(\mathbf{x}_1) - \psi_i(\mathbf{x}_2))^2$$

Diffusion coordinates (eigenfunctions of the MD backward propagator) can be used to define a "natural" distance metric



Courtesy of Ralf Banisch

Application to Alanine Dipeptide

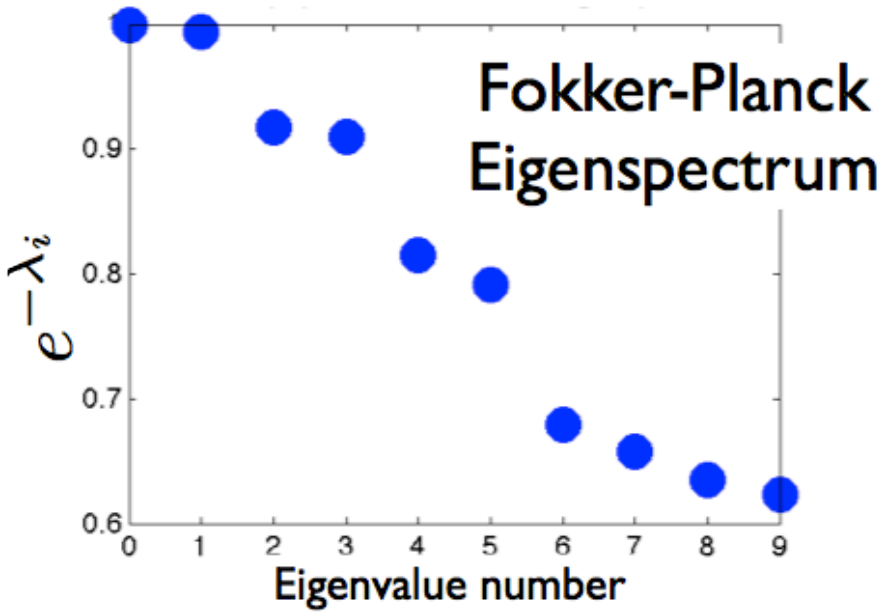


collapsed state (α)

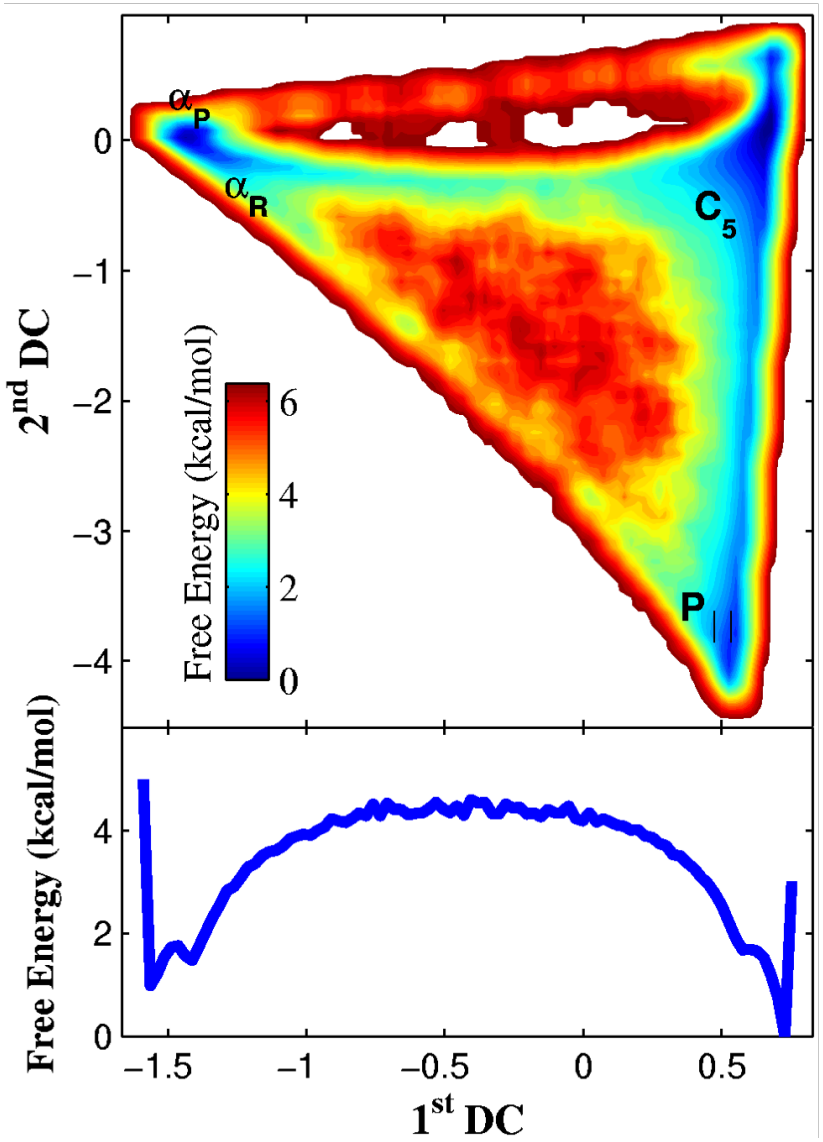
extended state (C5)

M.Rohrdanz, W. Zheng, M. Maggioni, C. Clementi *J. Chem. Phys.* 134(12), 124116, 2011

Application to Alanine Dipeptide

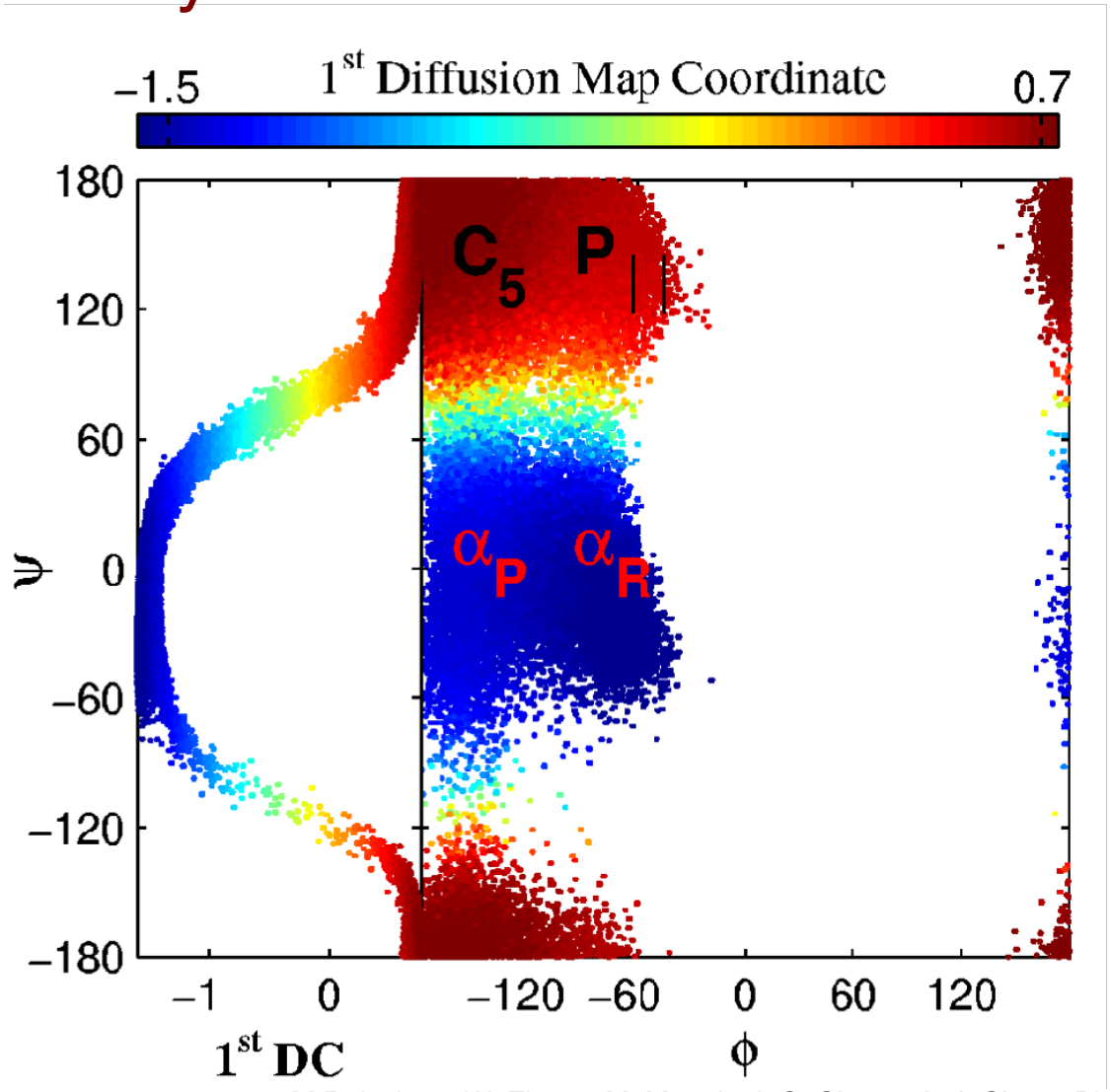


$$p(\boldsymbol{x}, t) = \phi_0(\boldsymbol{x}) + \sum_{i=1}^{k} c_i \phi_i(\boldsymbol{x}) e^{-\lambda_i t}$$



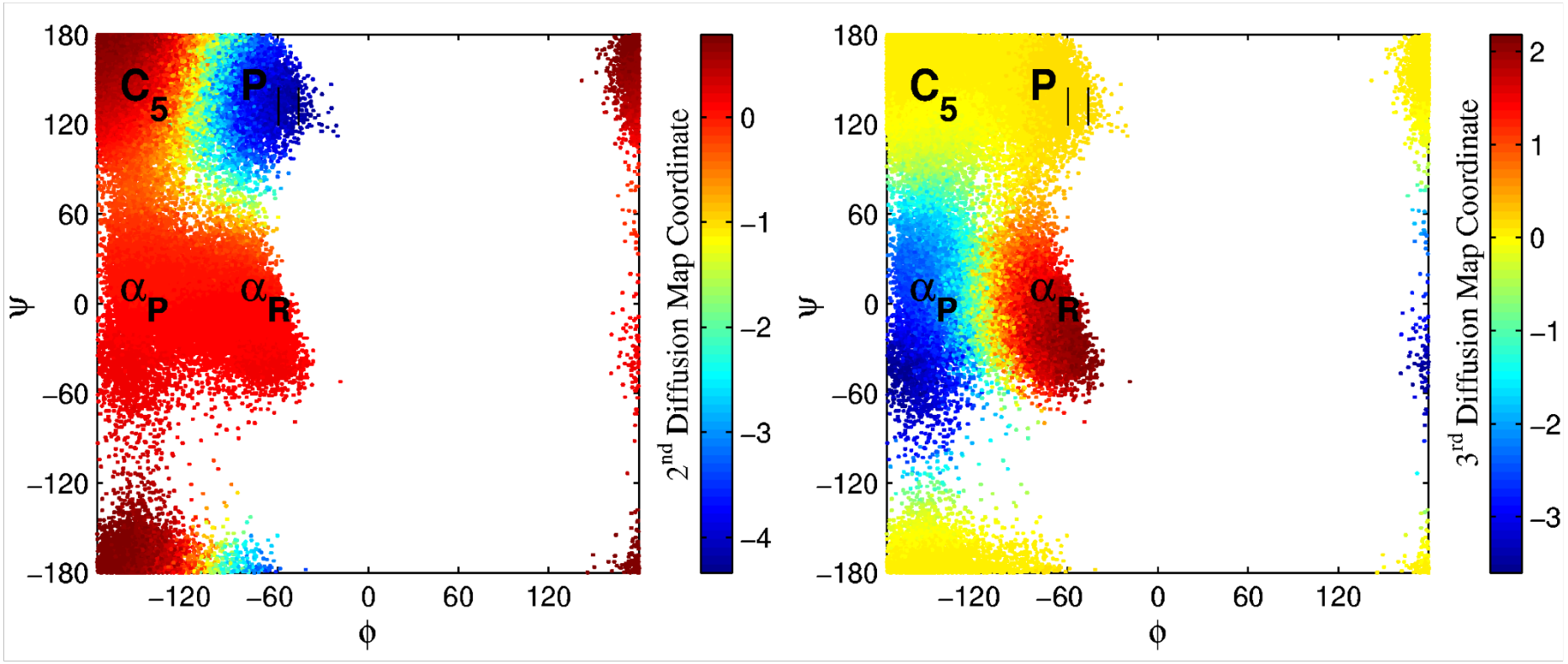
M.Rohrdanz, W. Zheng, M. Maggioni, C. Clementi *J. Chem. Phys.* 134(12), 124116, 2011

What motions are described by the diffusion coordinates?



