CE 530 Molecular Simulation

Lecture 19

Free-energy calculations: Distribution functions, precision and accuracy

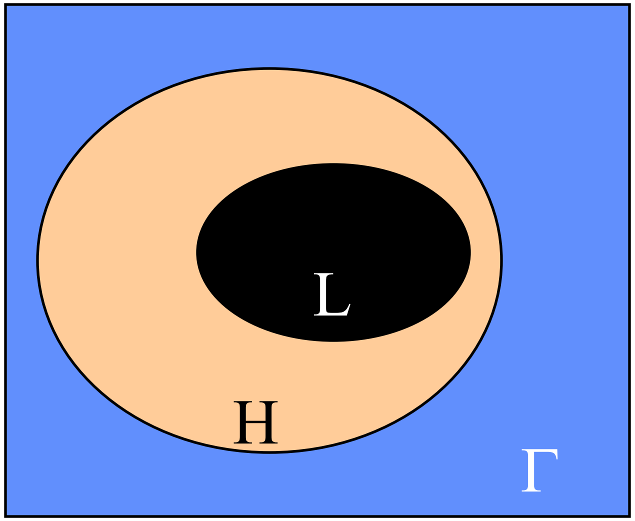
David A. Kofke
Department of Chemical Engineering
SUNY Buffalo
kofke@eng.buffalo.edu

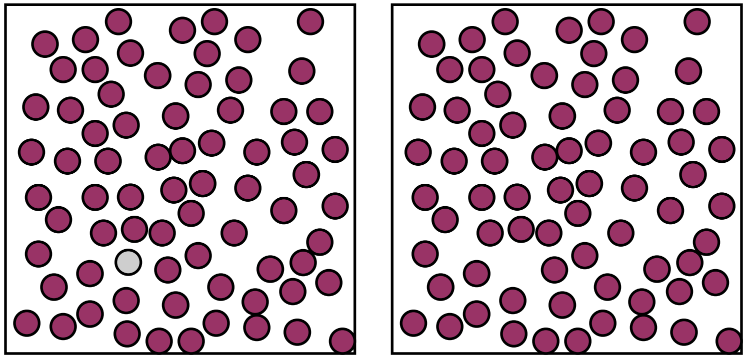
Review

- All useful free-energy methods compute free-energy differences
- Several approaches have been developed
- FEP gives free-energy difference via an ensemble average
  - *Asymmetric*
    - Deletion method is awful
- Four approaches to basic multistaging
  - *Umbrella sampling, Bennett's method, staged insertion/deletion*
- Thermodynamic integration uses $dA/d\lambda = \langle dU/d\lambda \rangle$
  - *Symmetric*
- Parameter hopping treats perturbation variable as an extension of phase space
  - *time spent at different values relates to their free-energy difference*

Free-Energy Perturbation

- Each stage of a FEP method should be used only for "encompassing systems"
  - *important configurations of one system form a subset of the configurations that are important to the other*
  - *staging may be needed to bring this about*
- The superset will have a higher entropy
  - *use "H" and "L" to distinguish high- and low-entropy systems*
- Remember hard-sphere with test particle
  - *Every configuration of non-overlap is important to the N+1 particle (L) system*
  - *But every one of these configurations is of uniform importance in the N-particle (H) system*





Distribution Functions

- The FEP average can be cast as a simple one-dimensional integral

$$e^{-\beta(A_L - A_H)} = \frac{\Lambda^{3N}}{Q_H}\int d\mathbf{r}^N e^{-\beta(U_L - U_H)} e^{-\beta U_H}$$

$$= \int du\, e^{-\beta u} \int d\mathbf{r}^N \delta\left(u - (U_L - U_H)\right) \frac{\Lambda^{3N} e^{-\beta U_H}}{Q_H}$$

$$\boxed{e^{-\beta \Delta A} = \int du\, e^{-\beta u} p_H(u)}$$

reference is the high-entropy system

- Likewise

$$\boxed{e^{+\beta \Delta A} = \int du\, e^{+\beta u} p_L(u)}$$

reference is the low-entropy system

- Energy distributions

$$p_H(u) = \left\langle \delta(u - \Delta U) \right\rangle_H$$

$$p_L(u) = \left\langle \delta(u - \Delta U) \right\rangle_L$$

normalized

$$\int_{-\infty}^{\infty} p(u)\, du = 1$$

Nomenclature

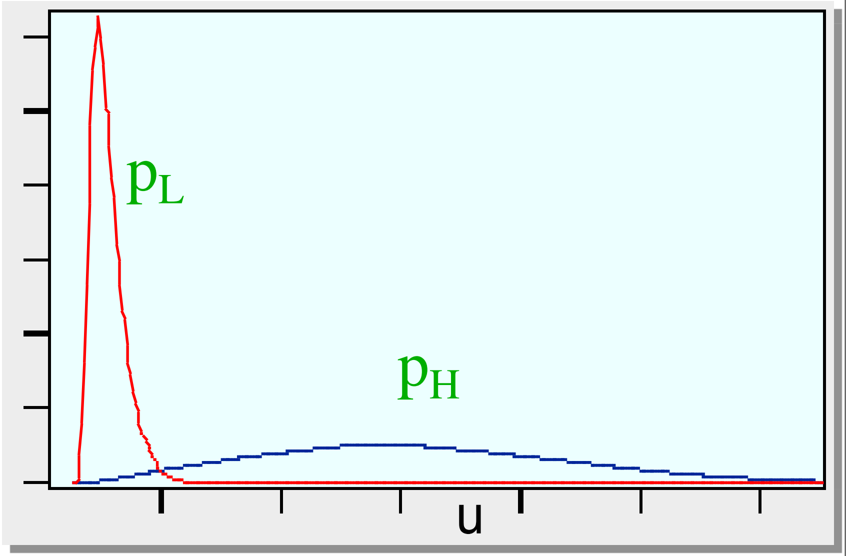
Free-energy difference: $\Delta A \equiv A_L - A_H$

