Enhanced Sampling

“When equilibrium MD is not enough”

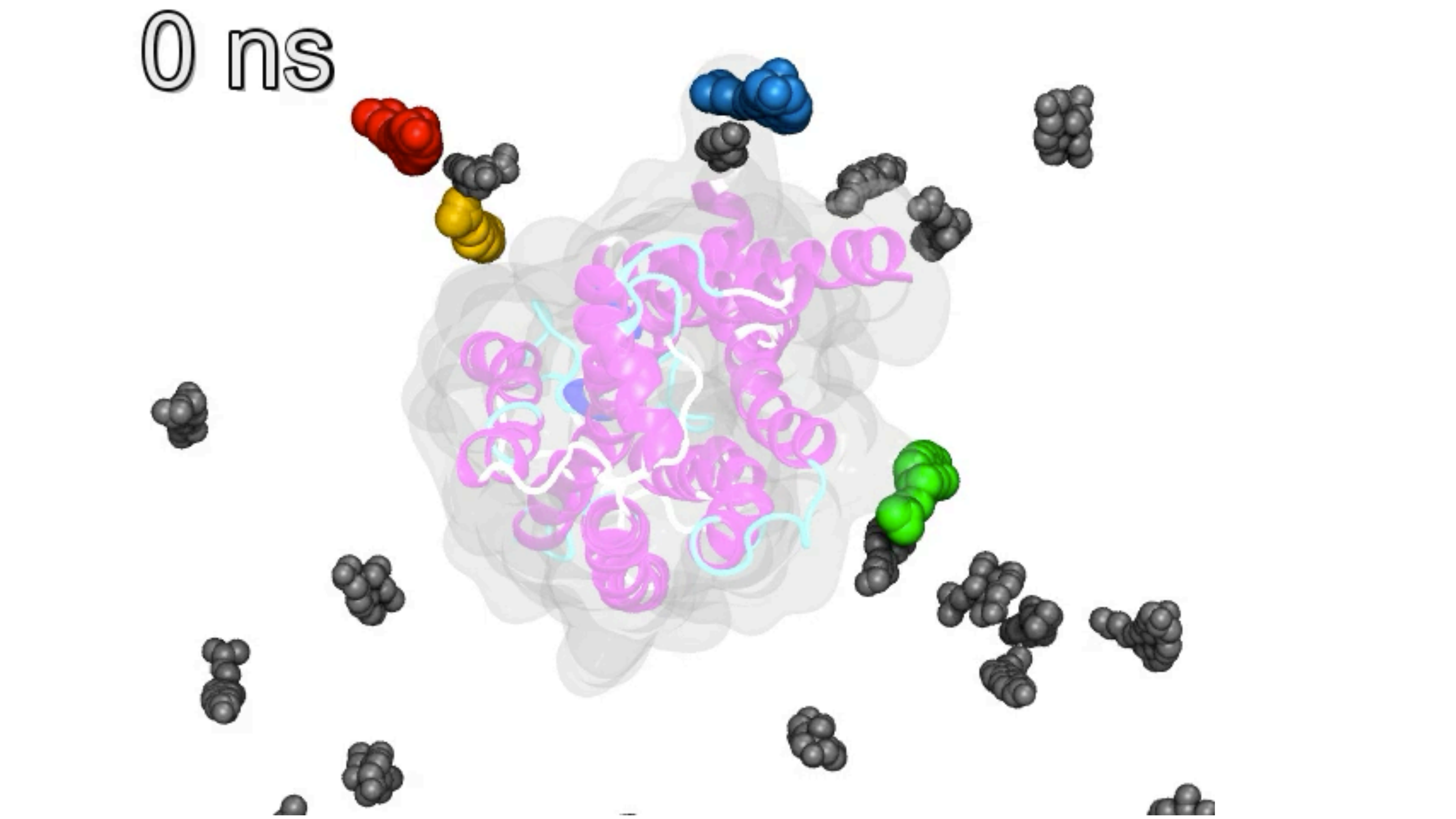
Matti Javanainen — CSC Spring School in Computational Chemistry, April 19th 2024

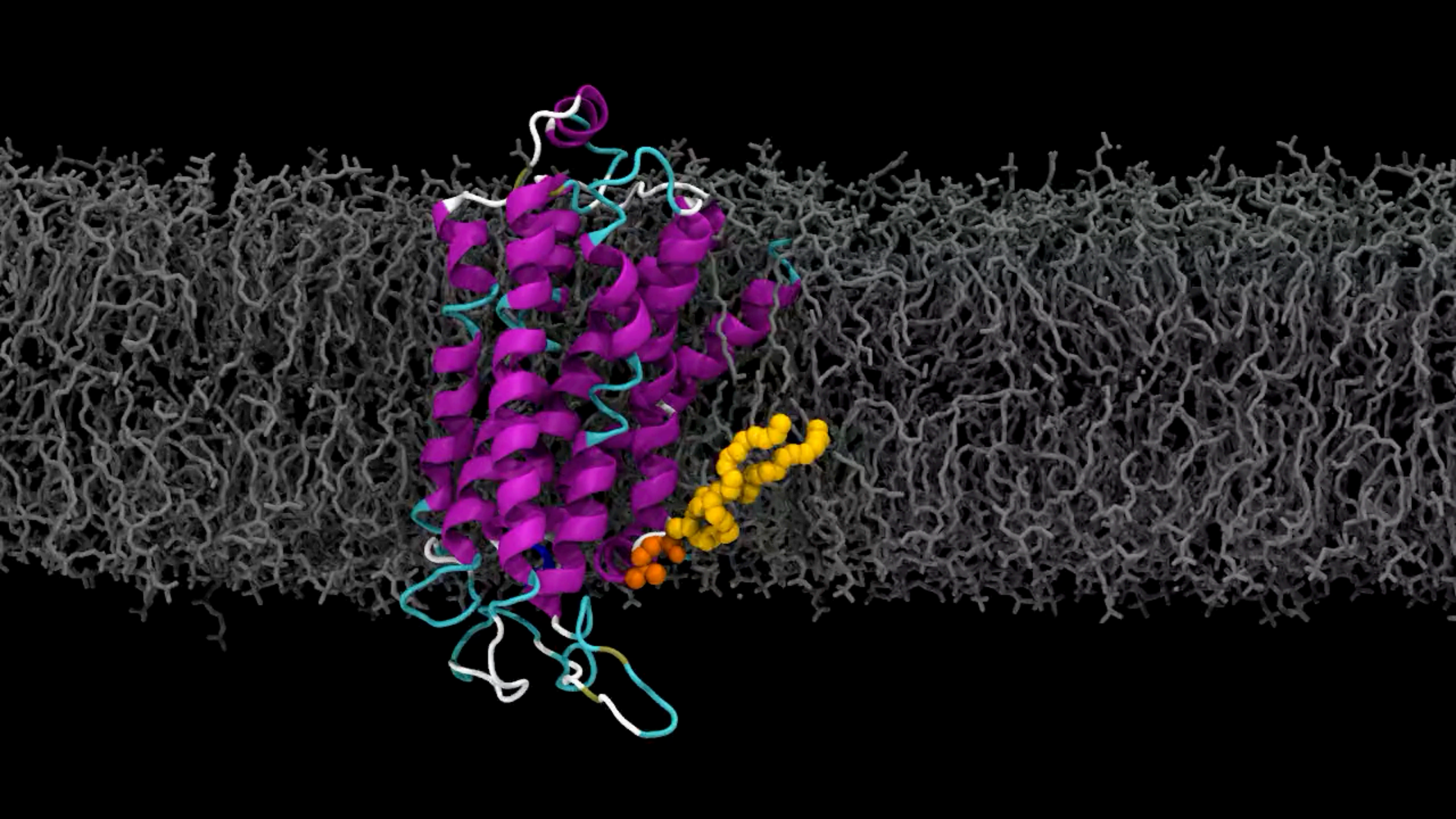
Outline for the morning

- **Lecture (~45 min):**
  - Use cases for enhanced sampling
  - Quick overview of common methods
- **Hands-on exercises (~90 min):**
  - Accelerated weight histogram (AWH) method:
    - Physical reaction coordinate: Permeation through a lipid bilayer
    - Alchemical reaction coordinate: Oil–water partition coefficient
  - Choose one or do both, depending on how fast you are

A very practical take from a user's perspective

Links to papers with theoretical concepts in the end


0 ns

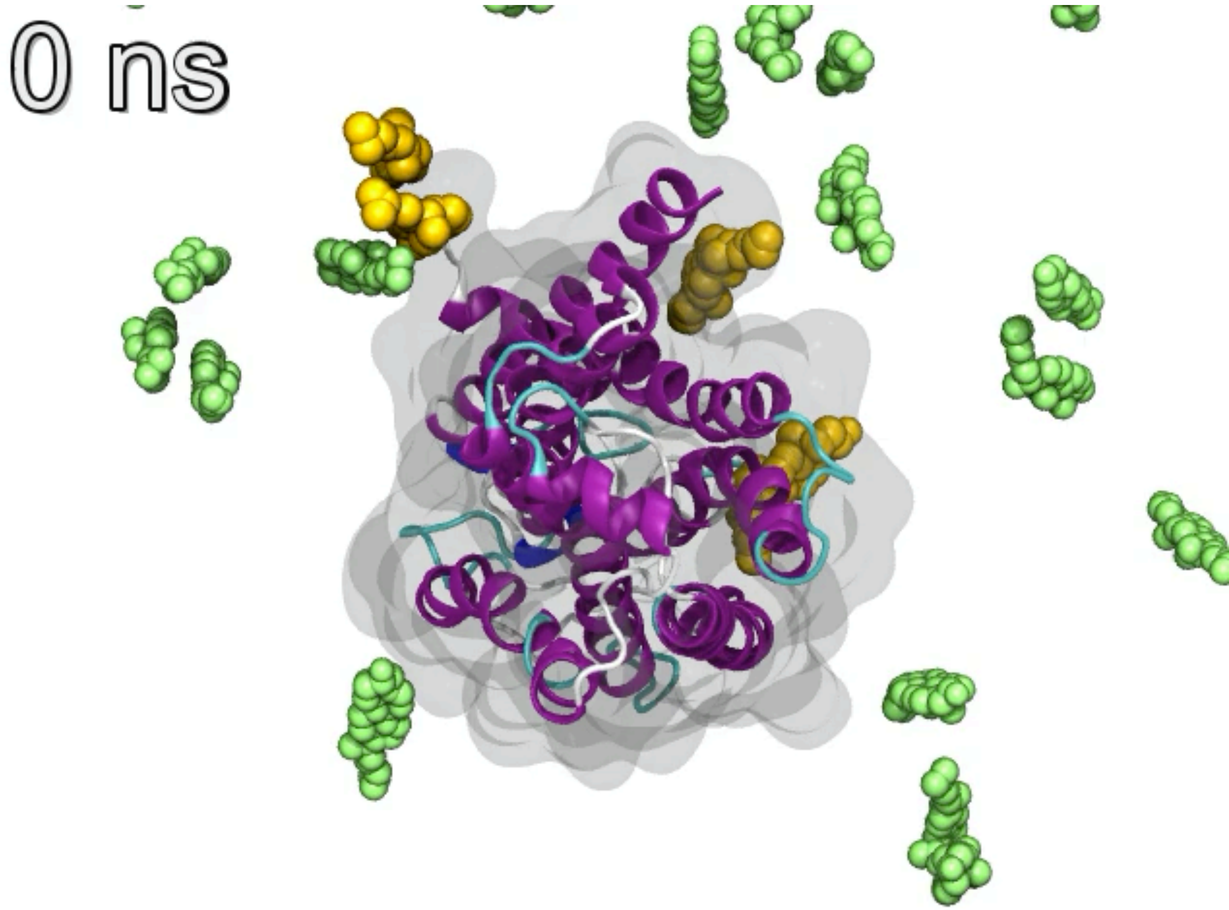


Can be done:

- Spontaneous events or rare events that are easy to force to happen

Cannot be done:

- Major conformational changes, binding to unknown sites
- Rare events or events that are hard to force to happen

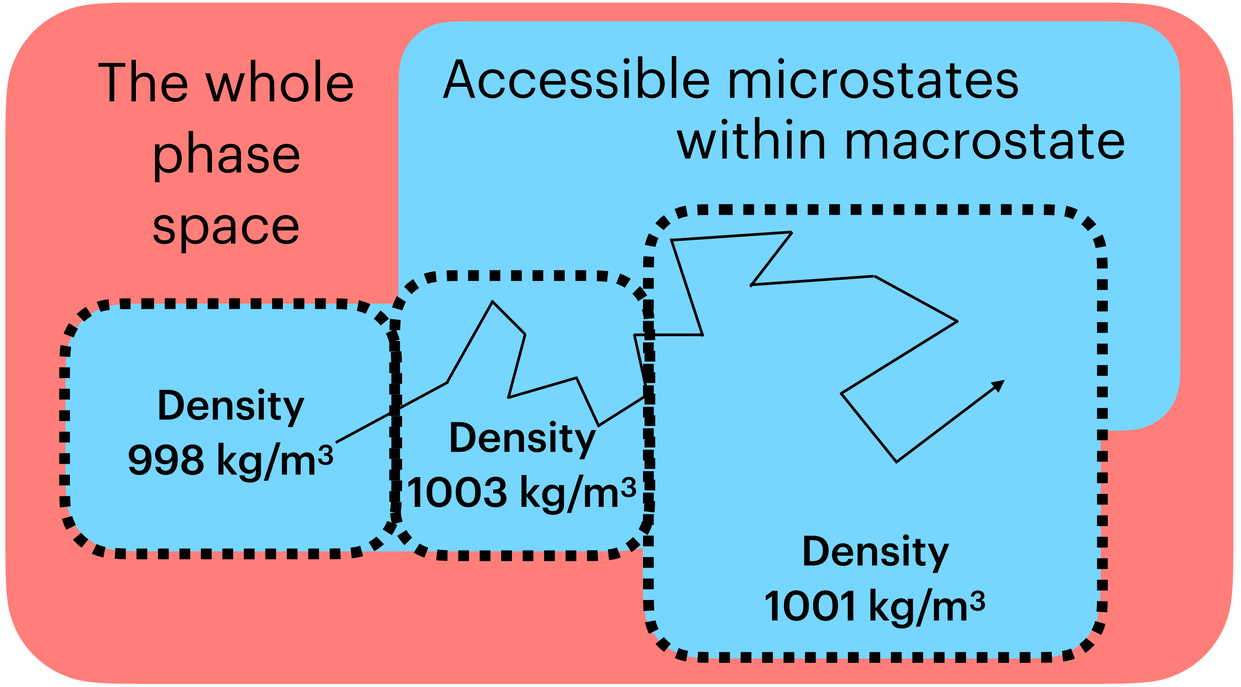


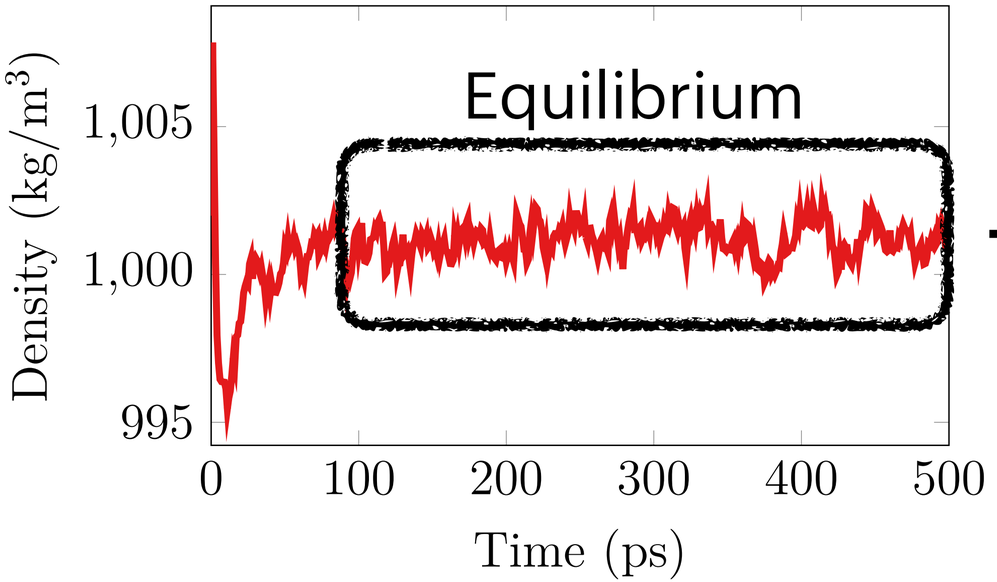
Cholesterol binding to beta-2 adrenergic receptor