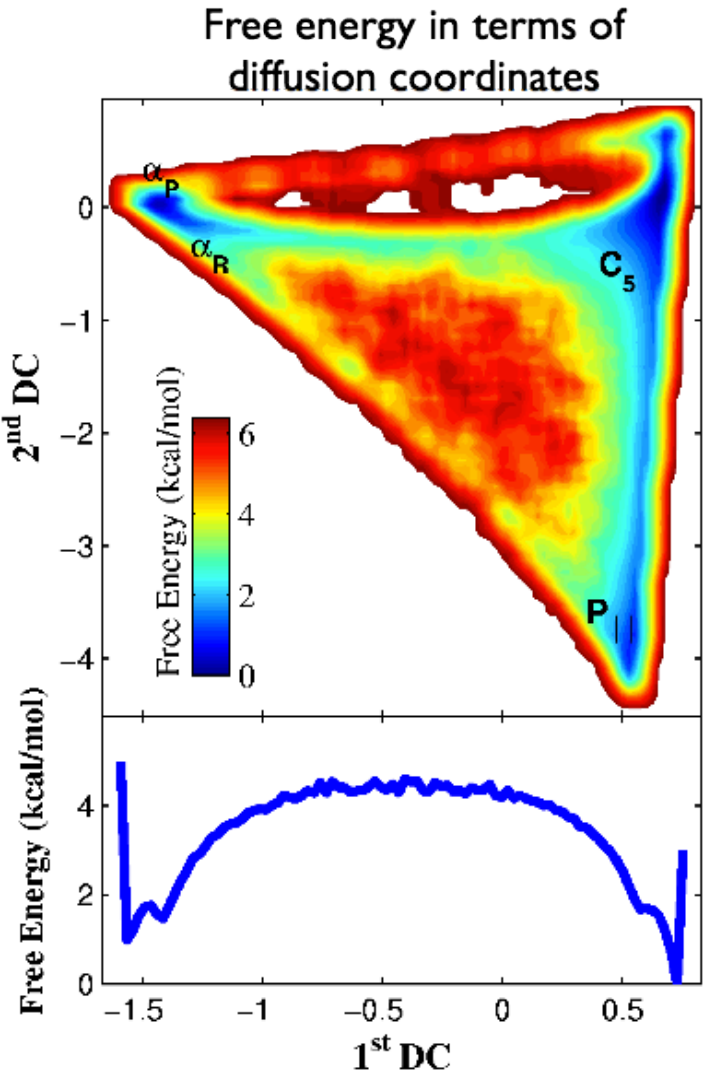
M.Rohrdanz, W. Zheng, M. Maggioni, C. Clementi *J. Chem. Phys.* 134(12), 124116, 2011

What motions are described by the diffusion coordinates?



M.Rohrdanz, W. Zheng, M. Maggioni, C. Clementi *J. Chem. Phys.* 134(12), 124116, 2011

Kramers' reaction rates



$$\text{rate} = \left(\int_{\text{barrier}} \frac{e^{\beta F(x)}}{D(x)} dx \int_{\text{well}} e^{-\beta F(x')} dx' \right)^{-1}$$

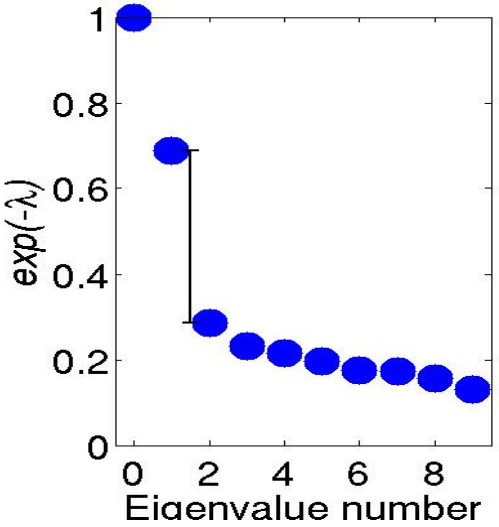
Coordinate	$C_5, P_{\|} \rightarrow \alpha_R\ \alpha_P$	$\alpha_R\ \alpha_P \rightarrow C_5, P_{\|}$
From simulation[b]	0.023	0.047
1st DC	0.023 ± 0.001	0.048 ± 0.003
Ψ	0.020 ± 0.001	0.040 ± 0.003

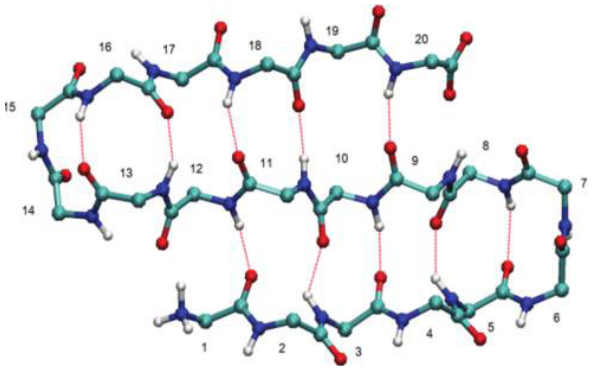
M.Rohrdanz, W. Zheng, M. Maggioni, C. Clementi *J. Chem. Phys.* 134(12), 124116, 2011