Entropy difference: $\Delta S \equiv S_L - S_H < 0$

Energy difference: $u \equiv U_L(\mathbf{r}^{(N)}) - U_H(\mathbf{r}^{(N)})$

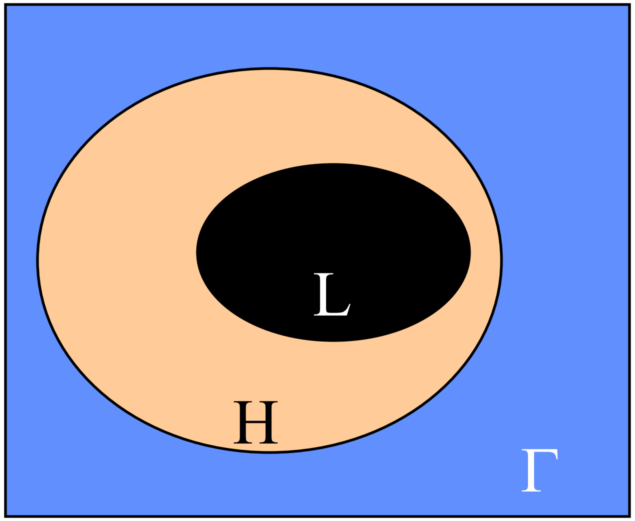
Interpretation

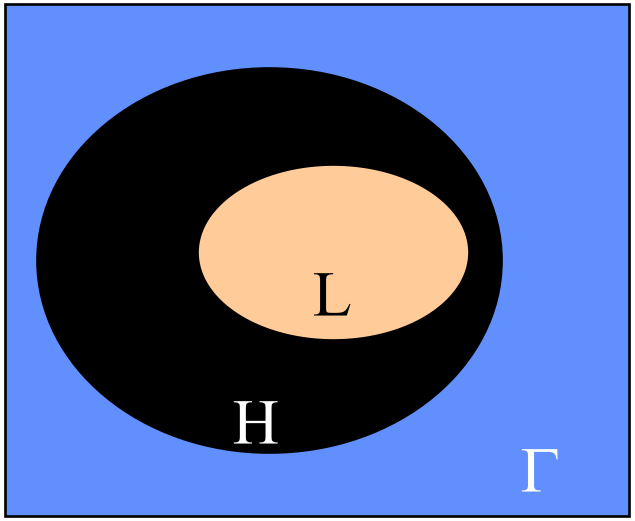
- Consider in the context of particle insertion (ghost → real)
- High-S system is ghost, Low-S is real
- u is the difference in energy $U_{real} - U_{ghost}$
- p_H is the distribution of energies (virtual energy changes) experienced by molecule acting as a ghost (insertion energy)
  - *many overlaps, so energy will tend to be large*
- p_L is the distribution of energies experienced by a molecule interacting with the others (deletion energy)
  - *no overlap, favorable interactions, so energy will be small*
- Typical behaviors



Generalized Insertion and Deletion

- Widom insertion samples high-S system, perturbs to low-S system
- Widom deletion does the opposite
- Define
  - Generalized insertion: *FEP calculation in which high-entropy system governs sampling*
  - Generalized deletion: *FEP calculation in which low-entropy system governs sampling*





Distribution-Function Relations

- Previously we derived this result

$$\langle M \rangle_L = \frac{\left\langle M e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

Non-Boltzmann averaging formula
(used 0 and W to designate systems)

Distribution-Function Relations

- Previously we derived this result

$$\langle M \rangle_L = \frac{\left\langle M e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

- Take M ≡ $\delta[u-(U_L-U_H)]$

$$\langle \delta \rangle_L = \frac{\left\langle \delta e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

Distribution-Function Relations

- Previously we derived this result

$$\langle M \rangle_L = \frac{\left\langle M e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

- Take $M \equiv \delta[u-(U_L-U_H)]$

$$\boxed{\langle \delta \rangle_L} = \frac{\left\langle \delta e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

This is definition of p_L

- Use definitions of p_H and p_L

$$p_L(u) =$$

Distribution-Function Relations

- Previously we derived this result

$$\langle M \rangle_L = \frac{\left\langle M e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

- Take $M \equiv \delta[u-(U_L-U_H)]$

$$\langle \delta \rangle_L = \frac{\left\langle \delta e^{-\beta \boxed{(U_L - U_H)}} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

This is u; the delta function lets us take it outside the average

- Use definitions of p_H and p_L

$$p_L(u) = \frac{p_H(u) e^{-\beta u}}{}$$

Distribution-Function Relations

- Previously we derived this result

$$\langle M \rangle_L = \frac{\left\langle M e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

- Take $M \equiv \delta[u-(U_L-U_H)]$

$$\langle \delta \rangle_L = \frac{\left\langle \delta e^{-\beta(U_L - U_H)} \right\rangle_H}{\boxed{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}}$$

This is the free-energy difference

- Use definitions of p_H and p_L

$$p_L(u) = \frac{p_H(u) e^{-\beta u}}{e^{-\beta \Delta A}}$$

Distribution-Function Relations

- Previously we derived this result

$$\langle M \rangle_L = \frac{\left\langle M e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

- Take $M \equiv \delta[u-(U_L-U_H)]$

$$\langle \delta \rangle_L = \frac{\left\langle \delta e^{-\beta(U_L - U_H)} \right\rangle_H}{\left\langle e^{-\beta(U_L - U_H)} \right\rangle_H}$$

- Use definitions of p_H and p_L

$$p_L(u) = \frac{p_H(u) e^{-\beta u}}{e^{-\beta \Delta A}}$$

rearrange

$$\boxed{p_L(u) e^{-\beta \Delta A} = p_H(u) e^{-\beta u}}$$

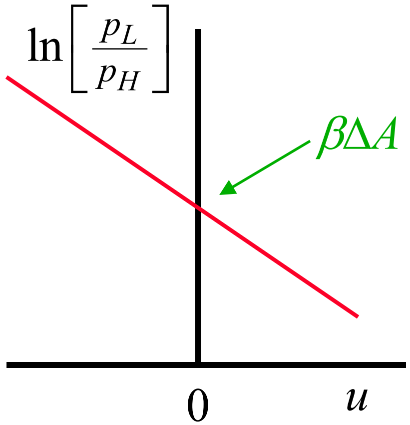
Using the p_H,p_L Relation

- Relation: $p_L(u)e^{-\beta\Delta A} = p_H(u)e^{-\beta u}$
- This can be used to obtain the free-energy difference
  - *Several equivalent formulations*

$$\ln\left[\frac{p_L(u)}{p_H(u)}\right] = \beta\Delta A - \beta u$$

Plot $\ln[p_L/p_H]$ vs u; slope -β, intercept βΔA

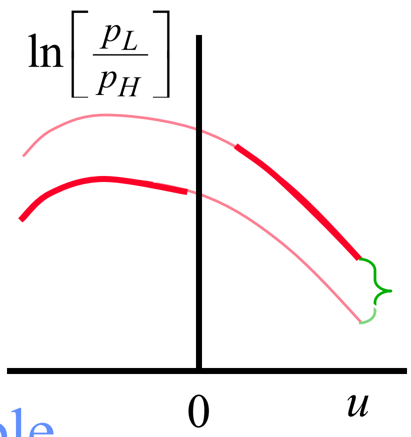
Requires that distributions have region of overlap



$$\ln[p_L(u)] + \tfrac{1}{2}\beta u = \ln[p_H(u)] - \tfrac{1}{2}\beta u - \beta\Delta A$$