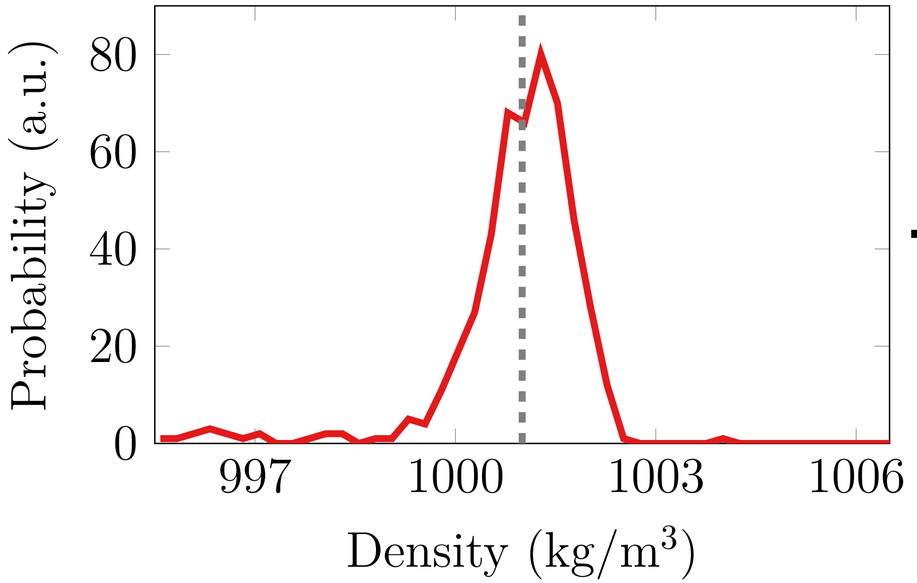
Manna *et al.*, eLife 5, e18432 (2016)

MD fundamentals

- Phase space with 6*N* dimensions!
- Macrostate set by *N,E,p,T*
  - Each has many microstates
- How system behaves (in a macrostate)?
  - Sample microstates & calculate average







(time) average 1001 kg/m³

Sampling problem(s)

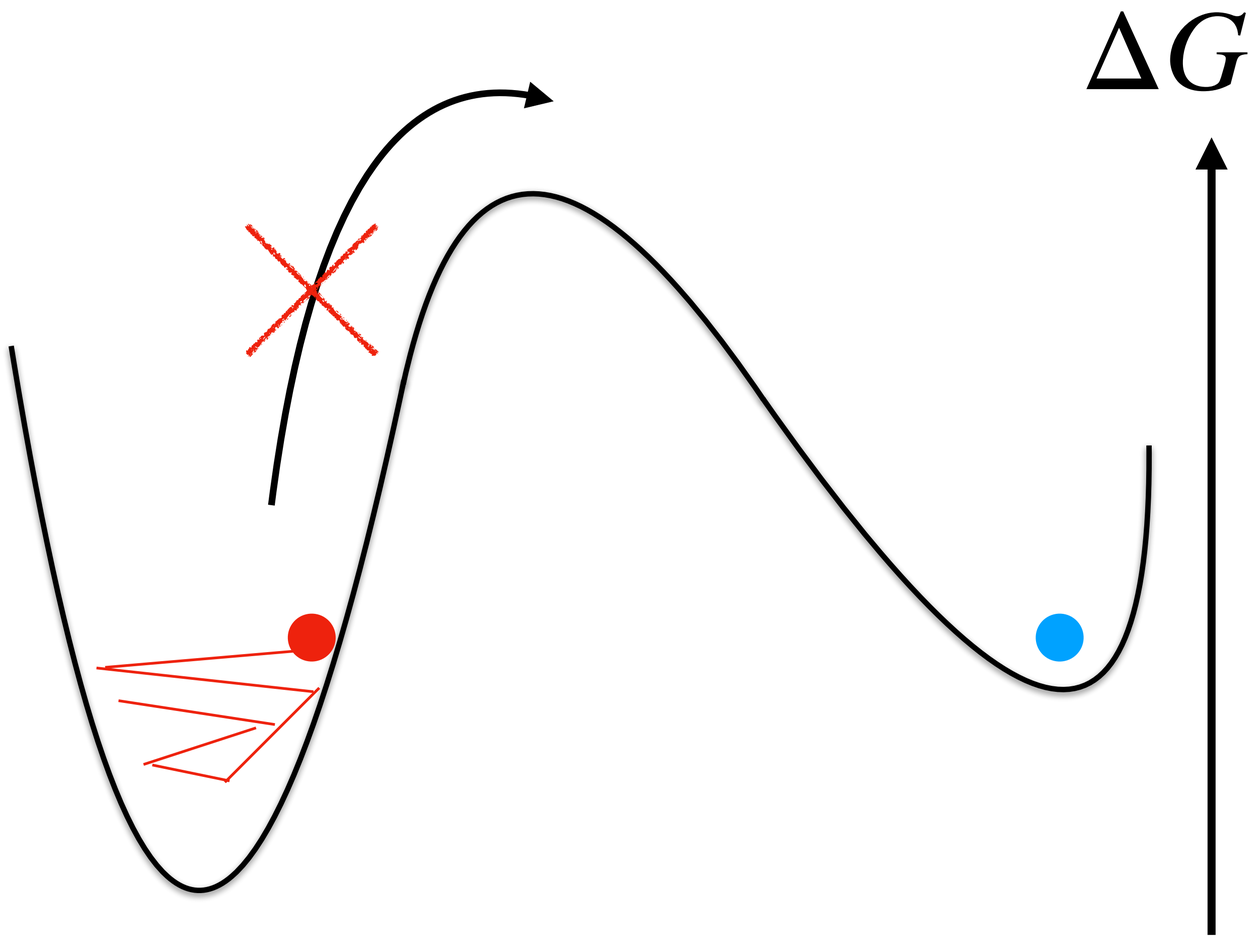
- Many slow processes with **large energy barriers** are not crossed in the (limited) simulation time scale

  → Often nothing happens *in silico*

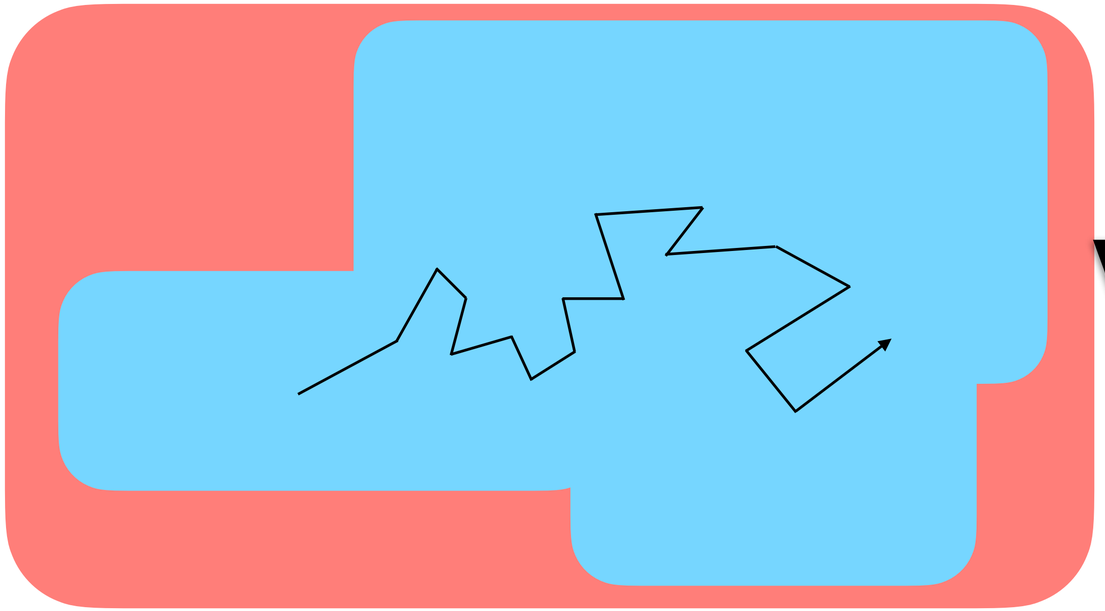
- MD aims to sample the macrostate to get an **ensemble average**

- Typically **time averages** are calculated (equal when ergodic)

  → Nonergodicity: Properties depend on initial conditions!



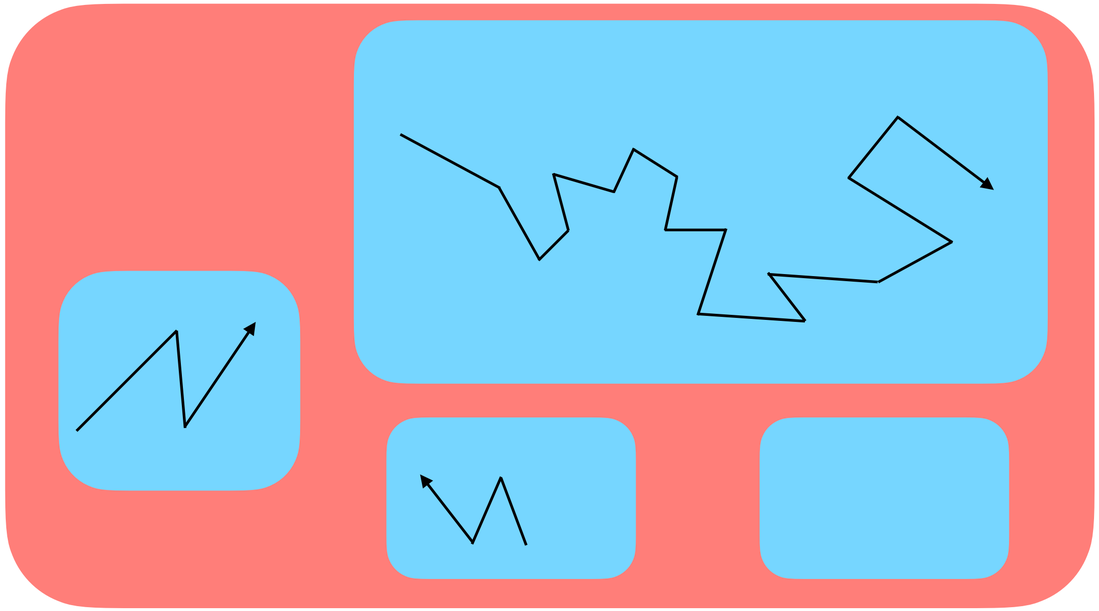
Sampling problem visualized



Microstates connected, all sampled within a long enough simulation

vs.

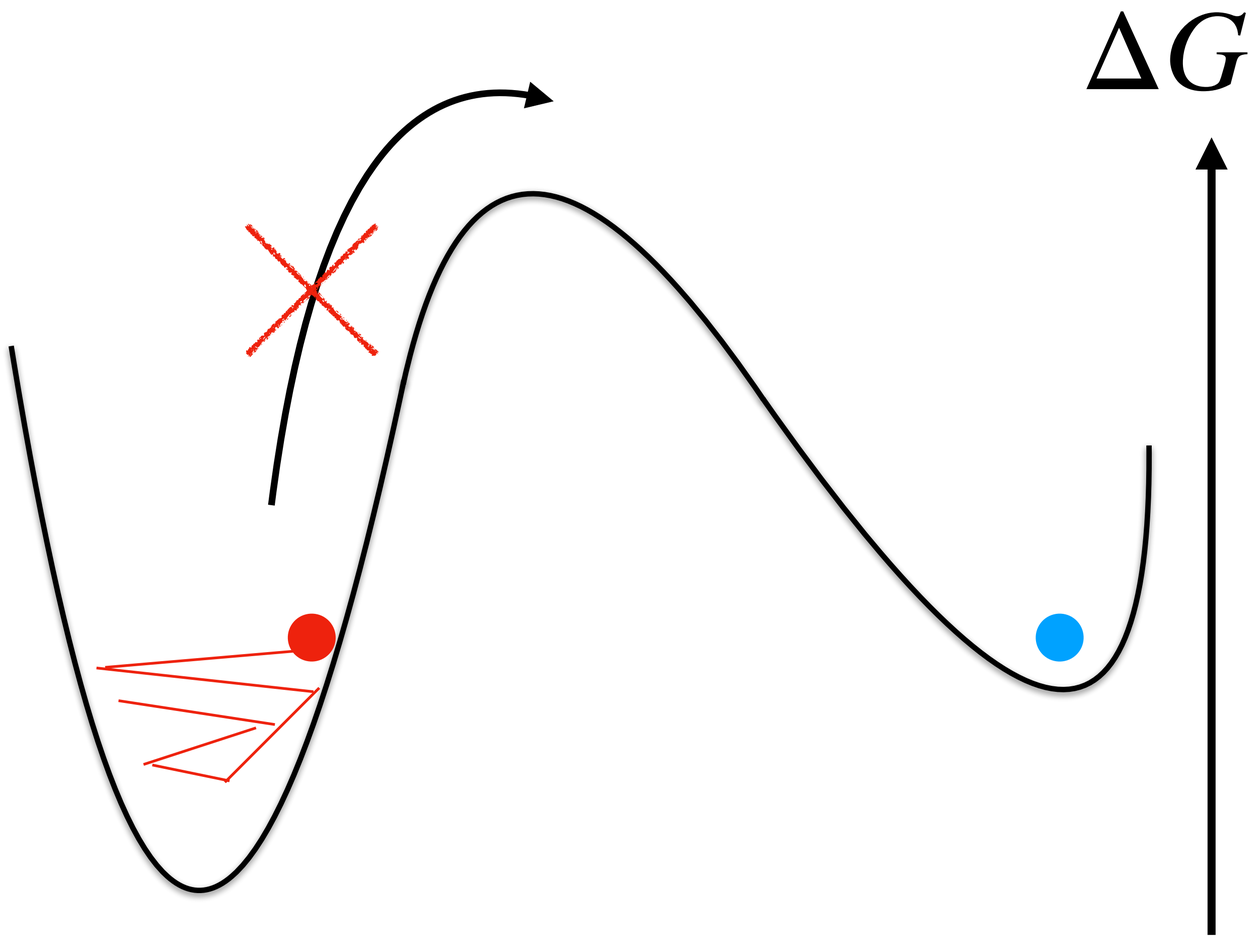
Microstates not connected, all not visited in a single simulation



Result depends on initial conditions

Solution

- **Bias** or **accelerate** the simulation for more efficient sampling
  - Reduce the energetic barriers or
  - Reduce the sampled phase space
  - Still sample all relevant states
- Finally unbias the result to recover the behavior of the original system
  - **Free energy surface (ΔG)**
  - Correctly averaged properties



