Testing for Physical Validity in Molecular Simulations

Michael R. Shirts

Department of Chemical and Biological Engineering
University of Colorado Boulder

Reproducibility of Molecular Modeling and Simulation Workshop
FOMMS 2018 July 19, 2018

Acknowledgements

Pascal Merz!!

You?

Can Pervane, U Southampton

Peter Kasson, Eric Irrgang, U Virginia

NIH National Institutes of Health
Turning Discovery Into Health

Grant R01GM115790

Outline

- Some physical assumptions and their validation
- Python module **`physical-validation`**
- PEBCAK
- Outlook & Conclusion

Downloads for tutorials

- Git clone data from:
- `https://github.com/shirtsgroup/physical_validation_workshop`
- OR download from: http://bit.ly/2LhjrEj
- ALSO run:
- `pip install physical_validation`
- (may need to do `pip install numpy scipy matplotlib`)

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*

Physically correct simulations are necessary for reproducibility

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

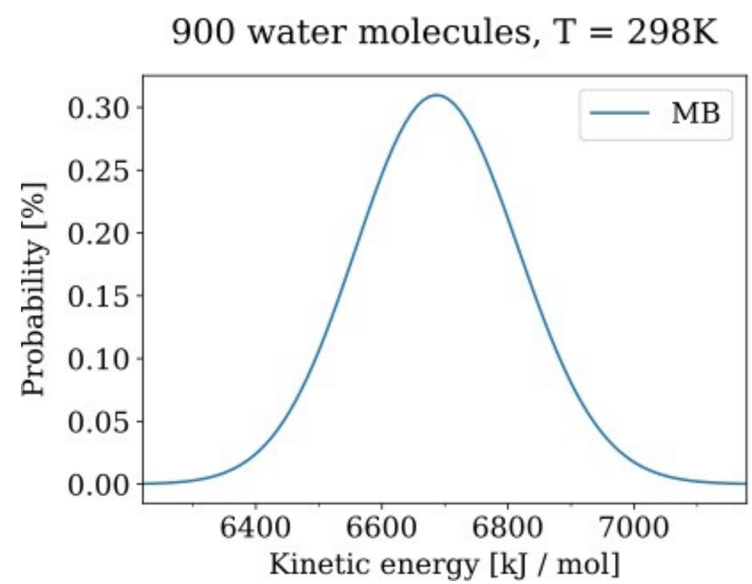
- There are many ways to not conserve energy
- There are many ways to not have equipartition of energy
- There are many ways to not be Boltzmann weighted
- There are many ways to create a steady state

Test: Is the kinetic energy Maxwell-Boltzmann distributed?

- *Expectation:* 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?



University of Colorado Boulder

Be Boulder.

Test: Is the kinetic energy Maxwell-Boltzmann distributed?

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

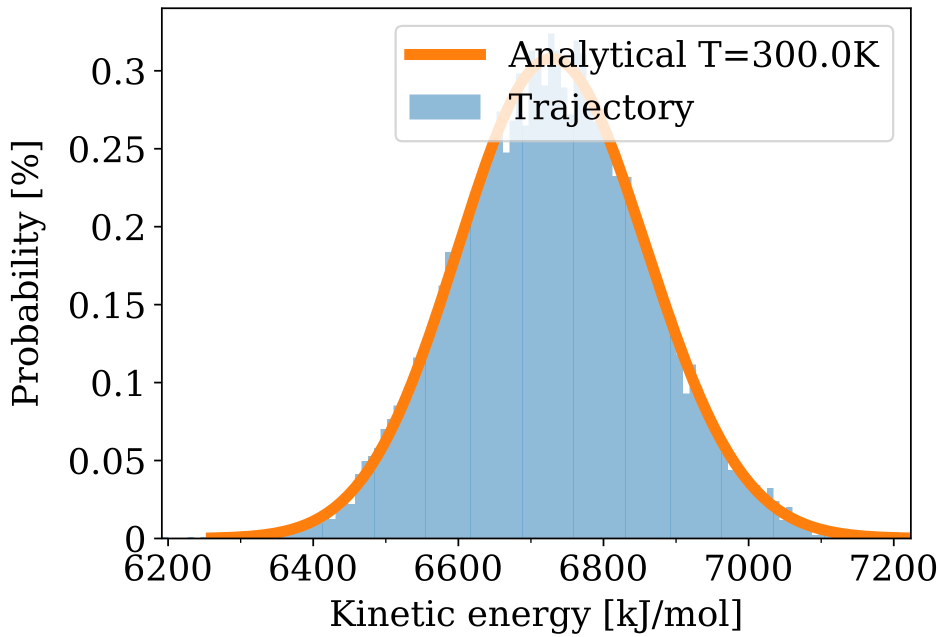
$$P(K) \propto K^{\frac{N-2}{2}} e^{-\beta K}$$

