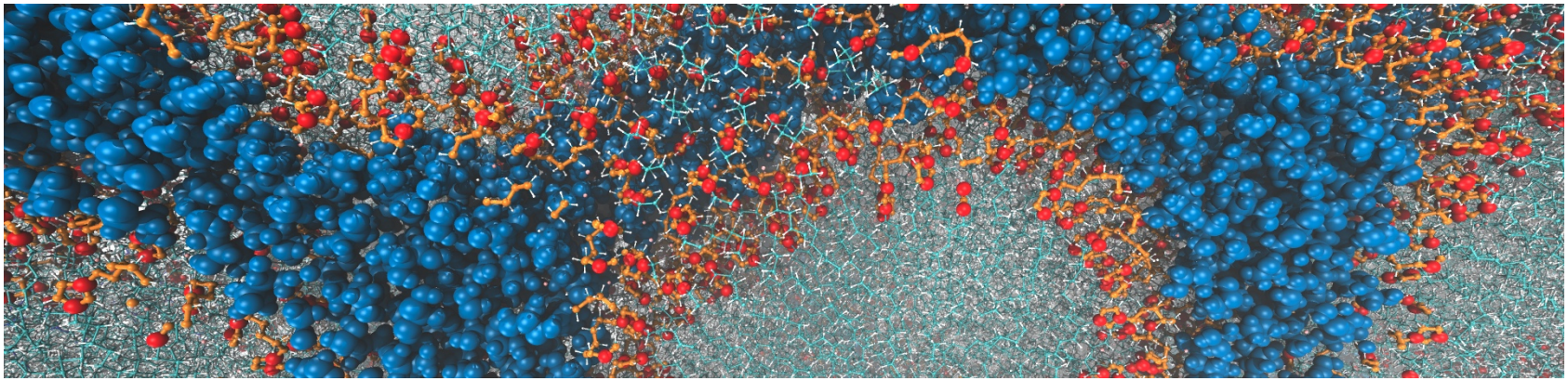
How do I know it's right? Physical validations of molecular simulations

Prof. Michael Shirts
University of Colorado, Boulder
i-CoMSE MD/MC Summer Workshop
Oklahoma State University, July 2022

What do we use molecular simulations to do?



- ASSUMING we are looking at properties of soft materials, liquids or solutions
- We are trying to model macroscopic behavior, usually thermodynamical, by modeling ensembles of molecules
- Results are statistical averages or other ensemble measures
- We try to get away with using classical mechanics

i-CoSE

Acknowledgements

Grant R01GM115790

NIH National Institutes of Health
Turning Discovery Into Health

i-CoSE

Physically correct simulations are necessary for reproducibility

Correct simulations

incorrect simulation

"Happy families are all alike; every unhappy family is unhappy in its own way."

incorrect

- Leo Tolstoy, first line of *Anna Karenina*

i-CoSE

Physically correct simulations are necessary for reproducibility

Correct simulations are all alike; every incorrect simulation is incorrect in its own way

- There are many ways to not conserve energy
- There are many ways to not have equipartition of energy
- There are many ways to not be Boltzmann-weighted
- There are many ways to create a nonequilibrium steady state

Testing of physical validity in molecular simulations is of great importance

Advances in recent years made MM simulations a *widely available and diversely used technique* (also by non-experts!)

Interpretation of MM results relies on the *validity of underlying physical assumptions* (thermodynamic boundary conditions, correct integration, …)

Main sources of violations include

- Unphysical models
- Incompatible choices of parameters
- Code bugs

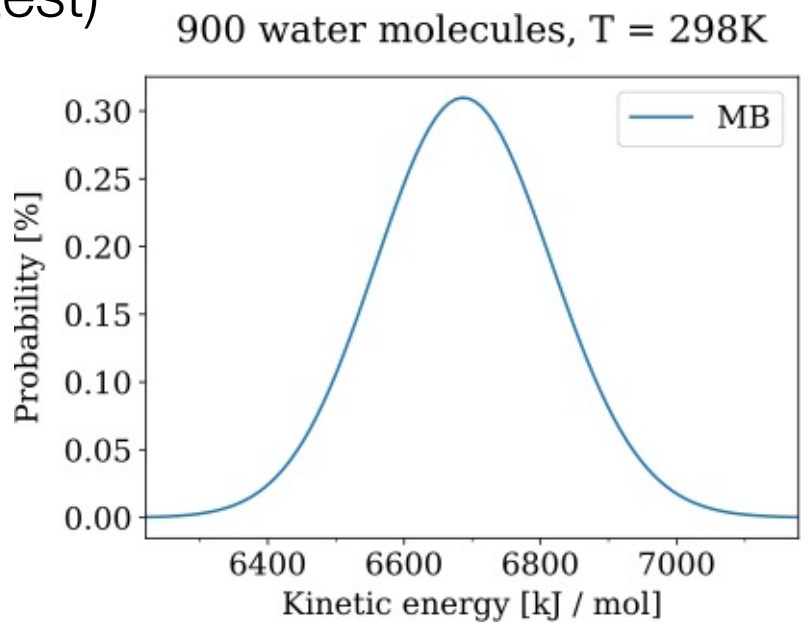
Need for robust and simple-to-use physical validation set, aimed at developers and end users!

