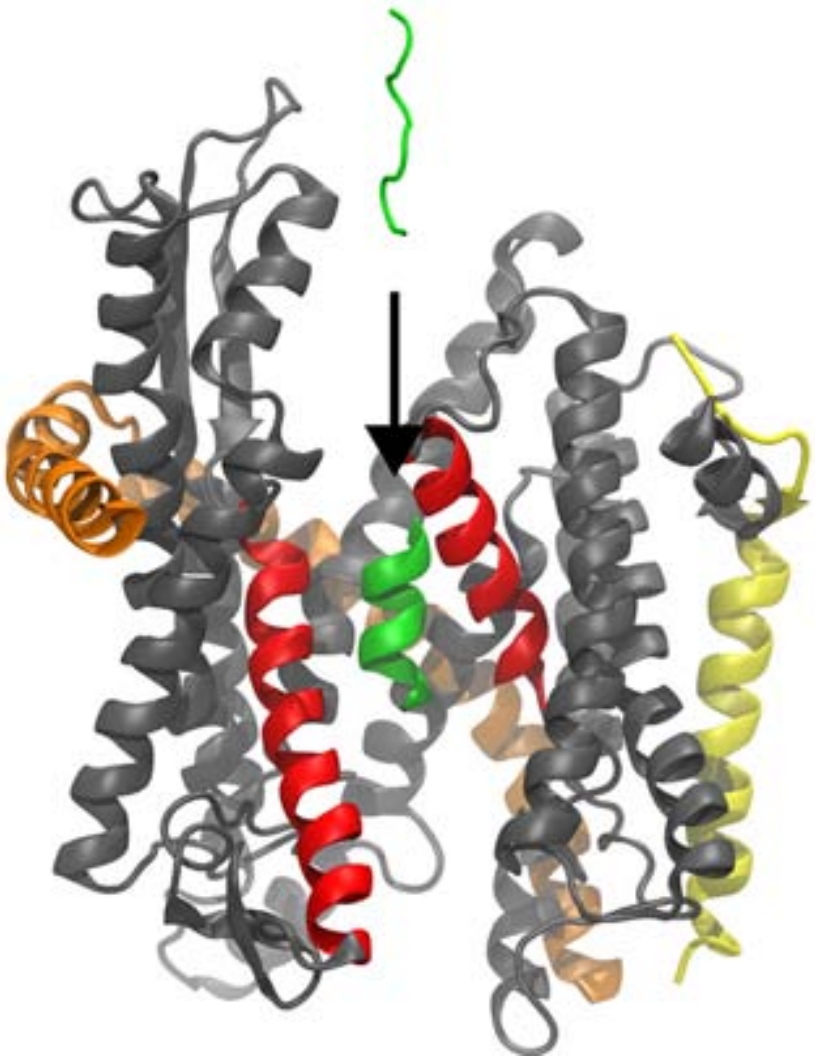
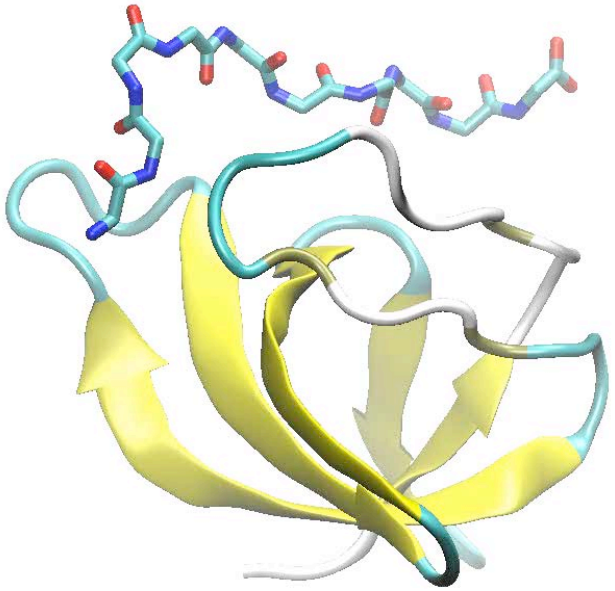
Application of free-energy methods in NAMD to complex biological systems



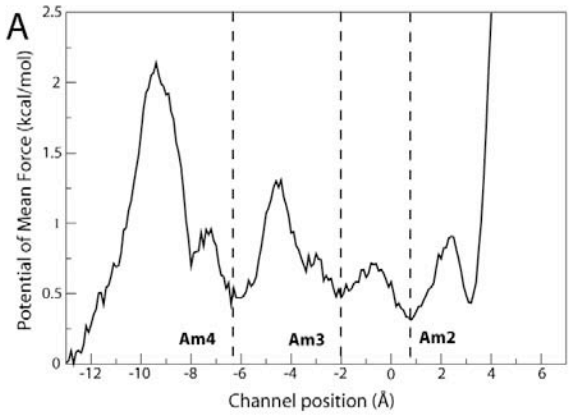


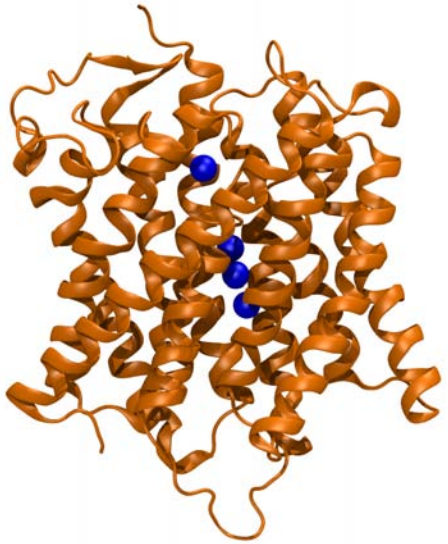
JC Gumbart
Argonne National Lab, Chicago IL
June 6, 2012

The importance of PMFs in biology

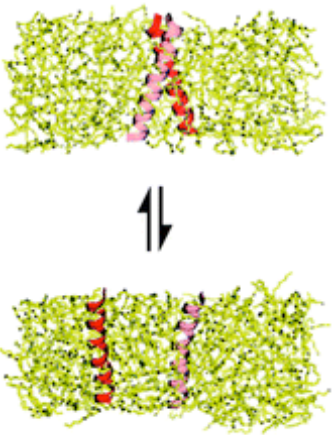
-potential of mean force[1] for a well chosen reaction coordinate can characterize many relevant processes including

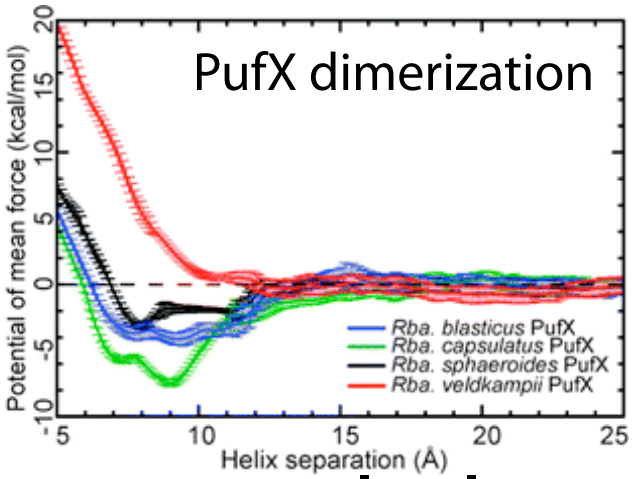
1: Kirkwood, J. G. (1935). *J. Chem. Phys*. **3**: 300–313.





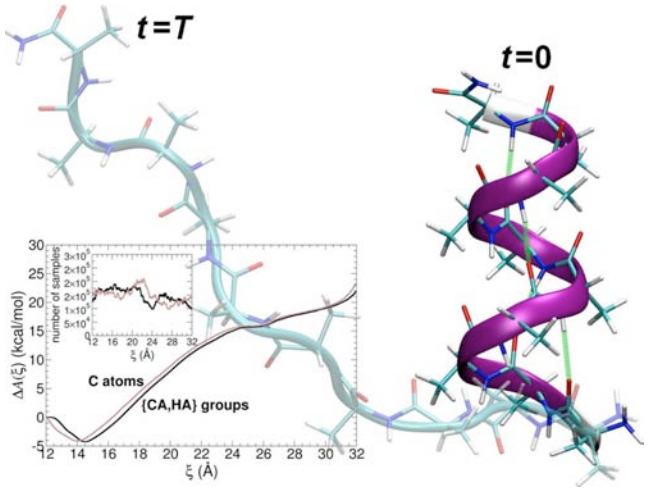
substrate translocation through a channel





interaction between two molecules, e.g., protein-protein

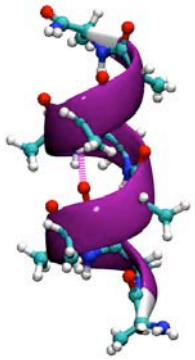
Hsin, J. et al. (2011) *J. Am. Chem. Soc.*, **133**: 14071–14081.



Henin, J. and Chipot, C. (2004). *J. Chem. Phys*. **121**: 2904.

conformational change, e.g., folding

Potential of mean force from ABF



$$A(\boldsymbol{\xi}) = -\frac{1}{\beta}\ln \mathcal{P}(\boldsymbol{\xi}) + A_0$$

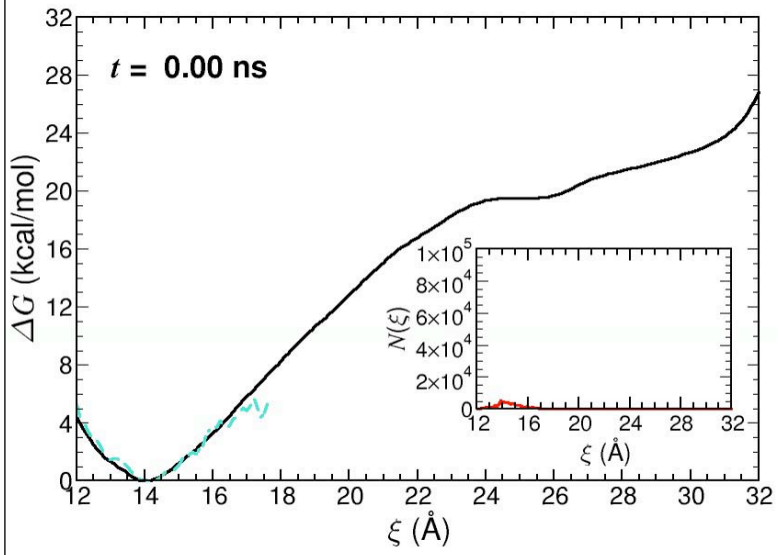
Free energy as function of ξ

$$\boldsymbol{\nabla}_{\xi} A(\boldsymbol{\xi}) = \langle -\boldsymbol{F}_{\xi} \rangle_{\boldsymbol{\xi}}$$

Relation to average force

$$\mathbf{F}^{\mathbf{ABF}} = \boldsymbol{\nabla}_{\boldsymbol{x}} \tilde{A} = -\langle F_{\xi} \rangle_{\xi} \boldsymbol{\nabla}_{\boldsymbol{x}} \xi.$$

Compute average force **adaptively** and apply **biasing force** to cancel it



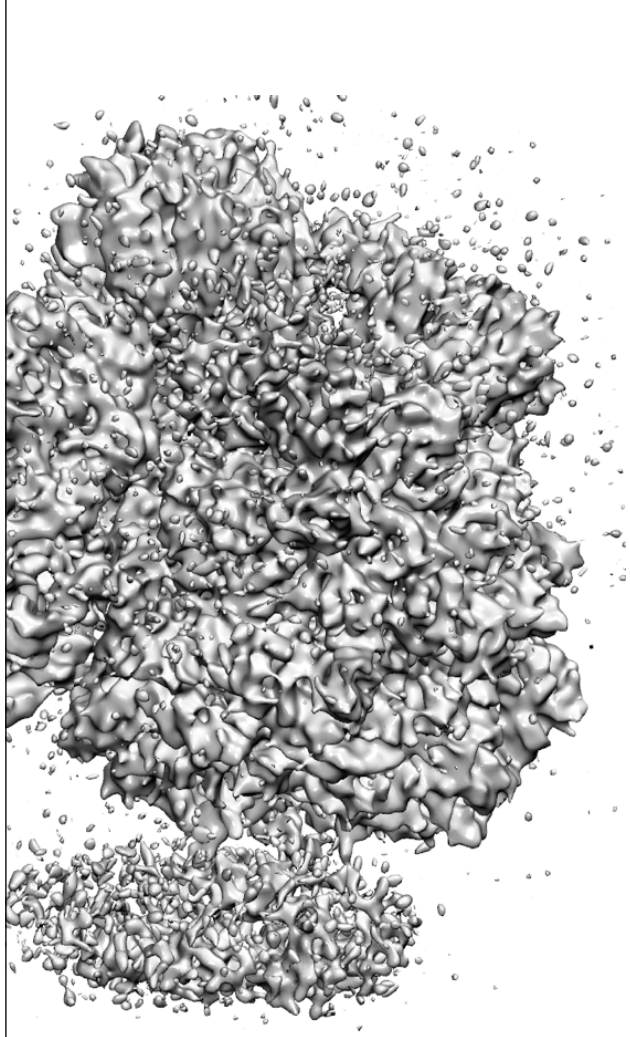
As the estimate of the PMF improves, the biasing forces *should* effectively cancel it, permitting the reaction coordinate to diffuse more easily (not always in practice though!)

