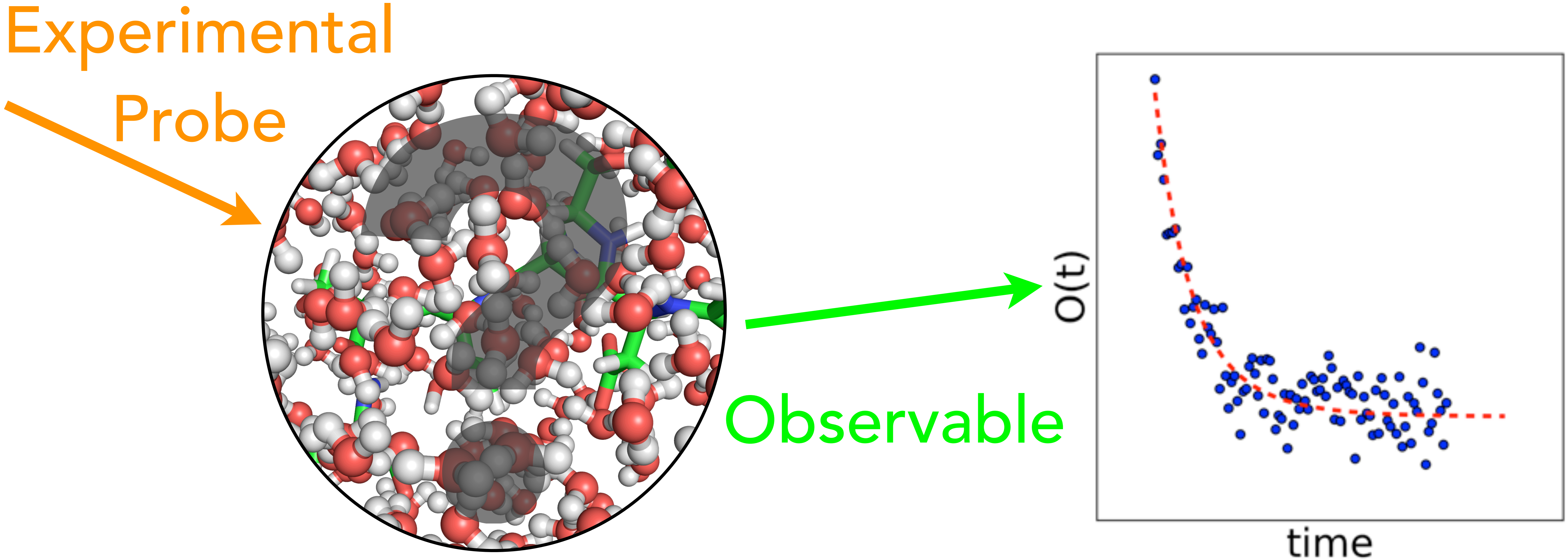
Molecular Dynamics Analysis

Robert McGibbon, Muneeb Sultan, and Christian Schwantes

March 29, 2014

Value of Simulation

- Experiments provide projections of the high-dimensional protein folding process
  - Determining the microscopic mechanism from these projections is difficult

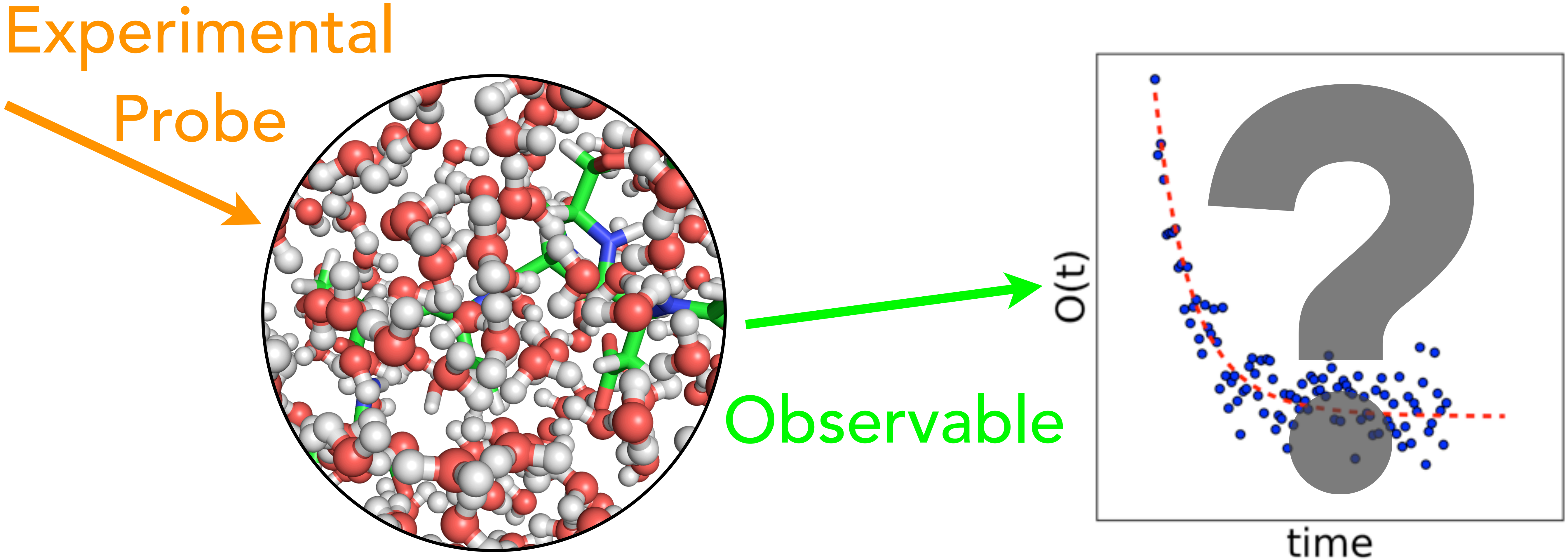


Experimental Probe

Observable

Value of Simulation

- Simulation can provide an atomic-level description that most experiments cannot
  - By predicting experimental observables, we can validate our models

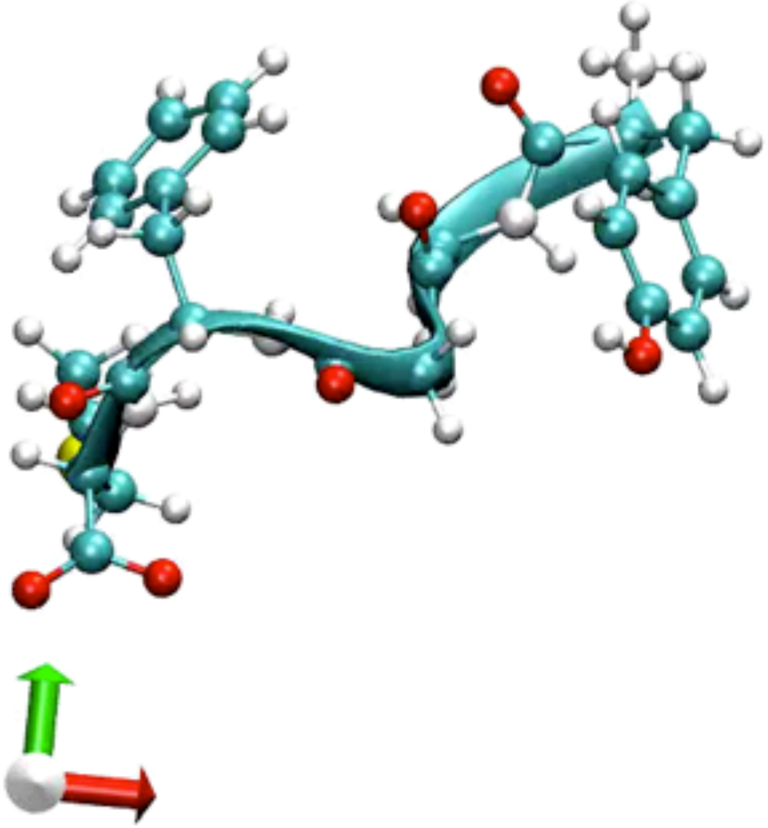


Experimental Probe

Observable

Molecular Dynamics (MD)

- Let’s say we’ve taken a lot of computer (and human) time to generate a large set of MD trajectories
- ***Now what?***
  - We can certainly make a pretty cool movie
  - But MD is so much more than a YouTube clip!
- We want to understand our results