Plot $\ln p_L + \beta u/2$ and $\ln p_H - \beta u/2$ vs. u on same plot; constant difference between is βΔA

Can be applied to nonoverlapping distributions if interpolating form is known



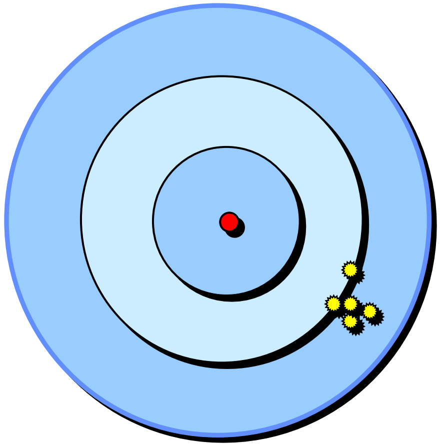
- More sophisticated methods are available
  - *Examine them later; present interest is using distributions to understand FEP performance*

Accuracy and Precision

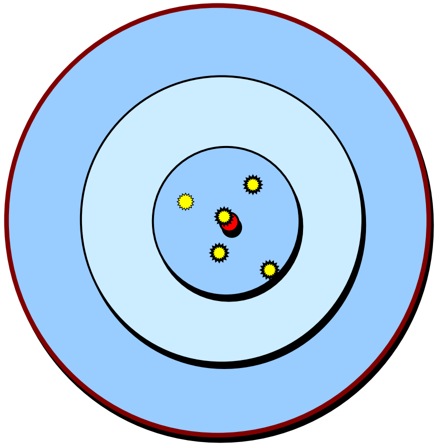
- Consider performance of FEP calculations from two perspectives
- Precision
  - *reproducibility of the result*
- Accuracy
  - *correctness of the result*
- Example
  - *hard-sphere deletion calculation*

    good precision

    terrible accuracy



Precise, but not accurate



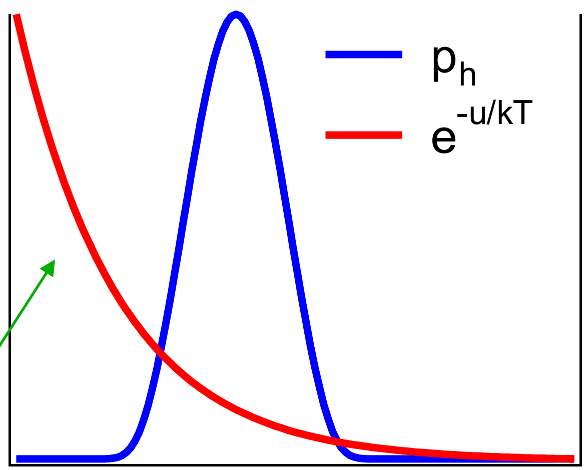
Accurate, but less precise

Tail Contributions in FEP Calculations

- Examine contributions to FEP averages

Generalized insertion

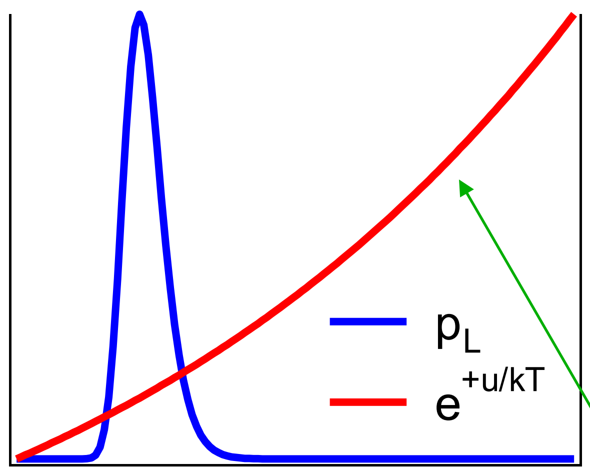
$$e^{-\beta\Delta A} = \int du\, e^{-\beta u} p_H(u)$$



Large contribution from tail at small u

Generalized deletion

$$e^{+\beta\Delta A} = \int du\, e^{+\beta u} p_L(u)$$

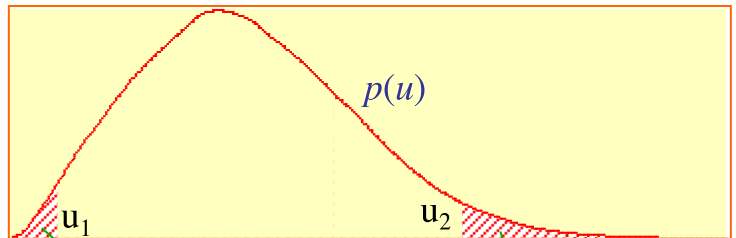


Large contribution from tail at large u

Inaccuracy in FEP Calculations 1.

- Main source of inaccuracy is inadequate sampling of tails

$$e^{A} = \int_{-\infty}^{+\infty} p(u)e(u)du$$



$$\int_{-\infty}^{u_1} p(u)e(u)du + \int_{u_2}^{\infty} p(u)e(u)du$$

Largest contribution to insertion inaccuracy

Largest contribution to deletion inaccuracy

- Model inaccuracy by assuming all error is due to missing tail contribution

$$e^{-\Delta A_H} - e^{-\Delta A_{exact}} = \int_{-\infty}^{u_H} p_H(u)e^{-\beta u}du = \boxed{e^{-\beta\Delta A}\int_{-\infty}^{u_H} p_L(u)du}$$

$$e^{+\Delta A_L} - e^{+\Delta A_{exact}} = \int_{u_L}^{\infty} p_L(u)e^{+\beta u}du = \boxed{e^{+\beta\Delta A}\int_{u_L}^{\infty} p_H(u)du}$$

Inaccuracy in each method is given by area under curve for other method

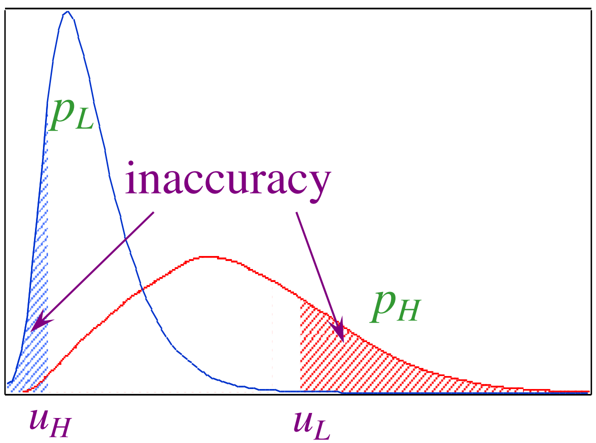
Inaccuracy in FEP Calculations 2.

