Computational Chemistry

Considered CHARMMful

Josh Mitchell

Why am I here

Why am I here

Computational methods can be...

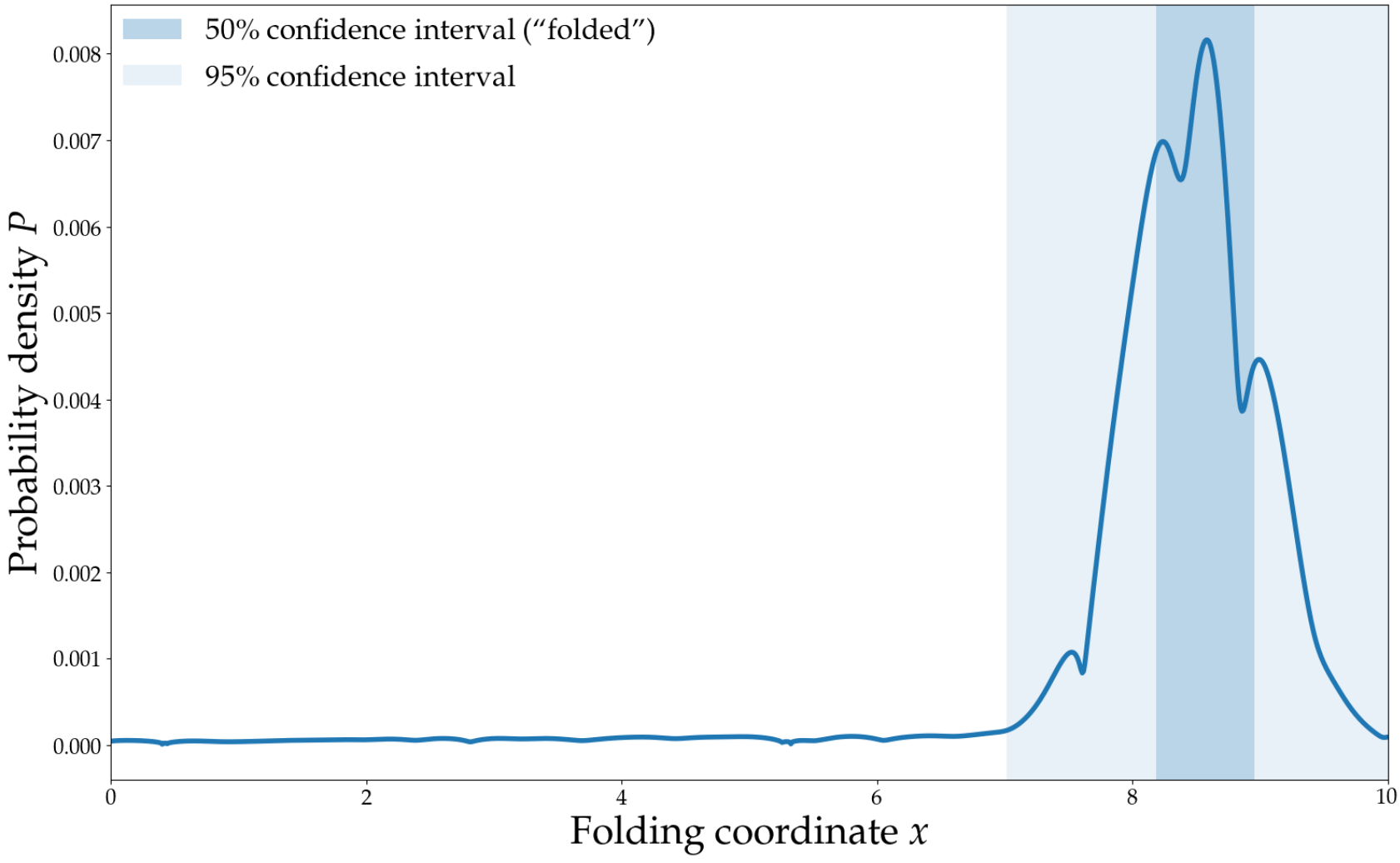
- Complementary to experiment
- Faster/cheaper than experiment
- More detailed than experiment

Though usually not at the same time

Lets think statistically

Protein structure forms according to a probability distribution

Folded proteins have a funnel-shaped probability distribution



Ready for some maths?