Free energy often of interest…

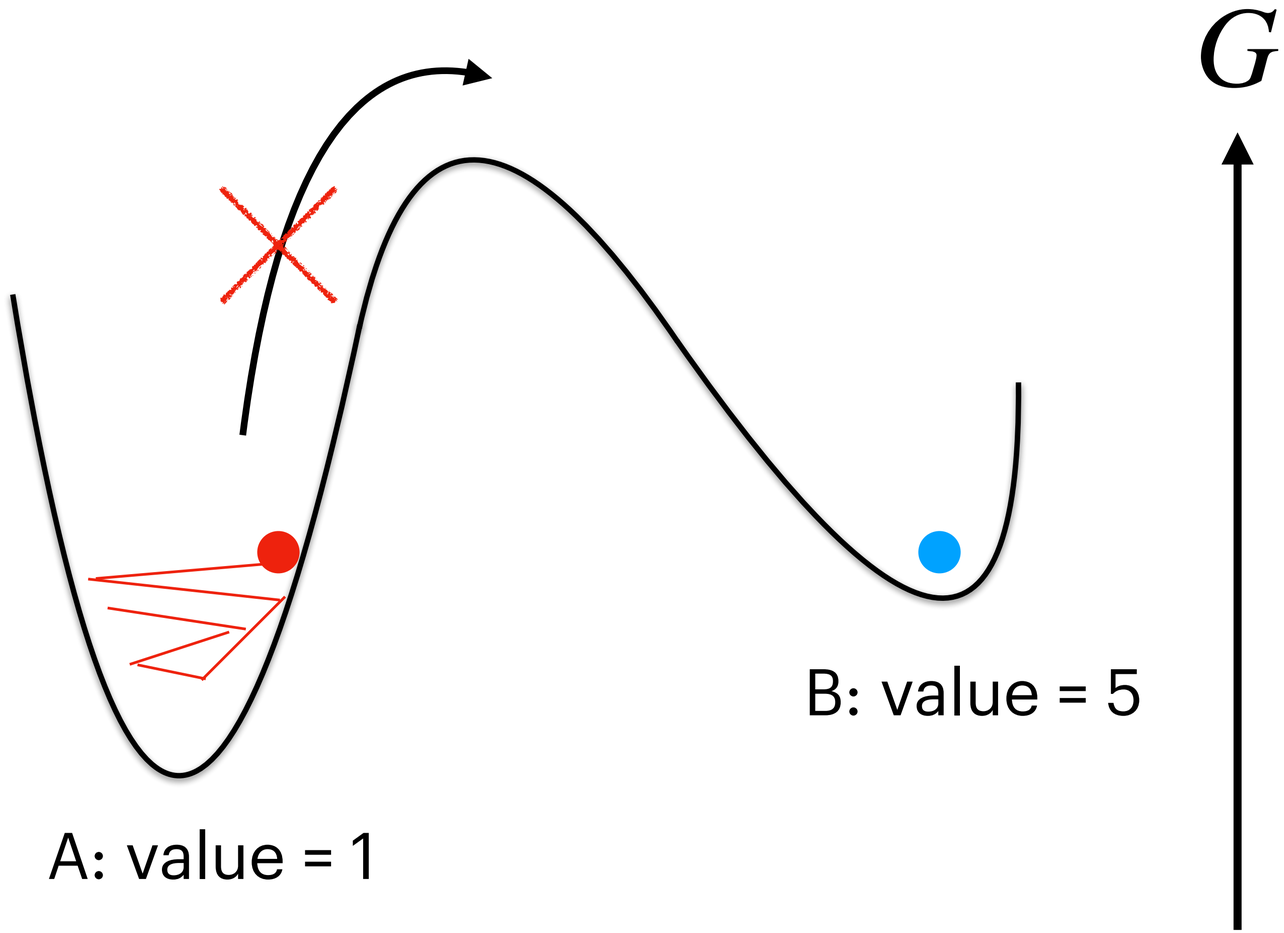
- Sometimes we want just the **free energy difference**:
  - Protein–ligand affinity for drug design
  - Effects of protein mutations on their interactions / structure
  - Phase diagrams (phase with smallest free energy)
- Sometimes the **free energy profile**:
  - Conformational landscapes of macromolecules
  - Energy barriers for various reactions/processes
    - Pharmacokinetics

… but it's not only about free energy.

- Free energy **differences** tell the relative probabilities of states

$$P_A/P_B = \exp\left[-\beta\left(G_A - G_B\right)\right]$$

- With large barriers, time-averaged quantities will not correspond to the ensemble average (non-ergodicity!)
- Correct averages obtained by **reweighting** the values by the respective probabilities

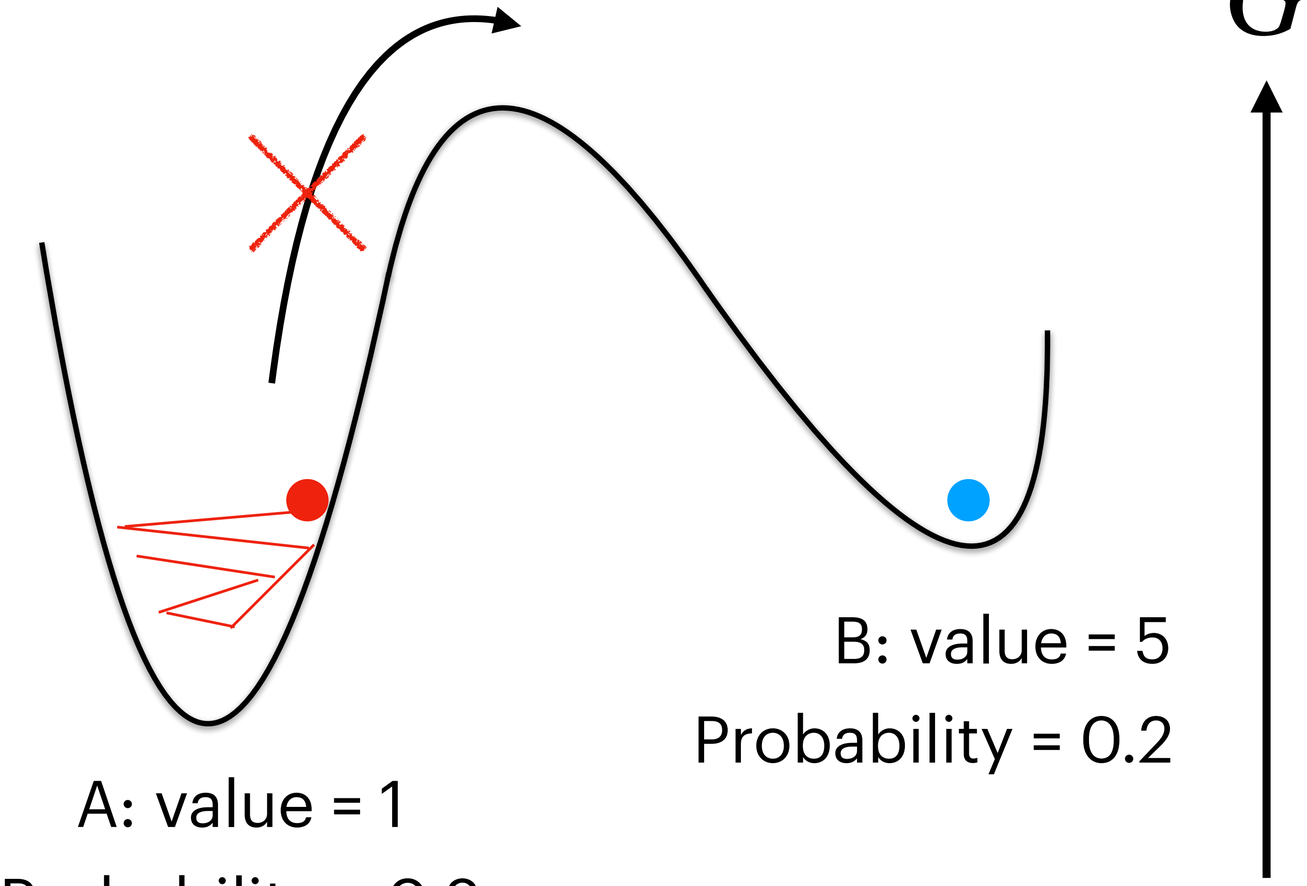


… but it's not only about free energy.

- Free energy **differences** tell the relative probabilities of states

$$P_A/P_B = \exp\left[-\beta\left(G_A - G_B\right)\right]$$

- With large barriers, time-averaged quantities will not correspond to the ensemble average (non-ergodicity!)
- Correct averages obtained by **reweighting** the values by the respective probabilities



After reweighting: proper average=1.8

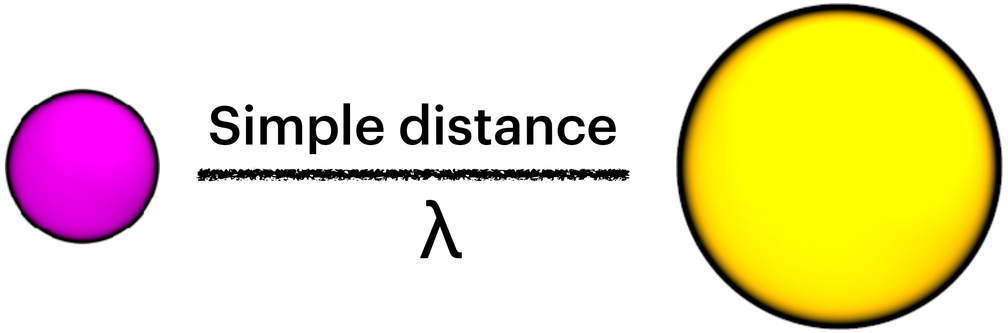
What to bias/accelerate?

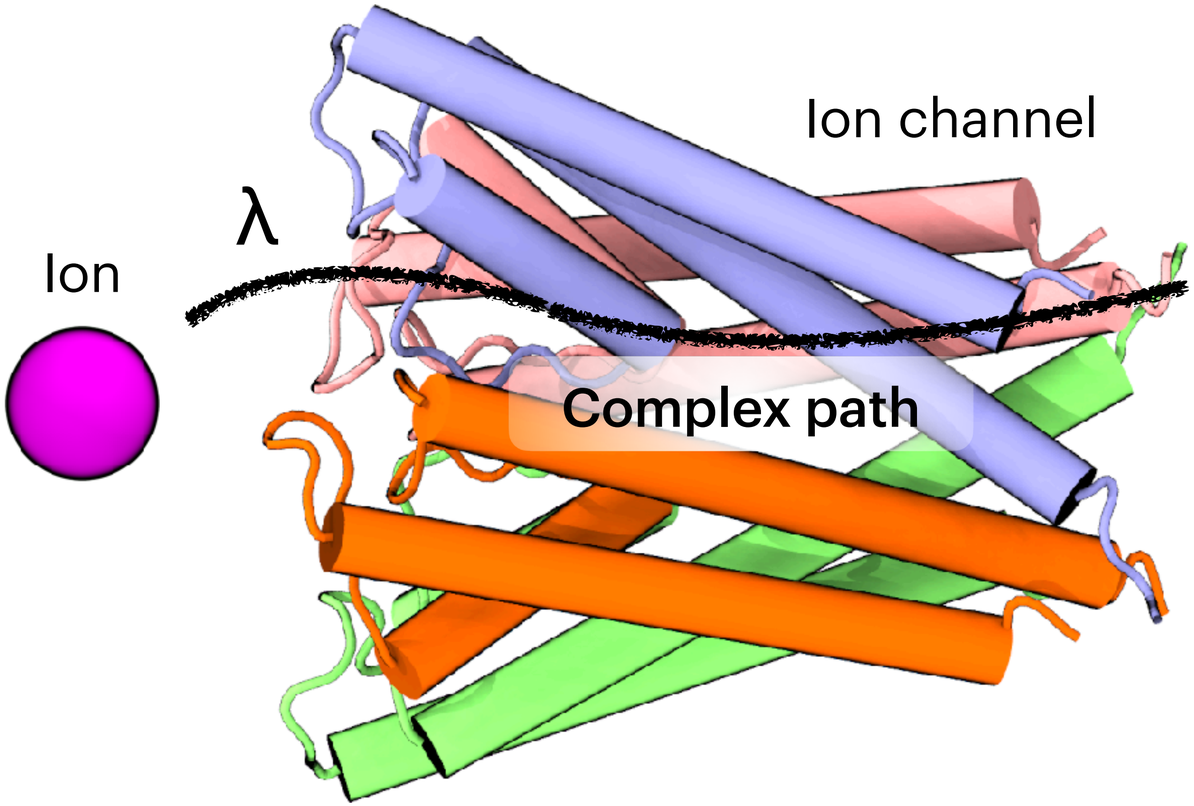
What is a reaction coordinate and how to select one?

Reaction coordinate
(= Collective variable, CV)

- Pre- and user-defined parameter λ
- Describes the state of the system linked to studied phenomenon (“*reactants to products*”)
- Physical or alchemical
- Differentiable for bias
- Low-dimension
- End points are easy to define, intermediates often not

Examples of physical 1D reaction coordinates

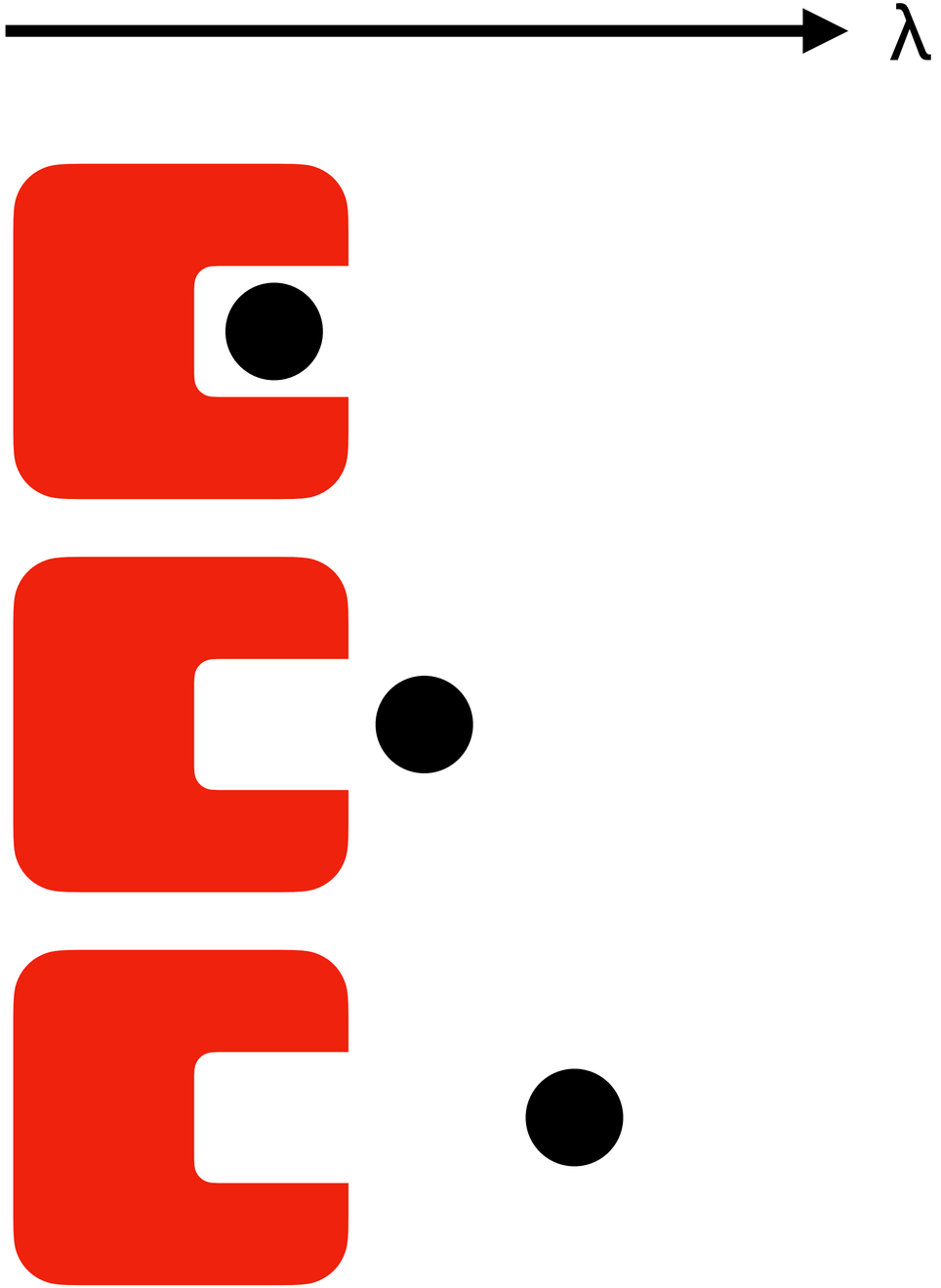




Physical ones

Based on atomic coordinates

- Complexity varies greatly:
  - Polymer end-to-end distance (1D)
  - Deviation from a reference structure (RMSD) of a protein (1D)
  - Center of mass distance and relative orientatition of two peptides (2D)
- Simple ones often from intuition
- Complex ones from machine learning

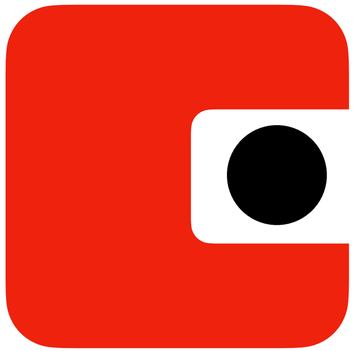


Alchemical ones

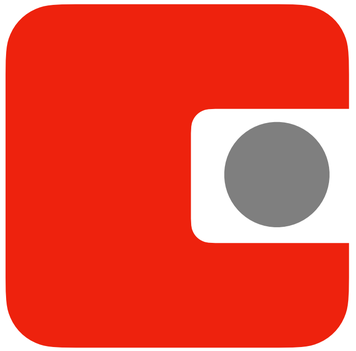
Tweaking the potential function

- The potential function depends on the reaction coordinate λ ∈ [0, 1]
- Molecules vanish (non-interacting with environment) by changing λ
- Molecules mutated by changing λ
- Only end points are meaningful
- We can simulate an amino acid that’s 30% Lys and 70% Arg!