Missing tail contributions

$$e^{-\Delta A_H} - e^{-\Delta A_{exact}} = \int_{-\infty}^{u_H} p_H(u) e^{-\beta u} du = \boxed{e^{-\beta \Delta A} \int_{-\infty}^{u_H} p_L(u) du}$$

$$e^{+\Delta A_L} - e^{+\Delta A_{exact}} = \int_{u_L}^{\infty} p_L(u) e^{+\beta u} du = \boxed{e^{+\beta \Delta A} \int_{u_L}^{\infty} p_H(u) du}$$

Inaccuracy in each method is given by area under curve for other method



Relative inaccuracy

$$\delta_H \equiv \frac{e^{-\Delta A_{sim,H}} - e^{-\Delta A_{exact}}}{e^{-\Delta A_{exact}}} = \int_{-\infty}^{u_H} p_L du \qquad \Delta A_{sim,H} - \Delta A_{exact} > 0$$

Insertion overestimates

$$\delta_L \equiv \frac{e^{-\Delta A_{sim,L}} - e^{-\Delta A_{exact}}}{e^{-\Delta A_{exact}}} = -\int_{u_L}^{\infty} p_H du \qquad \Delta A_{sim,L} - \Delta A_{exact} < 0$$

Deletion underestimates

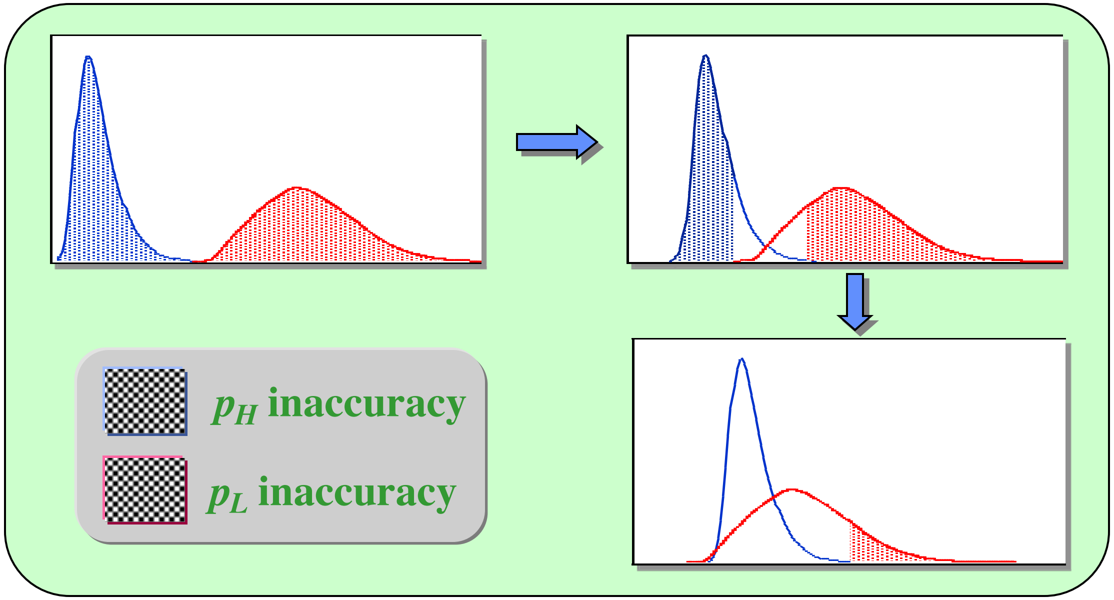
Asymmetry of the Inaccuracy 1.

- The opposite tendency of the insertion/deletion inaccuracy leads to statements like these
  - *"The forward and reverse [inaccuracy] should be of the same magnitude and opposite sign"* J. Phys. Chem., 98, 1487-1493,1999
  - *"The free energy change was taken as the average of the forward and reverse free energies."* J Comp. Chem., 20, 499-510, 1999
- Remember the asymmetry of the hard-sphere insertion/deletion methods
  - *for insertion, $e^{-\beta\mu}$ is zero until a non-overlap is completed*
  - *for deletion $e^{+\beta\mu}$ is always unity*
  - *averaging the insertion and deletion μ's would be bad*

Asymmetry of the Inaccuracy 2.

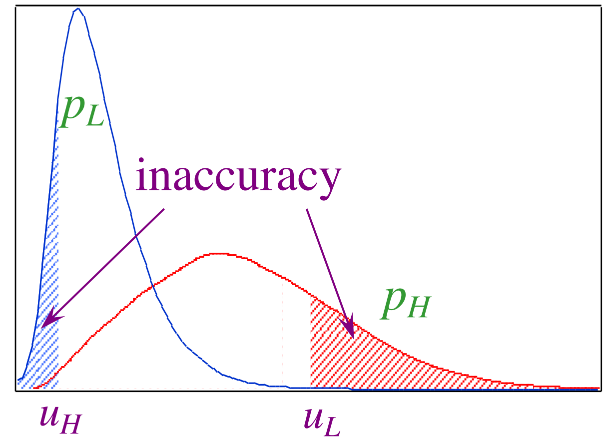
- p_H and p_L have different variances
- Reference with broader distribution gives more accurate result
- Large entropy reference has larger variance hence gives more accurate result
- Insertion is more reliable than deletion

Improvement of accuracies as length of simulation grows



Predicting Inaccuracy

- Maximum likelihood analysis
  - *consider most likely outcome for simulation with length M*
- Need most likely values for u_H, u_L
- Consider probability that largest deletion energy is some value, u*, after M attempted deletions



$$\mathrm{prob}(u_L = u^*) = \mathrm{prob}(\text{u* is sampled}) \times \mathrm{prob}(\text{u > u* is never sampled})$$

$$= p_L(u^*) \times \left[1 - \int_{u^*}^{\infty} p_L(u)du\right]^M$$

Similar formula derived for insertion

- Maximize with respect to u*

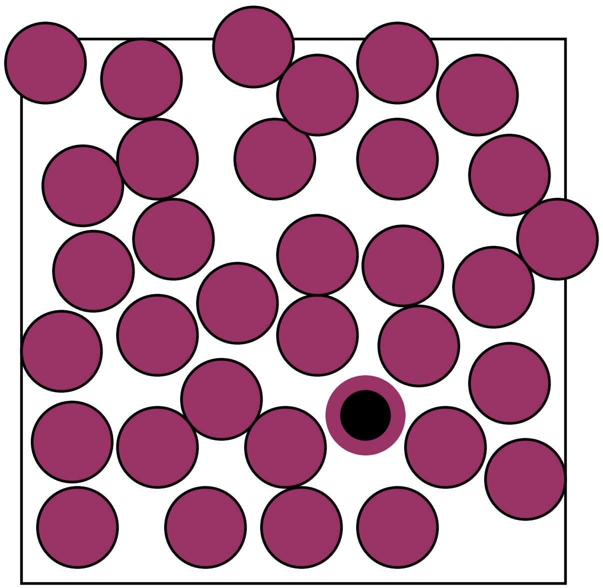
$$\left.\frac{\partial \ln p_H(u)}{\partial u}\right|_{u_H} = Mp_H(u_H) - \beta$$