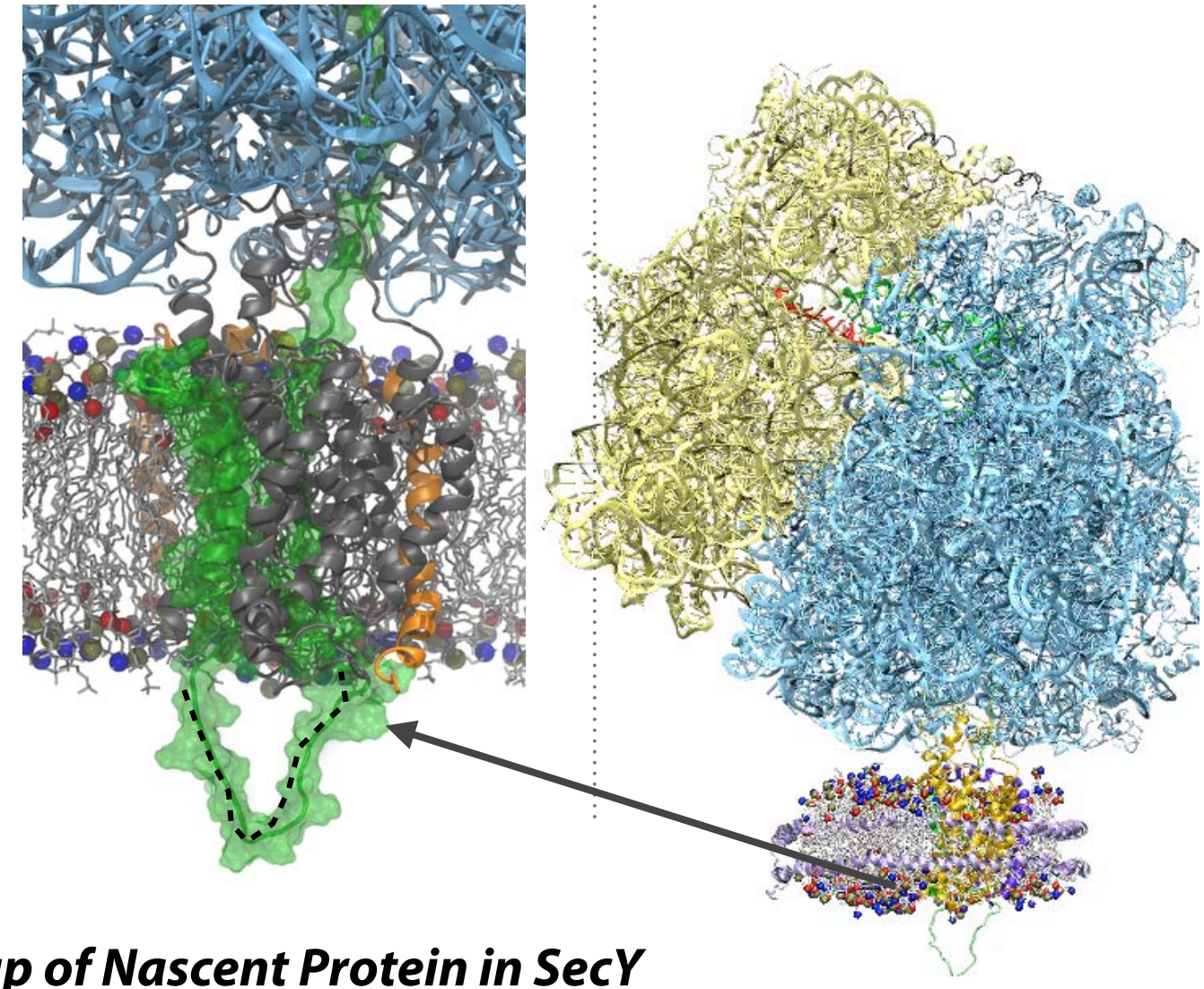
ABF takes advantage of **colvars** in **NAMD**, which includes many collective variables as possible reaction coordinates, e.g., distance, distanceZ, RMSD, etc. - **very versatile**

Challenge 1: Nascent (membrane) protein folding

Low-resolution Data

High-resolution Structure



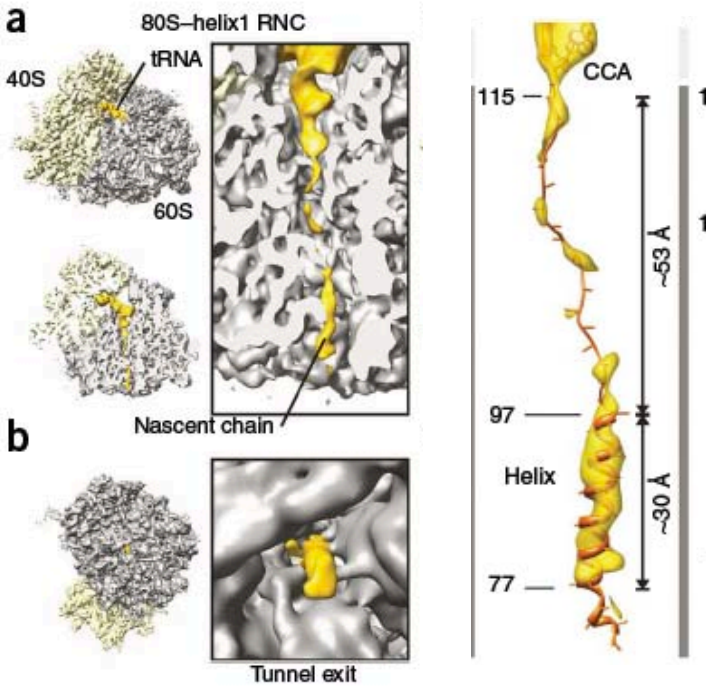


Close-up of Nascent Protein in SecY (already folded!)

J. Frauenfeld, J. Gumbart *et al*. (2011) *Nat. Struct. Mol. Bio.* **18**:614-621.

Proteins fold early in their development

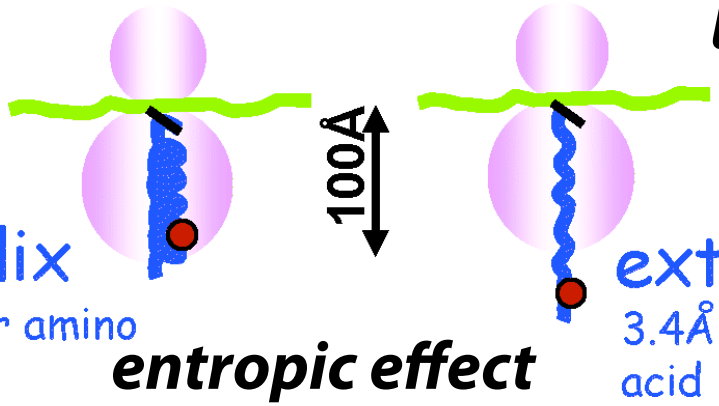
Many proteins can fold already within the ribosome and/or translocon



