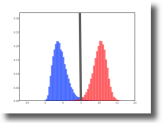
# Validation efforts and efficiency improvements in free energy calculations

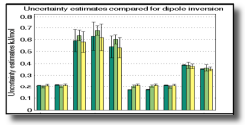
Michael R. Shirts
Department of Chemical Engineering
University of Virginia

CECAM: Free energy calculations: From theory to applications
Paris, France
June 7, 2012

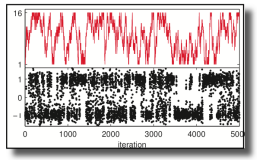
## One important project of the Shirts group: Making free energy calculations easier

More efficient analysis of samples collected from simulations





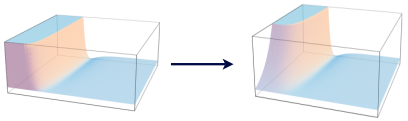

Benchmarking and validation of molecular simulation methods



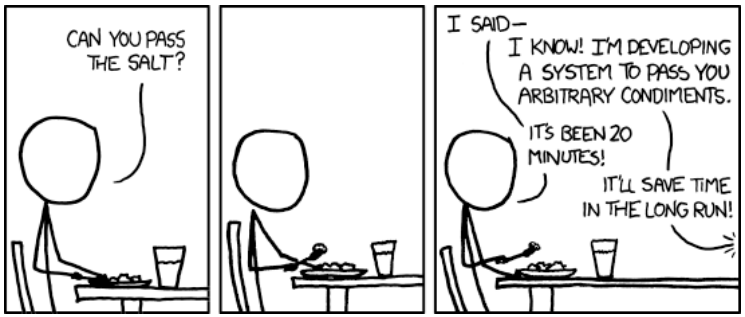
More efficient collection of samples from simulations


Ease of use of simulations

More efficient alchemical transformations



## It is hard and important to make molecular simulations easier and more rigorous




I find that when someone's taking time to do something right in the present, they're a perfectionist with no ability to prioritize, whereas when someone took time to do something right in the past, they're a master artisan of great foresight.

## How do we extract thermodynamic properties from simulations?


$$\langle X \rangle = \sum_i P_i X_i$$

$$\langle X \rangle = \sum_i e^{-\beta E_i} X_i$$

$$\langle X \rangle = \frac{1}{N} \sum_{i=1}^{N} X_i$$

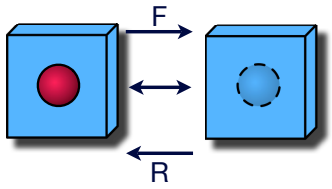
Sampled from the Boltzmann probability distribution

## The central idea: Statistical mechanics

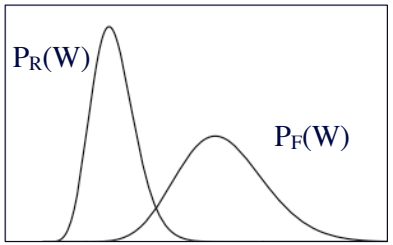
- We are sampling from a parametric distribution
- Possibly statisticians might have something to say about this problem . . . ?
- If they did, would we understand them?

$$\Delta A = \int_0^1 \left\langle \frac{dU}{d\lambda} \right\rangle d\lambda \qquad \lambda = \int_0^1 E_\theta[U(w,\theta)]d\theta$$

## How can statistics help? An example




- Sample problem: free energy of solvation

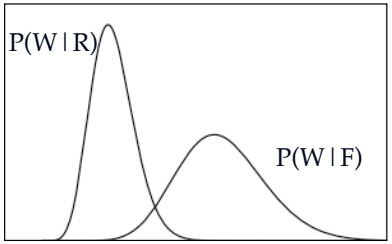


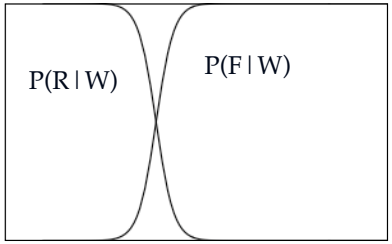

$$\frac{P_F(W)}{P_R(-W)} = e^{\beta(F-W)}$$

Logistic Regression: Given characteristics (F), what is the most likely response (W)?

Reverse Logistic regression: Given a response (W), what are the most likely characteristics (F)?