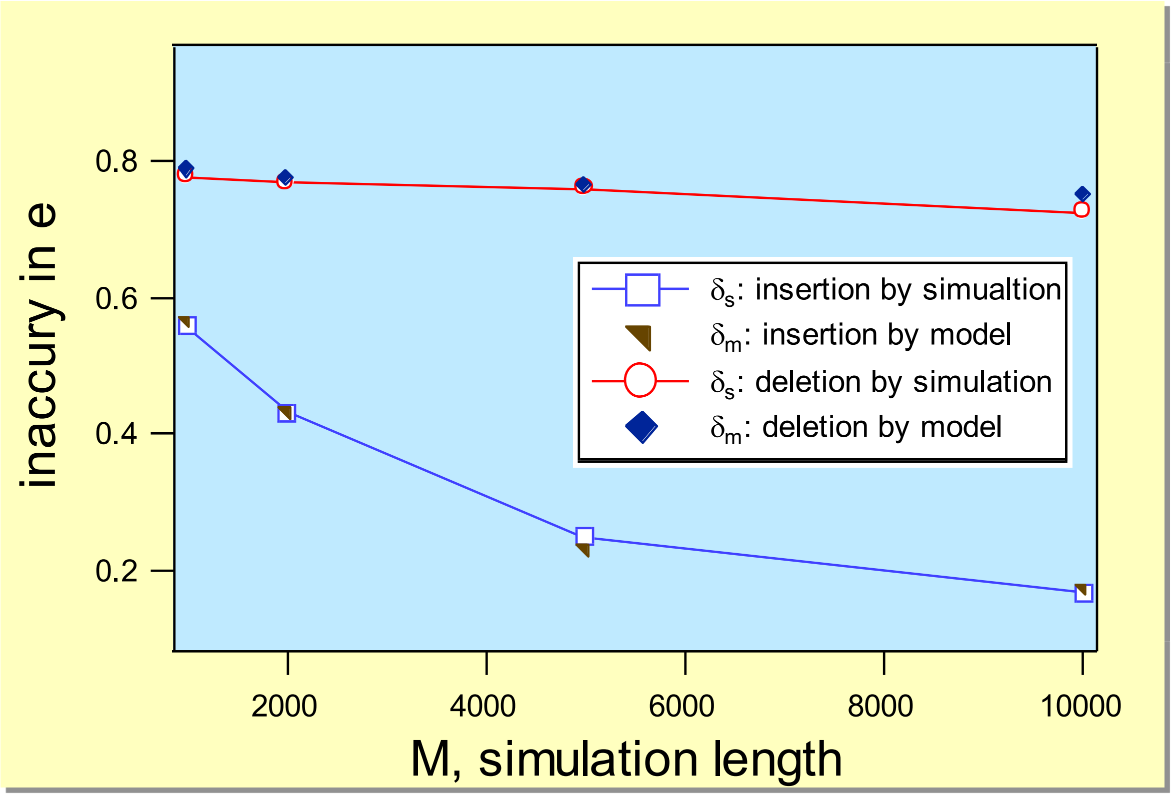
$$\left.\frac{\partial \ln p_L(u)}{\partial u}\right|_{u_L} = -Mp_L(u_L) - \beta$$

Testing Inaccuracy Model

MC Simulation

- *NVT*
- *(N-1) LJ + 1 HS* $\longleftrightarrow$ *N LJ*
- *HS diameter = 0.8*
- $T = 2.0;\ \ \rho = 0.9$
- *simulation repeats up to 200 runs*





Knowing Your Inaccuracy

- How can the accuracy of a simulation result be assessed if the simulation is inaccurate?
- Compare to precision calculation where simulation data (variance) are used to provide confidence limits
- Consider most-likely inaccuracy for HS insertion

$$\delta_{\Delta A} = \left(2Me^{\Delta S}\right)^{-1}$$

Insertion only ($\Delta S < 0$)

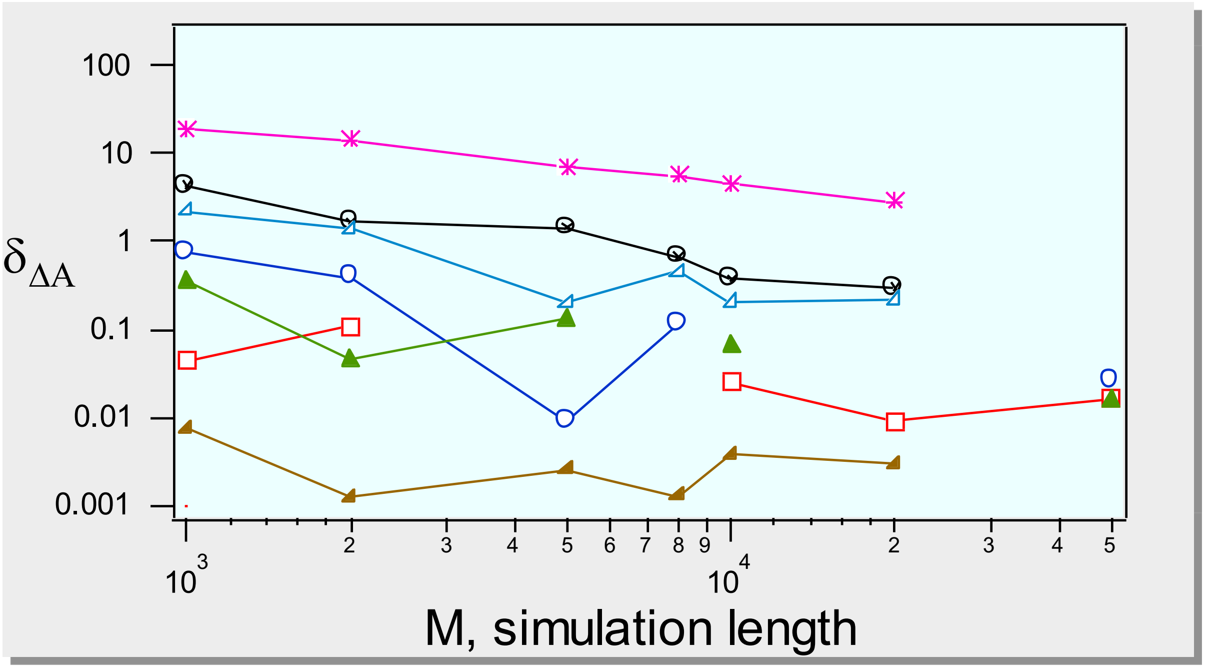
- Postulate $\delta_{\Delta A} \sim (Me^{\Delta S})^{-1}$ for continuous distributions
  - *evaluate ΔA accuracy using simulation ΔS*
- But simulation gives 'incorrect' ΔS
  - *generally, simulation ΔS < true ΔS* ($e^{-\Delta S(\mathrm{sim})} > e^{-\Delta S(\mathrm{true})}$)
  - *thus 'incorrect' ΔS indicates larger error*
  - *safe estimate of inaccuracy*
  - *gives (probabilistic) upper bound of ΔA inaccuracy*

Test of Postulated Form 1.