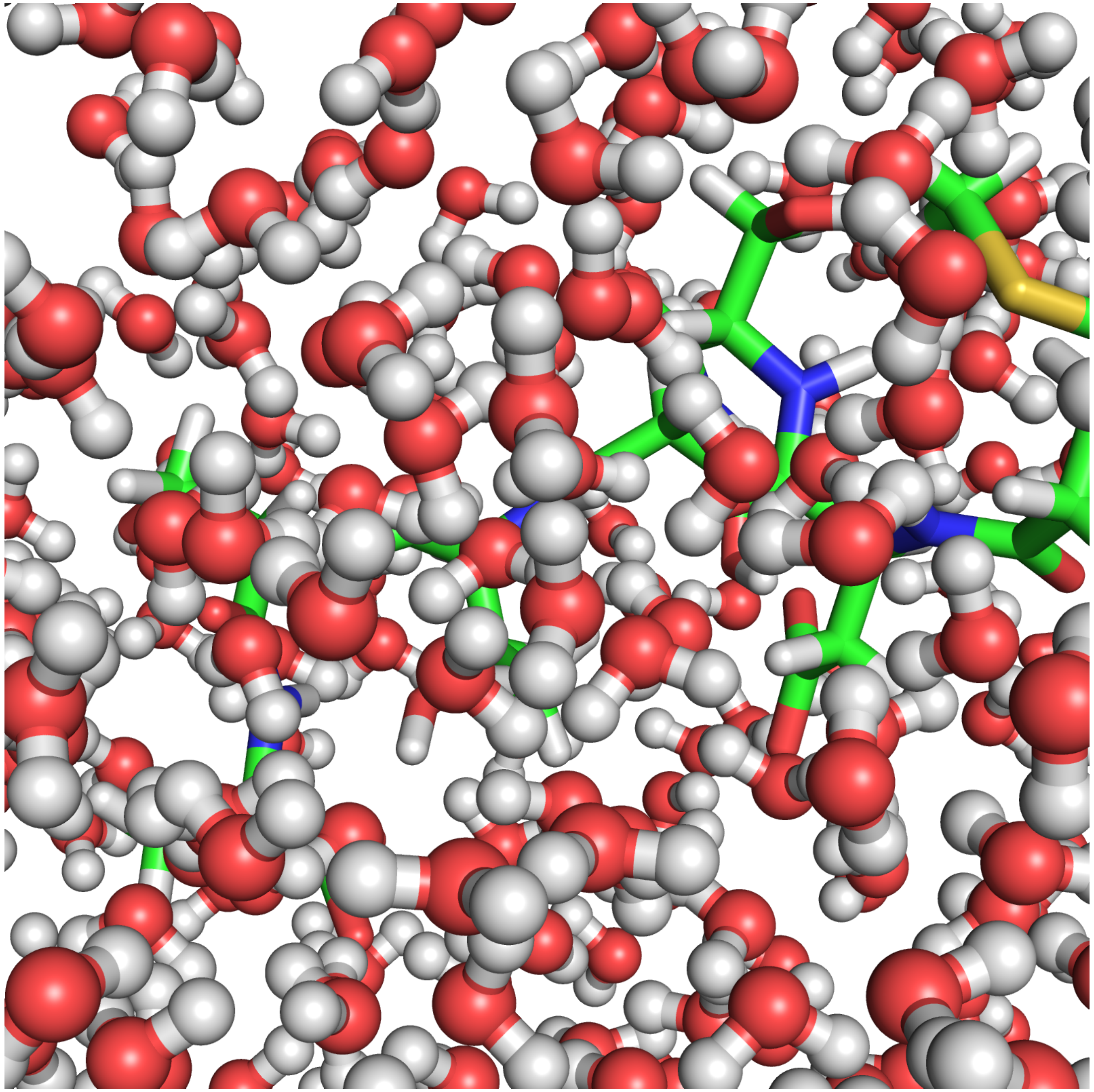
Quantitative Analysis

- MD datasets are too high-dimensional to simply make sense of out of the box
  - A typical molecular dynamics data set has 25,000+ atoms
  - We frequently have datasets that are hundreds of microseconds or even milliseconds (millions of frames)
- So we need to simplify!
  - But we also want to be sure that we don’t simplify in such a way that we lose important information

Dimensionality Reduction

- We need to simplify the picture in order to make sense of it!
- We already do this!
  - Throw out velocities

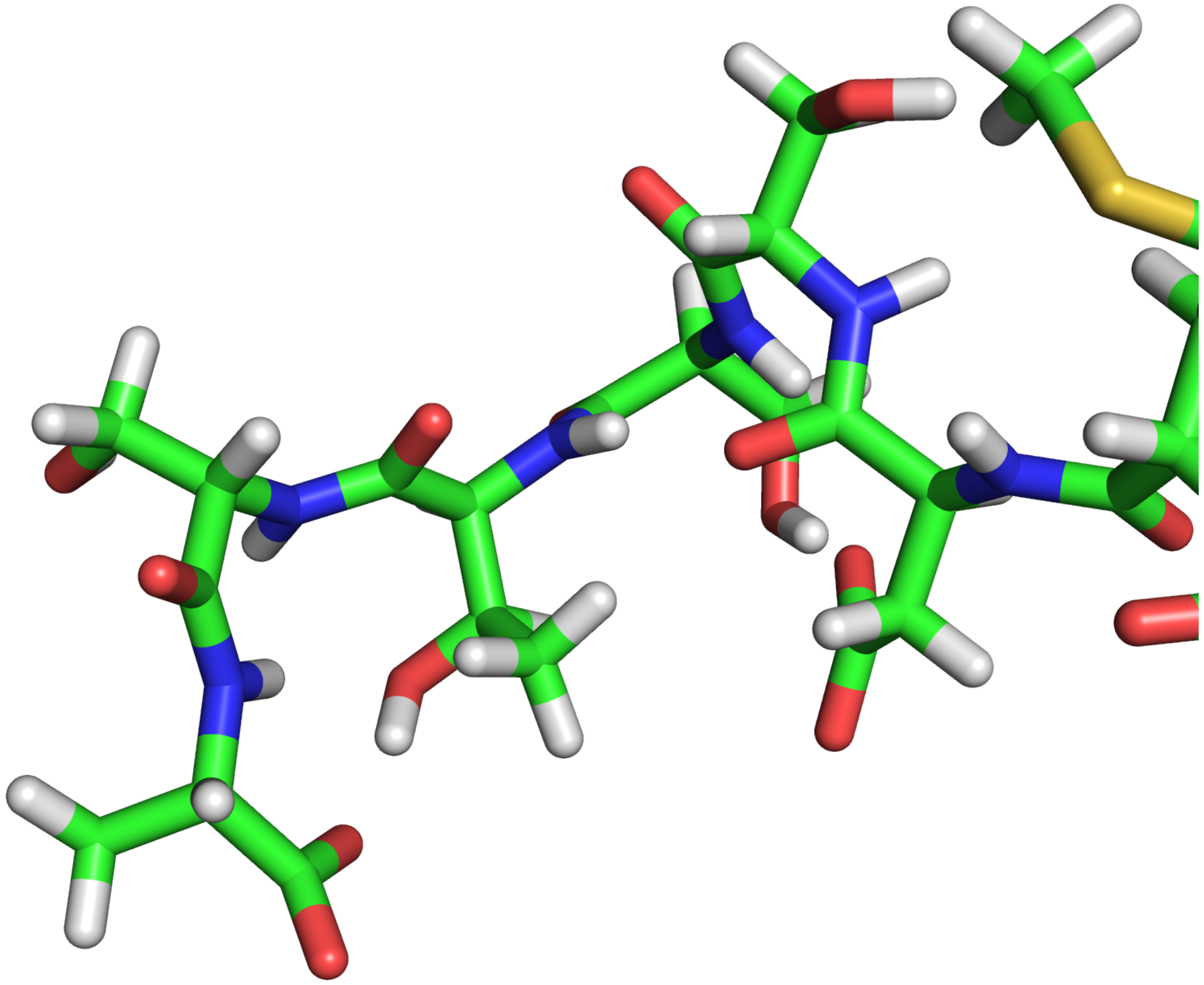
$$(\vec{\mathbf{x}}, \vec{\mathbf{v}}) \rightarrow (\vec{\mathbf{x}}, \xcancel{\vec{\mathbf{v}}})$$



Dimensionality Reduction

$$(\vec{\mathbf{x}}, \vec{\mathbf{v}}) \rightarrow (\vec{\mathbf{x}}, \xcancel{\vec{\mathbf{v}}})$$

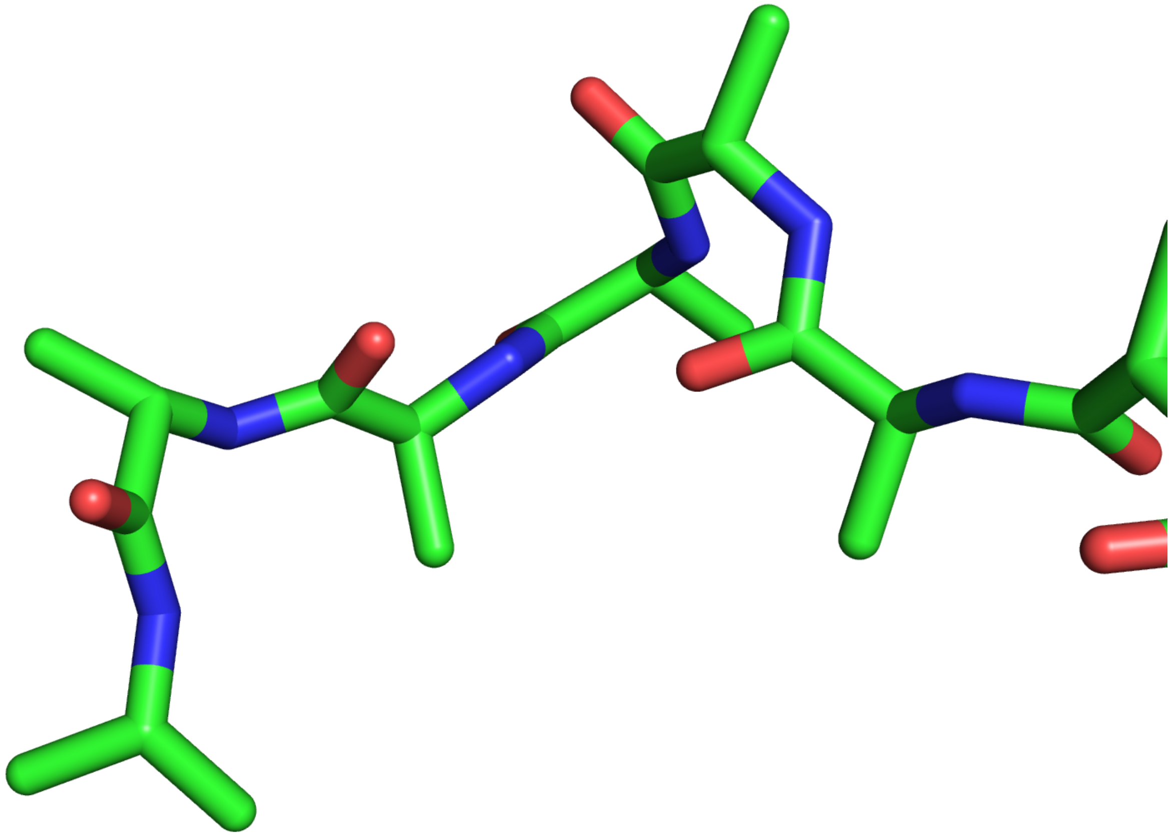
- We need to simplify the picture in order to make sense of it!
- We already do this!
    - Throw out velocities
    - Throw out solvent degrees of freedom