The Boltzmann Distribution

$$P_i = \frac{\exp(-\beta\epsilon_i)}{Z}$$

P_i is the probability of state i at equilibrium

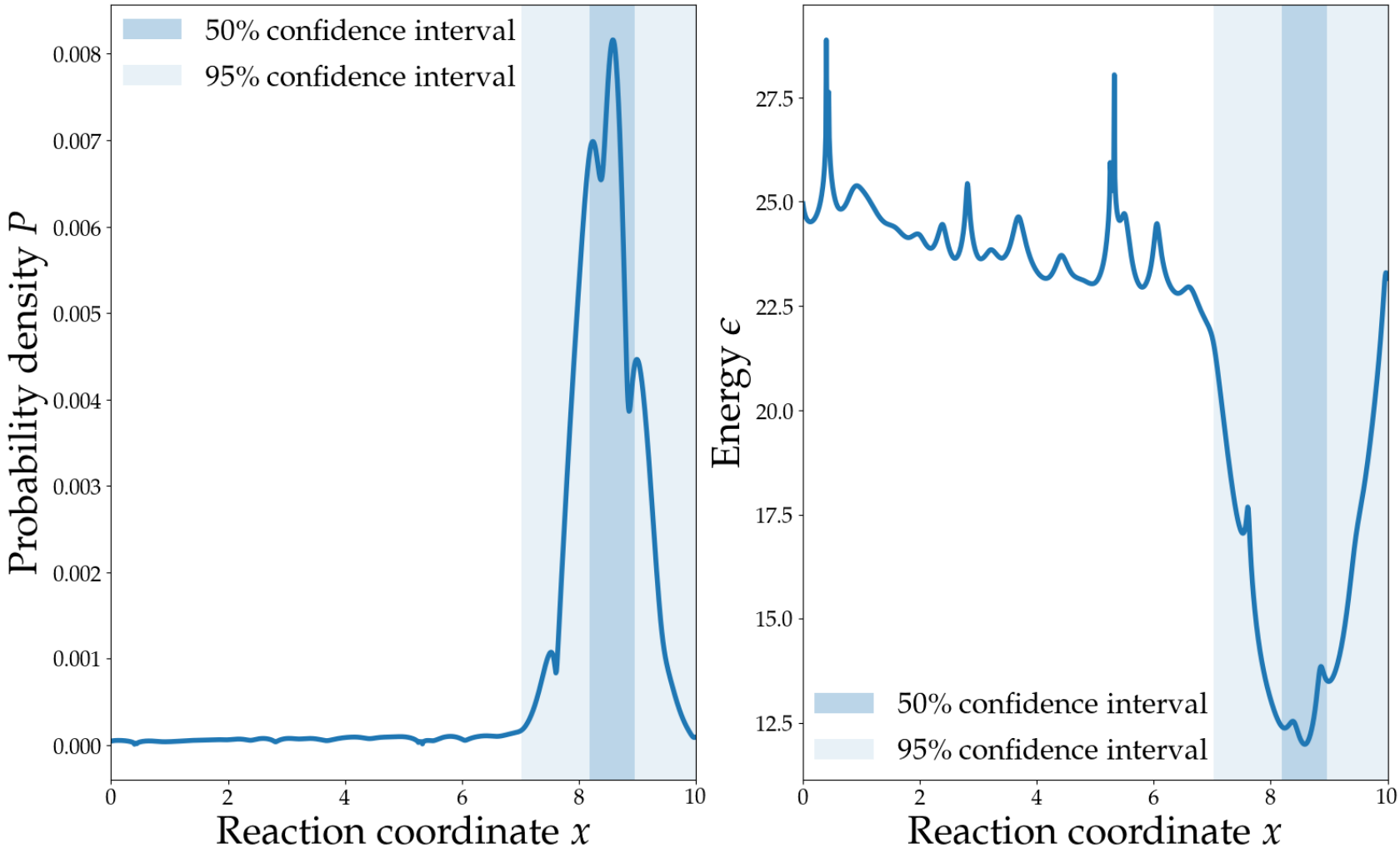
$\exp(x)$ is the exponential function e^x

$\beta = \frac{1}{K_B T}$ incorporates the temperature

ϵ_i is the energy of state i . . .

$$Z = \sum_j \exp(-\beta\epsilon_j)$$

The Boltzmann distribution



There are two problems

$$P_i = \frac{\exp(-\beta \epsilon_i)}{Z}$$

1. Which states are most important
2. How important are those states

Fast methods assume there is only one important state

So we just find the minimum energy state i

$$P_{\text{min}} = \frac{\exp(-\beta\epsilon_{\text{min}})}{\exp(-\beta\epsilon_{\text{min}})} = 1$$

Fast methods assume there is only one important state

So we just find the minimum energy state i

$$P_{\min} = \frac{\exp(-\beta\epsilon_{\min})}{\exp(-\beta\epsilon_{\min})} = 1$$

Entropic effects must be part of the energy ϵ

Energy function doesn't need much detail

What happens if we don't make that assumption?