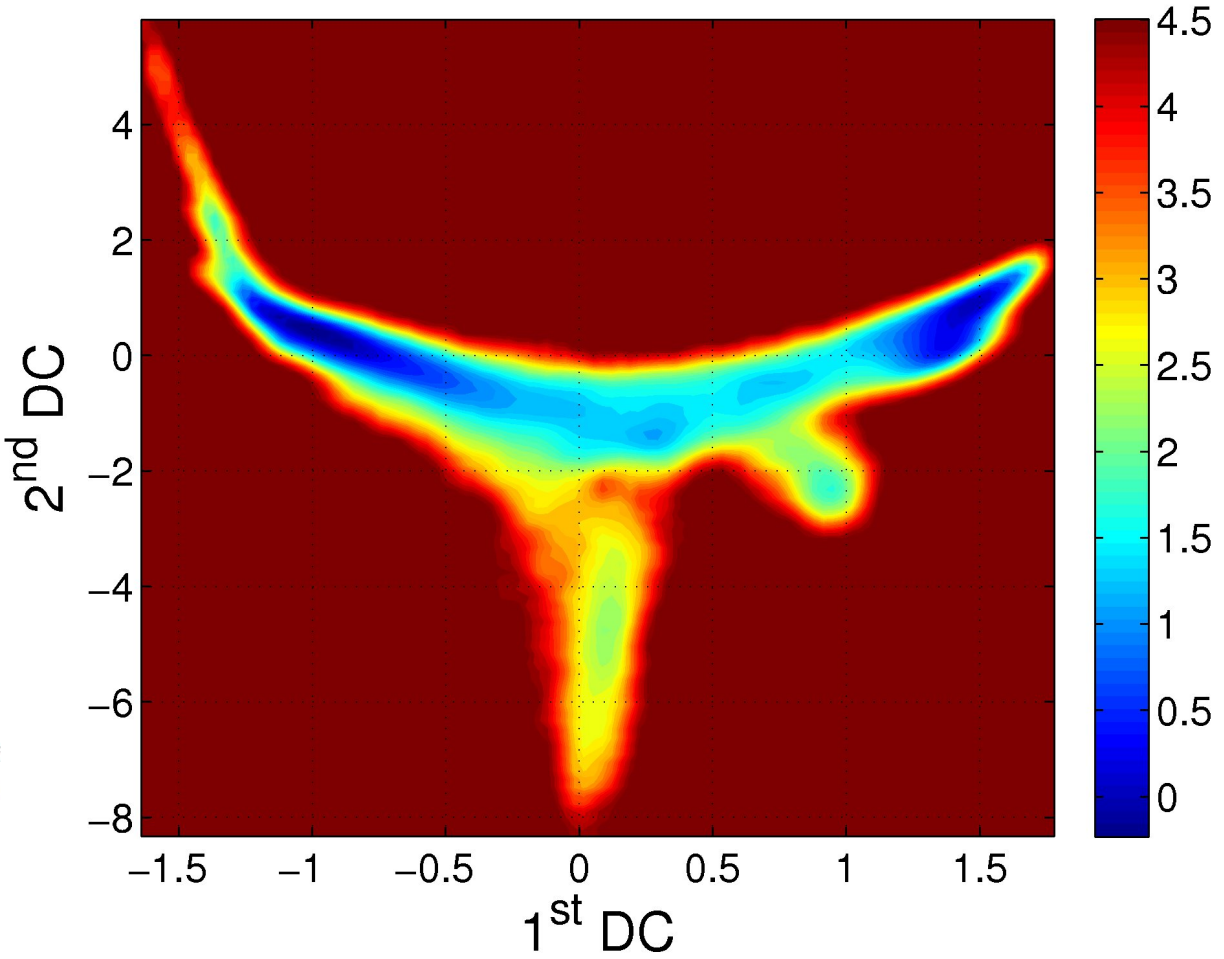
Application to β3s peptide

Free energy map as a function of the first two diffusion coordinates

Eigenvalue spectrum

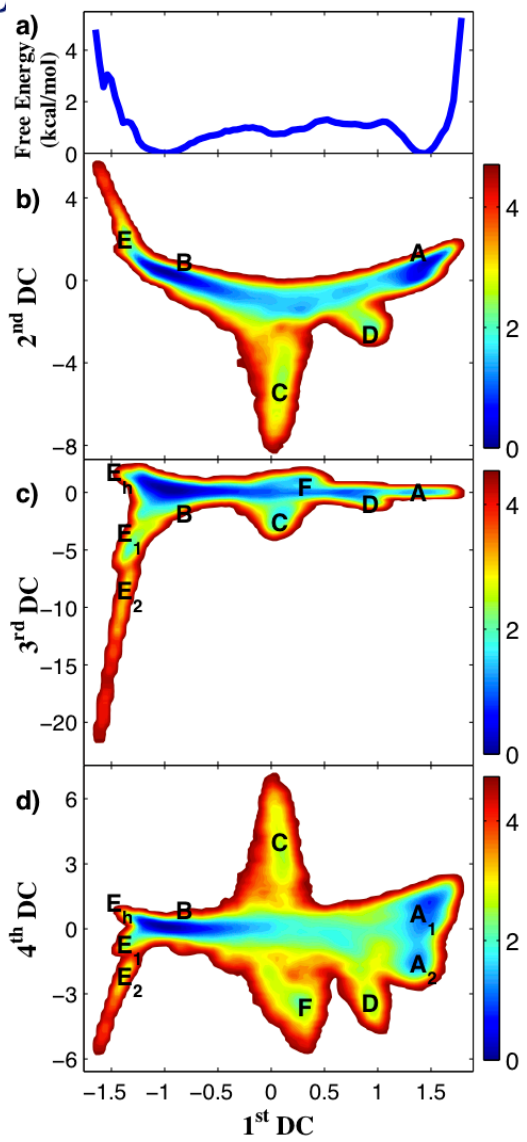


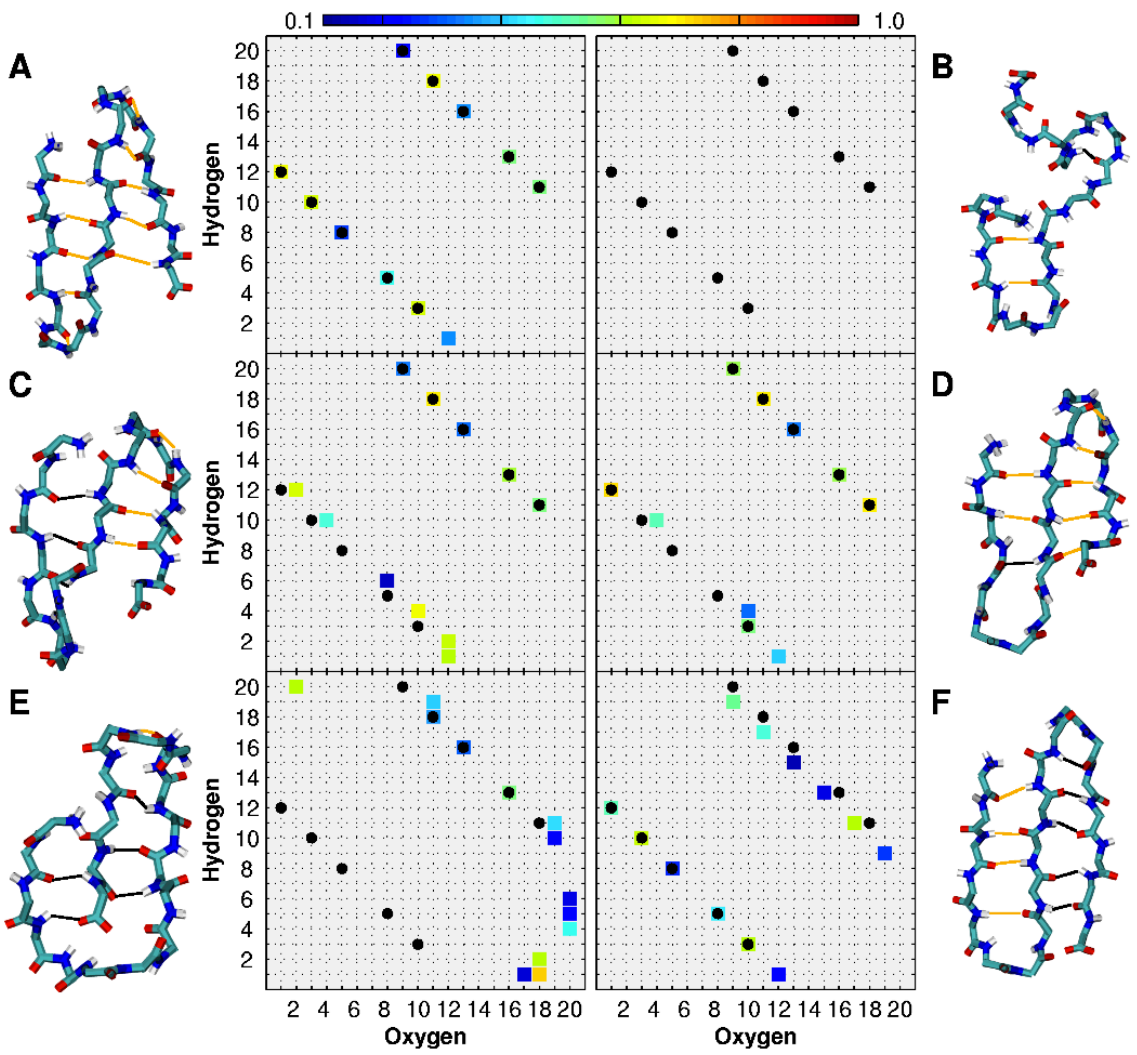




W. Zheng, B. Qi, M.Rohrdanz, A. Caflish, A. Dinner, C. Clementi *J. Phys. Chem. B,* 115(44) 13065-13074, 2011

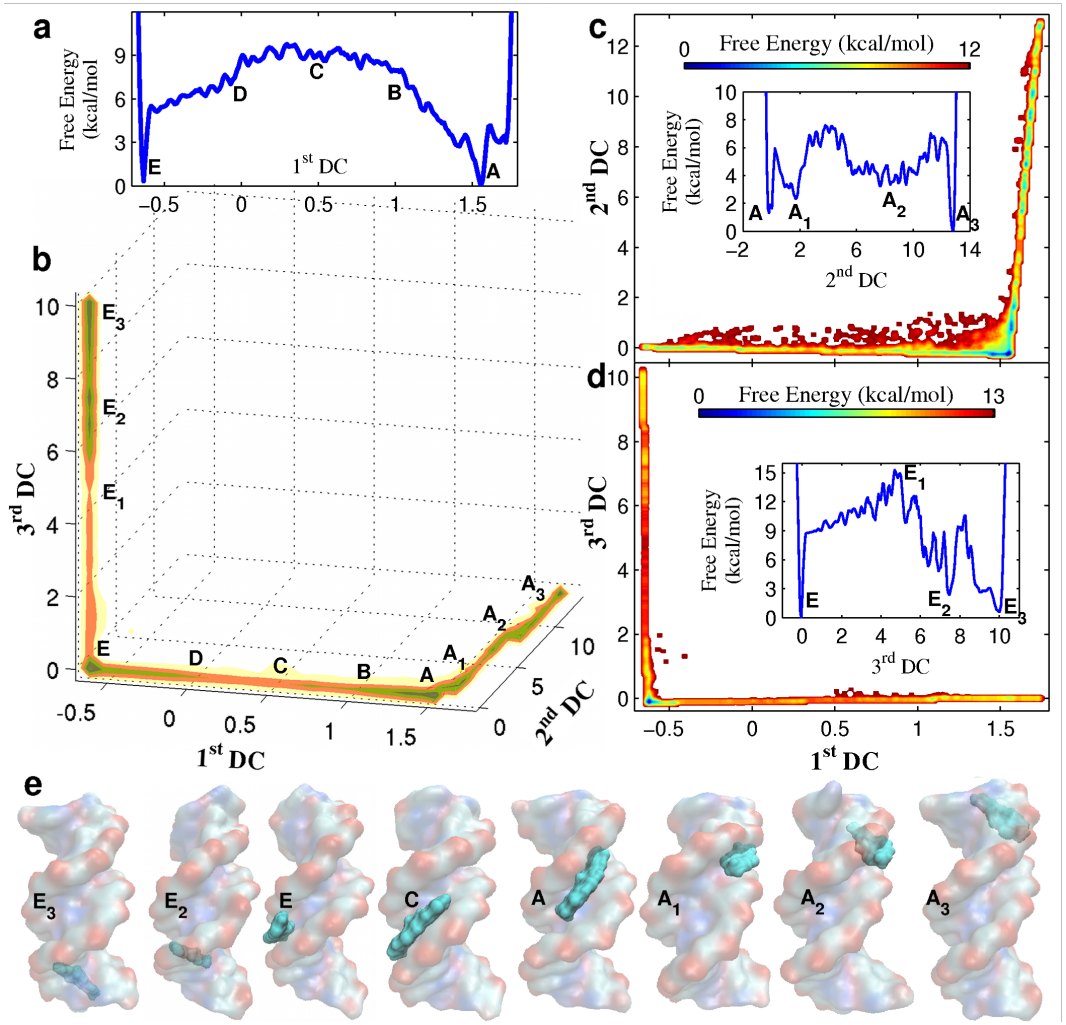
Application to β3s peptide

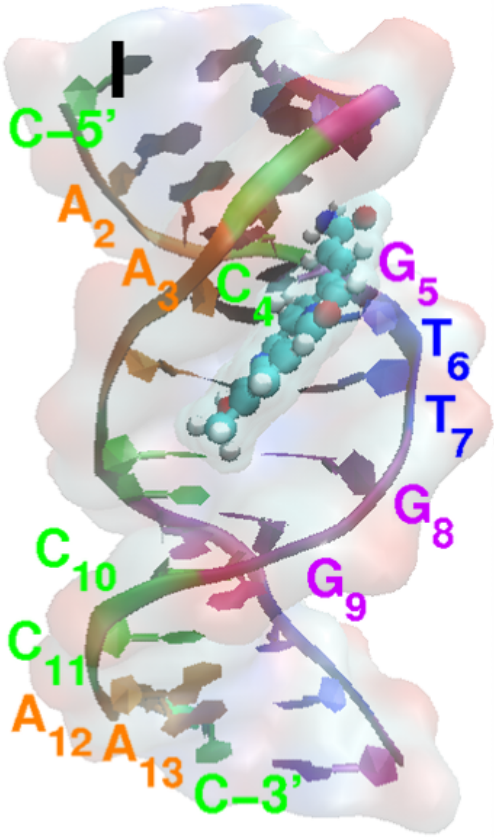




W. Zheng, B. Qi, M.Rohrdanz, A. Caflish, A. Dinner, C. Clementi *J. Phys. Chem. B,* 115(44) 13065-13074, 2011

Roughness and complexity of the free energy landscape associated with DNA/drug binding





W. Zheng, A.V. Vargiu, M.A. Rohrdanz, P. Carloni, C. Clementi J. Chem. Phys., 139, 145102 (2013)

Kinetic Distance

Given a (large) set of candidate coordinates, tICA provides their linear combination that best approximates the diffusion coordinates

Diffusion distance: $$D_\tau(x^\alpha, x^\beta) = \sqrt{\sum \lambda_i^2 (X^\alpha_{tica,i}(\tau) - X^\beta_{tica,i}(\tau))^2}$$

$$\hat{\lambda}_{tica,i} \lesssim \lambda_i$$

$$X_{tica,i} = \sum_j r_{ij} y_i \simeq \psi_i$$

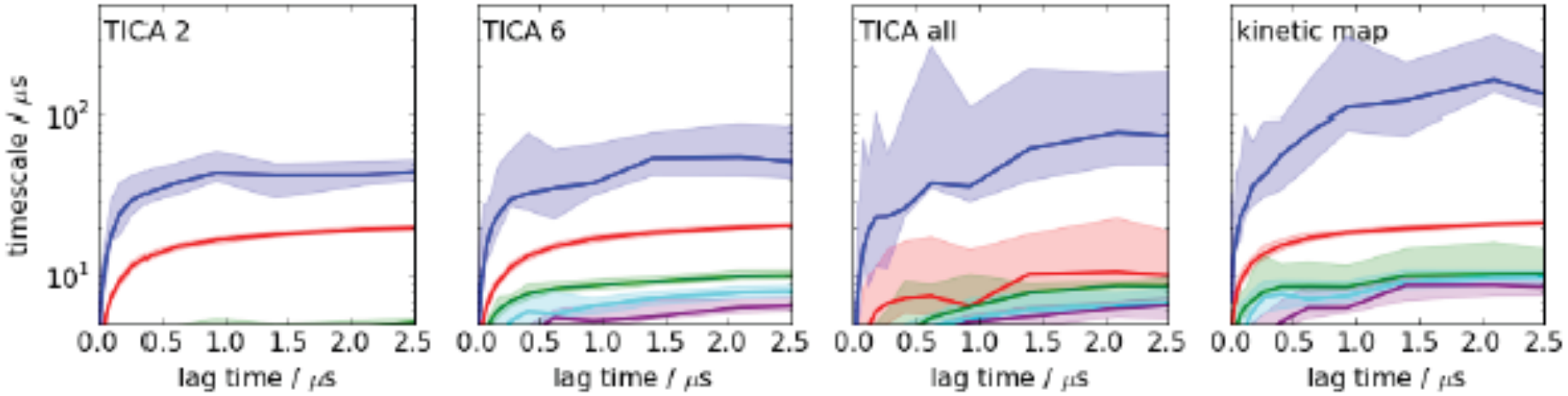
Kinetic distance: $$D^2_\tau(x^\alpha, x^\beta) = \sum \lambda_i^2 (X^\alpha_{tica,i}(\tau) - X^\beta_{tica,i}(\tau))^2$$