Dimensionality Reduction

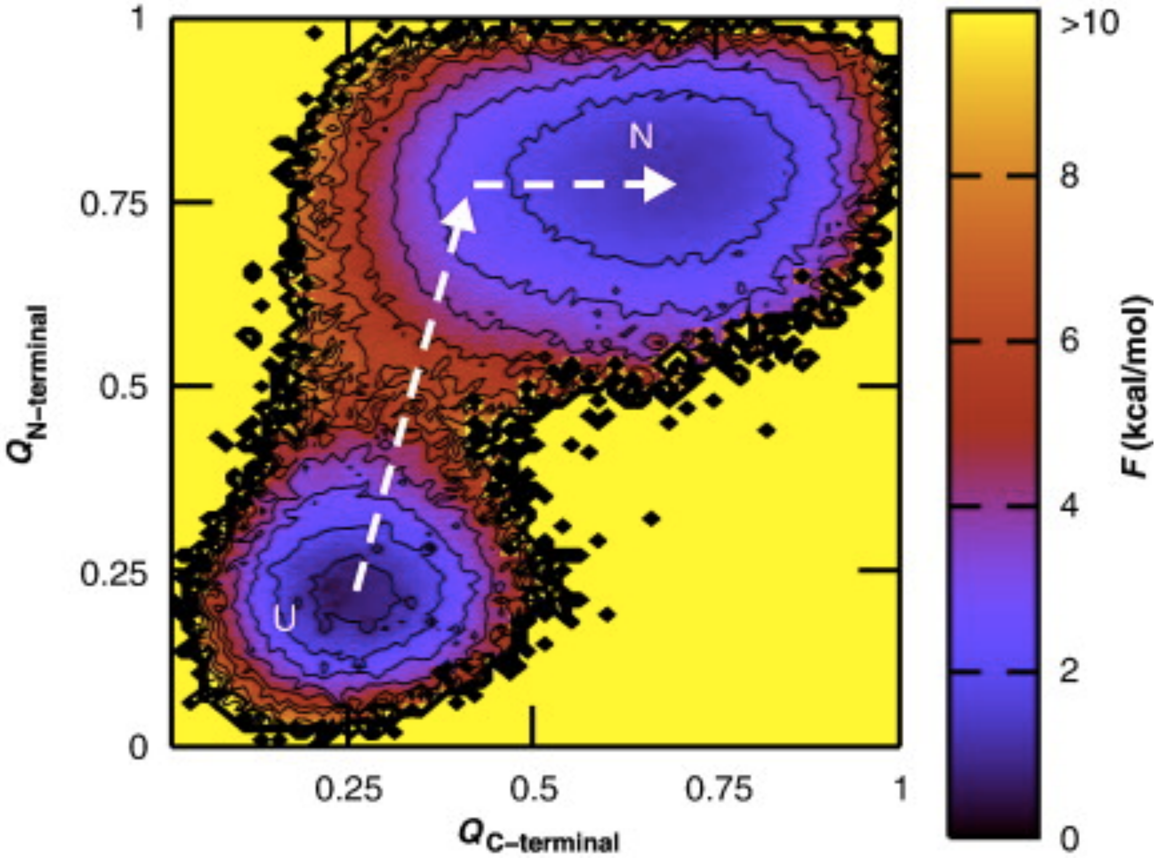
$$(\vec{\mathbf{x}}, \vec{\mathbf{v}}) \rightarrow (\vec{\mathbf{x}}, \cancel{\vec{\mathbf{v}}})$$

- We need to simplify the picture in order to make sense of it!
- We already do this!
  - Throw out velocities
  - Throw out solvent degrees of freedom
  - Only consider a subset of the atoms



Projection-Based Analysis

- Even if we just consider a subset of all of the atoms, our dataset is usually still very high-dimensional!
    - For example, a typical protein might have 500 atoms, which means we have a vector of length 500 x 3 that changes in time
- The solution: turn each high-dimensional vector into one or two projections



Hills, RD Jr. and Brooks, CL III. *J. Molec. Biol.* **2008**

Common Projections

- In biomolecule simulations, several projections (reaction coordinates) are very common:
  - RMSD to a crystal pose, radius of gyration
  - Fraction of “native contacts” formed
  - Important residue - residue distances
  - DSSP assignments (secondary structure)
- In the protein folding field, many people have their “favorite” version of one of the above
- Other structural characterizations exist for non-protein systems

Common Projections

- Root mean square deviations of atomic position (RMSD)

$$\mathrm{RMSD}(\mathbf{X},\mathbf{Y}) = \min_{\mathbf{R}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} ||X_i - (\mathbf{RY})_i||^2}$$

$$\text{s.t. } \mathbf{R} \text{ is a rotation matrix}$$

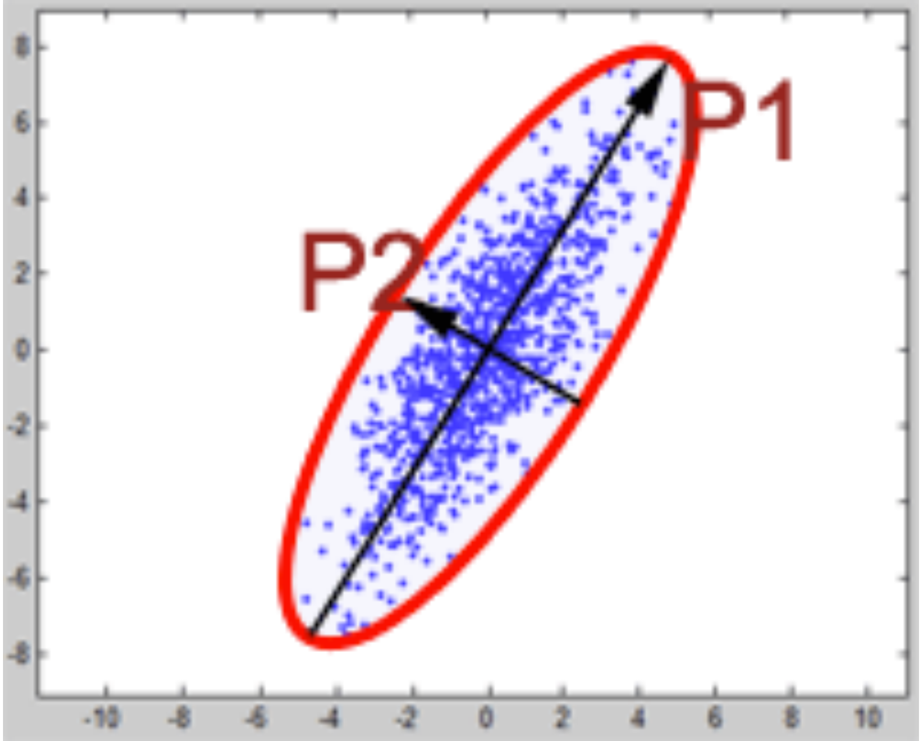
- Radius of gyration

$$R_g(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - X_{mean})^2}$$

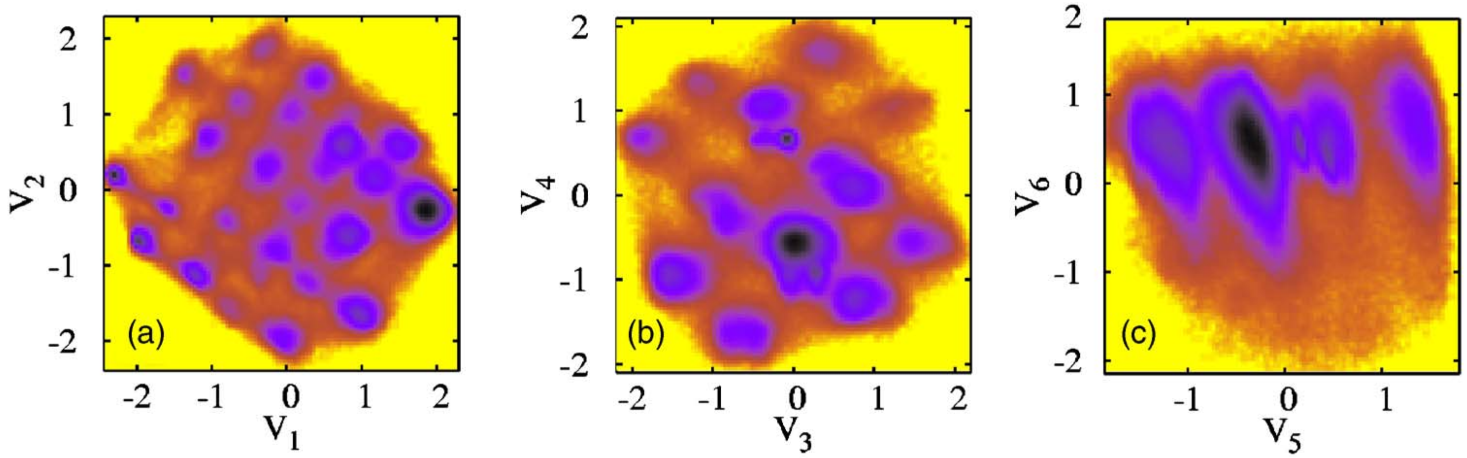
- Fraction of native contacts

$$q(\mathbf{X}) = \sqrt{\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} I_{(\text{c formed in } \mathbf{X})}}$$

Statistical Projections



Altis A, *et al. J. Chem. Phys.* **2008**

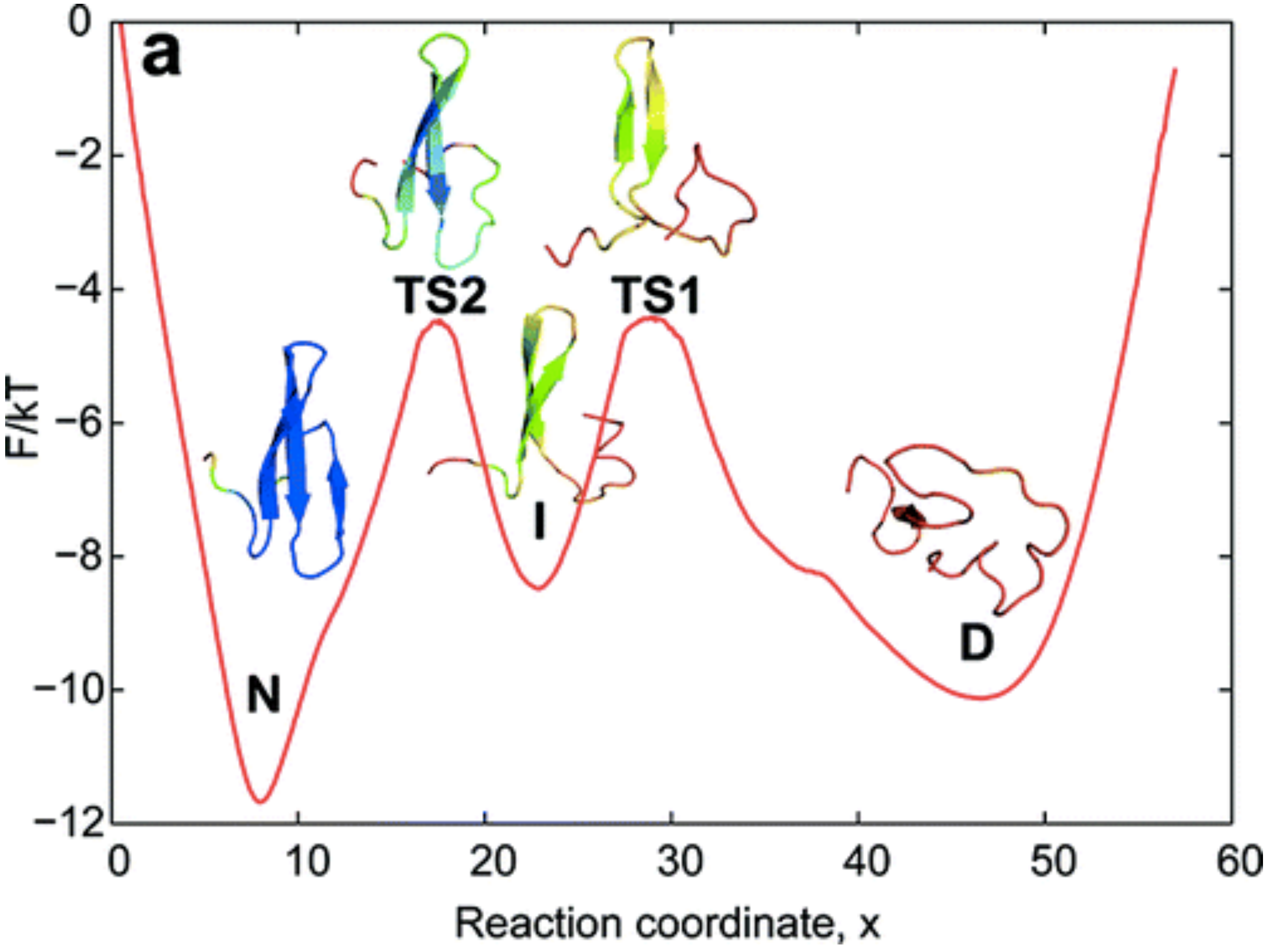


PCA applied to protein folding simulations shows many free energy minima in the PC space

