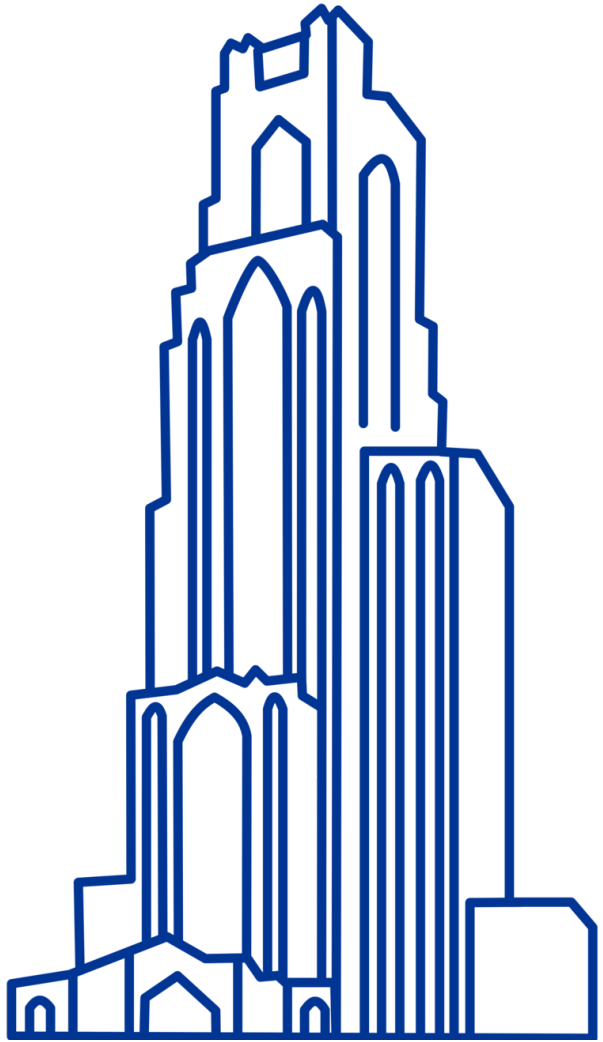


Computational Biology
(BIOSC 1540)

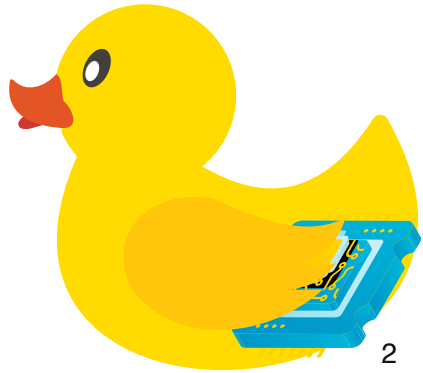
Lecture 15:
Ensembles and
atomistic insights

Oct 24, 2024

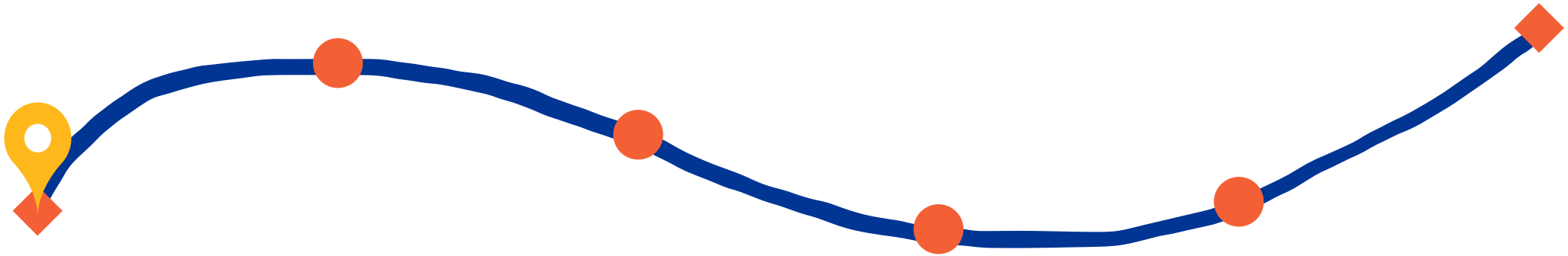
University of Pittsburgh

Announcements

- **No class on Nov 5** for election day
- A05 is **due tonight** by 11:59 pm
- A06 will be **released tomorrow**
- The next **exam is on Nov 14**
  - We will have a review session on Nov 12
  - Request DRS accommodations if needed



After today, you should better understand



Molecular ensembles

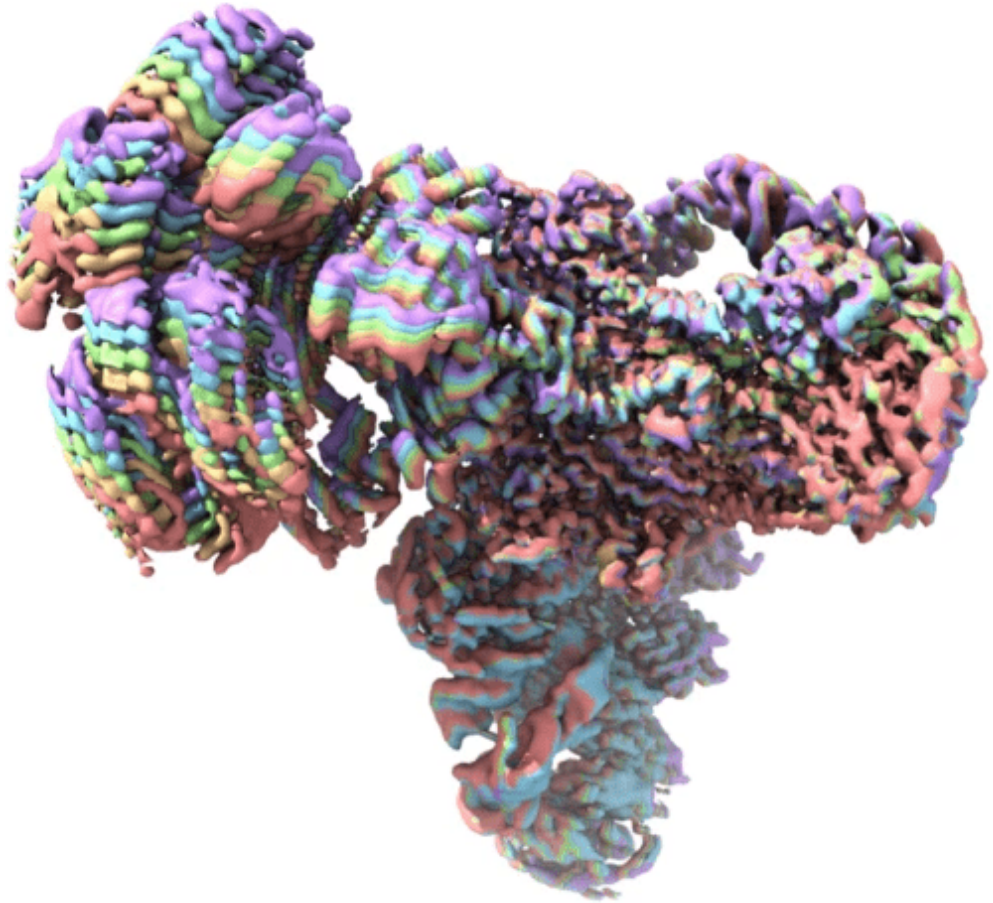
and their relevance

Physics is statistical at the molecular level is statistical

Number of Particles: Biological systems contain billions of atoms interacting simultaneously

Thermal Motion: Atoms and molecules are in constant motion due to thermal energy

Uncertainty and Variability: Exact positions and velocities of particles are inherently uncertain



Observable properties are averages of atomistic behaviors

Atomistic systems are **stochastic**, measurable properties are computed as averages

Microscopic level: Individual atoms and molecules

Macroscopic level: Bulk properties from collective behavior

Statistical mechanics: Uses statistical methods to relate microscopic properties to macroscopic observables

Relevance to biology: Helps in understanding the dynamics of proteins, DNA, and other biomolecules

What is a macrostate?

A macrostate specifies the temperature, pressure, volume, and number of particles of a molecular system