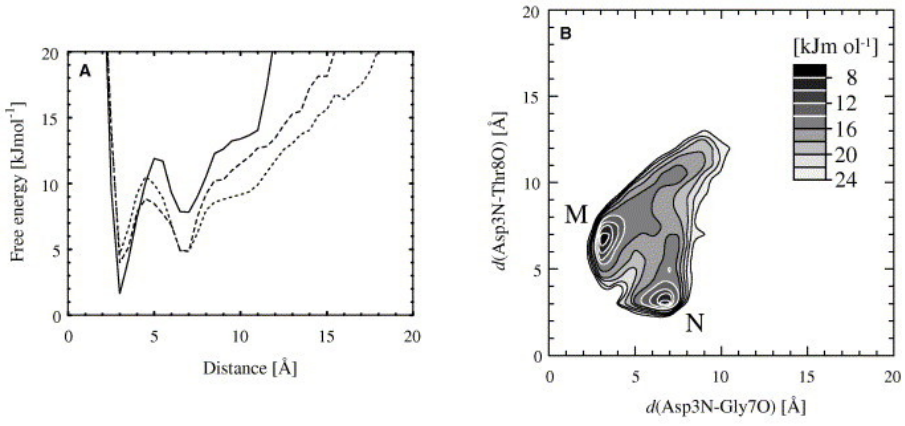
What happens if we don't make that assumption?

$$P_i = \frac{\exp(-\beta \epsilon_i)}{Z}$$

We need to *sample* from the entire Boltzmann distribution with a detailed energy function

Sampling is hard

The curse of dimensionality



(Satoh et al. 2006)

Levinthal's paradox

~50 KDa, 500 residue protein

Assume each residue is either -helix or -sheet

500 binary dimensions

Looking for 1 folded state in 2^{500} ($\sim 10^{150}$) possible states

Yet, nature can do it

Molecular dynamics

Calculate forces from the energy function ("force field")

Step the system through time with classical mechanics

Let the shape of the probability distribution guide your sampling

Simulation is the method, not the goal

Kinetics are unreliable

Produces lots of noisy data

Humans are good at finding patterns

Know what you're looking for

Just like experiments

I'm looking for a non-crystallographic, highly populated state

I'm looking for the important states in solution near the crystal structure

I'm looking for the folded state from an extended model or homology model

I wanna measure the free energy of this process

I wanna measure the FRET efficiency in this state

I wanna see if my crystal structure is stable in solution

Make sure it's computationally feasible

Check you've found it

Check everything seems reasonable

The good, the bad and the user in soft matter simulations (Wong-ekkabut and Karttunen 2016)

Make sure it's not just a statistical fluke

Best practices for quantification of uncertainty and sampling quality in molecular simulations (Grossfield et al. 2018)

Run replicas

Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas (Knapp, Ospina, and Deane 2018)

So

- We sample from the Boltzmann distribution
- Simulation is the method, not the goal
- Know what you're looking for
- Check you've found it

Questions about sampling?

Anyone want more maths?

Anyone want more maths?

A really nice way to think about entropy is only a few algebraic manipulations away!

What *is* energy

$$P_i = \frac{\exp(-\beta\epsilon_i)}{Z}$$

What *is* energy

$$P_i = \frac{\exp(-\beta\epsilon_i)}{Z}$$

$$\exp(-\frac{1}{K_B T}\epsilon_i) = P_i Z$$

What *is* energy

$$P_i = \frac{\exp(-\beta\epsilon_i)}{Z}$$

$$\exp(-\frac{1}{K_B T}\epsilon_i) = P_i Z$$

$$-\frac{1}{K_B T}\epsilon_i = \log(P_i Z)$$

What *is* energy

$$P_i = \frac{\exp(-\beta\epsilon_i)}{Z}$$

$$\exp(-\frac{1}{K_B T}\epsilon_i) = P_i Z$$

$$-\frac{1}{K_B T}\epsilon_i = \log(P_i Z)$$

$$\epsilon_i = -K_B T \log(P_i) - K_B T \log(Z)$$

What *is* energy

$$\epsilon_i = -K_B T \log(P_i) - K_B T \log(Z)$$

What *is* energy

$$\epsilon_i = -K_B T \log(P_i) - K_B T \log(Z)$$

$$\epsilon_{i\mathrm{rel}} = -K_B T \log(P_i)$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$P_i = \Omega_i p_i$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$P_i = \Omega_i p_i$$