S. Bhushan *et al.* (2010) *Nat. Struct. Mol. Bio.* 17:313-318.

Folding in channel reflects that in water, NOT membrane

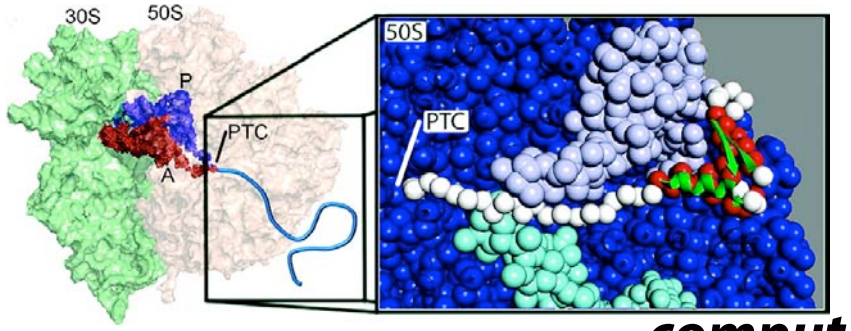
biochemical



J. Lu, C. Deutsch. *Biochemistry*. 44:8230-43.

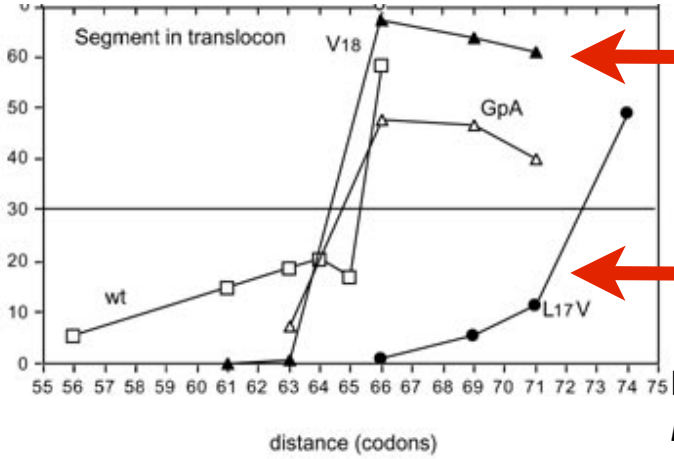
entropic effect

computational



E. P. O'Brien *et al.* (2010) *JACS*. 132:16928-37.

structural



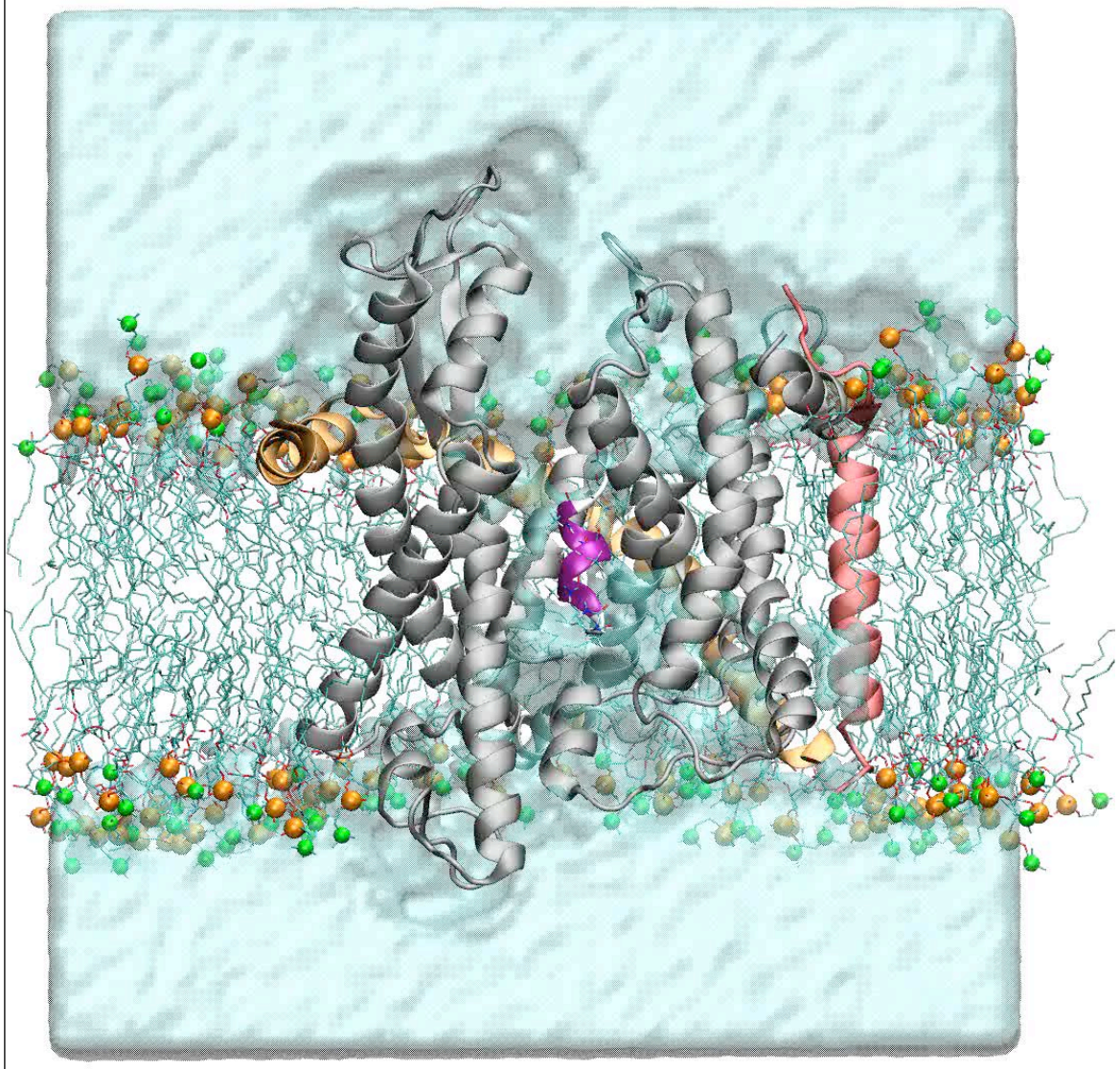
PolyVal: extended

PolyLeu: compact

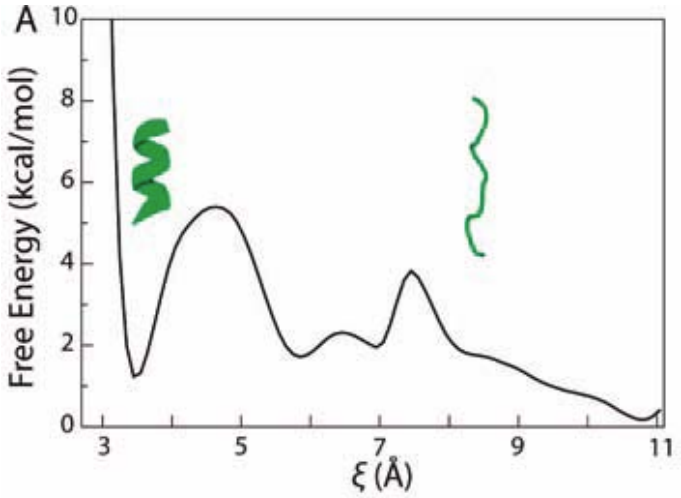
I. Mingarro *et al.* (2000) *BMC Cell Biol.* 1:3.

Folding at different positions in the channel

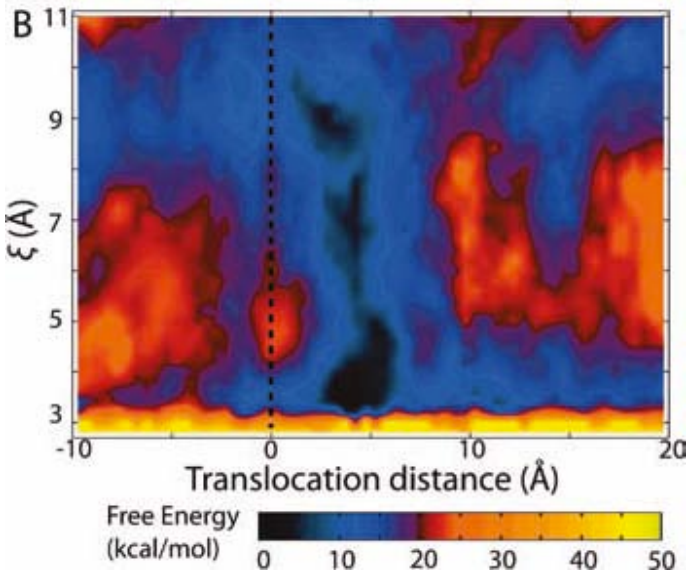
ABF in SecY



1D PMF for folding of Ala_{10} in water



2D PMF for folding of Ala_{10} in SecY

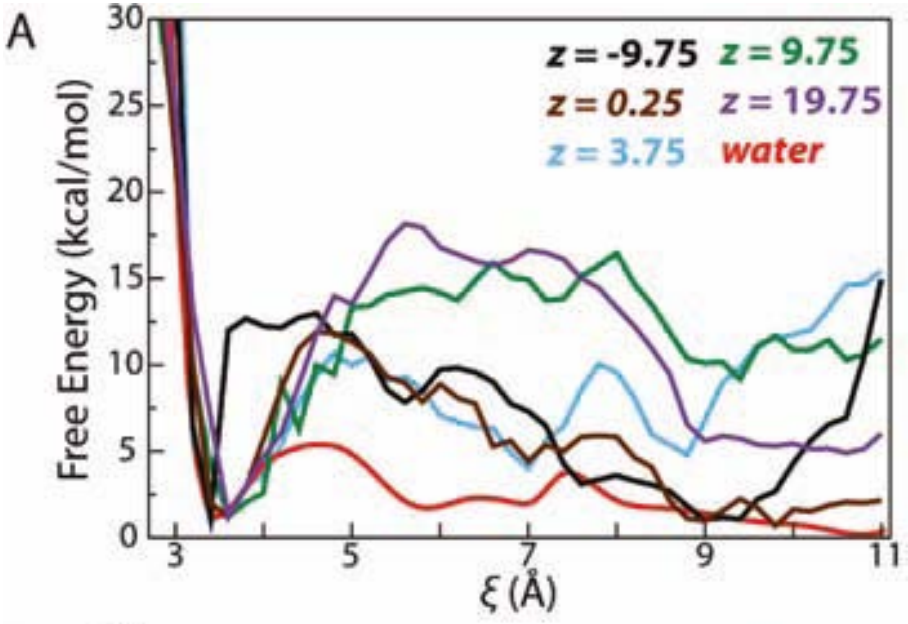


But what does it tell us?

Gumbart, Chipot, Schulten. (2011) *JACS*. **133**:7602-07.

Folding at different positions in the channel

1D projections

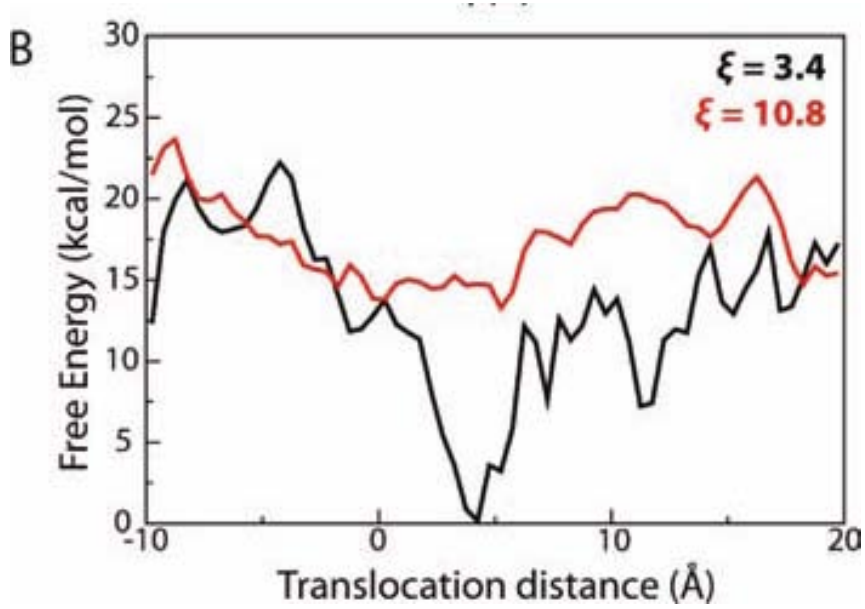


-barriers between folded/extended states larger in channel **(slow transitions)**

-states similar to those in water (as seen in experiments), but small helical bias appears

-helical state reaches minimum near channel center (also expected from experiments!)

-extended state unchanged throughout



Conclusion: SecY helps folding! (but slowly...)

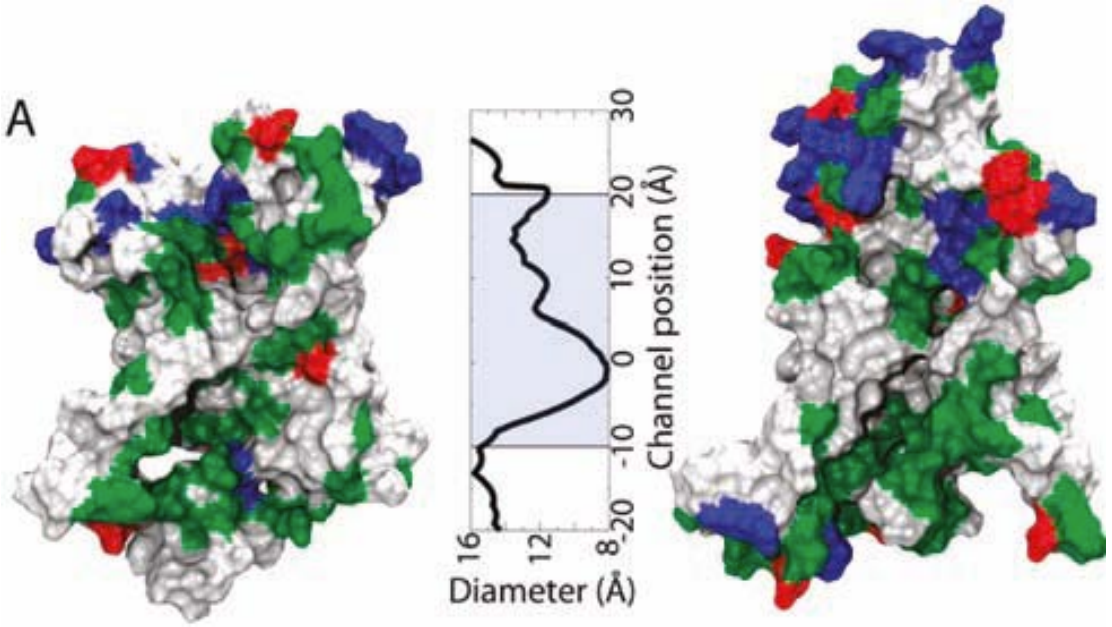
Gumbart, Chipot, Schulten. (2011) *JACS*. **133**:7602-07.

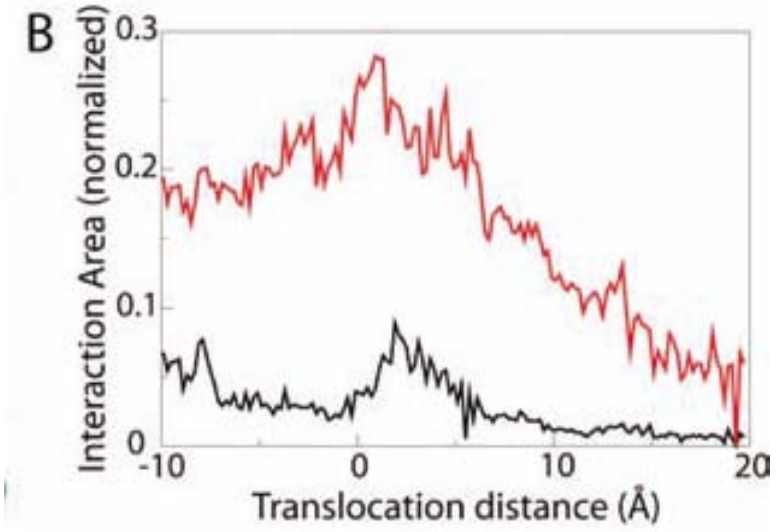
How does SecY participate in protein folding?