- Another common tool is Principal Components Analysis (PCA), which looks for a projection that has maximal variance
- This is useful for exploratory analysis, but assumes that high variance is an indicator of “importance”

Kinetic Analysis of Projections

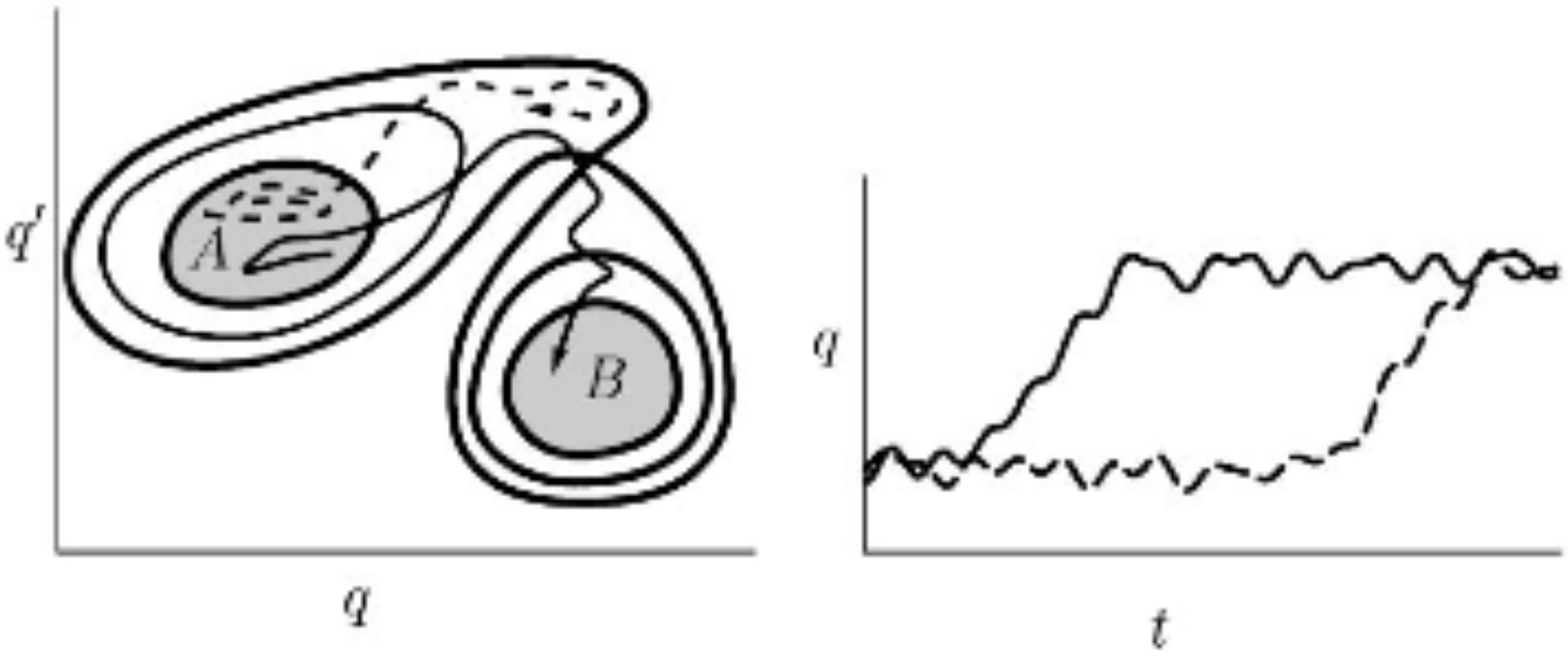
- If dynamics along the projection (reaction coordinate) are *slower* than dynamics in the orthogonal subspace, then dynamics can be modeled in the projection
  - The orthogonal subspace acts like a heat bath
- But the analysis will depend on how good your reaction coordinate is



Krivov S. V. *J. Phys. Chem. B* **2011**

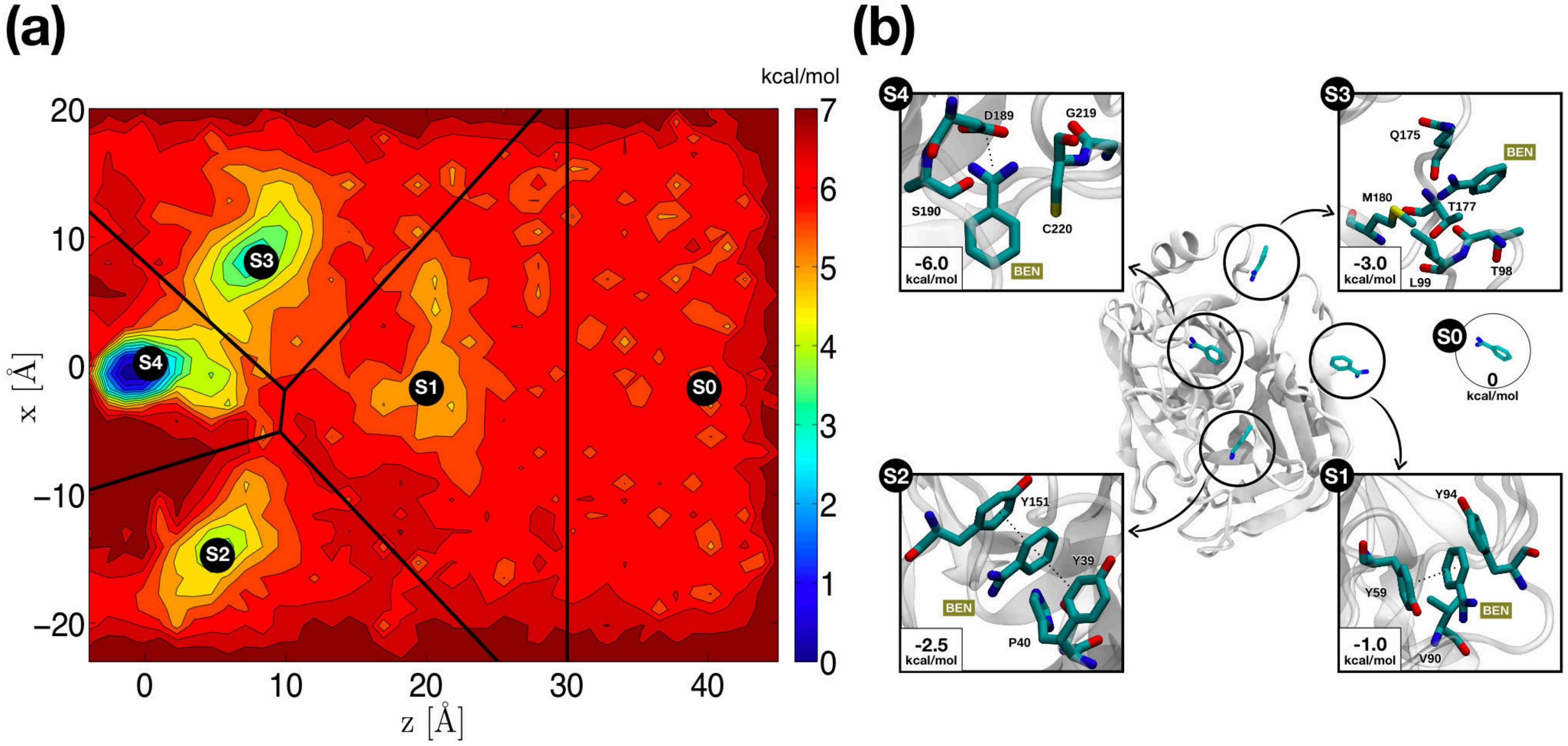
Why Not Stop There?

- Projections can filter out *critical information.*
- Say, you're analyzing the potential below and asking how long it takes to go from A to B
- *If you just monitor the variable q, then you may think you've transitioned to B when in fact you haven't!*



Boolhuis, P.G. et al. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291-318.

Motivated Projections



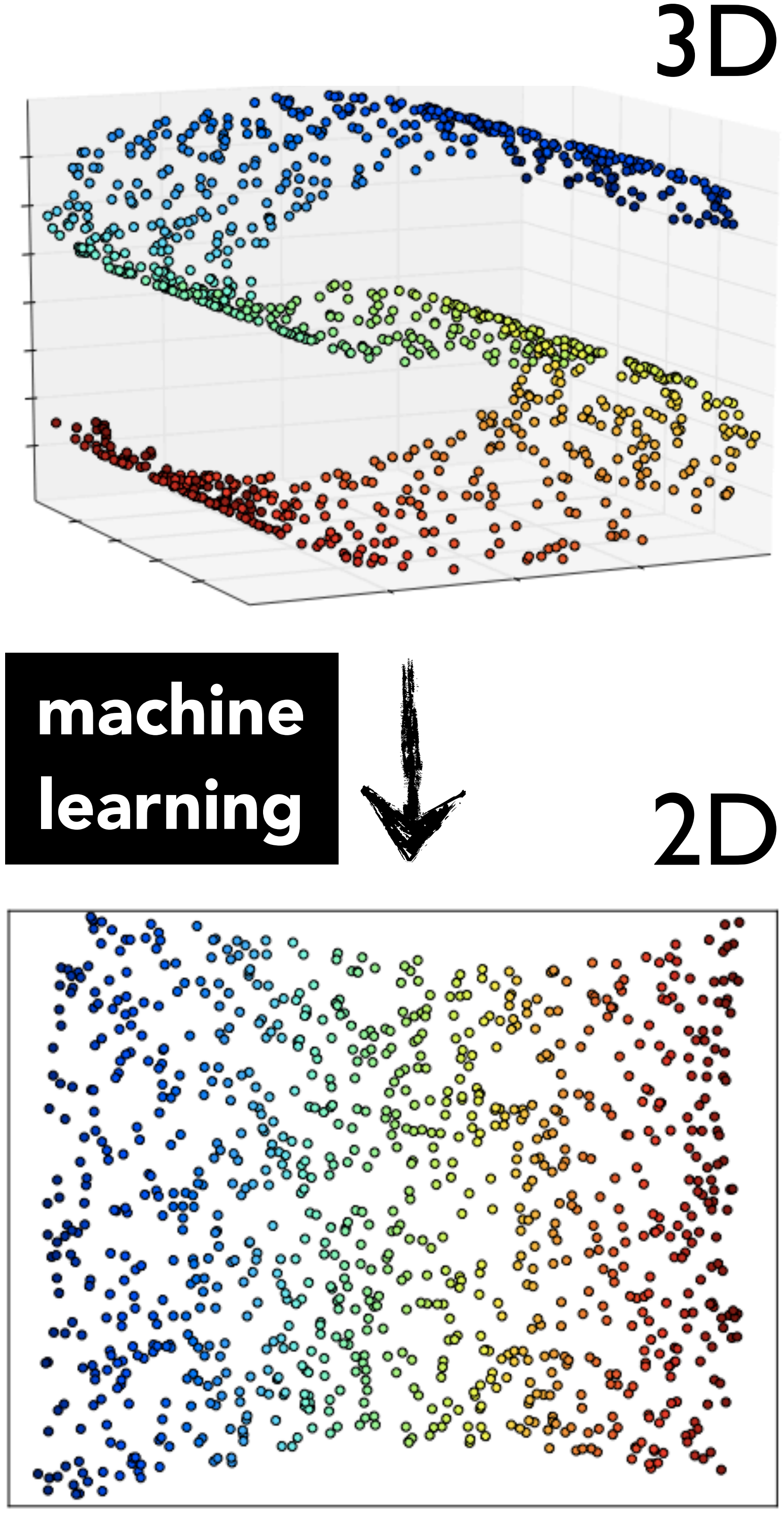
In protein ligand binding, a really easy projection that works well, is the location of the ligand relative to the protein

