Structural Biology - BIO315

Master SV - Spring Semester
Lecture 8

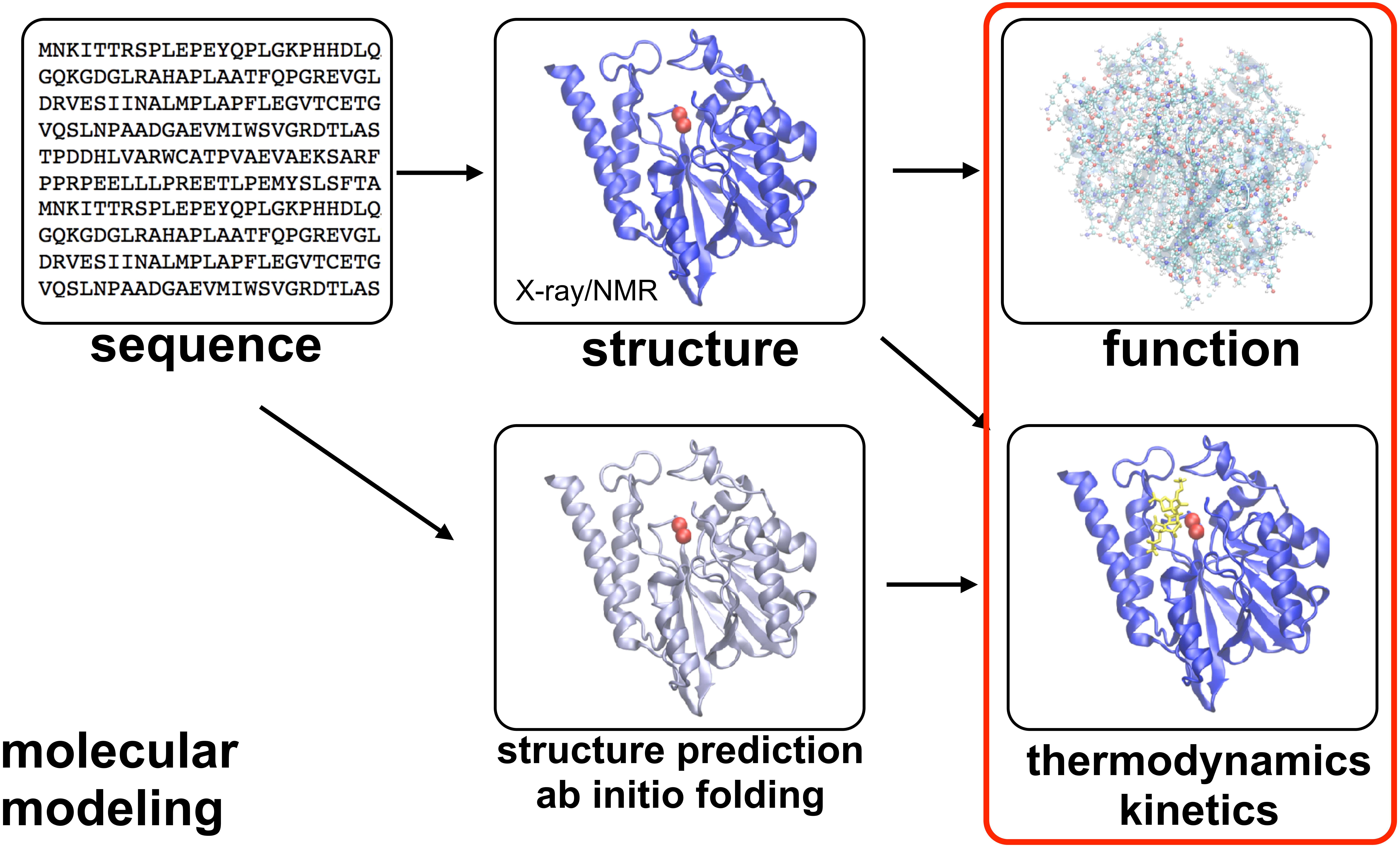
Matteo Dal Peraro
matteo.dalperaro@epfl.ch
AAB 048, phone: 31681

Outline of lecture 8:

- **energy minimization** techniques
  - steepest descent
  - conjugated gradient

- **introduction to Molecular Dynamics (MD)**
  - initialize the system
  - integration methods
  - choosing the correct time-step
  - calculation of relevant quantities
  - free-energy sampling
  - state-of-the-art of MD simulations
  - current limitations

Paradigm in Structural Biology



```
MNKITTRSPLEPEYQPLGKPHHDLQ
GQKGDGLRAHAPLAATFQPGREVGL
DRVESIINALMPLAPFLEGVTCETG
VQSLNPAADGAEVMIWSVGRDTLAS
TPDDHLVARWCATPVAEVAEKSARF
PPRPEELLLPREETLPEMYSLSFTA
MNKITTRSPLEPEYQPLGKPHHDLQ
GQKGDGLRAHAPLAATFQPGREVGL
DRVESIINALMPLAPFLEGVTCETG
VQSLNPAADGAEVMIWSVGRDTLAS
```

sequence

X-ray/NMR

structure

function

structure prediction
ab initio folding

thermodynamics
kinetics

molecular modeling

- knowledge-based: structural databases

- first principles:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{R},t) = \hat{H}\Psi(\mathbf{R},t)$$

What we need for modeling at the molecular mechanics (MM) level

For a molecular simulation or modeling one needs:

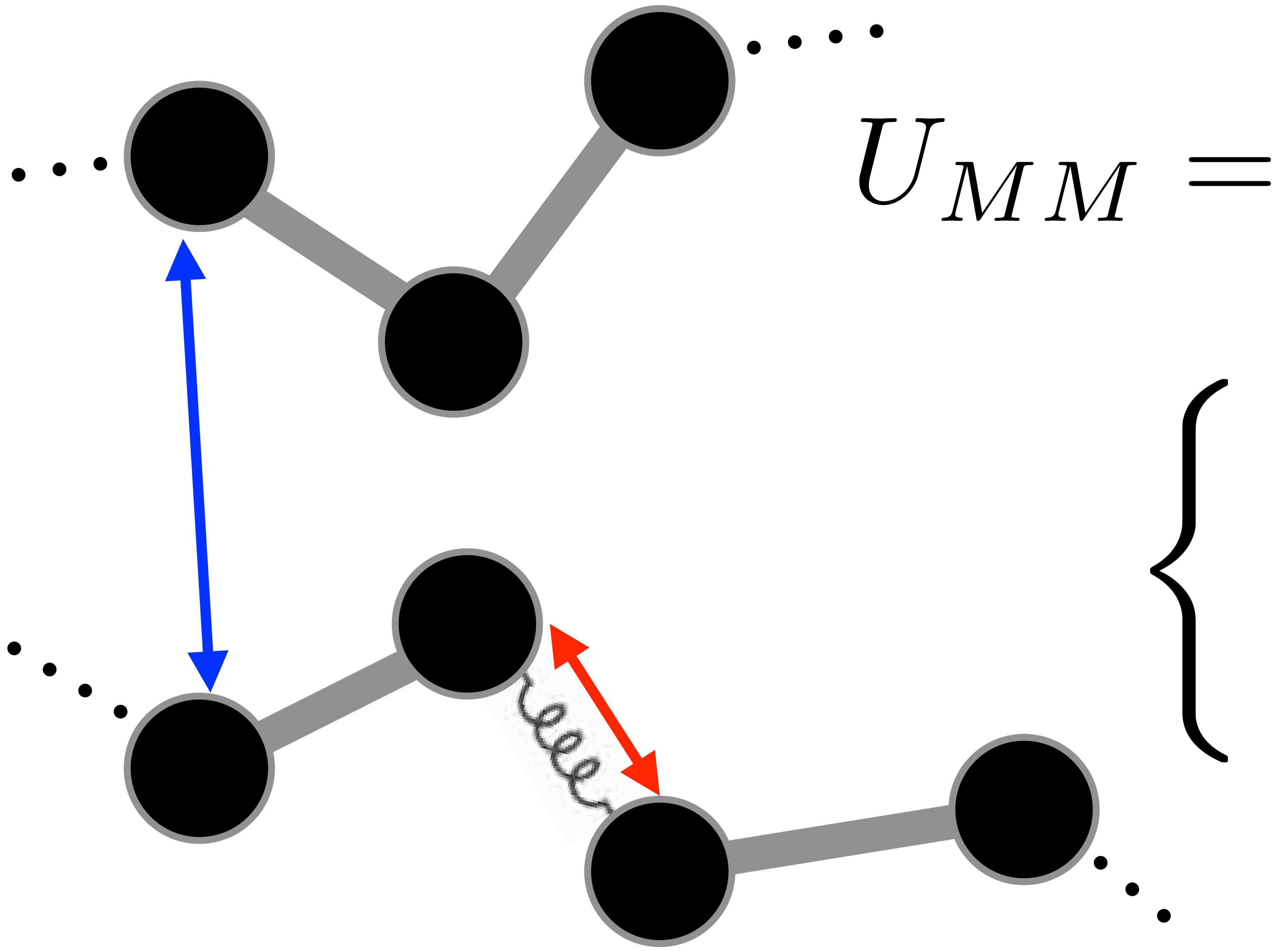
1. a **representation** of the biomolecules at a certain level of resolution (i.e. initial conditions)
2. a functional form for the **potential** energy for molecular mechanics (MM)
3. a search algorithm or optimizer/minimizer (**minimization** can be used to find favorable regions in the conformational space; **sampling** techniques to compute dynamics and thermodynamic quantities)

Molecular mechanics potentials

- **molecular mechanics** (MM) potential energy gives minimum-energy conformation of a molecule

- based on **physics**, but uses simplified “ball-and-spring” models (**classical** physics, *Newton equation*), which mask the quantum nature (*Schrodinger equation*)

- are **empirical**, i.e. calibrated to describe the quantum nature of chemical bonds and short-range interactions



$$U_{MM} = \boxed{U_{bonded}} + \boxed{U_{non-bonded}}$$

$$\begin{cases} U_{bonded} = U_{bond} + U_{angle} + U_{torsion} \\ U_{non-bonded} = U_{electrostatics} + U_{VdW} \end{cases}$$

Empirical potential energy function

$$U_{MM}(r) = \boxed{\sum_{bonds} \frac{k_b}{2}(r-r_0)^2 + \sum_{angles}\frac{k_\theta}{2}(\theta-\theta_0)^2 + \sum_{torsions,n}\frac{k_{\phi,n}}{2}[1+cos(n\phi-\delta)]} +$$

$$+\boxed{\sum_{i>j}^{N}\left(\frac{A}{r_{ij}^{12}}-\frac{C}{r_{ij}^{6}}\right)+\sum_{i>j}^{N}\frac{1}{4\pi\epsilon_0}\frac{q_iq_j}{r_{ij}}}$$

- large number of **parameters** fitted to represent experimental data or QM calculated quantities (usually structure and thermodynamic of small molecules)

- "trial and error" or least-squares fitting methods to converge to a consistent set of parameters

- coupling/correlation between parameters, thus parameterization of a **force field (FF)** is a global task

- assumption that parameters can be **transferable** to different contexts (specialized *vs.* generalized FF)

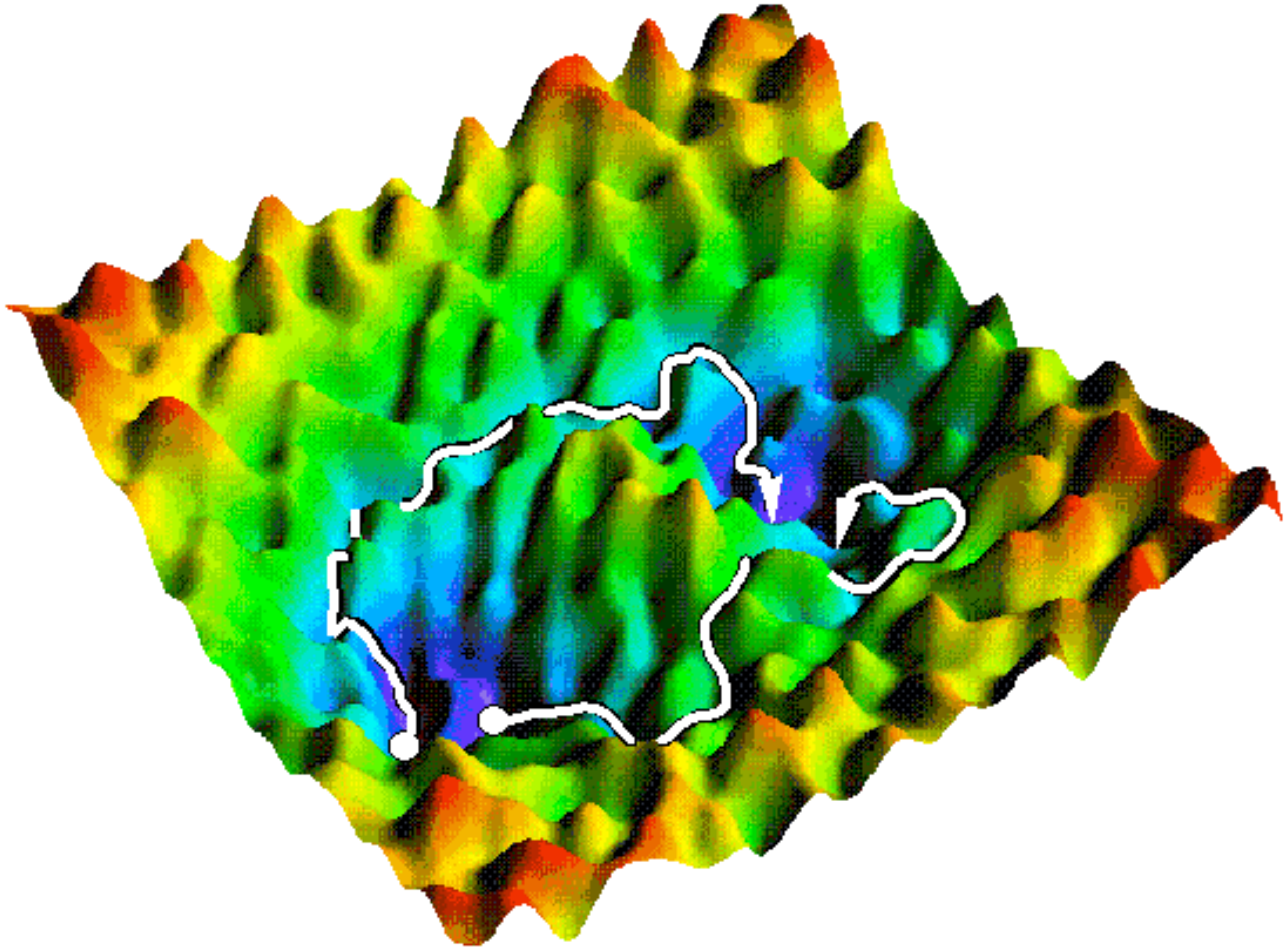
MM empirical potential

$$U_{MM}(r) = \boxed{\sum_{bonds} \frac{k_b}{2}(r - r_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta - \theta_0)^2 + \sum_{torsions,n} \frac{k_{\phi,n}}{2}[1 + cos(n\phi - \delta)]} +$$

$$+\boxed{\sum_{i>j}^{N} \left(\frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^{6}} \right) + \sum_{i>j}^{N} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}}$$

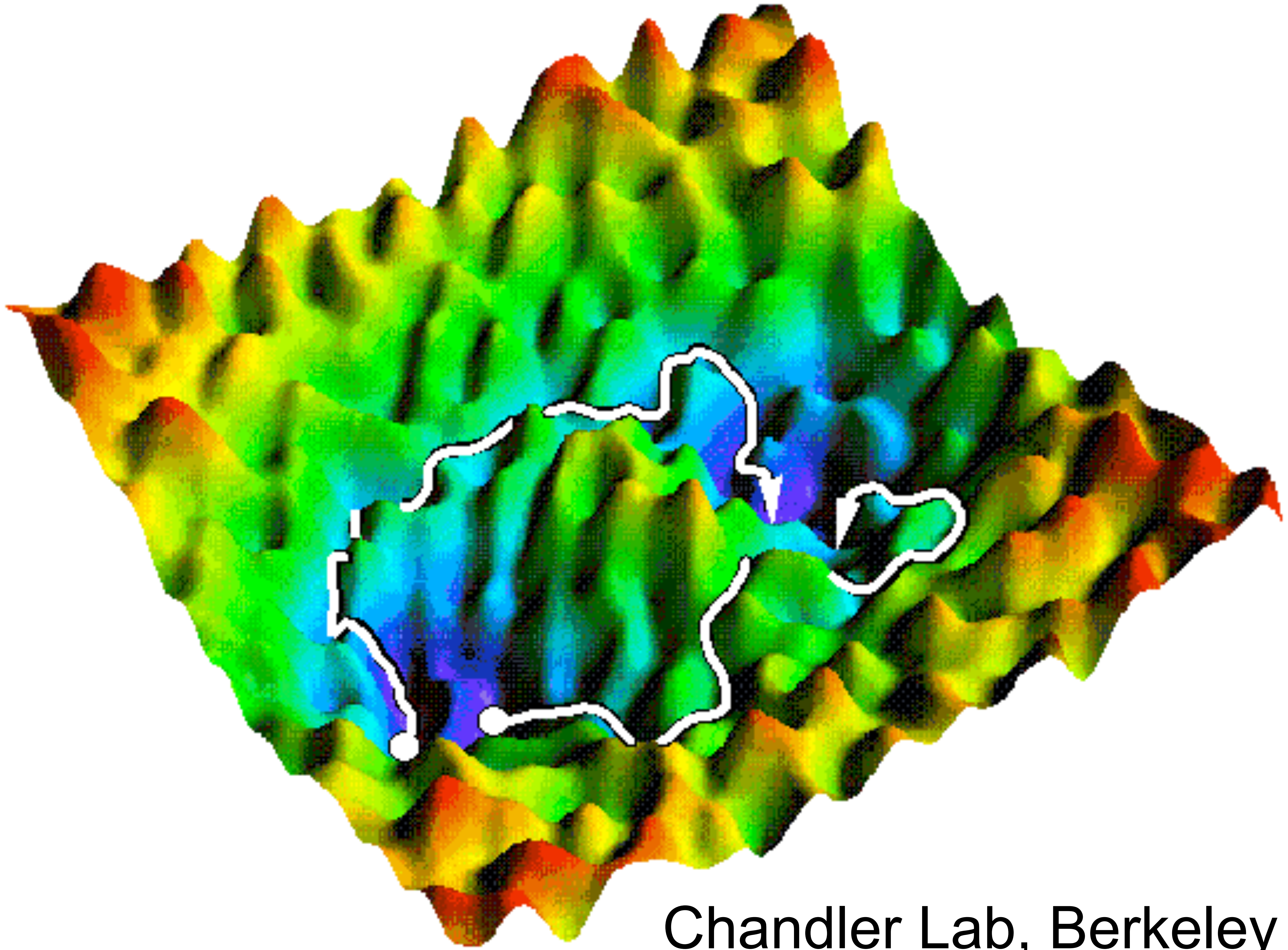
$$f(x_1, x_2, \dots, x_N) \qquad \mathbf{x} \in \Re^N$$

$$f(\mathbf{x}) : \Re^N \to \Re$$



Optimization

- it is a central problem in every science
- it can be final goal of modeling
- or starting point for more advanced calculations
- in **chemistry and biology**: determination of the low-energy conformation for a given energy function *U(r)*
- but also the search for maxima associated with chemical reactions, etc,
- in general used to describe the **energy landscape** of a system



Chandler Lab, Berkeley

Energy minimization

- global and local minima of a function *f(**x**)*
- stationary points: minima, saddle points, maxima
- landscape for a energy function *f(**x**)=U(**r**)*

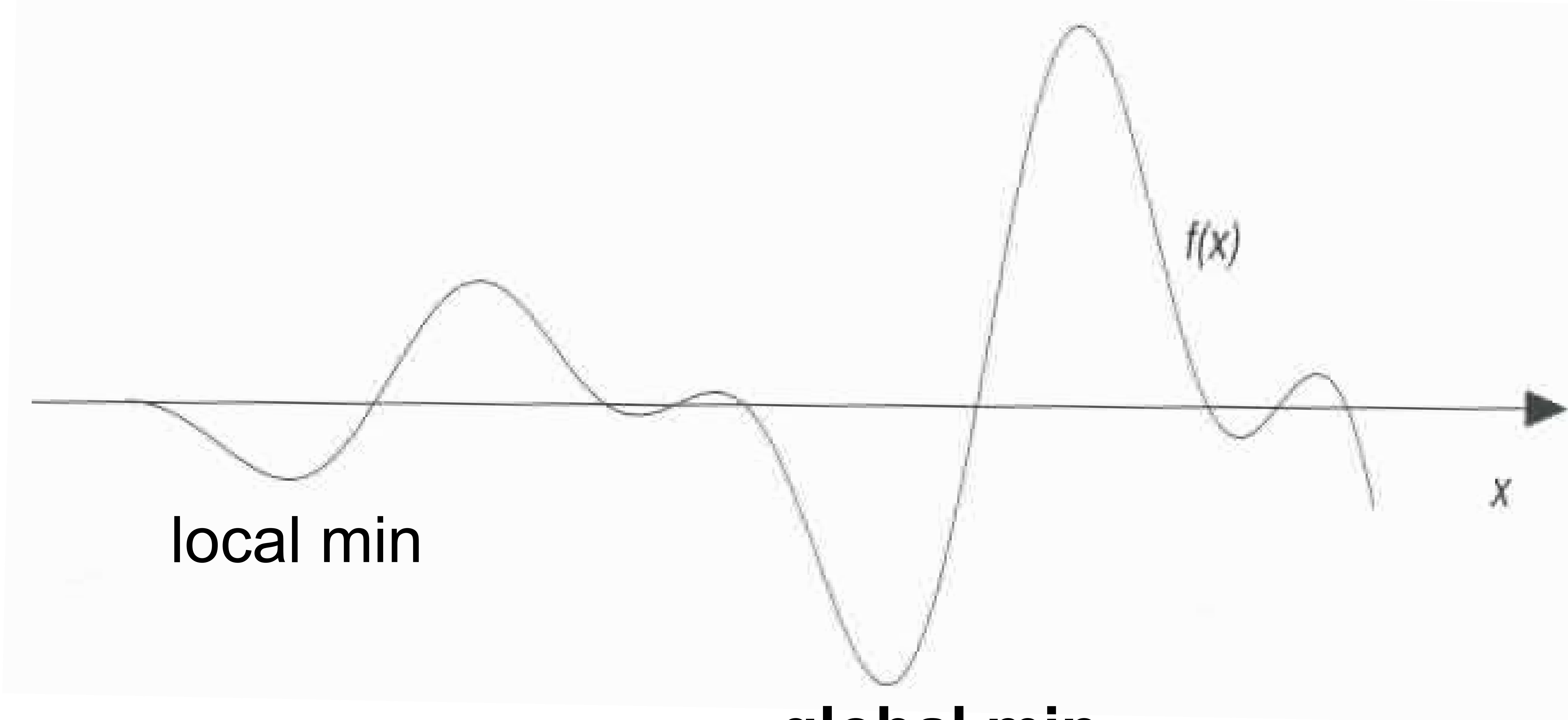
$$f(x_1, x_2, \dots, x_N) \qquad \mathbf{x} \in \Re^N$$

$$f(\mathbf{x}) : \Re^N \rightarrow \Re$$

$$min_{\mathbf{x}}\{f(\mathbf{x})\}$$

$$\frac{\partial f}{\partial x_i} = 0; \qquad \frac{\partial^2 f}{\partial x_i^2} > 0$$

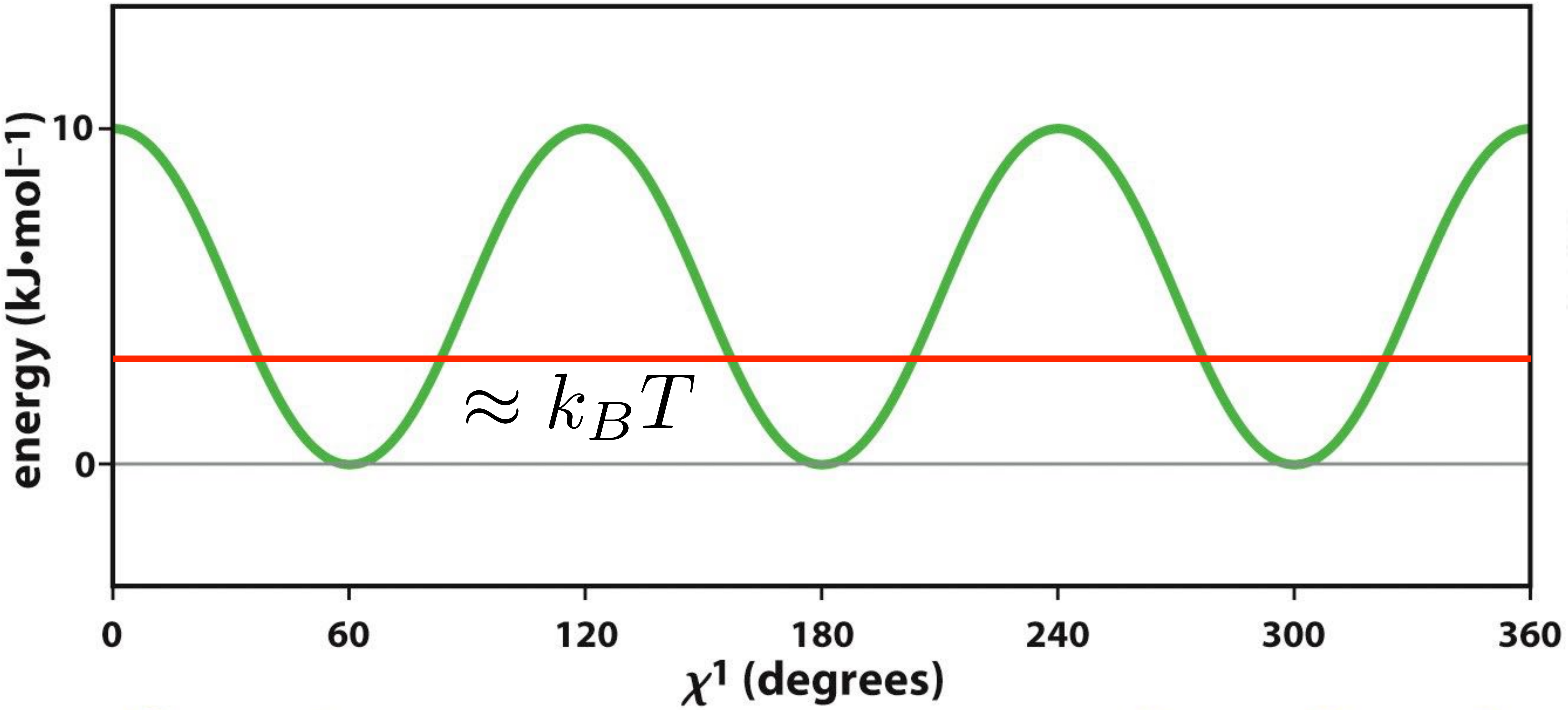
global max

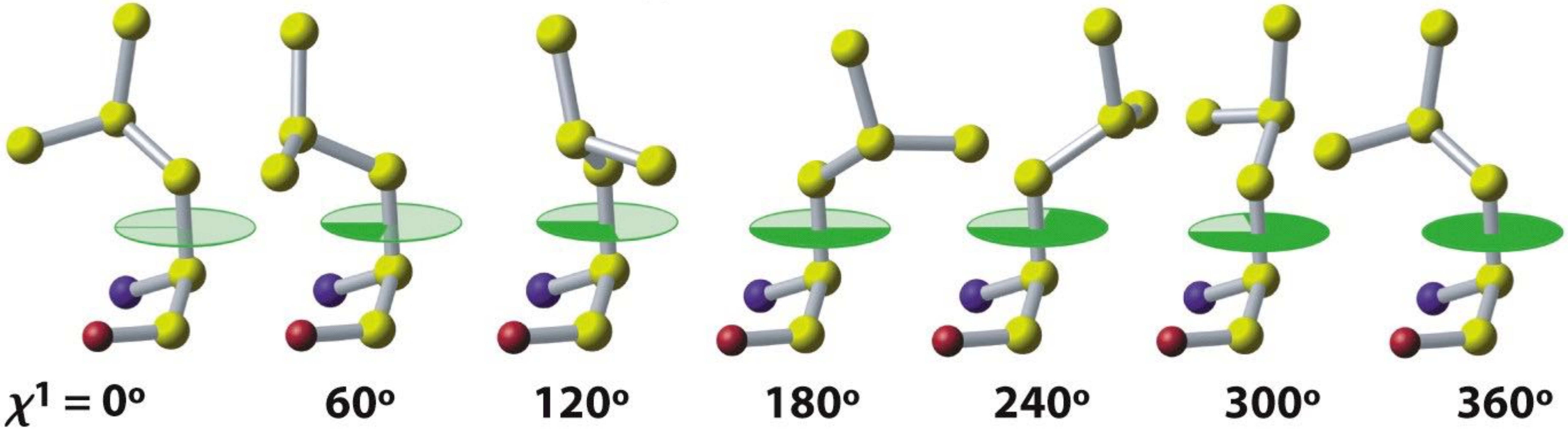


local min

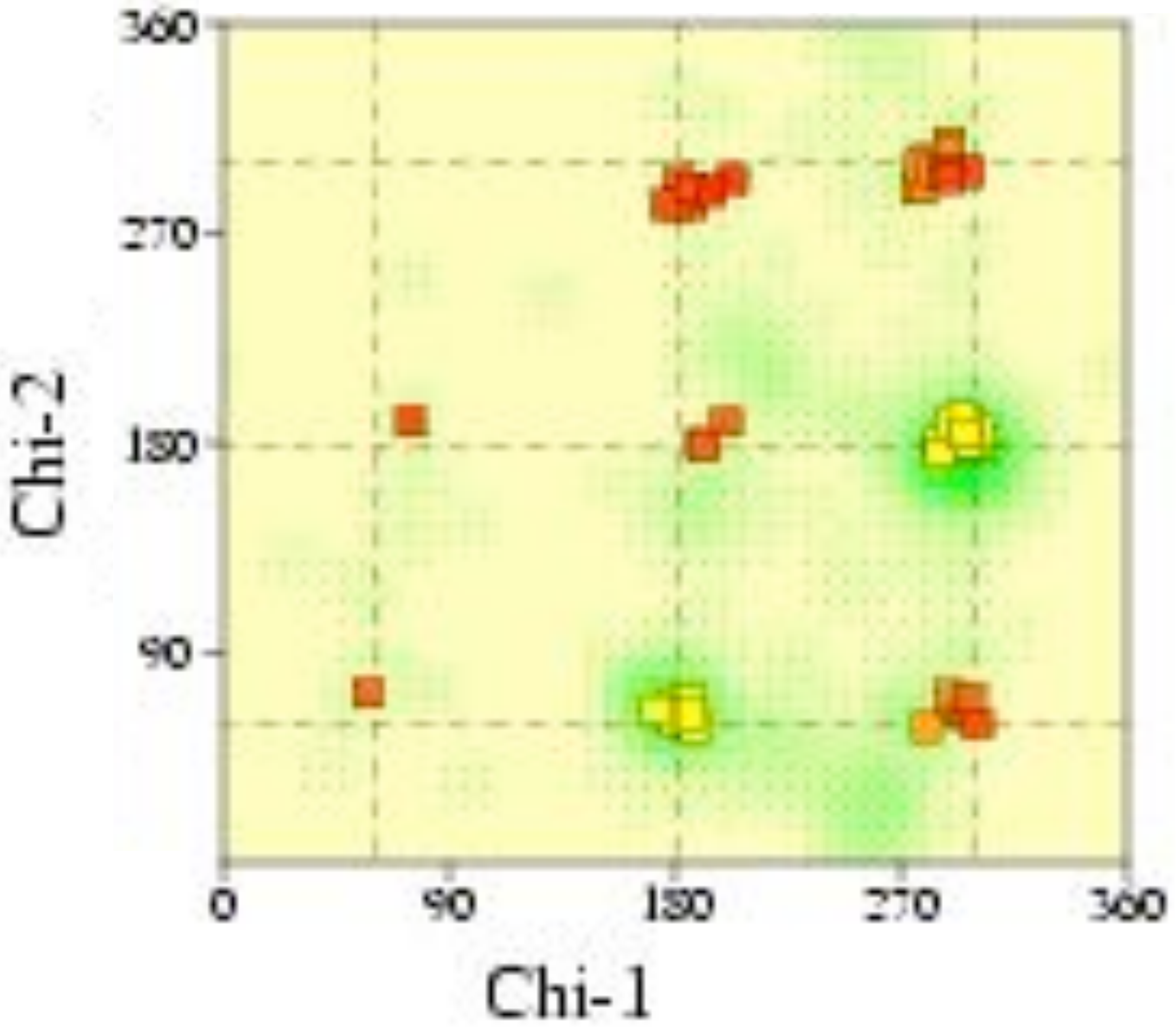
global min

$$U(\phi) = \frac{k_\phi}{2}(1 + cos3\phi)$$



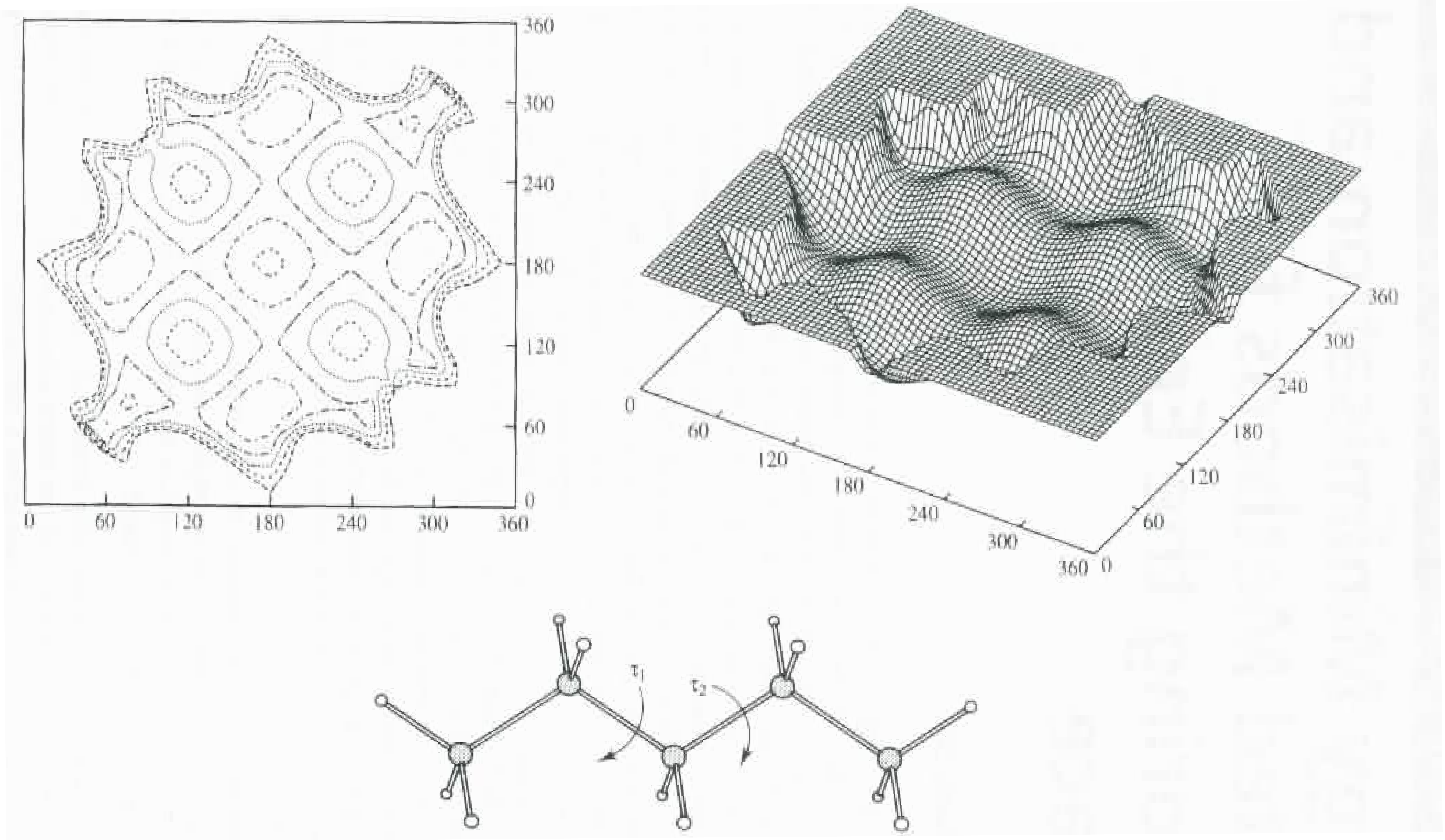


leucine (Leu, L)