Through two simple physical principles

Mechanism 1: entropy

-narrowing of channel induces formation of compact states (well documented effect)



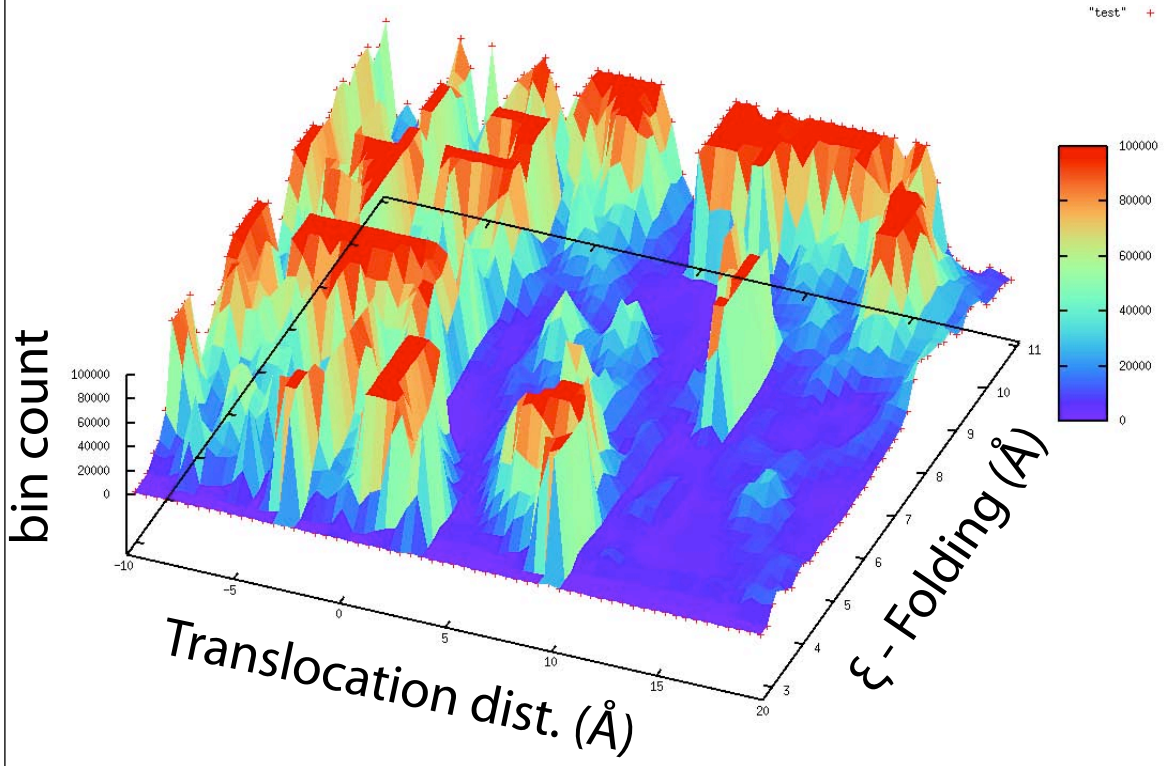


Mechanism 2: surface properties

-nascent protein interacts with lipids (black curve) along with other hydrophobic regions of SecY (red) near channel center

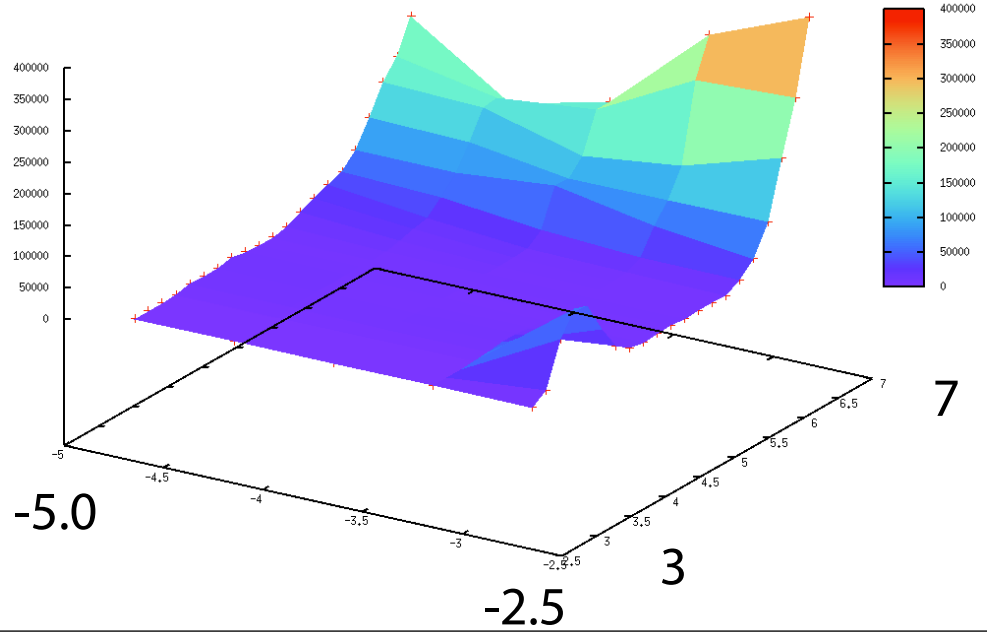
Gumbart, Chipot, Schulten. (2011) *JACS*. **133**:7602-07.

Problem: Sampling deficiencies

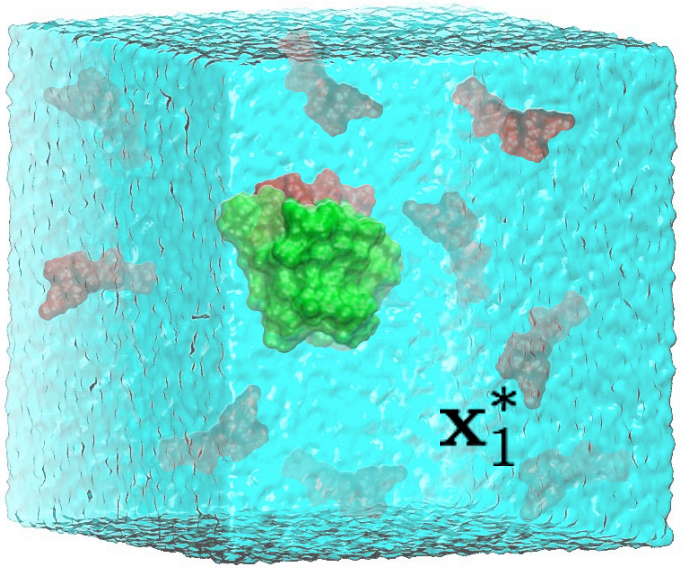


-counts/bin very non-uniform, even after dividing reaction coordinates into many windows

-increasing stratification required to overcome
-even with exceedingly small windows, samples pile up on one side



Challenge 2: Absolute binding free energies



$$K_{\mathrm{eq}} = \frac{[\mathrm{protein : ligand}]}{[\mathrm{protein}][\mathrm{ligand}]}$$

$$\mathrm{protein} + \mathrm{ligand} \overset{K_{\mathrm{eq}}}{\rightleftharpoons} \mathrm{protein : ligand}$$

$$K_{\mathrm{eq}} = \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\ \mathrm{e}^{-\beta U}}{\int_{\mathrm{bulk}} \mathrm{d1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta U}}$$

$$\Delta G^0 = -kT \ln(K_{\mathrm{eq}} C^\circ)$$

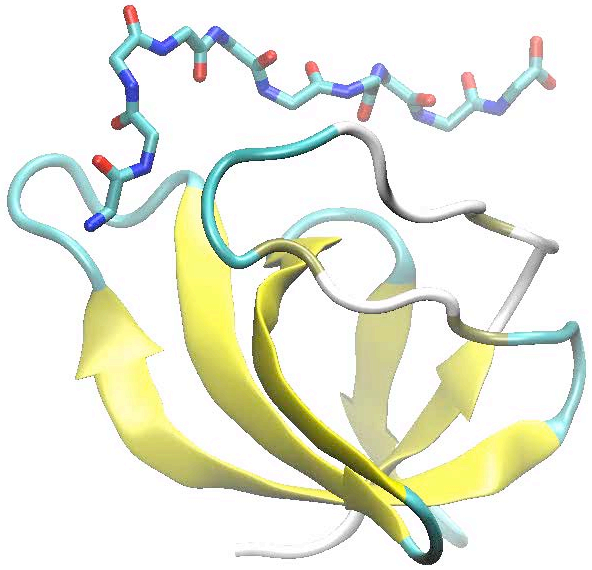
$$C^\circ = 1/1661 \text{Å}^3$$

Abl Src homology domain 3

Binder: APSYSPPPPP (flexible!)

ΔG^0 = **-7.94 kcal/mol (exp)**

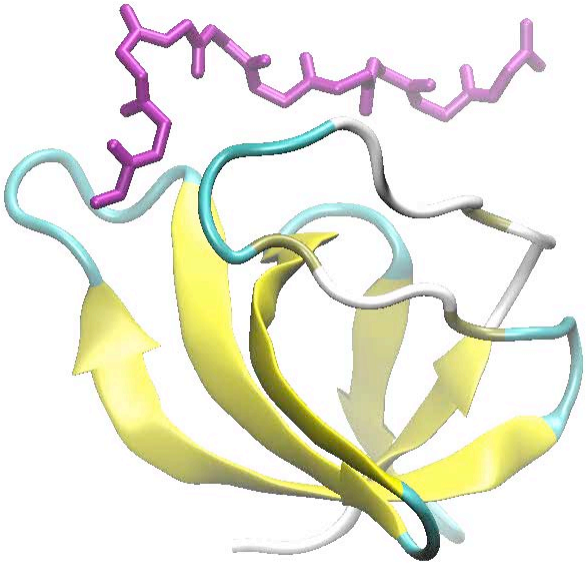
MM/PBSA estimate: **-2.6 kcal/mol** !



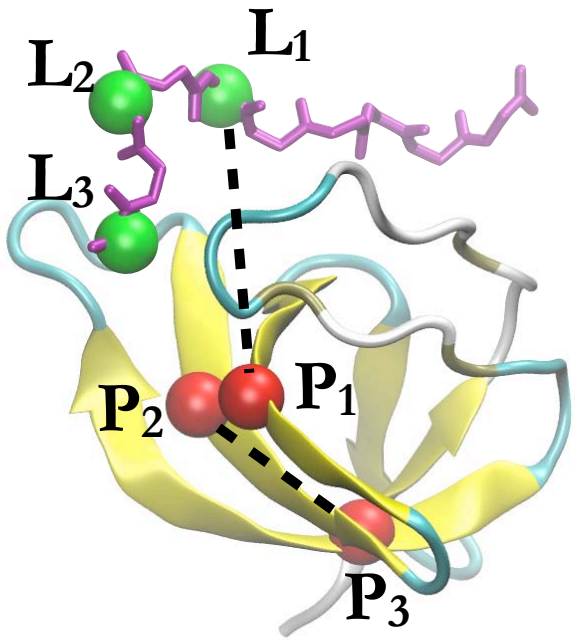
Pisabarro, M. T.; Serrano, L. *Biochemistry* **1996**, 35, 10634-10640
Hou, T. et al. *PLoS Comput. Biol*. **2006**, 2, 0046-0055

Inefficient sampling dominates

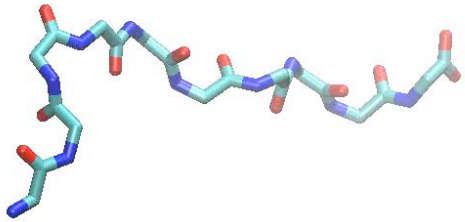
Overcoming sampling issues with restraints