Buch, I. *et al.* *PNAS* **2011**

- In order to be confident in a projection-based method you need to know that you're picking the right thing
  - In many systems, you actually already know the answer!
    - Conformational changes in kinases or well-studied enzymes
    - Protein-ligand distance

(Machine) Learning Projections

- There are other projection based methods that attempt to pick the *correct* degrees of freedom in an automated way
  - ISOMAP / Diffusion Maps (nonlinear)
  - tICA (use time)
- The usefulness of these techniques will depend on the properties of your data



3D

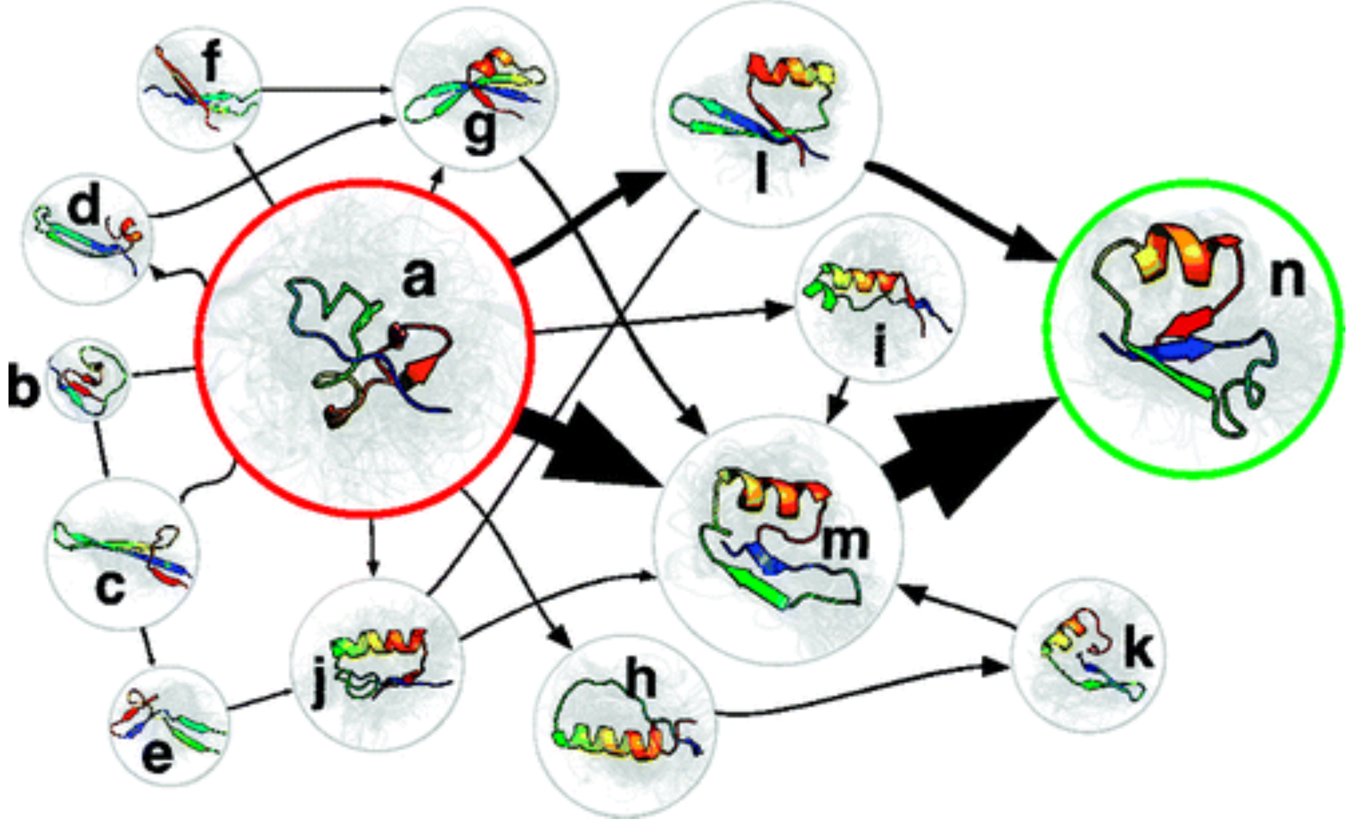
machine learning

2D

MSMs Move Beyond Projections

- Remember that projections were useful because they simplified the high-dimensional dataset into something that we could understand

- Master equations (Markov state models) approach this from a different perspective.



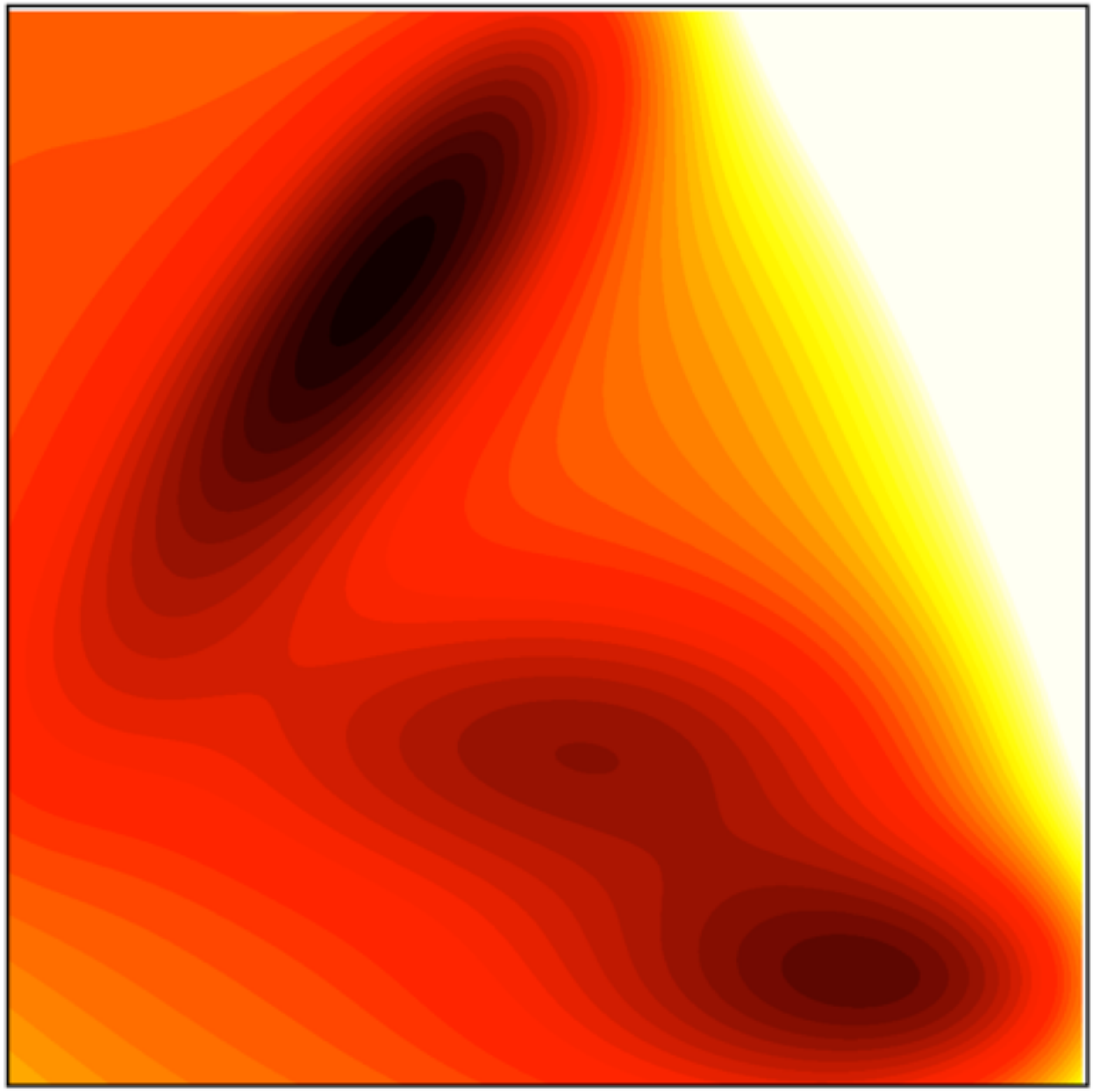
An MSM is a set of states and probabilities of transitioning between these states

It's extendable to any simulation in which you can define the states.

Voelz, V.A. *et al. J. Am. Chem. Soc.* **2010**, *132*, 1526-1528.