$$\epsilon_i = -K_B T \log(\Omega_i p_i) - K_B T \log(Z)$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$P_i = \Omega_i p_i$$

$$\epsilon_i = -K_B T \log(\Omega_i p_i) - K_B T \log(Z)$$

$$\epsilon_i = -K_B T \log(\Omega_i) - K_B T \log(p_i) - K_B T \log(Z)$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$\epsilon_i = -K_B T \log(\Omega_i) - K_B T \log(p_i) - K_B T \log(Z)$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$\epsilon_i = -K_B T \log(\Omega_i) - K_B T \log(p_i) - K_B T \log(Z)$$

$$S_i = K_B \log(\Omega_i)$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$\epsilon_i = -K_B T \log(\Omega_i) - K_B T \log(p_i) - K_B T \log(Z)$$

$$S_i = K_B \log(\Omega_i)$$

$$\epsilon_i = -T S_i - K_B T \log(p_i) - K_B T \log(Z)$$

What is entropy?

Suppose we have a state i that comprises Ω_i microstates, each of probability p_i

$$\epsilon_i = -K_B T \log(\Omega_i) - K_B T \log(p_i) - K_B T \log(Z)$$

$$S_i = K_B \log(\Omega_i)$$

$$\epsilon_i = -T S_i - K_B T \log(p_i) - K_B T \log(Z)$$

In reality, not all the microstates have the same energy, but this is the gist.

What's the difference?

$$\epsilon_i = -K_B T \log(\Omega_i p_i) - K_B T \log(Z)$$

$$\epsilon_j = -K_B T \log(\Omega_j p_j) - K_B T \log(Z)$$

What's the difference?

$$\epsilon_i = -K_B T \log(\Omega_i p_i) - K_B T \log(Z)$$
$$\epsilon_j = -K_B T \log(\Omega_j p_j) - K_B T \log(Z)$$

$$\epsilon_i - \epsilon_j = -K_B T \log(\Omega_i p_i) + K_B T \log(\Omega_j p_j)$$

What's the difference?

$$\epsilon_i = -K_B T \log(\Omega_i p_i) - K_B T \log(Z)$$
$$\epsilon_j = -K_B T \log(\Omega_j p_j) - K_B T \log(Z)$$

$$\epsilon_i - \epsilon_j = -K_B T \log(\Omega_i p_i) + K_B T \log(\Omega_j p_j)$$

$$\epsilon_i - \epsilon_j = -K_B T \log(p_i) + K_B T \log(p_j)$$
$$-K_B T \log(\Omega_i) + K_B T \log(\Omega_j)$$

What's the difference?

$$\epsilon_i = -K_B T \log(\Omega_i p_i) - K_B T \log(Z)$$
$$\epsilon_j = -K_B T \log(\Omega_j p_j) - K_B T \log(Z)$$

$$\epsilon_i - \epsilon_j = -K_B T \log(\Omega_i p_i) + K_B T \log(\Omega_j p_j)$$

$$\epsilon_i - \epsilon_j = -K_B T \log(p_i) + K_B T \log(p_j)$$
$$-K_B T \log(\Omega_i) + K_B T \log(\Omega_j)$$

$$\Delta\epsilon = \Delta[-K_B T \log(p)] - T\Delta[K_B \log(\Omega)]$$

Anyone recognise this?

$$\Delta\epsilon = \Delta[-K_B T \log(p)] - T\Delta[K_B \log(\Omega)]$$

Anyone recognise this?

$$\Delta\epsilon = \Delta[-K_B T \log(p)] - T\Delta[K_B \log(\Omega)]$$

$$G = \epsilon$$

$$H = -K_B T \log(p)$$

$$S = K_B \log(\Omega)$$

Anyone recognise this?

$$\Delta\epsilon = \Delta[-K_B T \log(p)] - T\Delta[K_B \log(\Omega)]$$

$$G = \epsilon$$

$$H = -K_B T \log(p)$$

$$S = K_B \log(\Omega)$$

$$\Delta G = \Delta H - T\Delta S$$

If you want more, check out stat mech!

No more maths

No more maths

I promise

A quick run down of force fields

What to look for in a force field

Can it model all the parts of my system?