Bound state RMSD restrained



Assorted spatial/rotational restraints

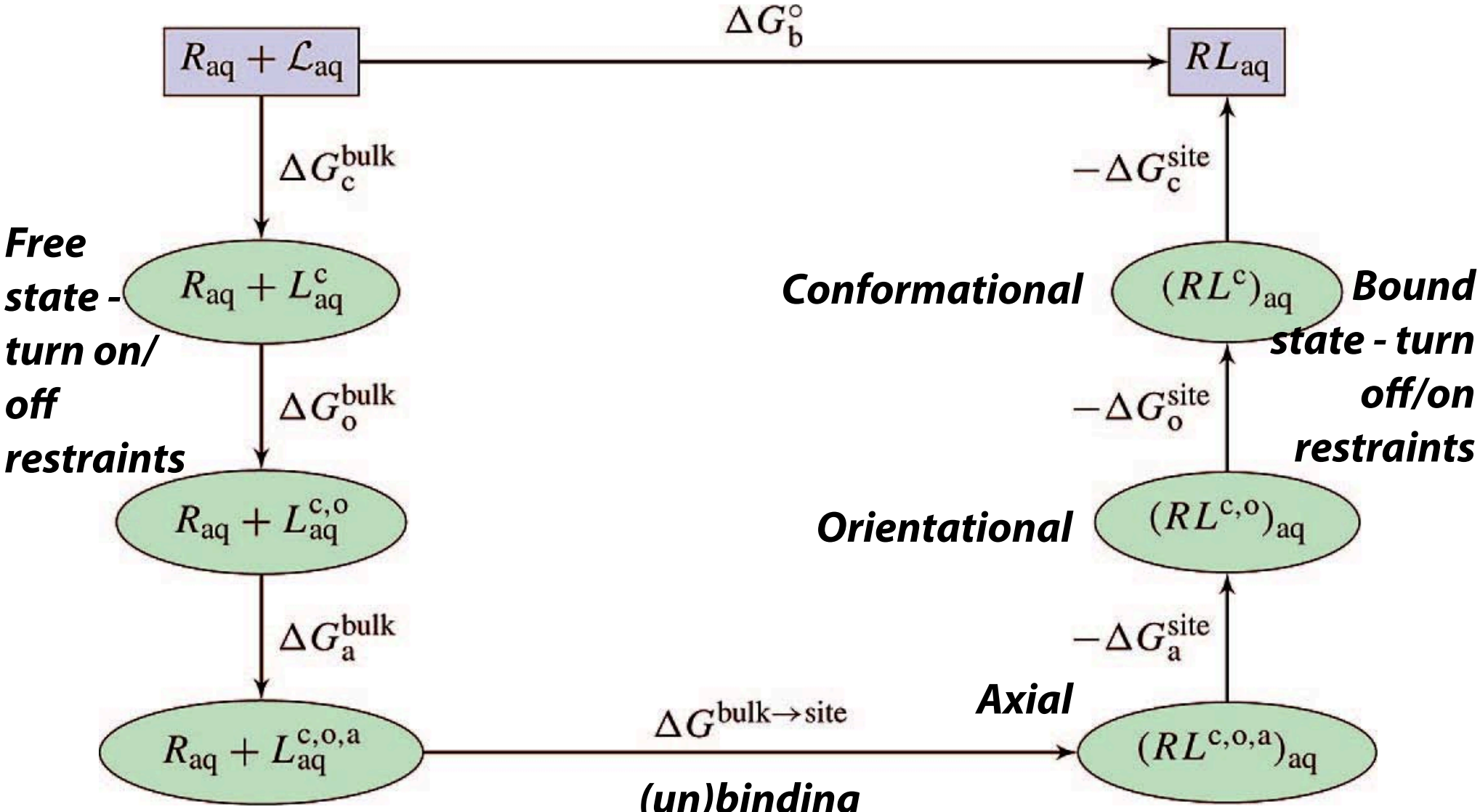


Free state RMSD restrained

-Design set of restraints to reduce conformational space needed to be sampled

-Contributions of each restraint to free energy need to be rigorously computed

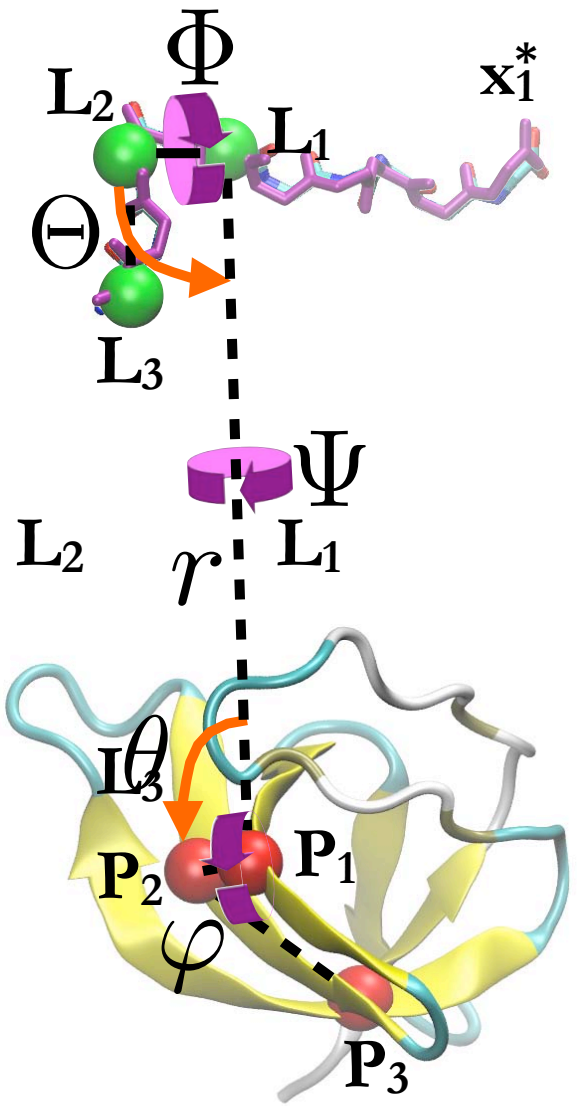
Overcoming sampling issues with restraints



Schematic of process

From: Deng and Roux. (2009)
J. Phys. Chem. **113**: 2234-2246.

Binding free energy from PMFs



$$K_{\mathrm{eq}} = \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta U}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c)}}$$

$$\times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}$$

$$\times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}$$

$$\times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi+u_\Psi)}}$$

$$\times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_o)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta)}}$$

$$\times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta)}}{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta+u_\varphi)}}$$

$$\times \frac{\int_{\mathrm{site}} \mathrm{d}\mathbf{1} \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_o+u_p)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_o)}}$$

$$\times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi+u_\Psi)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}$$

$$\times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}$$

$$\times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c)}}$$

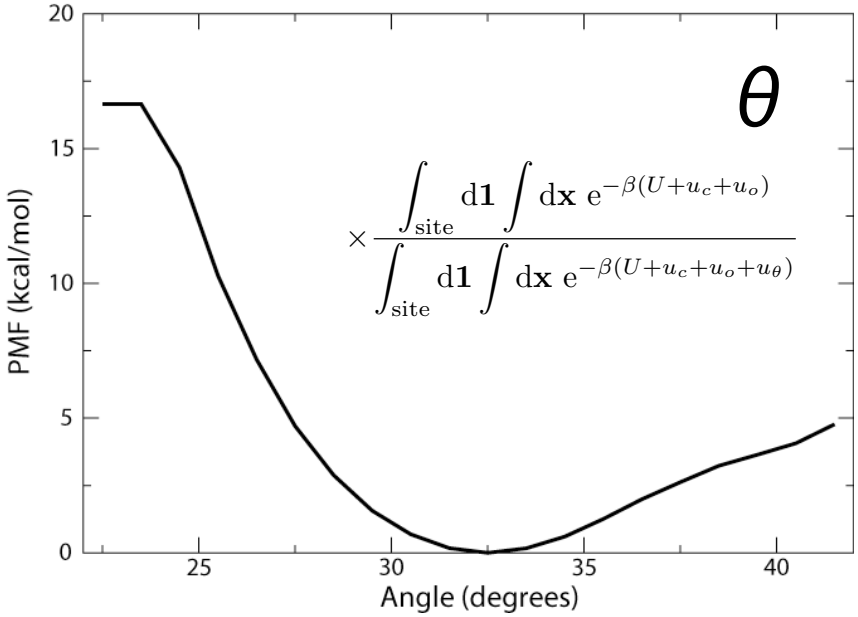
$$\times \frac{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta(U+u_c)}}{\int_{\mathrm{bulk}} \mathrm{d}\mathbf{1}\ \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\ \mathrm{e}^{-\beta U}}$$

Yu, Y. B. et al. (2001) *Biophys. J.*, **81**:1632-1642.

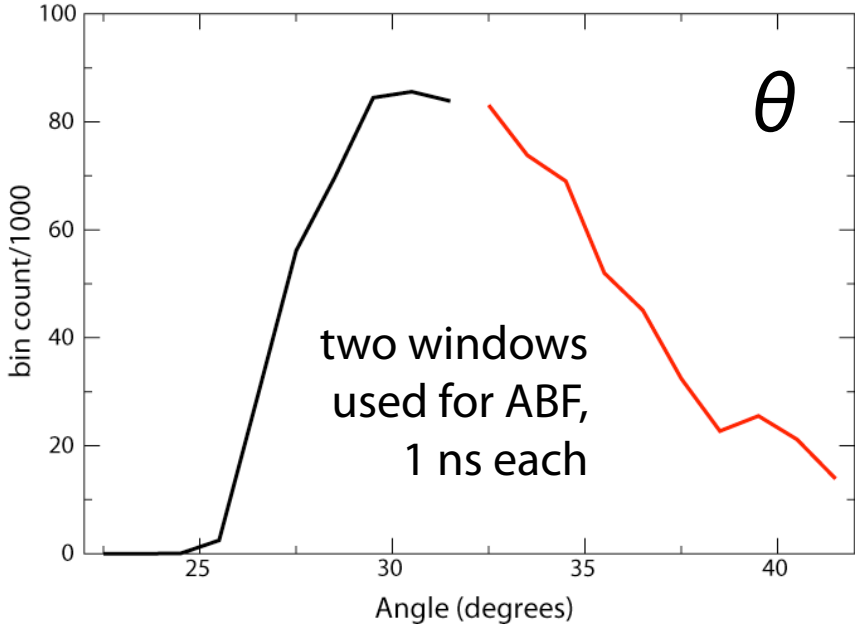
Woo, H. J.; Roux, B. (2005) *Proc. Natl. Acad. Sci. USA*, **102**:6825-6830.

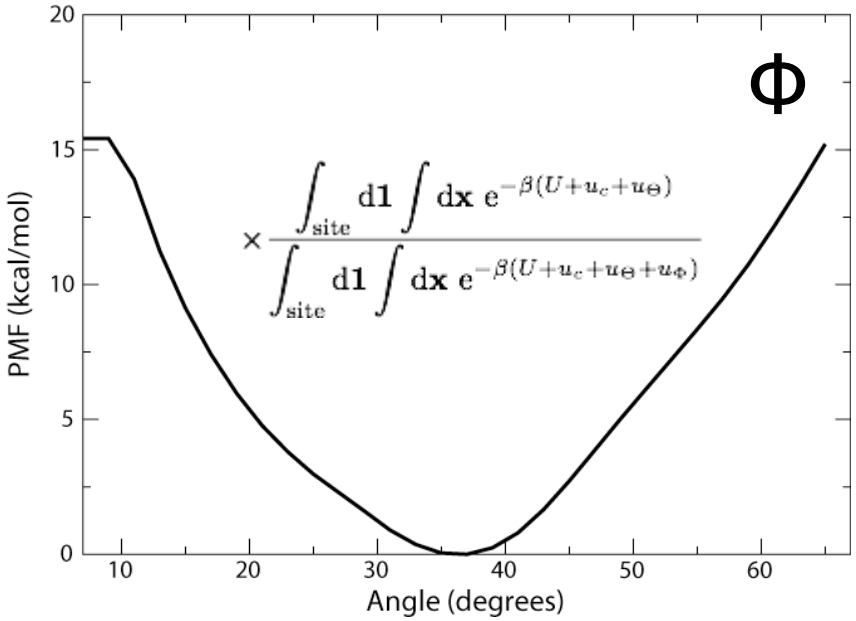
Maffeo, C., Luan, B., Aksimentiev, A. (2012) *Nucl. Acids Res.* **40**:3812-3821.