- *MC simulations with various conditions*
- *repeat simulations for up to 100 independent runs*
- *very long simulation generates pseudo true ΔA*
- *calculate entropy change, error-bar, inaccuracy etc.*

Series	reference	density	temperature	$\Delta S/k$
1	(N-1)LJ + 1 (LJ with α = 0.9)	0.9	2.0	-1.702
2	(N-1)LJ + 1 (LJ with α = 0.72)	0.9	2.0	-4.250
3	(N-1)LJ + 1 (LJ with α = 0.65)	0.8	1.0	-4.450
4	(N-1)LJ + 1 (LJ with α = 0.7)	0.9	1.0	-5.799
5	(N-1) LJ	0.8	1.0	-8.743
6	(N-1)LJ + 1 (soft with α = 0.3)	0.9	1.0	-9.504
7	(N-1) LJ	0.9	1.0	-12.179

$N = 108$



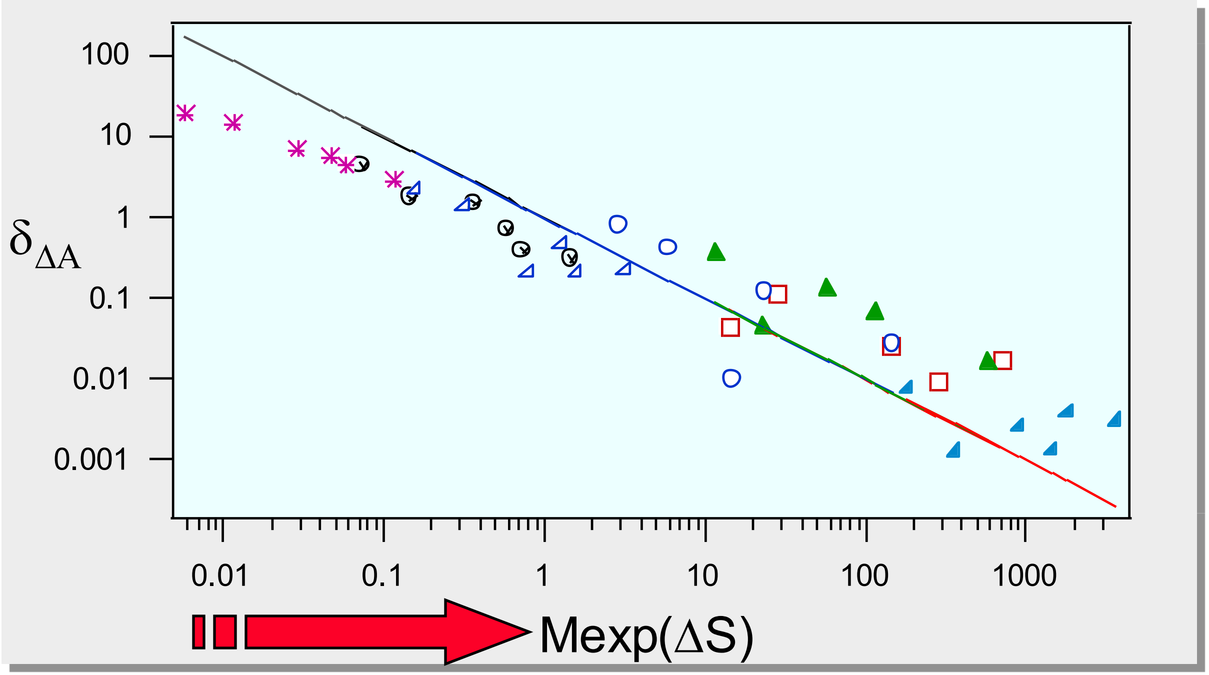
Test of Postulated Form 2.

- *MC simulations with various conditions*
- *repeat simulations for up to 100 independent runs*
- *very long simulation generates pseudo true ΔA*
- *calculate entropy change, error-bar, inaccuracy etc.*

Series	reference	density	temperature	$\Delta S/k$
1	(N-1)LJ + 1 (LJ with α = 0.9)	0.9	2.0	-1.702
2	(N-1)LJ + 1 (LJ with α = 0.72)	0.9	2.0	-4.250
3	(N-1)LJ + 1 (LJ with α = 0.65)	0.8	1.0	-4.450
4	(N-1)LJ + 1 (LJ with α = 0.7)	0.9	1.0	-5.799
5	(N-1) LJ	0.8	1.0	-8.743
6	(N-1)LJ + 1 (soft with α = 0.3)	0.9	1.0	-9.504
7	(N-1) LJ	0.9	1.0	-12.179

N = 108

$$\delta_{\Delta A} = \left(2Me^{\Delta S}\right)^{-1}$$



Appropriate group, perhaps incorrect exponent

Precision of FEP Calculations 1.