F. Noe, C. Clementi *JCTC 11, 5002 (2015)*

Kinetic Distance

When used to build MSMs, the kinetic distance yields significantly better results than working in a truncated TICA-space

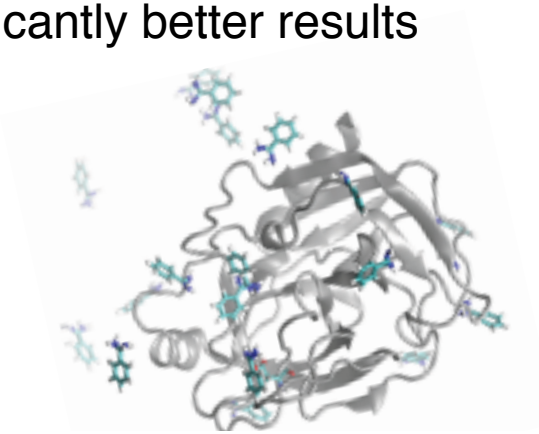
Example: comparison of MSMs of BPTI
(1 ms Anton trajectory)



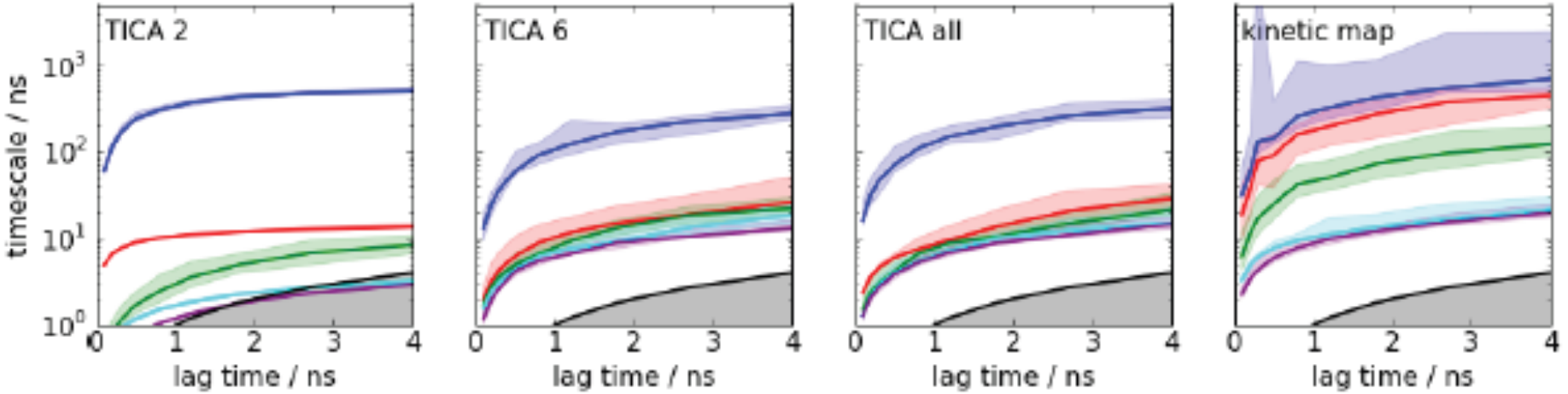
F. Noé, C. Clementi JCTC 11, 5002 (2015)

Kinetic Distance

When used to build MSMs, the kinetic distance yields significantly better results than working in a truncated TICA-space

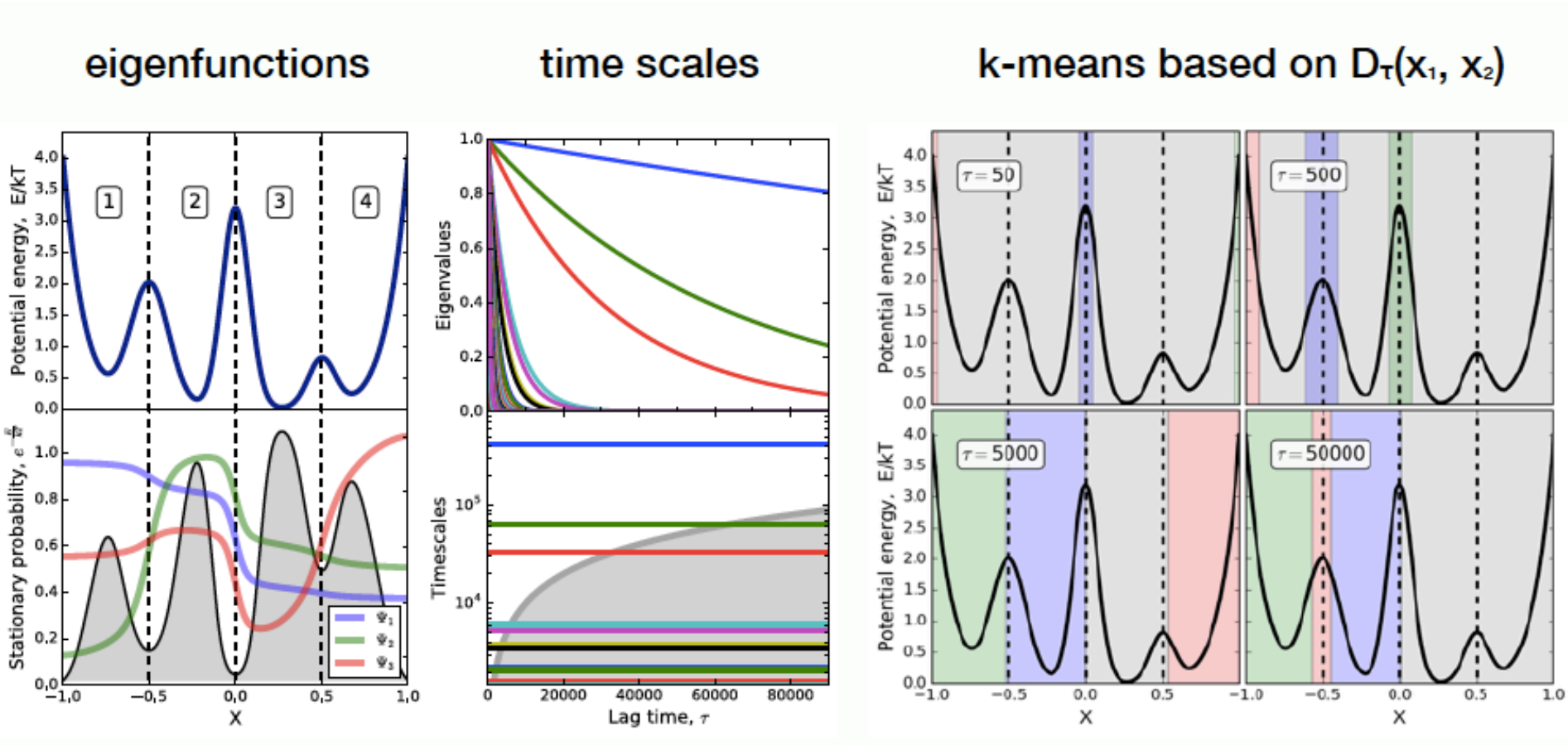


Example: comparison of MSMs of trypsin-benzamidine association
data from: *Buch, Giorgino, de Fabritiis, PNAS, 108, 10184-10189 (2011)*



F. Noé, C. Clementi JCTC 11, 5002 (2015)

Kinetic Distance



However, results are still strongly dependent on the choice of lagtime, τ !!

F. Noé, R. Banisch, C. Clementi JCTC 12(11) 5620 -5630 (2016)

Commute Distance

The same idea, but INTEGRATED over all timescales, provides a more robust distance metric

$$\mathrm{d}^2_{comm}(\mathbf{x}_1, \mathbf{x}_2) = \int_{\tau=0}^{\infty} \|p_\tau(\mathbf{y} \mid \mathbf{x}_1) - p_\tau(\mathbf{y} \mid \mathbf{x}_2)\|^2_{\pi^{-1}} \, \mathrm{d}\tau =$$

$$= \sum_{j=1}^{n} \left(\sqrt{\frac{t_j}{2}} \psi_j(\mathbf{x}_1) - \sqrt{\frac{t_j}{2}} \psi_j(\mathbf{x}_2) \right)^2 =$$

$$= \frac{1}{2} \sum_{j=1}^{n} t_j \left(\psi_j(\mathbf{x}_1) - \psi_j(\mathbf{x}_2) \right)^2$$

It also provides an estimate of the half round-trip time between two well-separated states:

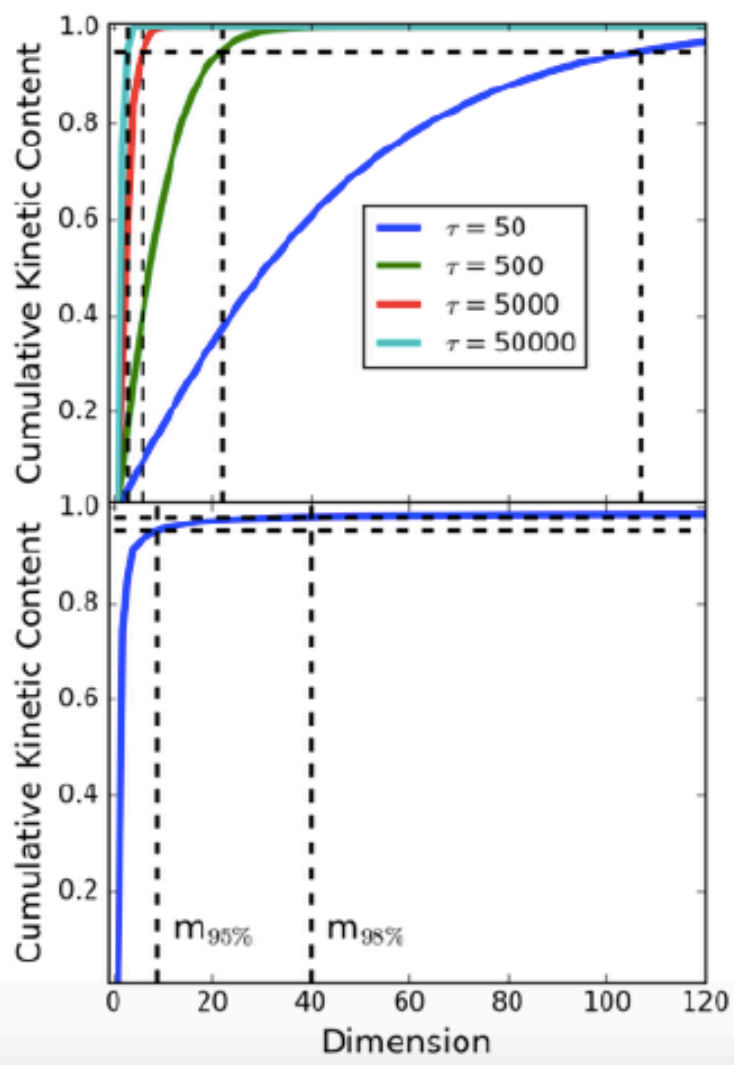
$$\mathrm{d}^2_{comm}(\mathbf{x}_i, \mathbf{x}_j) \lesssim \frac{t_{ij} + t_{ji}}{2}$$