Procheck: http://www.ebi.ac.uk/pdbsum/

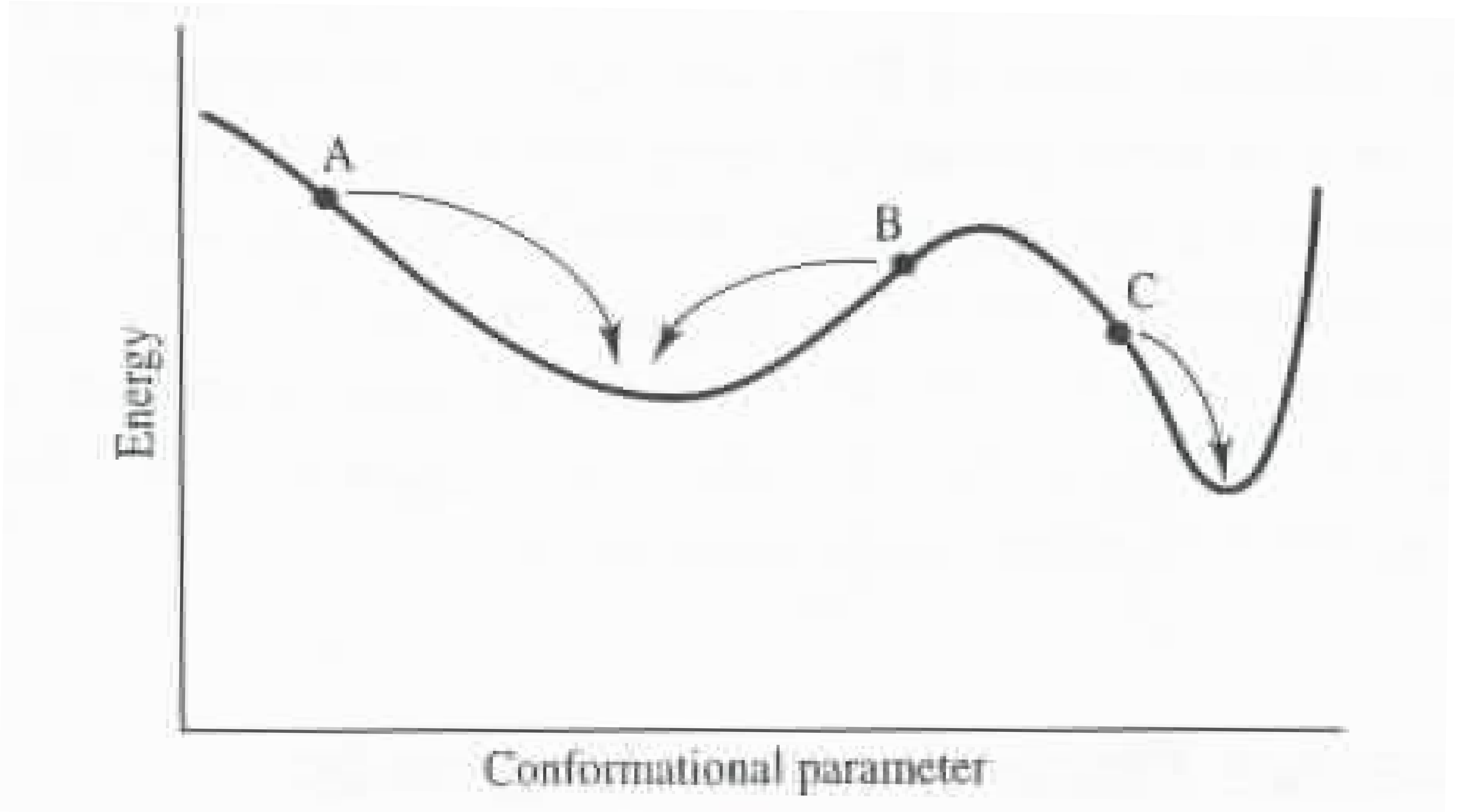
An example: alkanes



- 2 degrees of freedom for *U(**r**)*

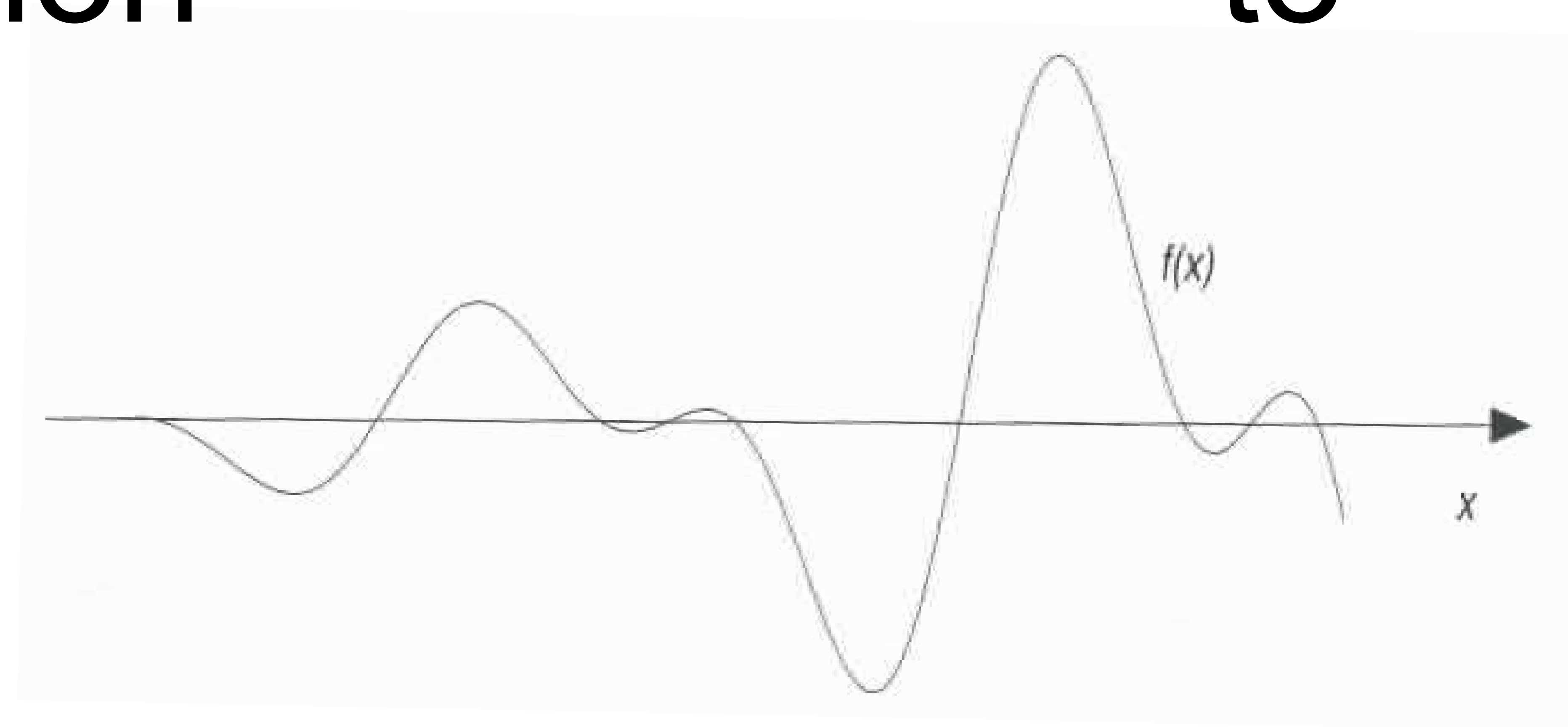
Minimization algorithms

- can make use of **derivatives** of *U(**r**)* or not
- **quick** answer, less time, less memory
- choice of method is problem-dependent
- most methods go **downhill**, multi initial starting points
- combination of experimental inputs and models for generating more initial states
- **no method** can surely locate global minimum from an arbitrary starting position



Derivative minimization methods

- direction of the **gradient** gives direction to search for the local/global minimum
- magnitude of the gradient gives the steepness of the local slope



- 1st and 2nd order methods (also 0th order methods)
- Taylor expansion of real *U(**x**)* introduces approximations

$$U(\mathbf{x}) = U(\mathbf{x_k}) + (\mathbf{x} - \mathbf{x_k})U'(\mathbf{x_k}) + (\mathbf{x} - \mathbf{x_k})^T \cdot U\text{”}(\mathbf{x_k}) \cdot (\mathbf{x} - \mathbf{x_k})/2 + ...$$

$$U'(\mathbf{x_k}) = \mathbf{g_k} \qquad g_i(\mathbf{x}) = \partial f(\mathbf{x})/\partial x_i$$

- **gradient**

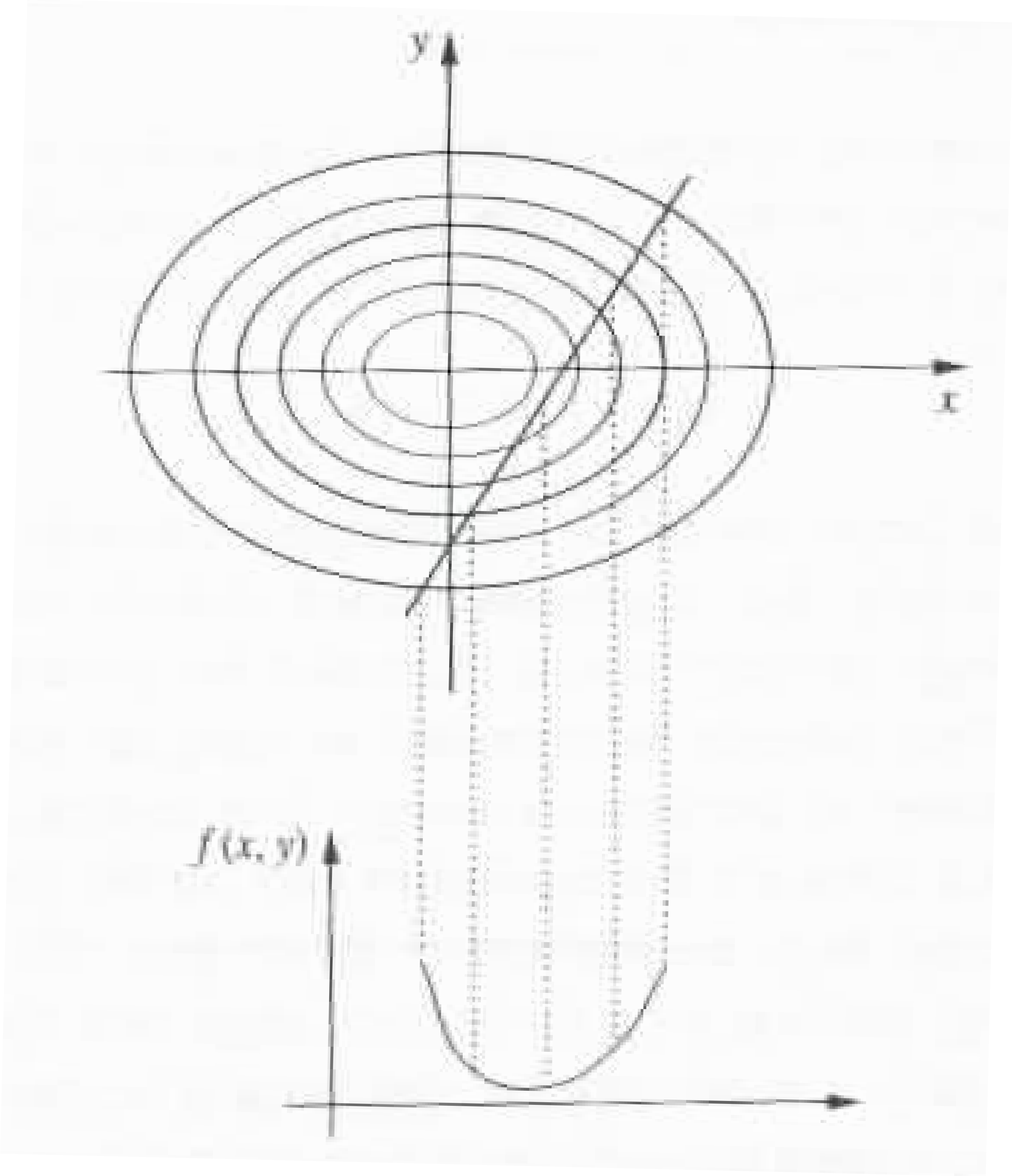
$$U''(\mathbf{x_k}) = H_{i,j}(\mathbf{x}) = \partial^2 f(\mathbf{x})/\partial x_i \partial x_j$$

- **hessian** or **force constant matrix**

First-order methods

- **Steepest descend (SD)**: move in the direction parallel to the net force (downhill), i.e.

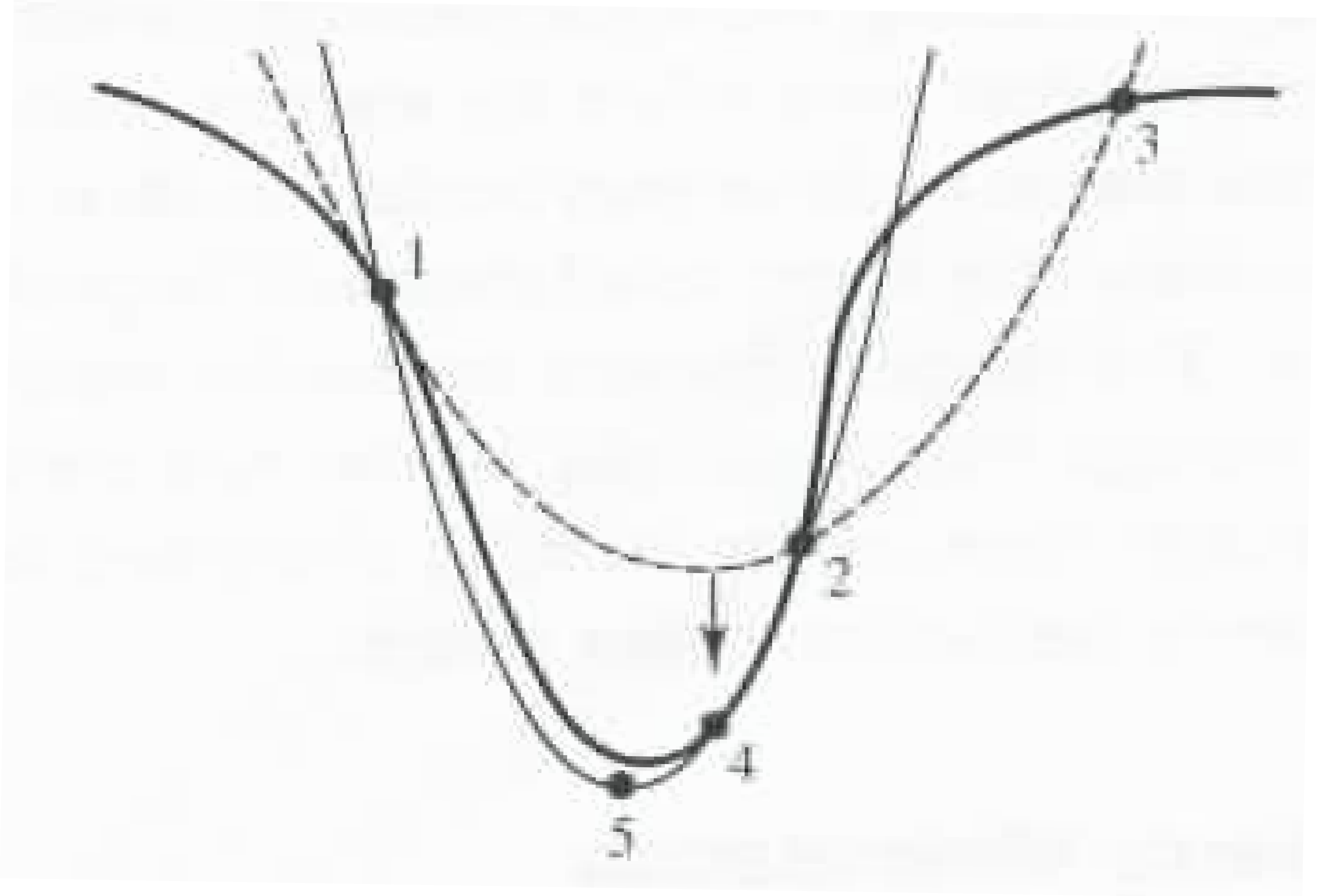
$$\mathbf{s_k} = -\mathbf{g_k}/|\mathbf{g_k}|$$



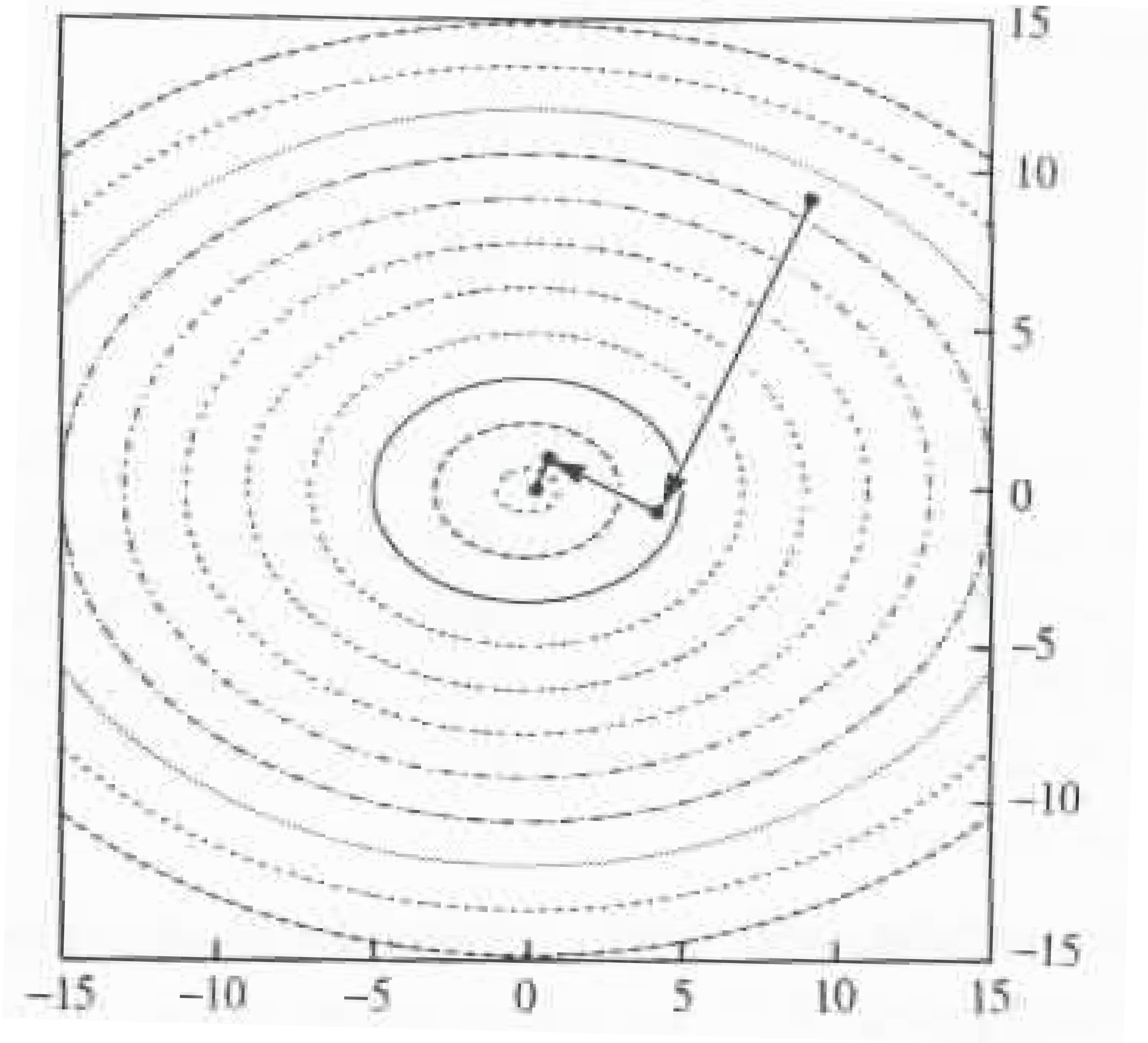
- how long should be the **step** along the gradient?

First-order methods

- 1. **line search**: bracket the minimum; gradient at the minimum will be orthogonal to the previous direction



$f(x,y) = x^2 + 2y^2; f_0 = f(9,9)$



$$\mathbf{g_k} \cdot \mathbf{g_{k-1}} = 0$$

- 2. **arbitrary step**: $\mathbf{x_{k+1}} = \mathbf{x_k} + \lambda_k \mathbf{s_k}$ consistently increased or reduced to minimize energy

- **SD** is good to relieve high-energy features, very robust far from minima, it has problems when close to them

Convergence criteria

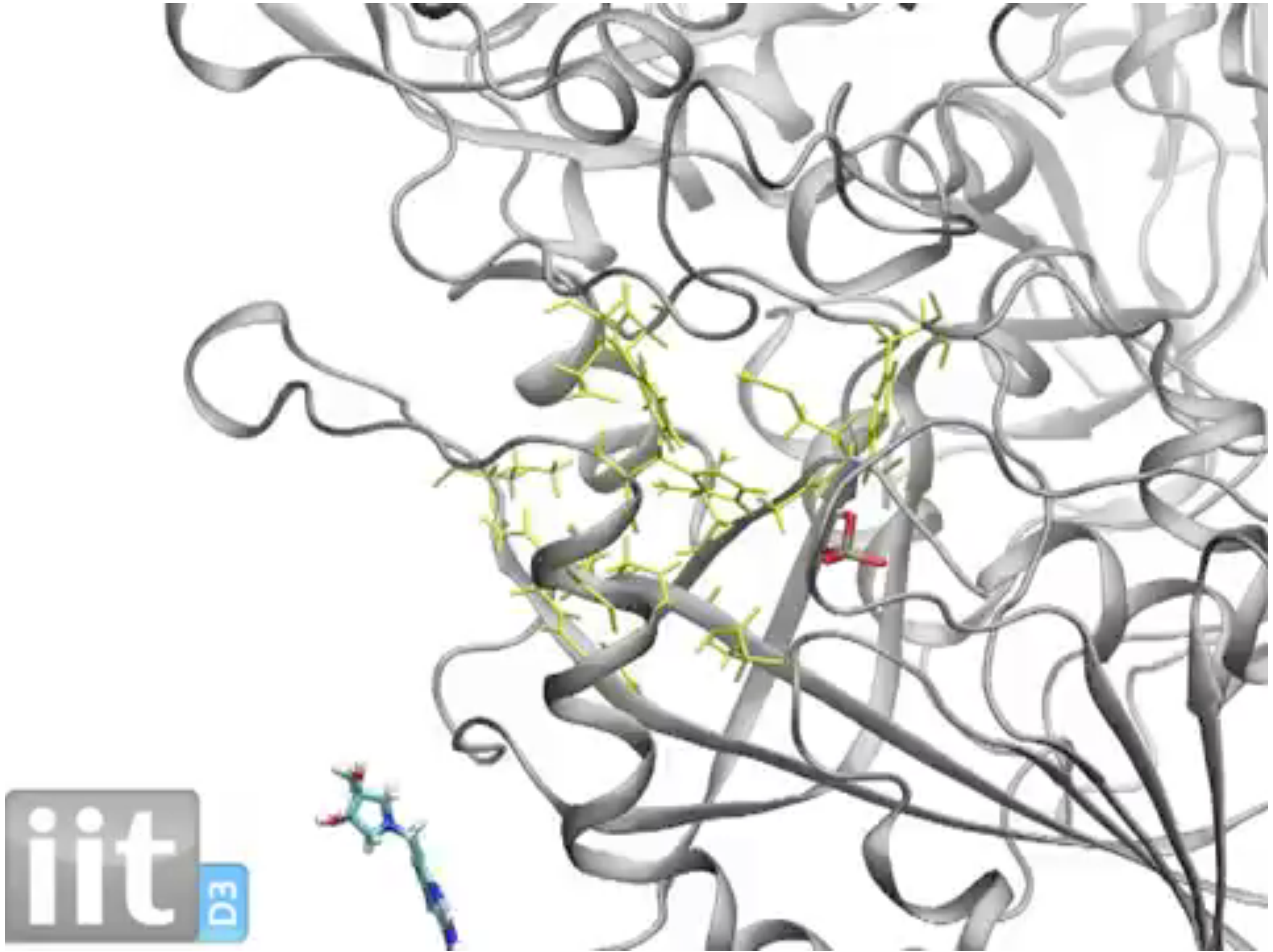
- **true** minimum is difficult to reach for any method
- consider machine precision: $1 + \epsilon_m = 1$
  (double precision ~10^{-15}, single precision 10^{-7})
- need for a convergence threshold to stop search
- monitor the energy drop and decide a threshold for energy between successive steps, or monitor change in coordinates, or the maximum value of the gradient in every dimension
- depend on the step following minimization, it can be **more or less stringent**

Caveats

- use different starting points; perturb your structure or use different models or experimental structures
- use different methods, and combination of methods
- use different force fields (e.g. for small molecules)
- check hessian eigenvalues (close to minimum all are positive, apart 6 zero terms)
- check artefacts from non-bonded cutoff methods
- use Monte-Carlo or heuristic search alternatively

Molecular Dynamics

- the motion of the particles is **realistic**, MD is able to get information about the mechanistic aspects of transformations undergone by the system (e.g., the mechanism of a chemical binding or the folding kinetic of a polymer).



Statistical mechanics in a nutshell

- relates **microscopic** to **macroscopic** observables
- gives a probability to find a given microstate with energy *E_i*

$$p(E_i) = \frac{1}{Z} e^{-E_i/k_B T}$$

$$Z = \sum_{i=1}^{N} e^{-E_i/k_B T}$$

- *$p(E_i)$* follows the Boltzmann distribution
- *Z* is called partition function (normalization)
- key thermodynamic quantities can be computed

$$\langle E \rangle = \sum_{i=1}^{N} E_i p(E_i)$$

Statistical mechanics

- we can express thermodynamic function in term of Z

$$\langle E\rangle = \frac{1}{Z}\sum_{i=1}^{N} E_i e^{-E_i/k_B T} = -\frac{1}{Z}\frac{\partial}{\partial\beta}Z = -\frac{\partial}{\partial\beta}lnZ$$

- or Gibbs free energy: $G = -k_B T lnZ$ $\qquad \beta = \frac{1}{k_B T}$
- derivation from second law of thermodynamics: dS>0
- maximization of Shannon entropy with the physical constraint, average E is constant by effect of thermal bath

$$S = -\sum_i p_i lnp_i - \gamma\left[\sum_i p_i - 1\right] - \beta\left[\sum_i p_i E_i - \langle E\rangle\right]$$

Molecular Dynamics

- the motion of the particles is **realistic**, MD is able to get information about the mechanistic aspects of transformations undergone by the system (e.g., the mechanism of a chemical reaction or the folding kinetic of a polymer).

- MD trajectories can be directly used to obtained **thermodynamically averaged quantities** (**ergodic theorem**: trajectory followed by a dynamical system explores the phase space according to its statistical probability):

$$\langle \mathcal{O} \rangle = \frac{1}{Z(T)} \int \mathcal{O}(\{p\},\{q\}) e^{-\beta \mathcal{H}(\{p\},\{q\})} d\Gamma = lim_{\mathcal{T} \to \infty} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} \mathcal{O}(s(t)) dt$$

Newton’s laws of motion

- 1. a body rests or moves at constant velocity unless a force acts upon it
- 2. force equals the rate of change of momentum (***F***=*m**a***)
- 3. to every action there is an equal and opposite reaction
- thus the **trajectory** of a particle is obtained by solving the differential equations derived from the Newton’s law (**equations of motion**):

$$\frac{d^2 x_i}{dt^2} = \frac{F(x_i)}{m_i} = -\frac{1}{m_i}\frac{dU(x_i)}{dx_i}$$

Integrating the equations of motion

- using realistic potentials the force on each particle x_i ($i=1,...,N$) changes whenever it moves (motion is coupled to all particles in the systems)

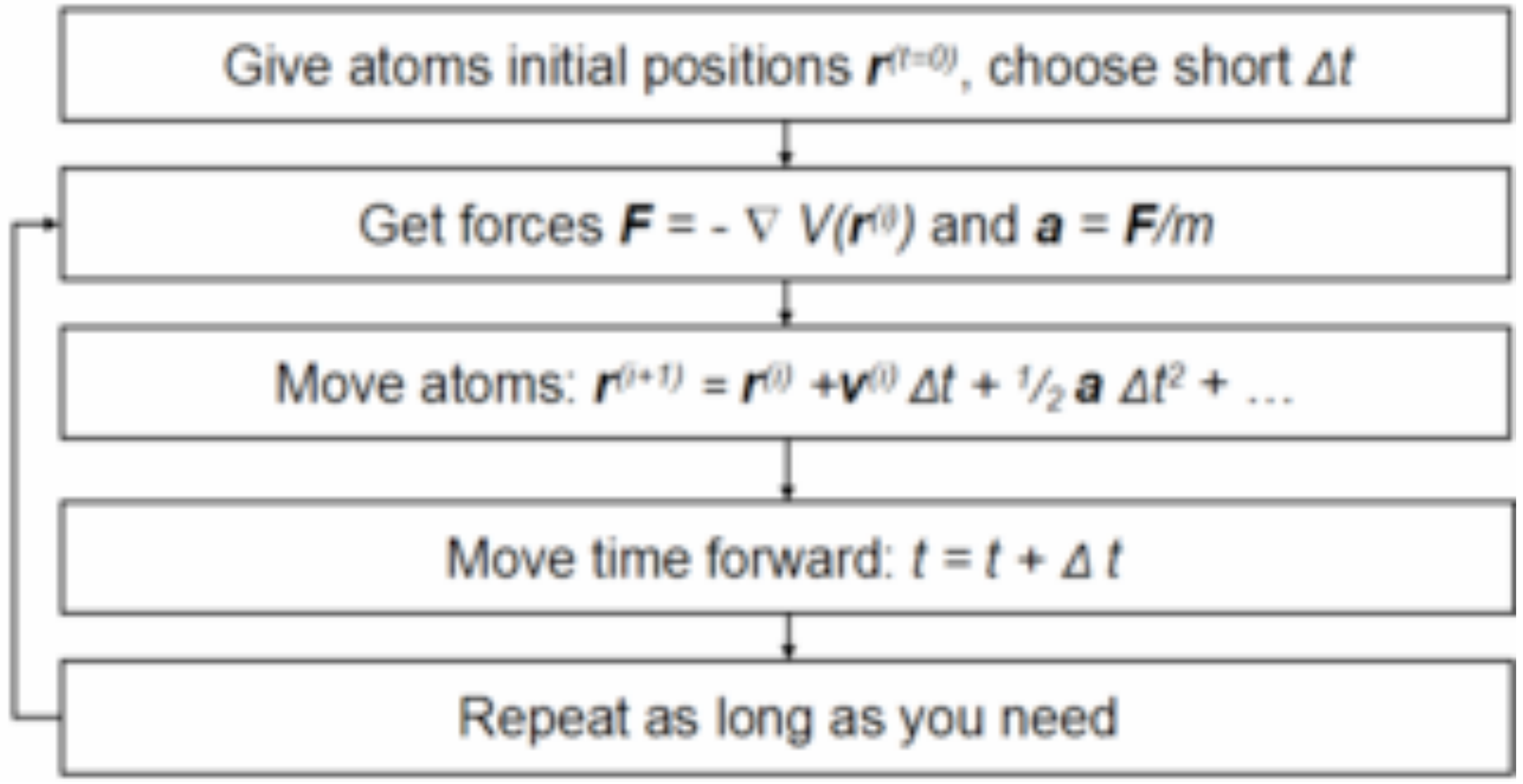
- need for **finite difference methods** to solve numerically the equations of motion

- the integration is broken down into many small steps, each of them separated by a fixed time, δt (**timestep**)

- **flow diagram for MD**:

- force calculation is the most cpu-demanding step

- various integrators to propagate atomic positions



Give atoms initial positions $\mathbf{r}^{(t=0)}$, choose short Δt

Get forces $\mathbf{F} = - \nabla V(\mathbf{r}^{(i)})$ and $\mathbf{a} = \mathbf{F}/m$

Move atoms: $\mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} + \mathbf{v}^{(i)} \Delta t + \frac{1}{2} \mathbf{a} \Delta t^2 + \ldots$