- Consider L simulations, each doing M insertions

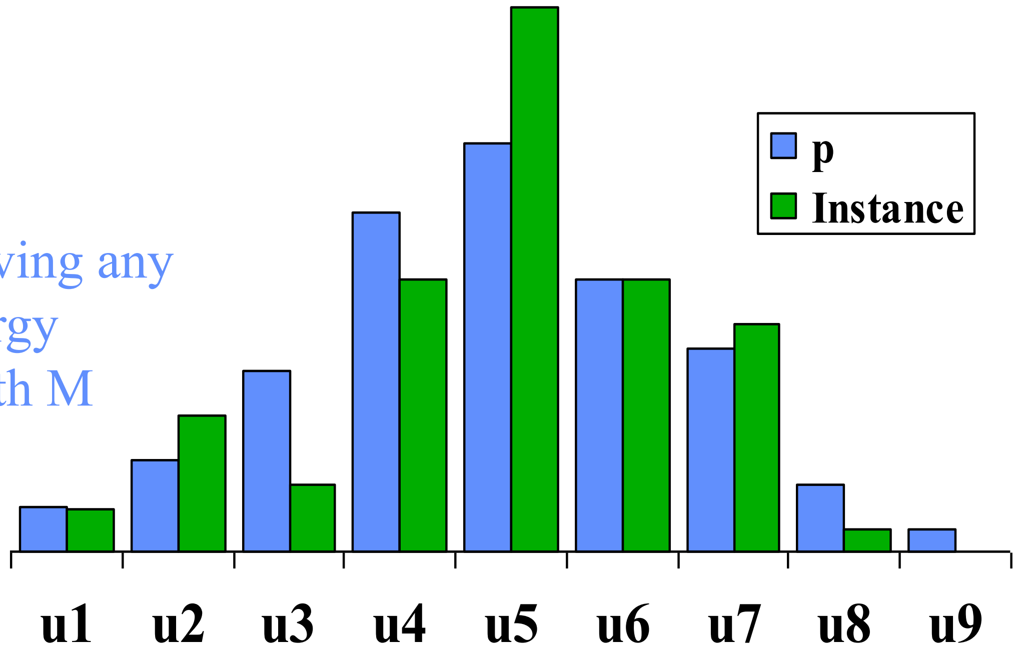
$$\left.\begin{array}{l} 1,2,3,4,5,\ldots,M \\ 1,2,3,4,5,\ldots,M \\ . \\ . \\ . \\ 1,2,3,4,5,\ldots,M \end{array}\right\} L \text{ times}$$

- Each M-length run gives a value for ΔA
- Variance of these averages for the L runs describes the precision of the calculation

Precision of FEP Calculations 2.

- Discretize p_H
- Consider probability of observing any given distribution of FEP energy values in a simulation of length M
- Follows binomial distribution

$$\Omega[\{m_i\}] = M!\prod \frac{p_i^{m_i}}{m_i!}$$



- Variance in FEP average given in terms of variance of this distribution

$$\sigma_{\Delta A}^2 \approx e^{\pm\beta\Delta A}\sum p_i(1-p_i)e^{\pm 2\beta u_i}$$

- Return to continuum formulation, rewrite

$$\left(M\sigma^2\right)_{del} = \exp(-\beta\Delta A)\int_{-\infty}^{\infty} p_H(u)e^{+\beta u}du$$

$$\left(M\sigma^2\right)_{ins} = \exp(+\beta\Delta A)\int_{-\infty}^{\infty} p_L(u)e^{-\beta u}du$$

Note that exponents have signs opposite those for averages

Precision of FEP Calculations 3.

Decompose into entropic and energetic contributions

- *focus on insertion form*

$$\left(M\sigma^2\right)_{ins} = e^{-\Delta S/k}\int_{-\infty}^{\infty} du\, p_L(u) e^{-\beta(u-\Delta U)}$$

- *expand*

$$\left(M\sigma^2\right)_{ins} = e^{-\Delta S/k}\int_{-\infty}^{\infty} du\, p_L(u)\left[1 - \cancel{\beta(u-\Delta U)} + \tfrac{1}{2}\beta^2(u-\Delta U)^2 + \ldots\right]$$

- *finally*

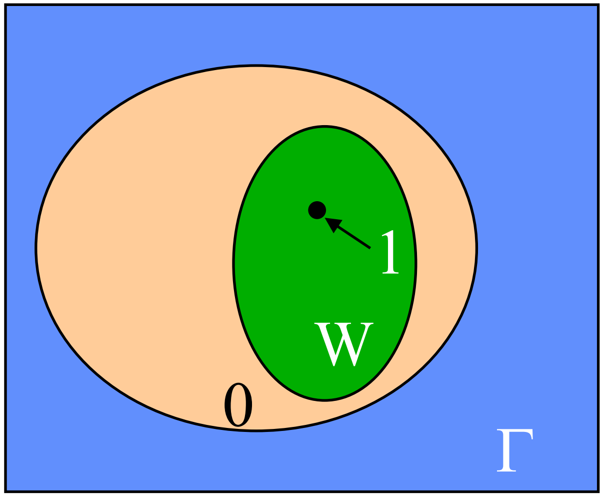
$$\boxed{\left(M\sigma^2\right)_{ins} = e^{-\Delta S/k}\left(1 + \tfrac{1}{2}\beta^2\hat{\sigma}^2_{\Delta U}\right) \equiv \zeta e^{-\Delta S/k}}$$