Test 1: Is the kinetic energy Maxwell-Boltzmann distributed?

- *Physical law:* Kinetic energy K is Maxwell-Boltzmann (MB) distributed:

$$P(K) \propto K^{\frac{N-2}{2}} e^{-\beta K}$$

- MB distribution is a χ^2 distribution → statistical tests available (*e.g.* Kolmogorov-Smirnov test)
- K is a trajectory / distribution
- Null hypothesis:
  K is MB distributed
- Given confidence level α, what is the probability the null hypothesis is violated?



i-CoSE

Test 1: Is the kinetic energy Maxwell-Boltzmann distributed?

- *Physical law:* Kinetic energy K is Maxwell-Boltzmann (MB) distributed: $$P(K) \propto K^{\frac{N-2}{2}} e^{-\beta K}$$

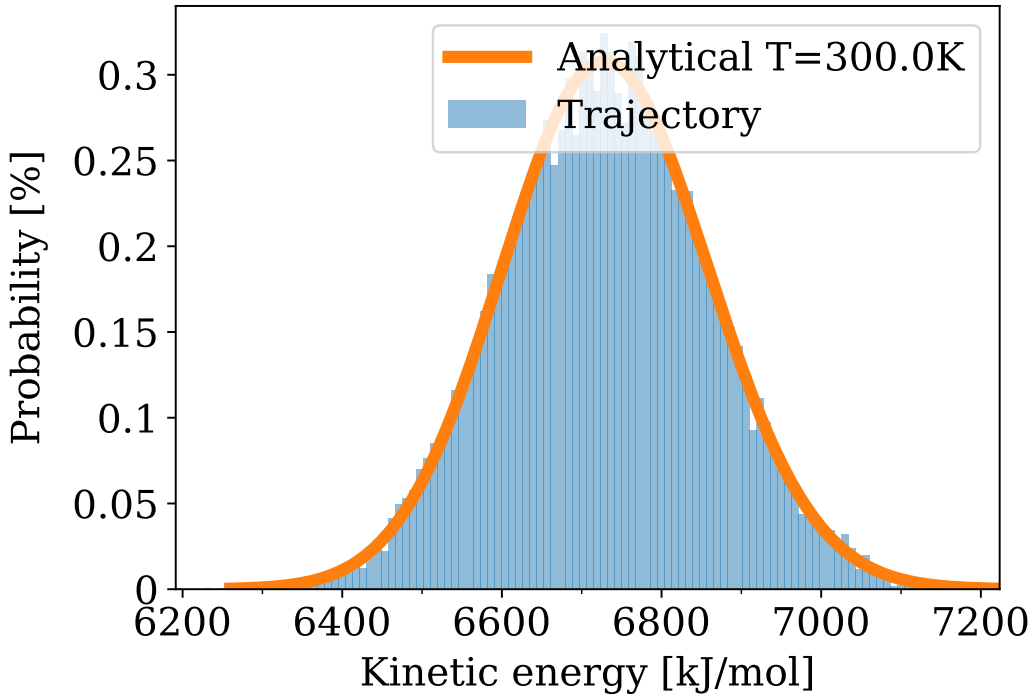
Simpler test, that can be less dependent on noise
Average KE must be 1/2 $k_B T$ x [# degrees of freedom]
Variance of KE must be: 1/2 $(k_B T)^2$ x [# degrees of freedom]

A given distribution of observed kinetic energy implies two temperatures: T_μ and T_σ

Are these both consistent with $T_{\text{what I meant to run}}$?