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 <u>observed</u> 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

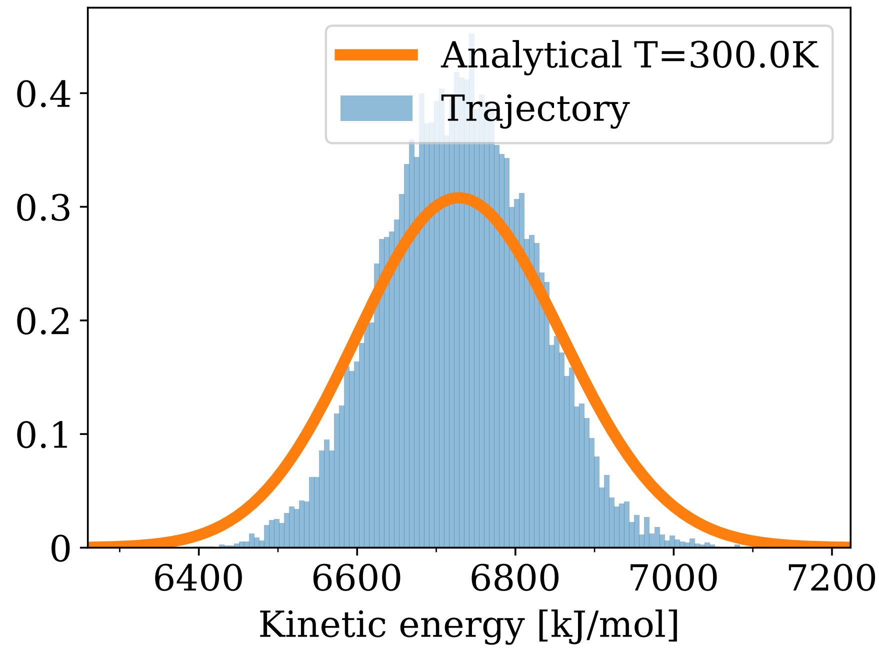


$T_\mu = 299.94 \pm 0.04$ K

$T_\sigma = 301.01 \pm 1.47$ K

$p = 0.214045$

Berendsen thermostat



$T_\mu = 299.99 \pm 0.03$ K

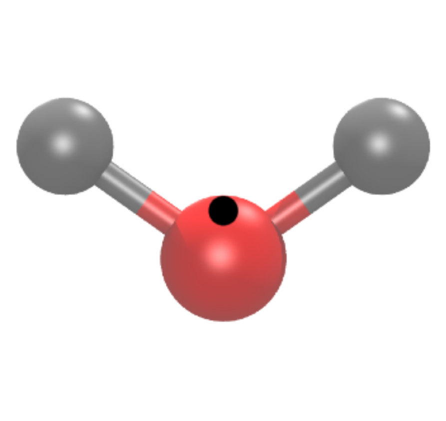
$T_\sigma = 222.60 \pm 1.16$ K

$p = 8.42116 \times 10^{-83}$

Test: 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)