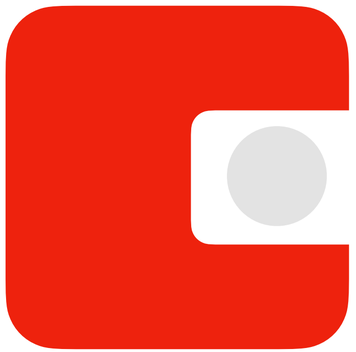
$$H(\lambda) = \lambda H_A(r) + (1 - \lambda) H_B(r)$$



$\lambda = 0$



$\lambda = 0.5$



$\lambda = 0.9$

How to bias/accelerate?

How to force the reaction coordinate to sample all desired values?

Types of enhanced sampling 1

- **Free sampling of reaction coordinate(s)**
  - Enhanced by higher temperature: **replica exchange MD (REMD)**
  - Enhanced by shallower potential: **accelerated MD (AMD)**
- **Biased sampling of reaction coordinate(s)**
  - Fixed or restrained to multiple reaction coordinate values (windows): **blue moon, umbrella sampling (US)**
  - Adaptive bias potential automatically samples the entire reaction coordinate in one simulation: **metadynamics (metaD), AWH**

Types of enhanced sampling 2

- **Alchemical methods**
  - Slightly different implementations to analyze the results: **Thermodynamic integration (TI)**, **free-energy perturbation (FEP)**, (multi-state) **Bennett acceptance ratio (BAR)**
- **Simulations guided by structural restraints (limited phase space)**
  - NMR-based **distance restraints** between certain atoms
  - **Density-guided simulations** based on (cryo-)electron microscopy

Types of enhanced sampling 3

- **Coarse-graining / United atom approaches**
  - Sacrifice chemical specificity for faster dynamics:
    - Smoother energy landscape, larger time step, fewer particles
- **Multi-scale simulations / hybrid approaches**
  - Resolution transformations / multiple resolutions in one system
- **Accelerating / removing certain degrees of freedom**
  - Constraints, virtual sites, hydrogen mass repartitioning

Free Sampling of Reaction Coordinate(s)

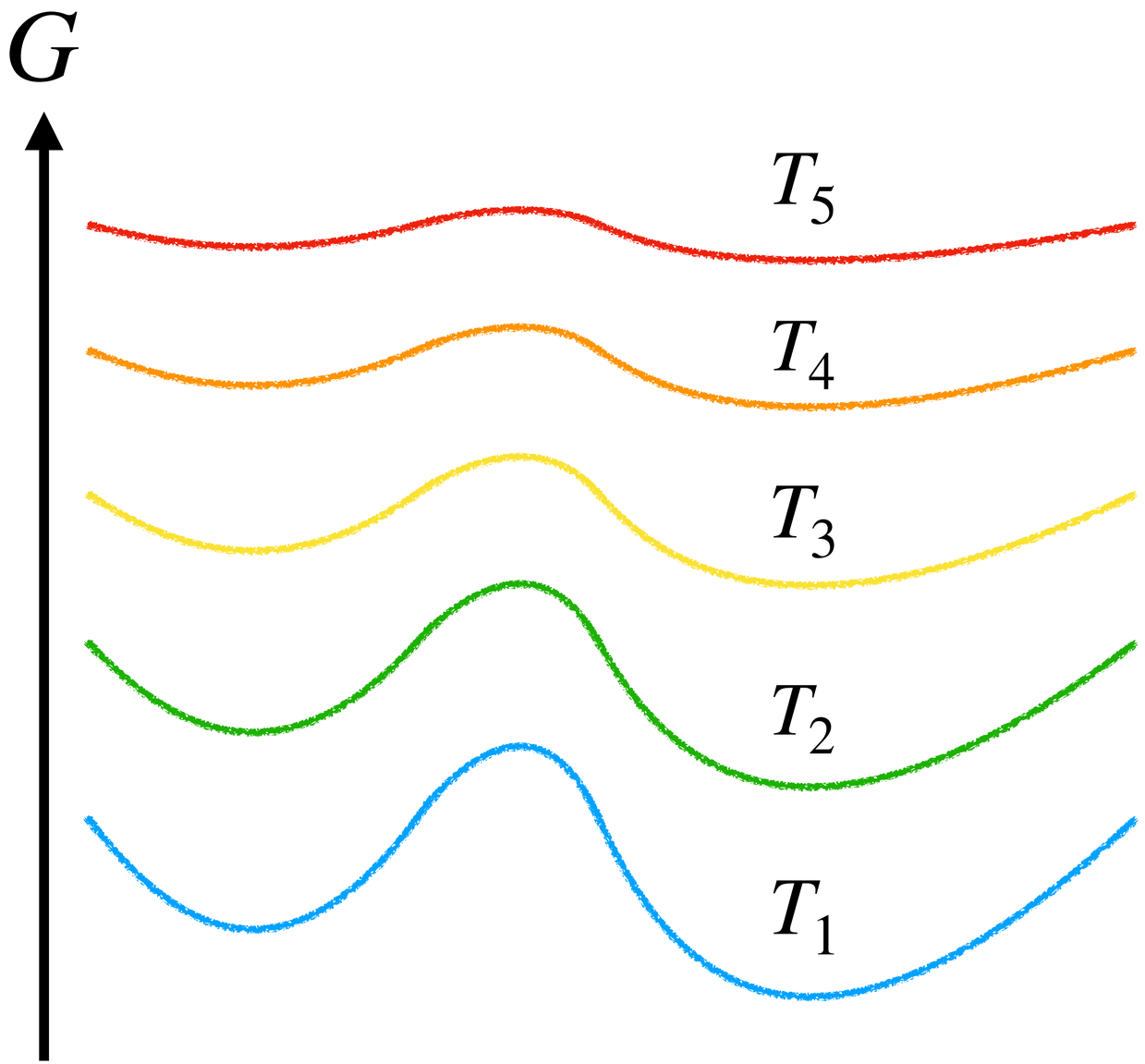
Replica Exchange 1

Heat up the system

- Free energy barriers crossed easier at higher temperatures
- We are usually interested in the behavior at physiological / room temperature

→ Perform multiple simulations at different temperatures!

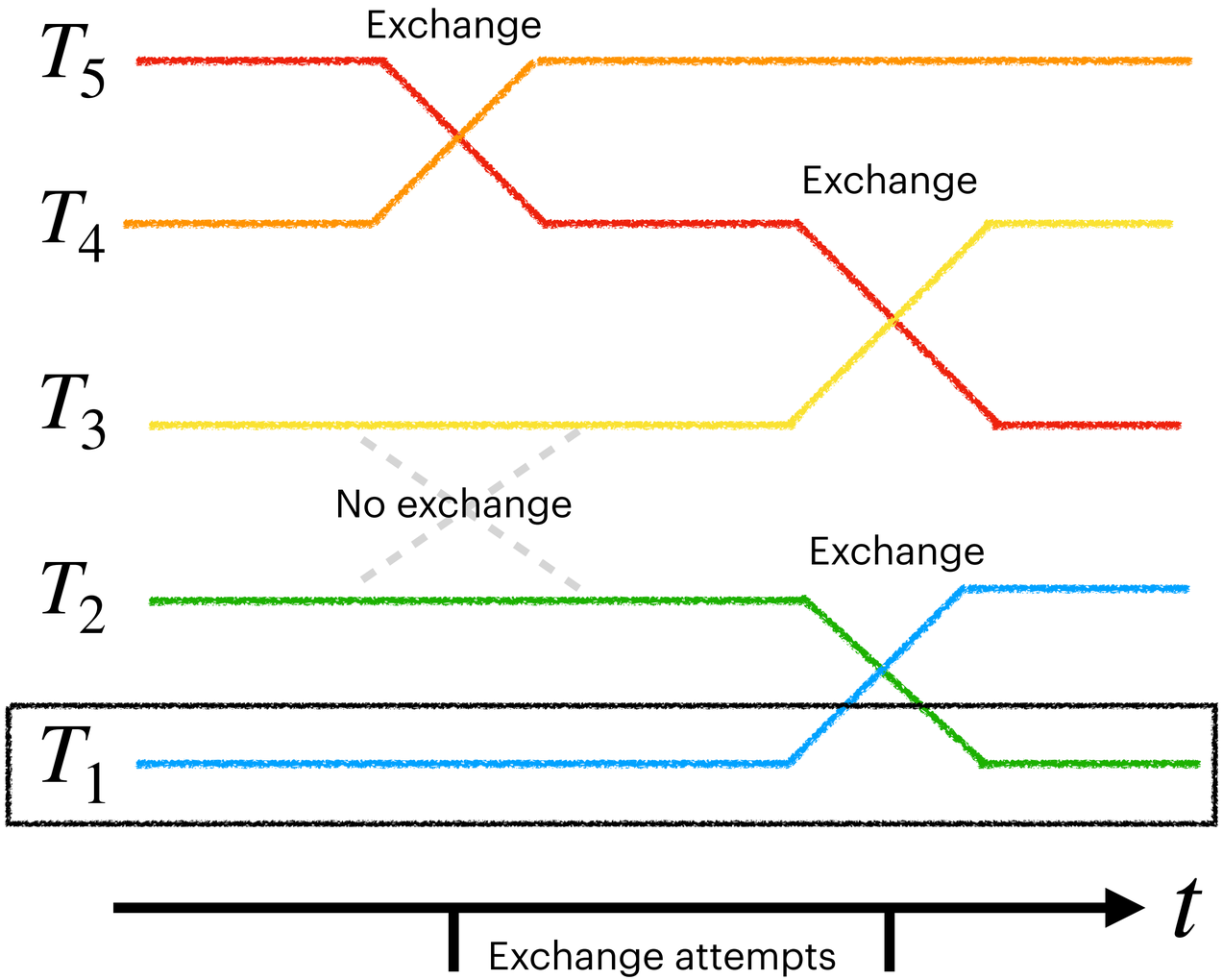
→ Feeds new conformations to the temperature of interest



Replica Exchange 2

Heat up the system

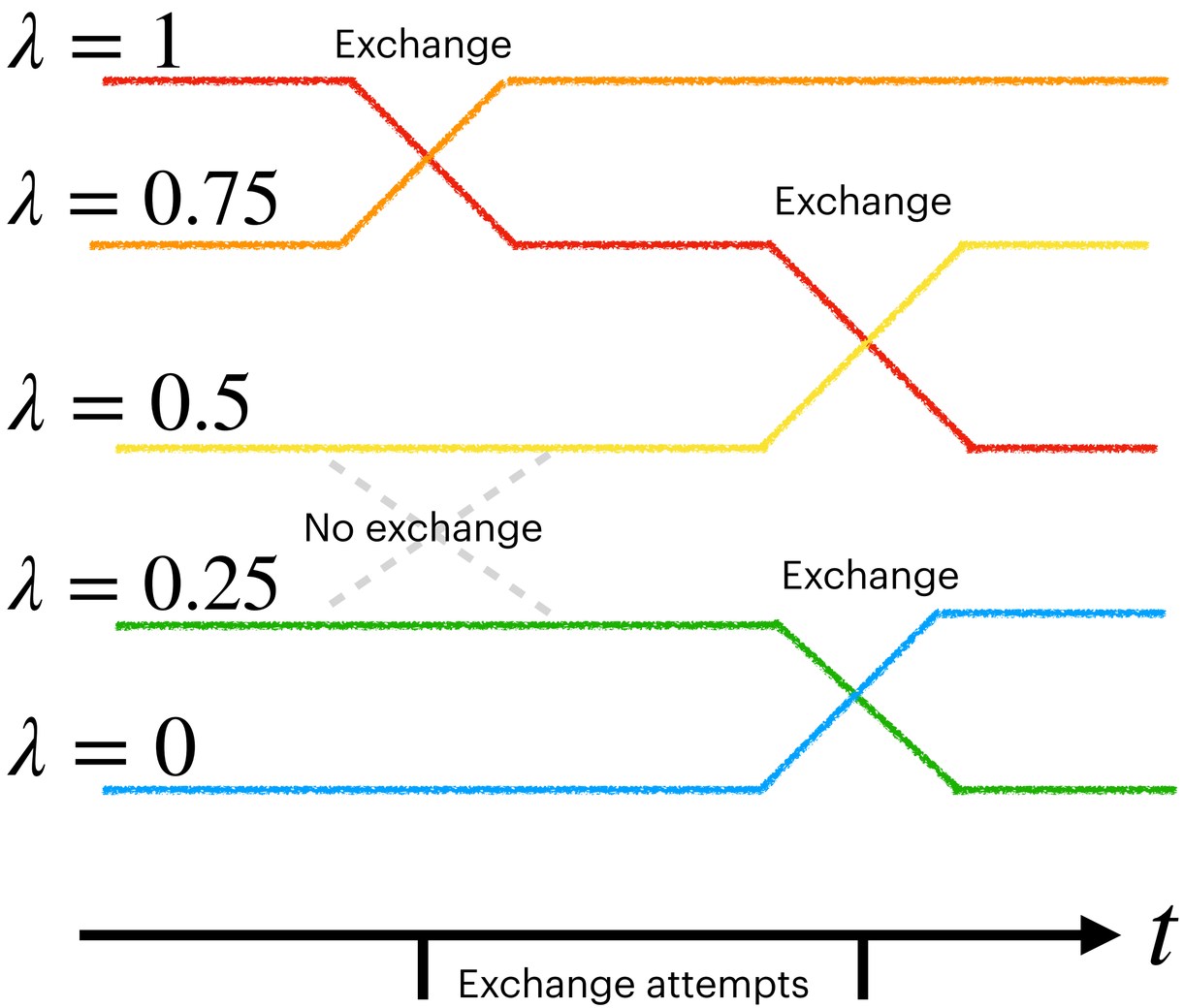
- Exchange coordinates if the energies are reasonably close
- Obtain equilibrium distribution at the temperature of interest
- For larger systems, need to simulate up to dozens of temperatures simultaneously
- Trajectories at each temperature will be discontinuous



Replica Exchange 3

Lower the interactions

- **Hamiltonian replica exchange:** Overcome energy barriers with scaled-down interactions
  - Alchemical reaction coordinate
  - This is also used to boost the convergence of alchemical calculations