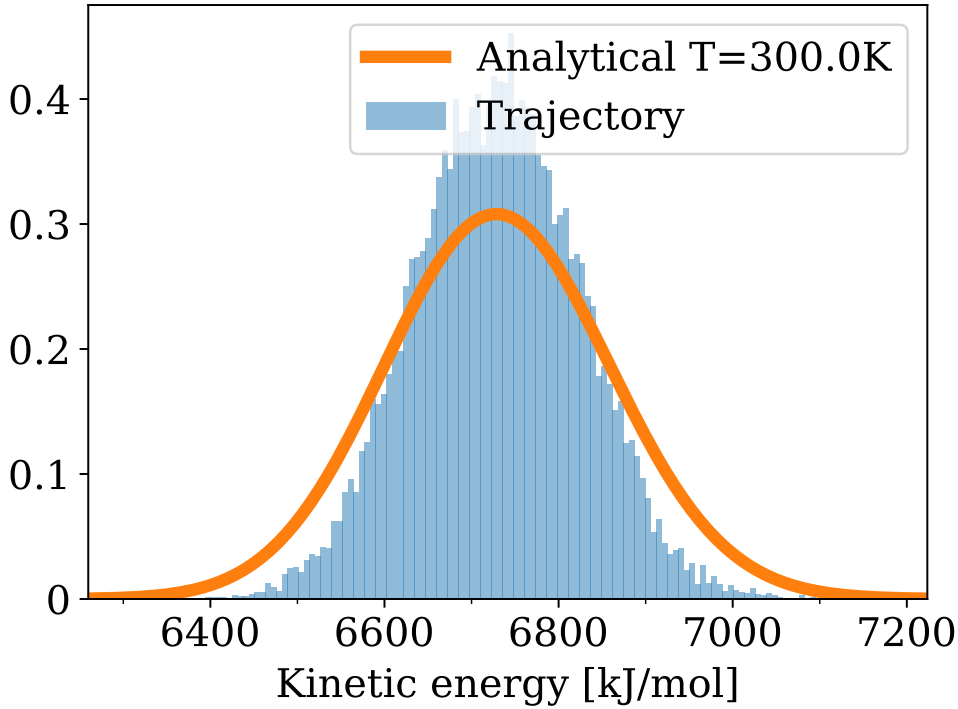
Example: Kinetic energy distributions

Velocity rescale thermostat



T_μ = 299.94 ± 0.04 K
T_σ = 301.01 ± 1.47 K
p = 0.214045

Berendsen thermostat



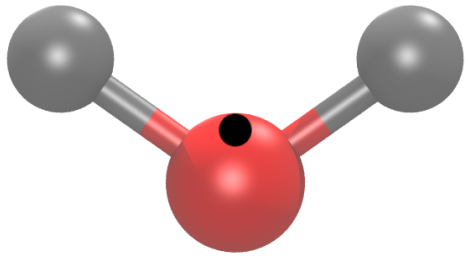
T_μ = 299.99 ± 0.03 K
T_σ = 222.60 ± 1.16 K
p = 8.42116×10^{-83}

i-CoSE

Test 1: Does equipartition hold for the kinetic energy?

Equipartition expected – homogeneous temperature!

- Parts of the system: Functional / arbitrary divisions
  - Randomly divide molecules in groups
  - Compare solute / solvent temperatures
  - Compare temperatures of components of liquid mixture
- Sets of degrees of freedom (DoF)
  Rigid body / internal DoF



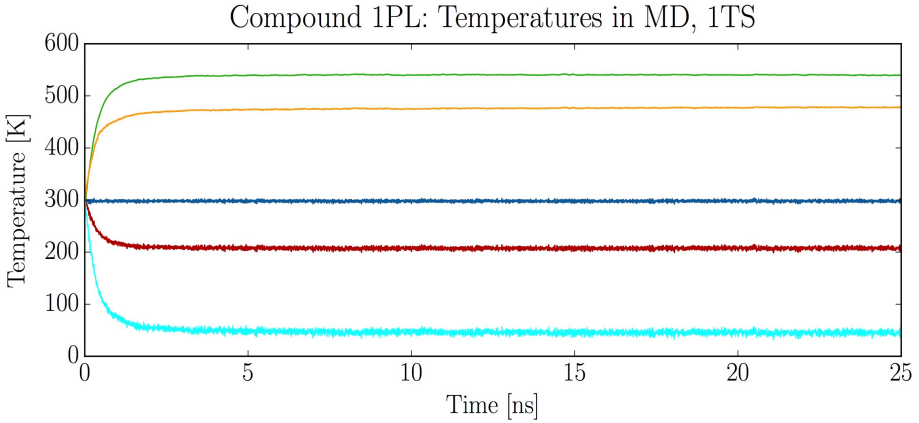
Jellinek & Li, Phys Rev Lett (1989)