Move time forward: $t = t + \Delta t$

Repeat as long as you need

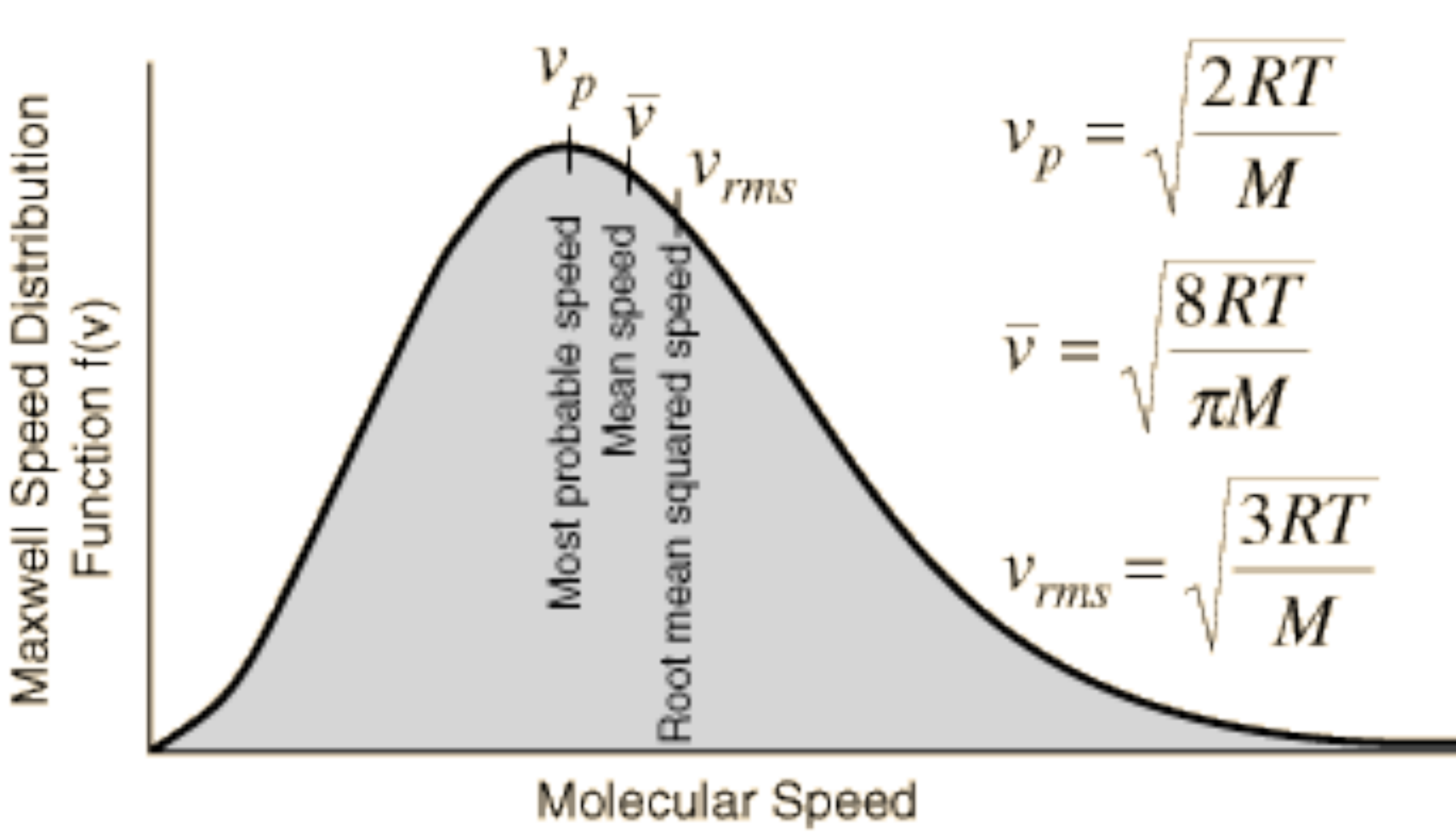
System initialization

- **positions** are derived from an experimental source (X-ray, NMR, etc.), or from a homology-based model that has been previously prepared and minimized

- **velocities** are assigned using a Boltzmann distribution:

$$p(v_x) = \sqrt{\frac{m}{2\pi k_B T}} exp\left(\frac{-mv_x^2}{2k_B T}\right)$$

$$p(v) = \sqrt{\frac{2}{\pi}\left(\frac{m}{k_B T}\right)^3} v^2 exp\left(\frac{-mv^2}{2k_B T}\right)$$



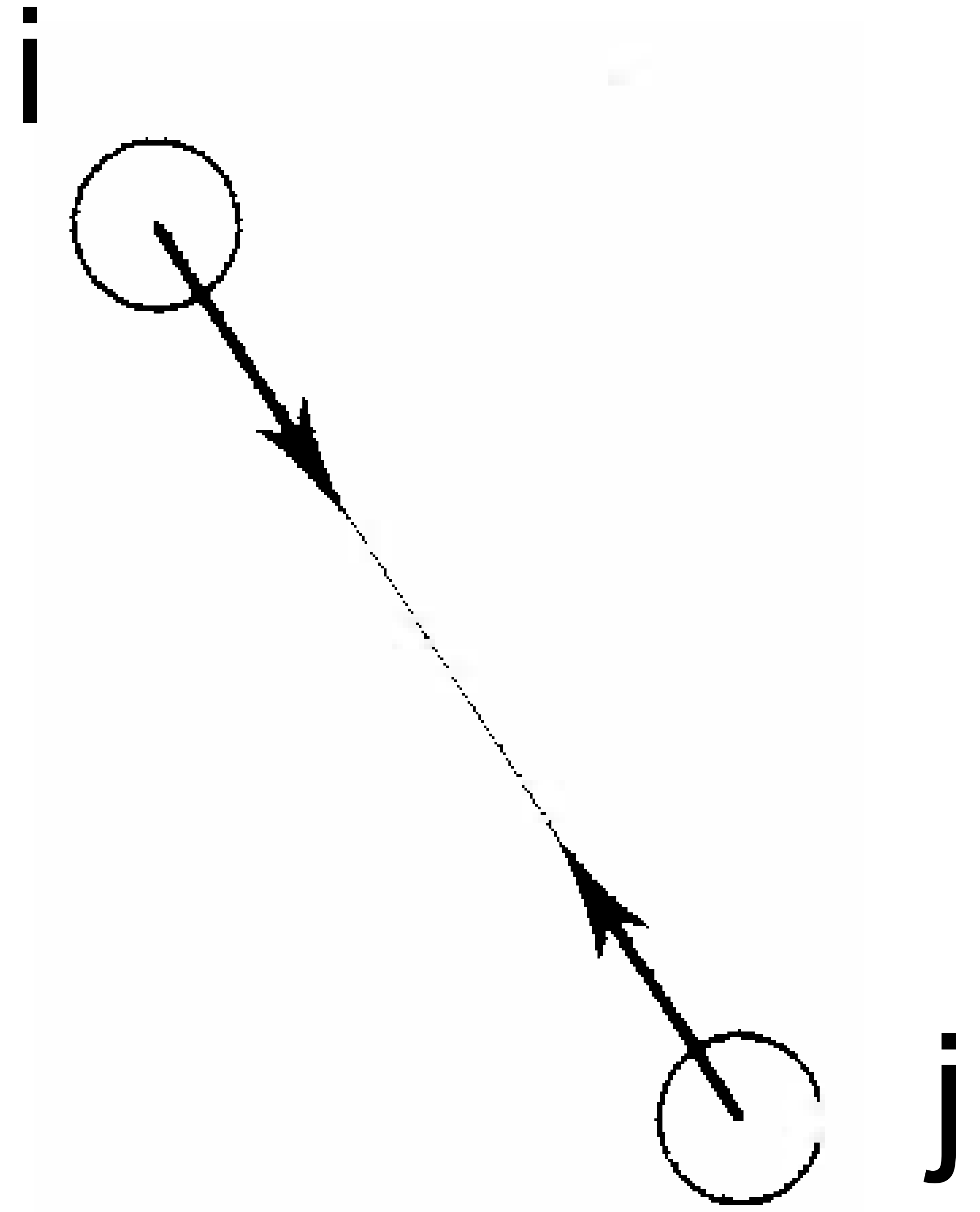
- $\langle v^2 \rangle = 3k_B T/m$: equipartition theorem (*k_BT/2* per DoF)

- from the velocities you have a way to measure the **temperature T** of your system

Force calculations

- from the potential *U(r)*, you can calculate **forces** and **acceleration** on the N atoms of your system:

$$\mathbf{F_i} = - \frac{\partial U(\mathbf{r_1}, ..., \mathbf{r_N})}{\partial \mathbf{r_i}}$$



- for instance, for the LJ potential part:

$$\mathbf{F_{ij}} = \frac{\mathbf{r_{ij}}}{|\mathbf{r_{ij}}|} \left[2 \left(\frac{\sigma}{r_{ij}} \right)^{13} - \left(\frac{\sigma}{r_{ij}} \right)^{7} \right]$$

- once you have the force contribution for each atom you can calculate its **trajectory** till the next timestep

Integration methods

- all use positions, velocities and accelerations of particles, and approximate them as a Taylor series:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + 1/2\delta t^2\mathbf{a}(t) + 1/6\delta t^3\mathbf{b}(t) + \delta t^4\mathbf{c}(t) + ...$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \delta t\mathbf{a}(t) + 1/2\delta t^2\mathbf{b}(t) + 1/6\delta t^3\mathbf{c}(t) + ...$$

$$\mathbf{a}(t+\delta t) = \mathbf{a}(t) + \delta t\mathbf{b}(t) + 1/2\delta t^2\mathbf{c}(t) + ...$$

- **Verlet algorithm** (1967) is the most widely used method for MD:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + 1/2\delta t^2\mathbf{a}(t) + ...$$

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \delta t\mathbf{v}(t) + 1/2\delta t^2\mathbf{a}(t) - ...$$

$$\boxed{\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2\mathbf{a}(t)}$$

Verlet algorithm

$$\mathbf{r}(t + \delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \delta t) + \delta t^2 \mathbf{a}(t)$$

- uses positions at time (t - δt) and accelerations, but no velocities, which can be derived from positions:

$$\mathbf{v}(t) = [\mathbf{r}(t + \delta t) - \mathbf{r}(t - \delta t)]/2\delta t$$

- easy implementation, memory needed is modest
- $\delta t^2 \mathbf{a}(t)$ is a small term, which can lead to loss of precision (i.e. no conservation of energy)
- velocity calculation is postponed
- it is not a self-starting algorithm

Velocity Verlet

- it give positions, velocities and accelerations at the same time and does not compromise precision

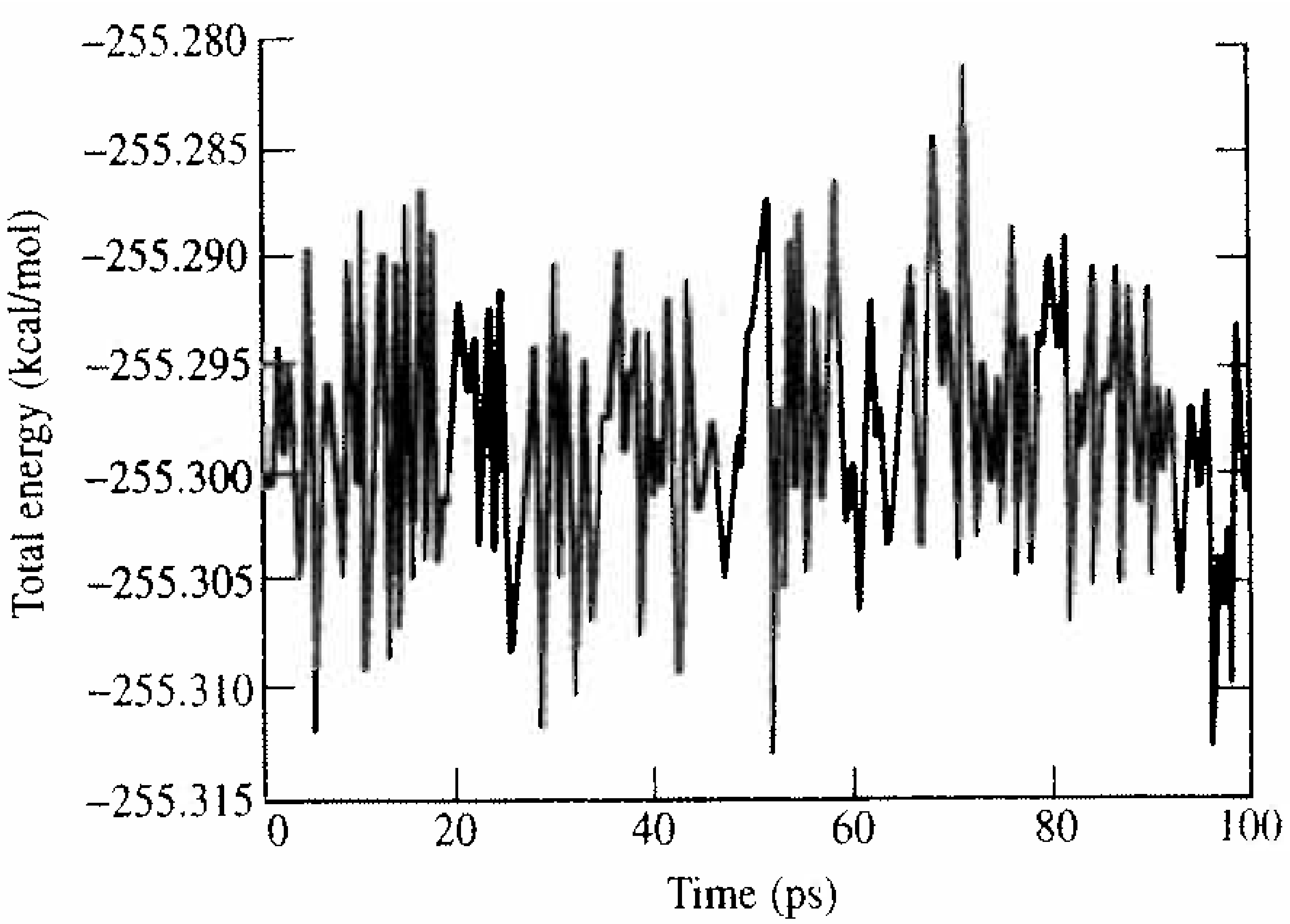
$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t \mathbf{v}(t) + 1/2 \delta t^2 \mathbf{a}(t)$$
$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + 1/2 \delta t [\mathbf{a}(t) + \mathbf{a}(t+\delta t)]$$

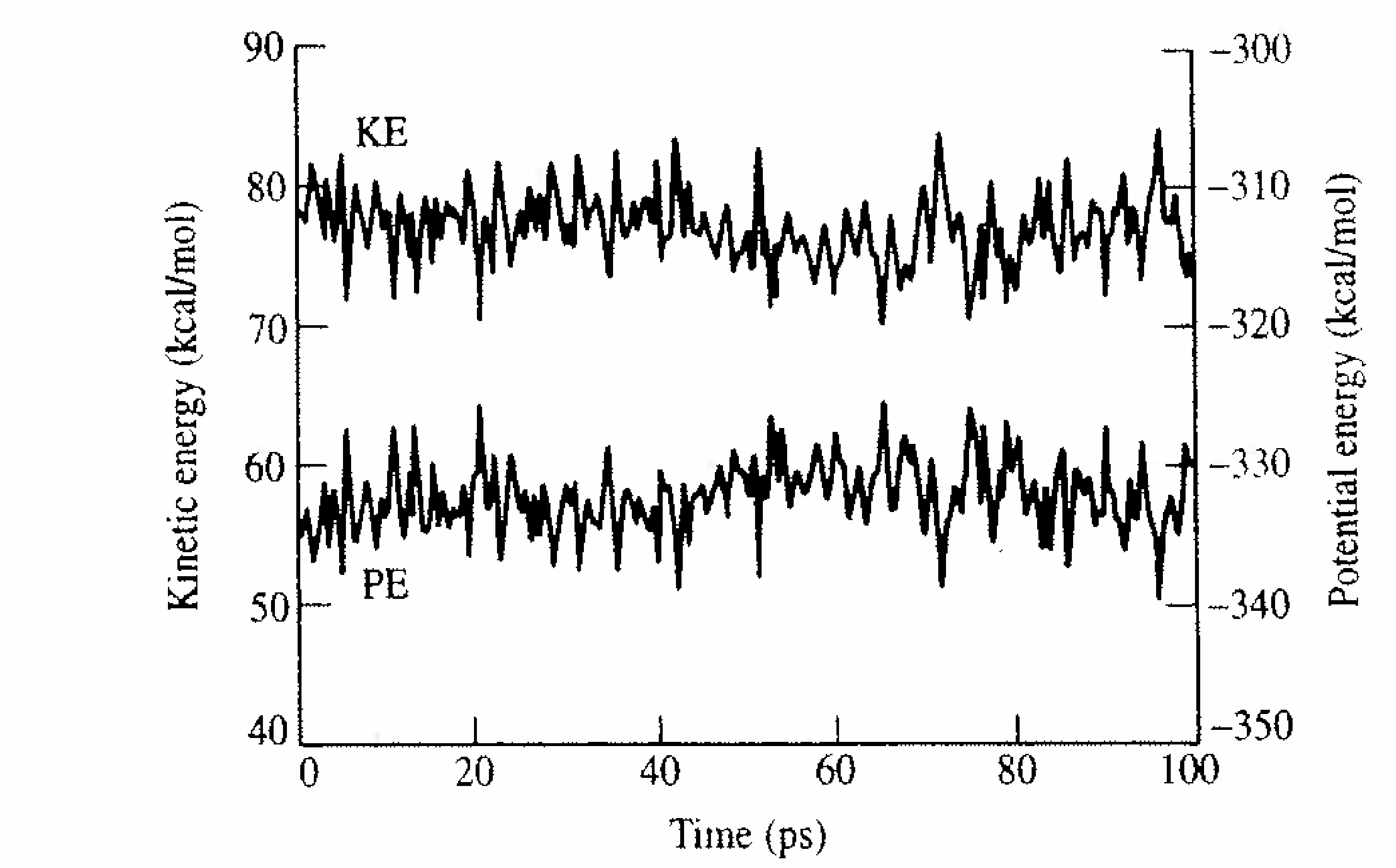
- 3-stage calculation: first positions at (t+δt), then forces at (t+δt) and finally velocities at (t+δt)

- **Leap-frog** is another common algorithm, where position and velocities are not synchronized though
- both are **time-reversible** and **symplectic** integrators

Choosing the integrator

- importance of energy conservation: **E~cost**; H=T+U in the **microcanonical** ensemble (NVE constant)



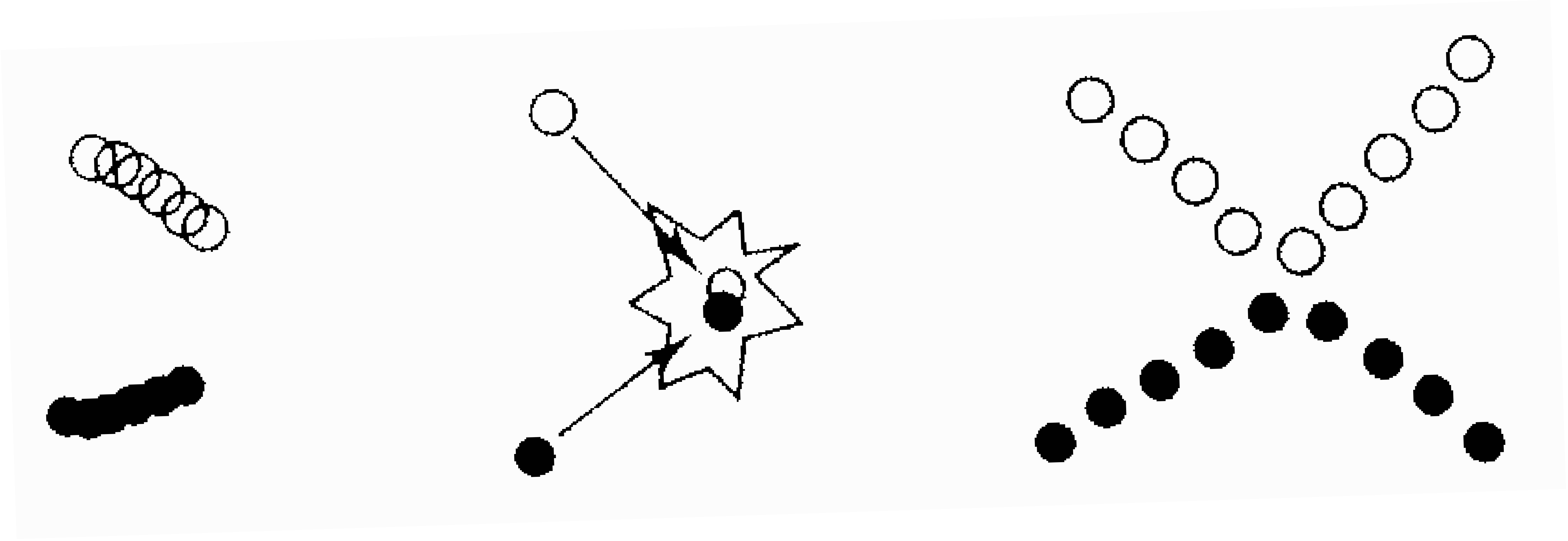


- as **timestep** increases, energy RMS fluctuations increase (tolerance: $\Delta E/E \sim 10^{-4}$)

- for a given timestep, the **drift** of energy in short or long trajectory can vary for different algorithms

Choosing the timestep

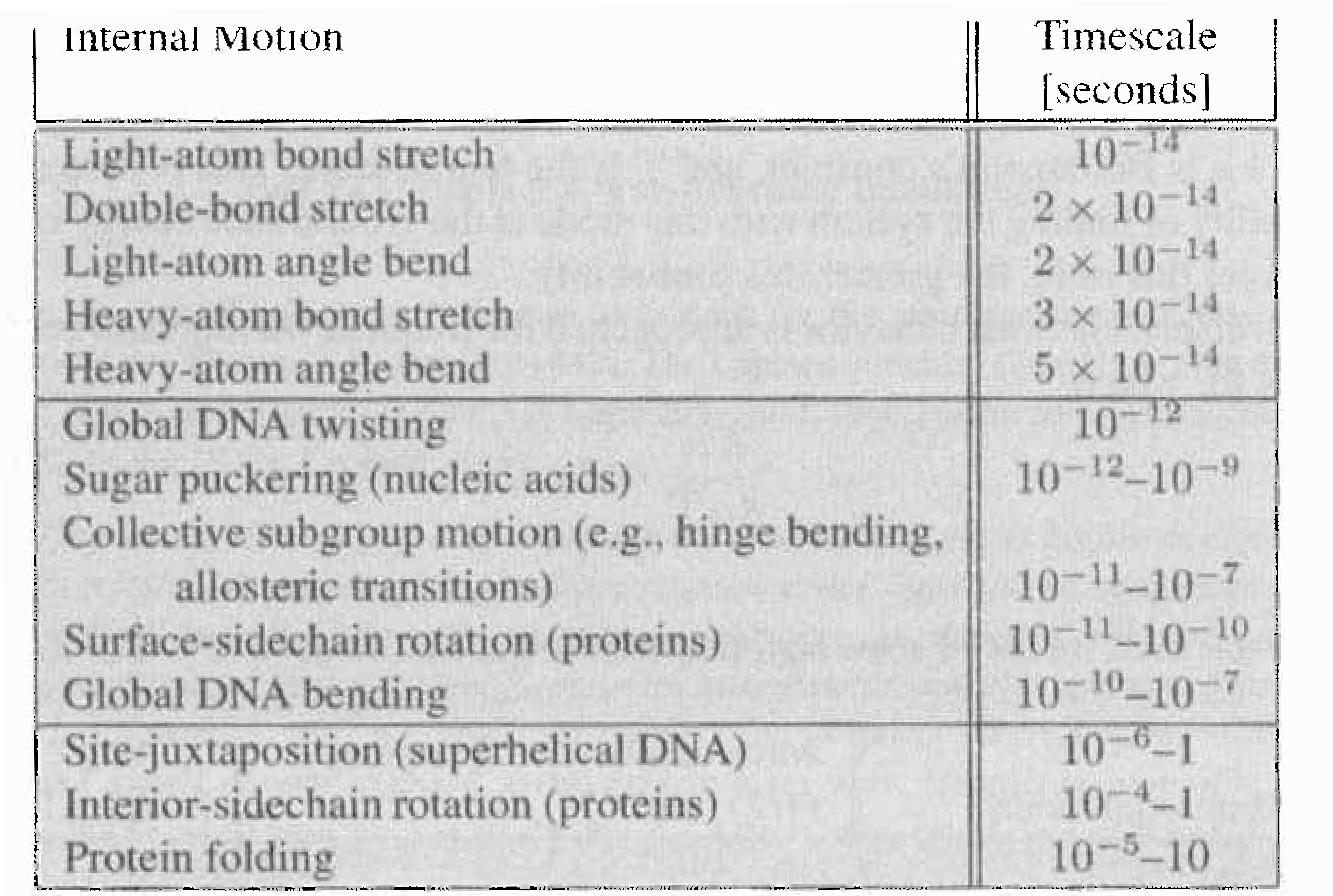
- timestep (δt) is crucial for MD: need for a **compromise**
- **too short:** the trajectory will cover only a limited region of the phase-space
- **too large:** integration of the equations of motion will produce instabilities and failure in energy conservation
- **rule of thumb**: 0.1*(shortest motion time in the system)



Choosing the timestep

- in practice for biomolecular systems ***δt*~1 fs** (shortest motion is bond fluctuations involving H atoms, for instance C-H bond: ~10 fs)

- **multiple time step** integration (RESPA), or **freezing** of fast fluctuations (all H-X or X-Y bonds with SHAKE, RATTLE, etc.) will permit a ***δt*~2 fs**



Internal Motion	Timescale [seconds]
Light-atom bond stretch	10^{-14}
Double-bond stretch	2×10^{-14}
Light-atom angle bend	2×10^{-14}
Heavy-atom bond stretch	3×10^{-14}
Heavy-atom angle bend	5×10^{-14}
Global DNA twisting	10^{-12}
Sugar puckering (nucleic acids)	10^{-12}–10^{-9}
Collective subgroup motion (e.g., hinge bending, allosteric transitions)	10^{-11}–10^{-7}
Surface-sidechain rotation (proteins)	10^{-11}–10^{-10}
Global DNA bending	10^{-10}–10^{-7}
Site-juxtaposition (superhelical DNA)	10^{-6}–1
Interior-sidechain rotation (proteins)	10^{-4}–1
Protein folding	10^{-5}–10

MD ensembles

- **microcanonical** (NVE), but thermodynamic variables **T** and **P** are more convenient, they are usually closer to the experimental setup

- in (NVE) from kinetic energy you can calculate T:

$$H = \sum_{i=1}^{\widetilde{N}} \frac{m_i v_i^2}{2} = \frac{3\widetilde{N} k_B T}{2} \qquad T = \frac{1}{2} \sum_{i=1}^{\widetilde{N}} \frac{2}{3} \frac{m_i v_i^2}{\widetilde{N} k_B}$$

- statistical ensembles connect microscopic to macroscopic quantities: **canonical** (NVT, Helmhotz free-energy); **isothermal-isobaric** (NPT, Gibbs free-energy)

- use of thermostats or barostats allows to control other quantities and to produce the appropriate ensemble

NVT: coupling thermostat

- **rescaling** of velocities: $\mathbf{v}_{n+1} = c_T \mathbf{v}_n;\ c_T = \sqrt{T_0/T}$

- more gently approach coupling to a **thermostat** of given temperature T, using a fictitious frictional coefficient (**Berendsen**)

$$m_i \dot{\mathbf{v}}_i(t) = -\nabla U(\mathbf{x}_i(t)) - \gamma_t m_i \mathbf{v}_i(t)$$

$$\gamma_t = \frac{1}{2\tau}\left(1 - \frac{T_0}{T}\right) \qquad c_T = \sqrt{1 - \frac{\delta t}{\tau}\left(1 - \frac{T_0}{T}\right)}$$

- the τ constant controls the strength of the coupling: when **large** (>1ps), c_T~1 (no scaling, microcanonical) when **small** (<0.01ps), the energy exchange between the system and the thermal bath is very significant (but this does not rigorously produce a **canonical** ensemble)

Canonical NVT ensemble

- **extended system** methods to produce rigorously thermodynamic ensemble

- additional degrees of freedom to the system H (e.g. **Nose´-Hoover**)

$$H^{NVT} = T + U + \frac{1}{2}(m_t \zeta_t^2) + \widetilde{N} k_B T_0 x_t$$

$$\begin{cases} m_i \dot{\mathbf{v}}_i(t) = -\nabla U(\mathbf{x}_i(t)) - \zeta_t m_i \mathbf{v}_i(t) \\ m_t \dot{\zeta}_i(t) = 2 m_i \mathbf{v}_i^2(t) - \widetilde{N} k_B T_0 \end{cases}$$

- x_t is the effective scaling parameter, ζ_t is the friction coefficient, m_t is a fictitious mass (control the rate of the thermalization process)

NPT ensemble

- more practical ensemble, closer to experimental setup
- controlling pressure, it is possible to equilibrated density of your system to target values (e.g. 1 g/cm^3 for water)
- scale the volume or couple a **pressure bath**:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p}(P_{bath} - P(t))$$

$$\lambda = 1 - \frac{\delta t}{\tau_p^{'}}(P(t) - P_{bath}) \ ; \quad \mathbf{r}_{i,n+1} = \lambda^{1/3}\mathbf{r}_{i,n}$$

- scaling can be applied isotropically or anisotropically
- extended methods to produce rigorous version of the **isothermal-isobaric NPT**

MD setup and production

- **check** and **prepare** your system (starting from experiments or predictions)

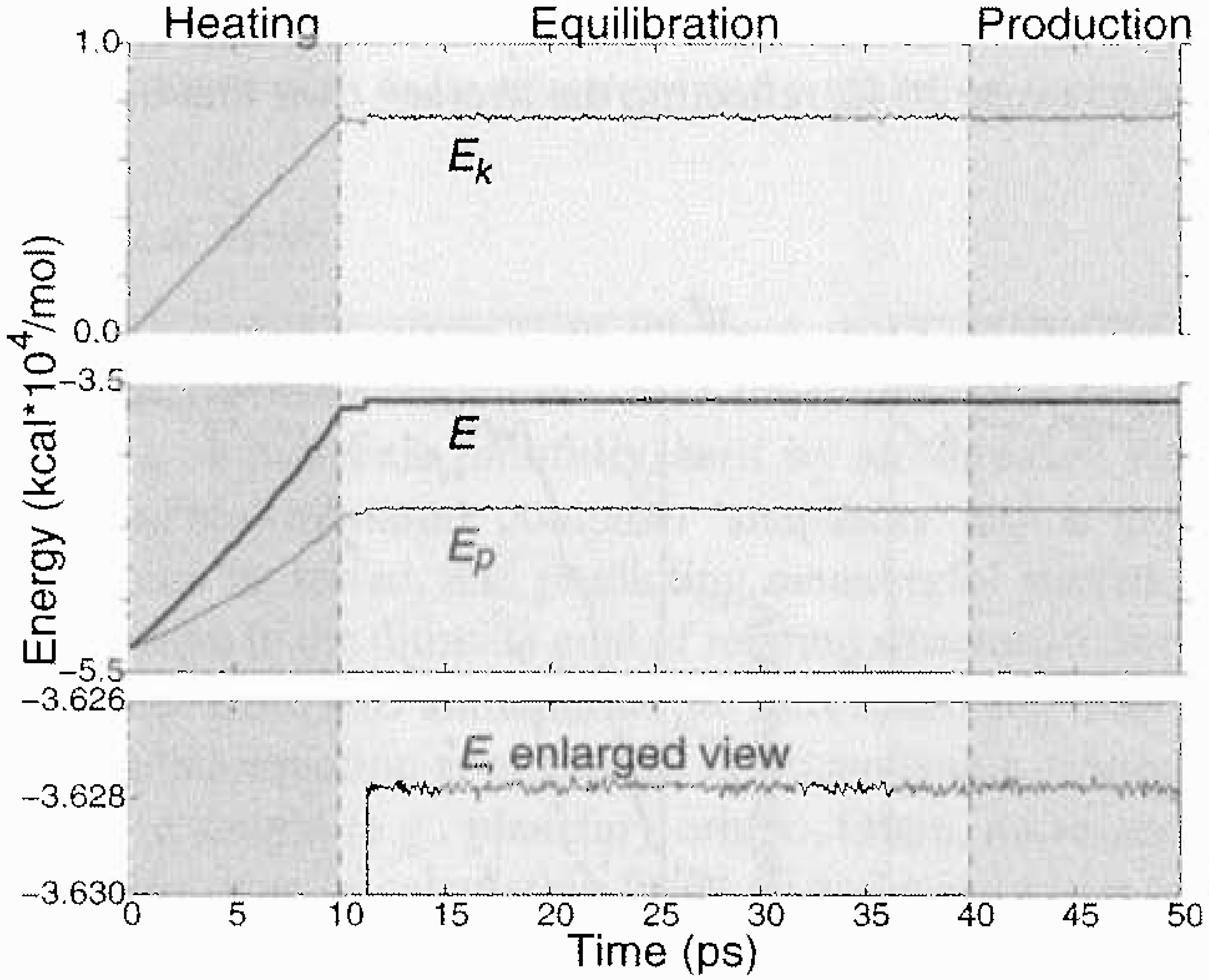
- define the simulation **cell**, **solvate**, add physiological concentration of **salt** (e.g. 150 mM of NaCl)

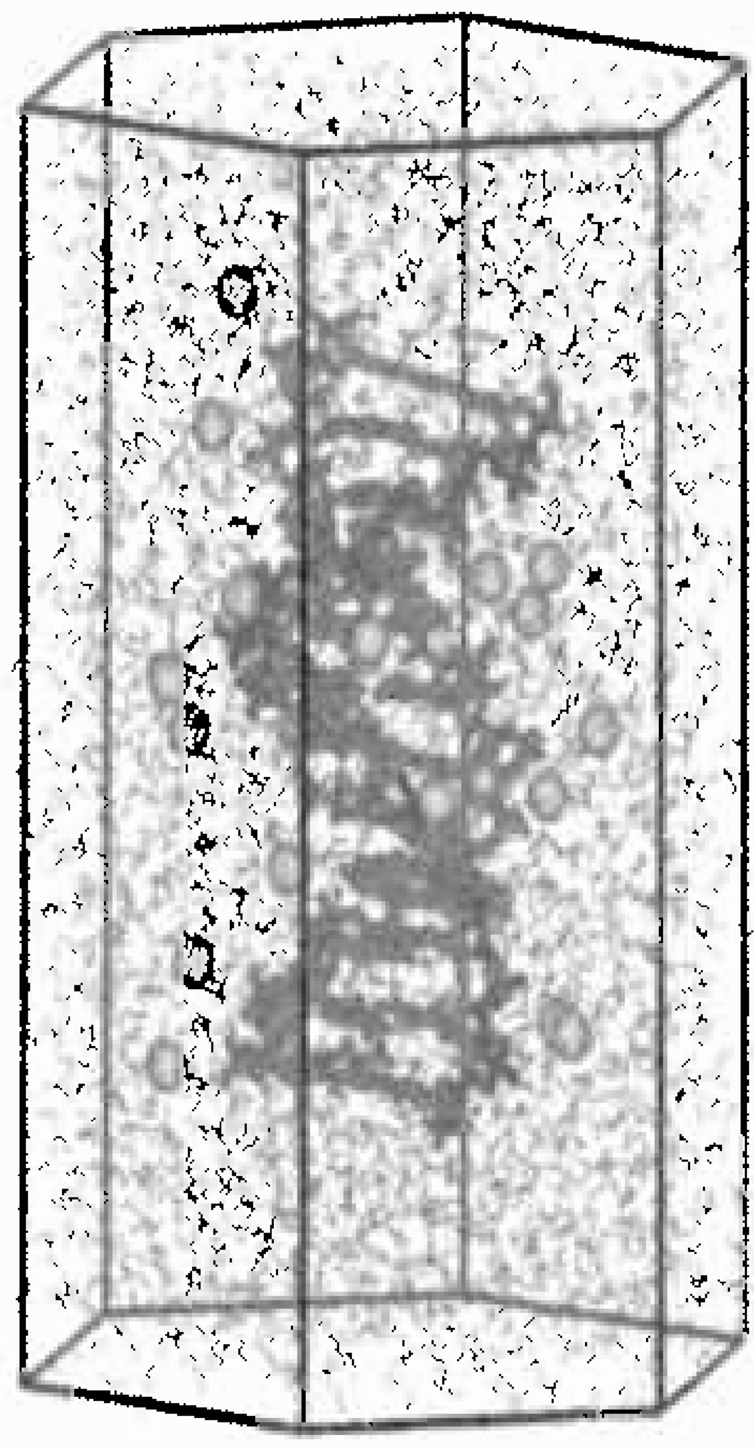
- **minimize** the energy to relax possible initial frustrations