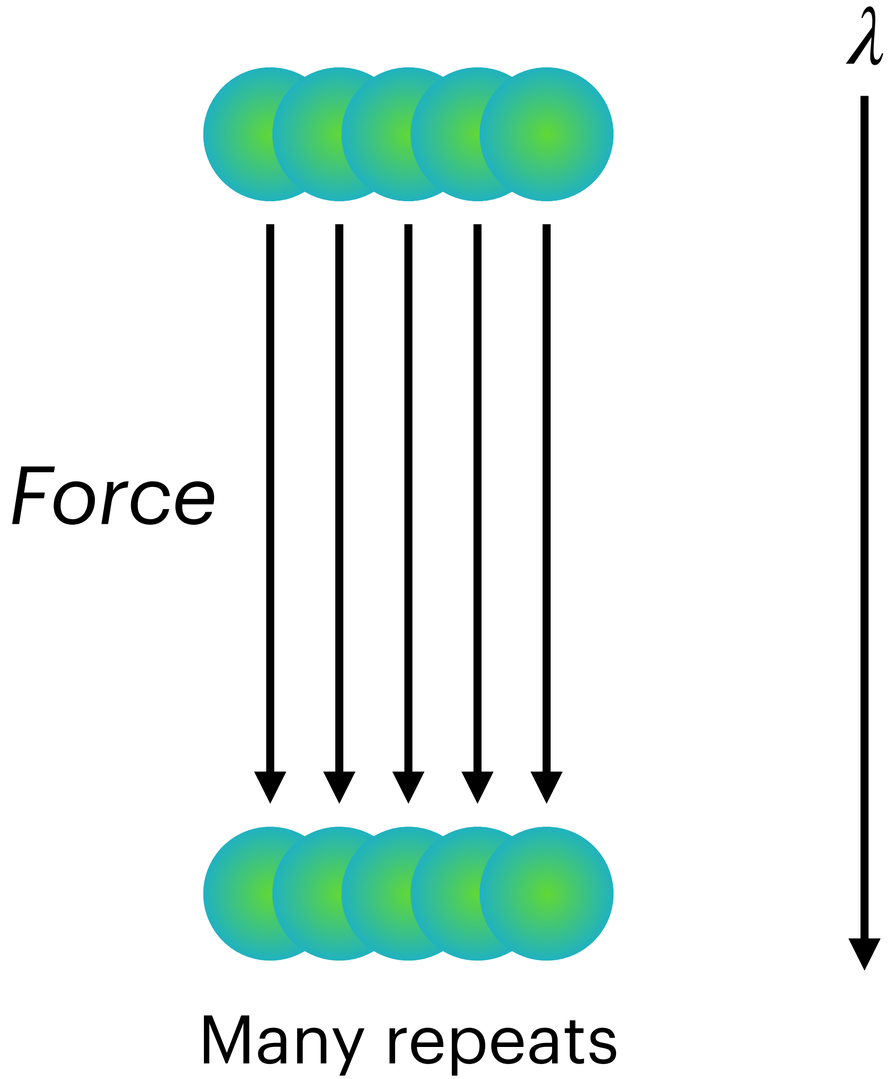
Biased Sampling of Reaction Coordinate(s)

Jarzynski equation

- Derived in 1996!
- Non-equilibrium technique
- Exponentially averaged work equals the exponential of ΔG
- Requires a large number of (short) simulations to converge

$$W \geq \Delta G, \text{ yet}$$

$$e^{-\beta \Delta G} = e^{-\overline{\beta W}}$$



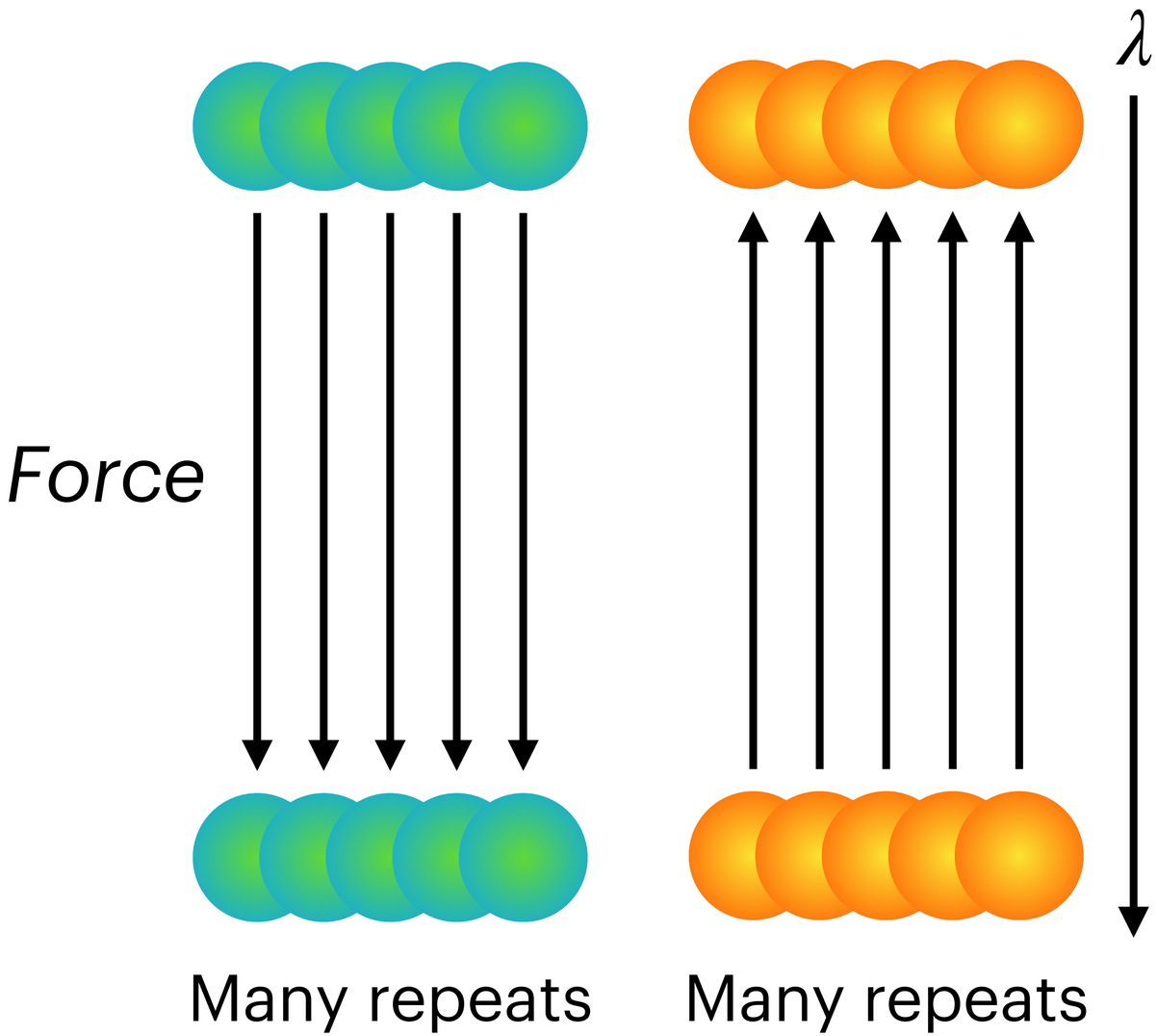
Crooks equation

- Derived in 1998!
- Non-equilibrium technique
- *W* in forward and backward directions is equal when $W = \Delta G$

$$\frac{P(W)_{A\to B}}{P(W)_{B\to A}} = e^{\beta\left(W_{A\to B} - \Delta G\right)}$$

- From a large number of simulations, intersect of the work distributions
- Also alchemical transitions

$$W \geq \Delta G$$

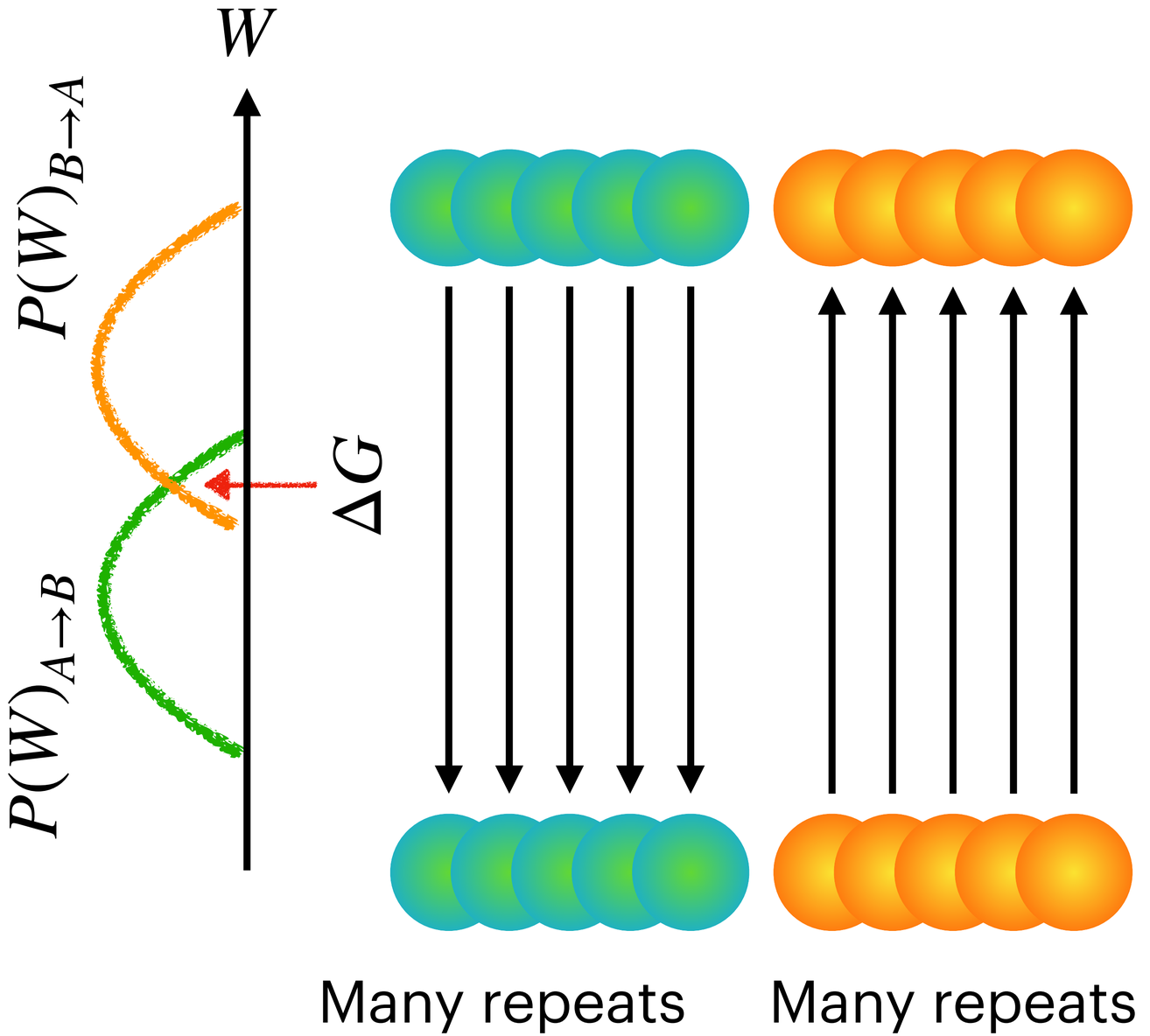


Crooks equation

- Derived in 1998!
- Non-equilibrium technique
- W in forward and backward directions is equal when $W = \Delta G$

$$\frac{P(W)_{A\to B}}{P(W)_{B\to A}} = e^{\beta\left(W_{A\to B} - \Delta G\right)}$$

- From a large number of simulations, intersect of the work distributions
- Also alchemical transitions