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

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

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

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

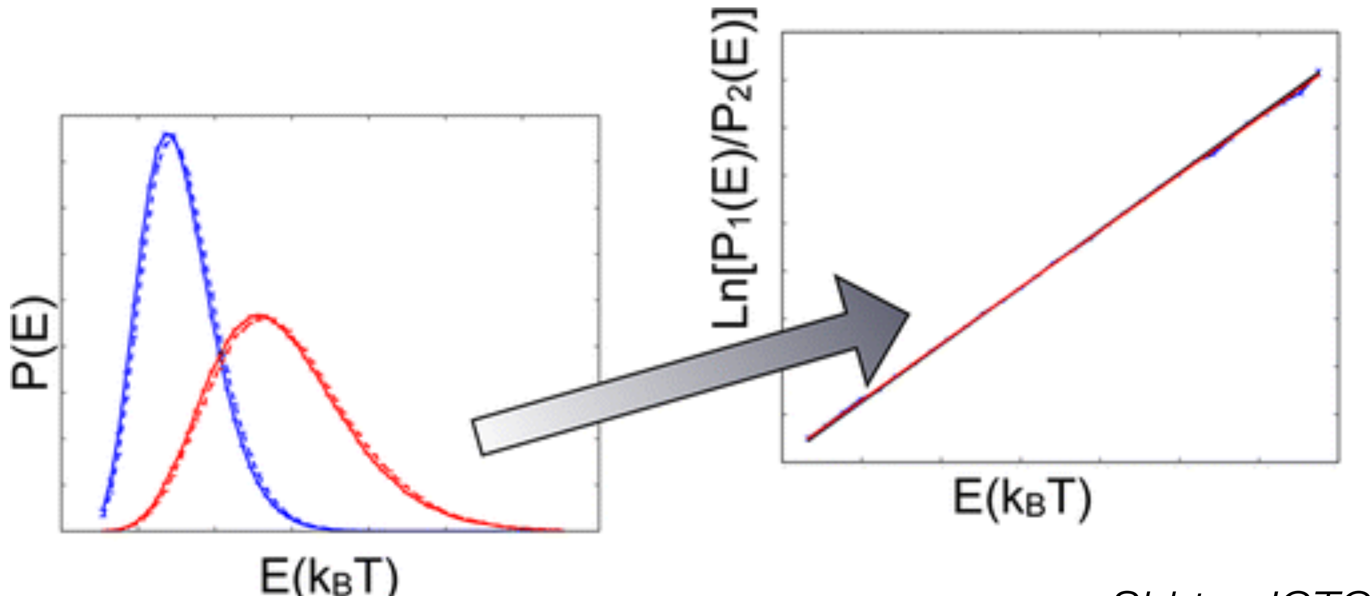
$$\frac{P_1(E)}{P_2(E)} = \frac{Q_2}{Q_1}e^{(\beta_2-\beta_1)E}$$

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

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

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

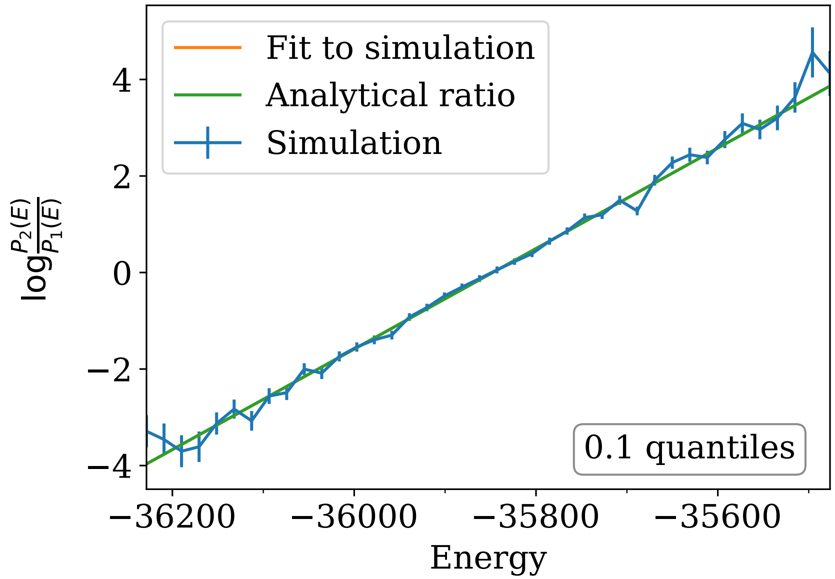
  Known: Ratio of these distributions between two state points (T, P, μ)



Shirts, JCTC (2013)