i-CoSE

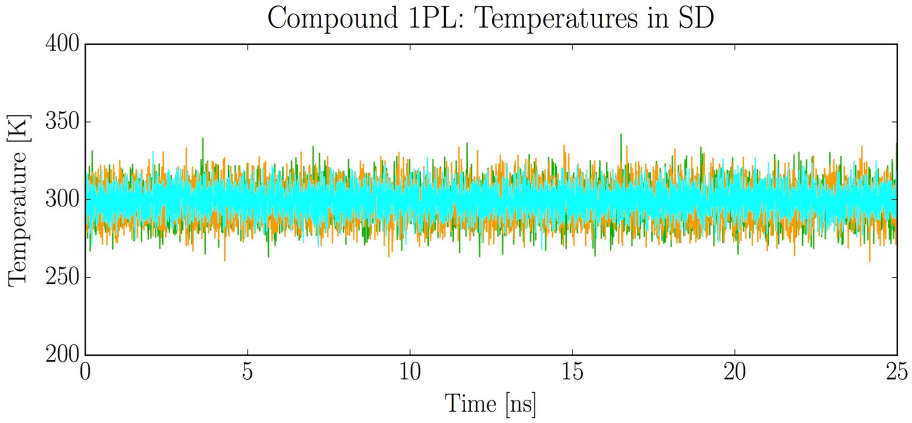
Equipartition, MD vs SD : Why are "ideal gas" potential energies so different?

Gas phase estimate on small organic molecule test set
Sparsely distributed "ideal gas" (distances ≫ cut-off)

Nosé-Hoover NVT



Langevin NVT



Scaling isolated oscillators (TI, inhomogenous systems, …)
by a single thermostat transfers heat to slow motions!

T_{tot} T_{tra} T_{rpi} T_{rot} T_{int}

Merz, PhD Thesis (2016)
Harvey, Tan & Cheatham, JCC (1998)

i-CoSE

Test 2: Is the expected statistical-mechanical ensemble (NVT / NPT / μVT) sampled?

- We need to do some statistical mechanics first.
- We know that we each configuration should appear *proportional* to $e^{-\beta U(x)}$, where $\beta = 1/k_B T$.
  - Low energies are more common
  - High energies are less common
  - The difference is more pronounced at low T, and less pronounced at high T.

- What is the probability of each configuration, where we insist that it's a true probability, so that $\sum P(x) = 1$?

Some statistical mechanics

- What is the actual probability of each configuration x?
- What if we sum up all the P(x), for all the x, then we divide by ∑P(x)?
  - Then, the probabilities will be normalized!
- Example:
  - I have 3 events with probability proportional to 1/10, 1/5, and 1/2.
  - 1/10+1/5 +1/2 = 8/10
  - Then divide the three by 8/10 and we get: 1/8, 1/4, and 5/8. Which is normalized (i.e. adds to 1)

Some statistical mechanics

- If it's a continuous space, we have to integrate. We need to integrate over all coordinates, and all velocities, since the total energy is potential energy U(x) + K(v).
- Probabilities are normalized if: $\frac{P(x)}{\int P(x)dx}$
- Call this integral Q = ∫P(x)dx, the **partition function**,
- and call A = -k_BT ln Q. The **free energy**!
- Then P(x) = $e^{\beta A - \beta U(x)}$. Note the **equals**.

Some statistical mechanics

- The partition function is:
- $Q = \int e^{-\beta U(x)} dx$
- The partition function can also be written as:
- $Q = \int \Omega(U) e^{-\beta U} dU$
- Now, Instead of summing over all states, we sum over energies TIMES Ω(U), the number of configurations with a given energy
- Summing over seconds is the same as summing over days TIMES the number of seconds each days.
- SO?? I don't know Ω(U) . . . we don't need to!

Test 2: Is the expected statistical-mechanical ensemble (NVT / NPT / μVT) sampled?