Example: Methanol and water

Temperature: 25 C

Pressure: 1.01325 bar

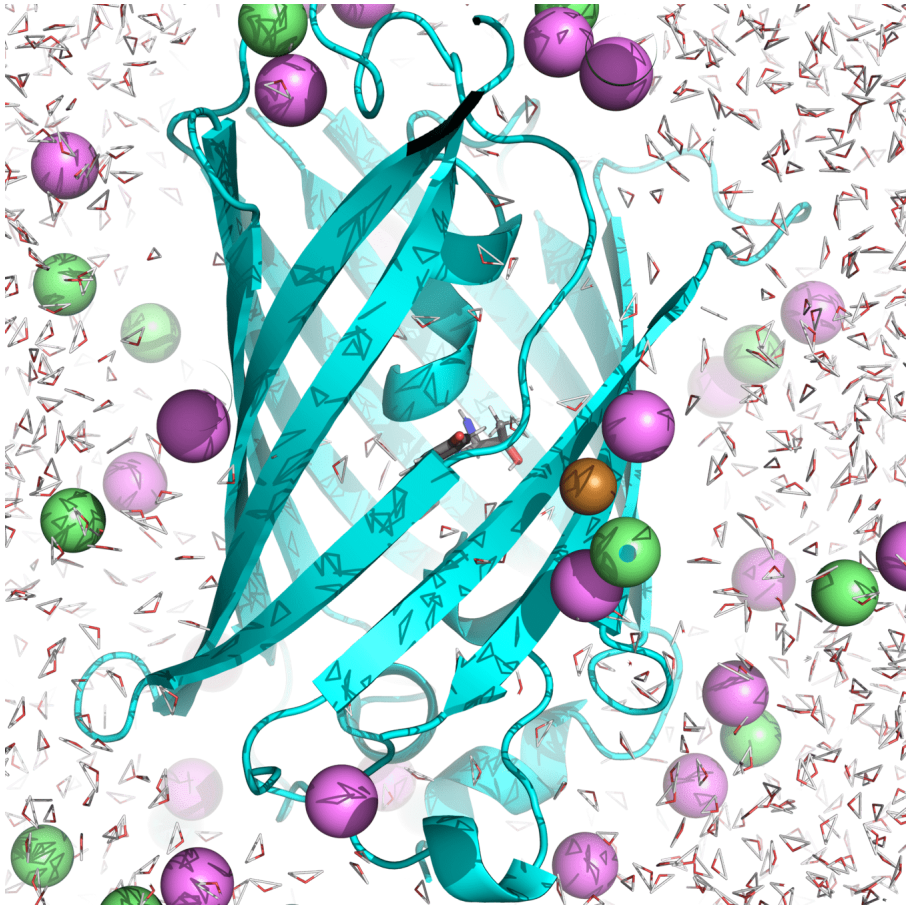
Volume: 100 mL

Composition: 70% methanol and 30% water by volume

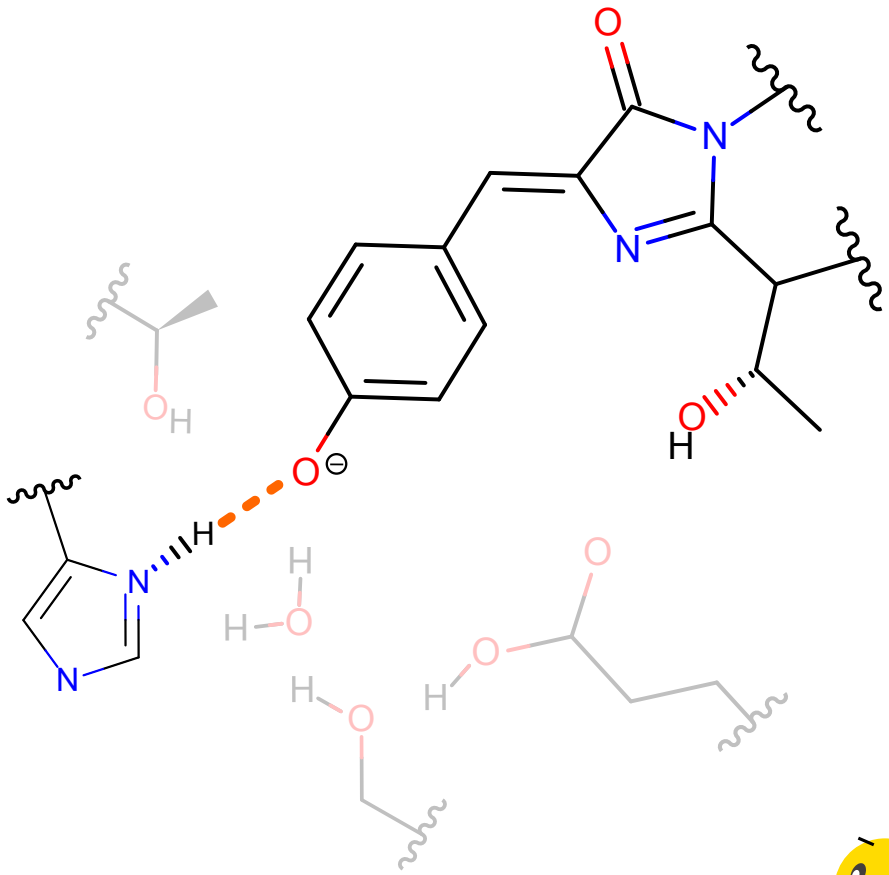
Changing any one of these values changes the macrostate

Ensemble example: roGFP2 hydrogen bonding

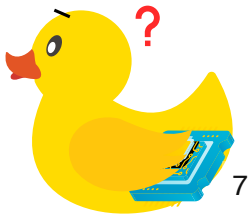
His148 in GFP stabilizes the anionic chromophore through a hydrogen bond



Let's use MD simulations to compute hydrogen bond length and energy

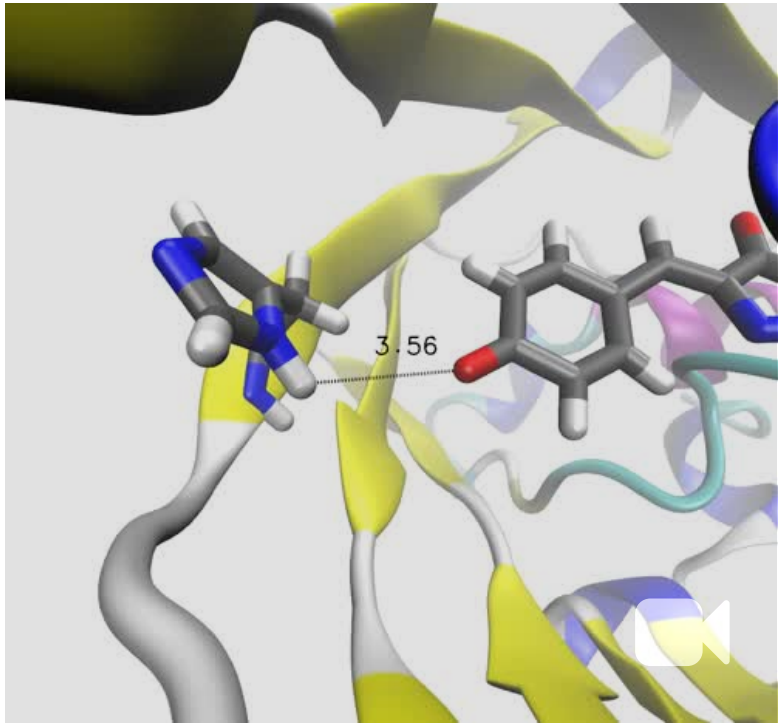


How would you approach this?



Compute the mean hydrogen bond length of the macrostate's ensemble

Our macrostate: roGFP2 in water, with 150 micromolar NaCl at 300 K and 1 atm



An **ensemble** is the collection of all possible microstates of a single macrostate

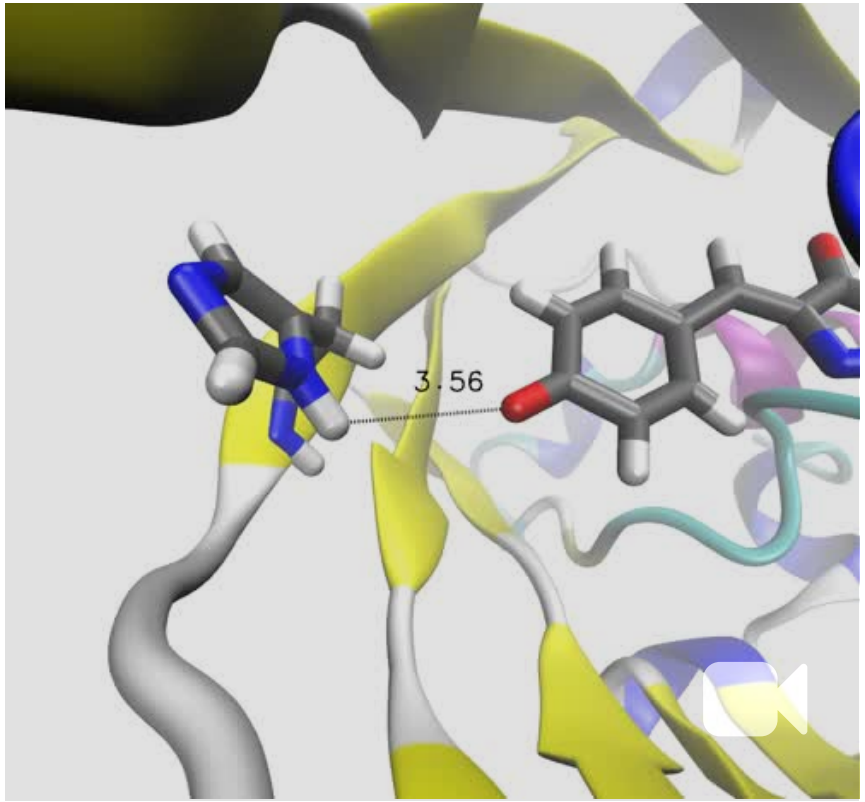
A **microstate** is a unique configuration defined by the positions and velocities of all particles

Here is the MD trajectory with a mean of 3.155 Å

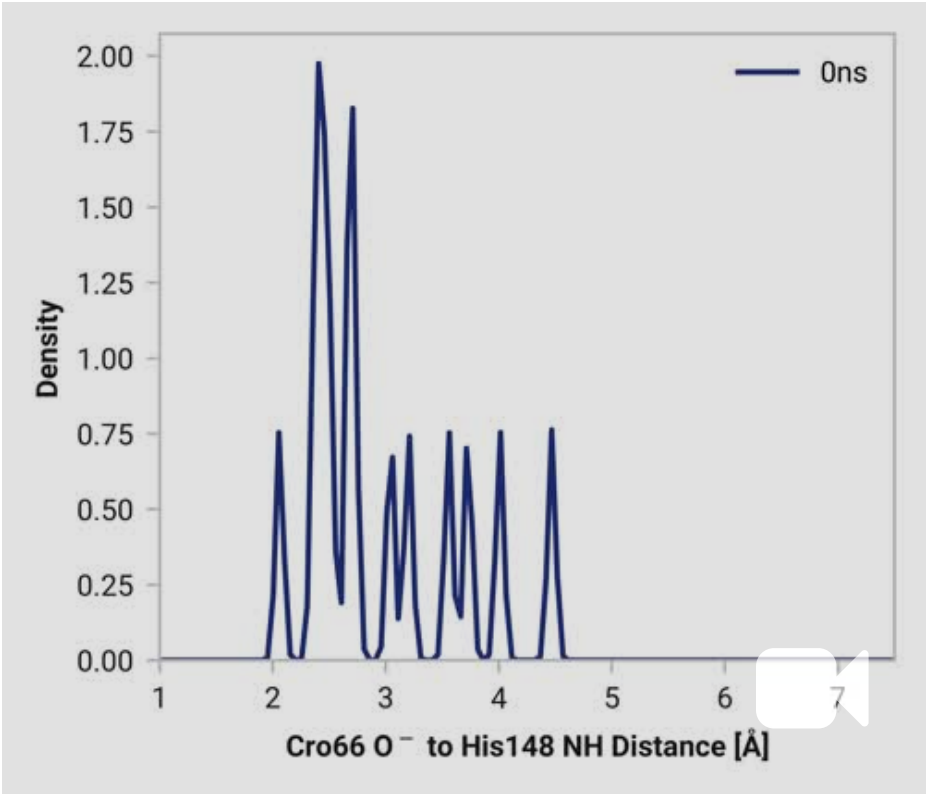
What is wrong with this? The MD simulation is extremely short