- gradually **heat up** the system to desired T

- **equilibrate** first the solvent, light atom, then the side chains, finally the backbone of your protein

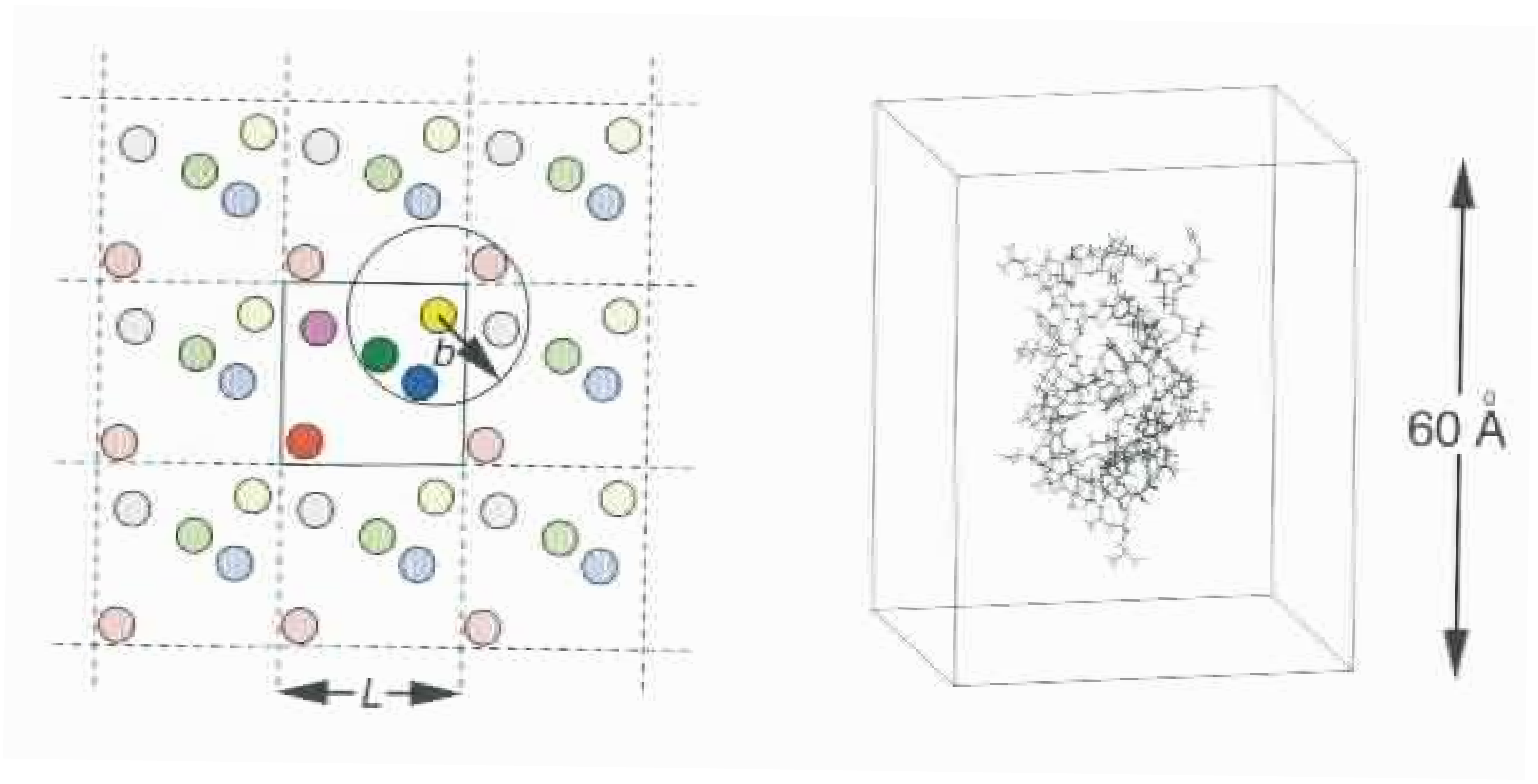
- complete equilibration and enter in **production** mode

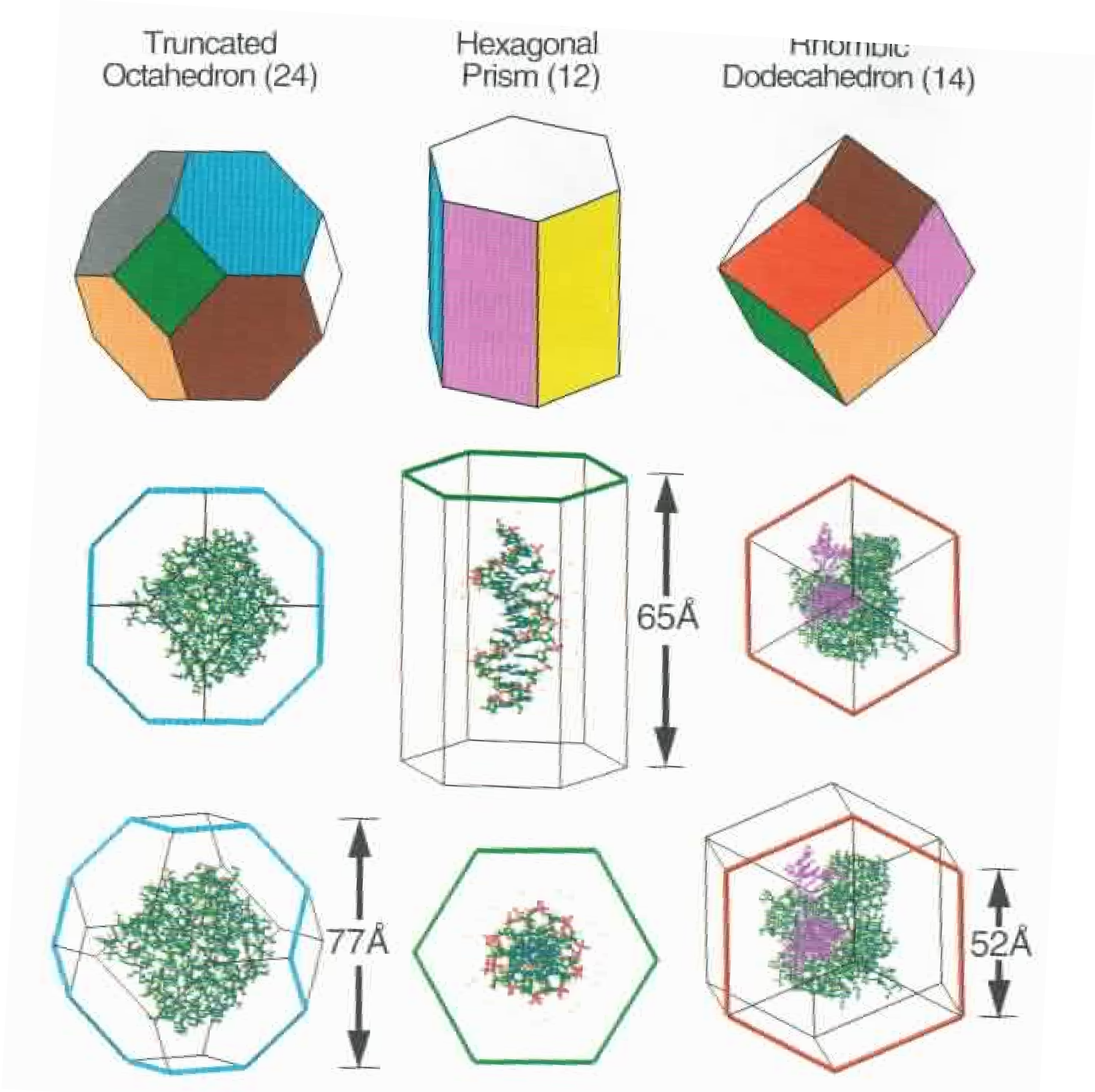




Ewald methods

- used for calculating electrostatic energy of systems in **periodic boundary conditions** (unit cell charge = 0)
- **minimum-image convention:** each atom interacts with the closest periodic image of the other N-1 atoms
- different unit cell lattice geometry
- use of **fast fourier transforms** to compute the electrostatic energy in the real and reciprocal lattice

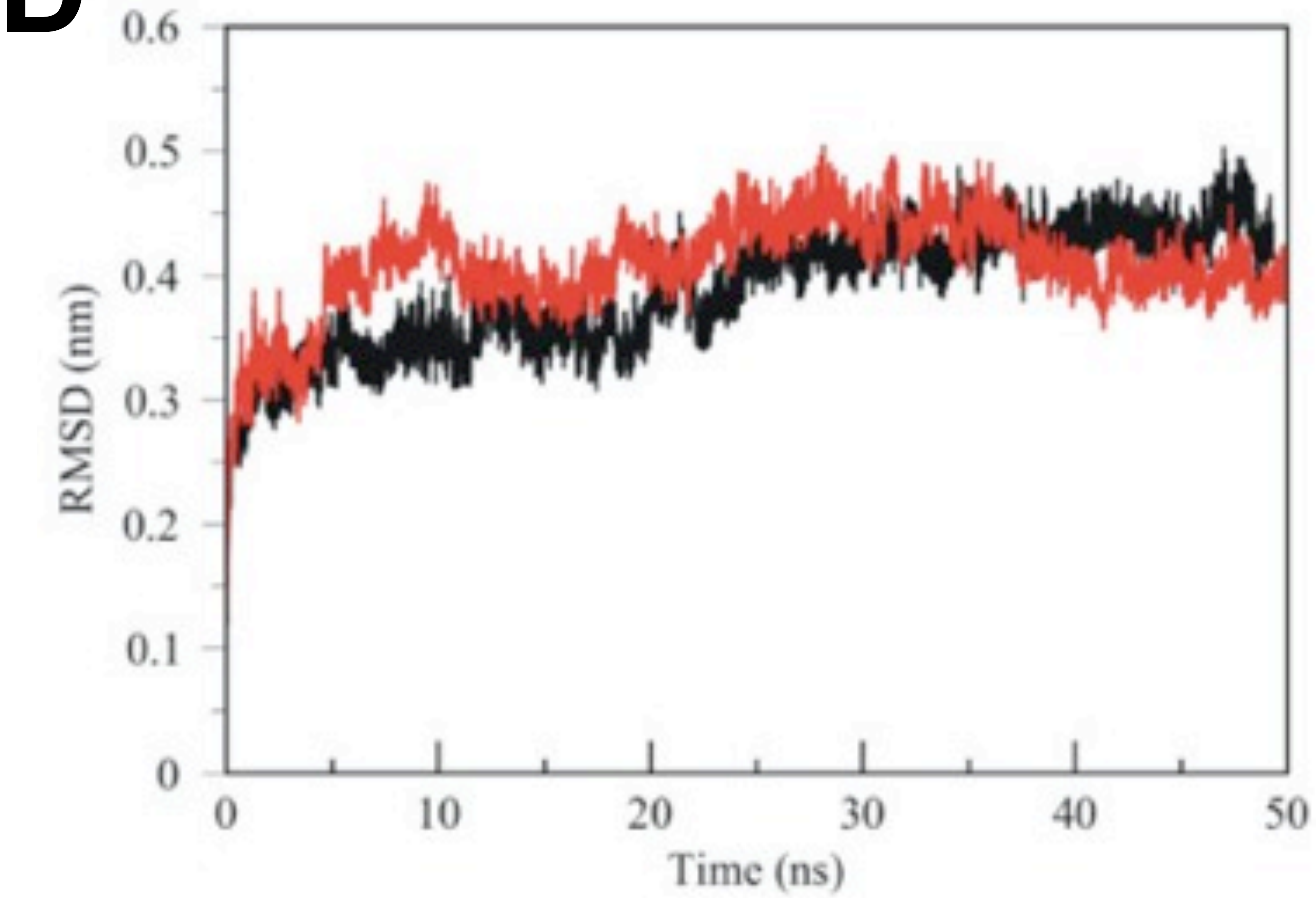




Deviation and fluctuation from reference

- **Root mean square deviation, RMSD**

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - \bar{r}_i)^2}$$



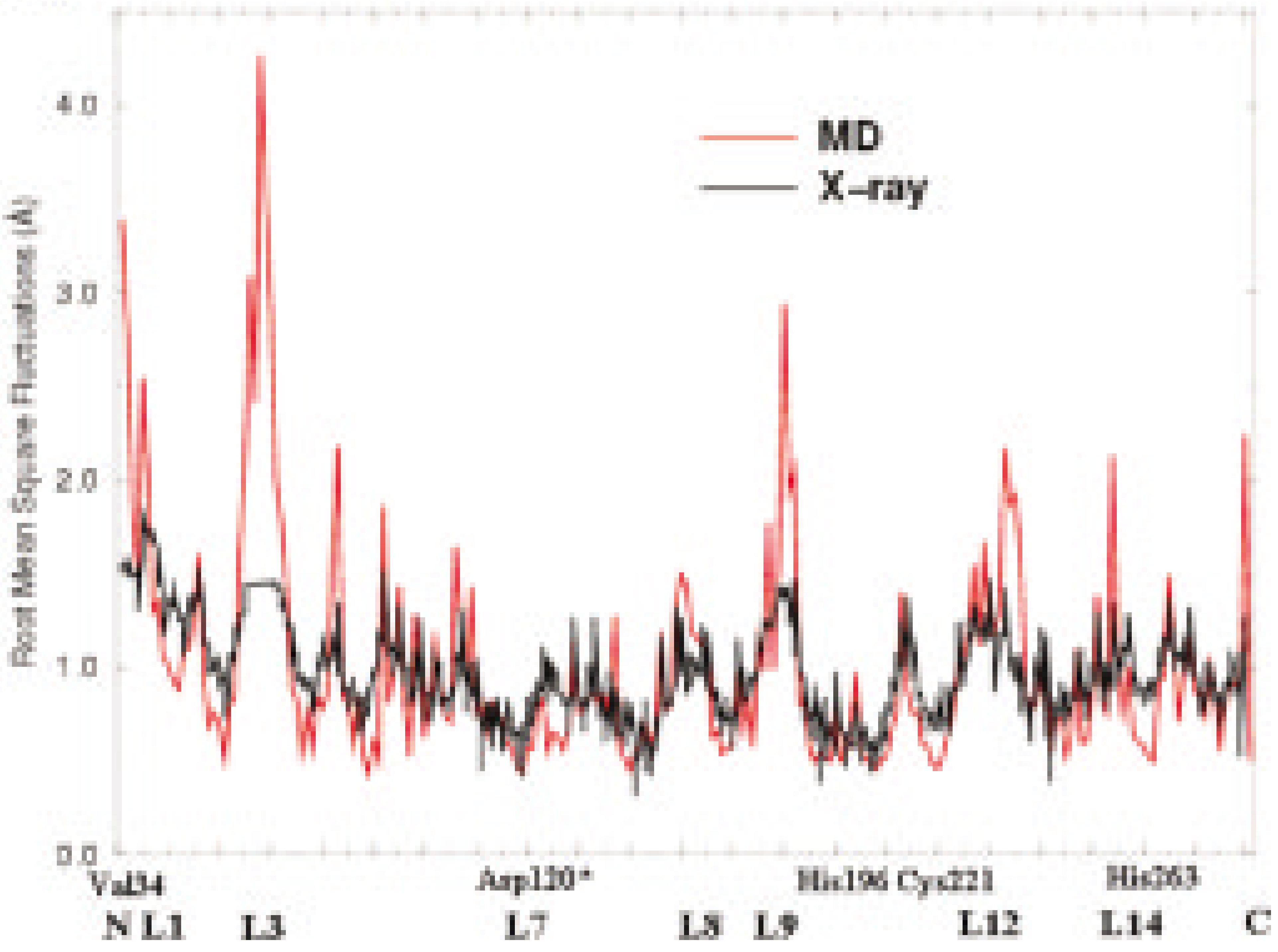
- **Root mean square fluctuations, RMSF**

$$RMSF = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(r_i(t) - \bar{r}_i)^2}$$

$$RMSF^2 = \langle u_r^2 \rangle$$

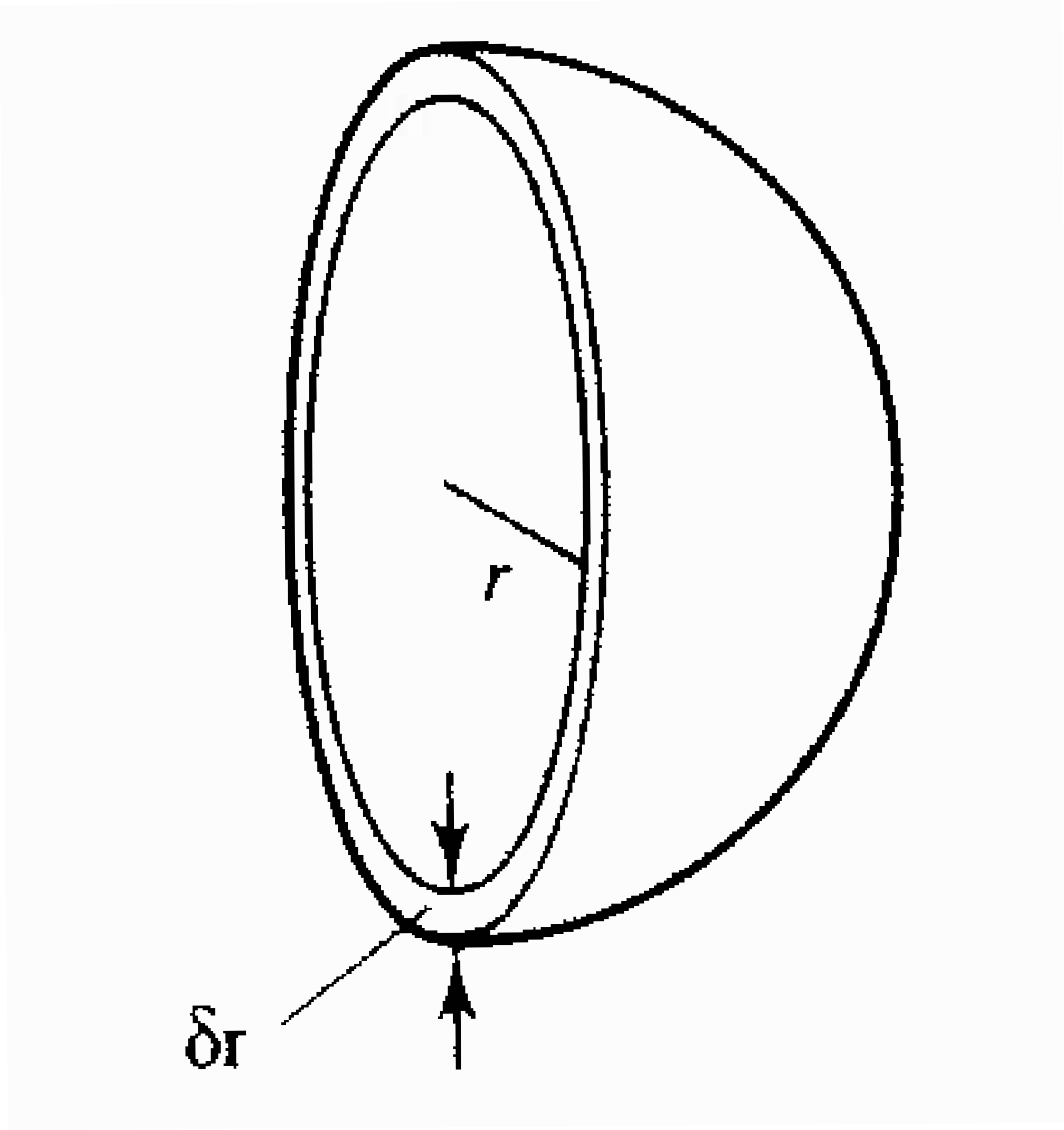
$$B = \frac{8\pi^2}{3} RMSF^2$$

atomic fluctuation ~0.25-0.60(Å)



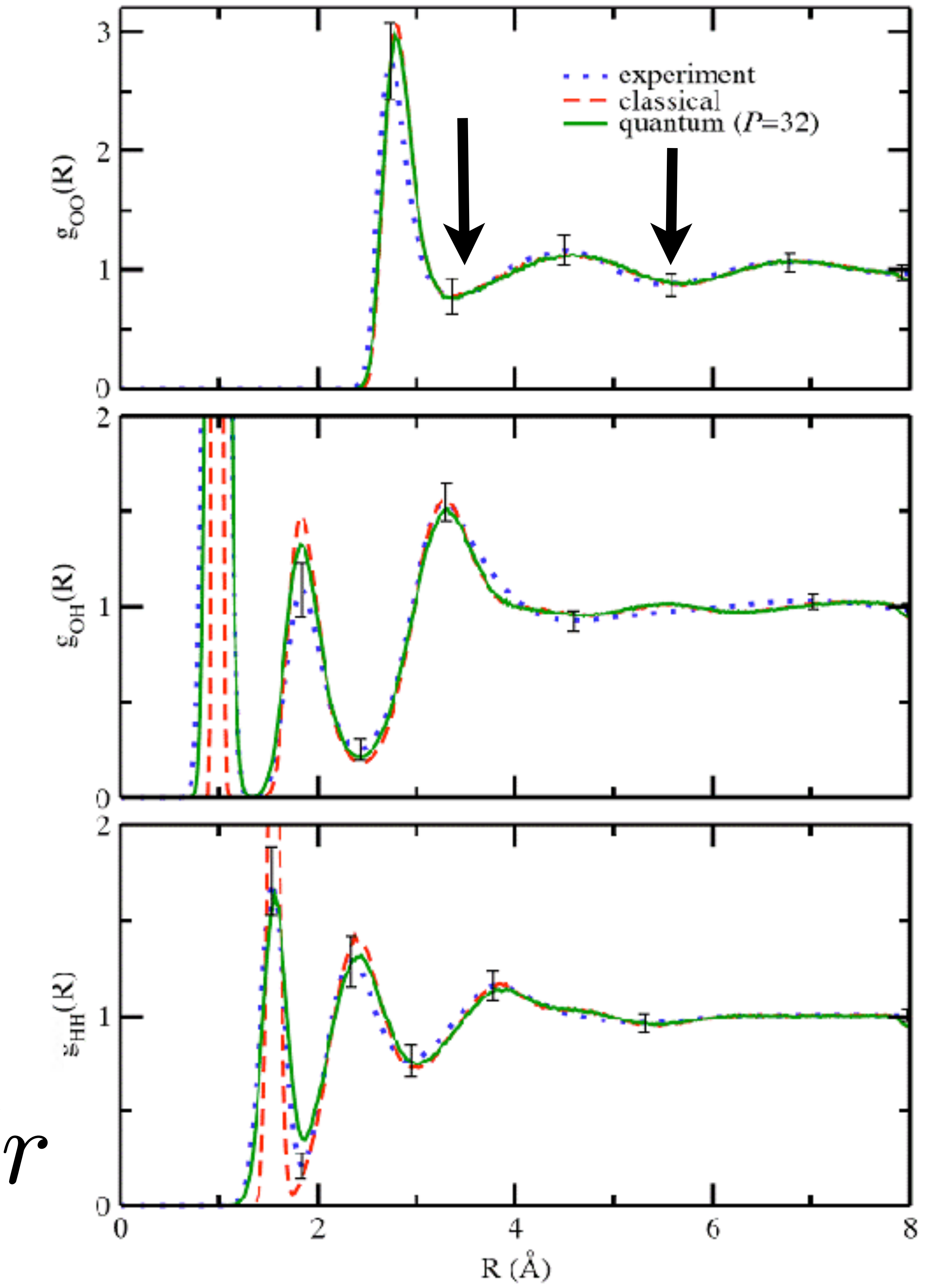
Radial distribution function

- **radial distribution function g(r):** describes the structure of a system (e.g. liquid water)



$$V = 4/3\pi(r+\delta r)^3 - 4/3\pi r^3$$
$$= 4\pi r^2\delta r + 4\pi r\delta r^2 + 4/3\pi\delta r^3 \approx 4\pi r^2\delta r$$

- if the number of particles per unit volume is ρ, then the total number in the shell is $4\pi\rho r^2\delta r$



- g(r) gives the probability to find an atom at a distance *r* from another atom (normalized to the ideal gas distribution)

- can be measured experimentally with X-ray diffraction

- **coordination number**: $CN(r) = 4\pi\rho\int g(r)r^2dr$

Time-dependent properties

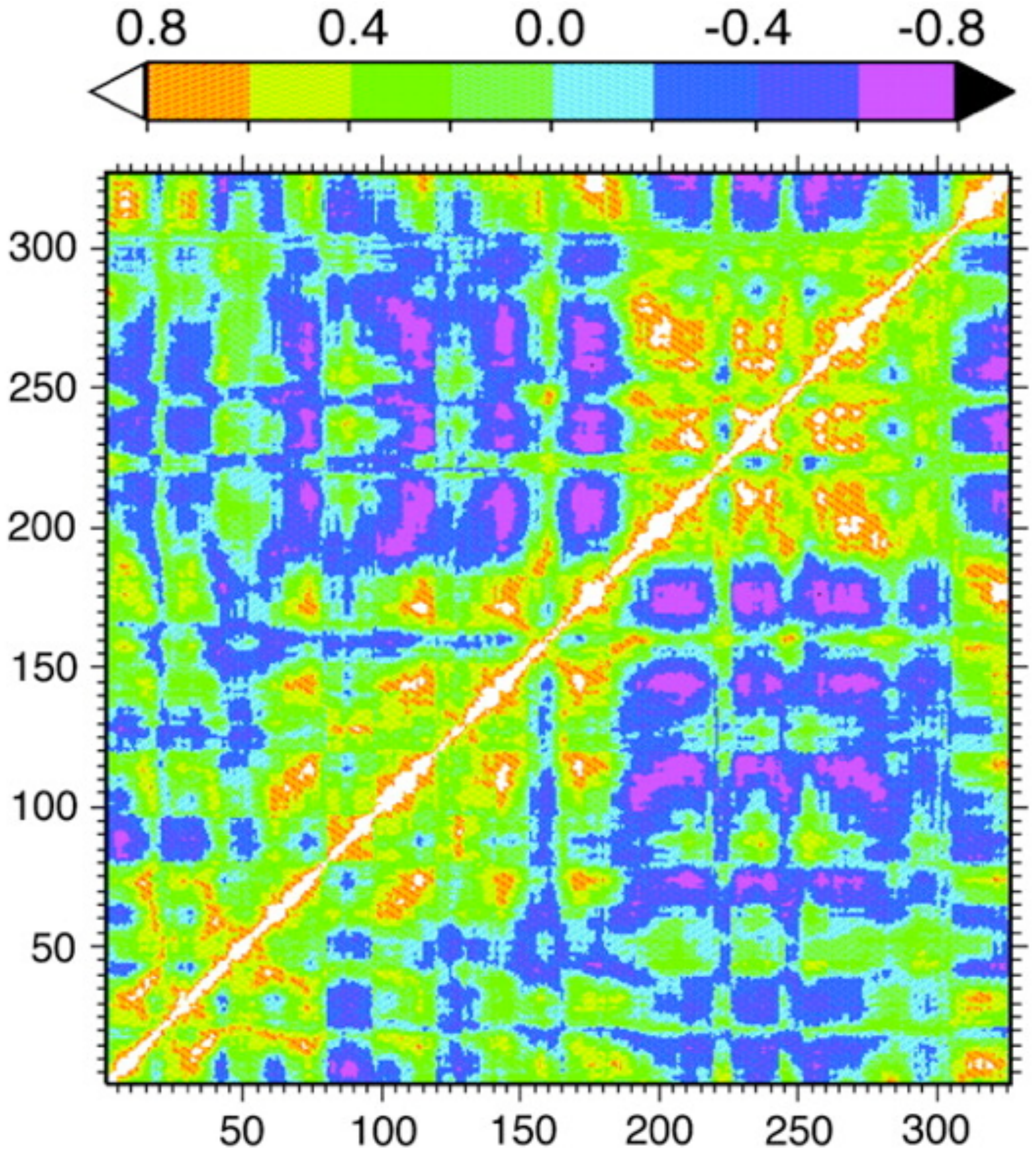
- **correlation function** between ***x*** and ***y*** dataset can be extracted from MD trajectories

$$C_{xy} = \frac{1}{M}\sum_{i=1}^{M} x_i y_i \equiv \langle x_i y_i \rangle$$

- if normalized you obtain data in [-1,1] range and w.r.t. mean values

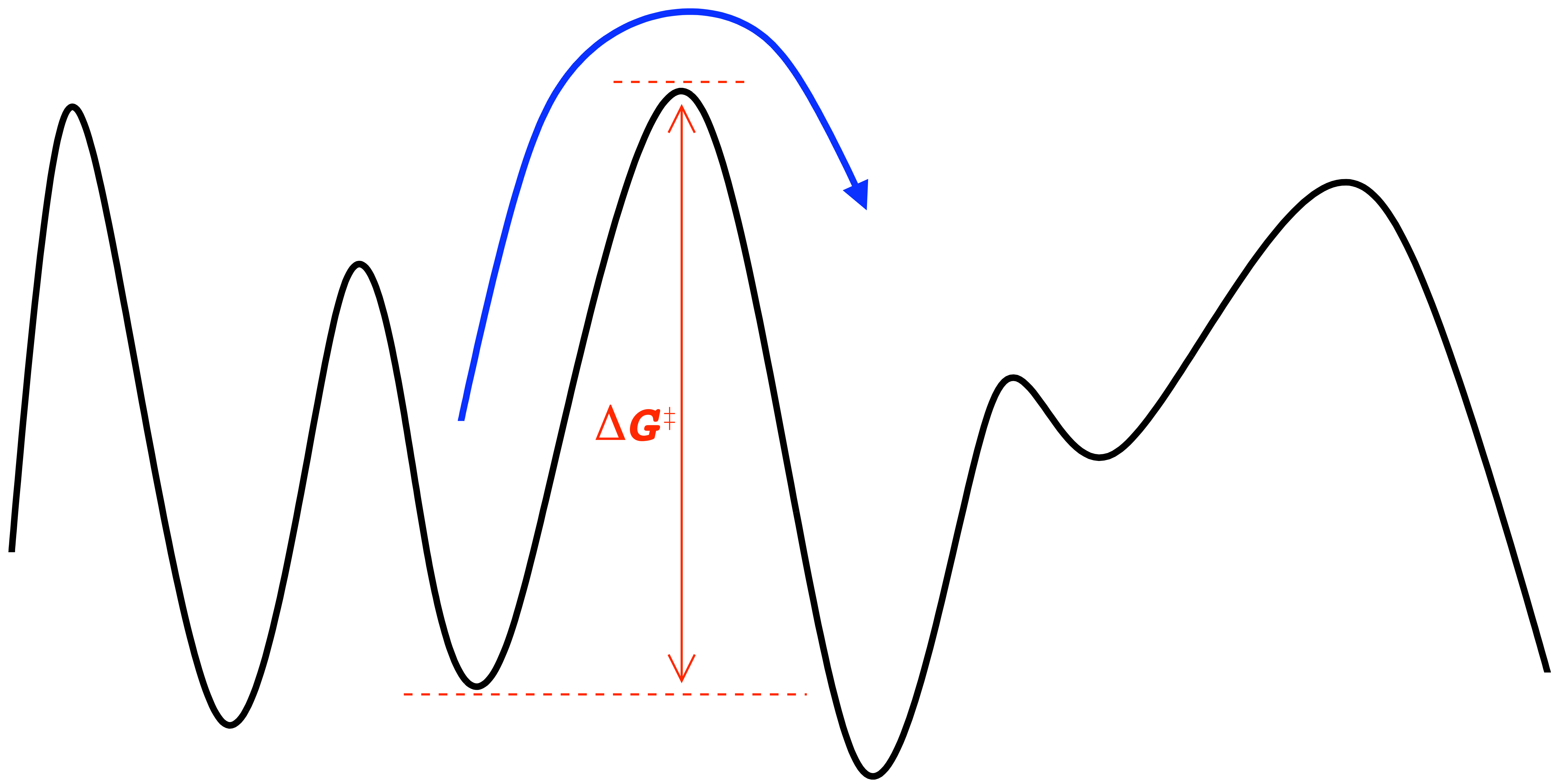
$$c_{xy} = \frac{\frac{1}{M}\sum_{i=1}^{M}(x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\left(\frac{1}{M}\sum_{i=1}^{M}(x_i - \langle x \rangle)^2\right)\left(\frac{1}{M}\sum_{i=1}^{M}(y_i - \langle y \rangle)^2\right)}} =$$

$$= \frac{\langle (x_i - \langle x \rangle)(y_i - \langle y \rangle) \rangle}{\sqrt{\langle (x_i - \langle x \rangle)^2 \rangle \langle (y_i - \langle y \rangle)^2 \rangle}}$$



Sampling the free energy landscape

- reaction coordinates, barrier crossing: ΔG or ΔF

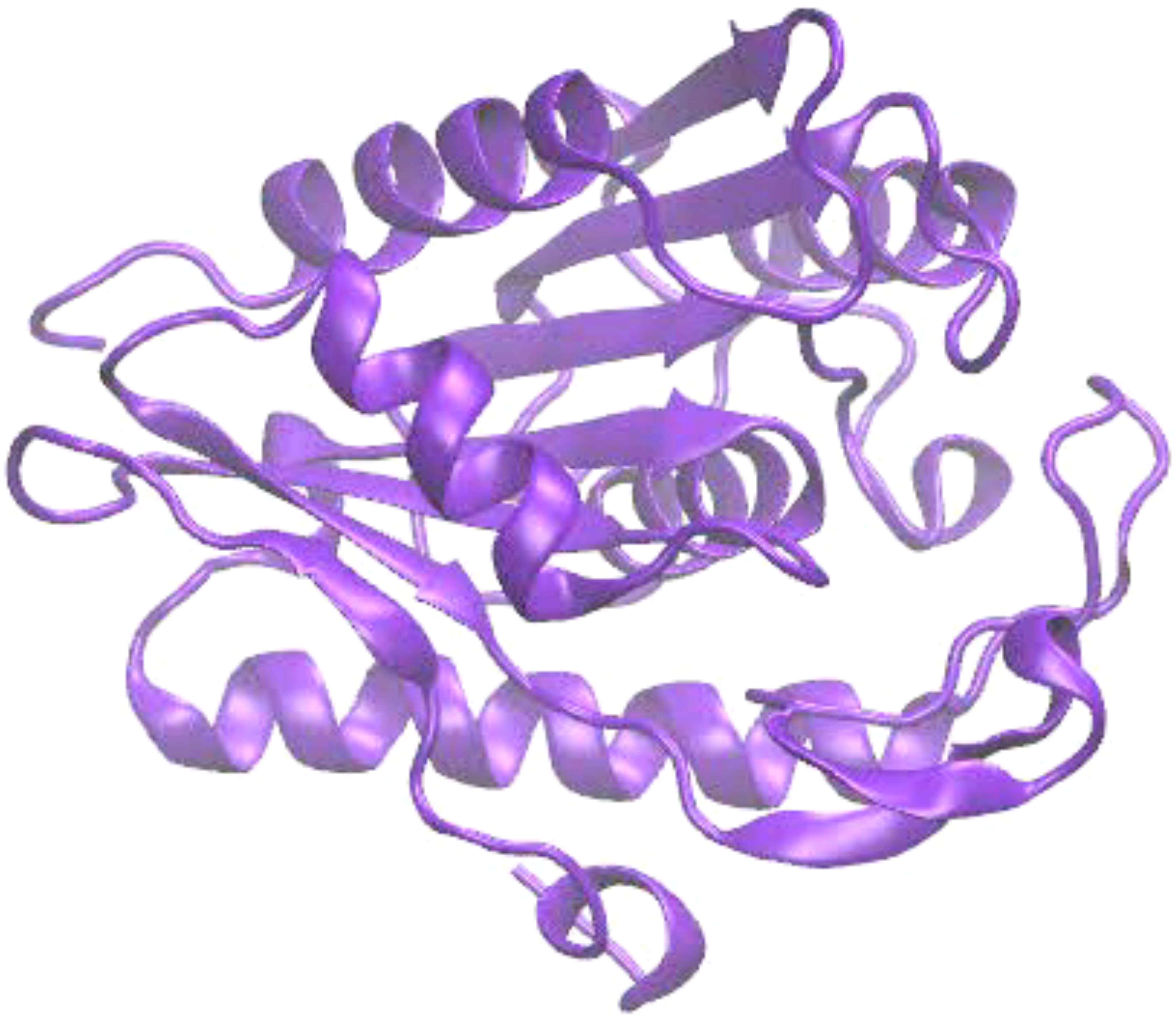


- to cross a free-energy barrier: $\tau = \tau_0 exp(\Delta G/k_B T)$ with τ_0 ~10^{-12} s: i.e. 1 kcal/mol barrier can be explored in ~ps; 5 kcal/mol in ~ns; 10 kcal/mol µs or longer

- rule of thumb: sampling should exceed timescales of interest by ~10-fold.