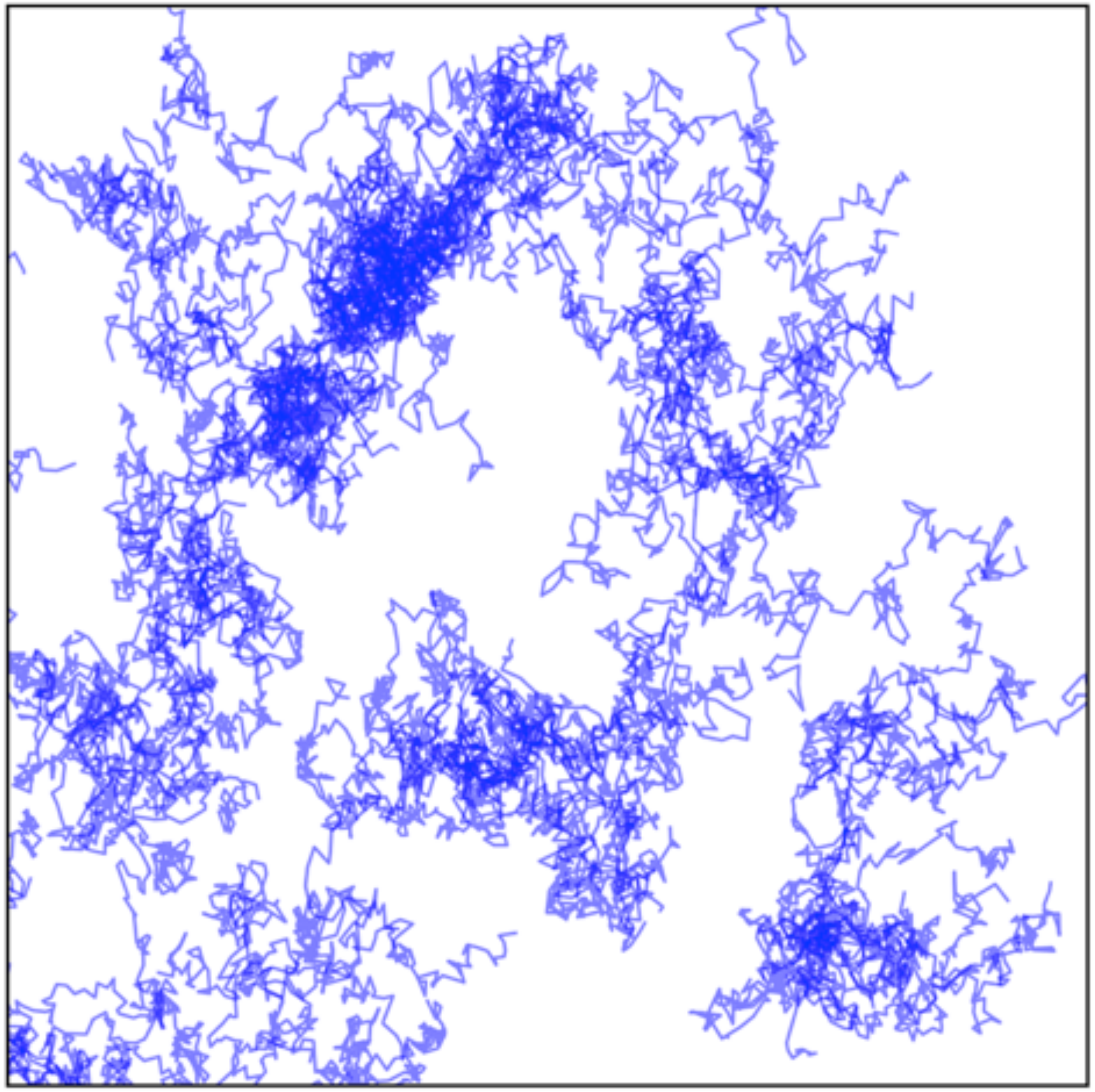
MSM Construction

- So how do we build an MSM?
- We start with some Hamiltonian, for example the Muller potential on the right
    - Sample the system with standard MD.
    - The goal is to describe the thermodynamics and kinetics in terms of a set of states and rates



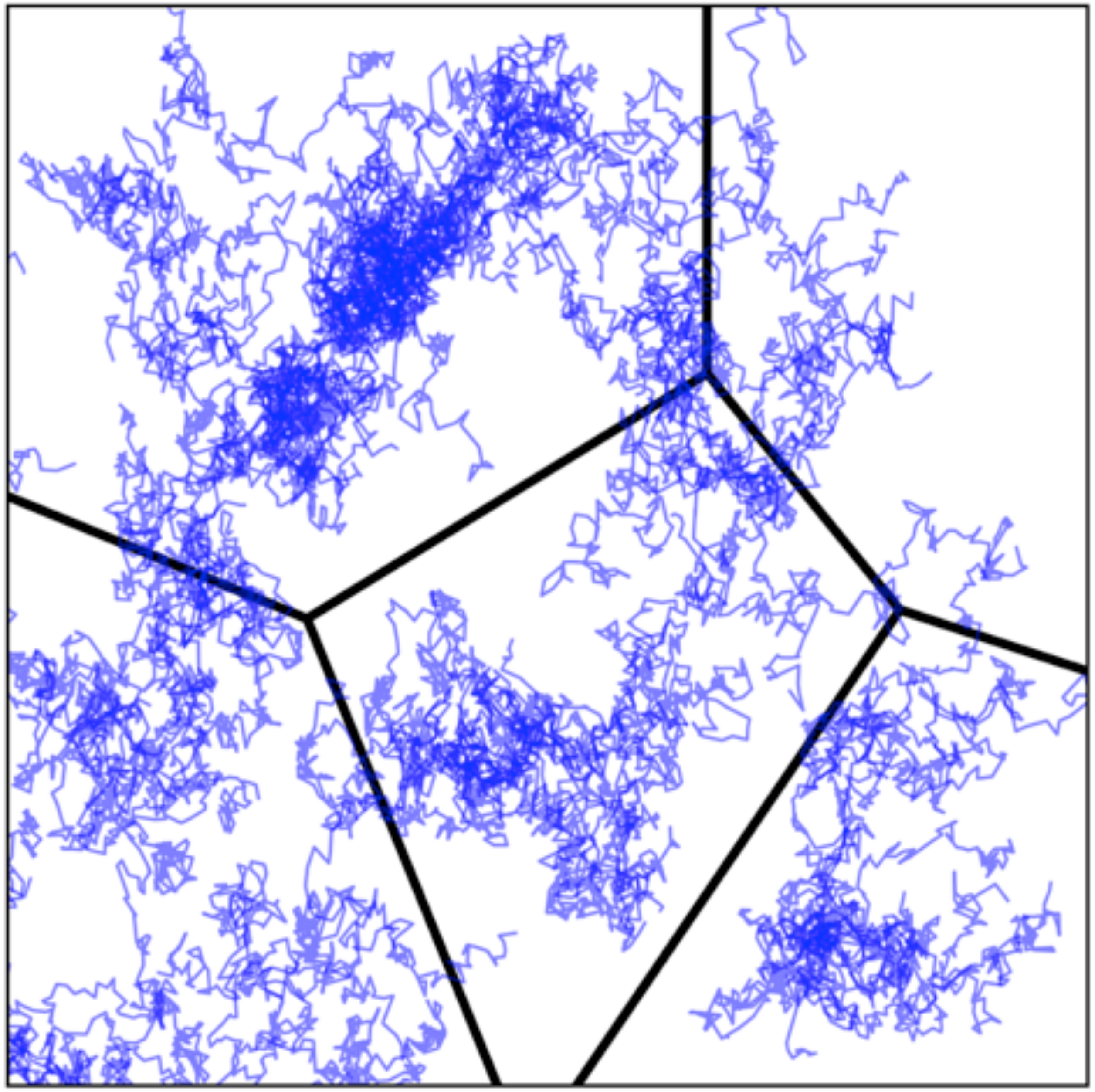
MSM Construction

- Sampling is not easy, but can be aided by:
  - Enhanced Sampling techniques
  - Lots of cores
  - Fast hardware (GPUs, Anton)



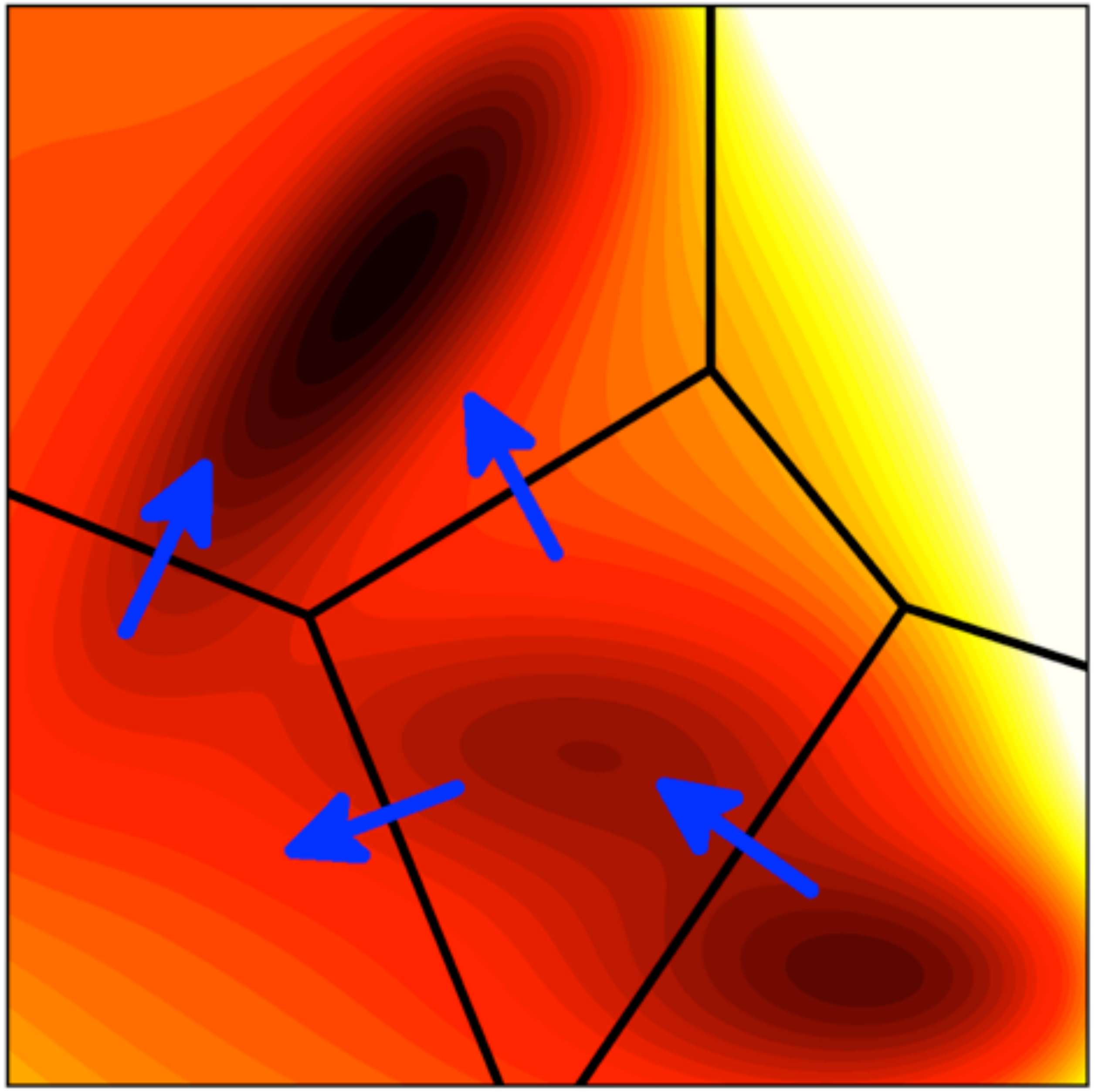
Markov State Models

- From the sampled data, we then:
  - Define a discrete set of states
  - Calculate the rates of transferring between them
- These states should consist of points that can interconvert rapidly
- Poor state decompositions can lead to poor MSMs