## Turns free energy calculations into a maximum likelihood problem




$$P(F|W_i) = \frac{1}{1+\exp(-\beta(M+W_i-\Delta F))} \qquad P(R|W_i) = \frac{1}{1+\exp(\beta(M+W_i-\Delta F))}$$

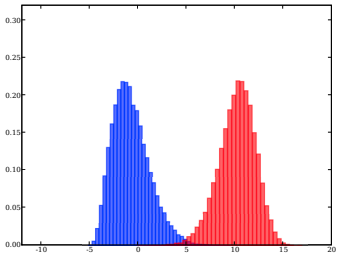
$$L(\Delta F) = \prod_{i=1}^{n_F} P(F|W_i) \prod_{j=1}^{n_R} P(R|W_j)$$

Maximize, setting derivative equal to zero

Error bounds by propagation of derivatives

## BAR is minimum variance free energy estimate of data extracted from two states

$$\left\langle \frac{1}{1+\frac{N_B}{N_A}e^{\Delta E(\mathbf{x})-\Delta F}} \right\rangle_A = \left\langle \frac{1}{1+\frac{N_A}{N_B}e^{\Delta F-\Delta E(\mathbf{x})}} \right\rangle_B$$



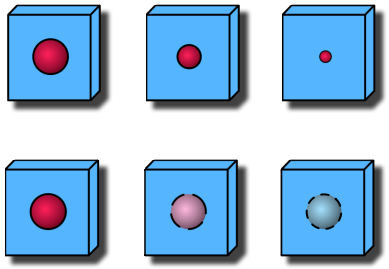

M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande. *Phys. Rev. Lett.* **119**:5740 (2003)

C. H. Bennett, *J. Comp. Phys*, **22**:245 (1976)

M. R. Shirts and V. S. Pande. *J. Chem. Phys*. **122**:144107 (2005)

## Frequently, multiple states are needed

$$\Delta F_{1\to N} = -\beta^{-1}\ln\frac{Z_N}{Z_1} = -\beta^{-1}\ln\frac{Z_2}{Z_1}\cdot\frac{Z_3}{Z_2}\cdots\frac{Z_N}{Z_{N-1}} = \sum_{n=1}^{N-1}\Delta F_{n\to n+1}$$



Shrink the size

Attenuate the interactions

If I run N simulations, can I optimally use data from all N to compute statistics?

## Maximum likelihood results: Multistate Bennett Acceptance Ratio

$$e^{-\beta F_i} = \frac{1}{N}\sum_{n=1}^{N}\frac{e^{-\beta E_i(x_n)}}{\sum_{k=1}^{K}\frac{N_k}{N}e^{\beta F_k - \beta E_k(x_n)}}$$

M. R. Shirts and J.D. Chodera, *J. Chem. Phys.* 129, 124105 (2008)

Vardi (1985), Gill et al. (1988), Kong et al. (2003), Tan (2004)

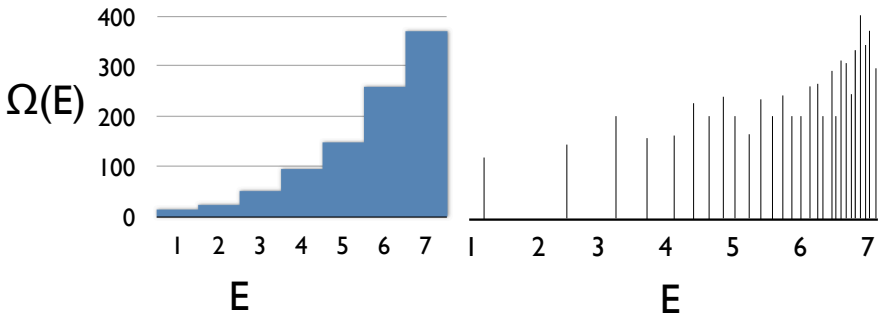
$$W_{nk} = \frac{P_k(\boldsymbol{x}_n)}{\sum_{k'=1}^{K} N_{k'} P_k(\boldsymbol{x}_n)}$$

$$\boldsymbol{\Theta} = [(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1} - \mathbf{N}]^{-1}$$

$$\delta\Delta F_{ij} = \beta^{-1}(\Theta_{ii} + \Theta_{jj} - 2\Theta_{ij})^{\frac{1}{2}}$$

- Reduce to BAR for two states
- In large N limit:
  - Error is normally distributed
  - Estimator is unbiased
- Provably the lowest variance reweighting estimator

## Histograms and MBAR




- Density of states Ω(U)
- Used to compute expectations
- Example: $\langle A\rangle = \sum A(E)\Omega(E)e^{-\beta E}$