X-ray crystallography

$$\{x_i, y_i, z_i\}_{i=1,\ldots,N}$$

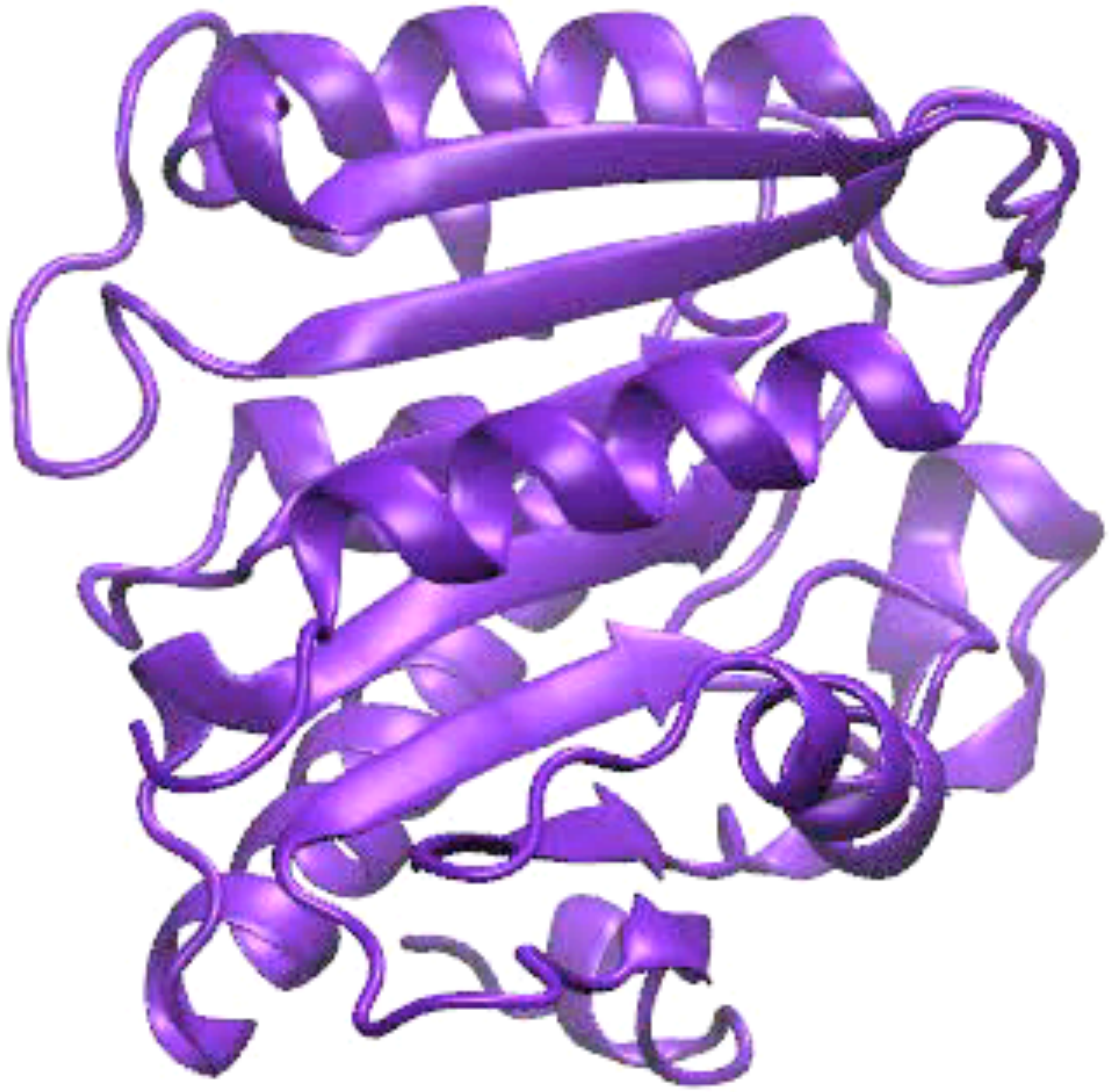


(human acyl-protein thioesterase)

molecular modeling and simulations

$$\{x_i(t), y_i(t), z_i(t)\}_{i=1,\ldots,N}$$

solvation
pH
post-translational modifications
interactions network
temperature effects (k_BT)
.....

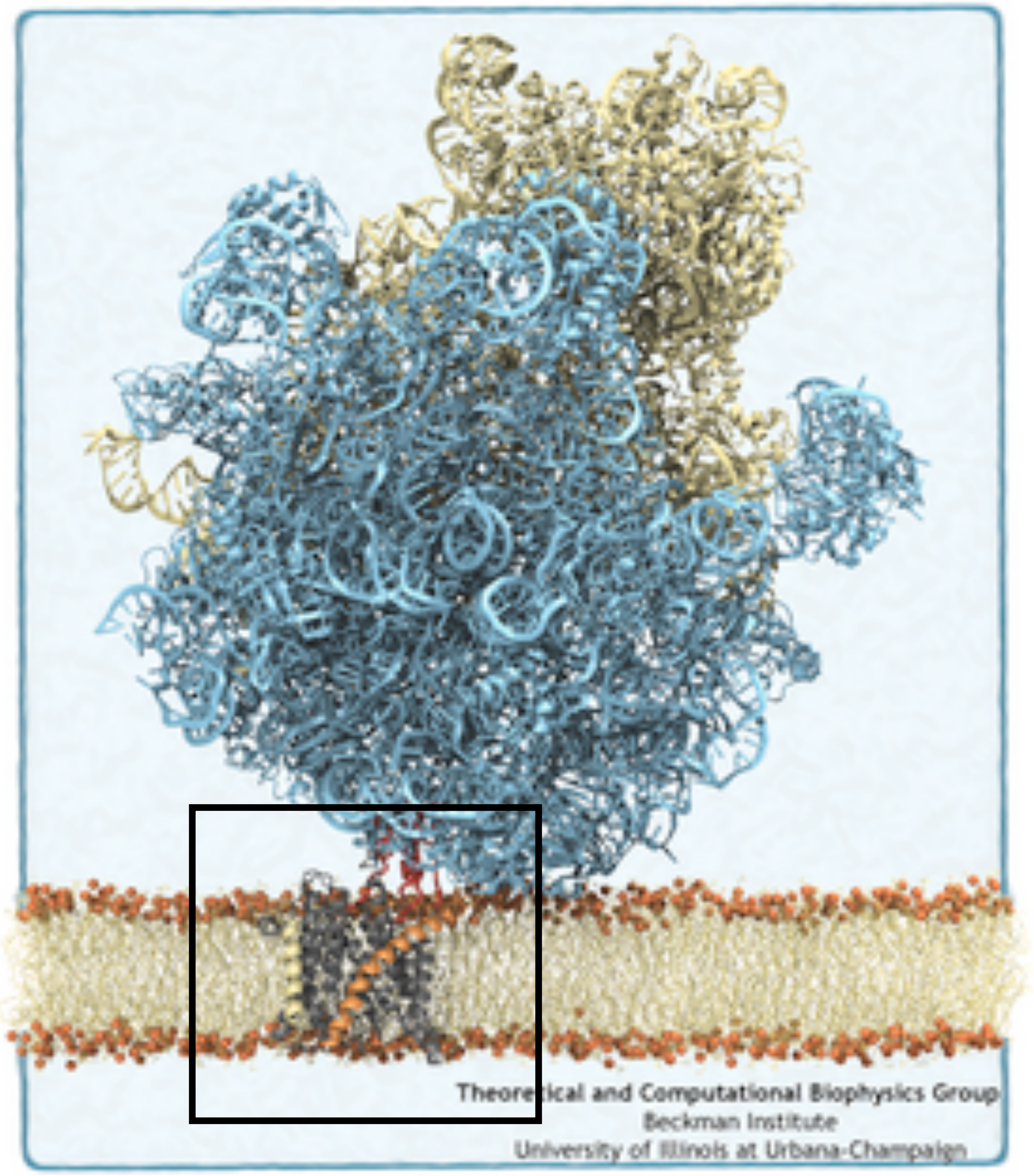


Laboratory for Biomolecular Modeling - LBM, http://lbm.epfl.ch

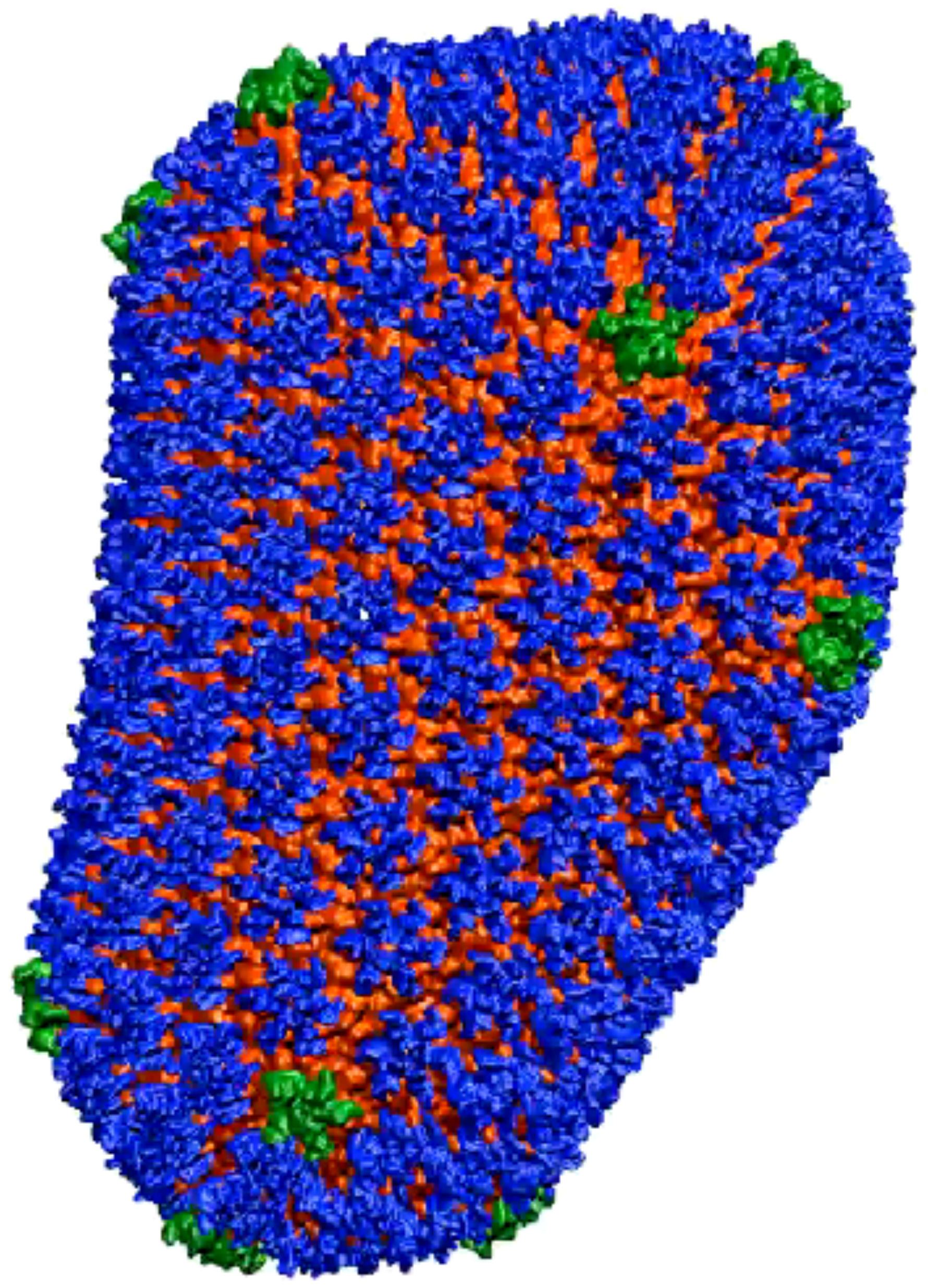
"Immune Recognition" by David Goodsell

State-of-the-art of molecular simulations

- up to **10^2 millions** of atoms (e.g. viruses, ribosome)

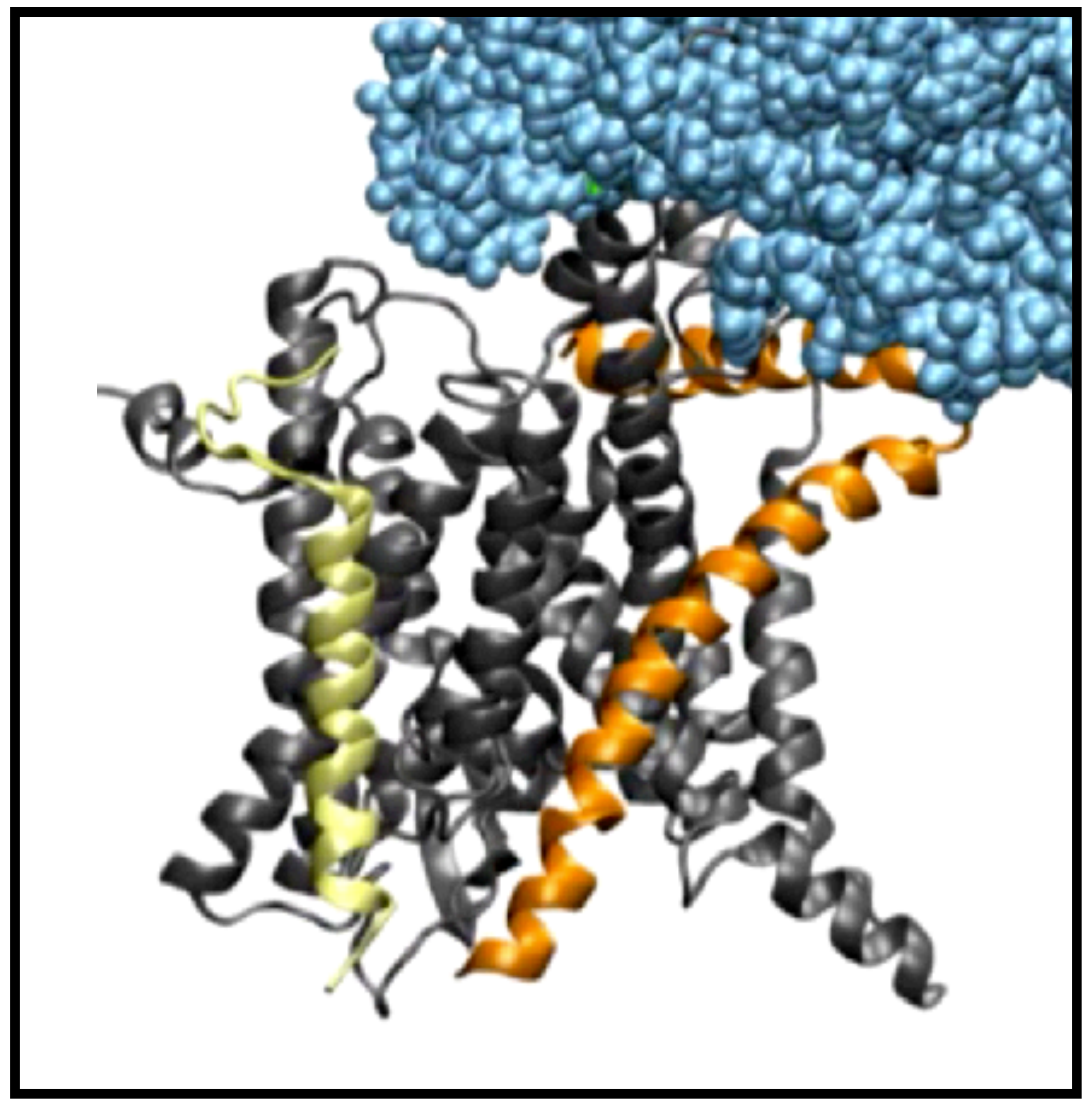

Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign

HIV-1 capsid



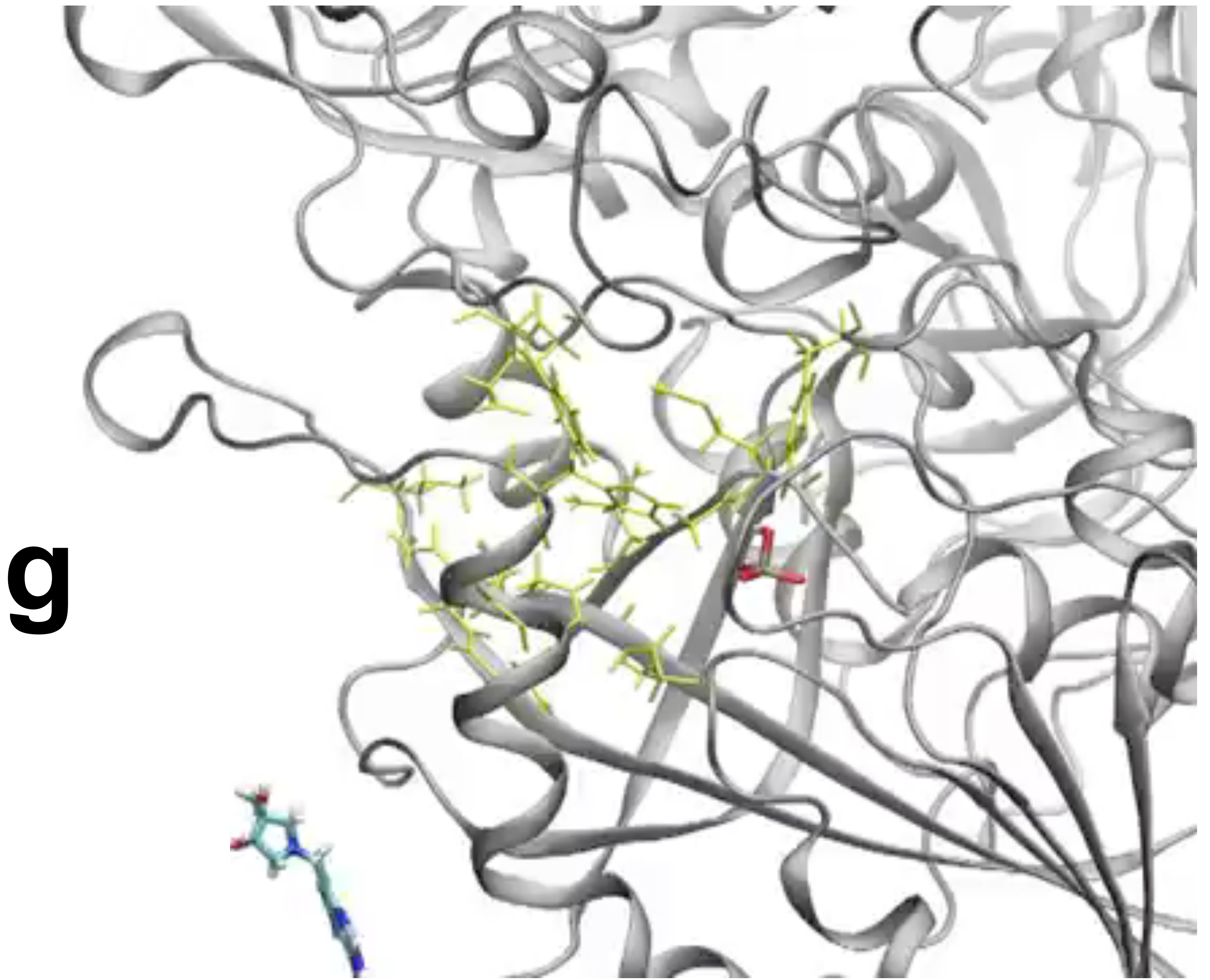
Zhao et al. *Nature*, 497:643-646, 2013
http://www.youtube.com/watch?v=pupVZI347H0

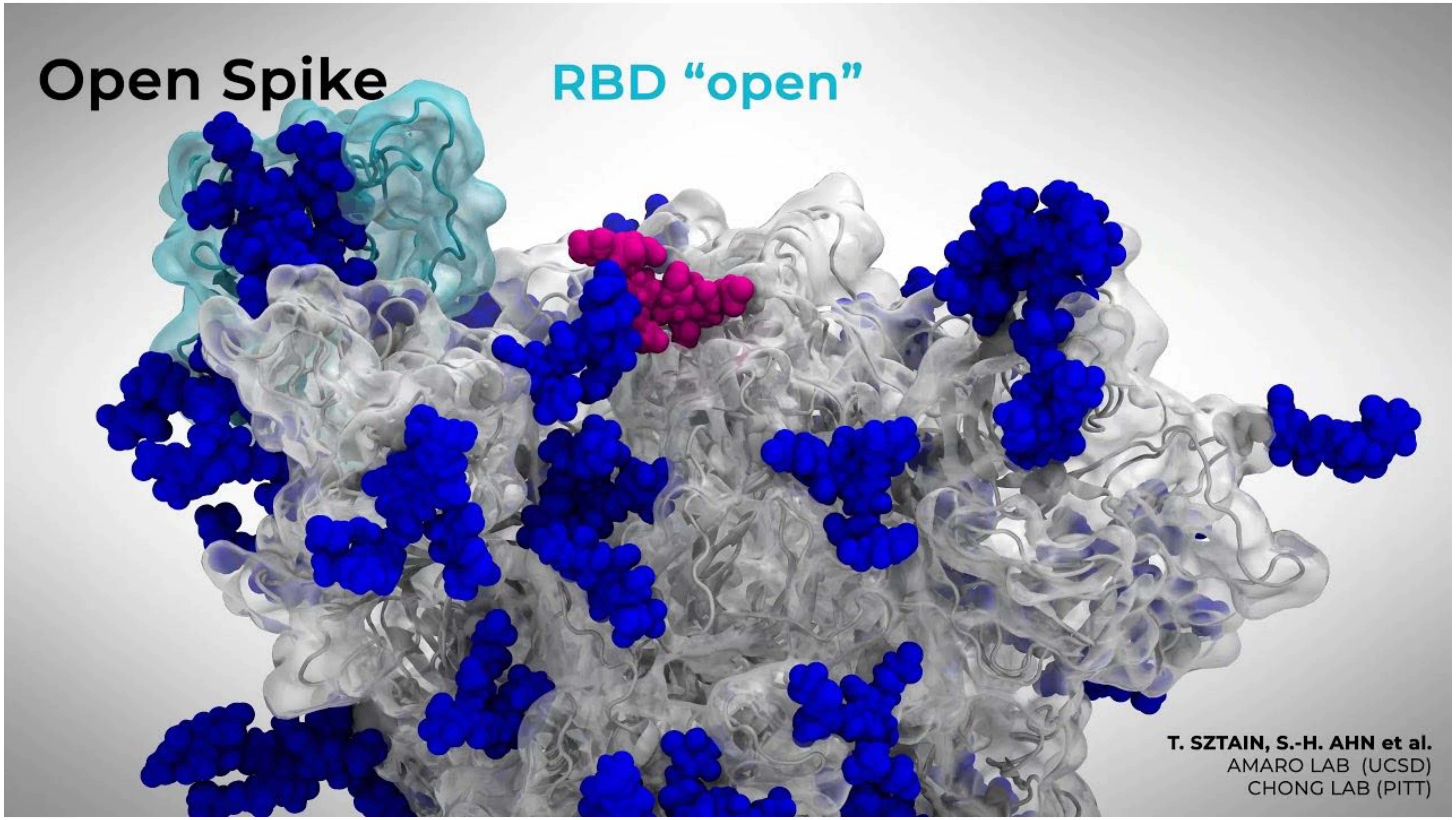
James Gumbart, et al.
Structure, 17:1453-1464, 2009.



protein translocation

drug binding on a kinase

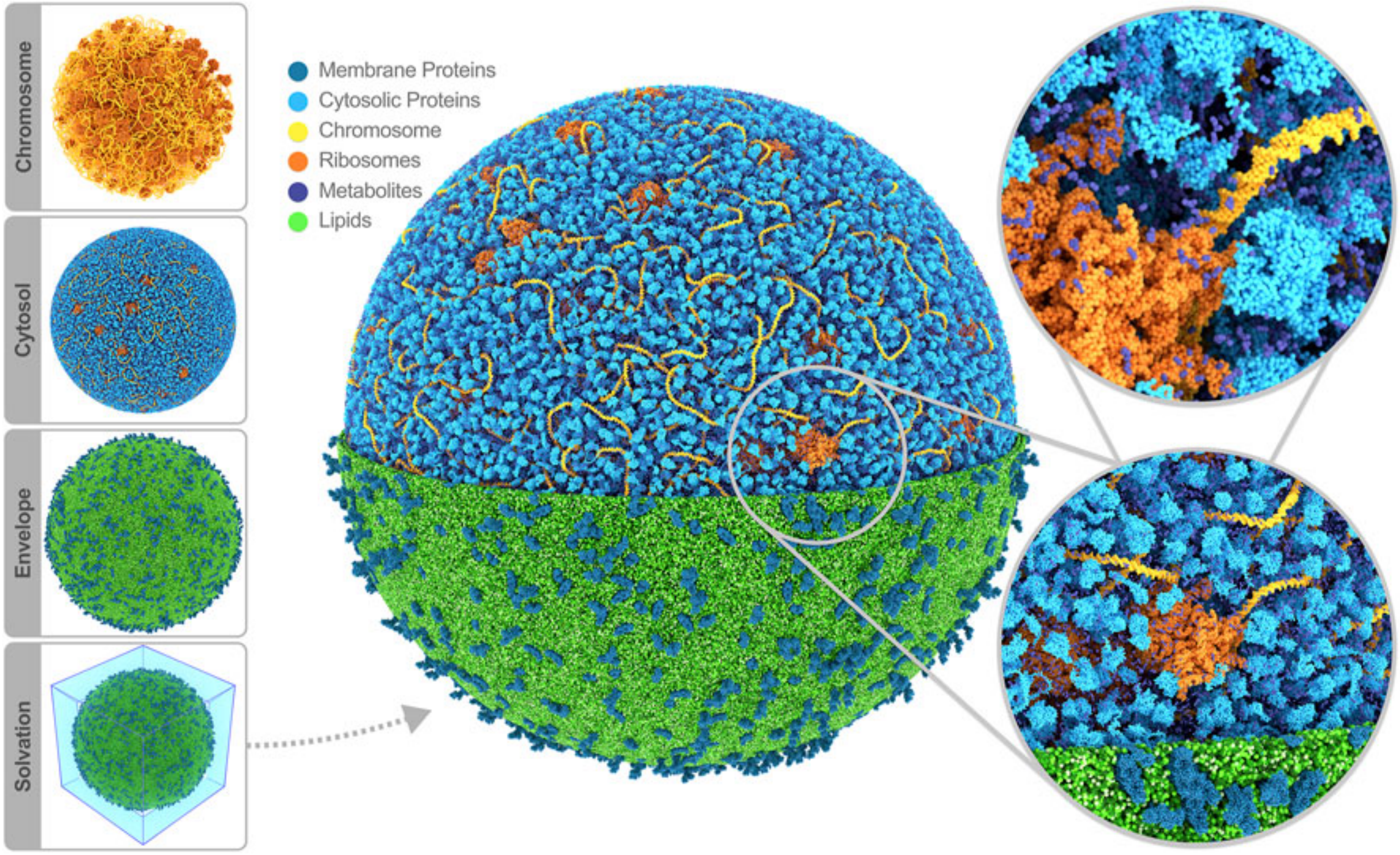




T. SZTAIN, S.-H. AHN et al.
AMARO LAB (UCSD)
CHONG LAB (PITT)

Molecular mechanism of SARS-CoV2

State-of-the-art of molecular simulations

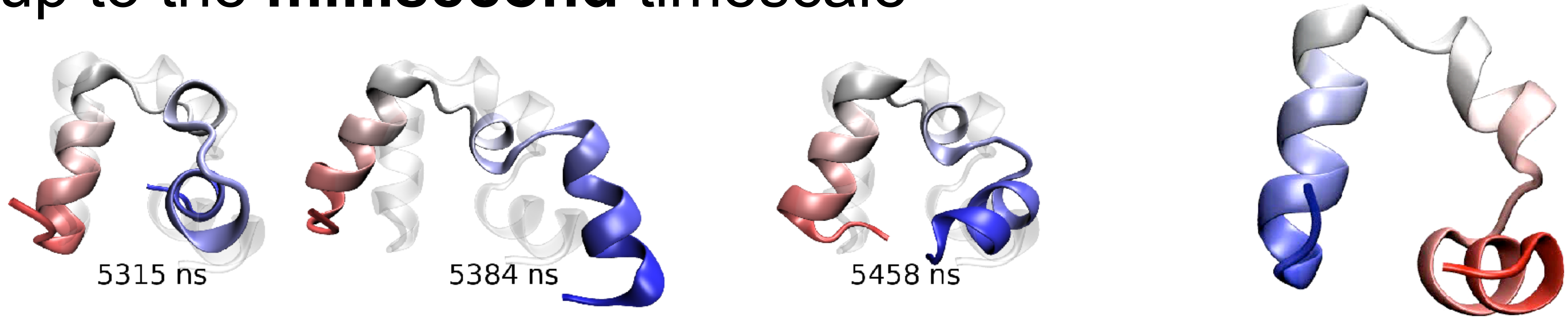


Whole-cell Martini model of **JCVI-syn3A**. The four stages of cell building are shown on the side. The final system contains 60,887 soluble proteins (light blue), 2,200 membrane proteins (blue), 503 ribosomes (orange), a single 500 kbp circular dsDNA (yellow), 1.3 million lipids (green), 1.7 million metabolites (dark blue), 14 million ions (not shown) and 447 million water beads (not shown) for a total of 561 million beads representing more than **six billion atoms**.

https://doi.org/10.3389/fchem.2023.1106495

State-of-the-art of molecular simulations

- up to the **millisecond** timescale



villin folding Freddolino, et al.. *Biophysical Journal*, 94:L75-L77, 2008.