Does it accurately produce the kind of data I'm looking for?

Has it been validated by people other than the authors?

Can I trust the parameters I have in the format I use?

WARNINGS

Don't confuse force field and software!

Don't mix parameters from different force fields!

Most force fields are not parameterised for kinetics!

Most force fields are only parameterised at one temperature!

CHARMM

CHARMM22* Extremely well validated force field for vanilla proteins. Reasonable accuracy even for IDPs and loop regions, despite not being parametrised for this. CHARMM22 modified by Shaw group.

CHARMM36m CHARMM36 modified for better performance with IDPs and loop regions. Distributed in many formats by the authors.

AMBER

Great supplementary DNA and metallic ion parameters

AMBER99SB-*disp* AMBER99SB-ILDN with torsion and protein-water VDW optimisations by Shaw group. 4-point water model (slow). State-of-the-art IDP/loop accuracy (according to authors).

AMBER99SB*-ildn AMBER99SB with optimised torsions. Solid, widely used AMBER force field. Predates the IDP revolution. Several daughter force fields with improved performance.

MARTINI

Coarse grained!

1000-fold faster than atomic MD

Converged MARTINI probably more accurate than unconverged atomic force fields

Other stuff

GROMOS 54a7 Great for arbitrary chemicals (ATB). Outdated for proteins.

OPLS 3e Schrödinger's proprietary force field. Claims experimental accuracy for binding free energies of arbitrary drug molecules to proteins. Expensive to license. Proteins under-validated.

Questions about force fields?

A few traps

Use the same parameters as your force field authors

Read the paper!

Use same VDW cutoff range and method.

PME is usually OK if force field uses other long-range electrostatic treatment.

Use same water model.

Use same constraints (everyone breaks this rule and it might be OK)

Thermostats and barostats

Don't ever use Berendsen thermostats or barostats in production!

Thermostat

Just use Bussi's stochastic velocity rescaling thermostat for everything (`v-rescale` in GROMACS). Use a Nosé–Hoover chain if that's unavailable.

Barostat

Use a Monte Carlo barostat if available (unless you really care about how the box changes shape). Coming soon to GROMACS.

If not, or if you care about box dynamics:

- Use Berendsen or Monte Carlo for equilibration
- Use Parrinello-Rahman or MTTK for production

Step size

It's OK to make it as big as possible, as long as your simulation doesn't crash.

If your simulation crashes, try reducing the step size.

Atomic production simulations should never need to be below 1 fs, or 2 fs with constraints.

Adjacent frames are very similar - don't be afraid to drop them!

Non-monovalent metal cations

Shape of orbitals is important IRL, but force fields are spherically symmetric!

Some work has been done on introducing virtual particles to correct this

Box size

Smaller box lets you have less water and faster simulation

Too small box introduces finite size artifacts

Keep your periodic image distance larger than VdW cutoff

Rhombic Dodecahedral boxes have about 0.707 times the volume of a cubic box

Orientation of protein can change PI distance!

Protonation state

Set at start of simulation.

If you're doing experiments near the PK_a of something, be careful!

Enhanced sampling

Enhanced sampling is good.

Replica exchange is good.

Biased sampling methods are good if you have a good reaction coordinate.

Ensemble sampling methods will probably be good very soon.

Be careful with accelerated MD.

Make sure you should know what you're doing!

Further reading and references

Further reading and references

Braun, Efrem, Justin Gilmer, Heather B. Mayes, David L. Mobley, Jacob I. Monroe, Samarjeet Prasad, and Daniel M. Zuckerman. 2018. "Best Practices for Foundations in Molecular Simulations [Article V1.0]." *Living Journal of Computational Molecular Science* 1 (1): 5957. https://doi.org/10.33011/livecoms.1.1.5957.

Eastman, Peter. 2015. "Introduction to Statistical Mechanics." 2015. https://peastman.github.io/statmech/.

Grossfield, Alan, Paul N. Patrone, Daniel R. Roe, Andrew J. Schultz, Daniel Siderius, and Daniel M. Zuckerman. 2018. "Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article V1.0]." *Living Journal of Computational Molecular Science* 1 (1): 5067. https://doi.org/10.33011/livecoms.1.1.5067.

Knapp, Bernhard, Luis Ospina, and Charlotte M. Deane. 2018. "Avoiding False Positive Conclusions in Molecular Simulation: The Importance of Replicas." *Journal of Chemical Theory and Computation* 14 (12): 6127–38.