- In MBAR, density of states is a sum of weighted delta functions
- $\Omega(E) = \sum_i w_i\delta(E-E_i)$
- Not good for visualizing, good for computing numbers
- Optimal weights $w_i$ from MBAR

## Yields lots of tricks to calculate other statistical quantities

$$\left.\begin{aligned} q_i(x) &= e^{-\beta E_i(x)} \\ q_j(x) &= e^{-\beta E_j(x)} \end{aligned}\right\} \longrightarrow \Delta F = \beta^{-1}\ln\frac{\int q_i(x)dx}{\int q_j(x)dx}$$

$$\left.\begin{aligned} q_k(x) &= e^{-\beta E_k(x)} \\ q_A(x) &= A(x)e^{-\beta E_k(x)} \end{aligned}\right\} \longrightarrow \langle A\rangle_k = \frac{\int A(x)e^{-\beta E_k(x)}dx}{\int e^{-\beta E_k(x)}dx}$$

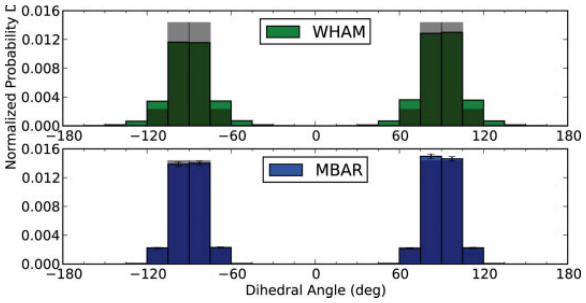
$$\delta\langle A\rangle_k = |\langle A\rangle_k|(\Theta_{AA} + \Theta_{kk} - 2\Theta_{Ak})^{1/2}$$

- Formulas also valid if $N_k$=0 (states at which no samples are taken)
- Can sample from K states, reweight at arbitrary other states

# MBAR has improved properties to WHAM

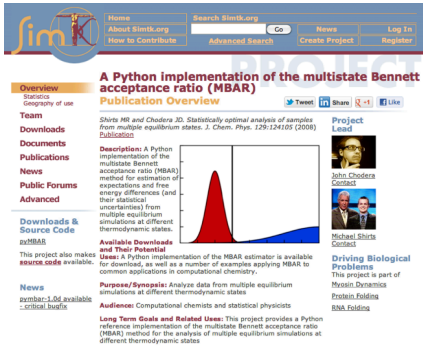

But you don't have to take my word for it!

Propane Dihedral Angle by Umbrella Sampling Replica Exchange

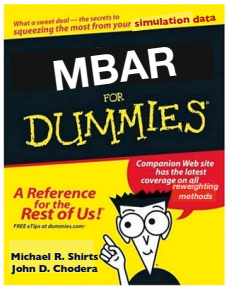



M. Fajer, R. V. Swift, J. A. McCammon *J. Comp. Chem,* 30, 719-1725 (2009)

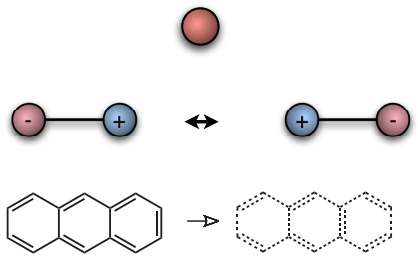
# pymbar 2.0




- Much more efficient: 1-3 orders of magnitude faster
- More examples
- Automated setup.py
- https://simtk.org/home/pymbar



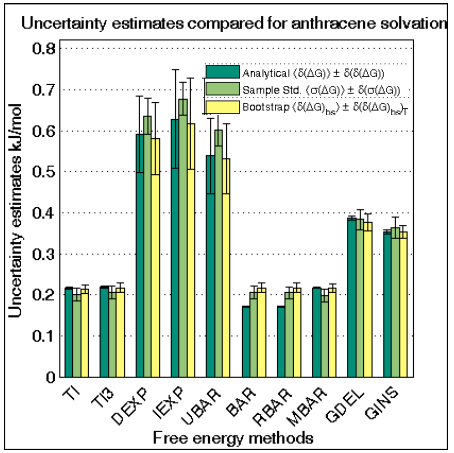

# We need benchmarks and standards for statistical calculations



- Paliwal and Shirts, *J. Chem. Theory Comput*, **7**, 4115 (2011)
- 100 starting configurations
- Input files and reference energies for Gromacs, Amber, Desmond