Some PMFs are simple to determine...

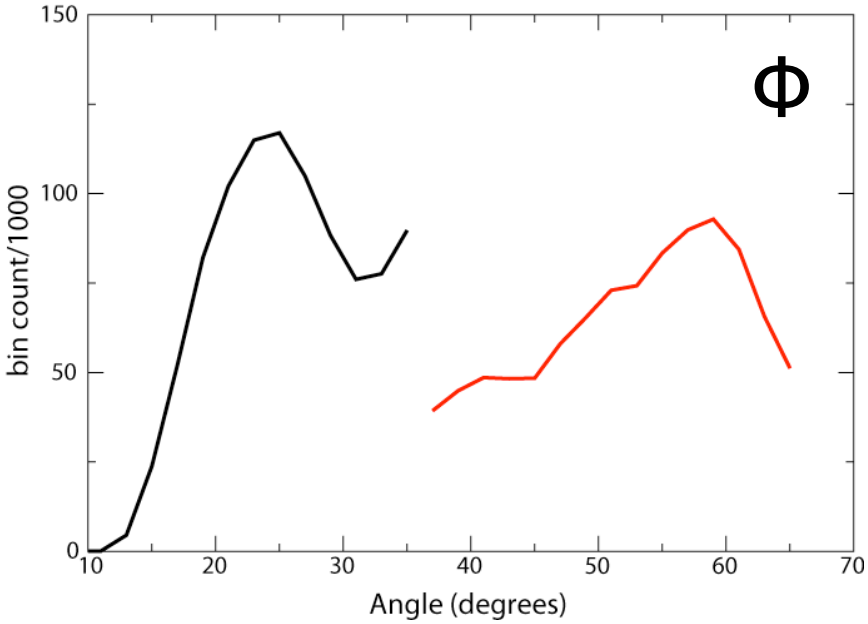


$$\times \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\ \mathrm{e}^{-\beta(U+u_c+u_o)}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\ \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta)}}$$

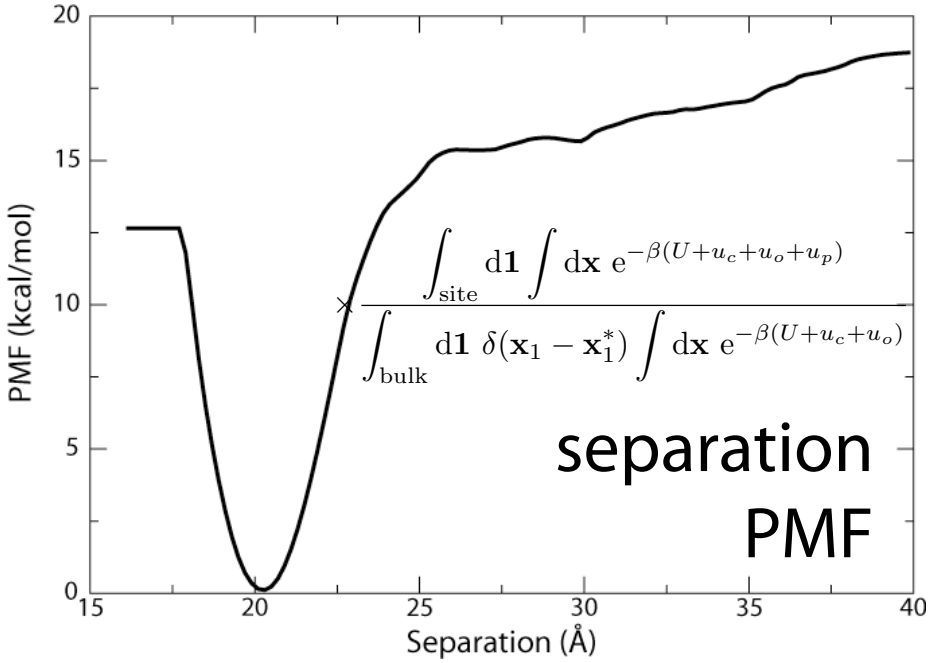


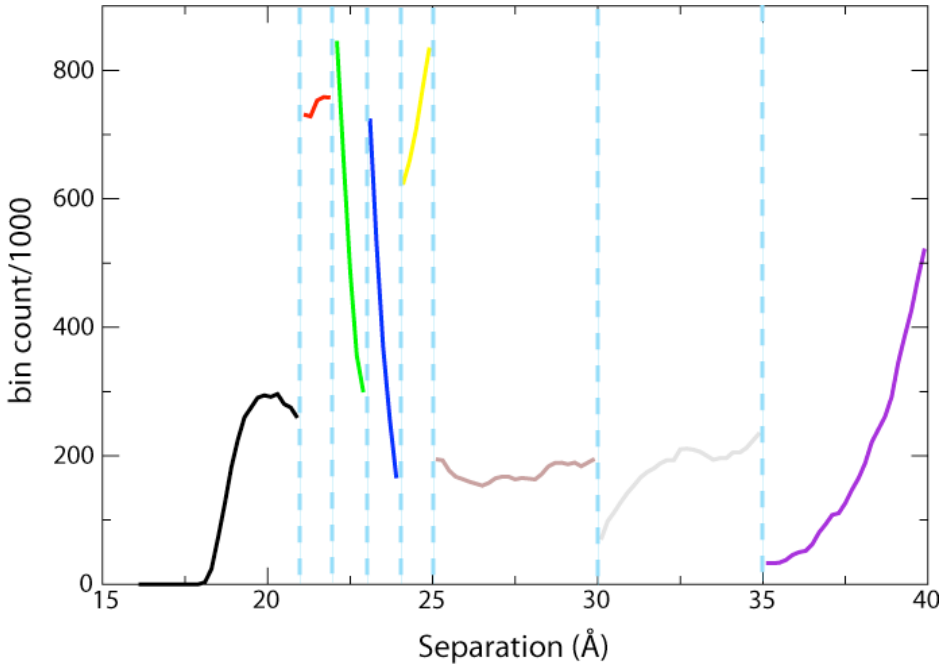


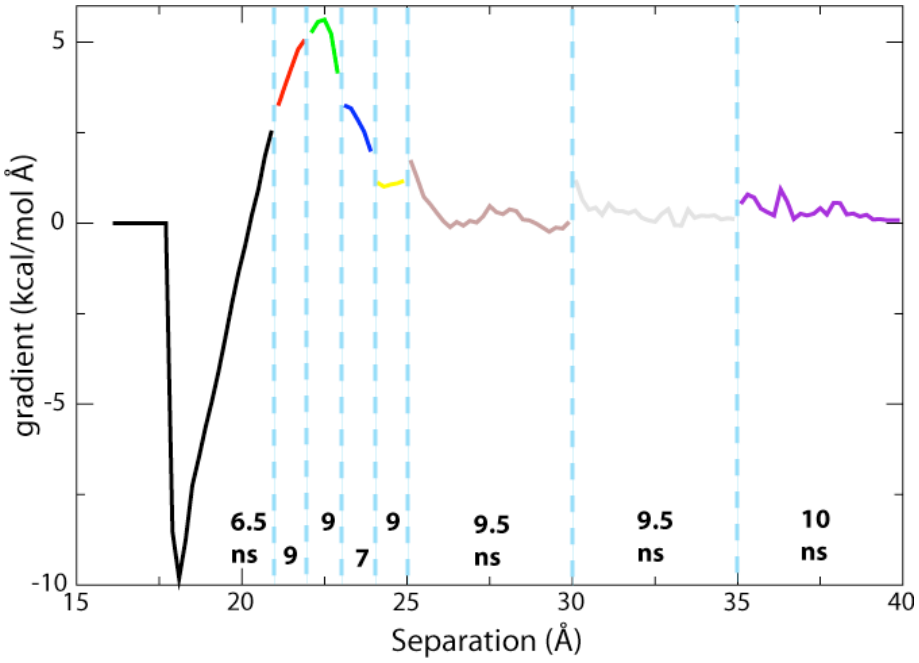
$$\times \frac{\int_{\text{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}{\int_{\text{site}} \mathrm{d}\mathbf{1} \int \mathrm{d}\mathbf{x}\ \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}$$



While others prove more difficult



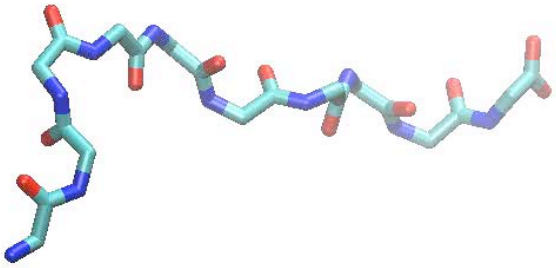


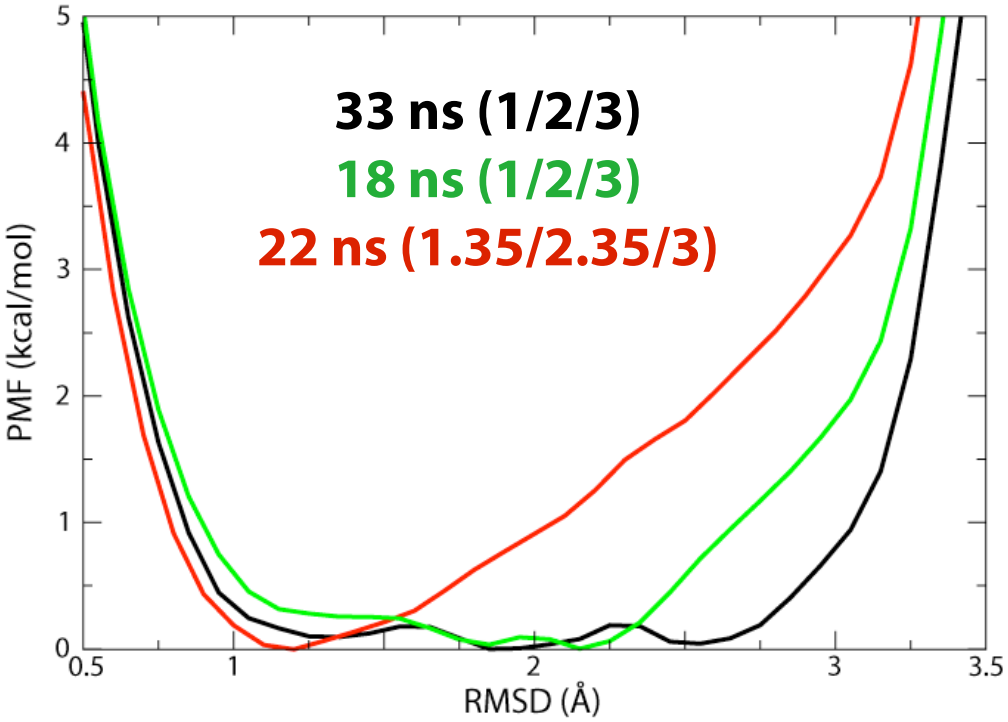


- even after dividing into 1-Å windows, sampling is still non-uniform

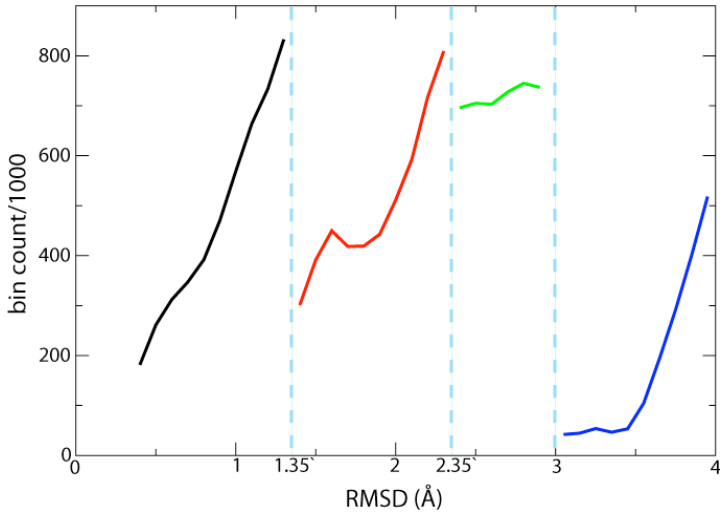
~70 ns total
gradients still not all aligned at boundaries