Accurate ensemble averages require sampling every possible microstate

Our previous MD simulation was very short

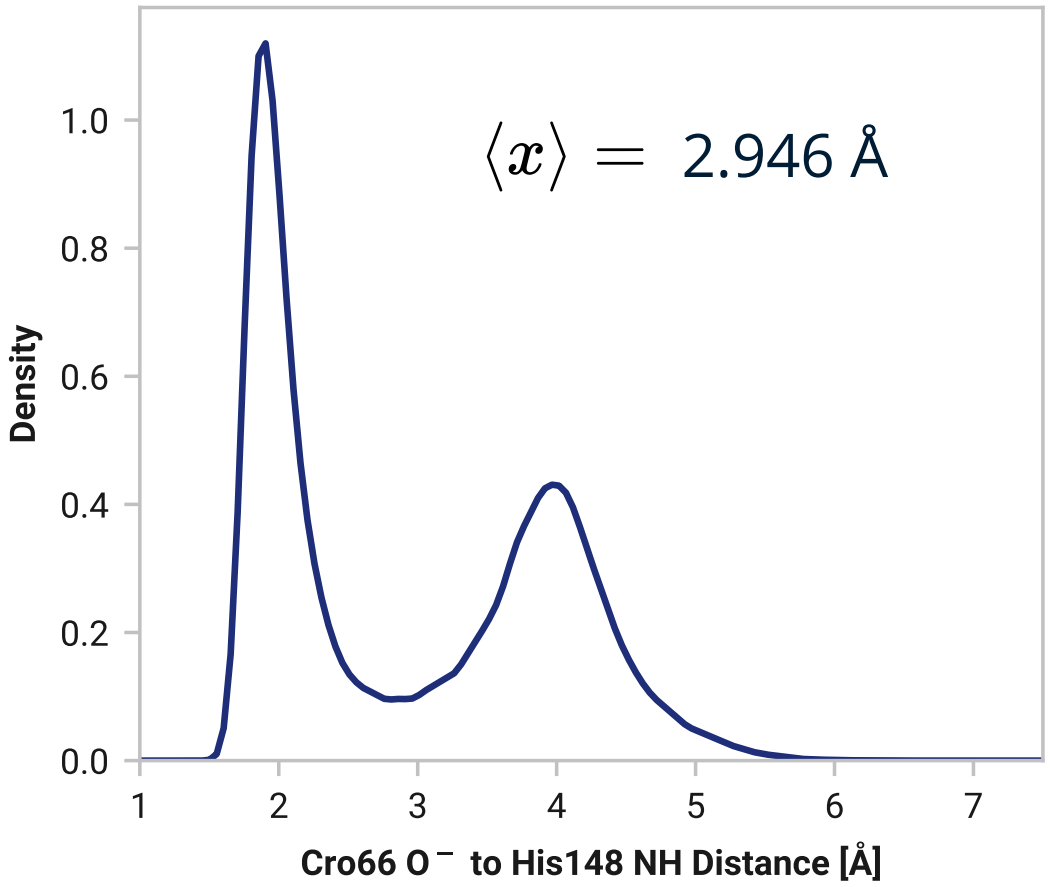


Longer simulations provide better sampling of microstates and their probabilities



More accurate hydrogen bond distance estimate!

Experiments measure the weighted mean of microstates



Remember: Multiple microstates (i.e., configurations) can have the same distance

We measure the ensemble probability of observing a microstate with value x_i

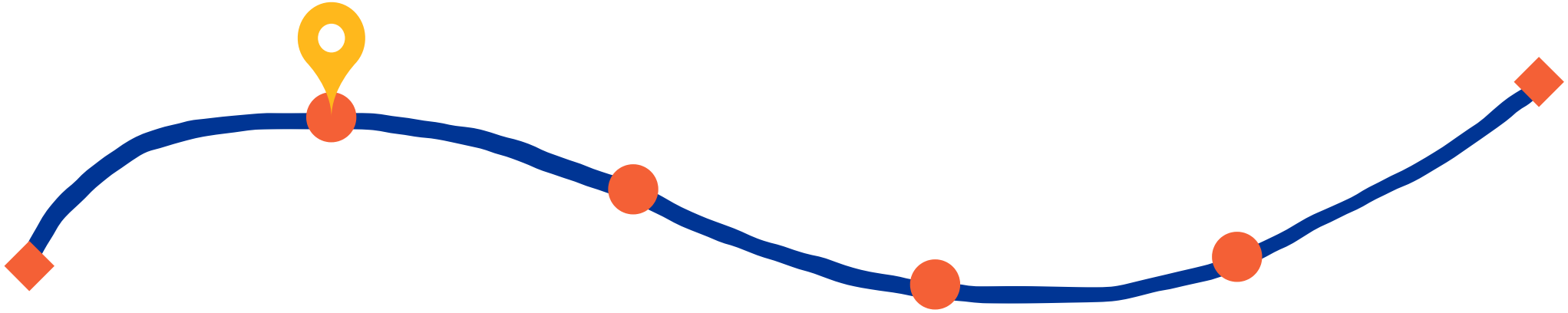
$$p(x_i)$$

Expected value of ensemble is computed by weighted mean

$$\langle x \rangle = \frac{\sum p(x_i) x_i}{\sum p(x_i)}$$

Note: Our denominator will always be 1 because we are not using actual partition function

After today, you should better understand



Maintaining thermodynamic equilibrium

Our molecular simulations need to reproduce the desired ensemble

Microcanonical Ensemble (NVE):

Fixed Number of particles (N), Volume (V), and Energy (E)

Canonical Ensemble (NVT):

Fixed Number of particles (N), Volume (V), and Temperature (T)

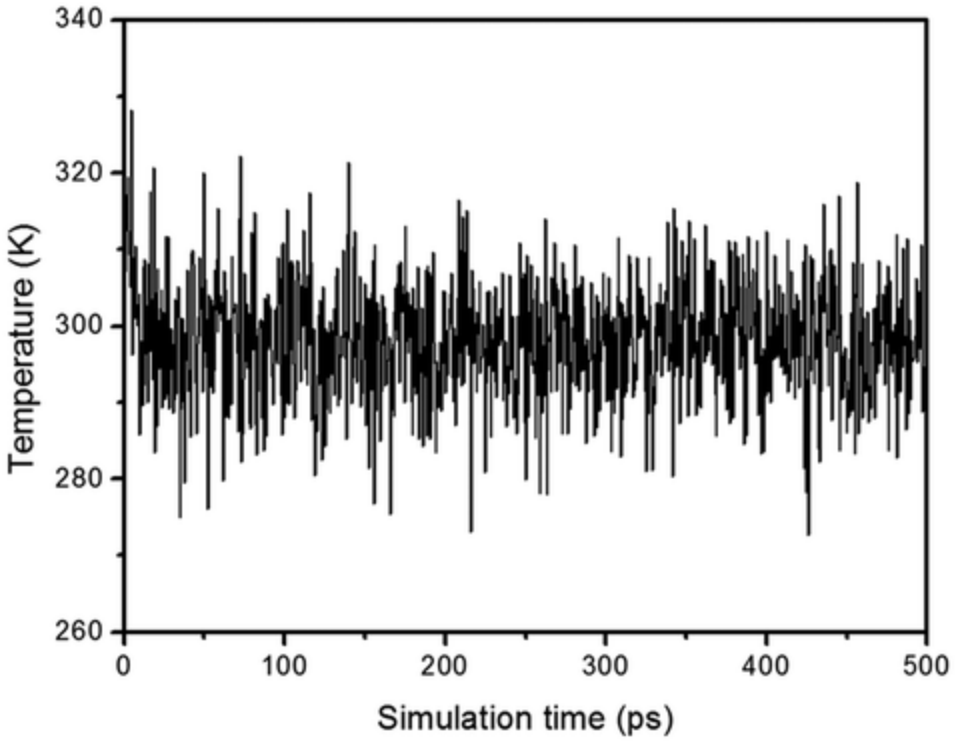
Isothermal-Isobaric Ensemble (NPT):

Fixed Number of particles (N), Pressure (P), and Temperature (T)

Most common

What does constant temperature mean?

Here is a plot of simulation temperature during a 500 ps MD simulation at 300 K



Is there something wrong with the simulation?

True:
Something is wrong

False:
Nothing is wrong

Talk with your neighbors on what could be wrong with the simulation

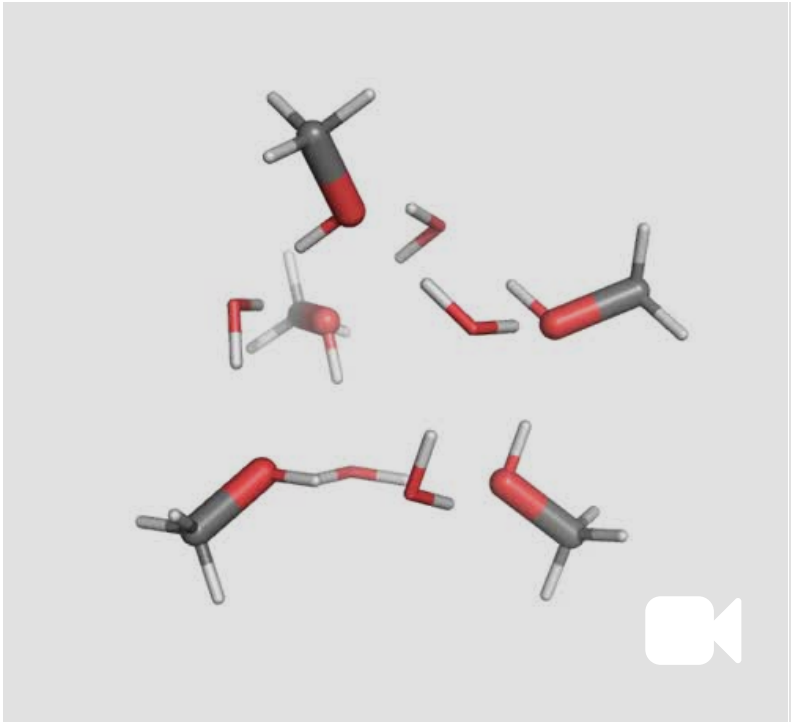
Remember: Macrostate observables are ensemble averages

The instantaneous temperature of microstates will fluctuate, but the ensemble average should be constant