# Validating the error estimates of free energy methods

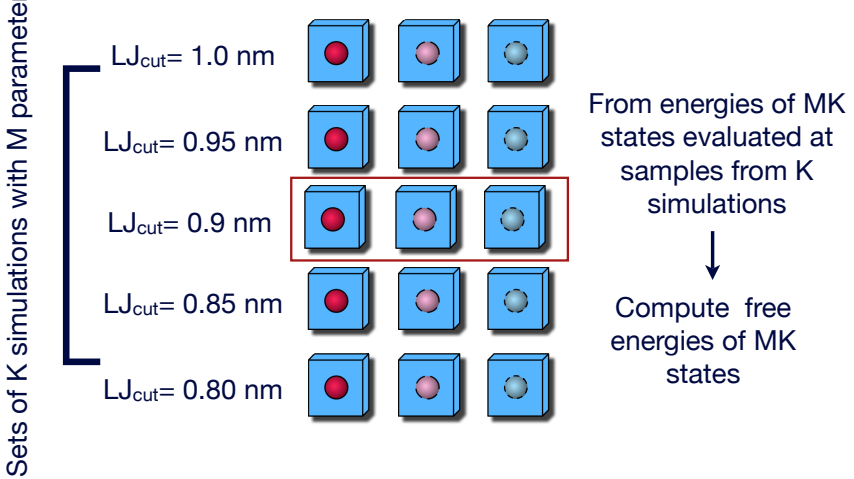
- 10 different free energy methods
- Repeat calculations 100 times
- Compare analytical uncertainties with actual sample variance
- Also use bootstrap error estimates



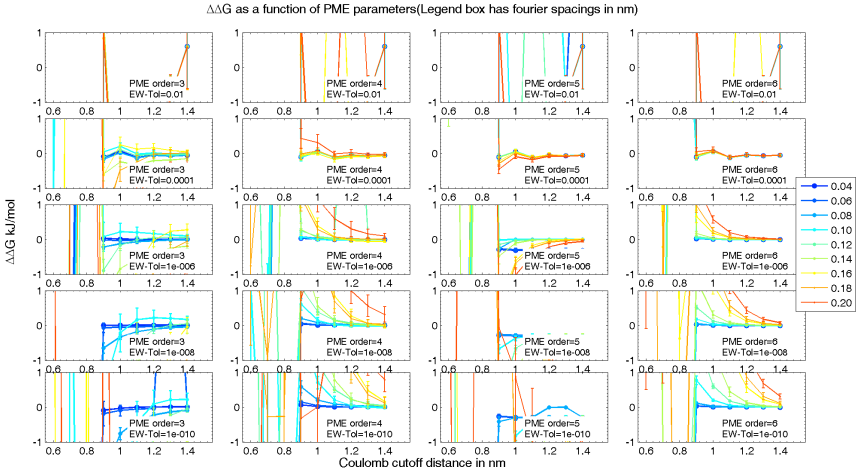
# Developing Benchmarks

- Effort started at Boston free energy application meeting (contact me if interested in partipating!)
- Data for users:
  - 40-100 starting coordinates
  - Files for multiple simulation programs
- Systems
  - T4 Lysozyme
  - FKBP
  - Other systems (Trypsin, DNA gyrase)
- Collect data from users
  - Binding data from users
  - Method details and time taken

# Using MBAR to validating simulation parameters for free energies




# Can use data from one parameter set to predict ΔΔG from converged parameters




# Allows us to identify fast simulation parameter choices with negligible difference from converged parameters

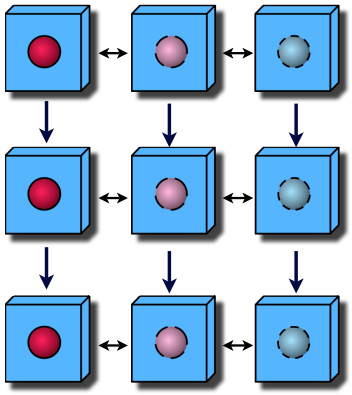
| $\Delta\Delta G$ for methane solvation (kJ/mol) | | | | | |
|---|---|---|---|---|---|
| $\Delta G_B - \Delta G_E$ | -0.110±0.021 | -0.213±0.089 | -0.202±0.013 | N.A | N.A |
| $\Delta G_B - \Delta G_O$ | -0.107±0.020 | -0.330±0.089 | N.A | -0.154±0.013 | N.A |
| $\Delta G_E - \Delta G_O$ | 0.003±0.004 | -0.117±0.096 | N.A | N.A | 0.007 ±0.002 |