Voelz *et al. J. Am. Chem. Soc.*, **2010**, *132*, 1526 http://www.youtube.com/watch?v=gFcp2Xpd29I

High-Performance Computing (HPC) resources

KUMA - EPFL HPC - 12 PetaFLOPS

HPC@EPFL BlueGene/P

TOP500

http://www.top500.org/lists

CSCS ALPS- 435 PetaFlops

Anton D.E. Shaw Research

Current limitations of MD simulation

- approximations and errors inherent to any force field
- systematic errors related with algorithm precision
- calculations of free energy differences are still very difficult to converge
- **time scale** and **sampling** problem → statistical error: conformational transitions that require >10 μs cannot be easily simulated by conventional molecular dynamics techniques (this is related to **sizescale** as well)
- some solution for sampling: **enhanced sampling** techniques, MD with implicit solvent (approximate) – **Brownian dynamics** – **Monte Carlo**, **coarse-grained MD** (see in the next lectures)

MM FF limitations

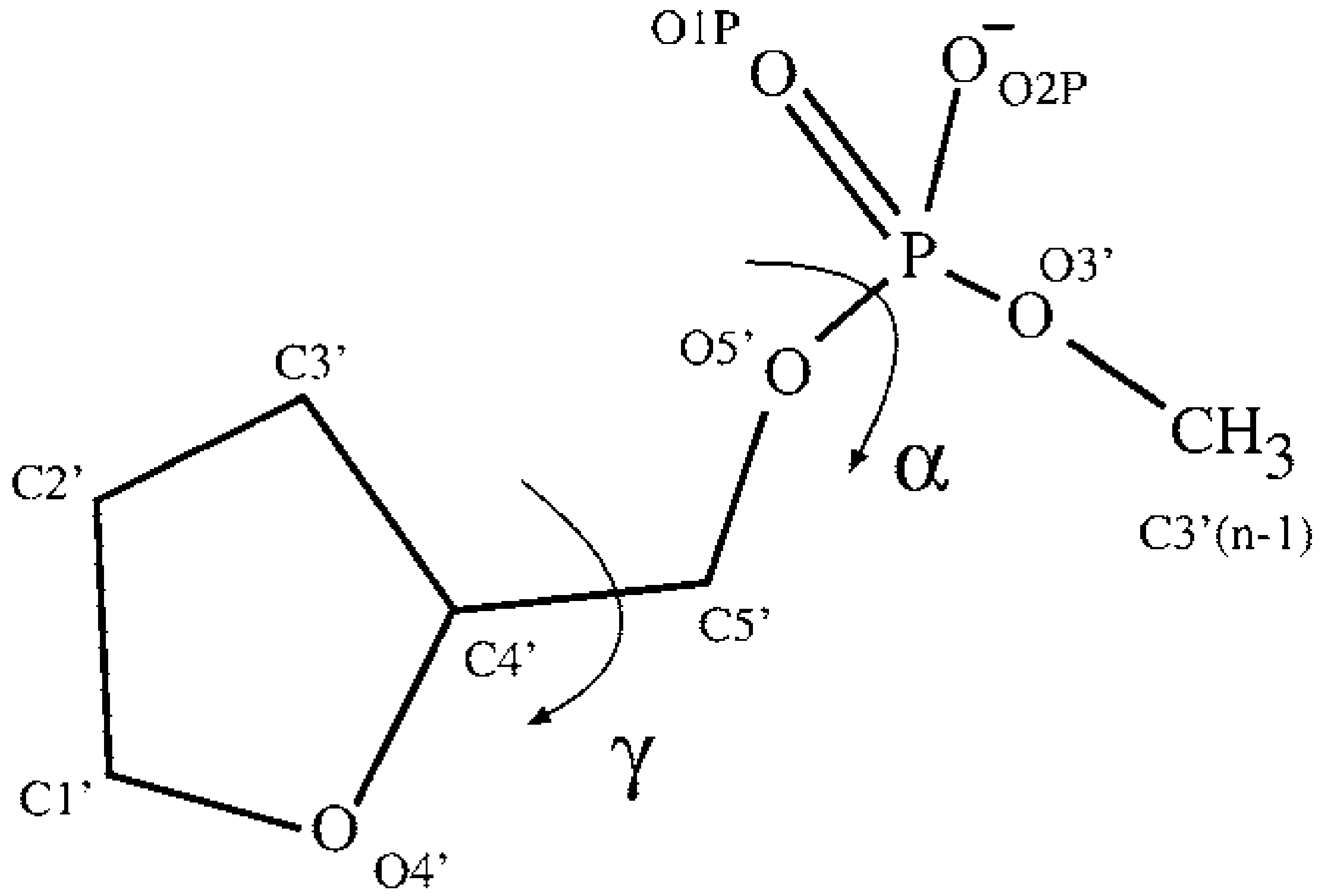
- transferability
- accuracy of parametrization
- functional form (e.g. can add polarizability)

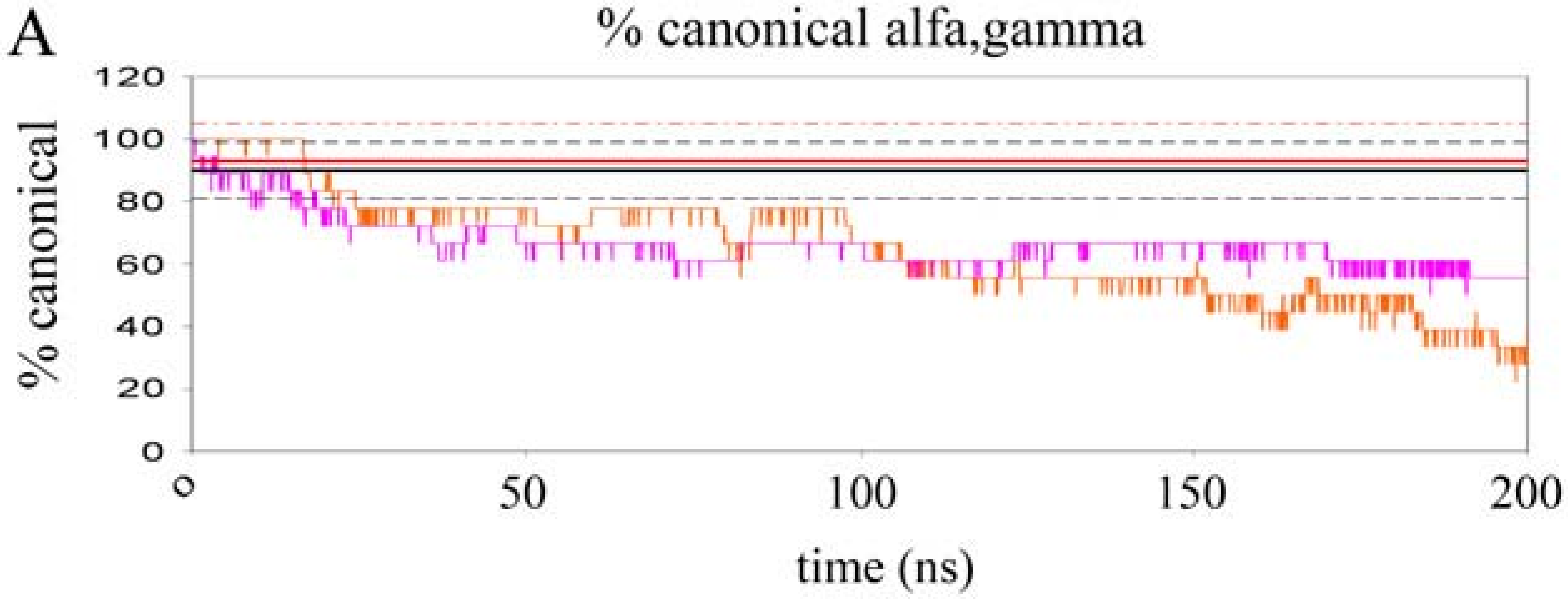
  $\mu_{ind} = \alpha \mathbf{E} \qquad \alpha : polarizability$

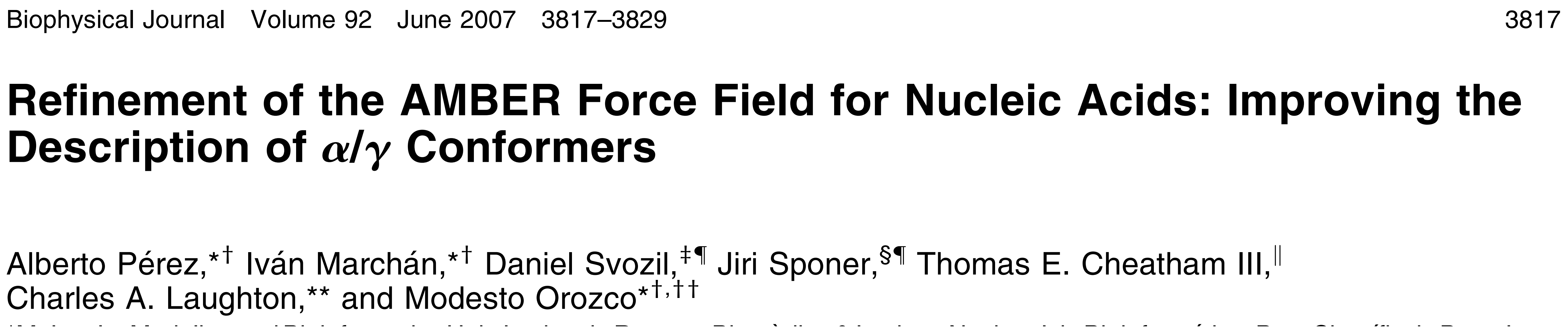
  or many-body terms
- many different force fields (specific vs. generalized)
- approximation in treating long-range interactions
- can be expensive for very large systems (e.g. ~10^6 atoms)

Failure of a force field

- enhanced computer power allows to run longer MD simulations, and to discover failures in the models

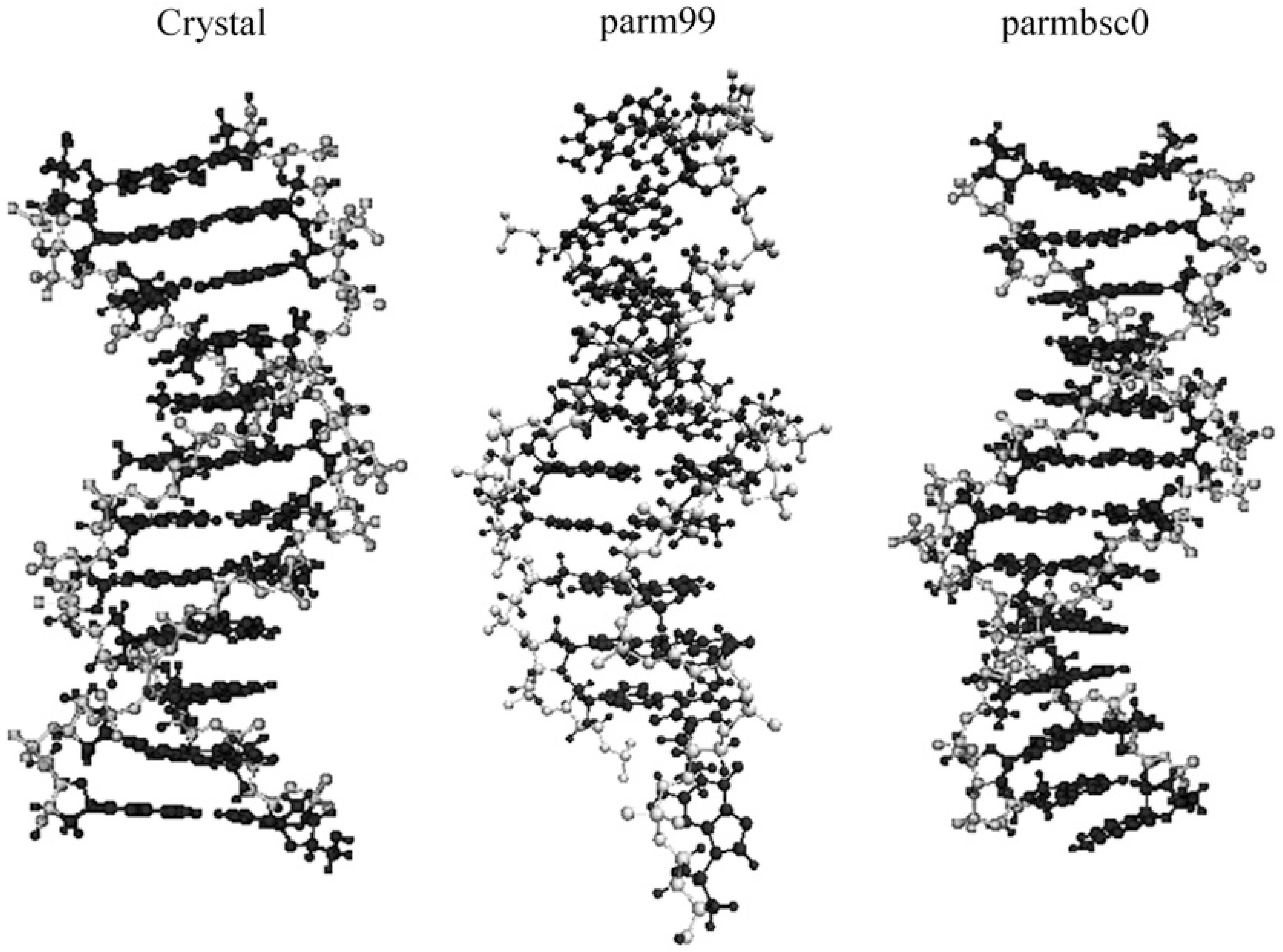





Biophysical Journal Volume 92 June 2007 3817–3829 3817

Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of α/γ Conformers

Alberto Pérez,*† Iván Marchán,*† Daniel Svozil,‡¶ Jiri Sponer,§¶ Thomas E. Cheatham III,‖ Charles A. Laughton,** and Modesto Orozco*†,††



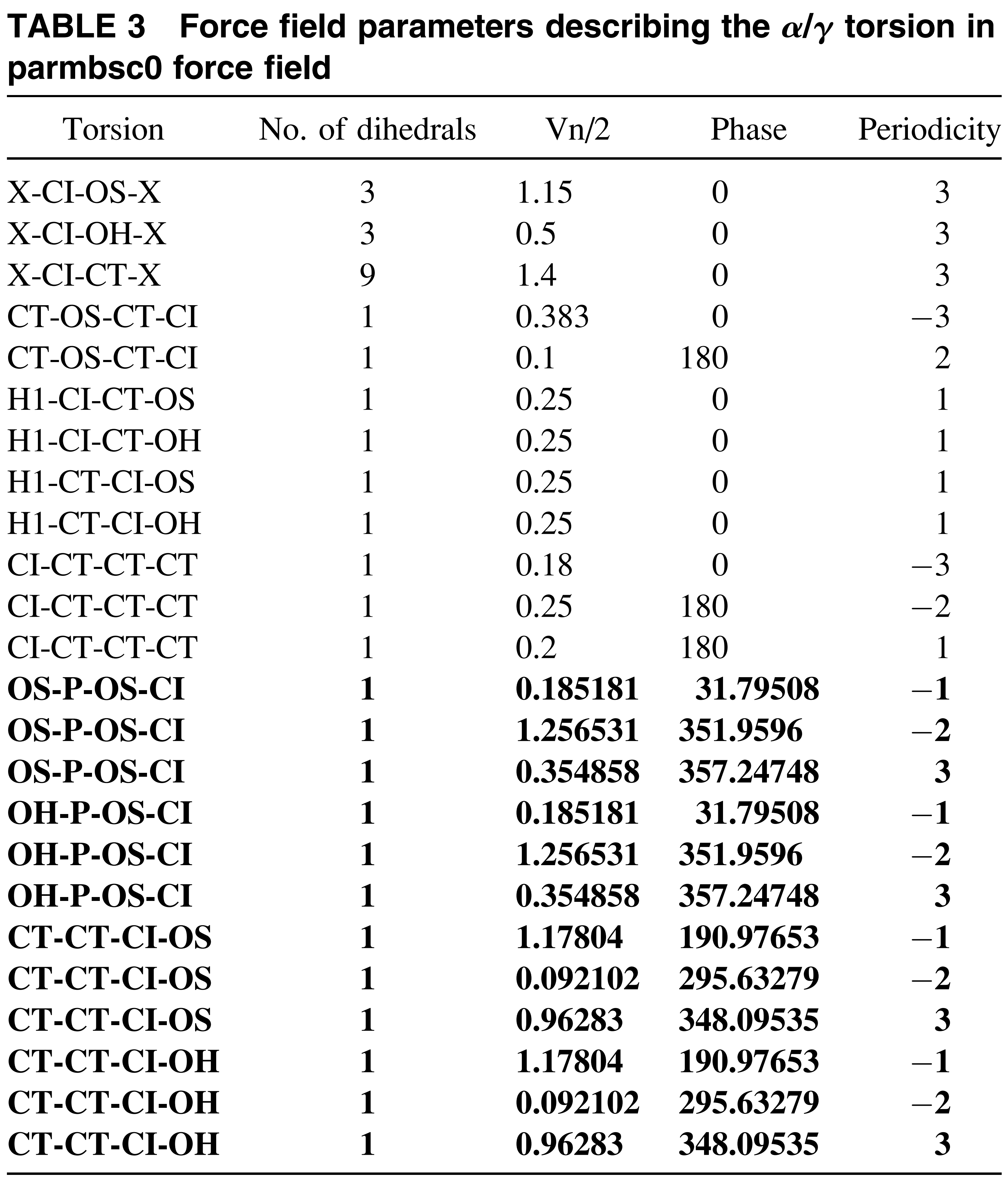


TABLE 3 Force field parameters describing the α/γ torsion in parmbsc0 force field

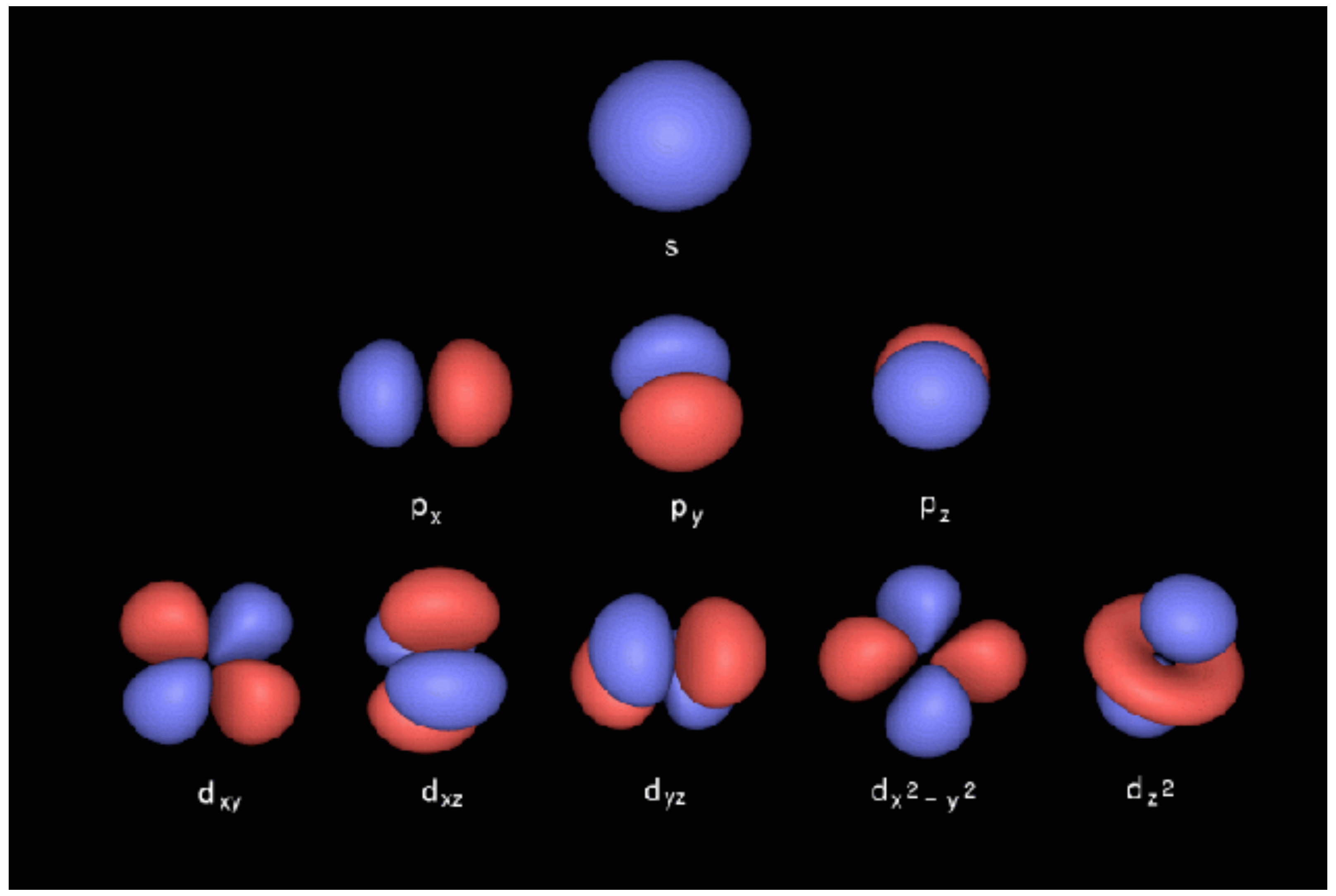
Torsion	No. of dihedrals	Vn/2	Phase	Periodicity
X-CI-OS-X	3	1.15	0	3
X-CI-OH-X	3	0.5	0	3
X-CI-CT-X	9	1.4	0	3
CT-OS-CT-CI	1	0.383	0	−3
CT-OS-CT-CI	1	0.1	180	2
H1-CI-CT-OS	1	0.25	0	1
H1-CI-CT-OH	1	0.25	0	1
H1-CT-CI-OS	1	0.25	0	1
H1-CT-CI-OH	1	0.25	0	1
CI-CT-CT-CT	1	0.18	0	−3
CI-CT-CT-CT	1	0.25	180	−2
CI-CT-CT-CT	1	0.2	180	1
OS-P-OS-CI	**1**	**0.185181**	**31.79508**	**−1**
OS-P-OS-CI	**1**	**1.256531**	**351.9596**	**−2**
OS-P-OS-CI	**1**	**0.354858**	**357.24748**	**3**
OH-P-OS-CI	**1**	**0.185181**	**31.79508**	**−1**
OH-P-OS-CI	**1**	**1.256531**	**351.9596**	**−2**
OH-P-OS-CI	**1**	**0.354858**	**357.24748**	**3**
CT-CT-CI-OS	**1**	**1.17804**	**190.97653**	**−1**
CT-CT-CI-OS	**1**	**0.092102**	**295.63279**	**−2**
CT-CT-CI-OS	**1**	**0.96283**	**348.09535**	**3**
CT-CT-CI-OH	**1**	**1.17804**	**190.97653**	**−1**
CT-CT-CI-OH	**1**	**0.092102**	**295.63279**	**−2**
CT-CT-CI-OH	**1**	**0.96283**	**348.09535**	**3**

Current common MD engines

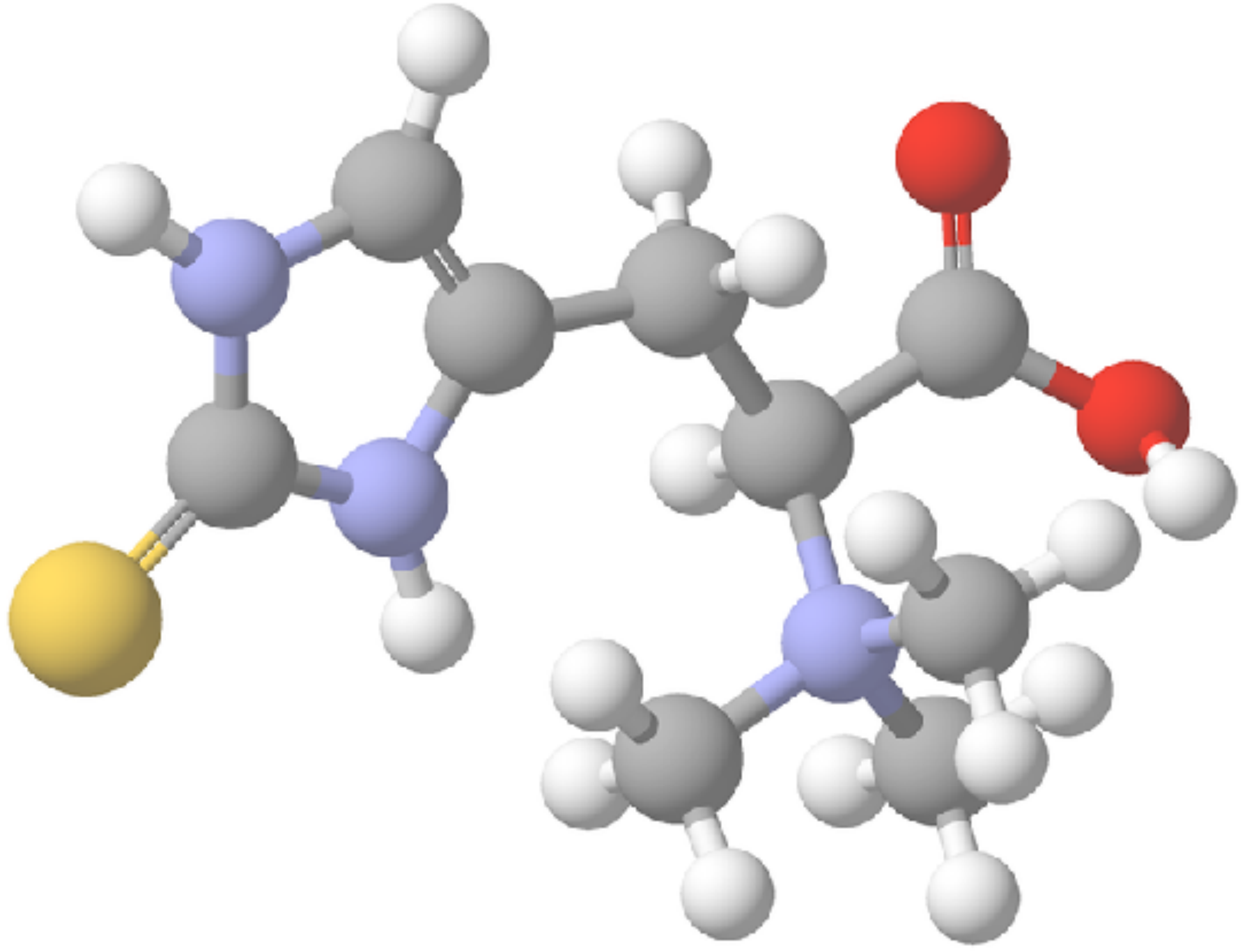
- **CHARMM**: Karplus Harvard, http://www.charmm.org/
- **AMBER**: Kollman UCSF, http://ambermd.org/
- **GROMOS**: van Gunsteren, ETHZ, www.igc.ethz.ch/GROMOS/index
- **DESMOND**: Shaw, http://www.deshawresearch.com/
- **GROMACS**: http://www.gromacs.org
- **LAMMPS**: http://lammps.sandia.gov
- **ACEMD**: http://multiscalelab.org/acemd
- **NAMD**: http://www.ks.uiuc.edu/Research/namd/

Multiscale resolution in modeling

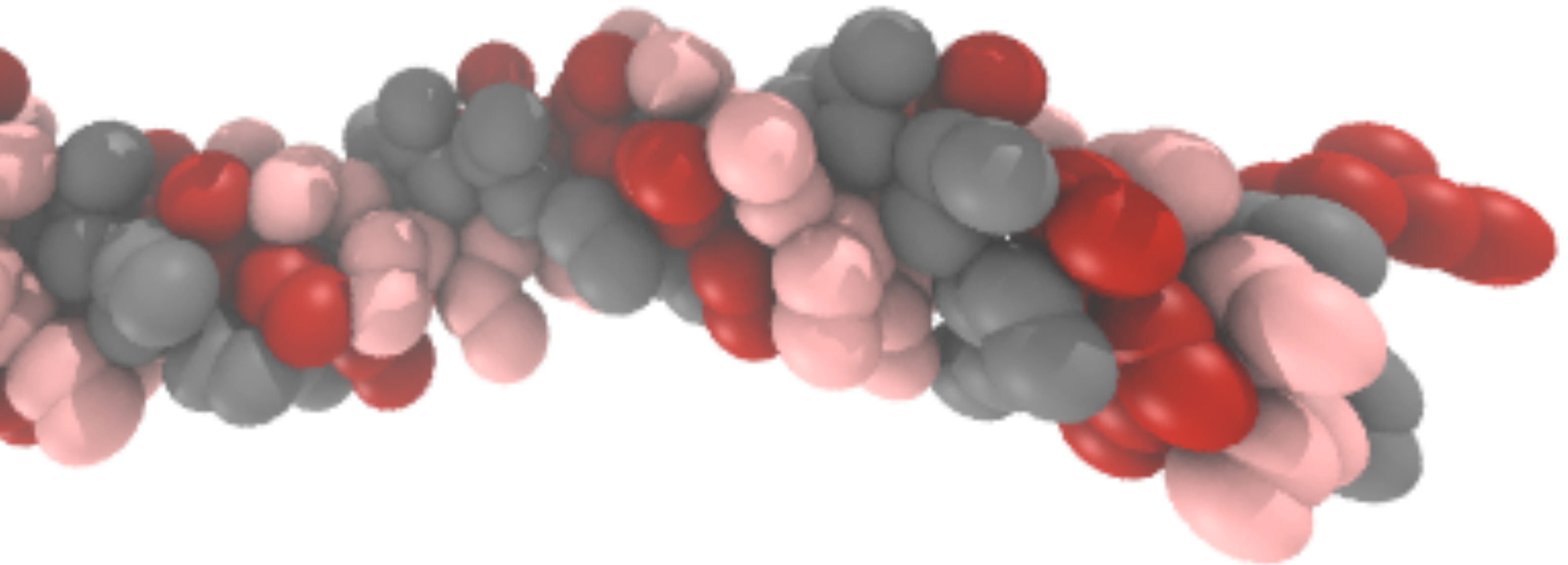
- electrons



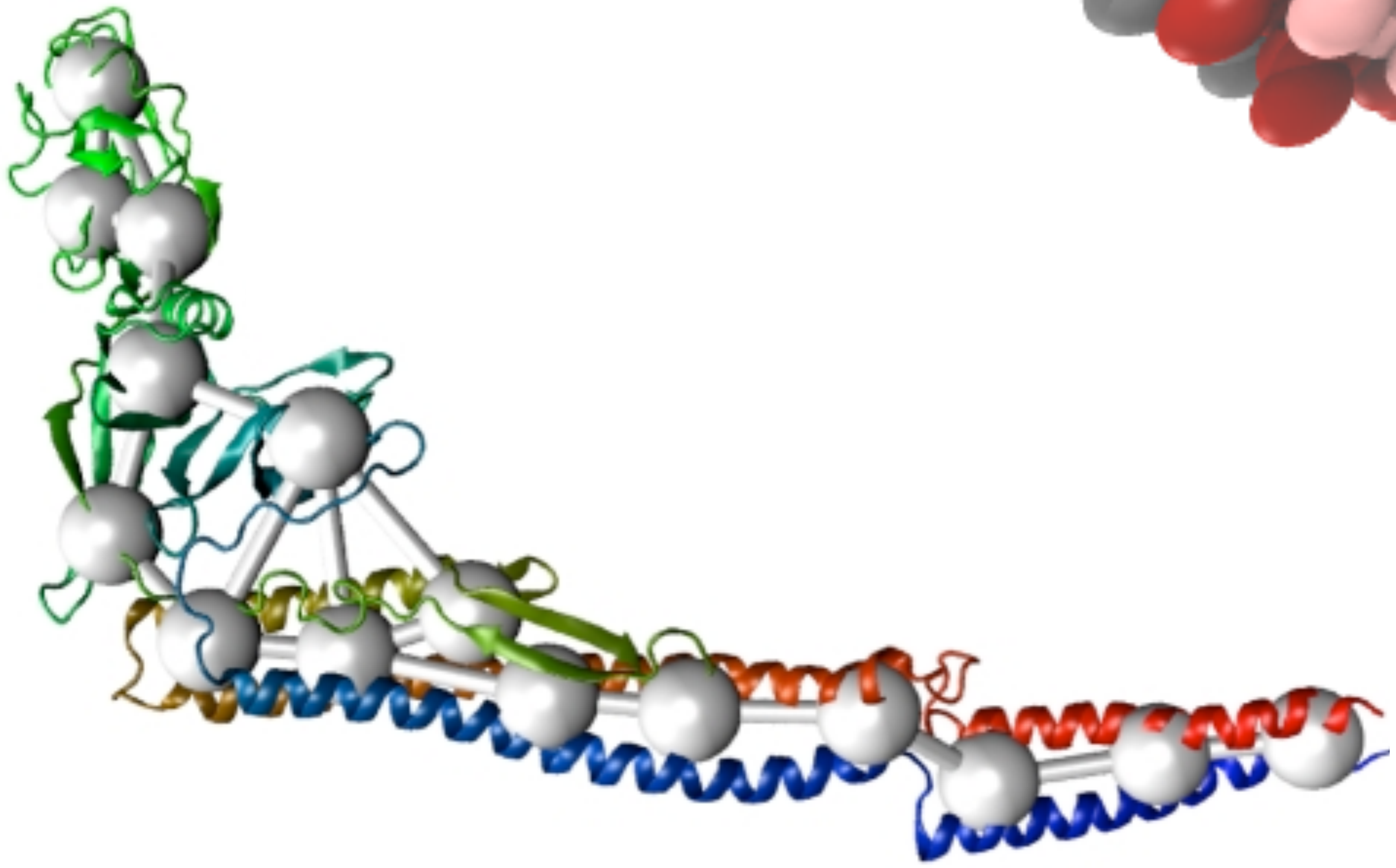
- atoms



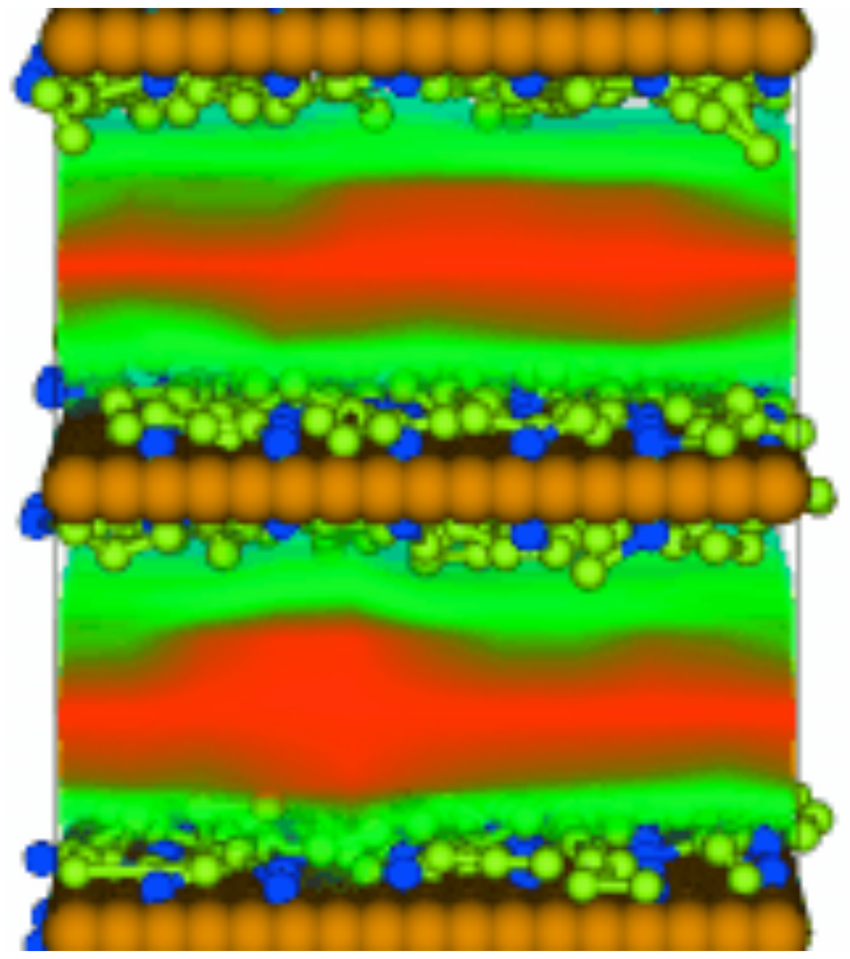
- amino-acids



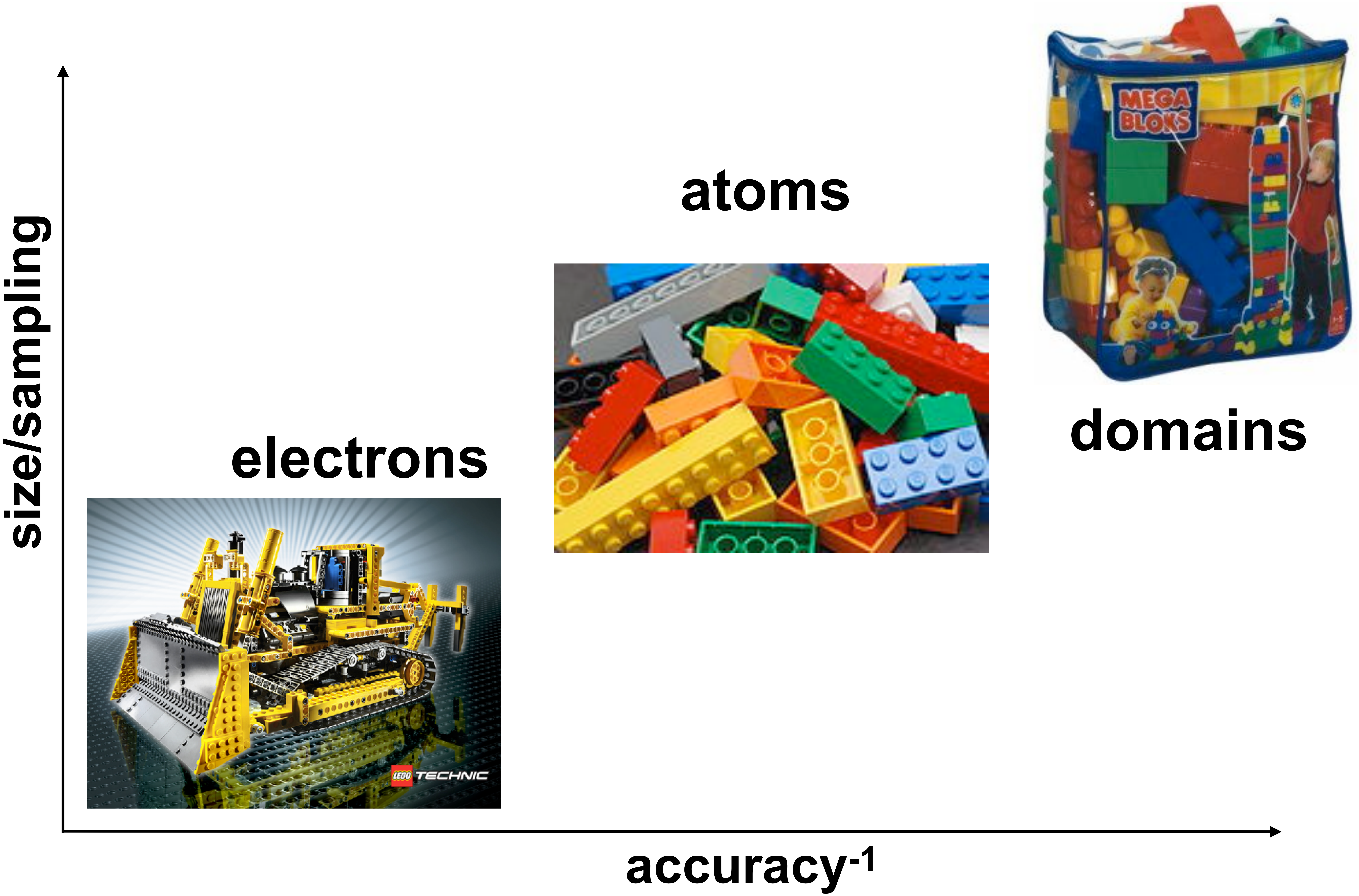
- domains



- mesoscopic to continuum



Building blocks



size/sampling

accuracy^{-1}

electrons

atoms

domains

chemical detail
sampling
size/atoms
10^8
10^6
10^4
10^2
enzymatic reactions
binding properties
folding
mesoscale
coarse-grained
MM-CG
QM-MM
MM
QM
structural fluctuations
self-assembly
fs
ps
ns
μs
ms
s
timescale