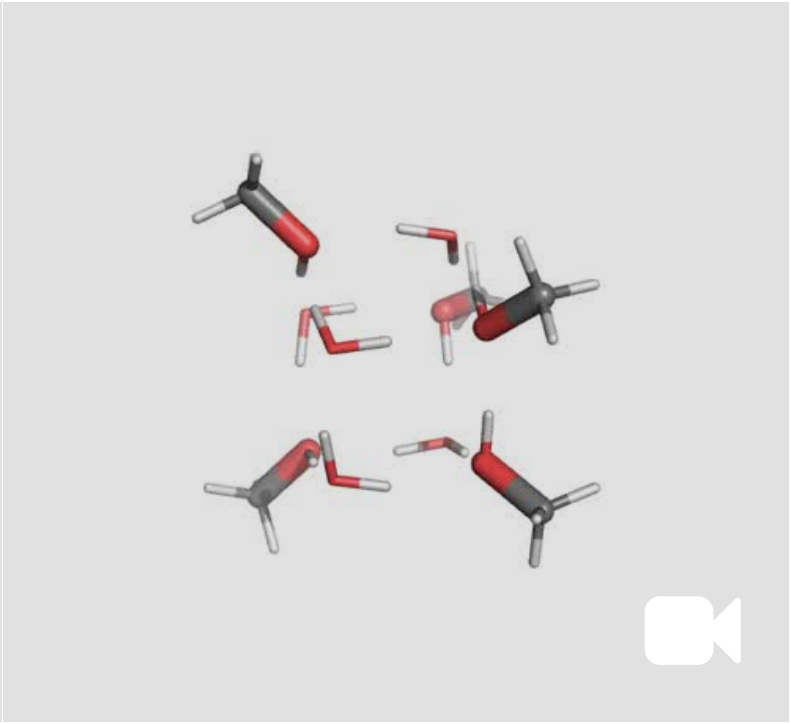
There should be no **net flow** of energy

Kinetic energy determines temperature

300 K



500 K



$$\langle KE \rangle = \frac{3}{2} k_B T$$

$\langle KE \rangle$	Ensemble average kinetic energy
k_B	Boltzmann constant
T	Temperature in Kelvin

Note: 3/2 comes from each degree of freedom (x, y, z)

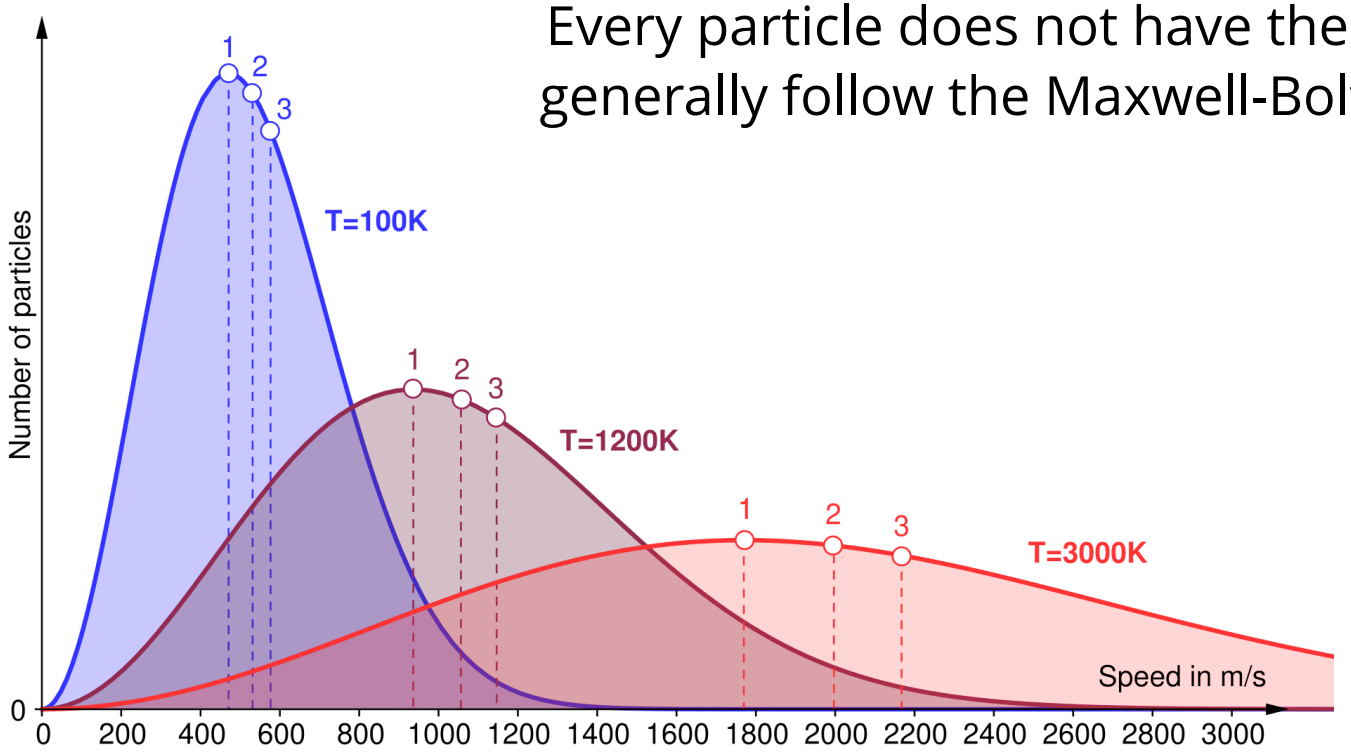
Particle velocities determine kinetic energy

$$KE = \frac{1}{2}mv^2$$

m Mass of each particle

v Velocity magnitude

Every particle does not have the same velocity; they generally follow the Maxwell-Boltzmann distribution



- **Most Probable Velocity:** The velocity at which the peak of the distribution occurs.
- **Average Velocity:** The mean velocity of all particles.
- **Temperature Dependence:** Higher temperatures shift the distribution towards higher velocities.

Thermostats adjust the velocities of particles to increase or decrease the system's kinetic energy

Thereby controlling the temperature

Berendsen thermostat: Adjusts the velocities of all particles uniformly based on the current temperature and the target temperature

$$\lambda = \left[1 + \frac{\Delta t}{\tau}\left(\frac{T}{T_0} - 1\right)\right]^{1/2}$$

Velocity scaling factor is computed by **slowly/carefully**

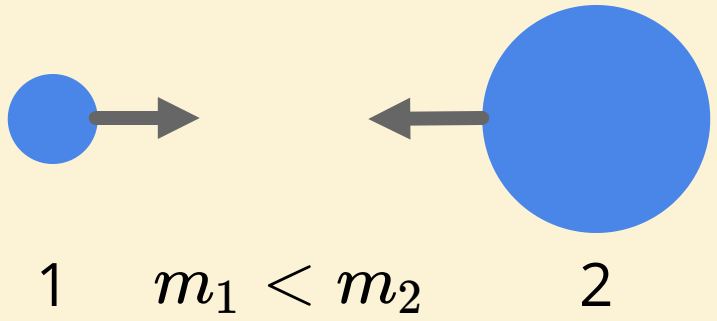
scaling the **current velocity** based on the **temperature deviation**

This prevents abrupt changes that could destabilize the simulation

Simple velocity scaling does not generate a true canonical (NVT) ensemble; it cannot reproduce realistic temperature fluctuations

Particle collisions are mass dependent

Berendsen thermostats inaccurately models thermal energy transfer via particle collisions



If two particles of different masses collide, will their velocities scale in the same way?

No

Momenta scaling provides realistic kinetic energy and thus temperature control

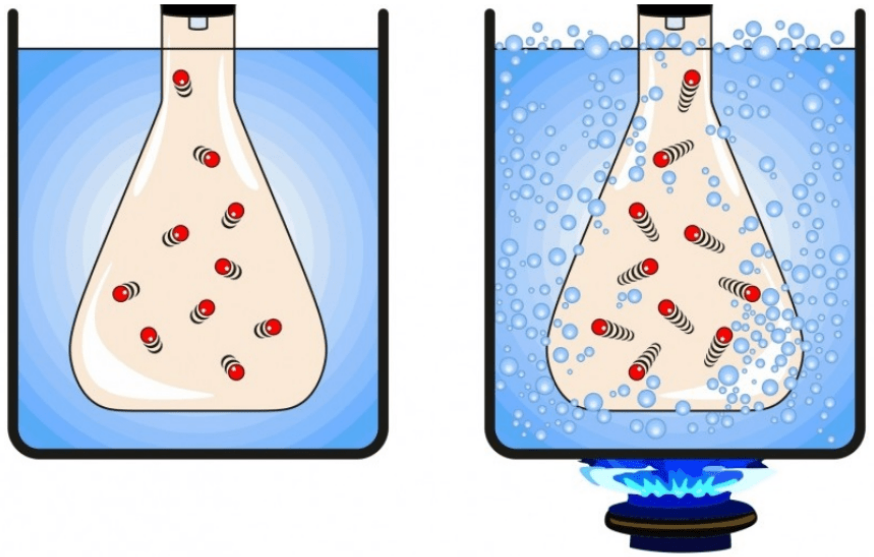
This is the principle behind the Nosé-Hoover thermostat

Nosé-Hoover thermostat connect particle momenta to a fictitious heat bath

This heat bath allows thermal energy to flow in and out of our simulation

Momenta adjustment:

$$\frac{d\vec{p}_i}{dt} = \vec{F}_i - \xi \vec{p}_i$$



"Friction" coupling constant:

$$\frac{d\xi}{dt} = \frac{1}{Q}\left(\sum_i \frac{\vec{p}_i^{\,2}}{m_i} - 3Nk_BT\right)$$