Variance of energy in L system (independent of M); O(1) quantity

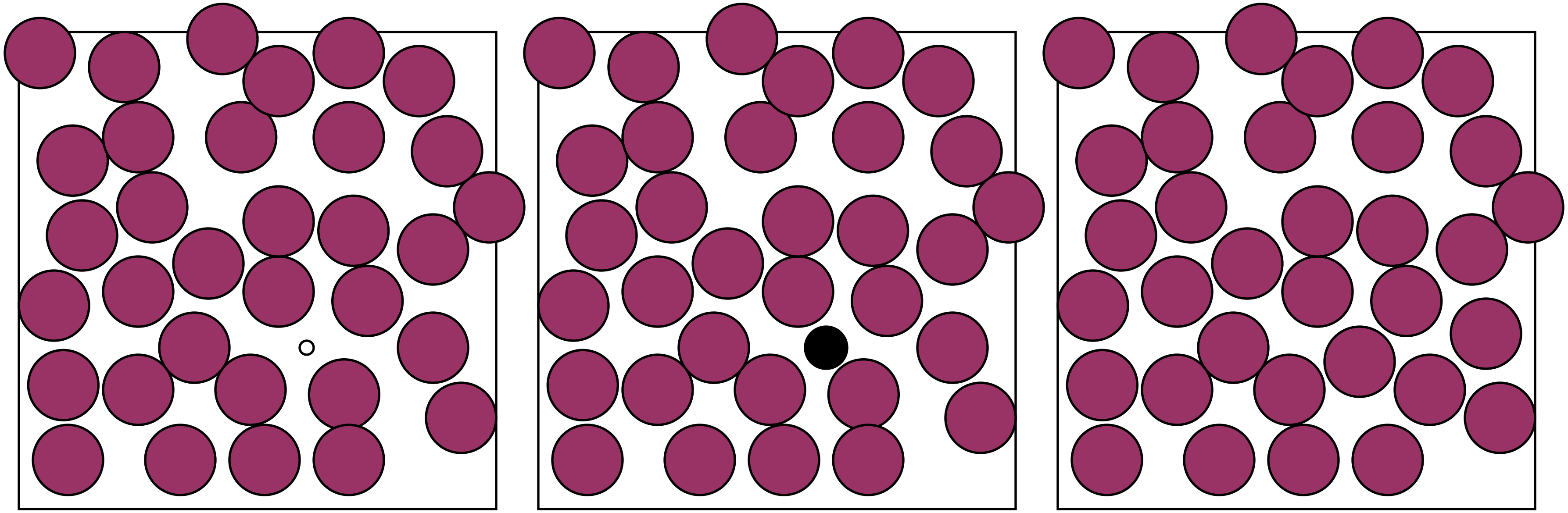
- *entropy difference is key*

Optimal Staging

- Apply precision model to optimize choice of intermediate in staged insertion
  - *how best to define U_W?* $e^{-\beta\Delta(A_L - A_H)} = \left\langle e^{-(U_w - U_H)} \right\rangle_H \left\langle e^{-\beta(U_L - U_w)} \right\rangle_w$
- Overall variance is the sum of variances of all stages
- Choose W:
  - *minimize* $\left(M\sigma^2\right)_{tot} = \sum_i \zeta_i \exp\left(-\Delta S_i / k\right)$
  - *subject to* $\sum \Delta S_i = \Delta S_{tot}$
  - *obtain* $\boxed{\Delta(\Delta S)_{ij} \equiv \Delta S_j - \Delta S_i = k \ln\left(\zeta_j / \zeta_i\right)}$
  - ζ_j/ζ_i: *order of unity*
- Heuristic: $\boxed{\Delta(\Delta S) = 0}$
  - *equal entropy difference*
  - *compare to unjustified rule-of-thumb*
    equal free-energy difference
    $\Delta(\Delta A/k) = 0$
  - *new rule greatly improves the precision*



Example Application



<u>System H</u>
N-1 LJ particles

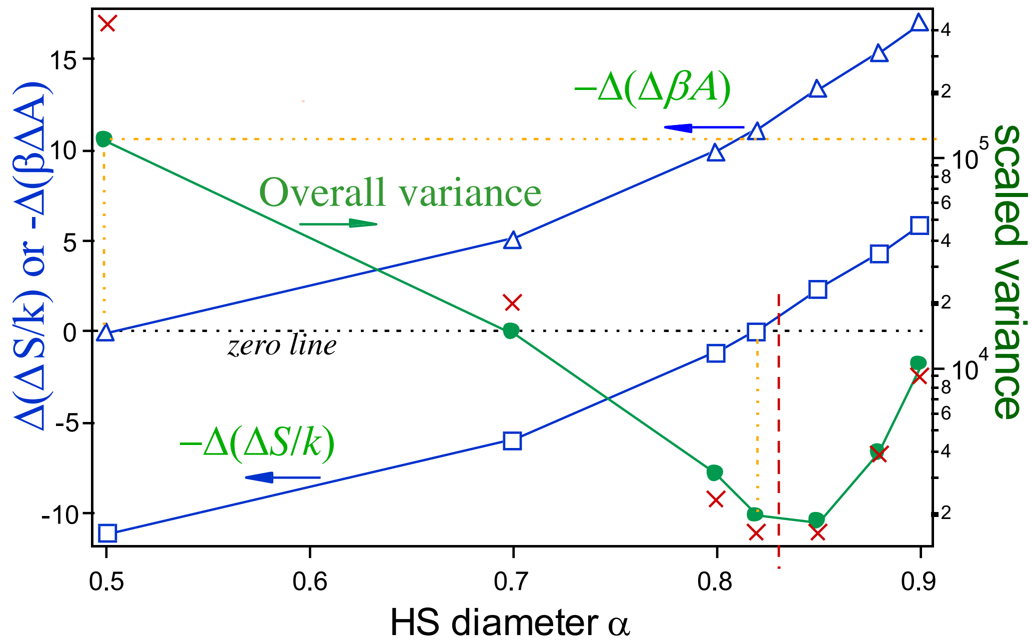
<u>System W</u>
N-1 Lennard-Jones particles
1 Hard sphere of diameter α

<u>System L</u>
N LJ particles

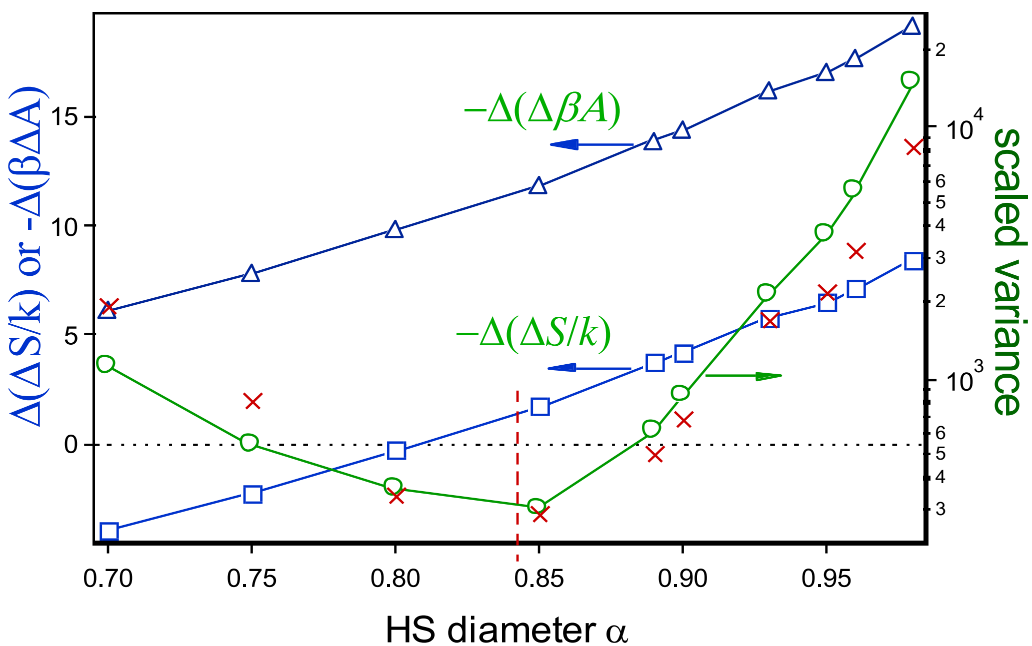
- Optimize with respect to intermediate-HS diameter α

Example Results

N=108
T=1.0
ρ=0.9



N=200
T=1.0
ρ=0.8



Summary

- FEP energy distributions provide detailed information regarding free-energy differences
- Relation between insertion and deletion distribution can be used to measure free-energy differences
- Distributions can be used to understand precision and accuracy of FEP calculations
- Insertion usually gives too-high value, deletion too low
  - *but not equally so; deletion is much worse*
- Formulated way to estimate inaccuracy using inaccurate data
- Both accuracy and precision strongly depend on entropy difference between states
- Can use precision analysis to optimize staged insertions