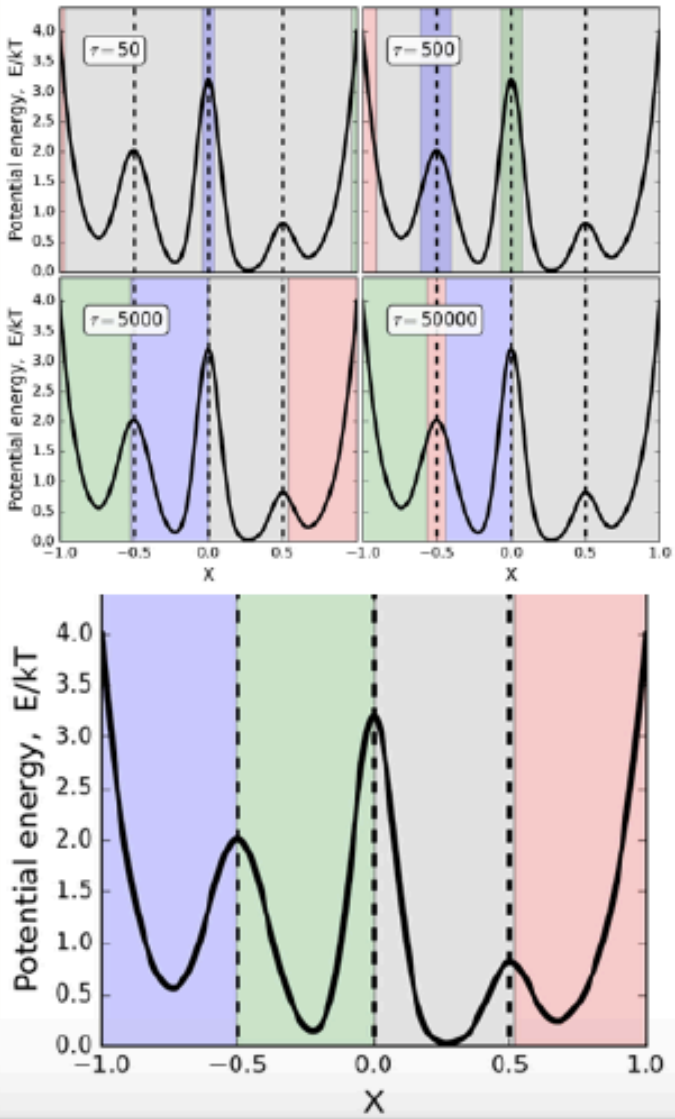
F. Noé, R. Banisch, C. Clementi JCTC 12(11) 5620 -5630 (2016)

Commute Distance

τ-kinetic distance

commute distance





F. Noé, R. Banisch, C. Clementi JCTC 12(11) 5620 -5630 (2016)

Commute Distance

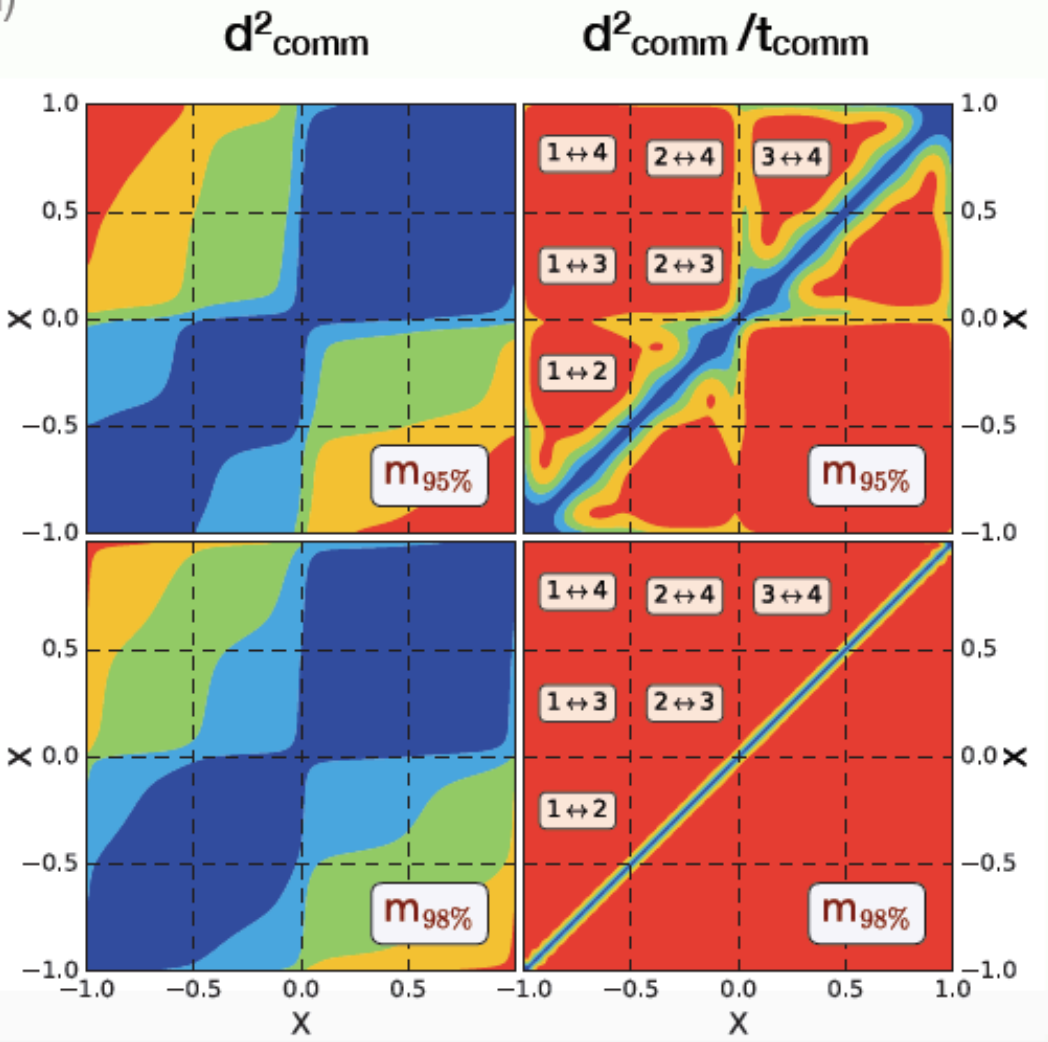
- in discrete state spaces, the (full) commute distance equals the commute time:

$$d^2_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{t_{ij} + t_{ji}}{2}$$

$$t_{ij} = \mathbf{E}[\text{time to hit } \mathbf{x}_j | \mathbf{x}_i]$$

- if we only use the largest m' eigenvalues, we get an upper bound:

$$\left[d^{(m')}_{\mathrm{comm}}(\mathbf{x}_i, \mathbf{x}_j)\right]^2 \leq \frac{t_{ij} + t_{ji}}{2}$$



F. Noé, R. Banisch, C. Clementi JCTC 12(11) 5620 -5630 (2016)

Commute Distance

simulation data

TICA

approximate eigenfunctions

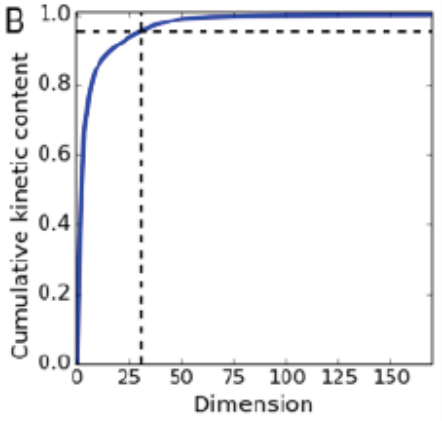
embedding: commute map, τ-kinetic map, …

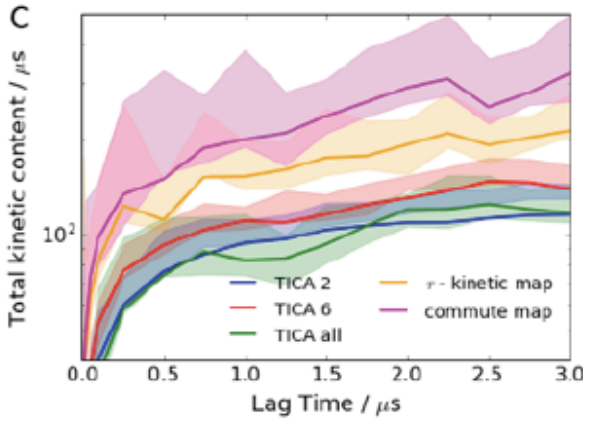
k-means

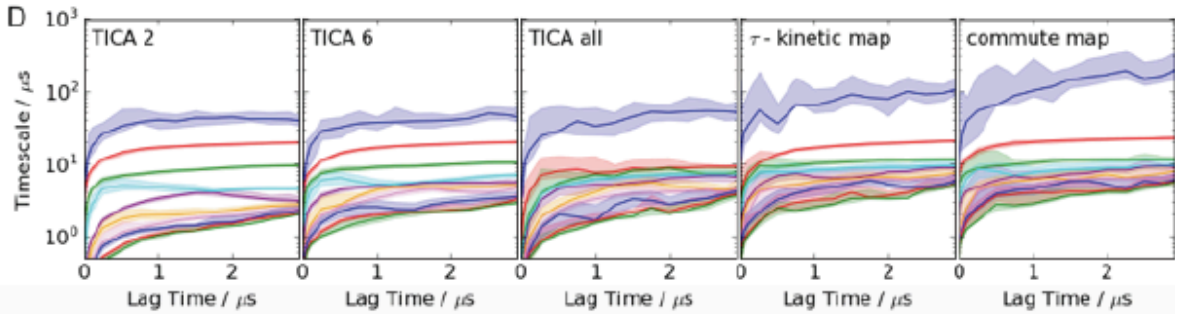
coarse-grained kinetic model (MSM)

data: 1ms simulation of BPTI produced by Anton supercomputer

variational principle $\hat{t}_i \leq t_i$





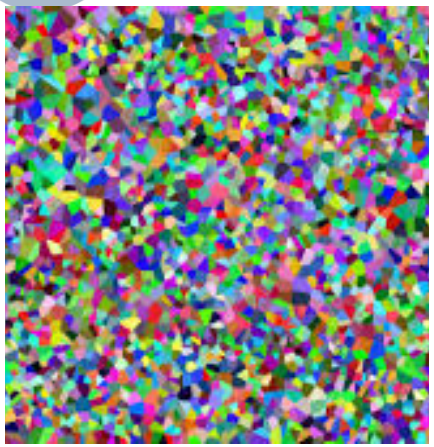


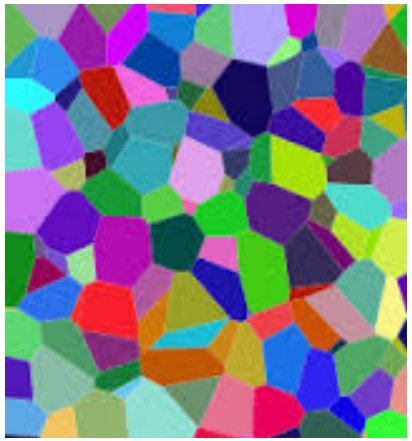
F. Noé, R. Banisch, C. Clementi JCTC 12(11) 5620 -5630 (2016)