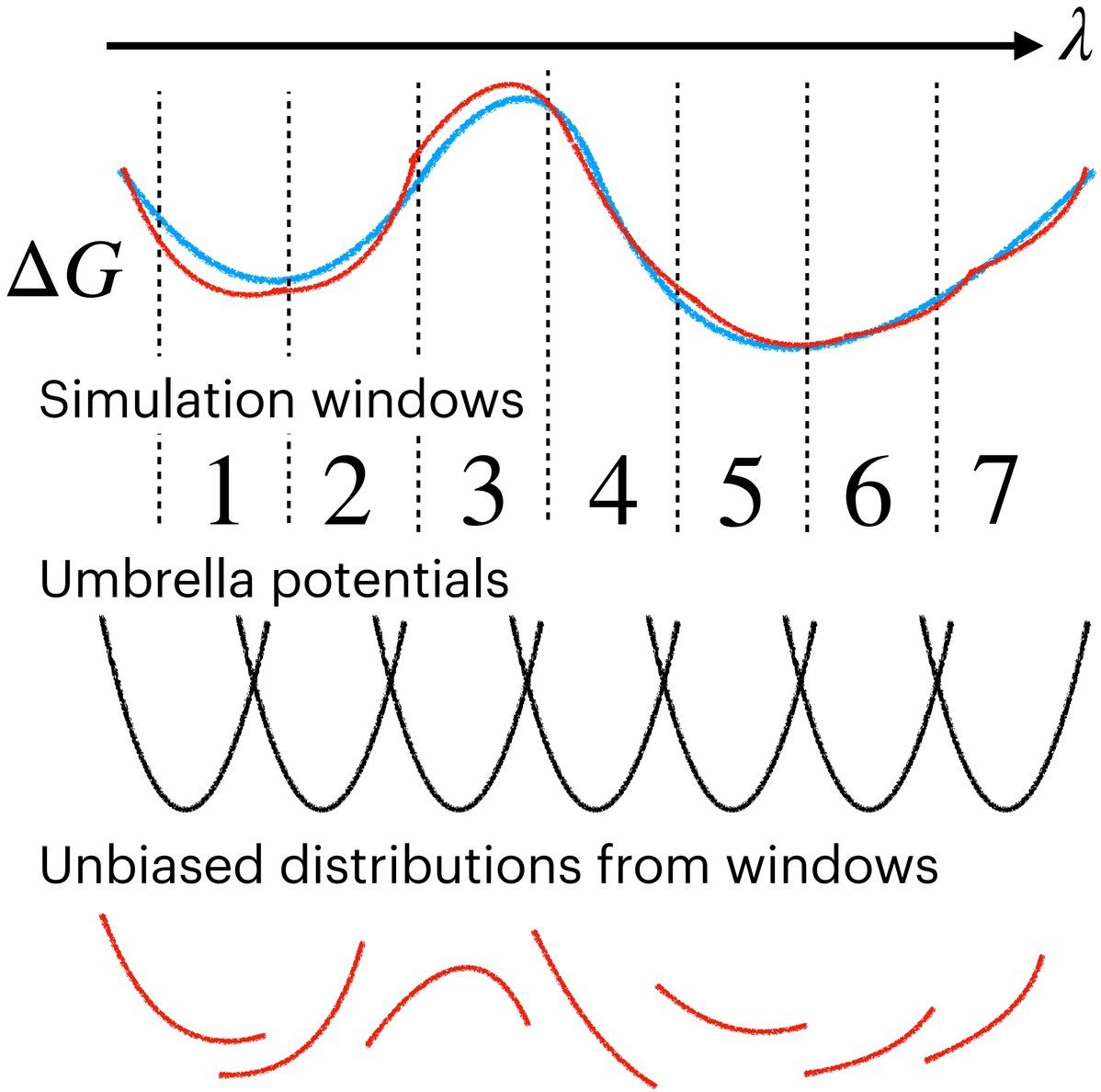


Umbrella Sampling

- Harmonic potential to keep system in desired region of the reaction coordinate

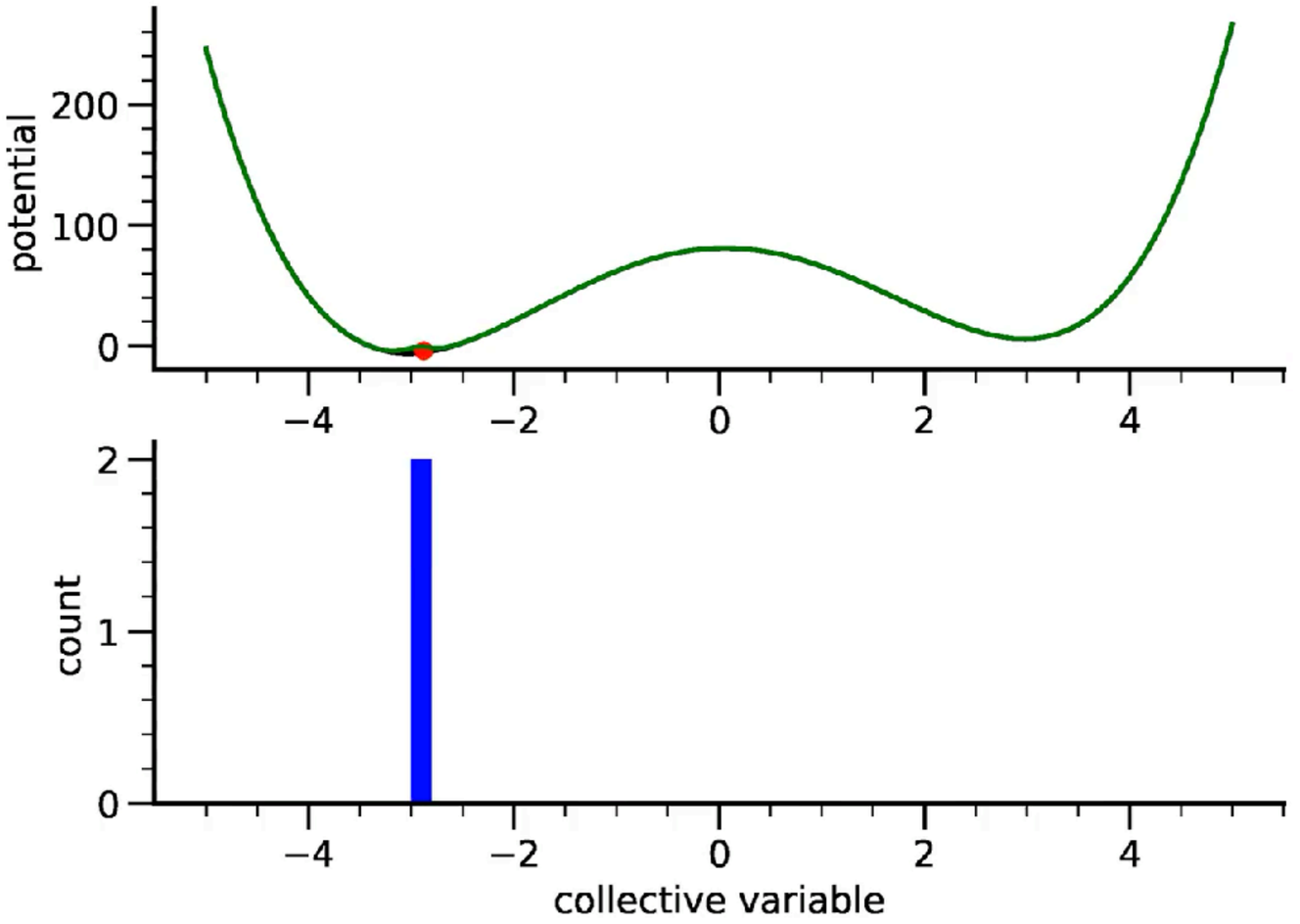
$$U(x) = k(x - x_0)^2$$

- Simulate a number of such windows at different reaction coordinate values
- Force constant important for computational efficiency:
  - Large k: no window overlap
  - Small k: insufficient sampling



Metadynamics 1

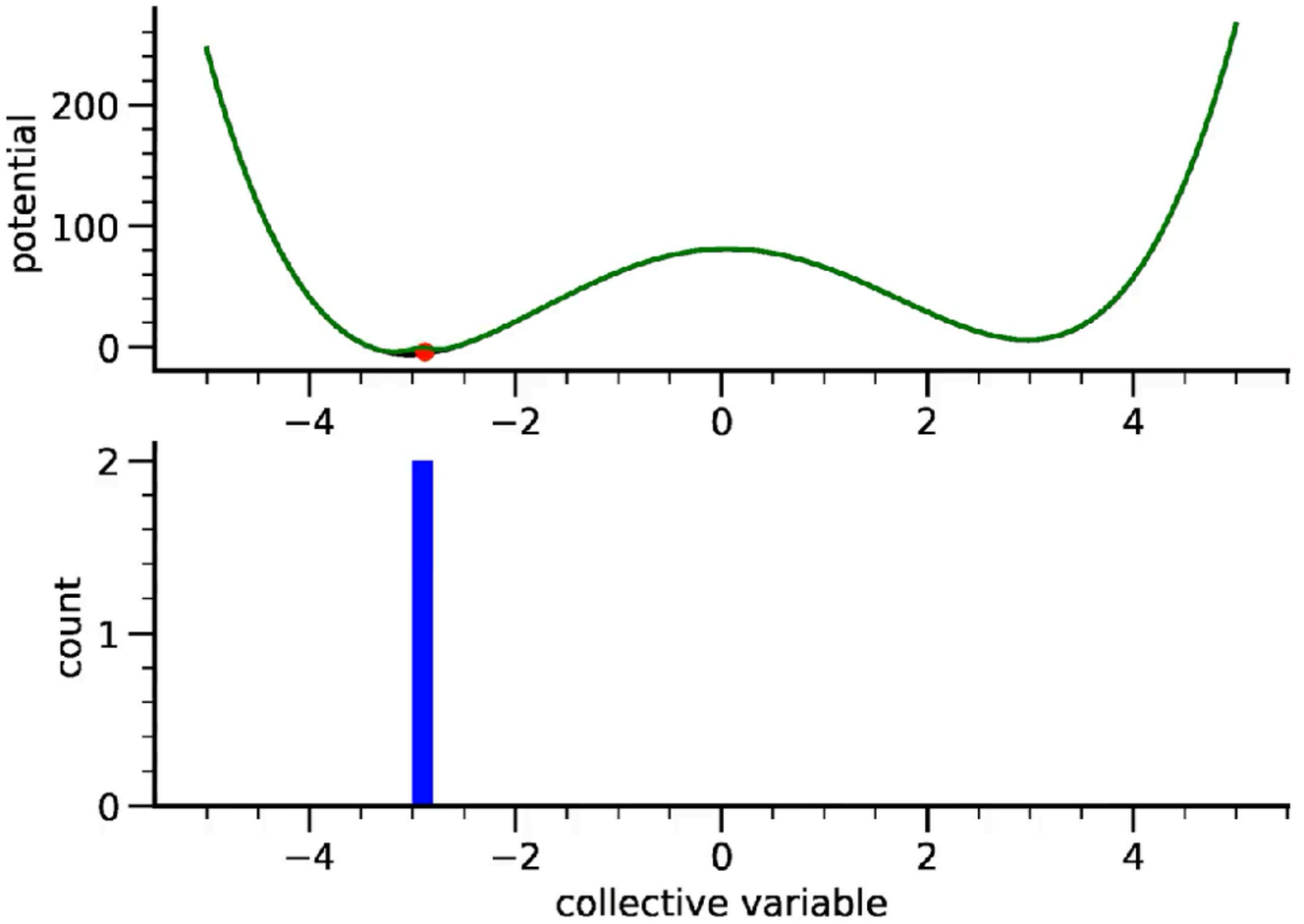
- Gaussian potentials are distributed to visited values of the reaction coordinate
- This forces the system to explore unvisited states
- The bias eventually mimics the inverse of the free energy
- Popular implementation in PLUMED compatible with many MD engines



Bougui505: Metadynamics toy model illustration, https://en.wikipedia.org/wiki/File:MetaD.webm

Metadynamics 2

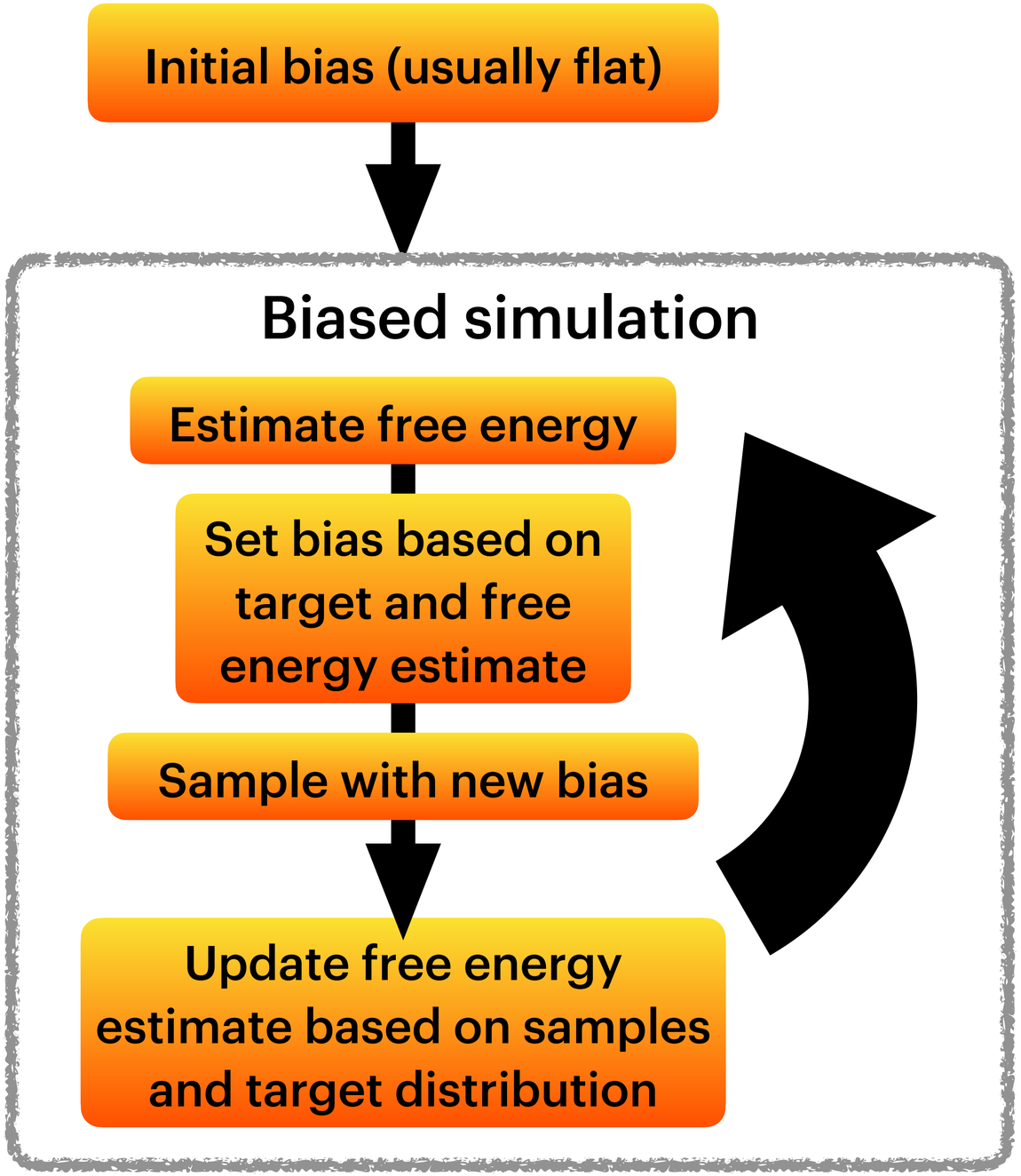
- Trivial parallelization using multiple walkers that accumulate a common bias
- Doesn’t converge!
- In **well-tempered metadynamics** the deposited Gaussians decrease in size
  - Convergence as bias eventually becomes flat



Bougui505: Metadynamics toy model illustration, https://en.wikipedia.org/wiki/File:MetaD.webm

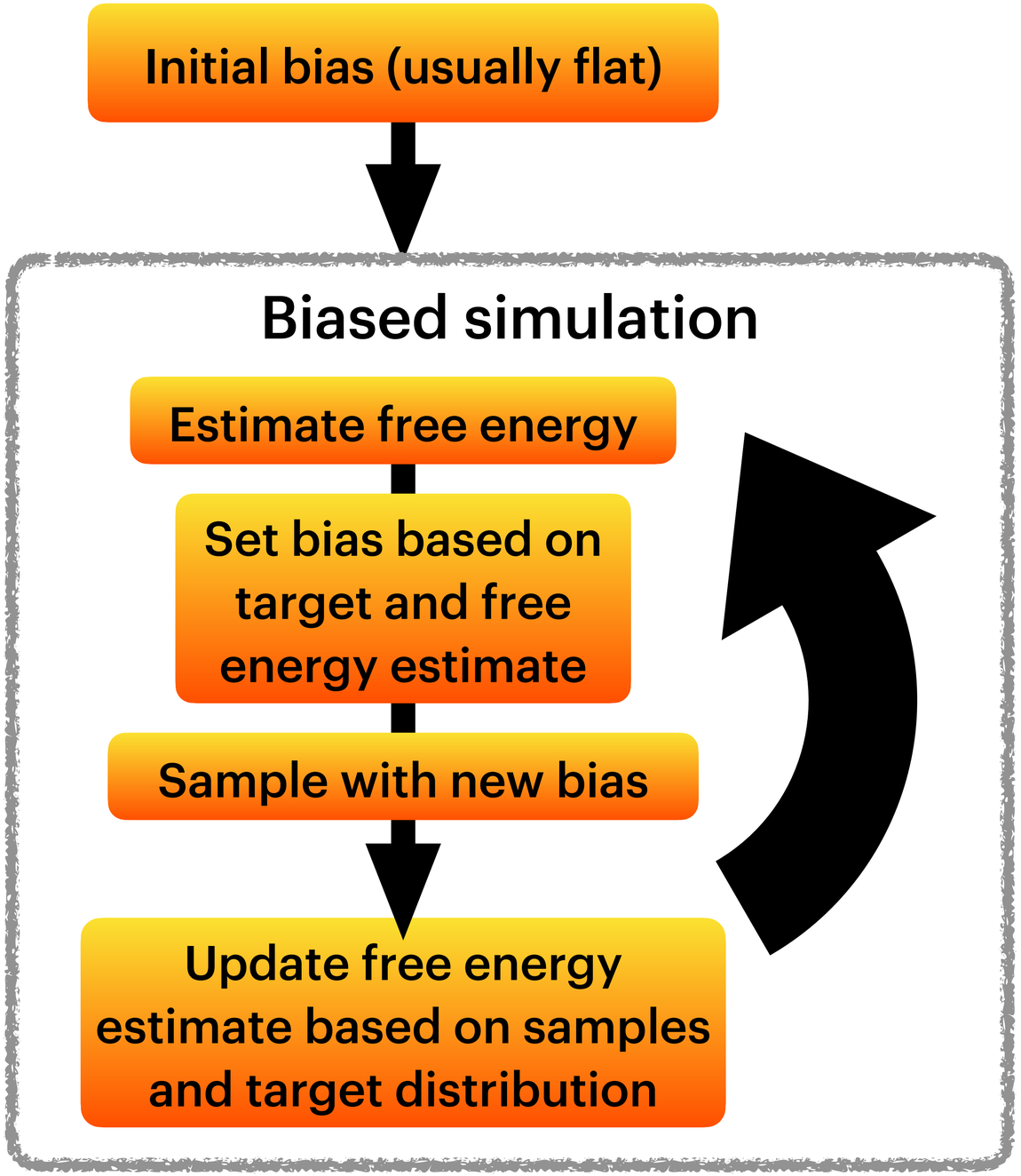
Accelerated Weight Histogram (AWH) 1

- Similar concept-wise to metaD: History-dependent bias
- Target distribution is an input (often flat, not necessarily)
- Bias potential adaptively determined based on target distribution and samples
- Always converges!
- Can be used for alchemical and physical reaction coordinates