MSM Construction

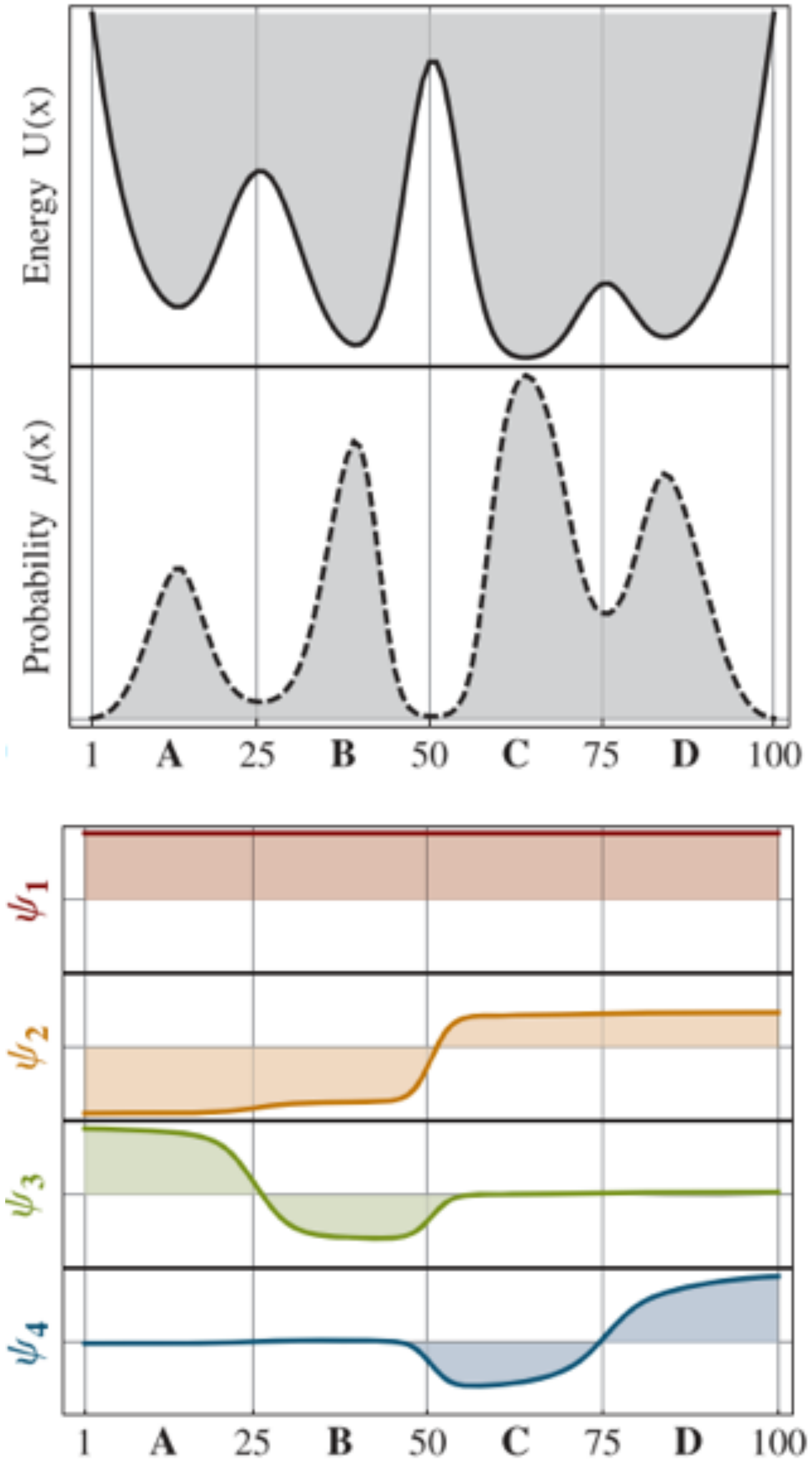
$$\vec{p}(t+\tau) = \vec{p}(t)\,\mathbf{T}$$



- We now have a model for the dynamics of our system
- There are many practical issues that come up in the process:
  - *How many states should we use?*
  - *How should the states be arranged spatially?*
  - *How do we validate an MSM?*

MSM Analysis

- Now that we have an MSM, what can we do with it?
  - Characterize long timescale dynamics (eigenspectum of **T**).
  - Find “macro-states” that are mutually kinetically separated.
  - Calculate the mean first passage time between sets of states
  - Quantitatively compare to / predict experimental results
  - Probe mechanism



Prinz, Jan-Hendrik, et al. JCP 134.17 (2011): 174105.

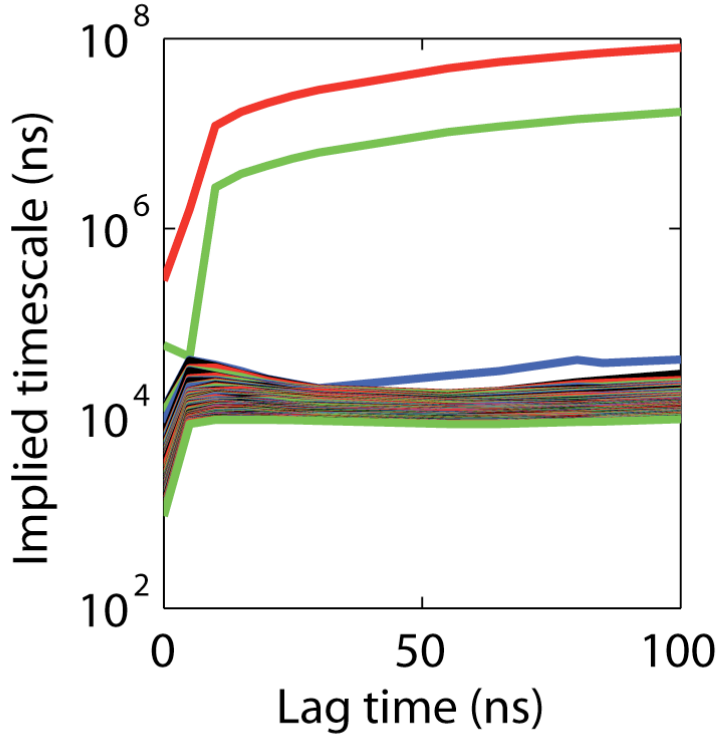
Slowest Timescales

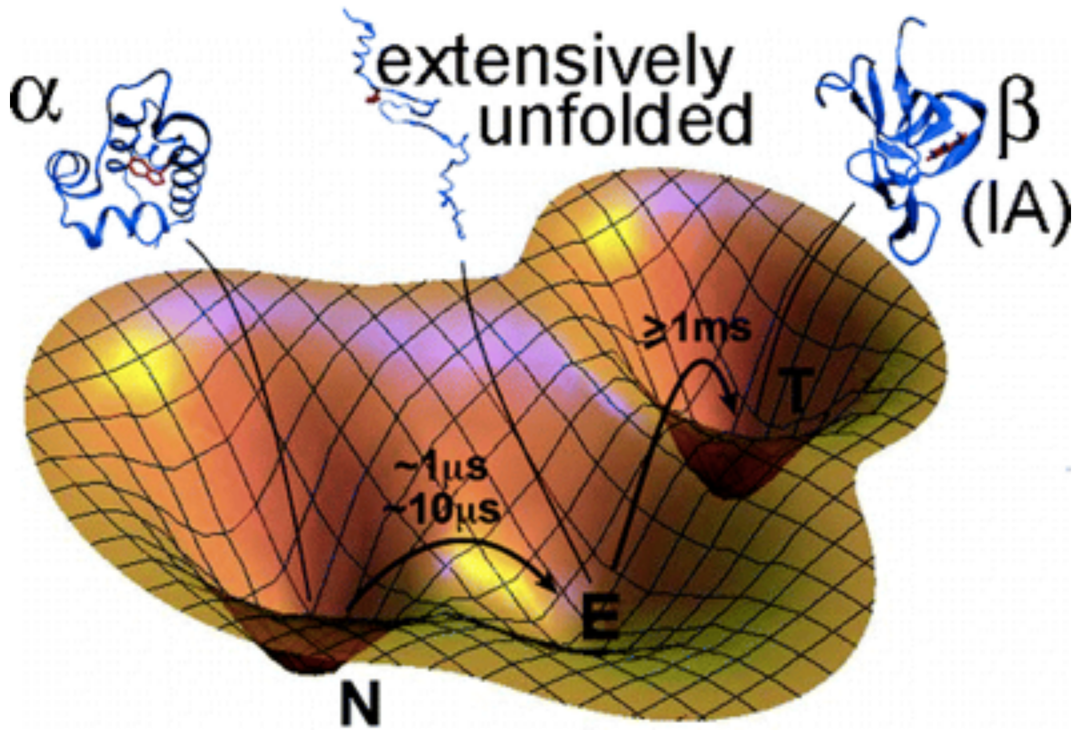
- The MSM's transition probability matrix can be decomposed into a sum of relaxation timescales:

$$\begin{aligned} p(t+n\tau)^T &= p(t)^T T \\ &= \sum_{i=1}^{\infty} \lambda_i^n \langle p(t), \psi_i \rangle \phi_i \\ &= \sum_{i=1}^{\infty} \exp\left(-\frac{n\tau}{t_i}\right) \langle p(t), \psi_i \rangle \phi_i \end{aligned}$$

  - Protein folding simulations typically have a slowest timescale corresponding to the folding transition