Prediction using only benchmark set

Tested results

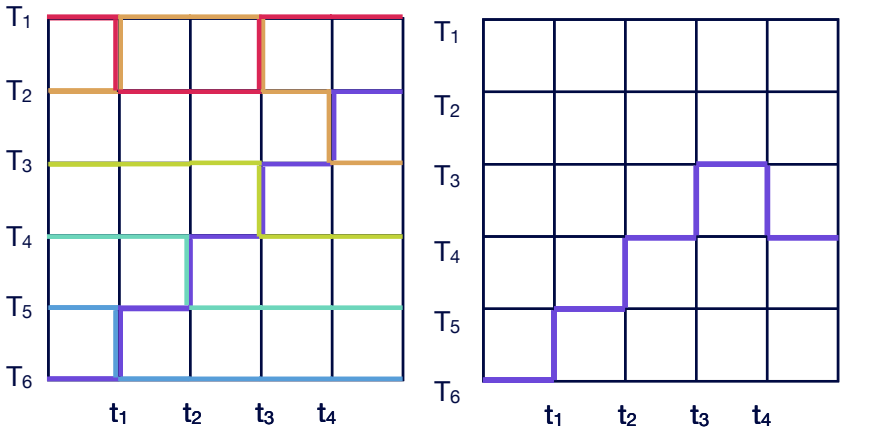
| $\Delta\Delta G$ for anthracene solvation with a vdw cutoff 1.3 nm (kJ/mol) | | | | | |
|---|---|---|---|---|---|
| $\Delta G_B - \Delta G_E$ | 0.168±0.585 | 0.314±0.155 | -2.035±0.021 | N.A | N.A |
| $\Delta G_B - \Delta G_O$ | 0.169±0.585 | 0.465±0.156 | N.A | -1.978±0.021 | N.A |
| $\Delta G_E - \Delta G_O$ | 0.001±0.005 | 0.151±0.170 | N.A | N.A | -0.002±0.0004 |

# Can also perform simulations in multiple chemical states



Improves sampling by swapping between intermediate states

# Replica exchange and expanded ensemble: two ways to use multiple states to improve sampling



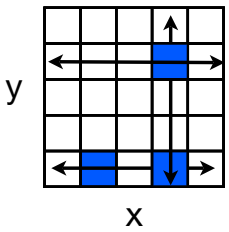

$\min\left[1, \exp\left(-(\beta_1 E_1 + \beta_2 E_2) + (\beta_2 E_1 + \beta_1 E_2)\right)\right]$ $\min\left[1, \exp\left(-\beta_1 E_1 + g_1 + \beta_2 E_1 - g_2\right)\right]$

# Gibbs sampling: a statistical technique for sampling multidimensional spaces

- Geman and Geman, 1984
  - 6,226 citations as of May 2012
- Given: a joint probability distribution P(x,y)
- Task: Generate samples from P(x,y)

- Algorithm

1. Start with a sample ($x_i$,$y_i$)
2. Sample from P(x|$y_i$) to obtain $x_{i+1}$
3. Sample from P(y|$x_{i+1}$) to obtain $y_{i+1}$
4. Rinse and repeat 2 and 3



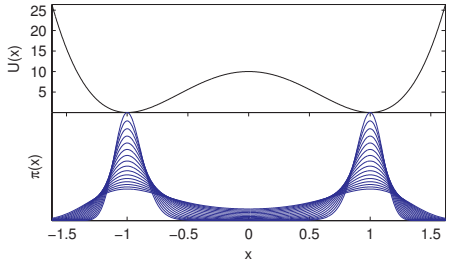

# Gibbs sampling to make rejectionless Monte Carlo for expanded ensemble

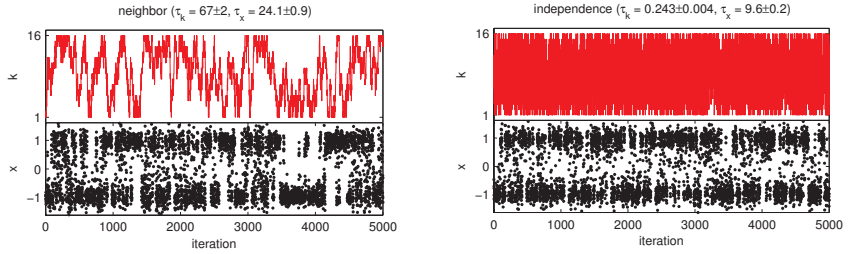
- Assume states $y_k$, k=1,N
- States have unnormalized conditional probability Q($y_i$|x)
- Normalize probabilities: P($y_i$|x) = Q($y_i$|x) / $\sum_k$ Q($y_k$|x)
- Propose new state j with probability P($y_j$|x)
- Accept with probability 1
- *Independence sampling*