Run the same system, same options,
except at two different temperatures

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

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

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

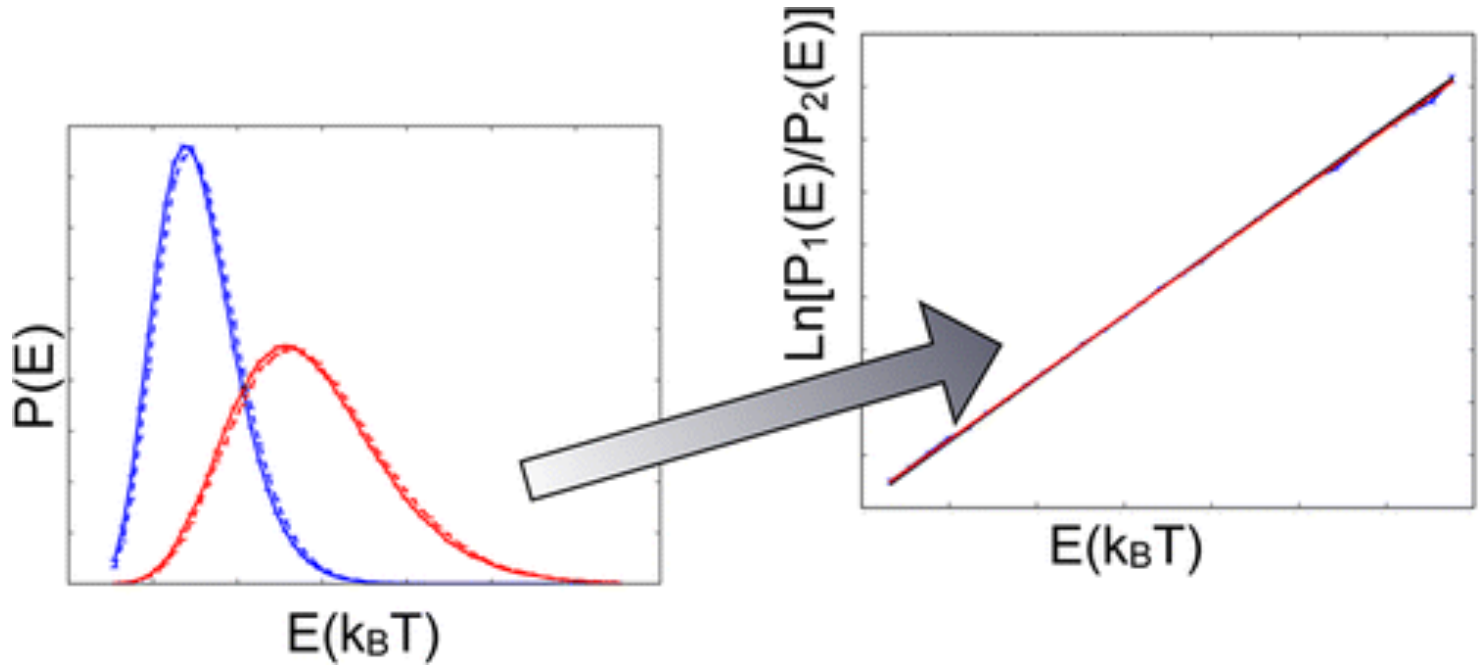
$$\ln\frac{P_1(E)}{P_2(E)} = \ln\frac{Q_2}{Q_1} + (\beta_2 - \beta_1)E$$

i-CoSE

Test 2: Is the expected statistical-mechanical ensemble (NVT / NPT / μVT)

Unknown: Analytical distribution of potential energy U, volume V, and/or chemical potential μ

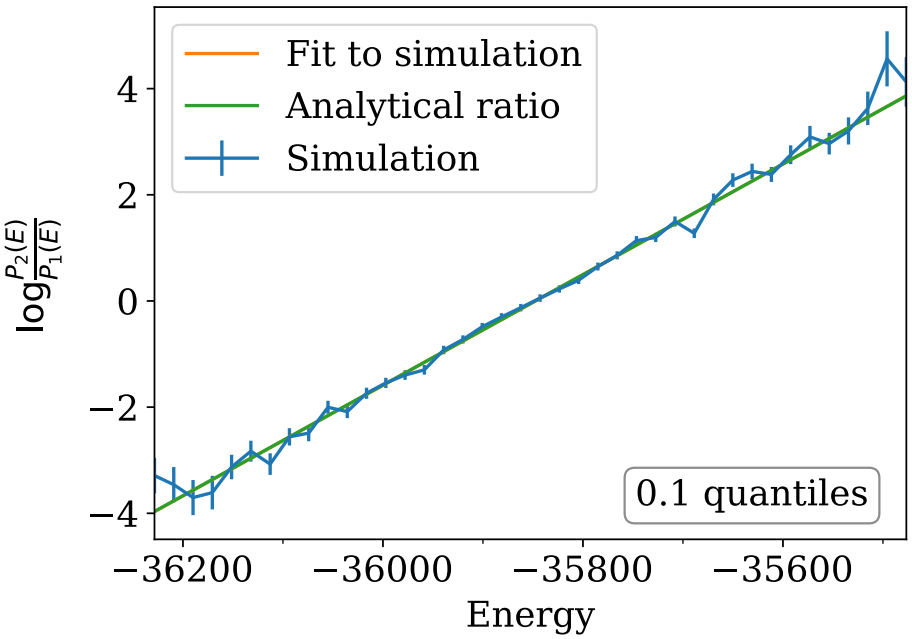
Known: Ratio of these distributions between two state points (T, P, μ)