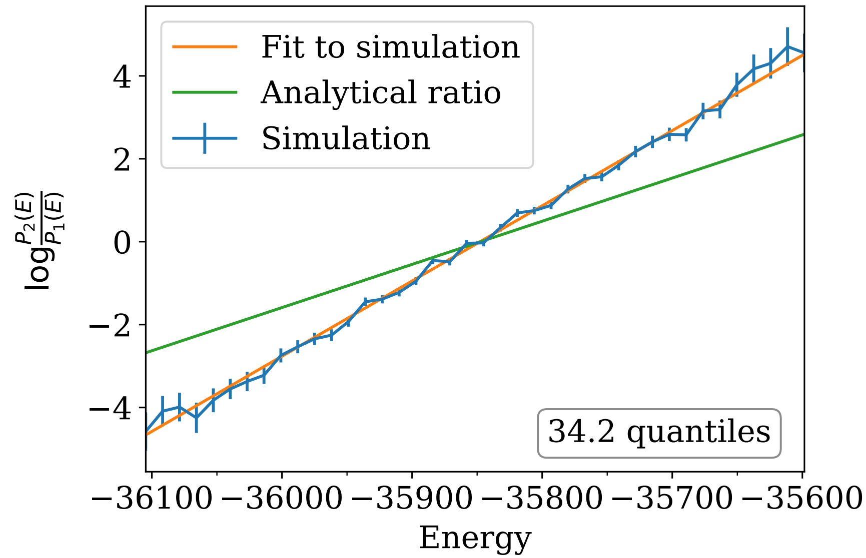
Example: ensemble checking of water potential energy

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 $P_1(E)/P_2(E) = 0.01815 \pm 0.00023$

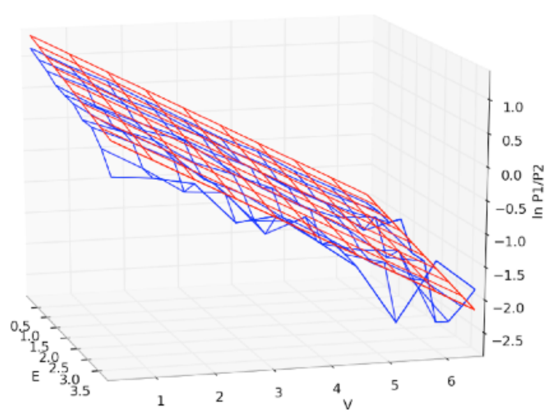
Most thermodynamic ensembles can be checked

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

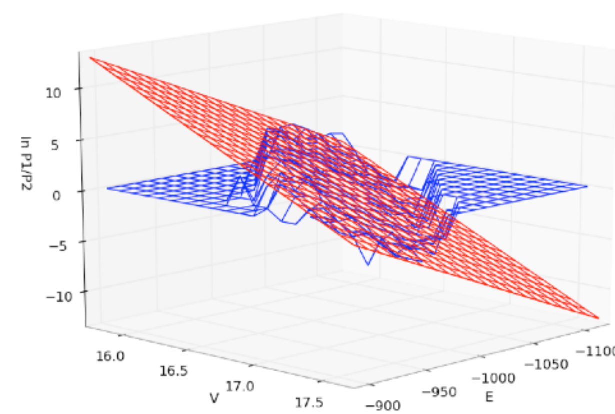
  Fixed ratio of U+PV at two T's

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

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



Model with K(V)



Lennard-Jones fluid

Blue = Data
Red = Fit

University of Colorado
Boulder

Be Boulder.

Monte Carlo can be checked

- Potential energy and kinetic energy are separable
- So formula is still valid if E = V, instead of E=U

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

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

- $\mu VT \rightarrow$ Fixed ratio of N distributions at two μ's
  Fixed ratio of $U - \mu N$ at two T's
  Fixed plane of (N,μ) joint distribution at pairs of T,μ
- $\Delta\mu VP \rightarrow$ Fixed ratio of N_1 distributions at two Δμ's
  Fixed ratio of $U - \Delta\mu N_1$ at two T's
  Fixed plane of (U, N_1) joint distribution at pairs of T,Δμ

Support for grand and emigrant canonical still in process in current version, but if you want to get it working let us know

Test: Is my integrator doing what I think it is?