- Features of sampling directly with the conditional distribution
  - ☑ Allows for nonlocal moves
  - ☑ No rejected moves
  - ☐ But requires enumeration of the conditional probabilities

Chodera and Shirts, *J. Chem. Phys*,135, 194110 (2011)

# Large speedup in state sampling, significant speedup in coordinate sampling



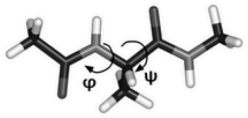




# A better strategy to improve replica exchange in state space

- Any set of moves in space of permutation of states preserving P(x,y) is valid:
- SO:
  - Perform multiple random swaps before continuing with coordinate sampling
  - OR: Perform multiple neighbors swaps before continuing with coordinate sampling
  - Need N log N to randomize for equal energy states
  - $10^3$-$10^5$ random swaps seems to work pretty well

# Replica exchange with alanine dipeptide

- OBC GBSA implicit solvent, Amber ff99sb
- 2000 iterations, sample from temperature distribution every 1 ps
- 20 temperatures, 270 K to 450 K
- Code: OpenMM on NCSA Lincoln GPU's
  - Approximately 80x speedup over single core AMBER
- Compare 2 types of swaps
  - One sweep of neighbor swaps
  - Random pairwise $N^3$ = 8000 times

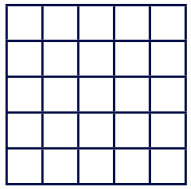


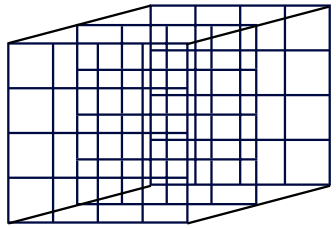

| | $(1-\lambda_2)^{-1}$ | $\tau_{ACF}$ | $\tau_{\sin(\phi)}$ | $\tau_{\sin(\psi)}$ |
|---|---|---|---|---|
| Neighbor swaps | 92±1 ps | 80±2 ps | 110±9 ps | 66±6 ps |
| All swaps | 2.62±0.01 ps | 1.60±0.06 ps | 8.7±0.4 ps | 9.1±0.5 ps |

Theoretical 36x speedup over Metropolis

Actual 7-12x speedup over Metropolis

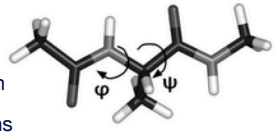
# Gibbs sampling in multidimensional parameter spaces





- What about swapping Hamiltonians *and* temperatures?
- What about restraint and alchemical Hamiltonians?
- Swapping neighbors becomes *complicated*
- Can perform Gibbs sampling over all states, or over each topologically compact region

# 2D replica exchange: alanine dipeptide with umbrella sampling



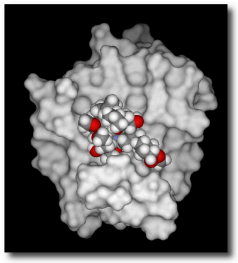

- 10 restraints in each direction in (ϕ,ψ) space
- 100 umbrella simulations total + one unbiased simulation
- Hamiltonian exchange between the restraint Hamiltonians
- 2000 iterations (swaps every 5 ps), $N^3 = 10^6$ swaps in λ

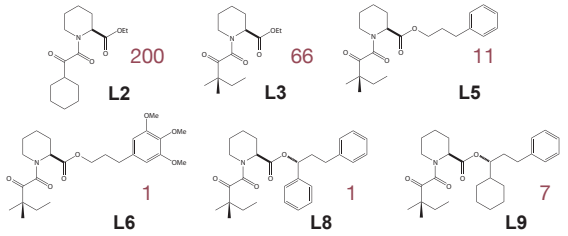
| | $(1-\lambda_2)^{-1}$ | $\tau_{ACF}$ | $\tau_{\sin(\phi)}$ | $\tau_{\sin(\psi)}$ |
|---|---|---|---|---|
| Neighbor swaps | 82±4 ps | 31±1 ps | 57±2 ps | 27.1±1 ps |
| All swaps | 24.2±0.3 ps | 5.5±0.1 ps | 9.9±0.1 ps | 6.1±0.1 ps |