Q is a "mass" coupling parameters that controls thermostat responsiveness

Barostats maintain desired pressure during simulations

Adjusts the volume of the simulation box to achieve and maintain target pressure

$$P = \frac{Nk_BT + \langle W \rangle}{V}$$

Represents thermal energy of ideal gas

$$Nk_BT$$

Assumes (1) non-interacting particles and (2) elastic collisions

Virial corrections to real gas

$$\langle W \rangle$$

Corrects for intermolecular forces

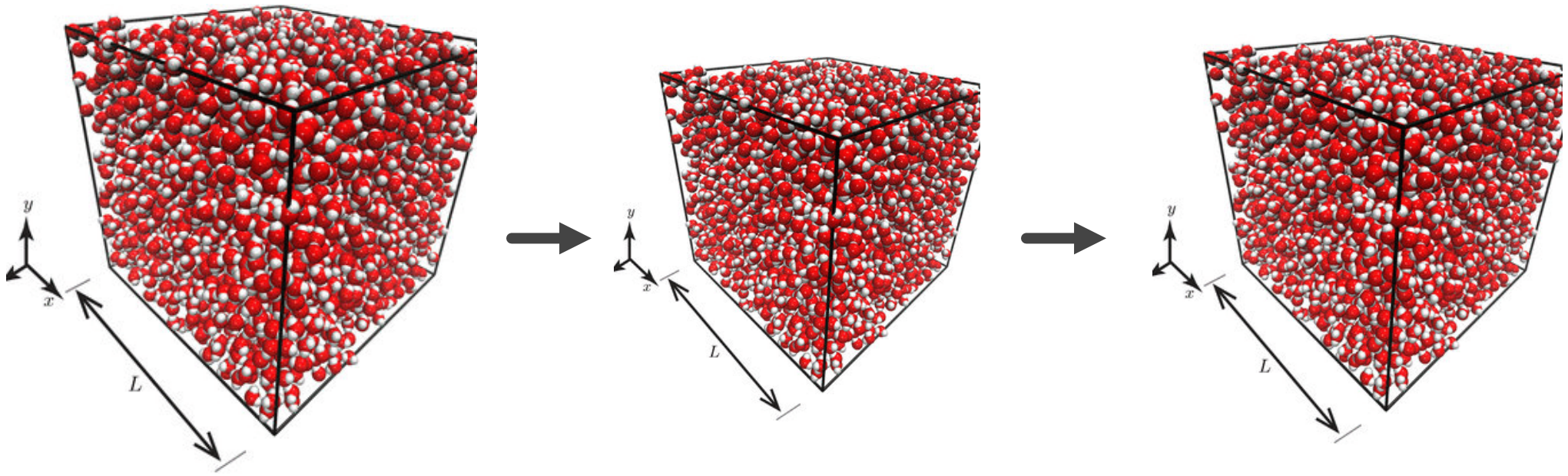
Pressure is directly proportional to density and temperature

$$\rho = \frac{Nm}{V} \Rightarrow P = \rho\frac{k_BT}{m} + \frac{\langle W \rangle}{V}$$

Berendsen Barostat: Gentle Pressure Stabilization

Same concept as Berendsen thermostat: Scale box volume based on pressure difference to target

$$\lambda = \exp\left[\frac{\Delta t}{\tau_P}\left(\frac{P_0}{P} - 1\right)\right]$$

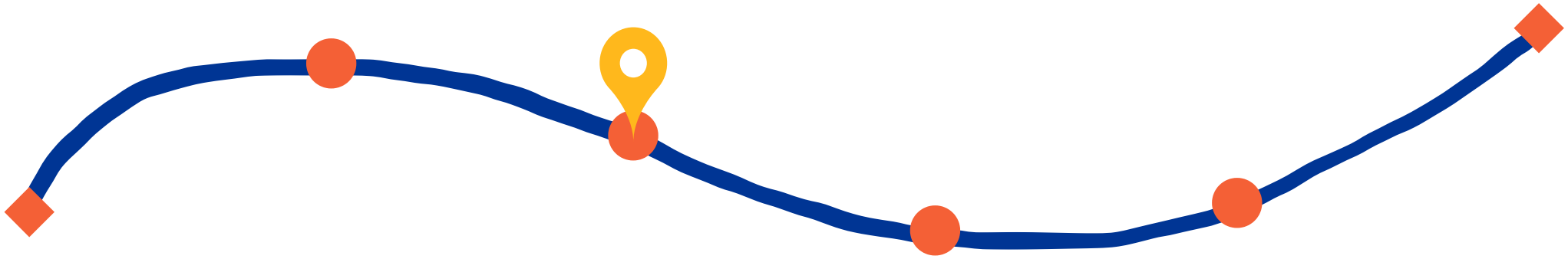


Atomic positions get scaled with box size

Velocities do not get affected

With thermostats and barostats we can keep a consistent macrostate

After today, you should better understand

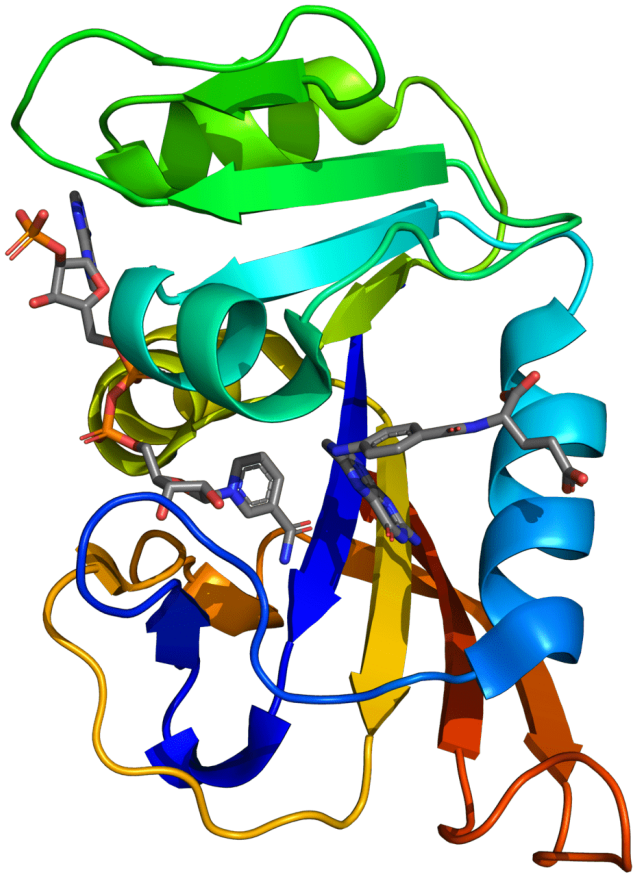


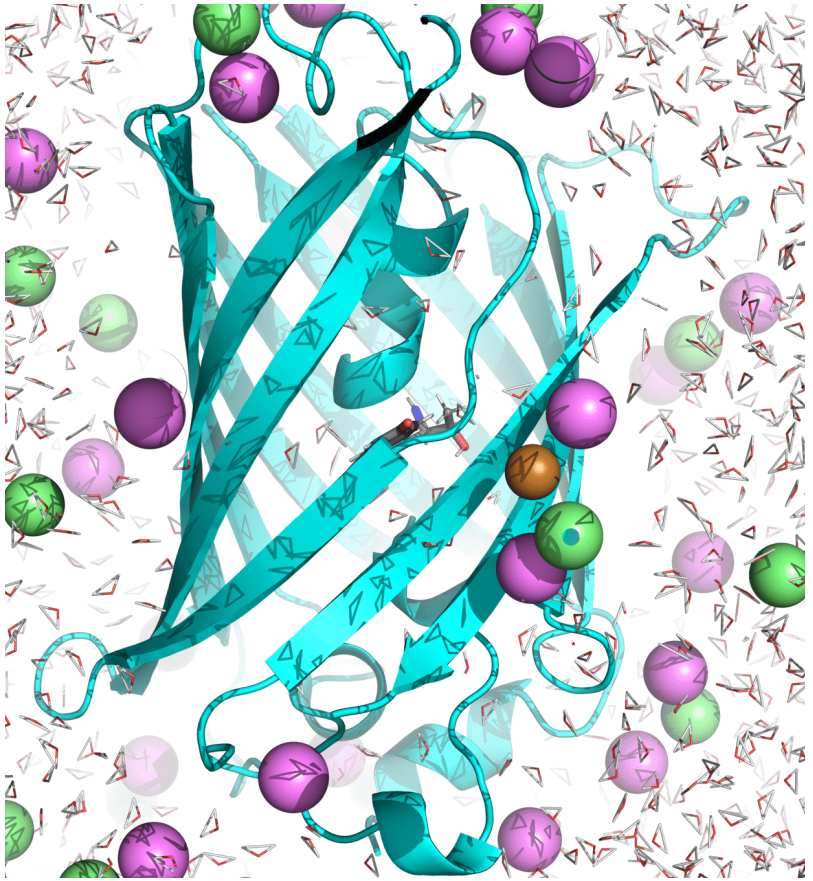
Relaxation and production

MD simulations

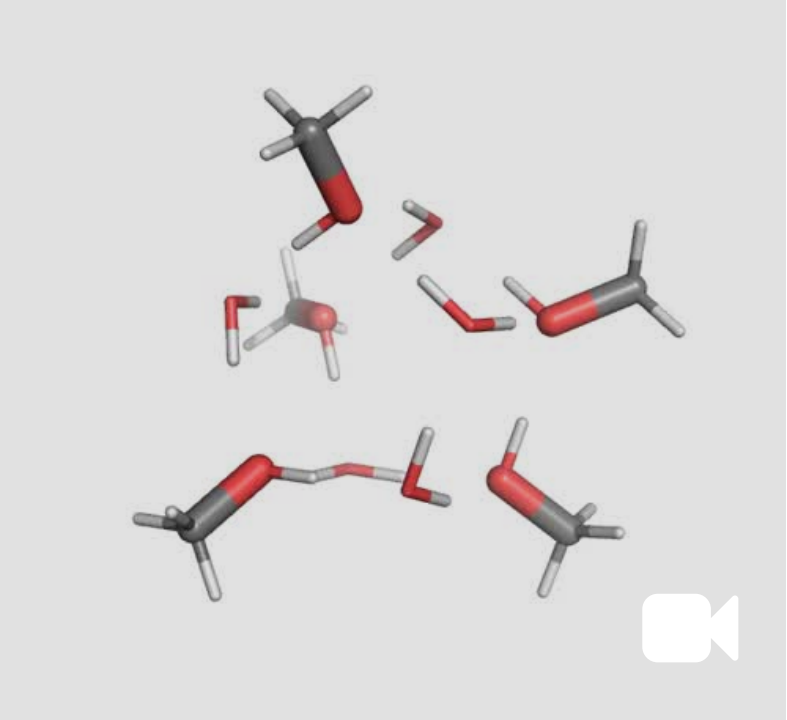
Initial configurations are not in true thermodynamic equilibrium

Remember: Starting structures often come from experiments not relevant for our simulation