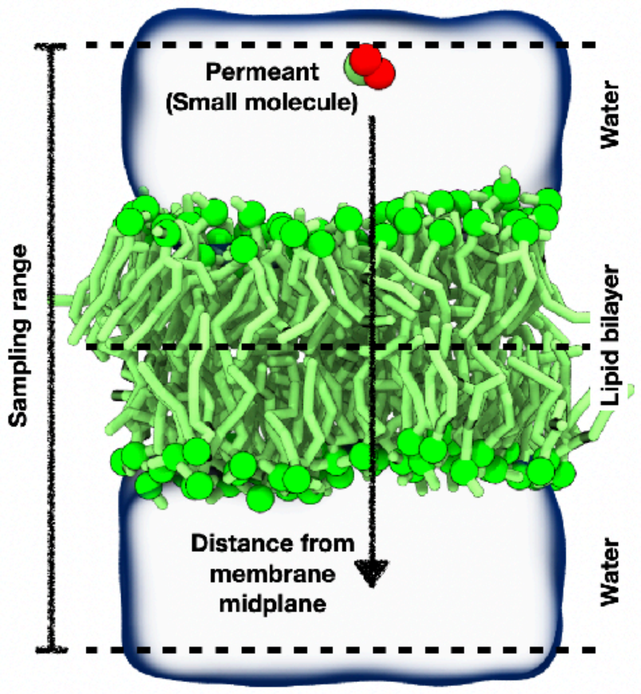


Accelerated Weight Histogram (AWH) 2

- Fast convergence and the independence of result on inputs
- Initially: exponential convergence
  - Covers reaction coordinate fast!
- Final state: converges as time$^{1/2}$
  - Ensures convergence
- Built into GROMACS: fast, parallelizes well, easy to install
- Parallelization with multiple walkers



AWH + Physical Reaction Coordinates




```
pull                    = yes          ; physical reaction coordinate
pull_ngroups            = 2            ; two groups: the permeant + reference
pull_ncoords            = 1            ; one coordinate defined by these two groups

pull-print-ref-value    = yes          ; we also print the coordinate of the reference group
pull-nstxout            = 5000         ; frequency of reaction coordinate output
pull-nstfout            = 0            ; no forces printed

pull_group1_name        = LIPIDS       ; reference group
pull_group2_name        = PERMEANT     ; permeant, renamed with a script

pull-group1-pbcatom     = 1

pull_coord1_type                = external-potential ; we use external potential from AWH
pull_coord1_potential_provider  = AWH
pull_coord1_geometry            = direction          ; direction allows negative distances
pull_coord1_groups              = 1 2                ; distance of reference + permeant
pull_coord1_dim                 = N N Y              ; we output the Z distance. For distance
                                                     ; pull geometry, this sets how distance
                                                     ; is calculated (1D, 2D, 3D)

pull_coord1_vec                 = 0 0 1              ; our pull direction is along Z axis
```

V. Lindahl, J. Lidmar, and B. Hess. *Accelerated weight histogram method for exploring free energy landscapes. J. Chem. Phys. 141, 044110 (2014)*

AWH + Physical Reaction Coordinates

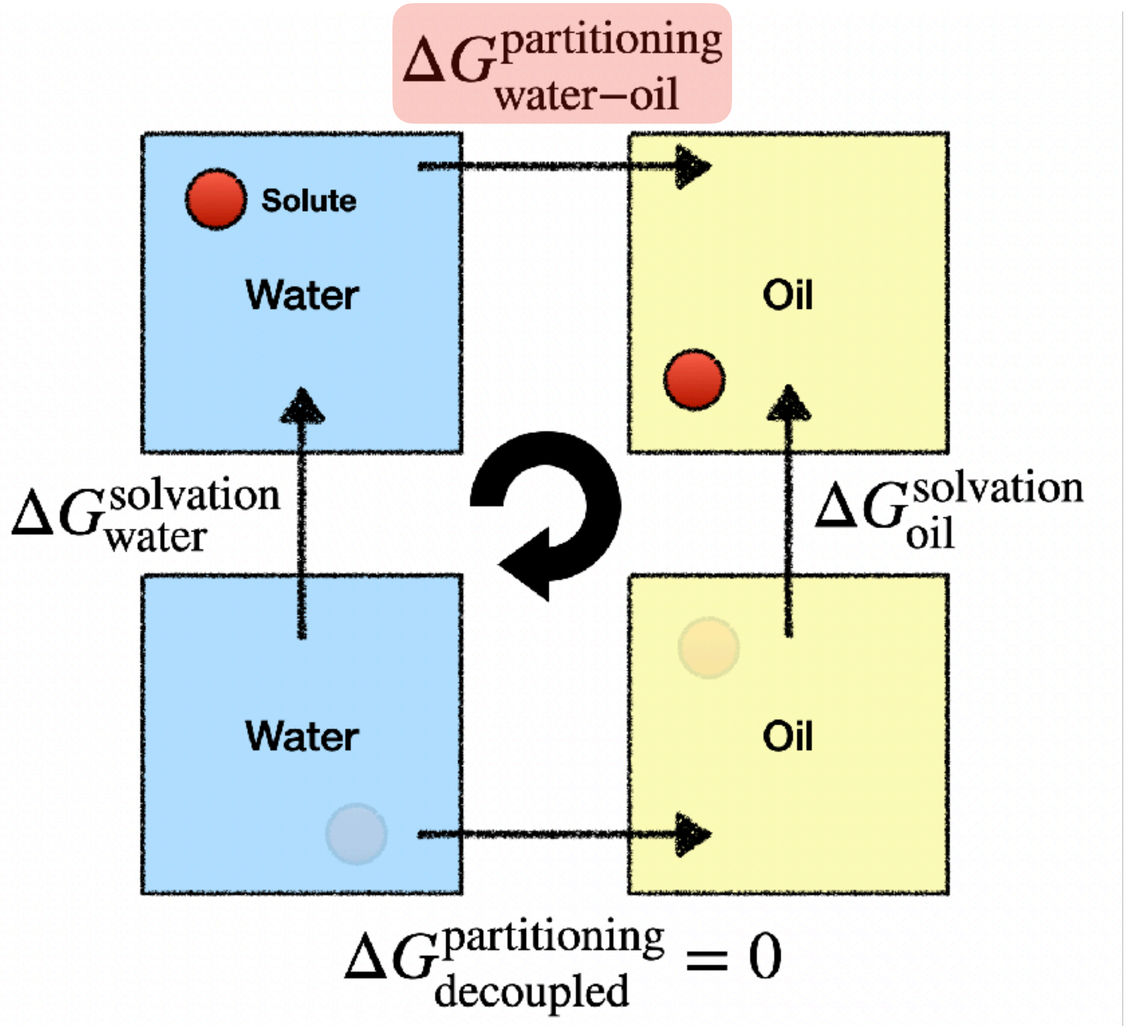
```
awh                                 = yes          ; turn AWH on
awh-potential                       = convolved    ; shape of AWH potential, default for
                                                   ; physical reaction coordinates
awh-nstout                          = 40000        ; frequency of xvg files from analysis
awh-nbias                           = 1            ; we only bias one coordinate

awh1-ndim                           = 1            ; dimensionality of the reaction coordinate
awh1-target                         = cutoff       ; we limit the sampling of very high barriers
awh1-target-cutoff                  = 40           ; very high barriers = 40 kJ/mol
awh1-error-init                     = 10           ; estimate of initial error, sets the initial
                                                   ; bias rate with diffusion parameter below
awh1-growth                         = exp-linear   ; two states for faster convergence

awh1-dim1-coord-provider            = pull         ;
awh1-dim1-coord-index               = 1            ; for the first (only) AWH potential,
                                                   ; the first (only) dimension is provided by
                                                   ; the first (only) pull coordinate
awh1-dim1-start                     = -4.5         ; range of sampled values (z coordinates)
awh1-dim1-end                       = 4.5          ;
awh1-dim1-force-constant            = 1e6          ;
                                                   ;
                                                   ;
awh1-dim1-diffusion                 = 1e-4         ; sets the initial bias rate together with
                                                   ; awh1-error-init
```

V. Lindahl, J. Lidmar, and B. Hess. *Accelerated weight histogram method for exploring free energy landscapes. J. Chem. Phys. 141, 044110 (2014)*

Alchemical Methods

Key Principle

- Alchemical reaction coordinate λ (potential function depends on it)
- Host-guest problems
- Free energy differences:
  - Solvation free energies
  - Relative binding affinities
  - Effects of mutations
- **Thermodynamic cycles** often used to design simulations

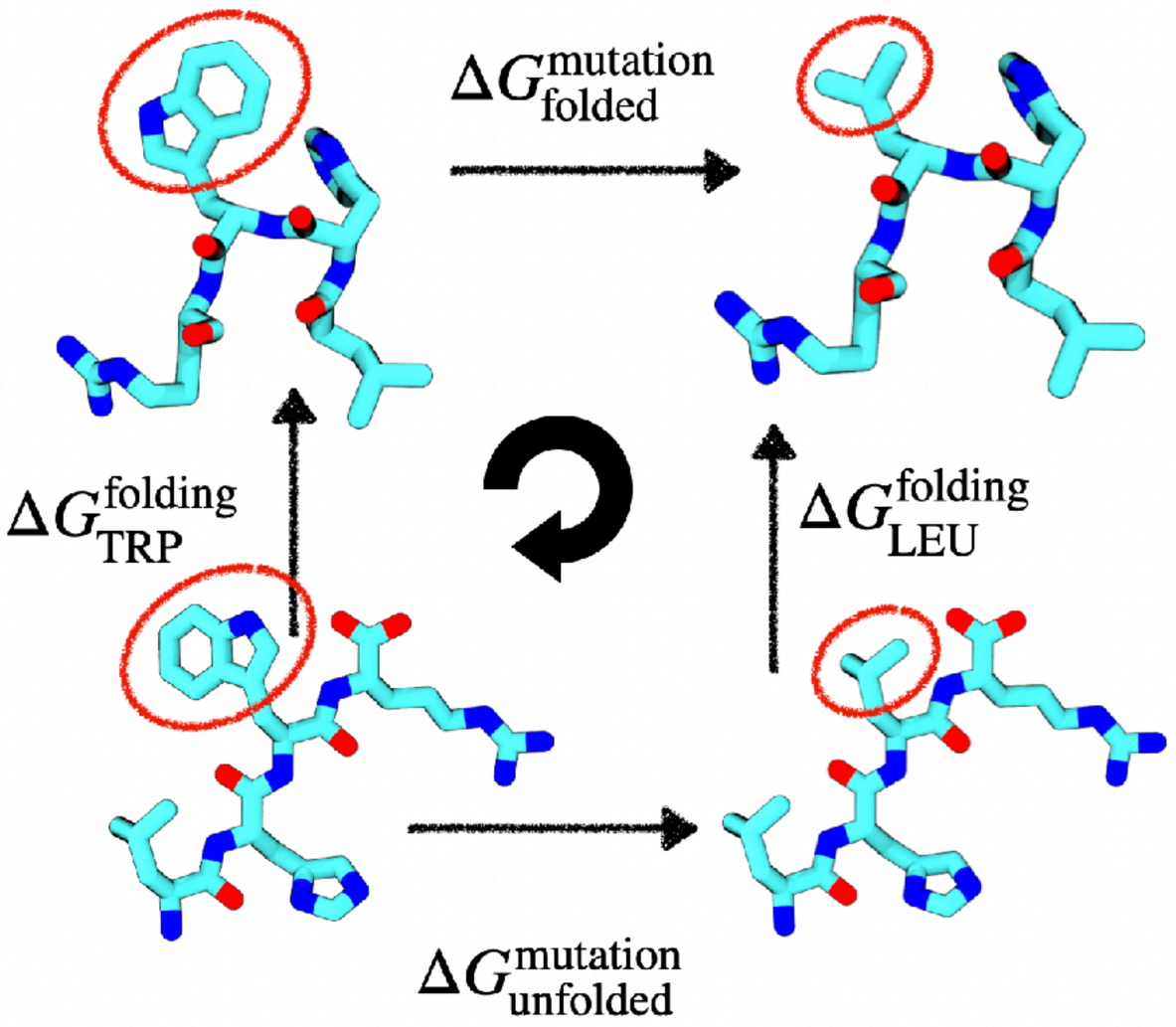


$$\Delta G^{\text{solvation}}_{\text{water}} + \Delta G^{\text{partitioning}}_{\text{water–oil}} - \Delta G^{\text{solvation}}_{\text{oil}} + \Delta G^{\text{partitioning}}_{\text{decoupled}} = 0$$

$$\Delta G^{\text{partitioning}}_{\text{water–oil}} = \Delta G^{\text{solvation}}_{\text{water}} - \Delta G^{\text{solvation}}_{\text{oil}}$$

Key Principle

- Alchemical reaction coordinate λ (potential function depends on it)
- Host-guest problems
- Free energy differences:
  - Solvation free energies
  - Relative binding affinities
  - Effects of mutations
- **Thermodynamic cycles** often used to design simulations

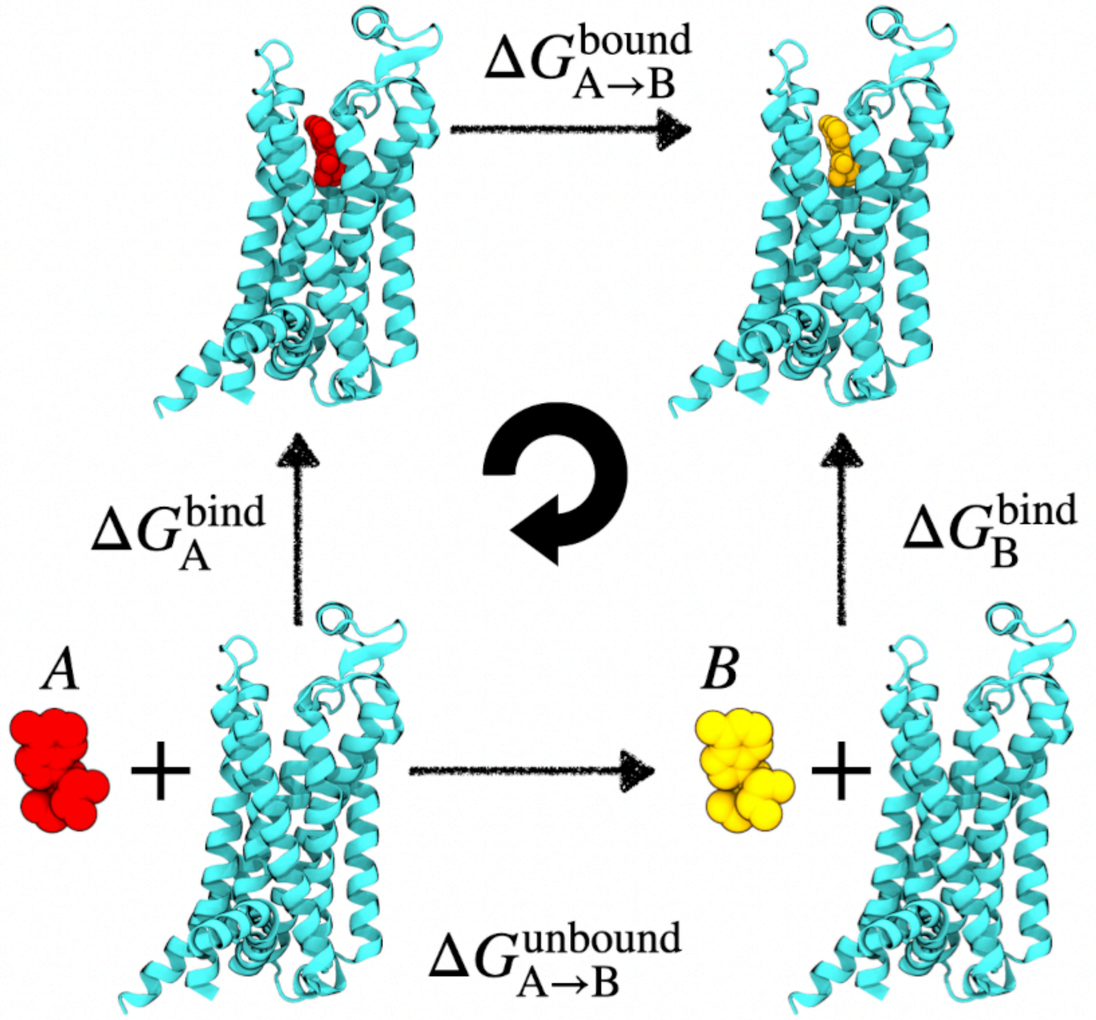


$$\Delta G^{\text{mutation}}_{\text{folded}} - \Delta G^{\text{folding}}_{\text{LEU}} - \Delta G^{\text{mutation}}_{\text{unfolded}} + \Delta G^{\text{folding}}_{\text{TRP}} = 0$$

$$\Delta G^{\text{folding}}_{\text{TRP}} - \Delta G^{\text{folding}}_{\text{LEU}} = \Delta G^{\text{mutation}}_{\text{folded}} - \Delta G^{\text{mutation}}_{\text{unfolded}}$$