i-CoSE

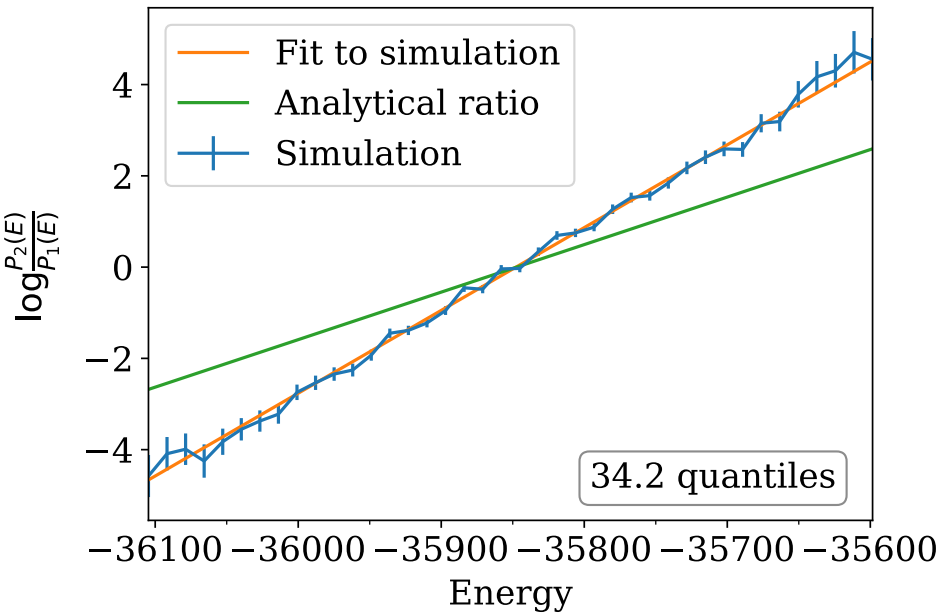
Example: ensemble checking of water potential energy

Velocity rescale thermostat



Analytical: $\beta_1/\beta_2 = 0.010413$
Slope of $P_1(E)/P_2(E) = 0.010420 \pm 0.000127$

Berendsen thermostat



Analytical: $\beta_1/\beta_2 = 0.010413$
Slope of $P_1(E)/P_2(E) = 0.01815 \pm 0.00023$

i-CoSE

Most thermodynamic ensembles can be checked

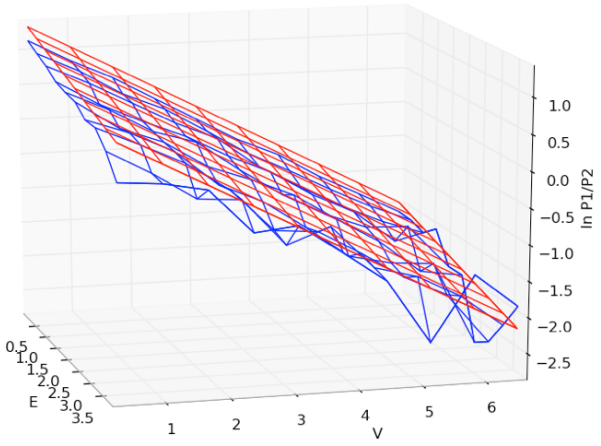
- NVT → Fixed ratio of U distributions at two T's
- NPT → Fixed ratio of V distributions at two P's

  Fixed ratio of H = U+PV at two T's

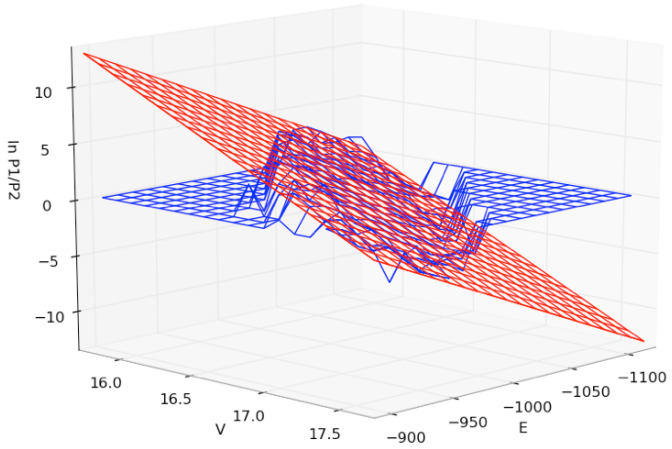
  Fixed plane of U,V joint distribution at T,P

Ratio of
$P_1(U,V)/P_2(U,V)$

Blue = Data
Red = Fit



Model with K(V)



Lennard-Jones fluid

i-CoSE

Monte Carlo can be checked

- Potential energy and kinetic energy are separable
- So formula is still valid if E = V, instead of E = U
- Works for Monte Carlo, too.