Speeding up timescales of Chemical Reactions

- **Enzymes** enhance the rate of chemical reactions by several orders of magnitude (e.g. arginine decarboxylase, alkaline phosphatase, staphylococcal nuclease **up to 10^{14} fold**)

- the transition rate depends on the activation barrier

$$\Gamma_{reactants \rightarrow products} \propto e^{-G_{barrier}/k_B T}$$

- and enzymes affect this, not the R and P states

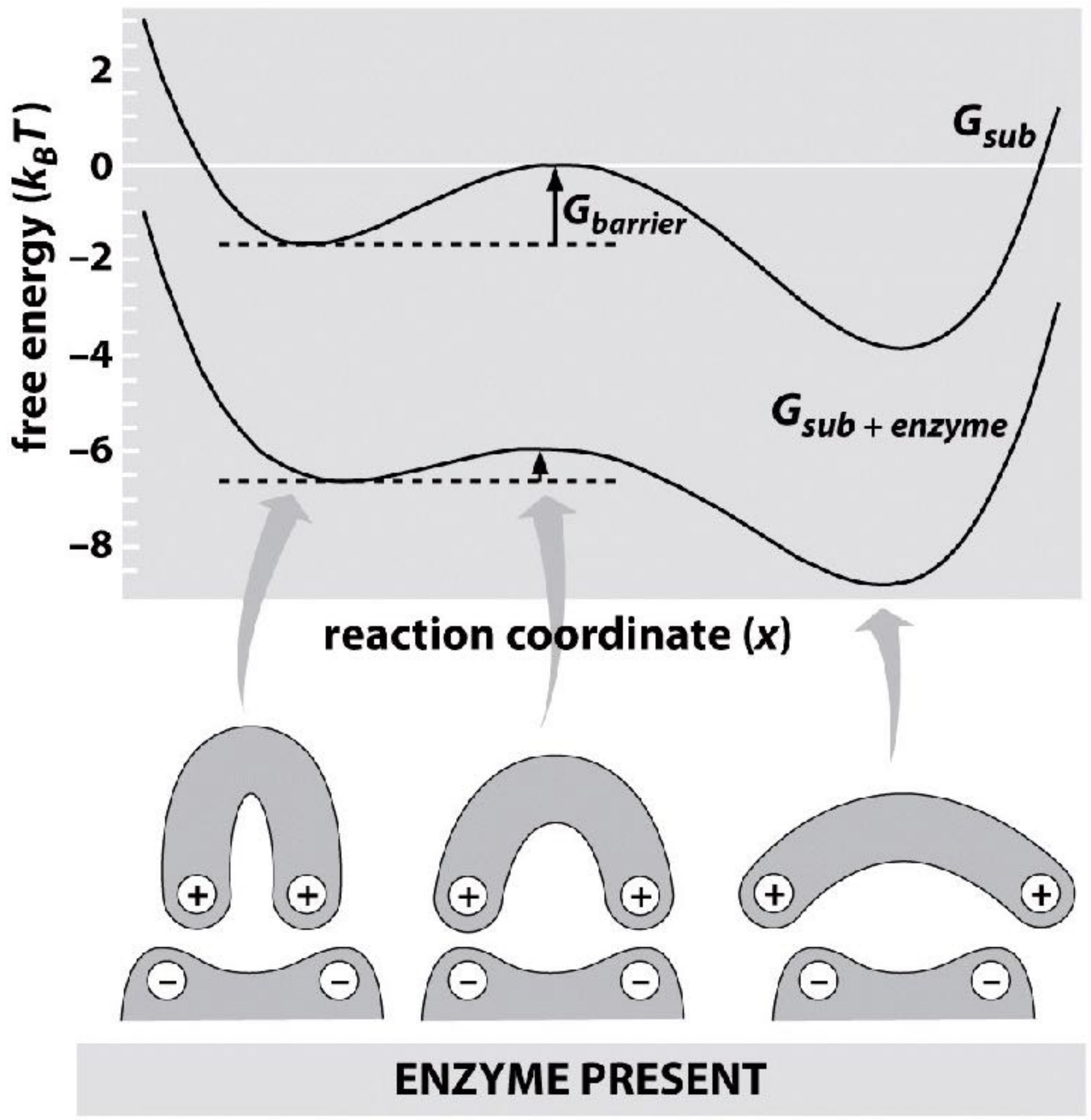
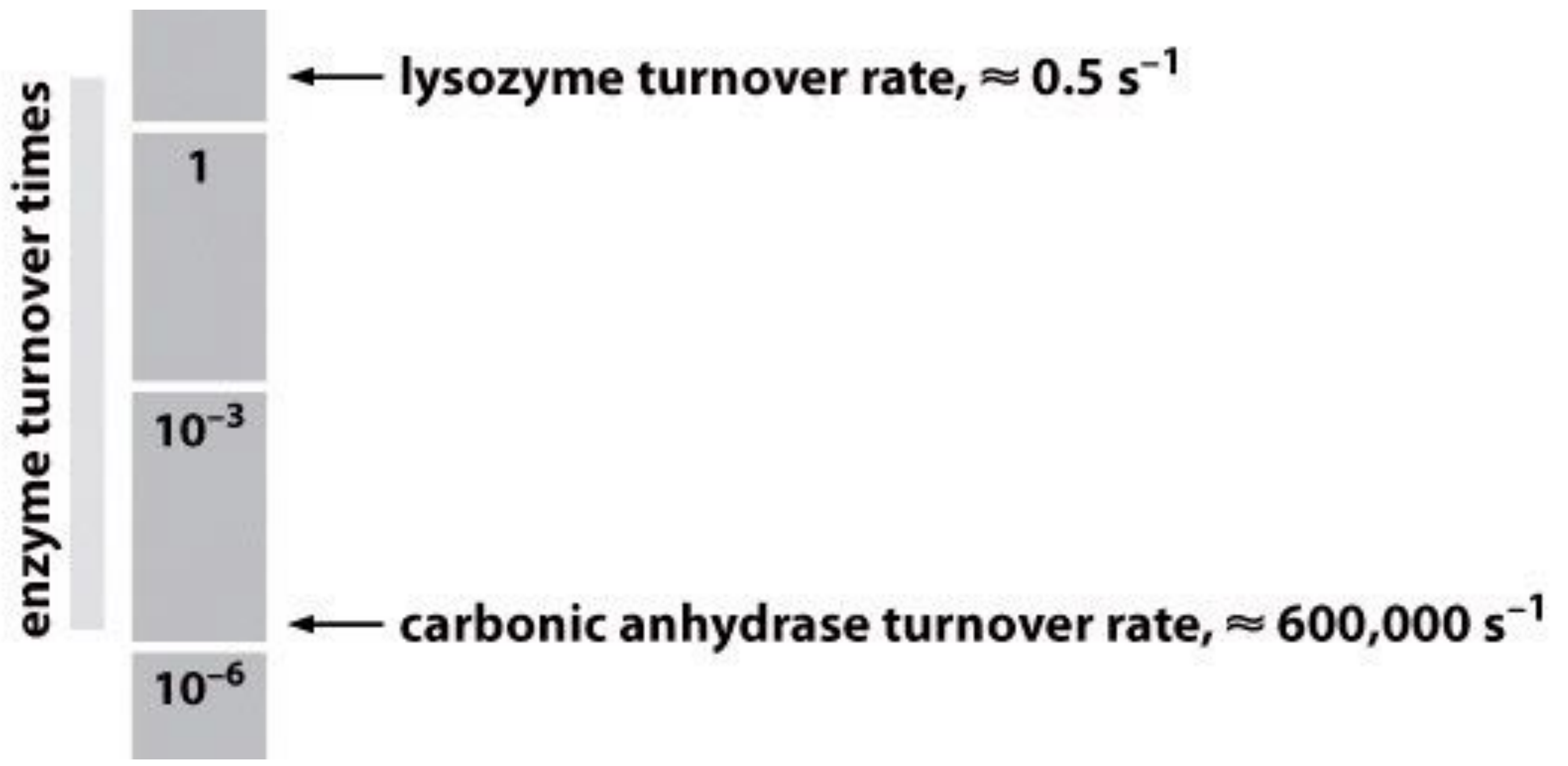
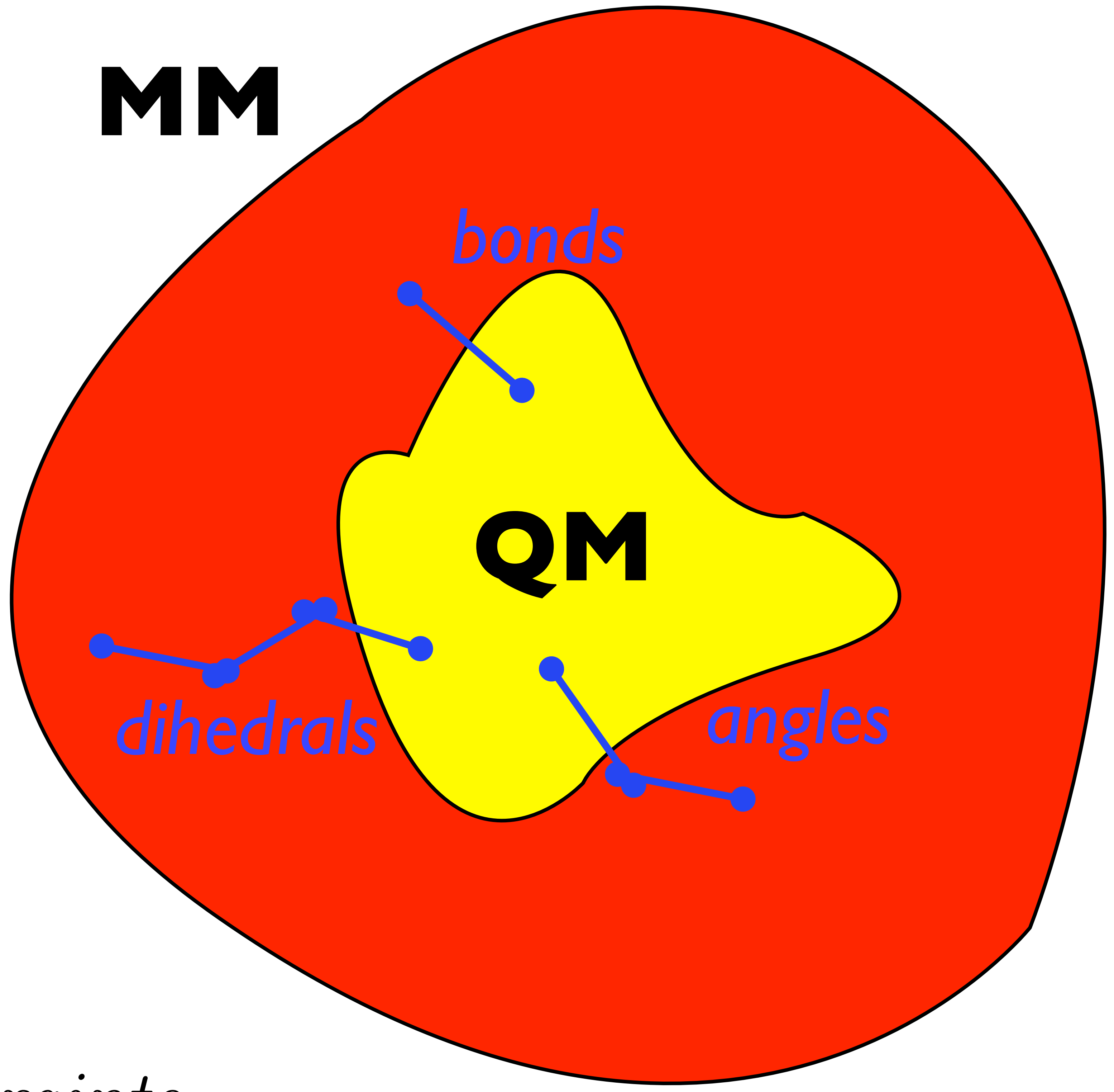


Figure 3.24b Physical Biology of the Cell (© Garland Science 2009)



Hybrid QM/MM molecular dynamics

$$H = H_{QM} + H_{MM} + \underbrace{H_{QM/MM}}_{\textit{coupling term}}$$



QM: First principles Density functional theory MD

$$\mathcal{L}_{\mathrm{CP}} = \underbrace{\sum_I \frac{1}{2} M_I \dot{\mathbf{R}}_I^2 + \sum_i \frac{1}{2} \mu_i \left\langle \dot{\psi}_i \middle| \dot{\psi}_i \right\rangle}_{\text{kinetic energy}} - \underbrace{\langle \Psi_0 | \mathcal{H}_\mathrm{e} | \Psi_0 \rangle}_{\text{potential energy}} + \underbrace{\textit{constraints}}_{\text{orthonormality}}$$

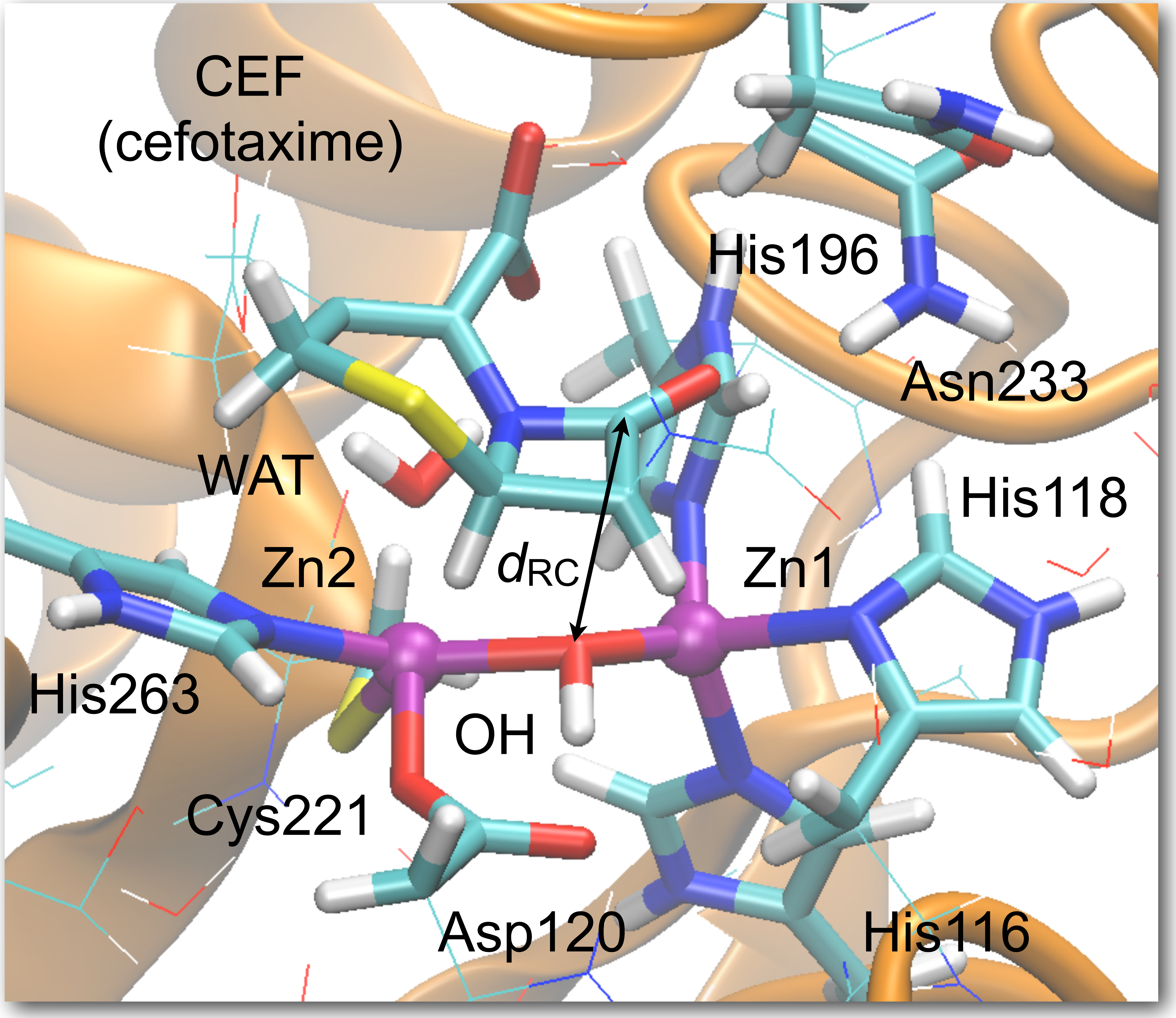
MM: Classical molecular dynamics (*e.g.* AMBER, Gromos force fields)

QM/MM: - boundary atom (*ad hoc* monovalent pseudopotential or H capping)

- hierarchical scheme to compute Coulomb interactions

Car, Parrinello, *PRL* 1985, Laio, Vandevondele, Rothlisberger *JCP* 2004, Dal Peraro *et al.*, *Curr. Opin. Struct. Biol.* 2007

CcrA MβL from *Bacteroides fragilis*



Thermodynamic integration along the reaction coordinate d_{RC}
DFT-BLYP, Martins-Troullier PPs, 70 Ry cutoff,
Nose’ thermostat at 300 K,
2 reactions pathways for a total of ~150 ps trajectory

Reactant state

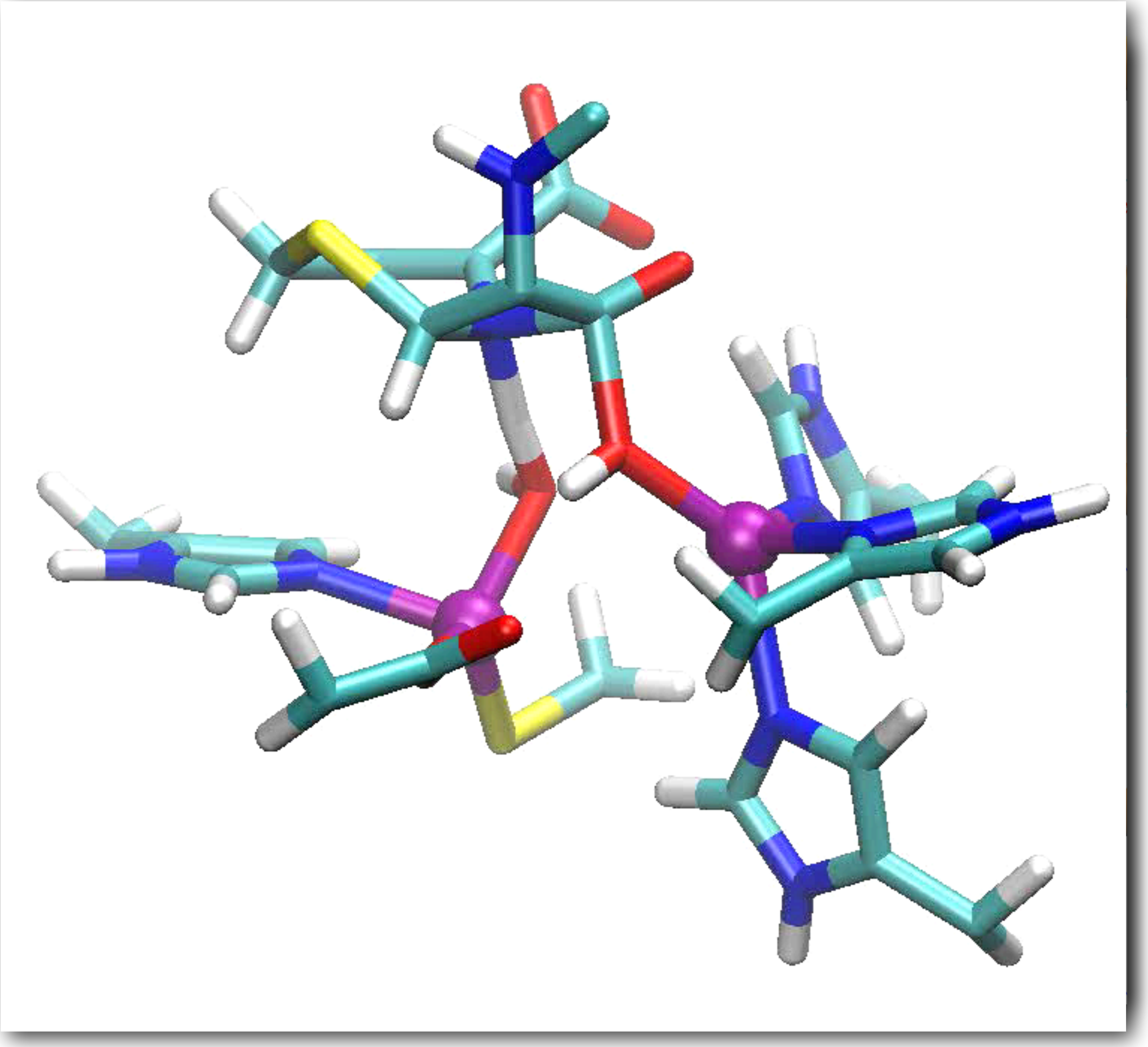
CcrA complexed with cefotaxime

- stable Michaelis complex OH-β-lactam distance=3.3(2)Å during 5 ns MD and 20ps QM/MM

- Zn2-bound WAT is the only water between the zinc center and CEF in 5Å

➡Classical force-field based MD is used as a tool to sample conformational space within the nanosecond timescale

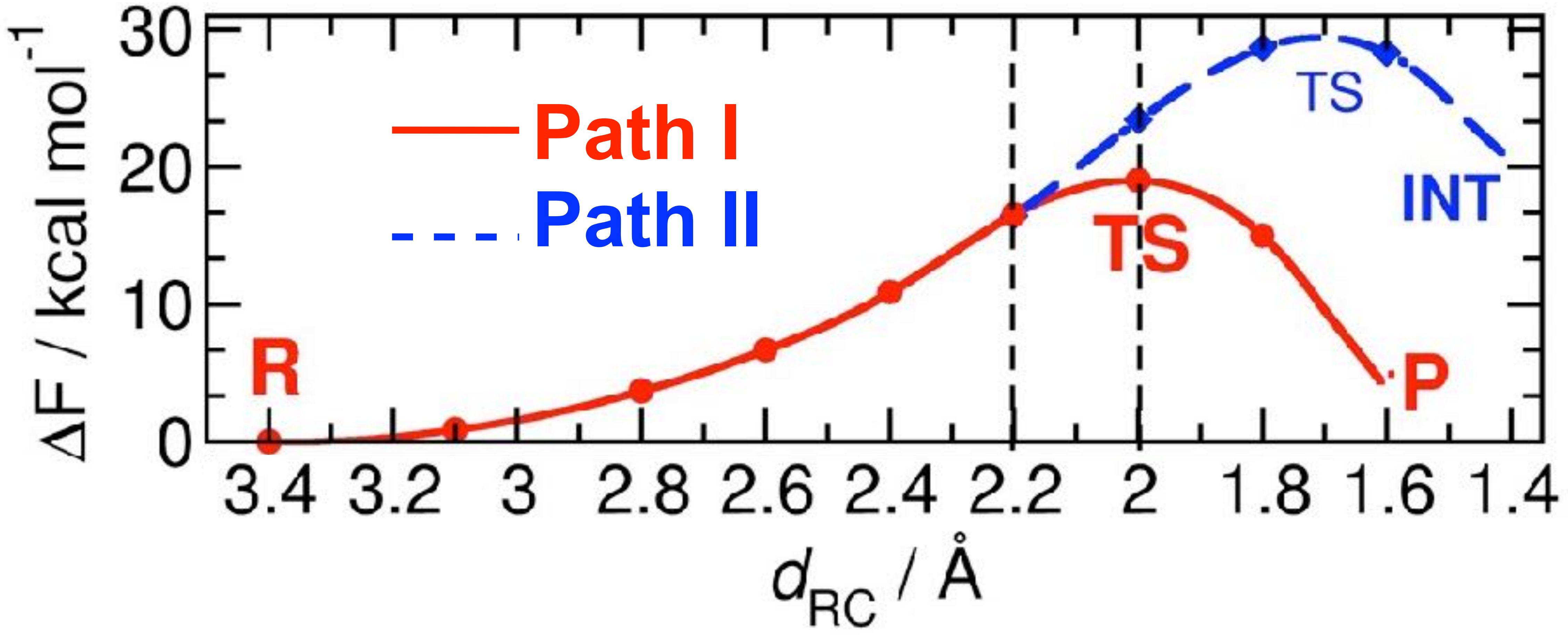
... from transition state to products



water-mediated single-step

- OH^- loses Zn2 coordination
- Zn1, Zn2 flexibility
- WAT protonates β-lactam N
- N-C β-lactam bond breaks
- WAT replaces OH^- as an hydroxide

- **ΔF = 18(2) kcal/mol** is in good agreement with experiments
- if Asn233 *does* H-bond β-lactam: formation of a high unfavorable intermediate (Path II)



Dal Peraro *et al*., *JACS* 2007

chemical detail
sampling
binding properties
mesoscale
coarse-grained
enzymatic
reactions
MM-CG
QM-MM
MD
QM
structural
fluctuations
self-assembly
size/atoms
10^8
10^6
10^4
10^2
fs
ps
ns
μs
ms
s
timescale

Coarse-graining degrees of freedom

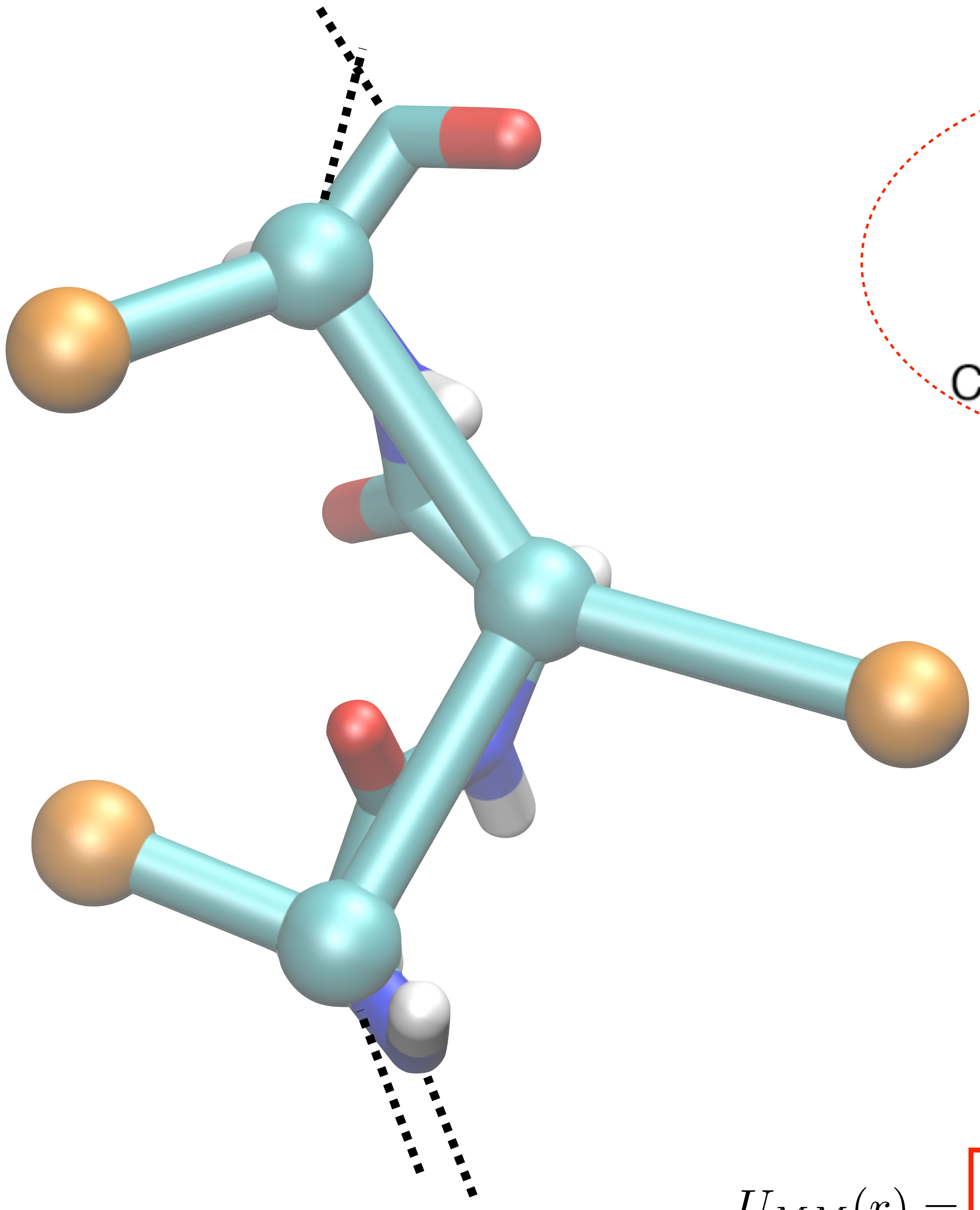
- **CG** is the process of consistently reduce the complexity of your problem integrating out degrees of freedom which can be in principle neglected for your system.