Key Principle

- Alchemical reaction coordinate λ (potential function depends on it)
- Host-guest problems
- Free energy differences:
  - Solvation free energies
  - Relative binding affinities
  - Effects of mutations
- **Thermodynamic cycles** often used to design simulations

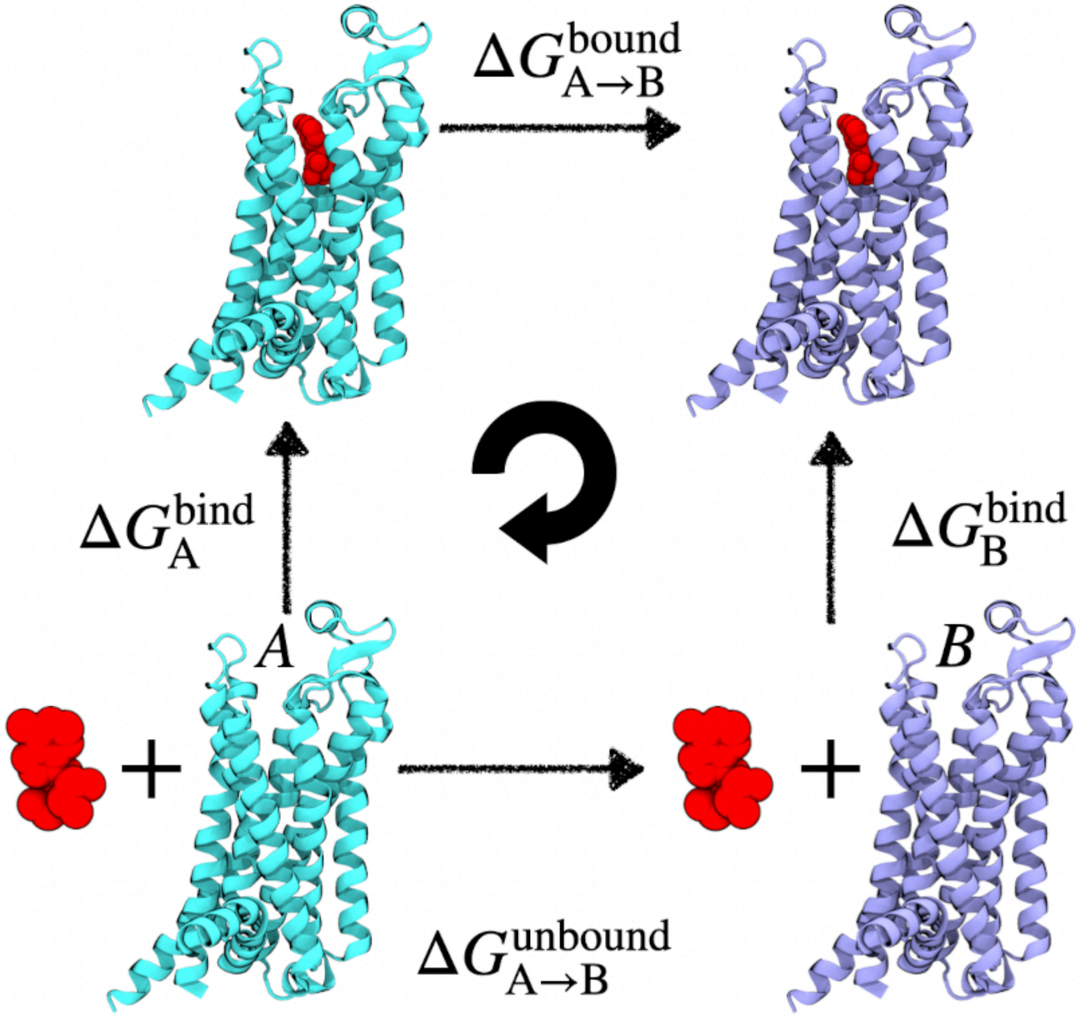


$$\Delta G_A^{\text{bind}} + \Delta G_{A\to B}^{\text{bound}} - \Delta G_B^{\text{bind}} - \Delta G_{A\to B}^{\text{unbound}} = 0$$

$$\Delta G_A^{\text{bind}} - \Delta G_B^{\text{bind}} = \Delta G_{A\to B}^{\text{unbound}} - \Delta G_{A\to B}^{\text{bound}}$$

Key Principle

- Alchemical reaction coordinate λ (potential function depends on it)
- Host-guest problems
- Free energy differences:
  - Solvation free energies
  - Relative binding affinities
  - Effects of mutations
- **Thermodynamic cycles** often used to design simulations



$$\Delta G_A^{\text{bind}} + \Delta G_{A\to B}^{\text{bound}} - \Delta G_B^{\text{bind}} - \Delta G_{A\to B}^{\text{unbound}} = 0$$

$$\Delta G_A^{\text{bind}} - \Delta G_B^{\text{bind}} = \Delta G_{A\to B}^{\text{unbound}} - \Delta G_{A\to B}^{\text{bound}}$$

Basic idea the same, different ways to analyze

Thermodynamic integration (TI):

Simulate at multiple λs and store the values of the analytical derivative of the dH/dλ.
Numerically integrate <dH/dλ > over λs.

Slow growth:

Same as TI but with λ(t) instead of multiple λs.

Accelerated weight histogram (AWH):

Adaptive biasing potential applies weights to already visited λ states to push the system to sample the entire range of λs

Free energy perturbation (FEP):

A.k.a. Zwanzig equation / Exponential averaging.
Estimate free energy difference from the exponential average of energy differences of two states.
Needs overlap: split λ to smaller intervals of states.

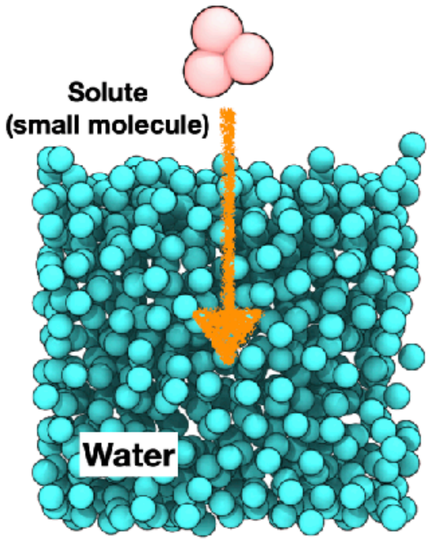
Bennett Acceptance Ratio (BAR):

Maximum likelihood estimator.
Free energy difference from energies sampled in one neighbouring state.

Multistate BAR (MBAR):

... energies sampled in all other states (λs).

AWH + Alchemical Reaction Coordinates




```
free-energy              = yes          ; turn on alchemistry
couple-lambda0           = none         ; no non-bonded interactions at lambda=0
couple-lambda1           = vdwq         ; LJ + electrostatics on at lambda=1
couple-moltype           = TOCOUPLE     ; molecule whose interactions coupled to lambda
couple-intramol          = no           ; no decoupling of bonded terms
                                        ; thus lambda=0 corresponds to the molecule in vacuum
init-lambda-state        = 30           ; we start sampling at lambda index 30 (see below)
                                        ; the values below define the lambdas for the 30 windows
                                        ; LJ and electrostatics are decoupled separately
vdw_lambdas              = 1   1   1   1   1   1   1   1   1   1 1 0.95 0.9 0.85 0.8 0.75 0.7 0.65 0.6
coul_lambdas             = 1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0   0   0   0   0   0   0   0   0
calc-lambda-neighbors    = -1           ; calculate energy differences to all neighbours
sc_alpha                 = 0.5          ; soft-core interactions to avoid overlap issues ...
sc_sigma                 = 0.3          ; when the molecule is barely present ...
sc_power                 = 1            ; (LJ lambda is close to 0)
sc_coul                  = no           ;
```

M. Lundborg, J. Lidmar, and B. Hess. *The accelerated weight histogram method for alchemical free energy calculations. J. Chem. Phys.* 154, 204103 (2021)

AWH + Alchemical Reaction Coordinates

```
awh                         = yes          ; enable AWH
awh-potential               = umbrella     ; harmonic umbrella with Monte Carlo sampling
                                           ; (must be used now with alchemistry)
awh-nstout                  = 50000        ; frequency of updating the edr file
awh-nbias                   = 1            ; we have 1 bias (lambda)
awh-nstsample               = 100
awh-nsamples-update         = 10
awh1-error-init             = 10           ; dictates initial convergence together with
                                           ; awh1-dim1-diffusion
awh1-equilibrate-histogram = no
awh1-target                  = constant    ; target distribution is flat
awh1-growth                  = exp-linear  ; we use the two-stage approach for convergence
awh1-ndim                    = 1
awh1-dim1-coord-provider   = fep-lambda ; AWH uses alchemical reaction coordinate
awh1-dim1-coord-index      = 1
awh1-dim1-start              = 0           ; we sample lambdas from 0 to 30
awh1-dim1-end                = 30
awh1-dim1-diffusion         = 0.001        ; sets the convergence together with awh1-error-init
```

M. Lundborg, J. Lidmar, and B. Hess. *The accelerated weight histogram method for alchemical free energy calculations. J. Chem. Phys.* 154, 204103 (2021)

Summary

Slow or energetically costly process

Host/quest problem?

Only $\Delta\Delta G$ important?

- Draw a thermodynamic cycle and think what to simulate
- Alchemical reaction coordinate to decouple/mutate
- Sample and analyze the result with **AWH** / BAR / TI similar

Physical change?

Profile of ΔG important?

- Design a rection coordinate that describes the change
- Choose a biasing technique (**AWH**, US, non-equilibrium)
- Sample and unbias the result to recover the ΔG profile

Things worth checking out 1

- **Recent reviews on enhanced sampling/free energy calculations**
  - An extremely thorough review of the methods, including derivations.
    Living Journal concept so it hopefully gets updated with new methods:

    J. Hénin *et al.*, ***Enhanced Sampling Methods for Molecular Dynamics Simulations***,

    Living J. Comp. Mol. Sci. 4(1), 1583 (2022). DOI: 10.33011/livecoms.4.1.1583

  - A comprehensive review on the pitfalls in free energy calculations:

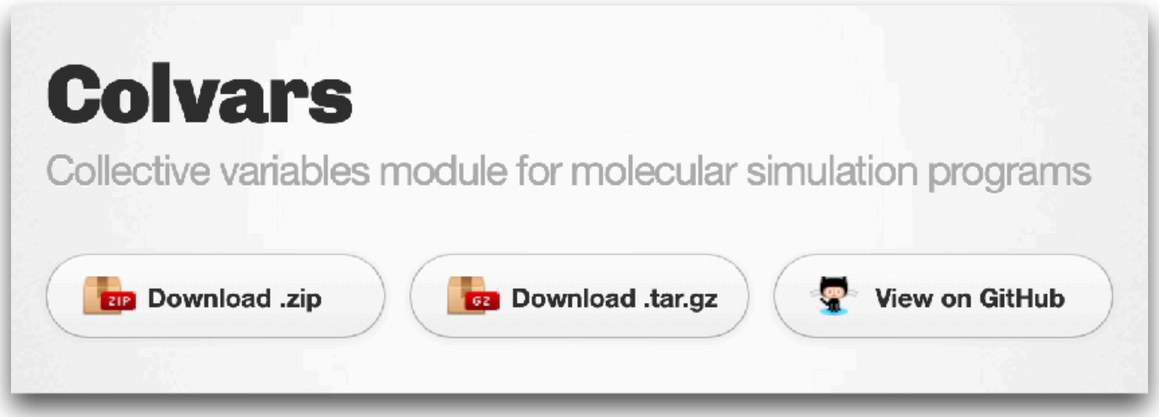
    E. Duboué-Dijon & J. Hénin, ***Building intuition for binding free energy calculations: Bound state definition, restraints, and symmetry,***

    J. Chem. Phys. 154, 204101 (2021). DOI: 10.1063/5.0046853

Things worth checking out 2

The Colvars module built into GROMACS, https://colvars.github.io/

- Recent demonstration: https://www.youtube.com/watch?v=-8l1Mt4XpVw
- Enables new collective variables to be used
  - Path collective variables (*e.g.*, for conformation changes)
  - Collective metrics (RMSD, radius of gyration,...)
  - Contacts (coordination number,...)
  - Tilt angles, rotations



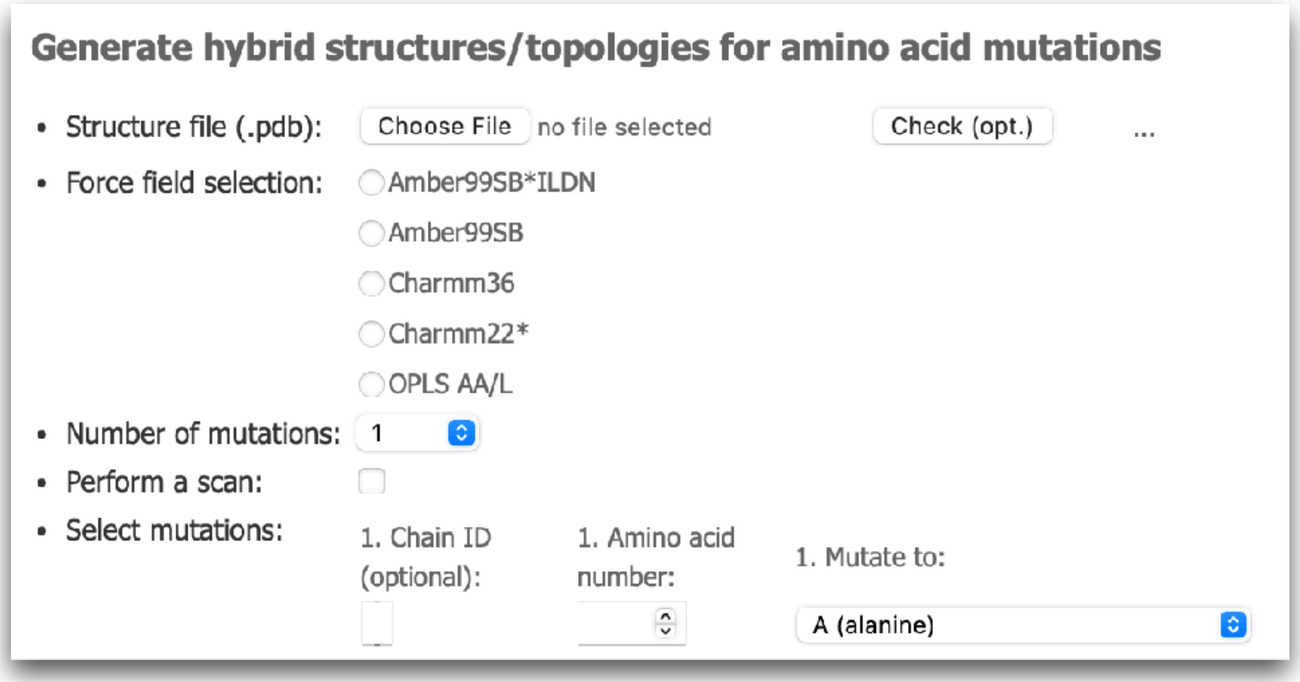
Giacomo Fiorin *et al.*, ***Using collective variables to drive molecular dynamics simulations***, Mol. Phys. 11, 3345–3362 (2013)

Things worth checking out 3

- **BioExcel webinars** on YouTube at https://www.youtube.com/@BioExcelCoE
- **pmx:** tool to set up alchemical calculations (protein mutations) in GROMACS
  - Great set of up-to-date tutorials available at http://pmx.mpibpc.mpg.de/

pmx protein

web server



Questions?