- Symplectic NVE integrators do not sample the true Hamiltonian, but a related "shadow Hamiltonian"

$$H = \tilde{H} + \mathcal{O}(\Delta t^2)$$

- Fluctuation δE of constant of motion H(x) = E depend on time step ΔE of integration

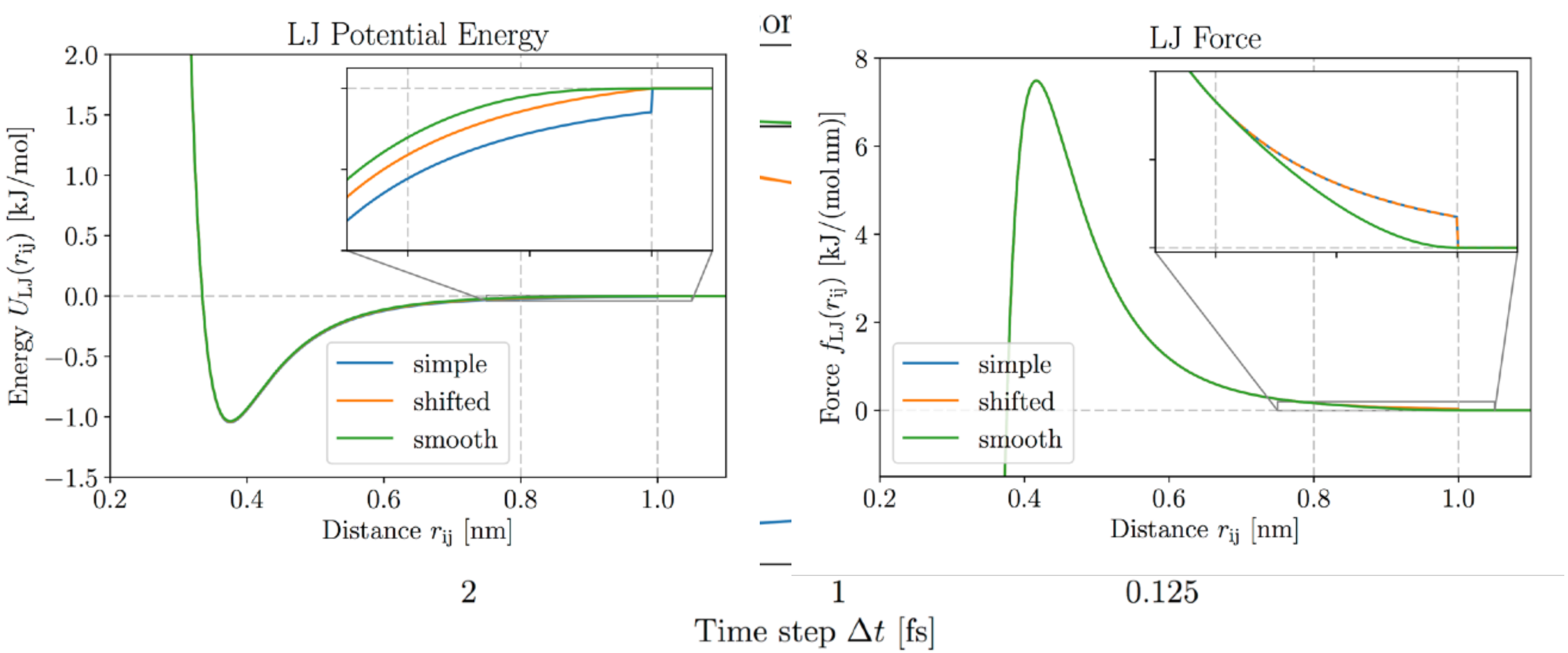
$$\delta E \propto \Delta t^2 \rightarrow \left(\frac{\Delta t_1}{\Delta t_2}\right)^2 = \frac{\delta E_1}{\delta E_2}$$

- non-quadratic behavior indicates lack of symplecticity
  - Integrator algorithm
  - Interaction function
  - Constraint tolerance
- **Powerful bug detection**

Gans & Shalloway, Phys Rev E (2000)

Example: Does my cutoff-scheme violate symplecticity?