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

Grand canonical and semigrand canonical ensembles can also use the same theory

- μVT → Fixed ratio of N distributions at two μ's
  Fixed ratio of U - μN at two T's
  Fixed plane of (N,μ) joint distribution at pairs of T,μ
- ΔμVP → Fixed ratio of N_1 distributions at two Δμ's
  Fixed ratio of U - ΔμN_1 at two T's
  Fixed plane of (U, N_1) joint distribution at pairs of T,Δμ

Python module: `physical-validation`

Tests:

```
kinetic_energy.py
  distribution()
  equipartition()

ensemble.py
  check()

integrator.py
  convergence()
```

Class representing simulation results, communicates with tests

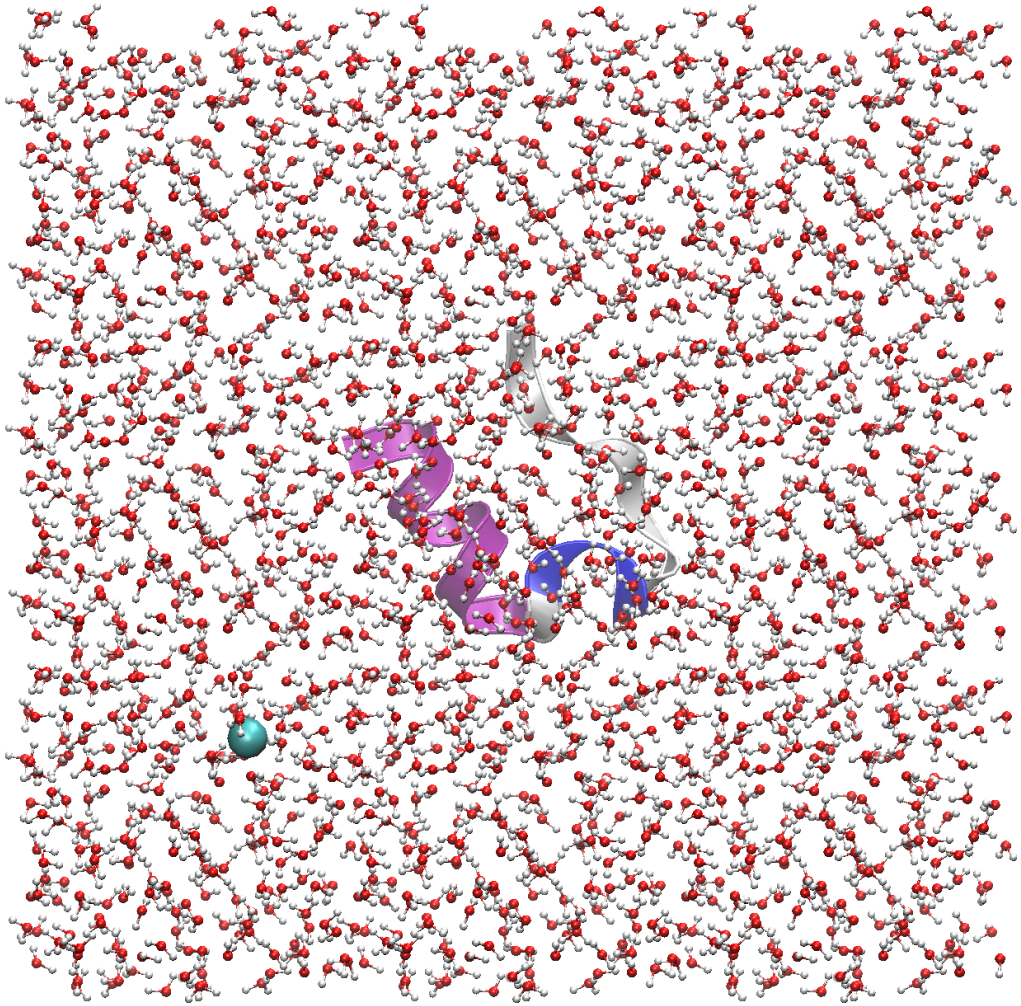
`SimulationData`

Parsers:

- GROMACS
- LAMMPS
- plain text

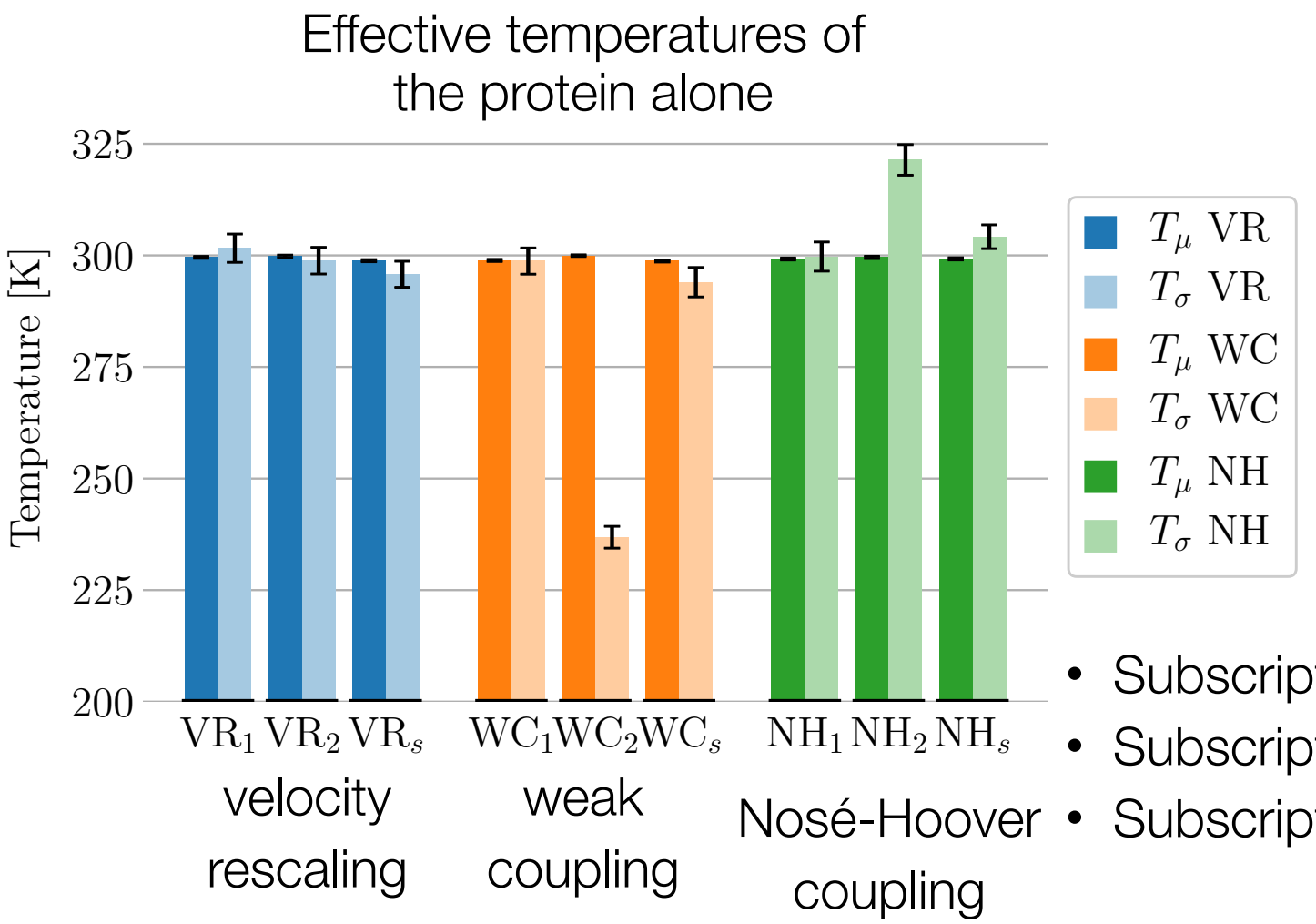
Equipartition: A solvated protein

- Trp-cage miniprotein
- Solvated in water
- Different thermostats, different coupling schemes:
  - system
  - protein, solvent separately
  - solvent alone



i-CoSE

How do multiple thermostats on the system affect the KE distribution of the protein?



- Subscript 1 = (system)
- Subscript 2 = (protein/solvent)
- Subscript S = (solvent only)

i-CoSE

When to use physical-validation

- Use it when you have a new method or procedure to check to make sure it behaves properly!
- You can check kinetic energy distributions with no extra work, and check ensembles with just 1 extra simulation!