And yet others are even stranger

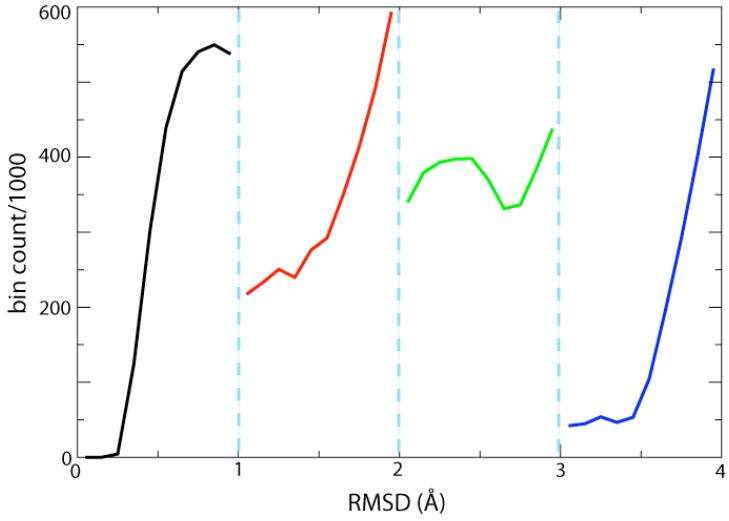




RMSD pmf for free state



counts for windows divided at 1.35, 2.35, and 3.0



counts for windows divided at 1.0, 2.0, and 3.0

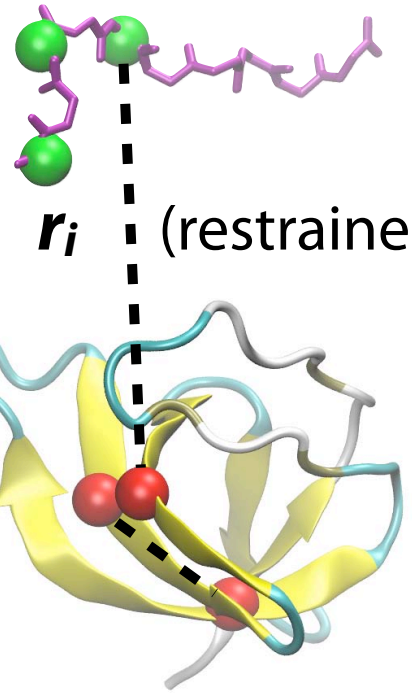
Profile is highly dependent on choice of window limits (artifact of bin size? 0.1 Å)

Potential of mean force from umbrella sampling

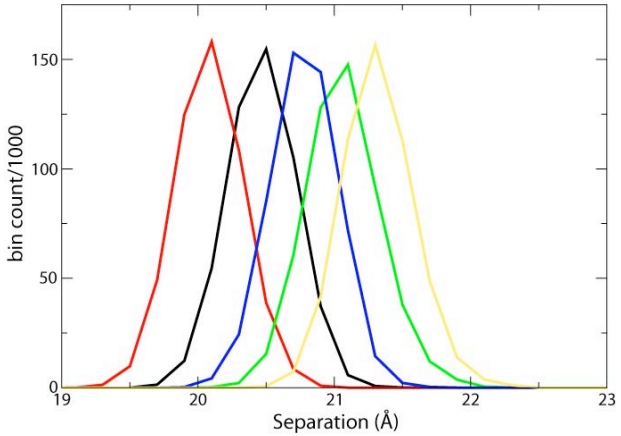
- the ultimate stratification

1) Apply restraining potential (bias) on reaction coordinate for a series of closely spaced windows

$$w_i(\xi) = \frac{1}{2}K(\xi - \xi_i)^2$$



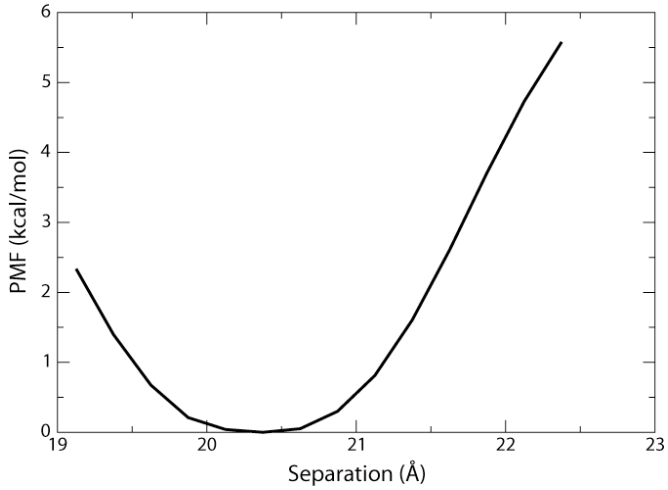
2) Track the fluctuations of the RC for each window, compute the histograms (i.e., the probabilities)



3) Combine the individual windows' PMFs and unbias using, e.g., WHAM (weighted histogram analysis method)

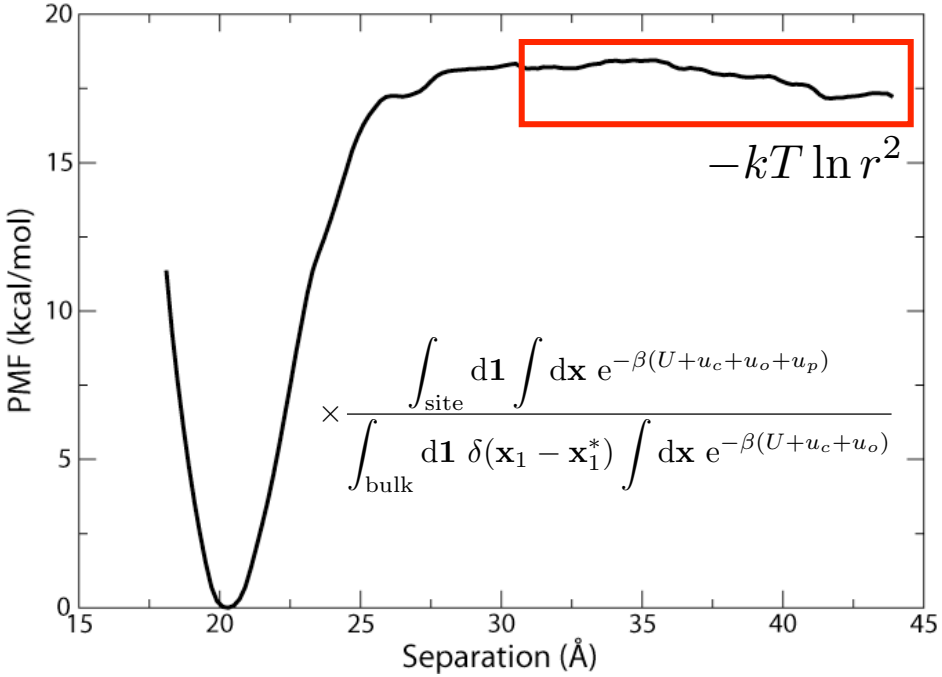
$$\mathcal{W}_i(\xi) = \mathcal{W}(\xi^*) - k_b T \ln[\frac{\langle \rho(\xi) \rangle_{(i)}}{\langle \rho(\xi^*) \rangle}] - w_i(\xi) + F_i$$

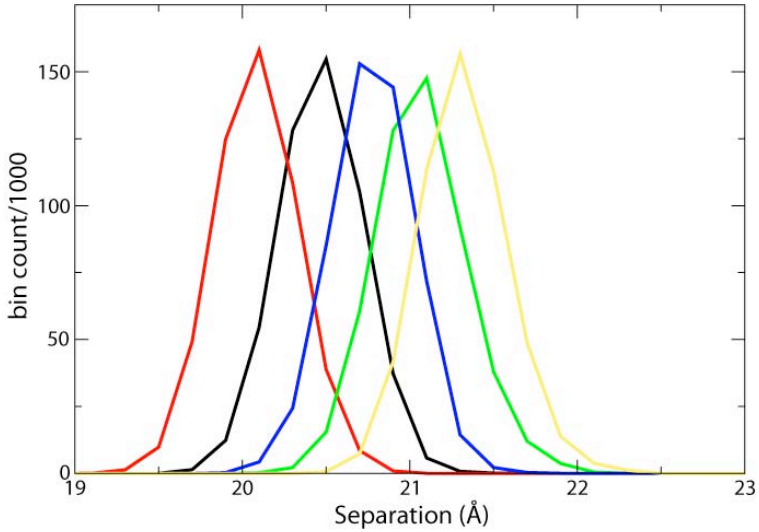
unbiased PMF for a single window[1]



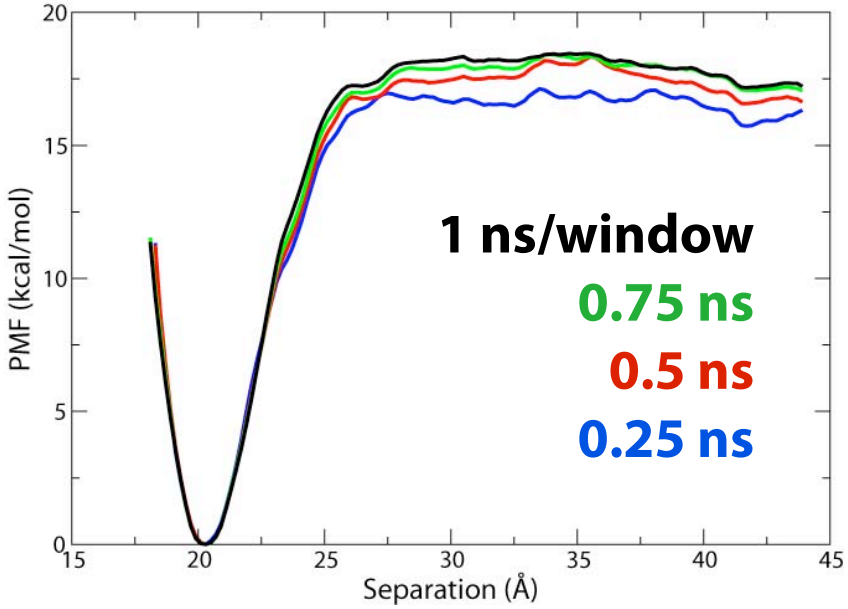
1: Roux, B. (1995). *Comp. Phys. Comm.* **91**: 275-282.

Separation PMF from umbrella sampling





37 windows used, spaced
0.5 - 1 Å apart
-histograms are overlapping



PMF was already converged within ~20 ns (compare to **70 ns** for ABF!)