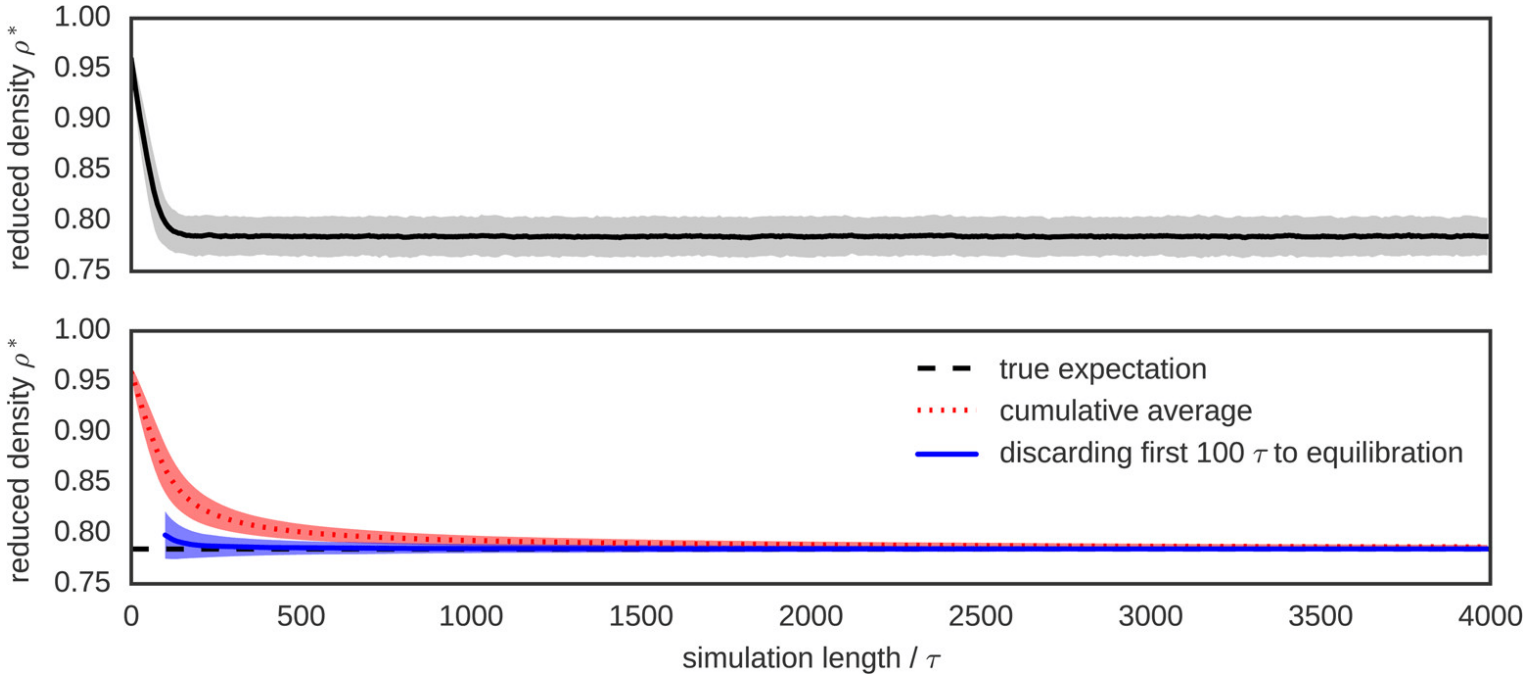


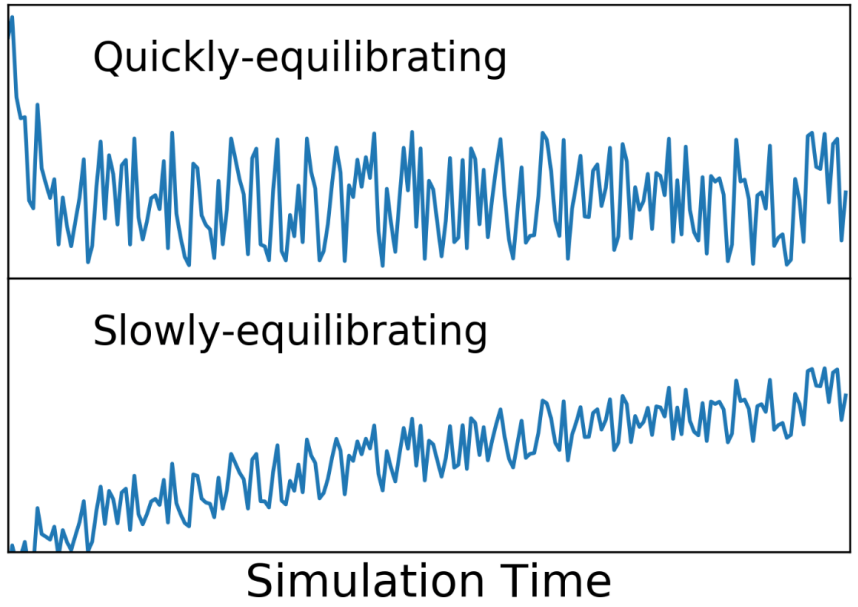
After minimization, we run a short simulation to let the system adjust to the desired macrostate



We discard the initial relaxation as it is not our desired macrostate

Once our macrostate variable(s) reach steady state, we are now sampling valid microstates



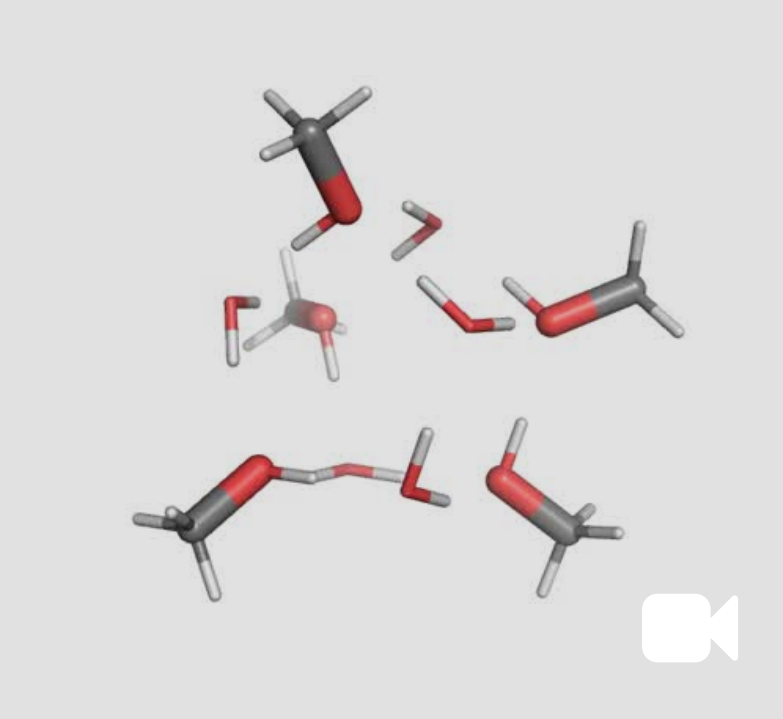


Production simulations sample microstates from our desired macrostate

NVT → NPT → NPT

(short simulation to relax to temperature of interest)

(short simulation to relax to density of interest)



Remember: Ensemble averages improve with more simulation time by sampling additional microstates

"Replicates" do not exists as it does in experimental biology and chemistry

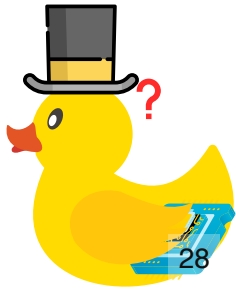
Multiple shorter simulations or one long one?

Which simulation protocol provides better sampling of microstates?

Option 1	**Option 2**
Three simulations of 500 ns each	One simulation of 1,500 ns

Assume: Each simulation starts with the same structure, but different initial velocities

Option 1 is correct



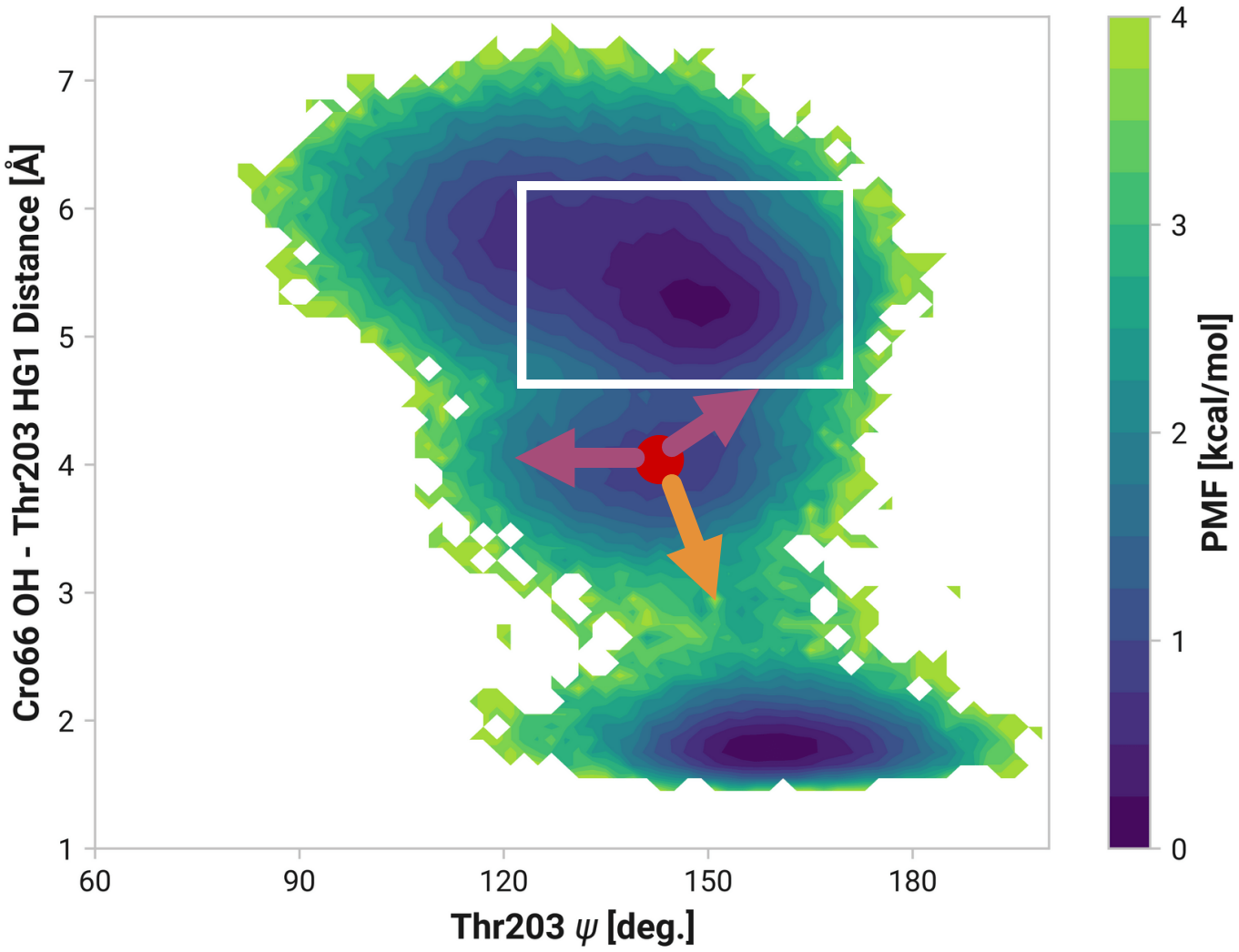
Random initial velocities provide better chance of sampling different microstates