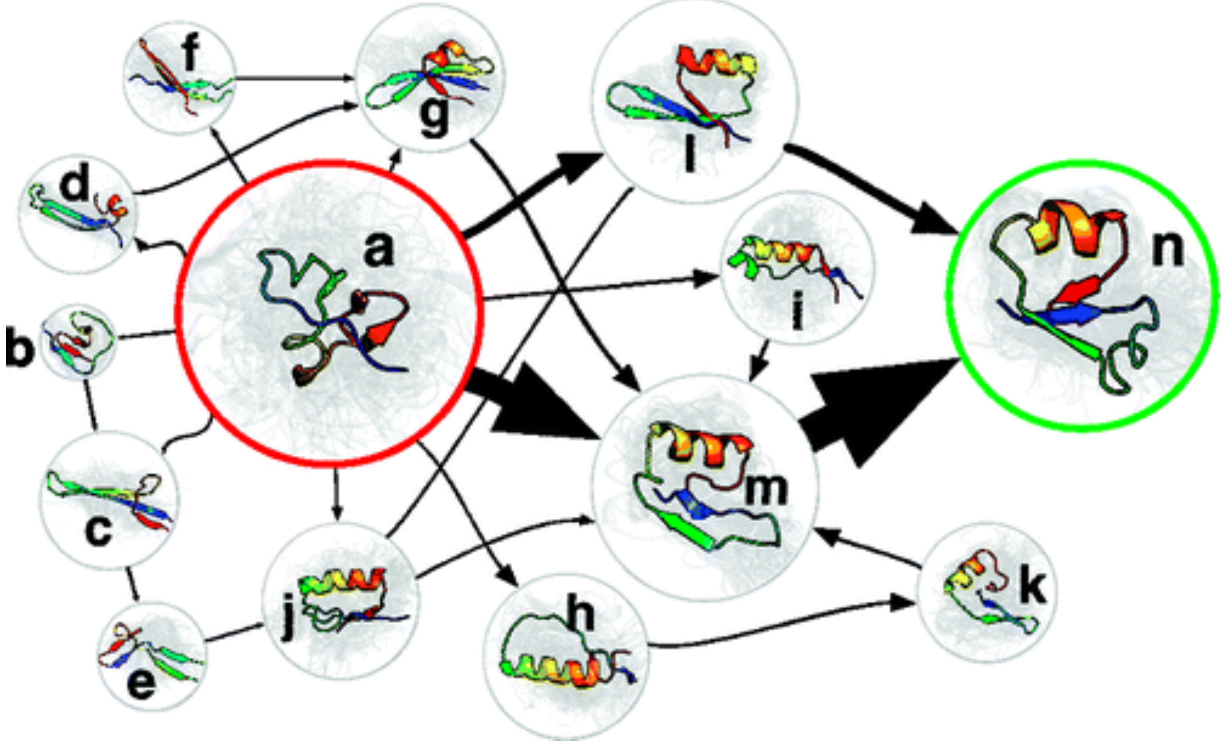
Bowman, G. *et al. JACS* **2011**





Prigozhin, M. B. *et al. JACS* **2011**

Lump it!

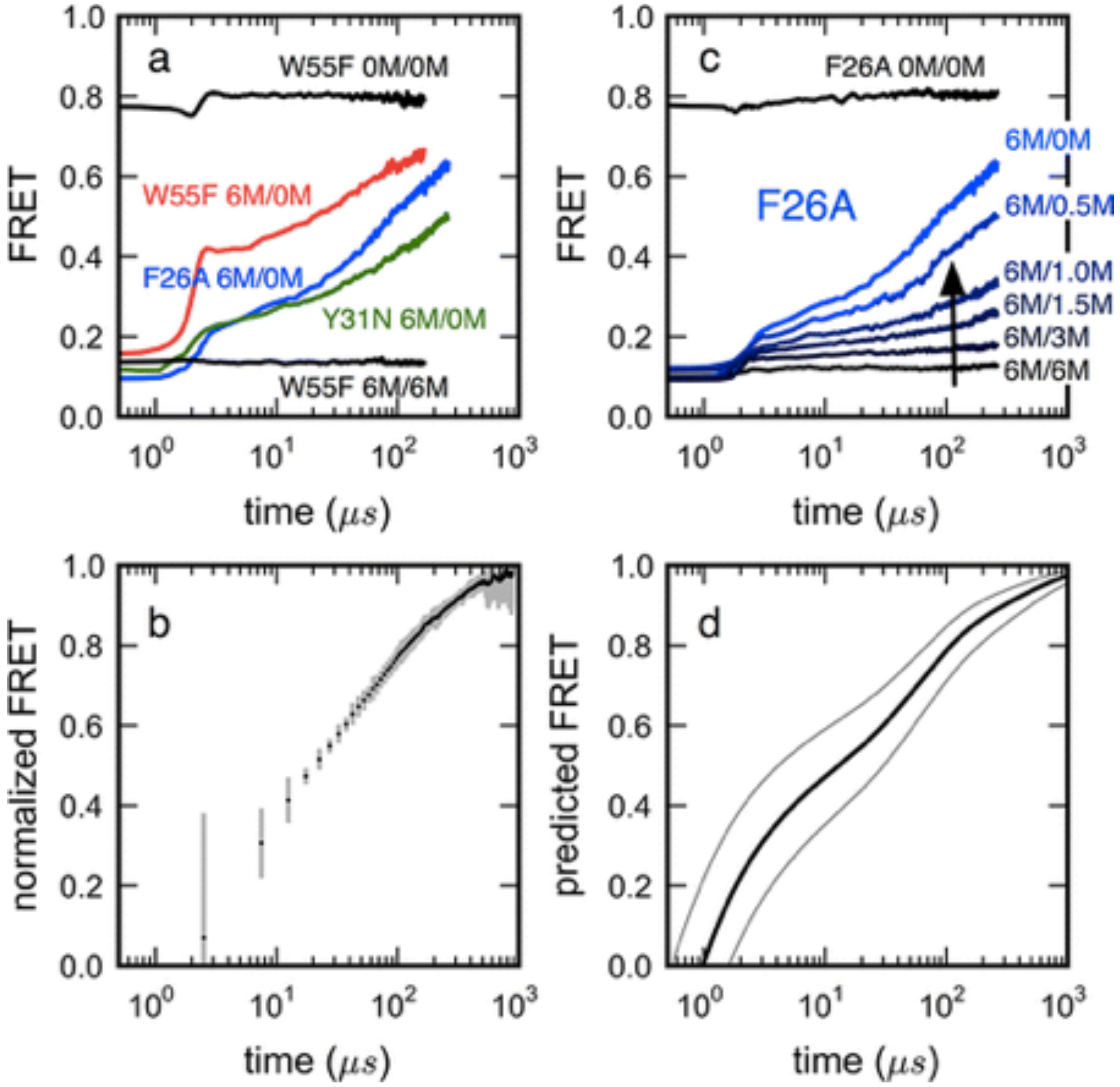


Voelz, V.A. *et al*. *JACS* **2010**, *132*, 1526-1528.

- Typical MSMs contain thousands of states
  - This is simpler than what we started with, but it's still not simple...
- There are many schemes for "lumping" a given MSM such that the slowest timescales are preserved
- This can be used to build a smaller model so that you can understand the qualitative features

Compare to Experiment

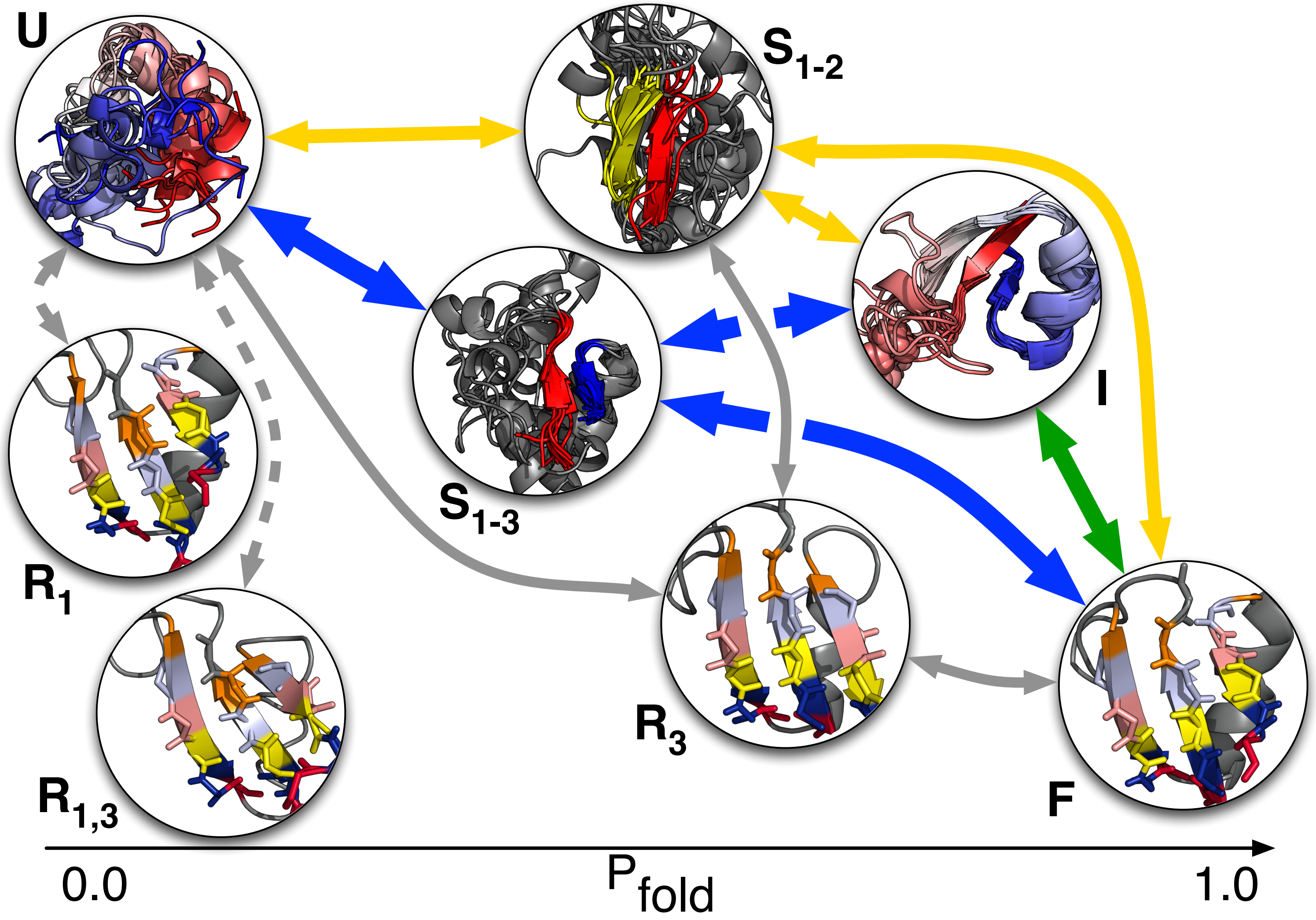
- Since we can propagate any trajectory in the MSM, it's trivial to calculate experimental observables along the way!
  - We can do this in a quantitative fashion
  - What you'll typically find is that one or two eigenvectors end up being prominent features in certain experiments
    - This allows you to provide a molecular interpretation of an experiment



Voelz, *et al. JACS* **2012**

Determine the Mechanism

- For protein folding, at least, many are interested in how a protein goes from unfolded to folded
- Within the MSM framework, you can calculate the most probable transition paths (via Transition Path Theory [TPT])



TPT often reveals many on-pathway intermediates that would be difficult to pick out while watching a movie

Schwantes, C.R. *et al. JCTC* **2013**