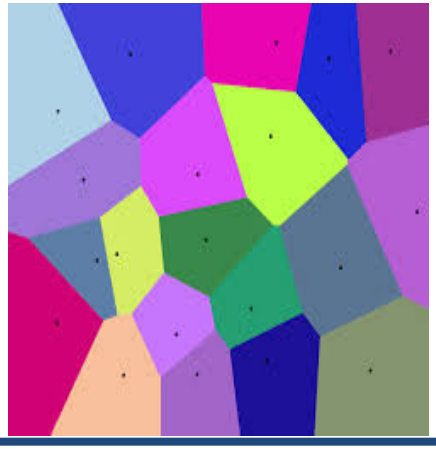
Coarse-Graining

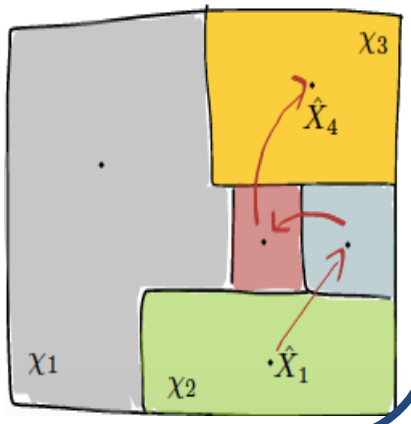
$\mathbb{R}^{3N}$

coarse-graining in conformation space



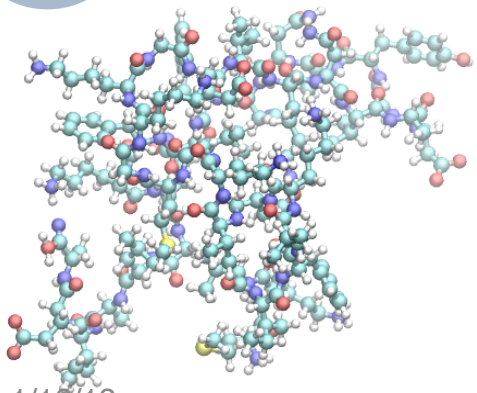


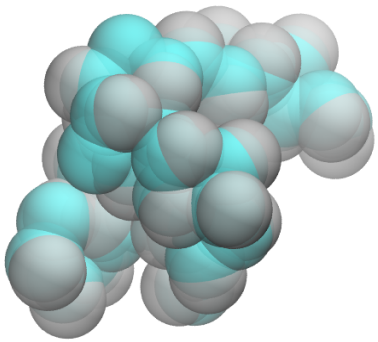


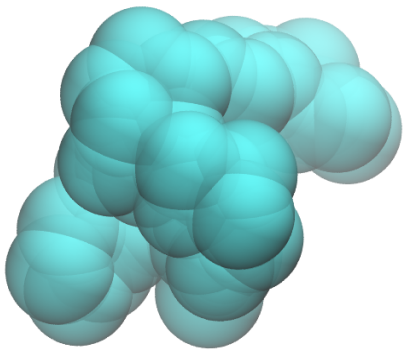


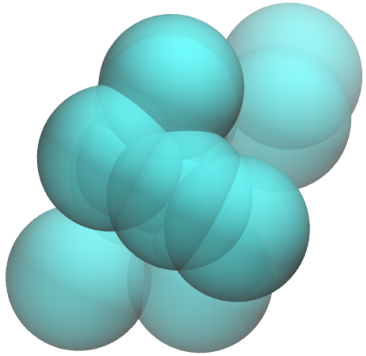
$\mathbb{R}^{3}$

coarse-graining in structural space









Coarse-graining in physical (structure) space: common challenges

1. Definition of coarse variables

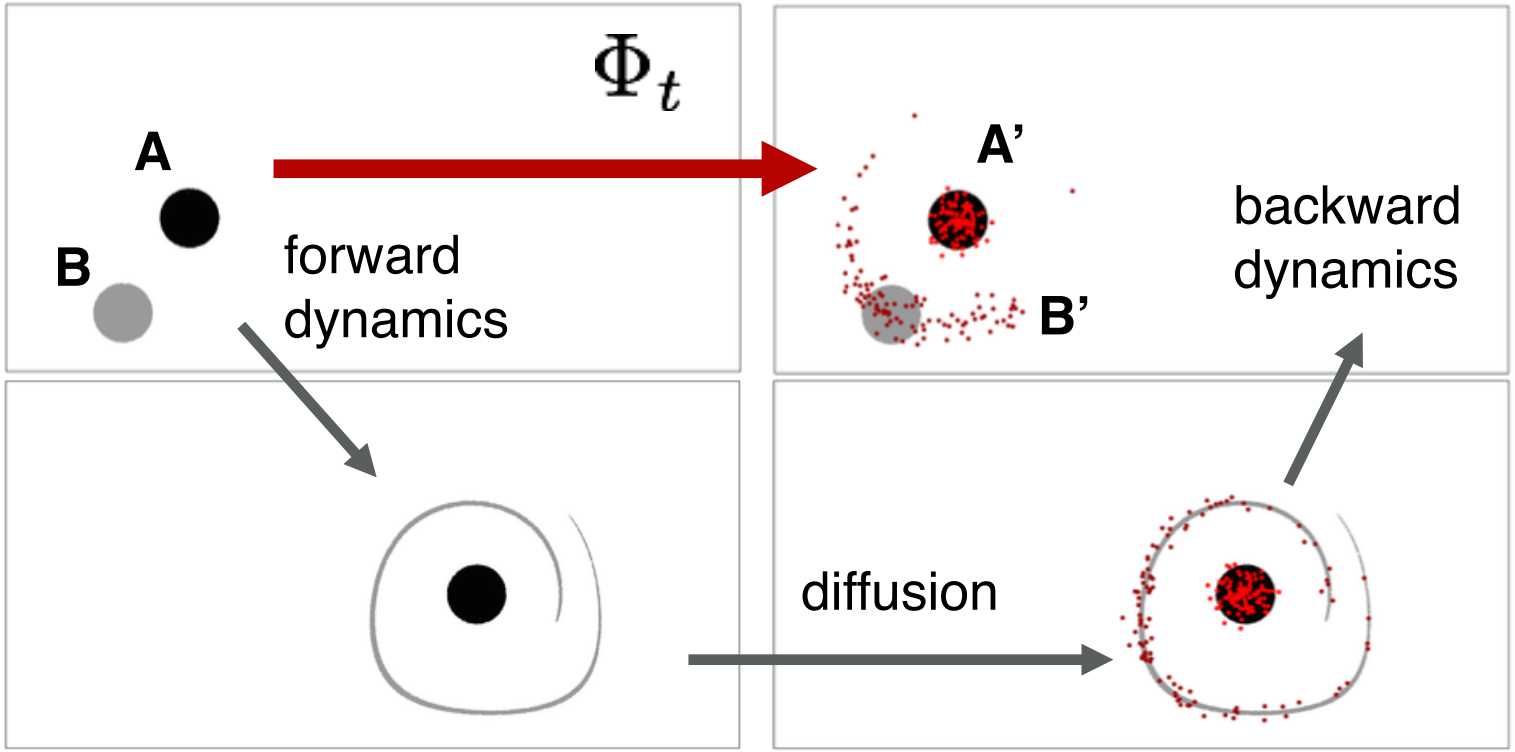
2. Definition of effective energy function (and dynamic equations)

3. Incorporation of experimental data

Coarse-graining in physical (structure) space: What groups of atoms ”belong” together?

Idea: use coherent state analysis

Coherent state region in state space that keep their geometric integrity, allowing very little transport in and out of themselves (i.e. atmospheric vortices)



R. Banisch, P. Koltai *Chaos* 27, 035804 (2017)

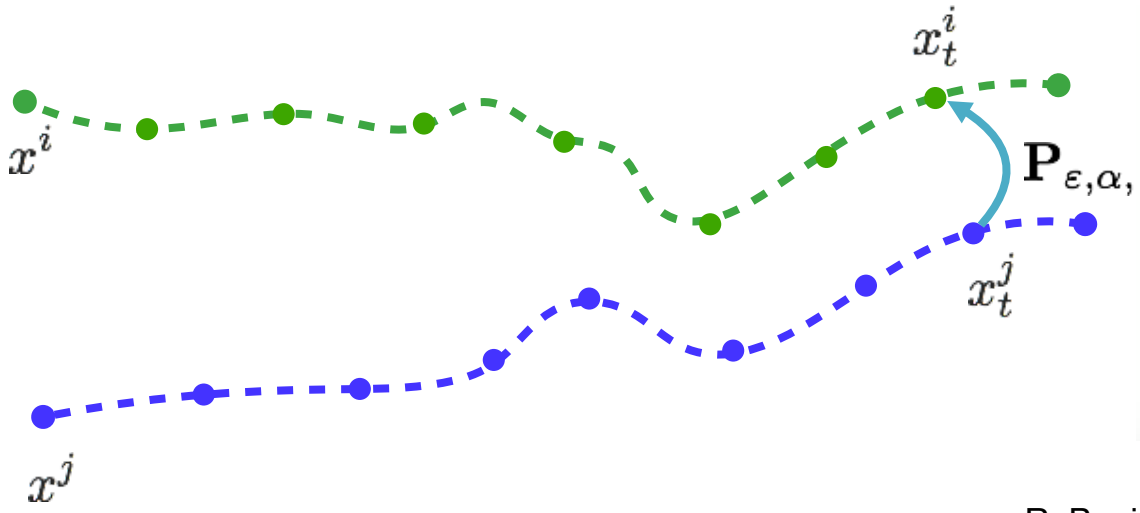
SpaceTime diffusion maps

Given *mT* data points from *m* trajectories sampled at *T* time points

$$I_t = \{t_0, \ldots, t_{T-1}\}$$

$$X = \{x_t^i := \Phi_t x^i \ : \ i = 1, \ldots, m; \ t \in I_t\},$$

Build a Markov chain on the trajectory data and use diffusion map to jump between points in the same time slice



Average over time slices, α = 1/2:

$$\mathbf{Q}_\varepsilon = \frac{1}{T} \sum_{t \in I_t} P_{\varepsilon,t}$$

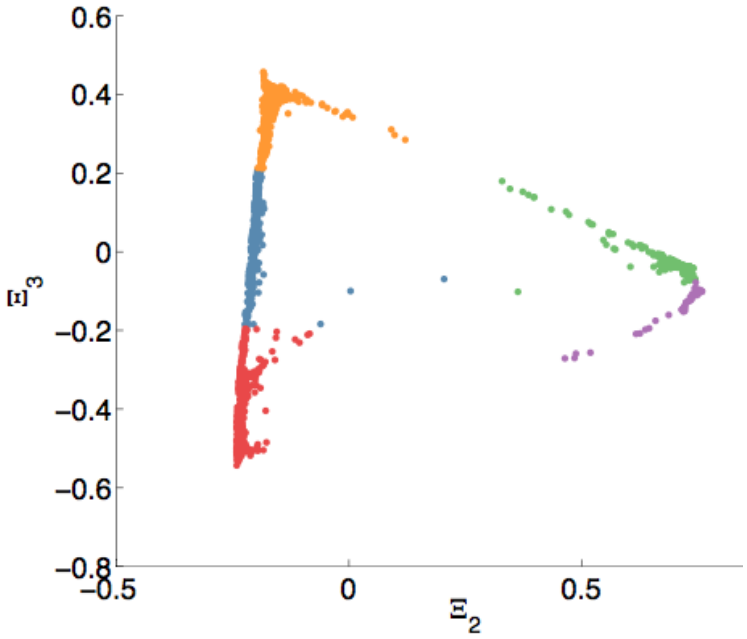
average over time slices

diffusion maps at time slice t

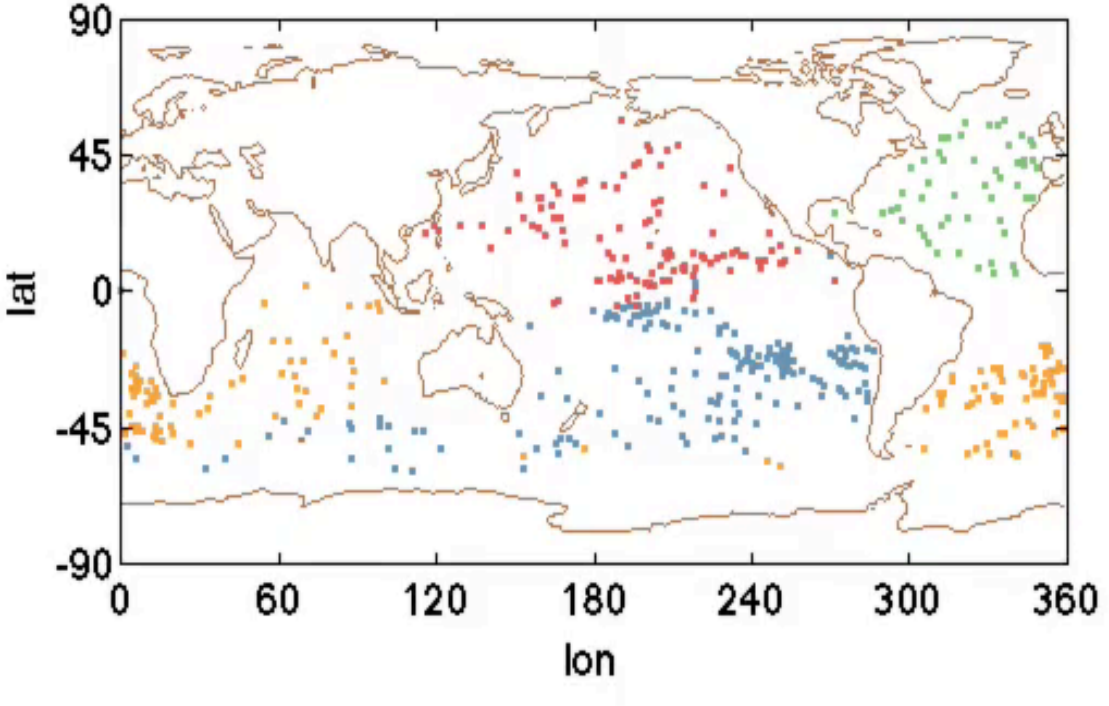
R. Banisch, P. Koltai *Chaos* 27, 035804 (2017)

SpaceTime diffusion maps: Ocean Drifter Data

embedding



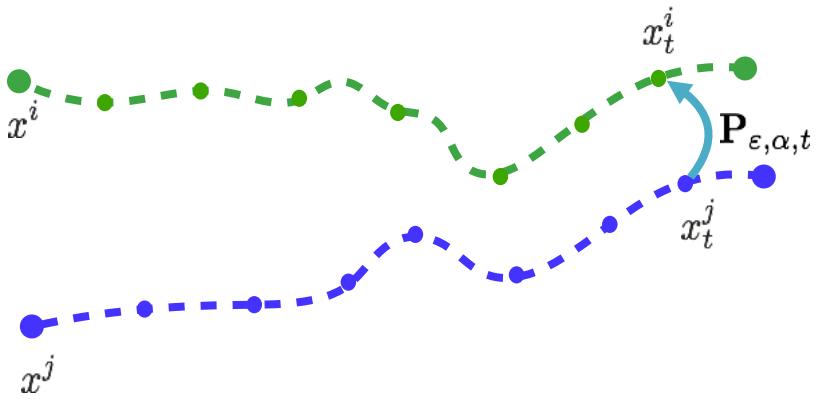
5-clustering

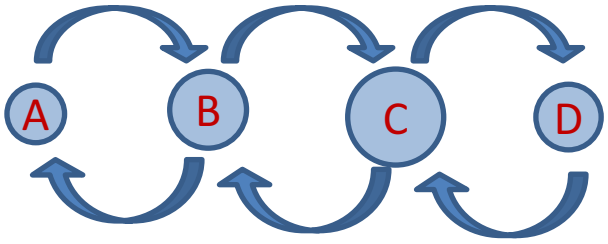


R. Banisch, P. Koltai *Chaos* 27, 035804 (2017)

Structural and State Space Decomposition (S^3D)

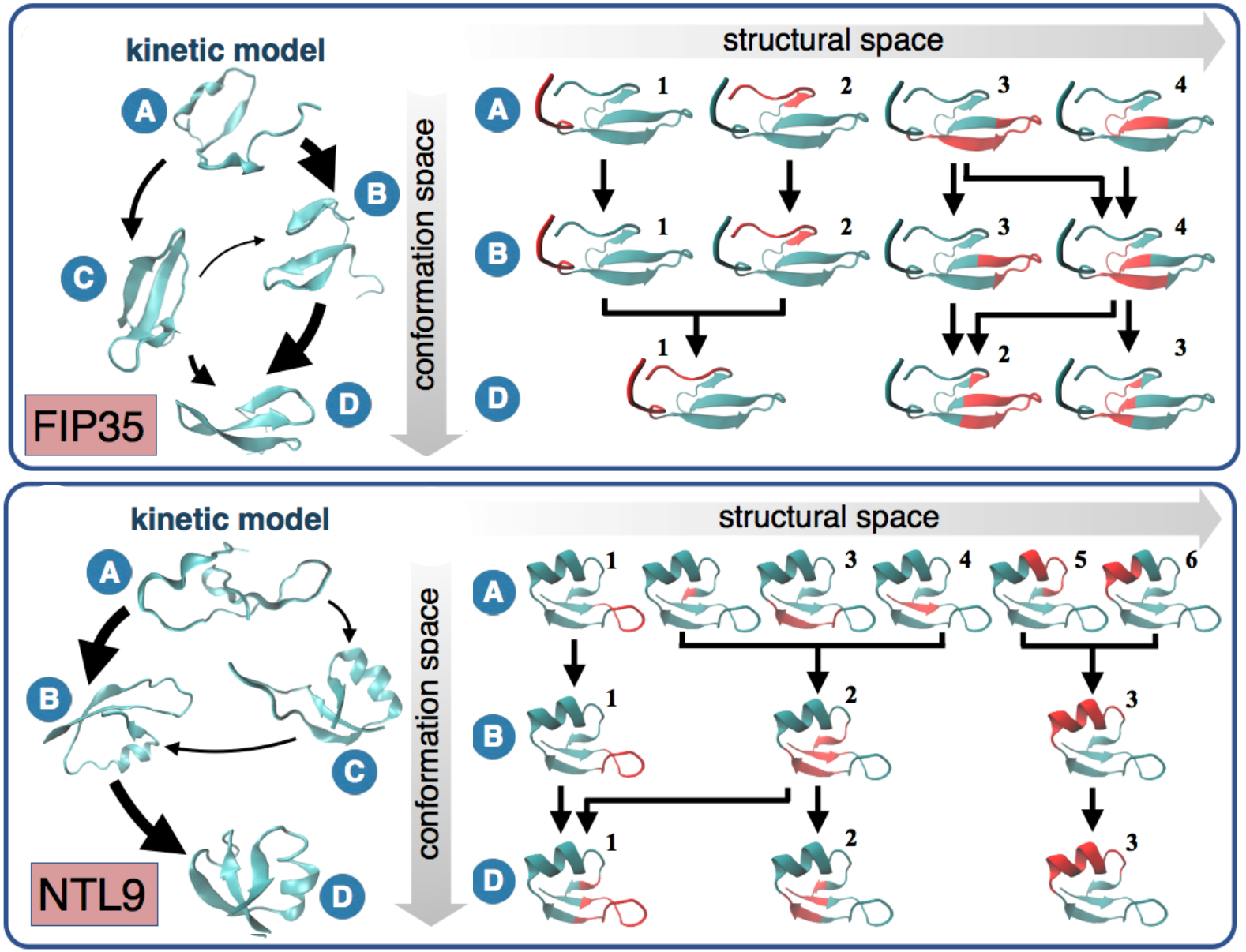
Combine:
Coherent State Analysis (Structural Space, in R^3)
and
Markov State Modeling (State Space, in R^{3N})





L. Boninsegna, R. Banisch, C. Clementi JCTC 14,453–460 (2018)

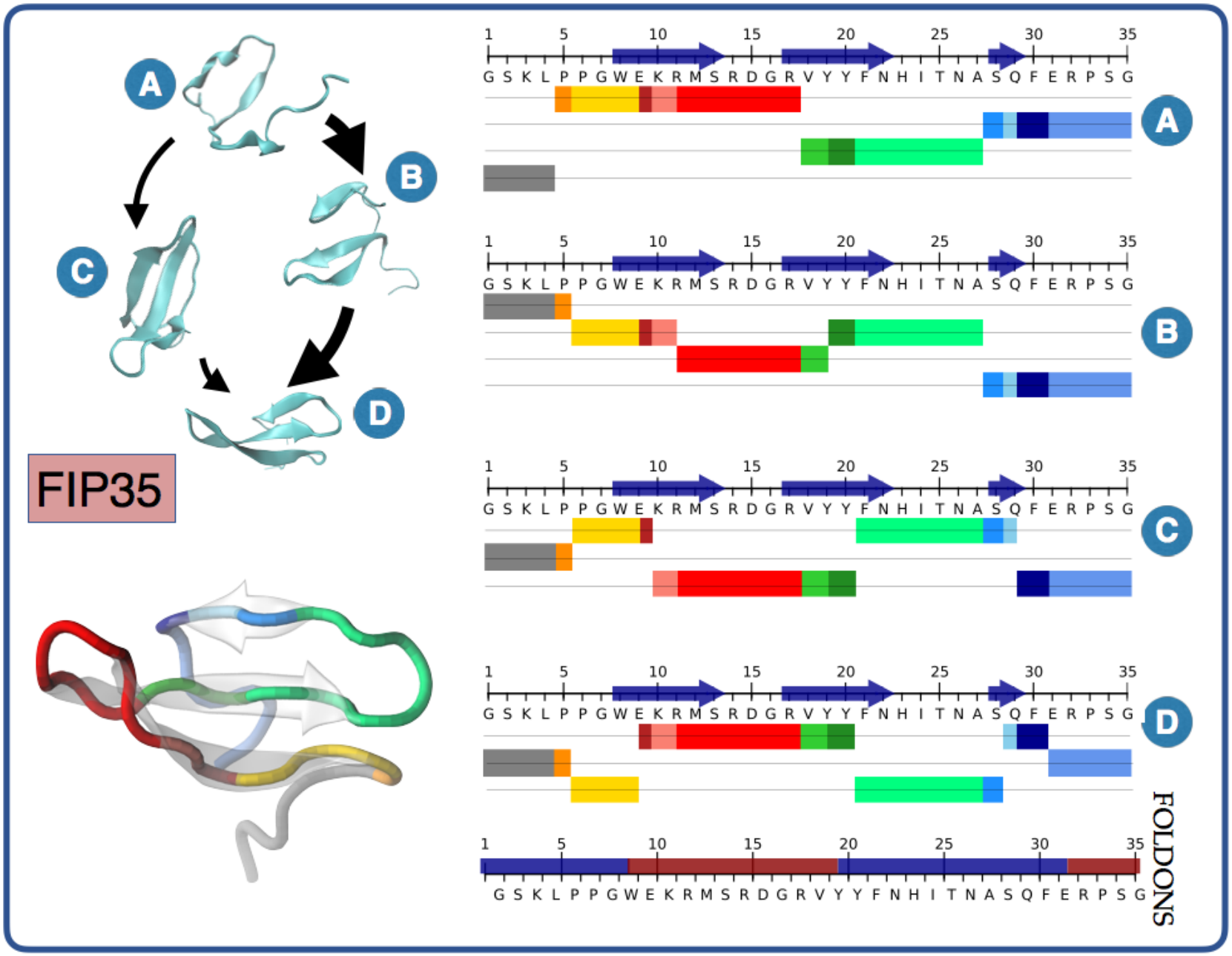
S^{3}D



all-atom trajectories from: Shaw DE *et al.* Science 330:341–346 (2010)
Lindorff-Larsen K, *et al.* Science 334:517–520 (2011)