Theoretical 3.4x speedup over Metropolis

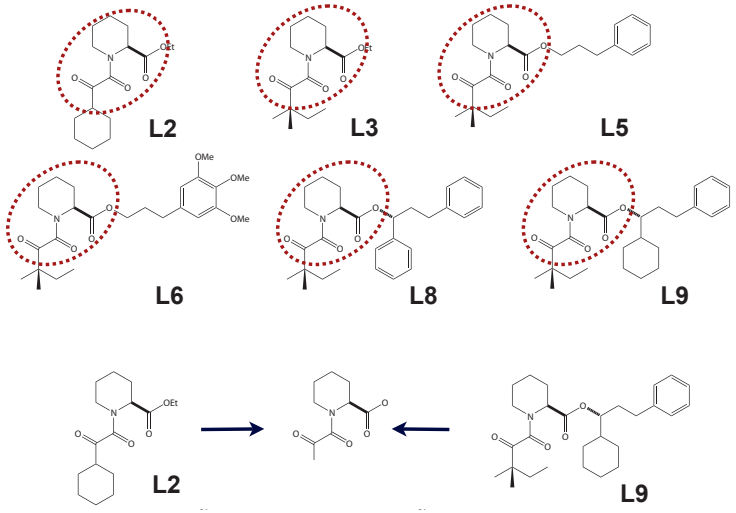
Actual 4-5 x speedup over Metropolis

# Expanded ensembles for

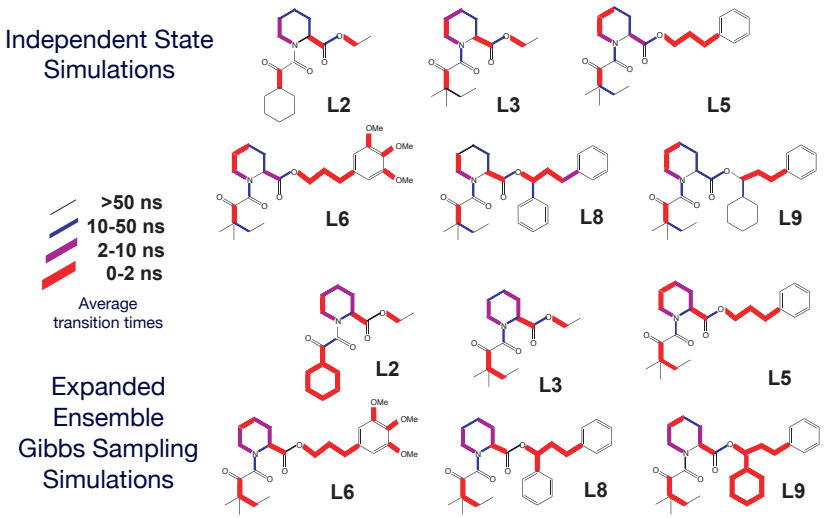





# "Mutate" ligands to common




# Ligand binding sampling accelerated



# How do we know if we're sampling from the correct distribution?

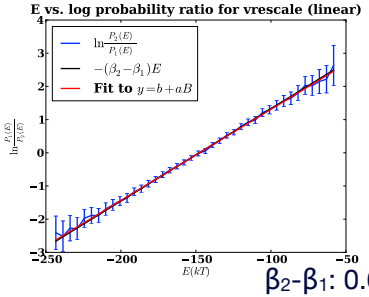
Run the same system, same options, but two different temperatures

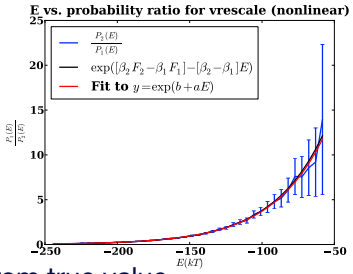
$$P_1(E) = Q_1^{-1}\Omega(E)e^{-\beta_1 E}$$
$$P_2(E) = Q_2^{-1}\Omega(E)e^{-\beta_2 E}$$

$$\frac{P_1(E)}{P_2(E)} = \frac{Q_2}{Q_1}e^{(\beta_2-\beta_1)E}$$

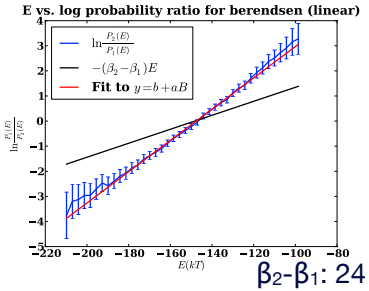
$$\ln\frac{P_1(E)}{P_2(E)} = \ln\frac{Q_2}{Q_1} + (\beta_2-\beta_1)E$$