- Pure LJ liquid, exact pair list
- Δt = [4,2,1,0.5,0.25,0.125] fs
- 1 nm cutoff: cutoff, shifted, "smooth"



Python module: **`physical-validation`**

Two major parts:

Tests

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

ensemble.py
  check()

integrator.py
  convergence()
```

Go to code!

Class representing simulation results, communicates with tests

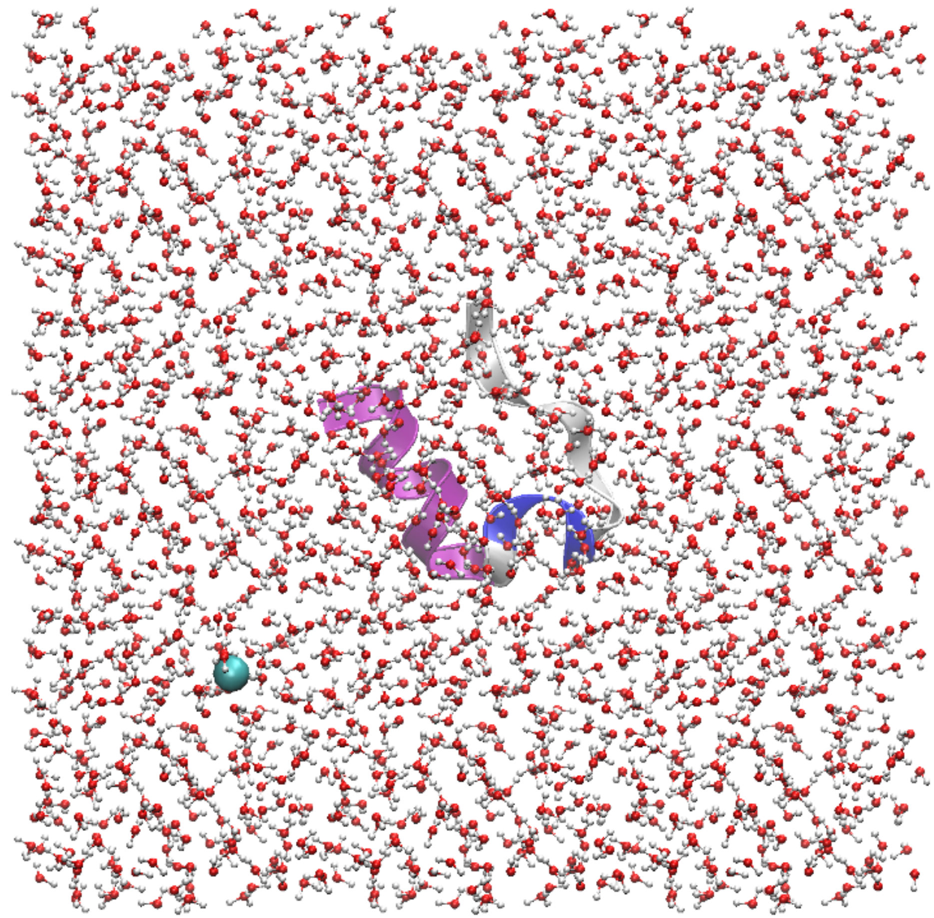
`SimulationData`

Parsers:

- GROMACS
- LAMMPS (beta)
- Flat-file or numpy array
- More to come!

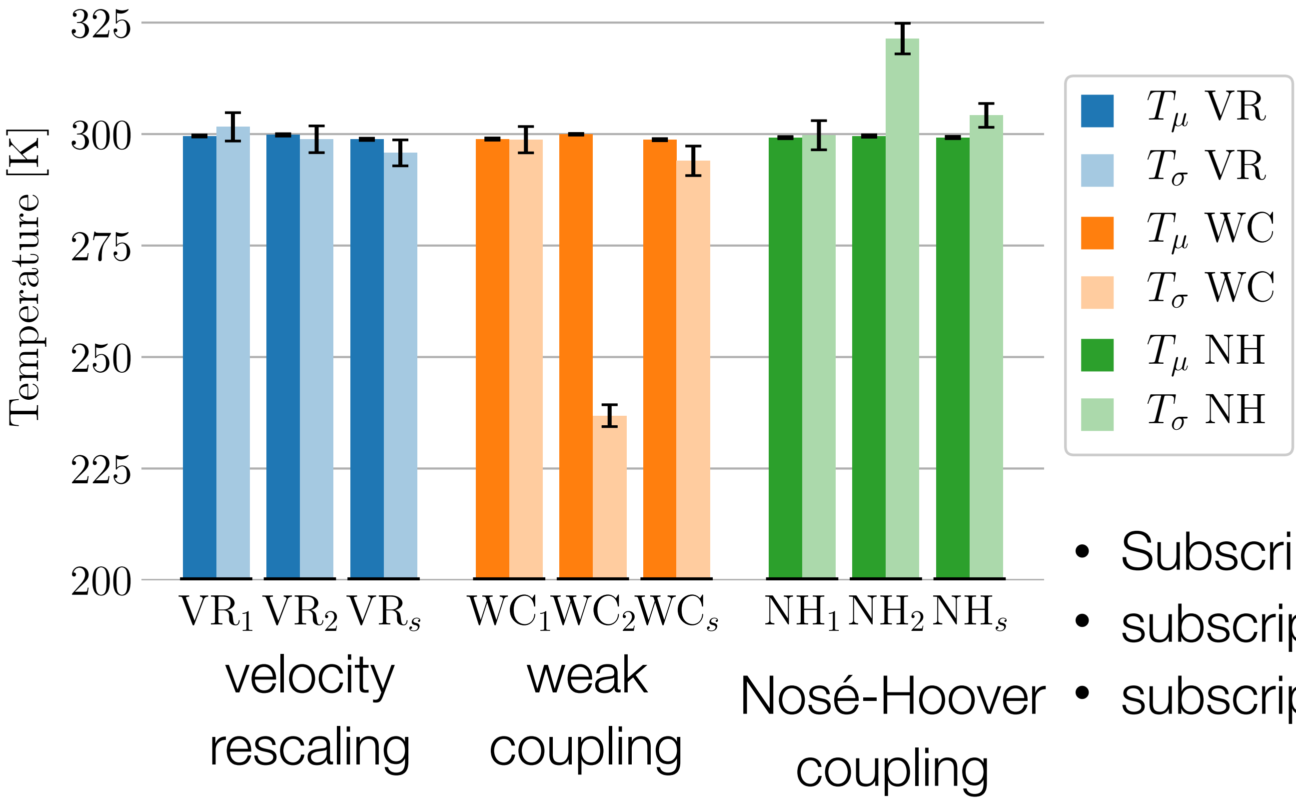
Equipartition: A solvated protein

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



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

Effective temperatures of the protein alone



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

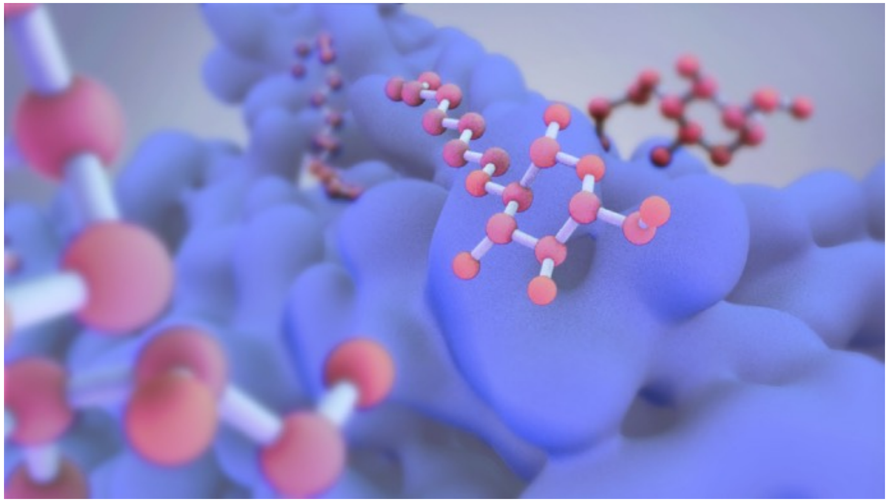
University of Colorado Boulder

Be Boulder.

PEBCAK: What about the people?

Problem exists between chair and keyboard

The Living Journal of Computational Molecular Science
http://www.livecomsjournal.org



- A journal for articles that can and *should* be updated
- Article types include best practices for molecular simulation, high quality tutorial, and perpetual reviews
- Will help improve reproducibility and quality of simulations eliminating error
- Please look for the first issue this fall, and consider submitting!

Outlook and Conclusions

- Wide usage of molecular simulation calls for simple-to-use physical validation tools that can be automated.
  - Integration in GROMACS testing to harden code base
- We can validate kinetic energy, ensemble and integrator
- **`physical-validation`** analysis, useable for many codes, at http://physical-validation.readthedocs.io.
- Interested in extensions? → michael.shirts@colorado.edu

Make physical validation a standard MM sanity check for developers and users!