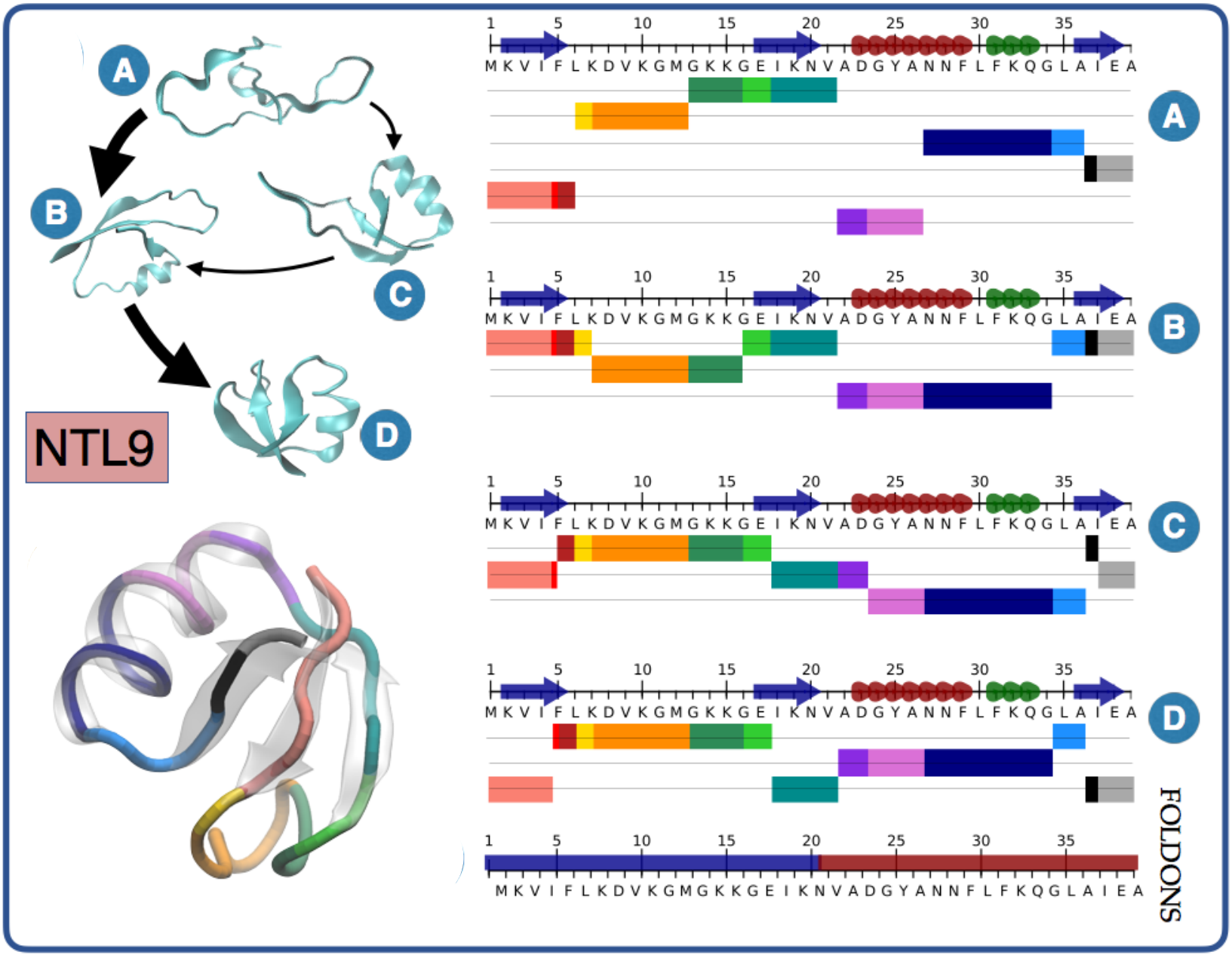
L. Boninsegna, R. Banisch, C. Clementi JCTC 14,453–460 (2018)

S^{3}D



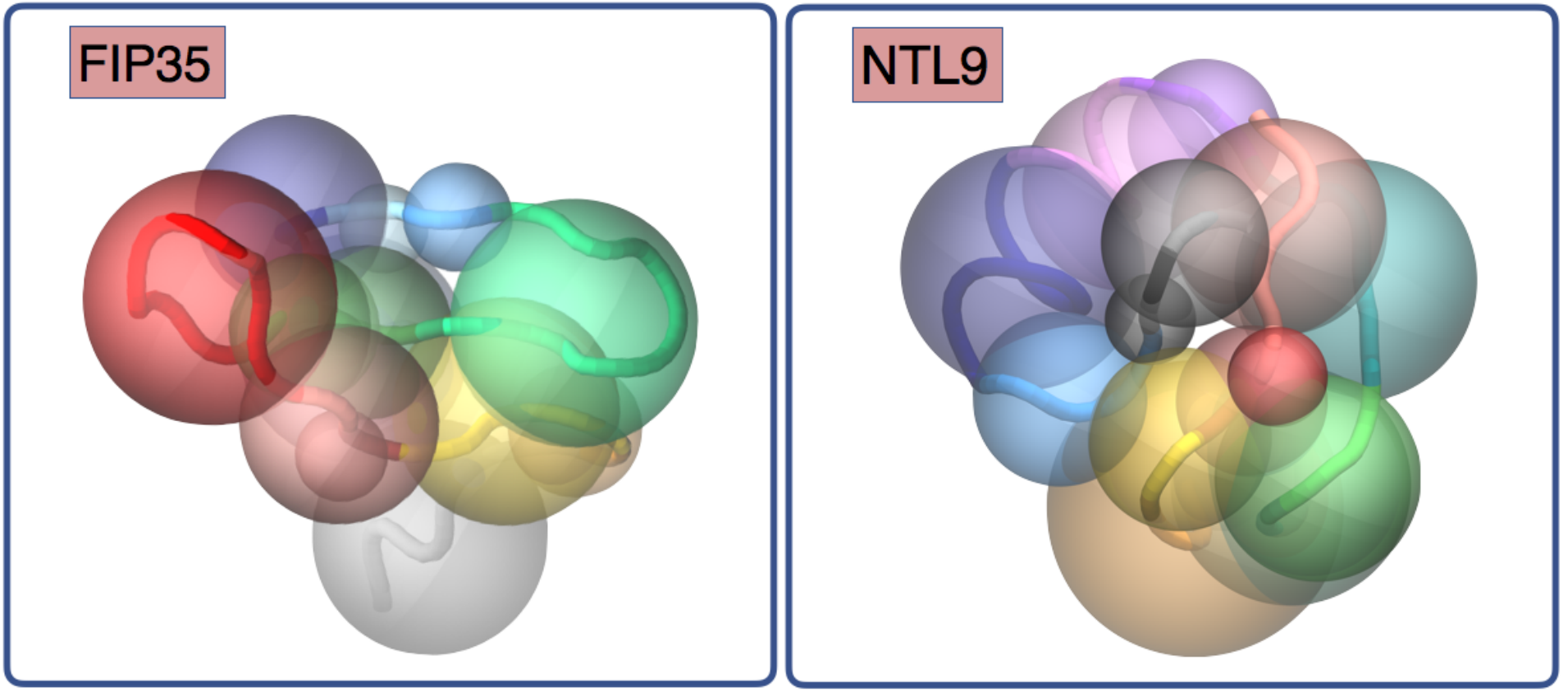
L. Boninsegna, R. Banisch, C. Clementi JCTC 14,453–460 (2018)

S^{3}D



L. Boninsegna, R. Banisch, C. Clementi JCTC 14,453–460 (2018)

Structural and State Space Decomposition (S^3D)



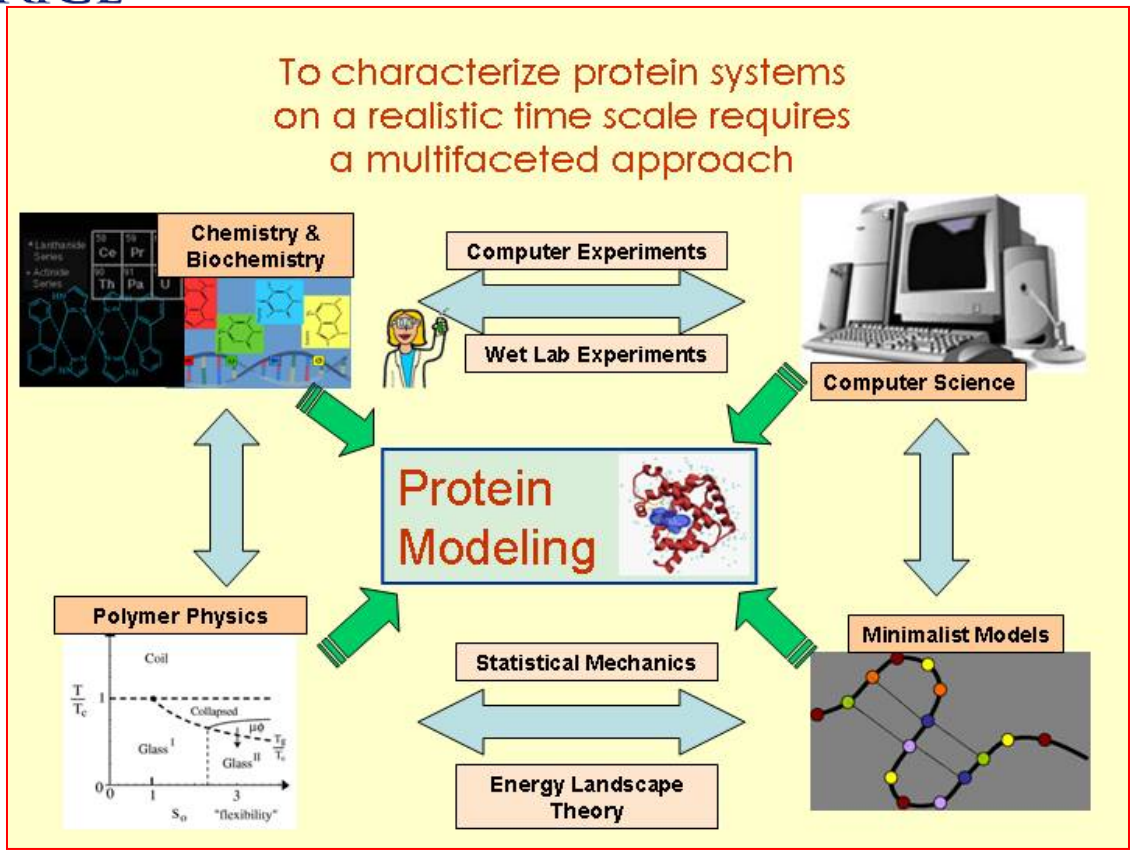
L. Boninsegna, R. Banisch, C. Clementi JCTC 14,453–460 (2018)

Coarse-graining in physical (structure) space: outstanding challenges

- Definition of coarse variables
- Definition of effective energy function
  (and dynamic equations)
- Incorporation of experimental data

Cecilia Clementi's research group

http://clementiresearch.rice.edu



Clementi's group

Dr. Feliks Nüske
Dr. Giovanni Pinamonti
Dr. Fabio Trovato
Lorenzo Boninsegna
Alex Kluber
Justin Chen
Eugen Hruska
Wangfei Yang

previous:

Dr. Fernando Yrazu	
Dr. Jordane Preto	(U Alberta)
Dr. Mary Rohrdanz	(MD Anderson)
Dr. Wenwei Zheng	(Arizona State U)
Dr. Amarda Shehu	(GMU)
Dr. Payel Das	(IBM)
Dr. Silvina Matysiak	(U. Maryland)
Dr. Brad Lambeth	(Shell)

$$ NSF

$$ Welch Foundation

$$ Einstein Foundation

Einstein Stiftung Berlin
Einstein Foundation Berlin