Suppose my simulation starts **here** on my potential energy surface (PES)

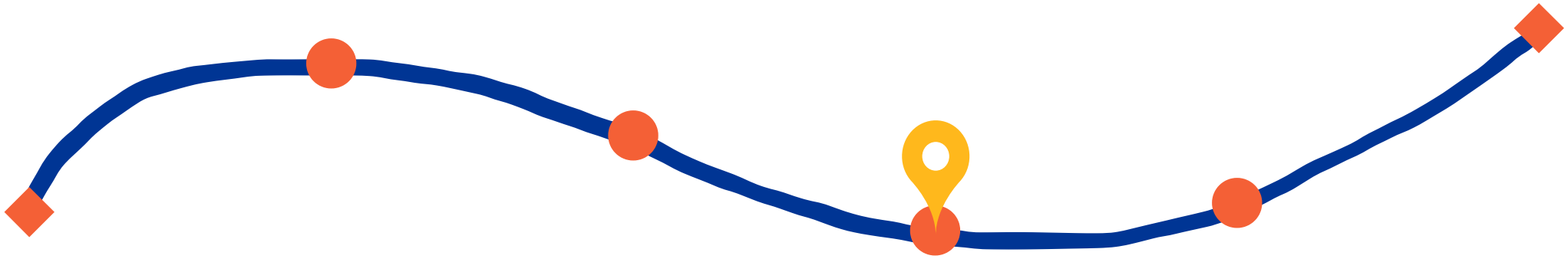
Suppose the initial velocities send it in **this direction**

There is a chance that it never samples **this minima**

Multiple simulations with random velocities reduces this chance



After today, you should better understand



RMSD and RMSF as conformational changes and flexibility metrics

Root Mean Square Deviation (RMSD): Monitoring Global Conformational Changes

RMSD measures the overall change in the structure during a simulation, tracking deviations from the starting conformation

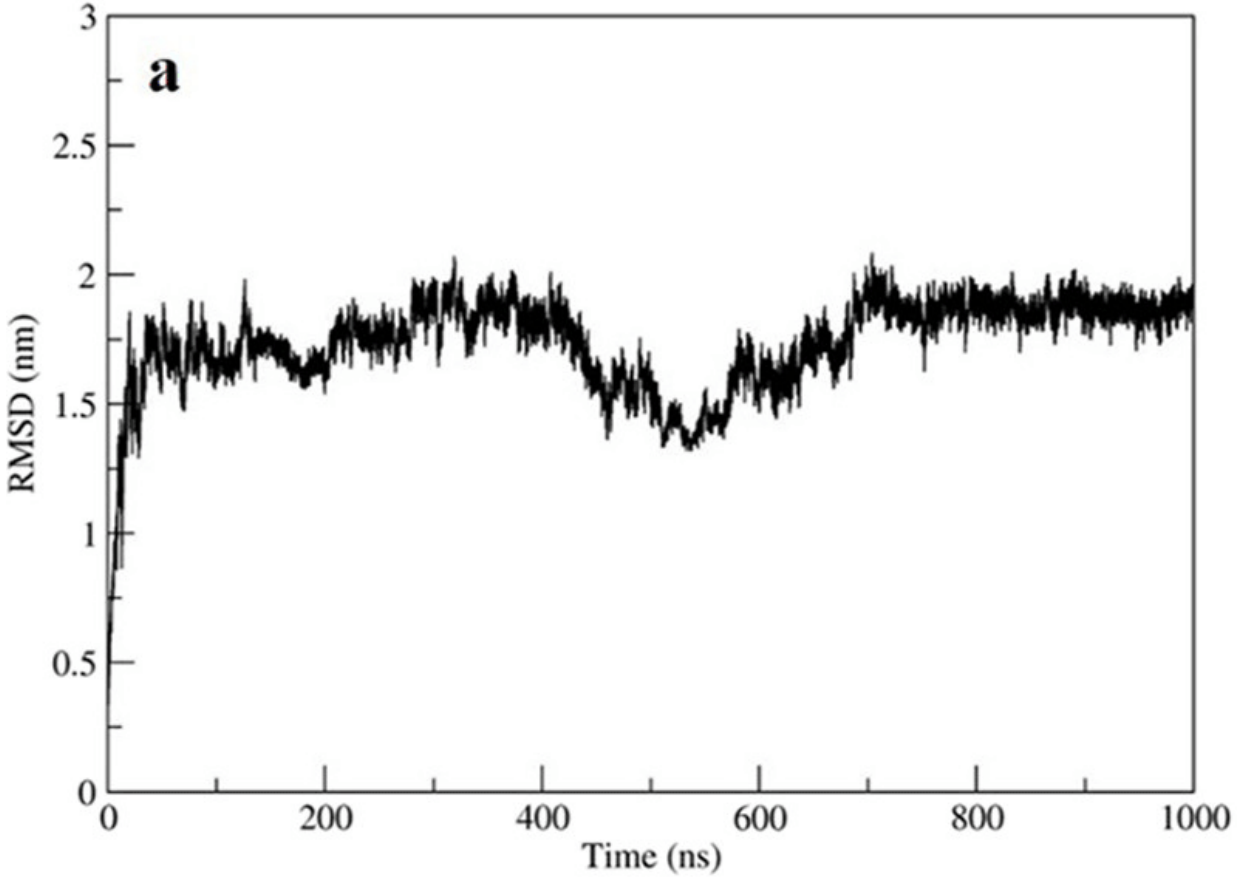
The **difference** between the coordinates represents the displacement of atom *i* from its reference position at time t

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{r}_i(t) - \mathbf{r}_i^{\mathrm{ref}}\right)^2}$$

N	Number of atoms to compare
$\mathbf{r}_i(t)$	Position of atom *i* at time *t*
$\mathbf{r}_i^{\mathrm{ref}}$	Reference position of atom *i*

- A **low RMSD** means the structure is very similar to the reference structure (e.g., stable conformation)
- A **high RMSD** indicates significant deviation, suggesting large structural changes or flexibility over time

Example RMSD plot



Root Mean Square Fluctuation (RMSF): Tracking local flexibility

RMSF identifies regions of flexibility in the protein by calculating the fluctuation of each atom or residue

This measures how much the atom is fluctuating around its mean, not relative to a reference structure

$$\mathrm{RMSF}(i) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle\right)^2}$$

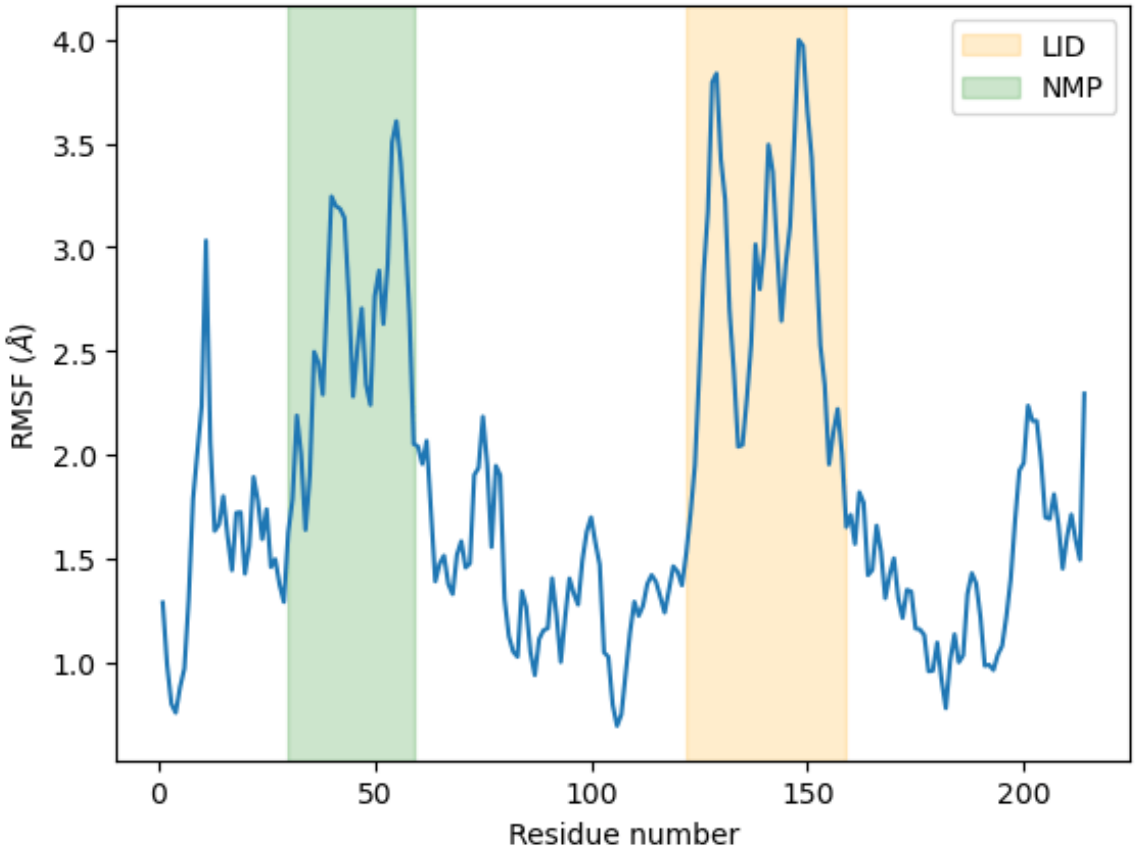
T	Total number of time frames
$\mathbf{r}_i(t)$	Position of atom *i* at time *t*
$\langle \mathbf{r}_i \rangle$	Average position of atom *i*

- A **high RMSF** value for an atom means that it fluctuates a lot, indicating flexibility (often seen in loops or solvent-exposed regions)
- A **low RMSF** value means the atom remains relatively fixed in place, suggesting rigidity (common in well-ordered regions like helices or beta-sheets)