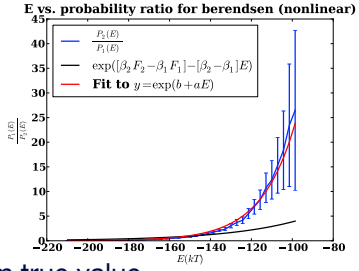
# Quantitative visualization of ensembles




β₂-β₁: 0.66 σ from true value

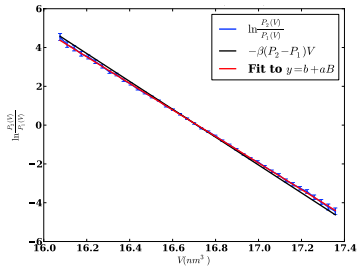


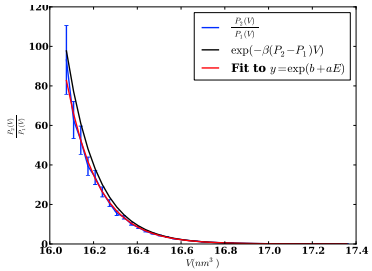




β₂-β₁: 24 σ from true value

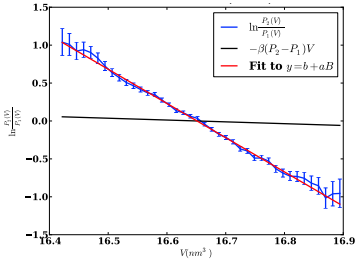
# Validation of Volume Fluctuations in NPT

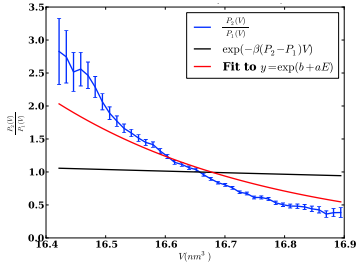
Parrinello-Rahman




Berendsen

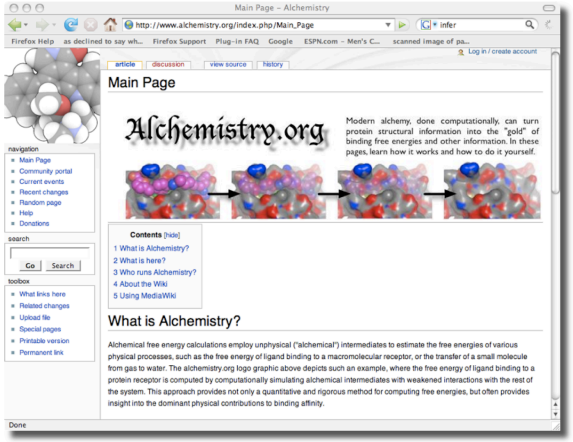



# Other improvements

- Can separate kinetic and potential energies
  - Means can use for MC algorithms as well
- NPT simulations
  - Can look at distribution of E + PV
  - Can look at distribution of V alone
  - Can look at joint distribution of E and V
- Grand canonical simulations (fluctuating N)
- Python implementation
  - https://simtk.org/home/checkensemble

## Towards best practices?




- Collaborative project for improved simulation setup
- Consortium gradually forming for simulation validation and benchmarking

## Take home messages

- Statisticians are frequently inspired by physical scientists: we really need to check back to see where they extended and fixed what we do.
- Maximum likelihood methods are powerful ways to calculate unknown parameters from statistical samples
- Statistical validation of simulation methods is required!
- Gibbs sampling provides an efficient way to get fast sampling in p(x,y) by sampling (in an independent or correlated manner) from $p(y|x_i)$ and $p(x|y_i)$
- Passing arbitrary condiments rigorously is a good idea

## Thanks!


- Y'all
- Free energy analysis: John Chodera (Berkeley), Eric Bair (UNC), Giles Hooker (Cornell), Vijay Pande (Stanford)
- Free energy validation: Himanshu Paliwal (UVa)
- Improved Sampling: John Chodera (Berkeley)
- Money:
  - NIH NRSA
  - Ralph E. Powe Jr. Faculty Enhancement Award
  - University of Virginia FEST fund
  - NSF CHE-1152786