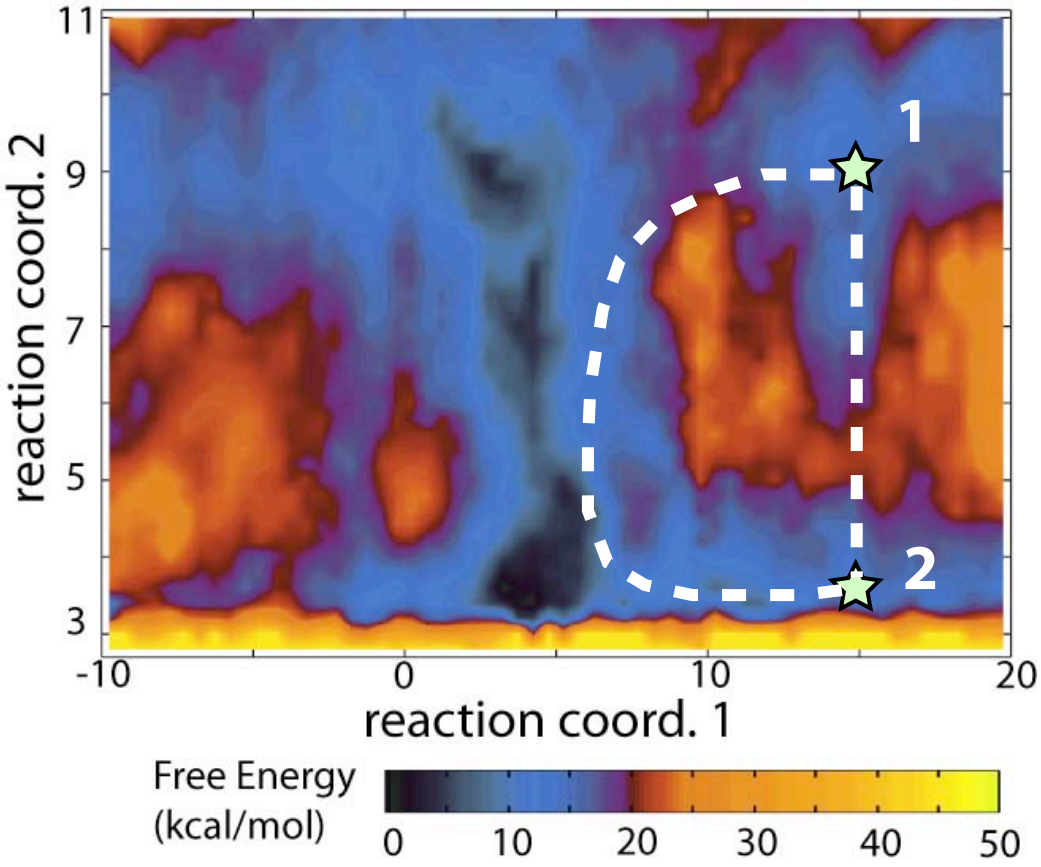
Limitations of umbrella sampling (US)

-common complaint about US: poor sampling of orthogonal degrees of freedom - system cannot evolve naturally

Example:

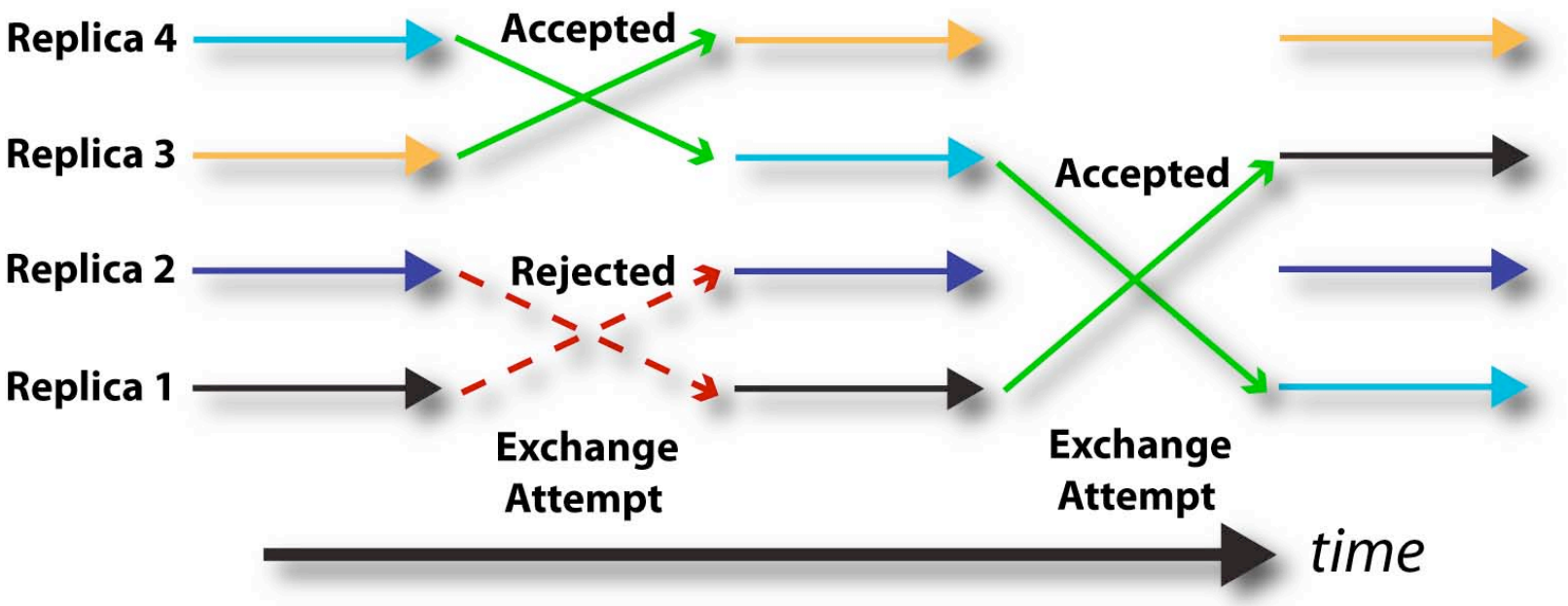
-if RC1 is restrained at 15 using US, a large barrier prevents it from fully sampling the orthogonal RC2

-however, if RC1 were free to diffuse, the system could take an alternate, lower energy path to reach state 2 from state 1



Note that ABF suffers from a similar problem: slow relaxation in degrees of freedom orthogonal to the RC prohibit full sampling - RC gets "stuck"

Replica-exchange umbrella sampling



-helps to circumvent sampling limitations by exchanging coordinates periodically between different windows

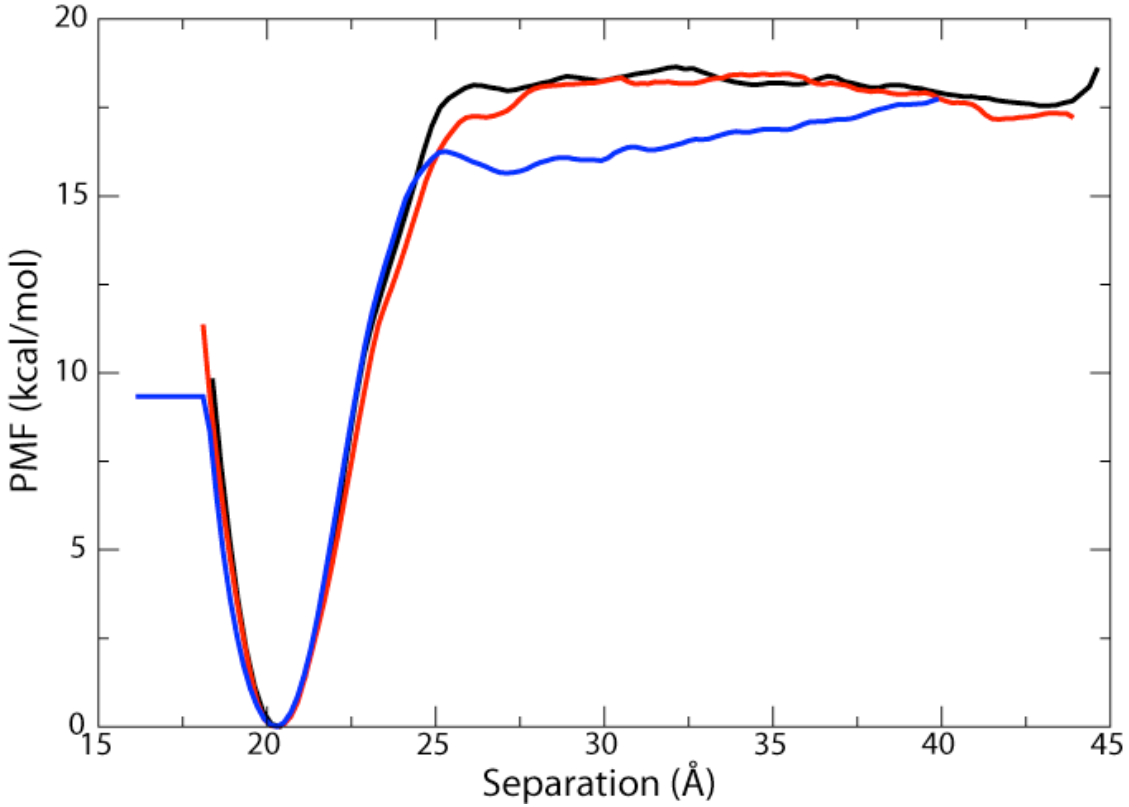
-exchanges accepted with some probability: $\min(1, e^{-\Delta E/kT})$

where $\Delta E = (w_i(\xi_j) - w_i(\xi_i)) + (w_j(\xi_i) - w_j(\xi_j))$

(swapped) (original) (swapped) (original)

Implemented in NAMD 2.9 for colvars

Replica-exchange umbrella sampling



REMD-US - 37 ns
US - 37 ns
ABF - 70 ns

-for this problem, **REMD-US** does not converge notably faster than standard **umbrella sampling**

-however, both fare significantly better than **ABF**

Back to the Abl kinase story...

$$K_{\mathrm{eq}} = \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta U}}{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c)}} \times \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c)}}{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta)}} \times \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}} \times \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi+u_\Psi)}} \times \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_o)}}{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta)}} \times \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta)}}{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_\theta+u_\varphi)}} \times \frac{\int_{\mathrm{site}} \mathrm{d1} \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_o+u_p)}}{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_o)}} \times \frac{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi+u_\Psi)}}{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}} \times \frac{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta+u_\Phi)}}{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta)}} \times \frac{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c+u_\Theta)}}{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c)}} \times \frac{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta(U+u_c)}}{\int_{\mathrm{bulk}} \mathrm{d1}\, \delta(\mathbf{x}_1 - \mathbf{x}_1^*) \int \mathrm{dx}\, \mathrm{e}^{-\beta U}}$$

~30 ns	$\Delta G_c^{\mathrm{site}} = 3.52$ kcal/mol
6 ns	$\Delta G_o^{\mathrm{site}} = 0.71$ kcal/mol
4 ns	$\Delta G_a^{\mathrm{site}} = 0.20$ kcal/mol
37 ns	$\Delta G_r^{\mathrm{sep}} = -14.47$ kcal/mol
(analytical)	$\Delta G_o^{\mathrm{bulk}} = 5.77$ kcal/mol
30 ns	$\Delta G_c^{\mathrm{bulk}} = 5.43$ kcal/mol

$$\Delta G^o = (\Delta G_c^{\mathrm{bulk}} - \Delta G_c^{\mathrm{site}}) + (\Delta G_o^{\mathrm{bulk}} - \Delta G_o^{\mathrm{site}})$$

$$+\Delta G_r^{\mathrm{sep}} - \Delta G_a^{\mathrm{site}} = \boxed{-7.7 \text{ kcal/mol}}$$

~ 100 ns required

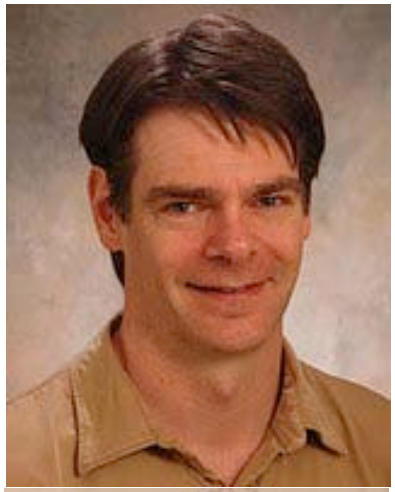
ΔG^0 = -7.94 kcal/mol (**exp**) **Agreement within 0.25 kcal/mol!**

final thought: Is (plain) ABF impractical for certain classes of problems?

Possible resolutions	*Potential issues*
Biasing the intransigent degrees of freedom and integrating over them after the fact	Requires knowing *a priori* what degrees to target
multiple-walker ABF Minoukadeh et al. (2010) *J. Chem. Theory Comput.* **6**:1008–1017.	Still diffusion limited - sampling distribution not under user control
Orthogonal space random walk Zheng, Chen, Yang. (2008) *PNAS.* **105**:20227-32.	???
New artificial bias to induce more distributed sampling (at least initially)	A Maxwell demon in disguise?

Acknowledgements

Chris Chipot,
Nancy U.



Benoit Roux,
U. Chicago

Klaus Schulten, UIUC

XSEDE
Extreme Science and Engineering Discovery Environment

NATIONAL INSTITUTES OF HEALTH

Argonne NATIONAL LABORATORY

NSF