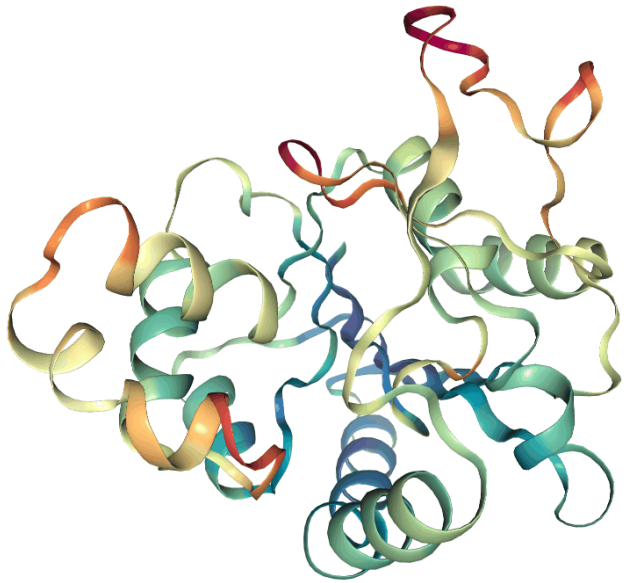
Example RMSF plot

LID is an ATP binding domain

NMP is an ADP binding domain

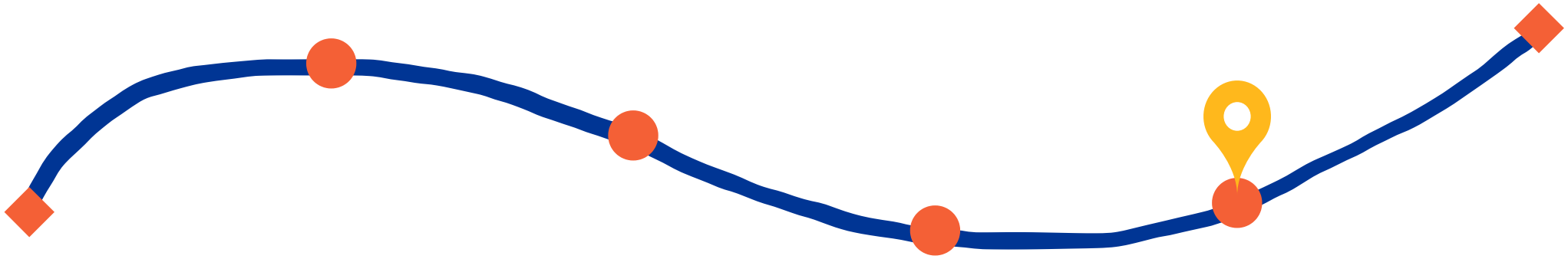


Adenylate kinase (AdK), a phosophotransferase enzyme



Regions in red have high flexability

After today, you should better understand



Relationship between probability

and energy in simulations

What is Potential of Mean Force (PMF)?

PMF represents the **effective potential** that governs the behavior of a system along a **collective variable**

- $W(x)$ is the PMF as a function of the collective variable x.
- k_B is the Boltzmann constant.
- T is the temperature.
- $P(x)$ is the probability of observing a microstate with the collective variable value of x.

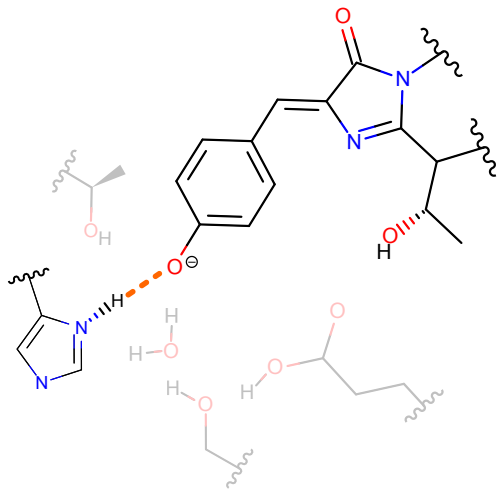
$$W(x) = -k_B T \ln P(x)$$

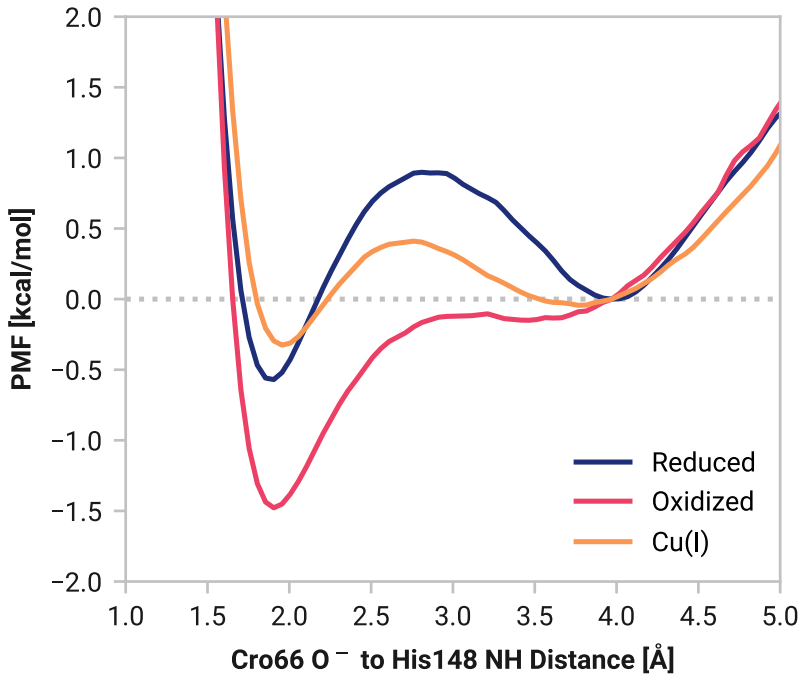
A collective variable defines the progress of an interaction or molecular reaction

Common collective variable include **distances between atoms**, **bond angles**, or **dihedral angles**.

Interpreting a 1D potential energy surface (PES)

His148 in GFP stabilizes the anionic chromophore



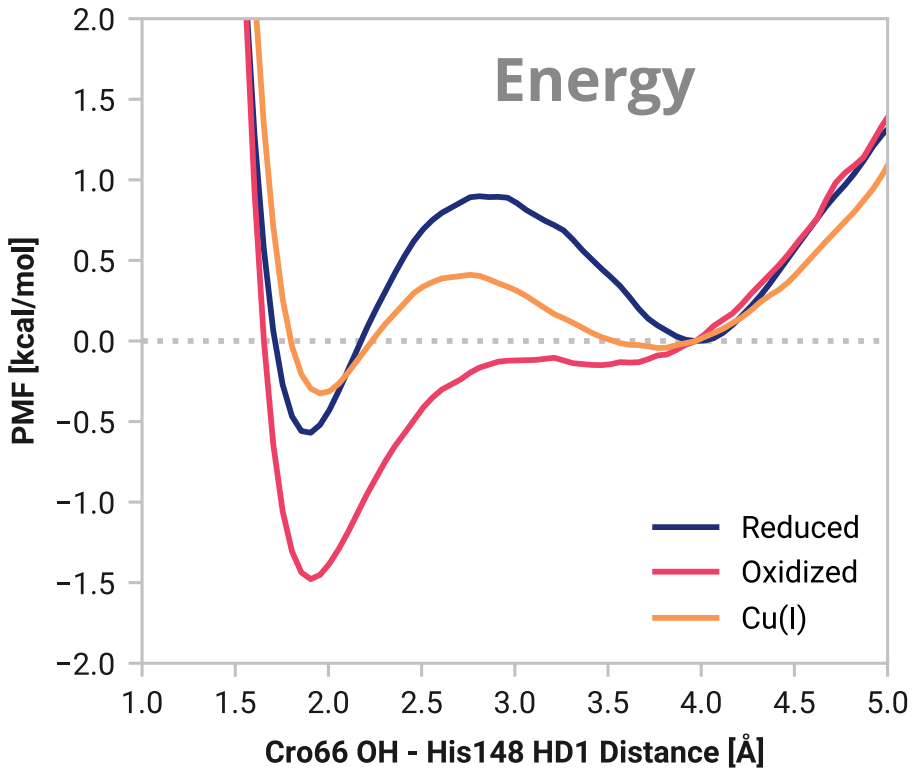


System	ΔG [kcal/mol]
Reduced	-0.559
Oxidized	-1.329
Cu(I)	-0.282

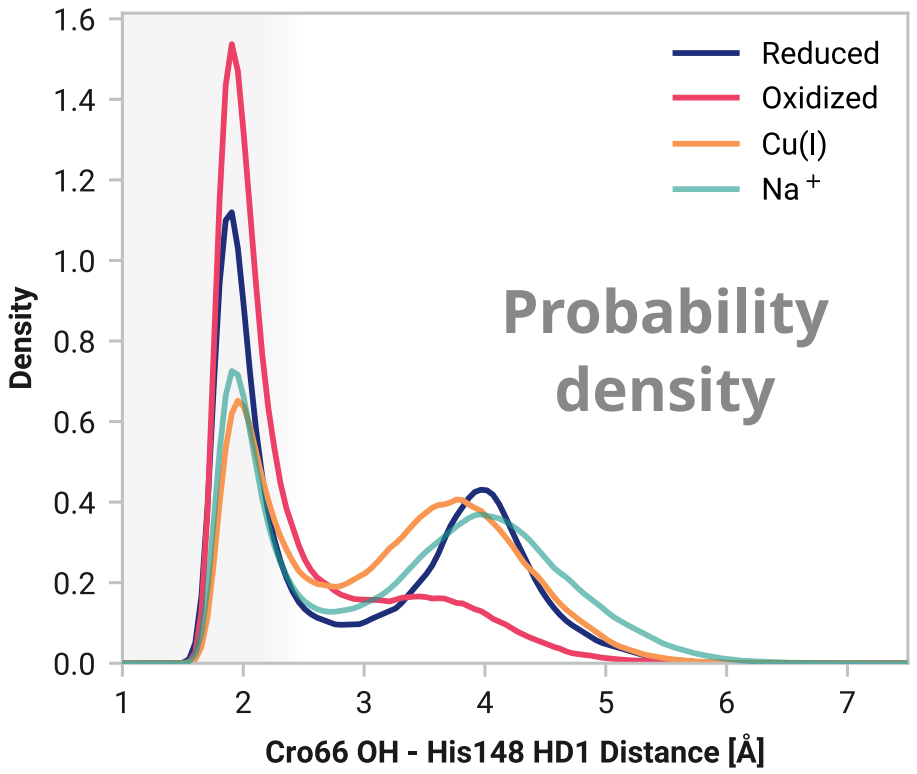
This 1D PES comes from 1500 ns of roGFP2 simulation data

Probability and energy are intricately linked

$W(x)$



$P(x)$



Before the next class, you should

Lecture 15:
Ensembles and atomistic insight

Lecture 16:
Structure-based drug design

Today

Tuesday

- Turn in A05
- Start A06