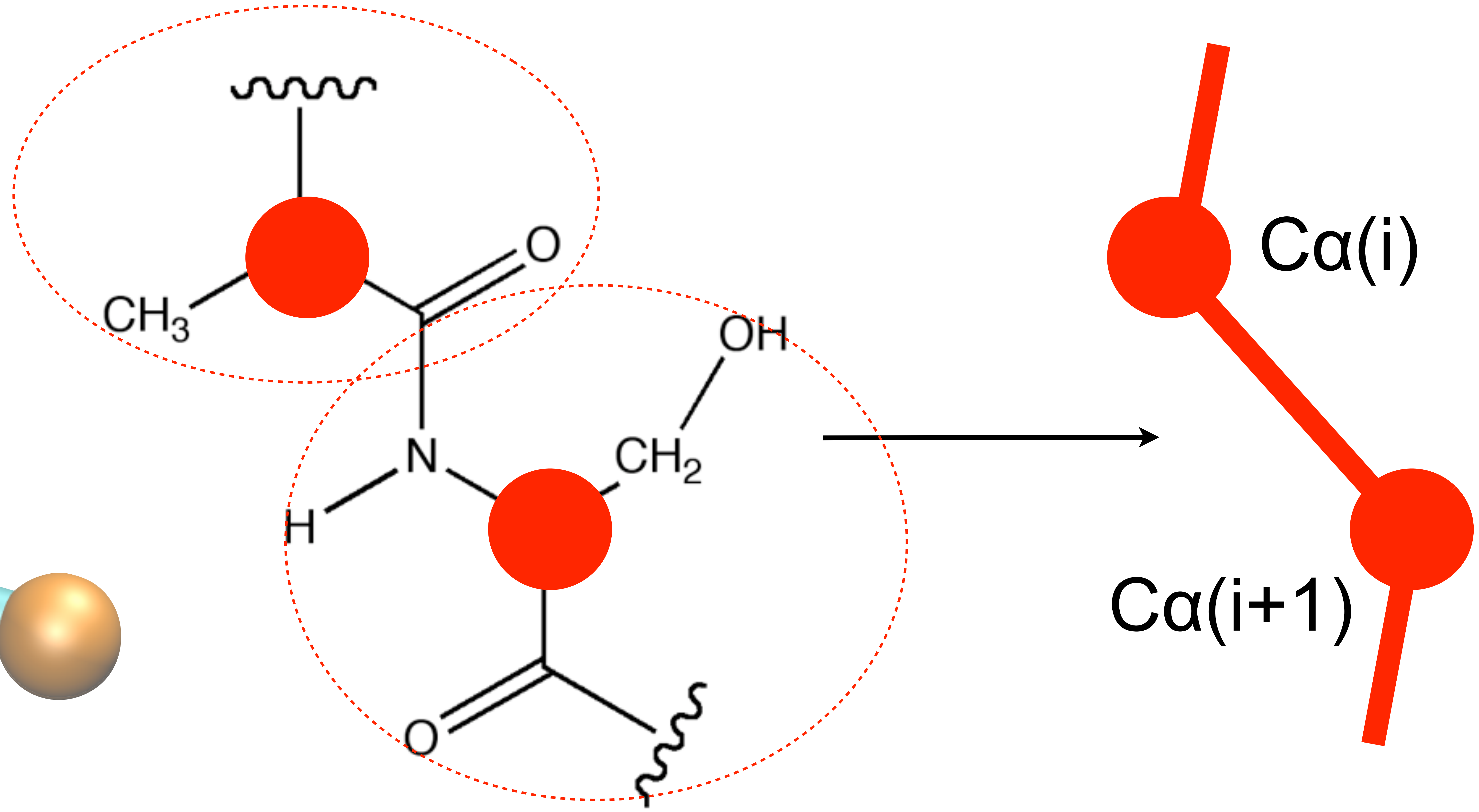
$$V_{QM} \rightarrow V_{MM} \rightarrow V_{CG-MM} \rightarrow V_{mesoscopic}$$

- the CG process implies a **simplification** of your potential that is not always rigorous and includes **approximations**

- what you obtain is an **effective** potentials which is parametrized to reproduce given properties

Coarse-graining degrees of freedom





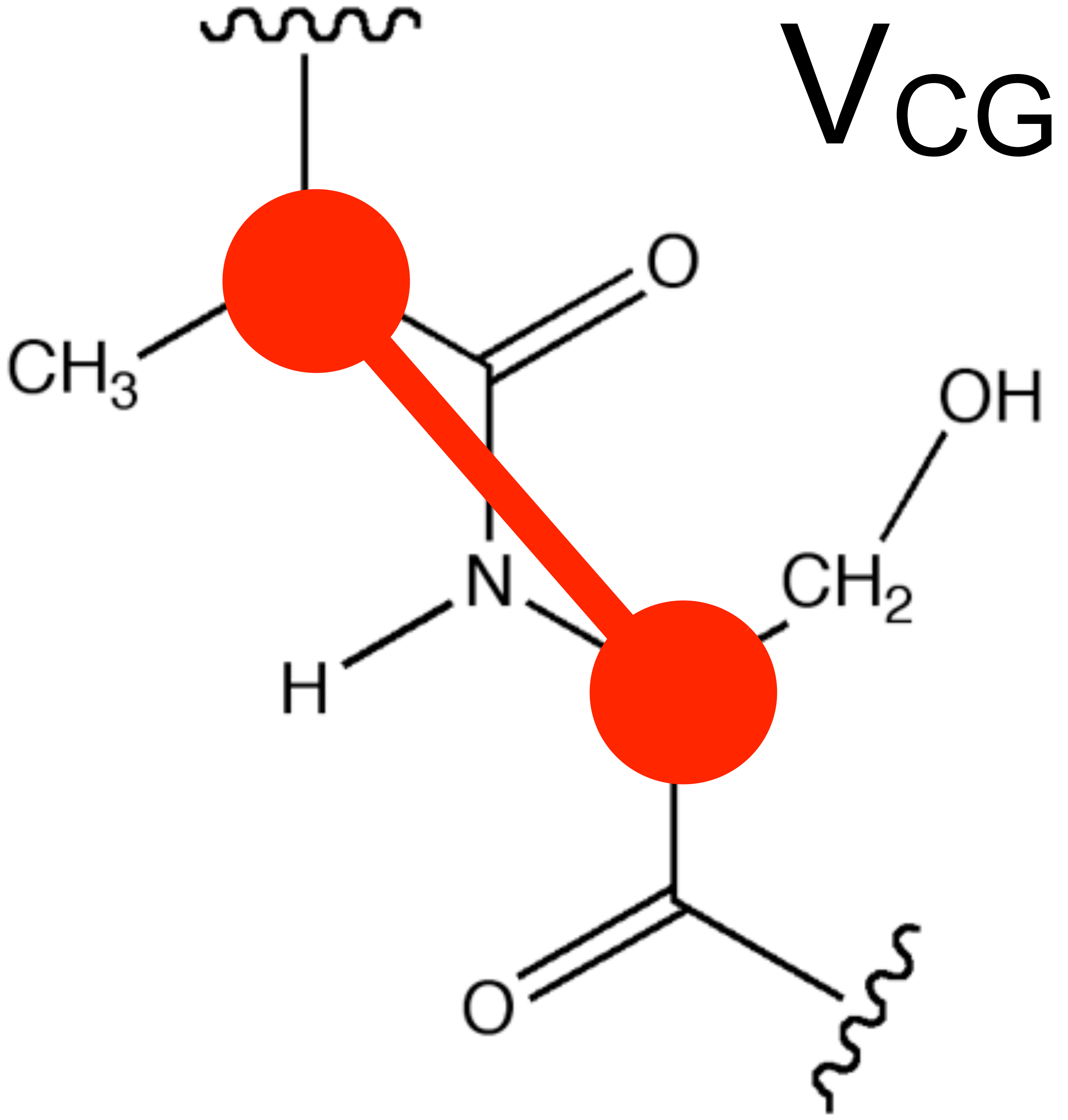
U_{MM}

U_{CG}

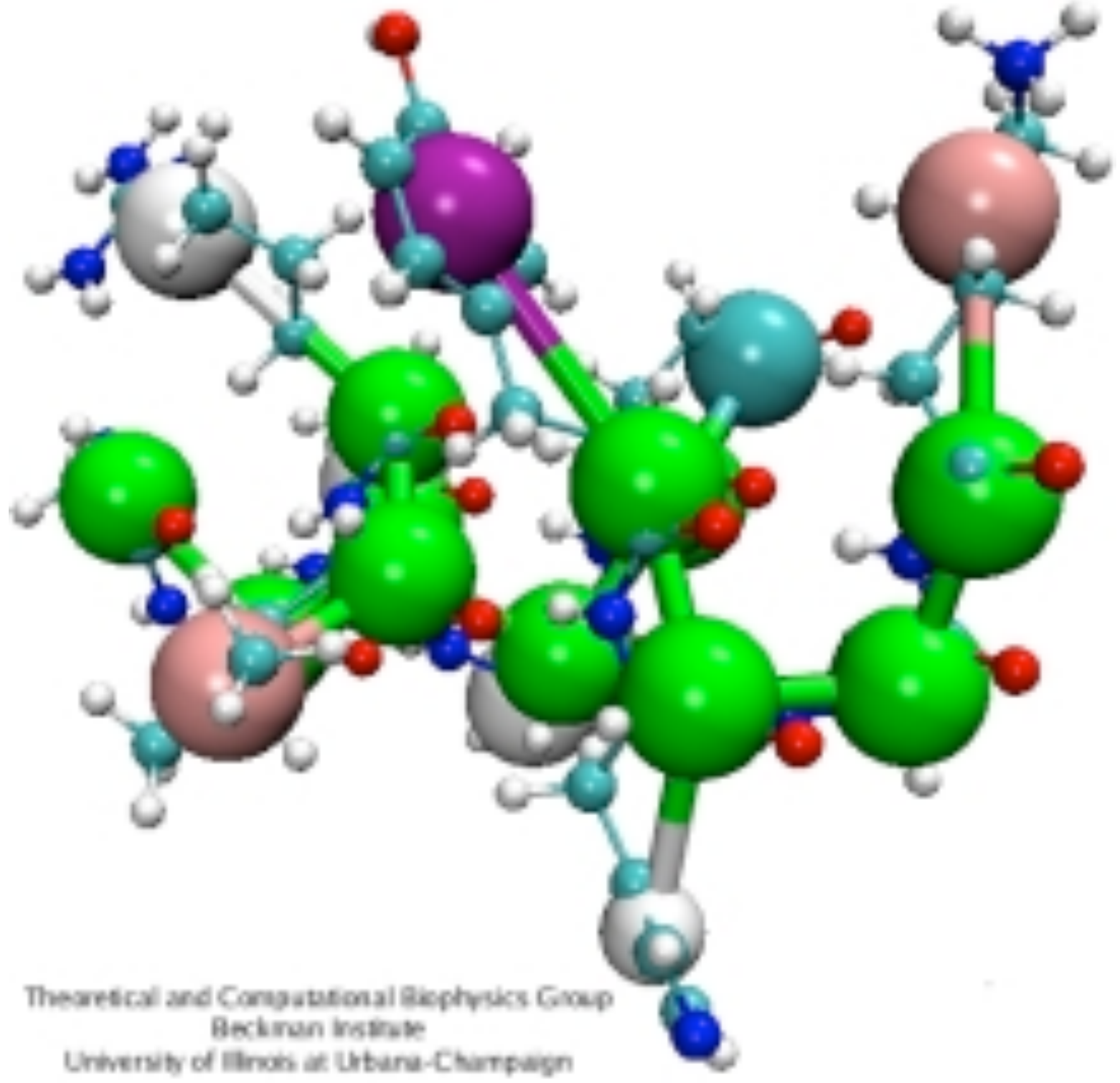
$$U_{MM}(r) = \boxed{\sum_{bonds} \frac{k_b}{2}(r-r_0)^2 + \sum_{angles} \frac{k_\theta}{2}(\theta-\theta_0)^2 + \sum_{torsions,n} \frac{k_{\phi,n}}{2}[1+cos(n\phi-\delta)]} +$$

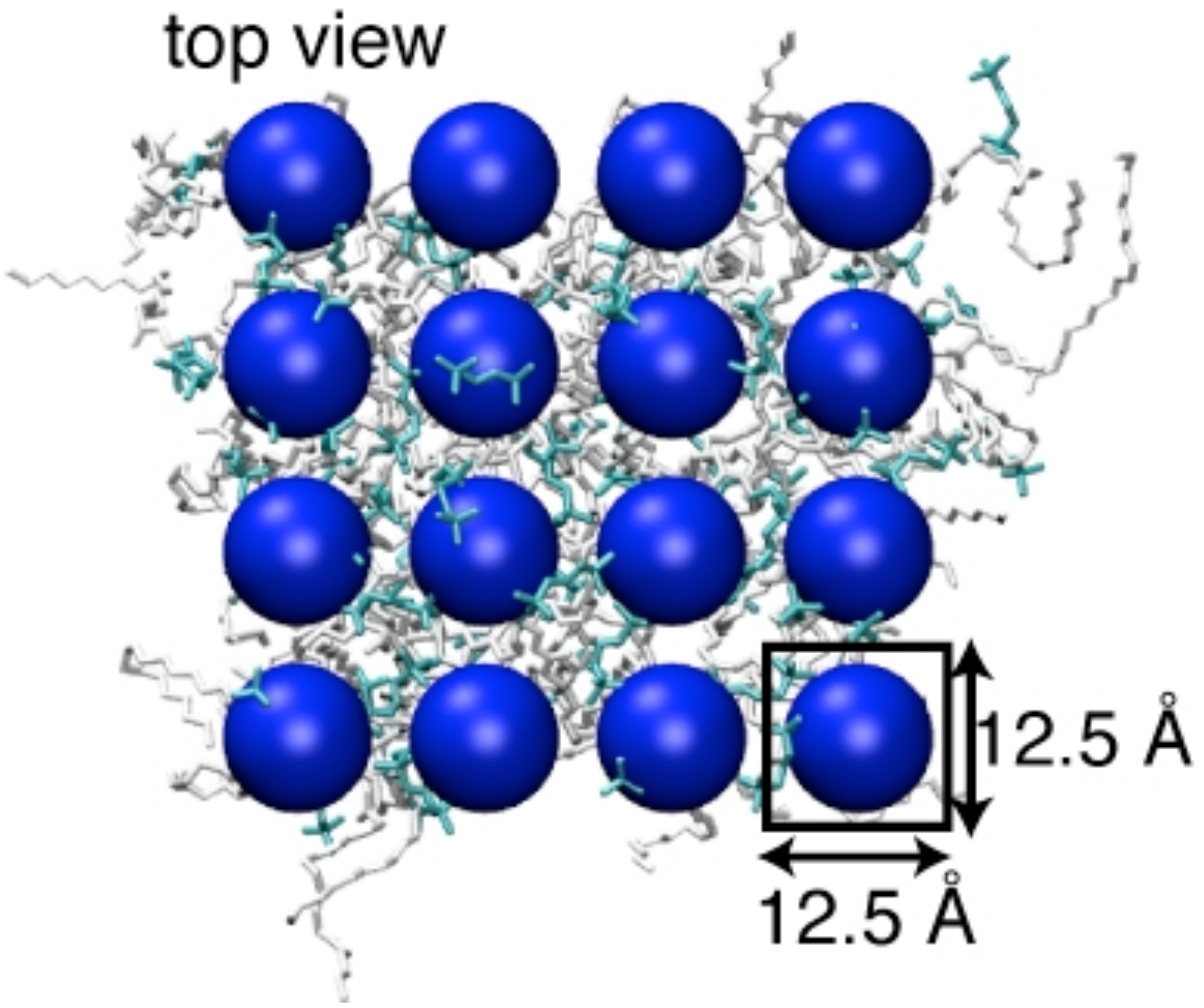
$$+\boxed{\sum_{i>j}^{N}\left(\frac{A}{r_{ij}^{12}} - \frac{C}{r_{ij}^{6}}\right) + \sum_{i>j}^{N} \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{r_{ij}}}$$

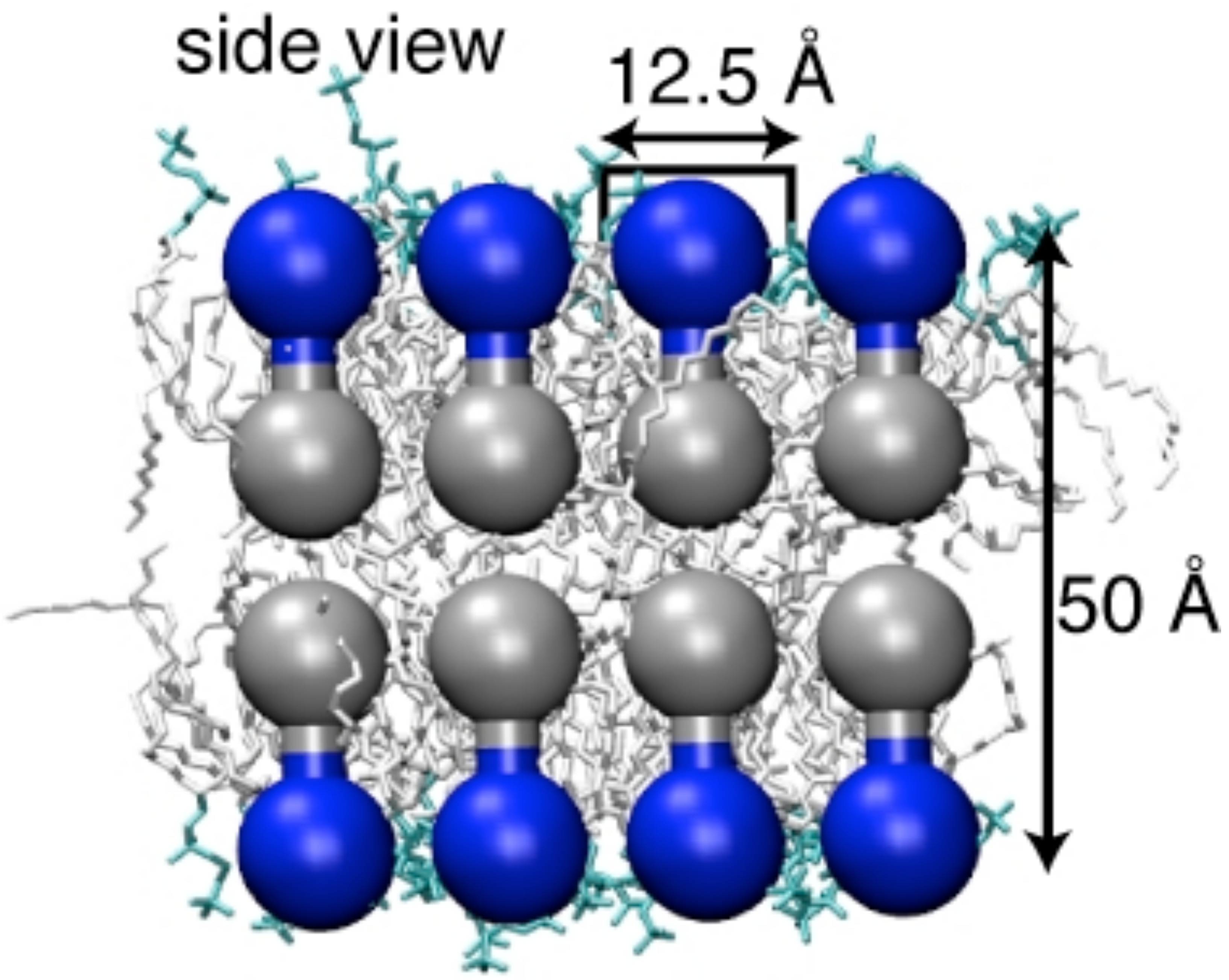
Coarse-grained force fields



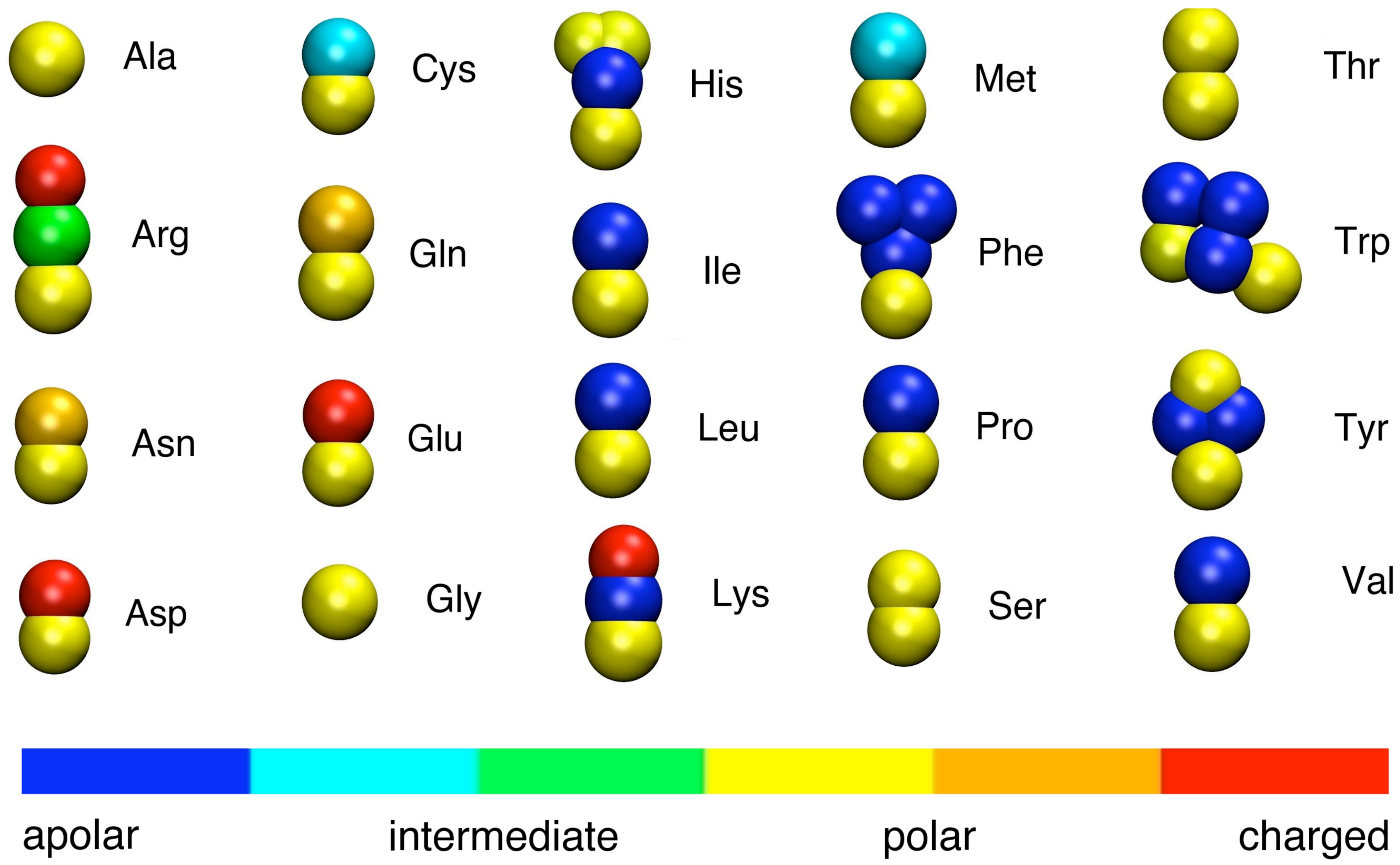
- CG FF models are not topologically biased on the native structure
- softer interactions allow for **longer** timestep in MD simulations
- sampling on the **millisecond** timescale
- accuracy can be a problem (e.g. **no explicit electrostatic** contribution)
- biases on the secondary structures


Theoretical and Computational Biophysics Group
Beckman Institute
University of Illinois at Urbana-Champaign



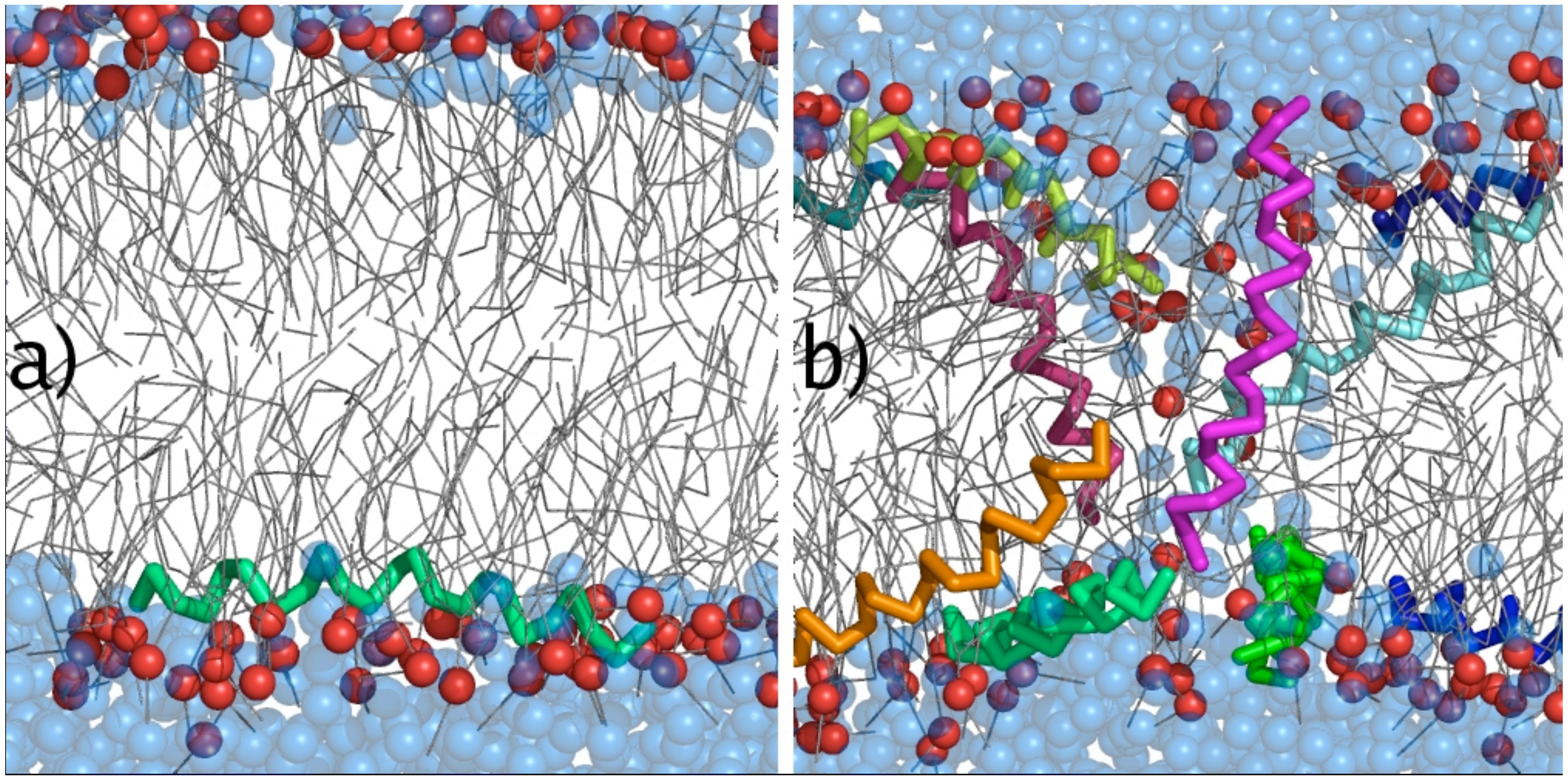


Coarse-grained MARTINI FF

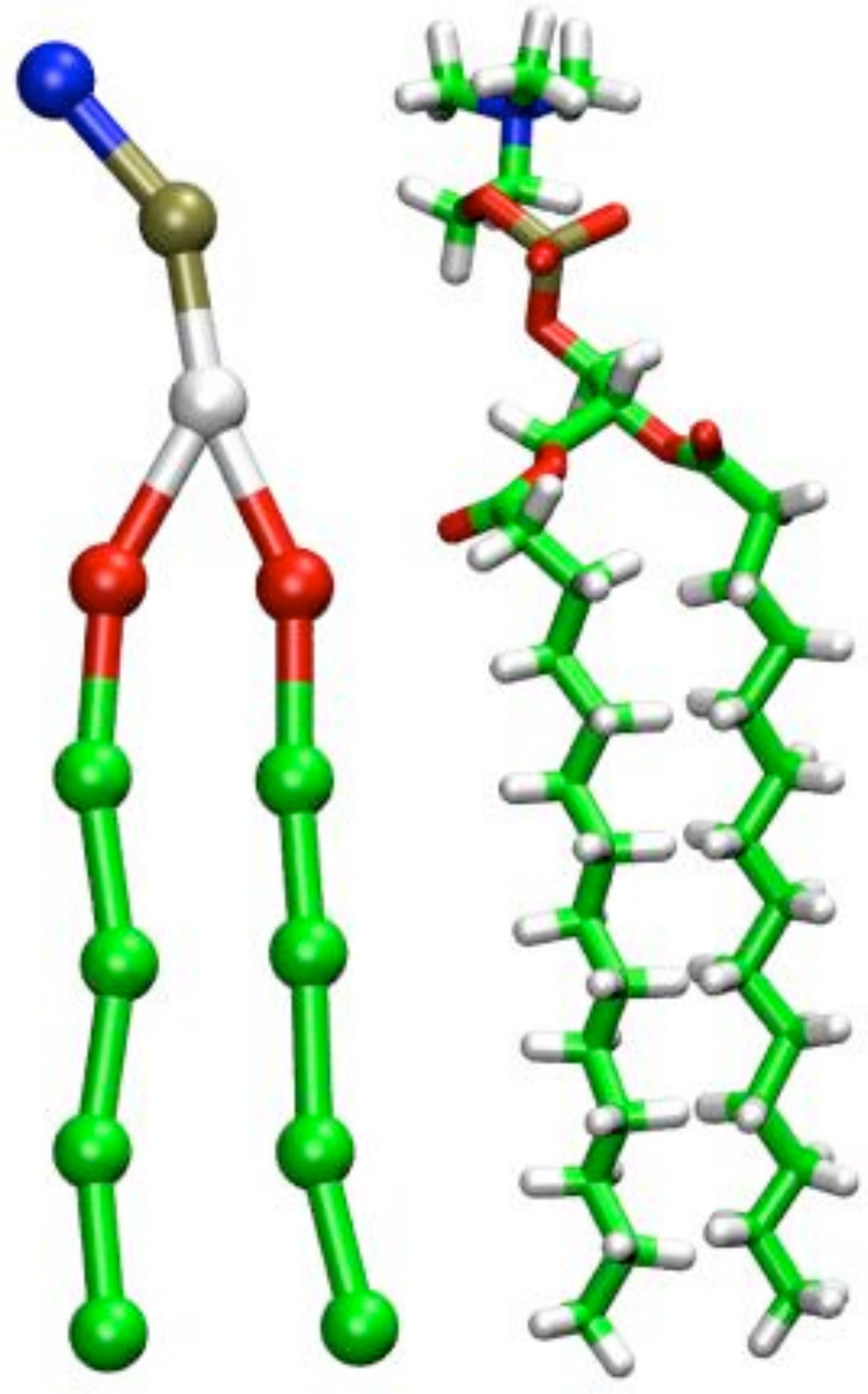


- MARTINI CG FF has functional form similar to MM FF
- 4-to-1 mapping from MM to CG
- very convenient for membranes and peptide-membrane interactions

Monticelli et al, JCTC 2008
Klein and coworkers



Magainin H2 in a DPPC bilayer, at low concentration (a) and high concentration



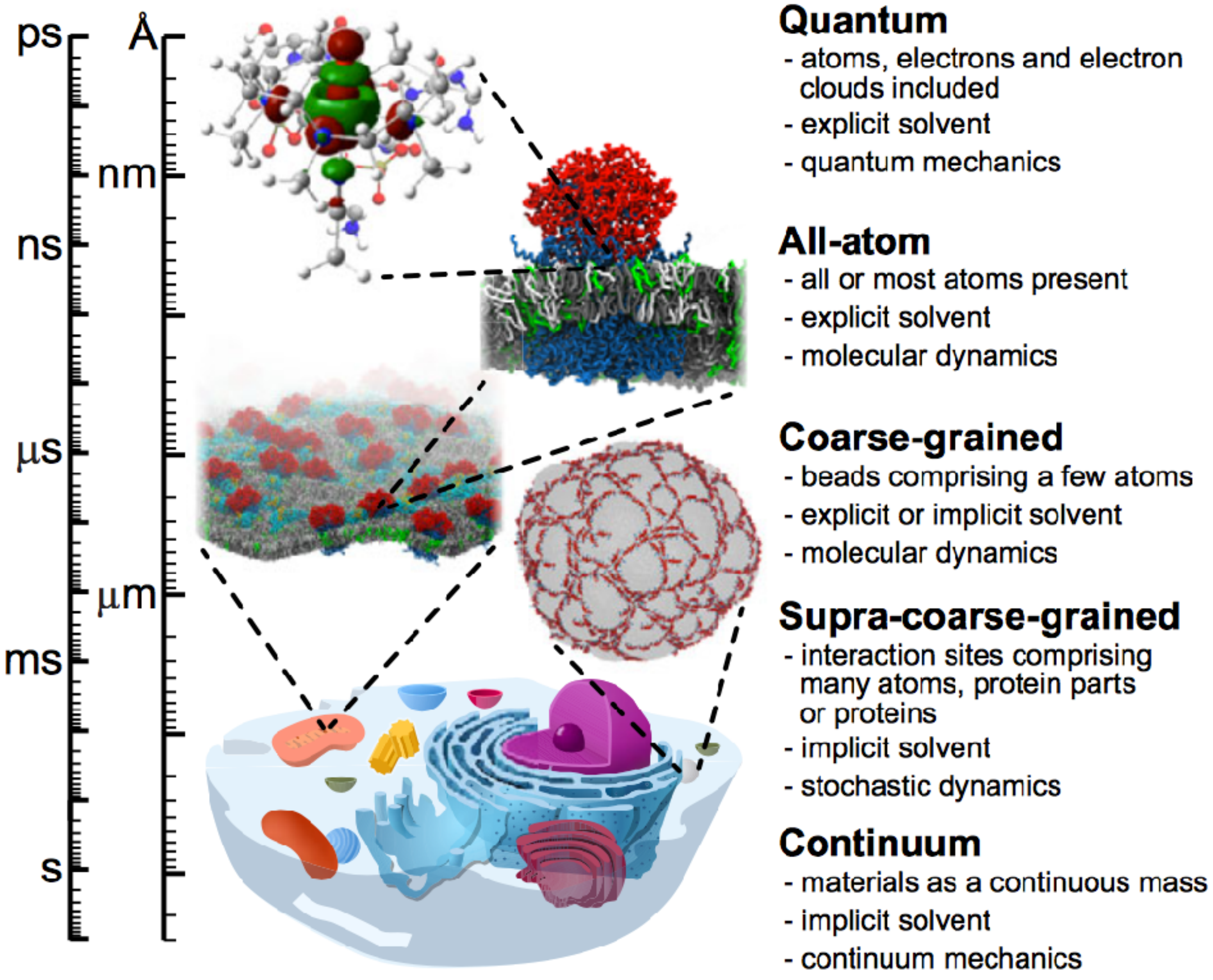
© 2016. Published by The Company of Biologists Ltd | Journal of Cell Science (2016) 129, 257-268 doi:10.1242/jcs.176040

COMMENTARY

ARTICLE SERIES: IMAGING

Computational ‘microscopy’ of cellular membranes

Helgi I. Ingólfsson, Clément Arnarez, Xavier Periole and Siewert J. Marrink*



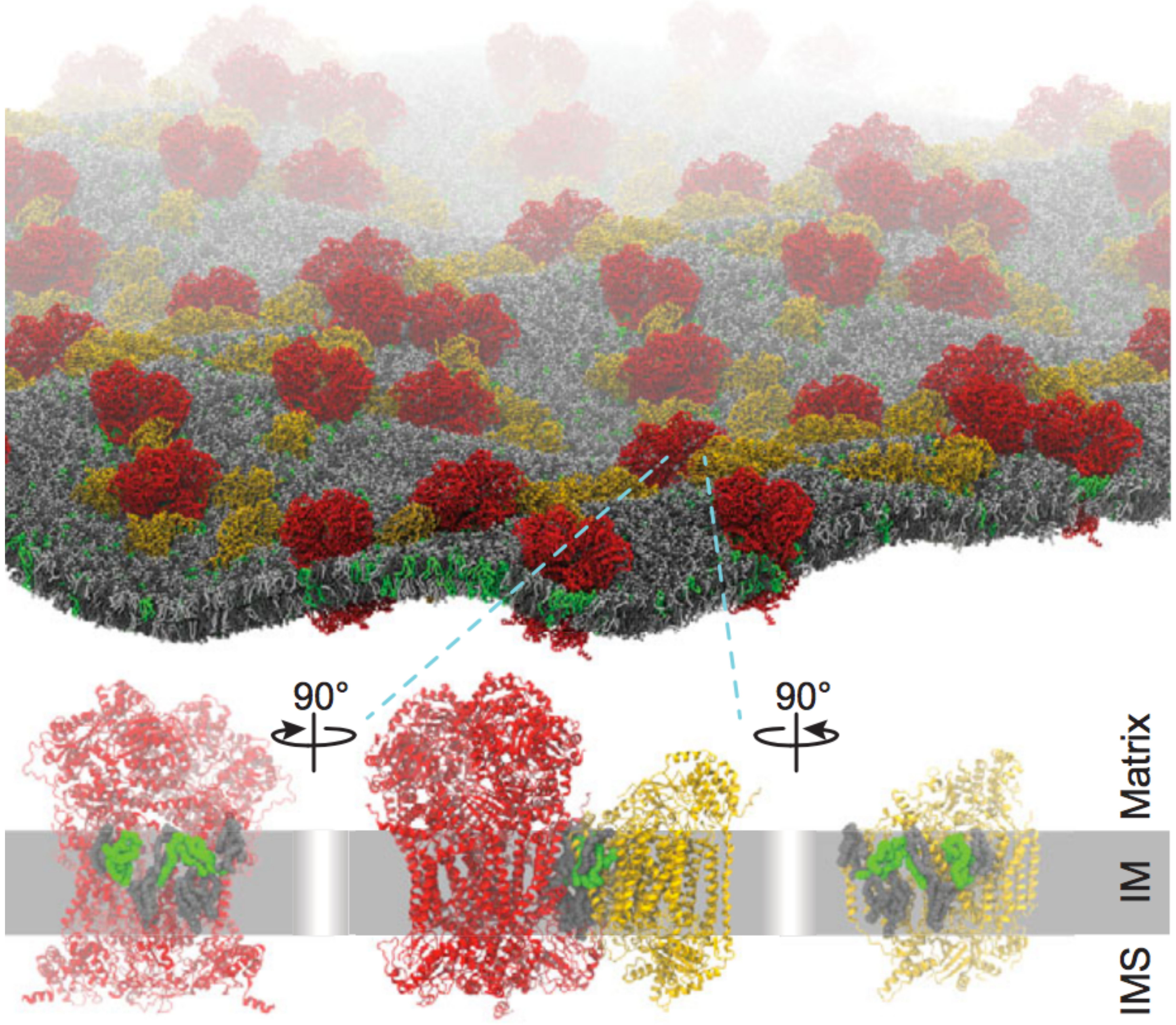
Quantum
- atoms, electrons and electron clouds included
- explicit solvent
- quantum mechanics

All-atom
- all or most atoms present
- explicit solvent
- molecular dynamics

Coarse-grained
- beads comprising a few atoms
- explicit or implicit solvent
- molecular dynamics

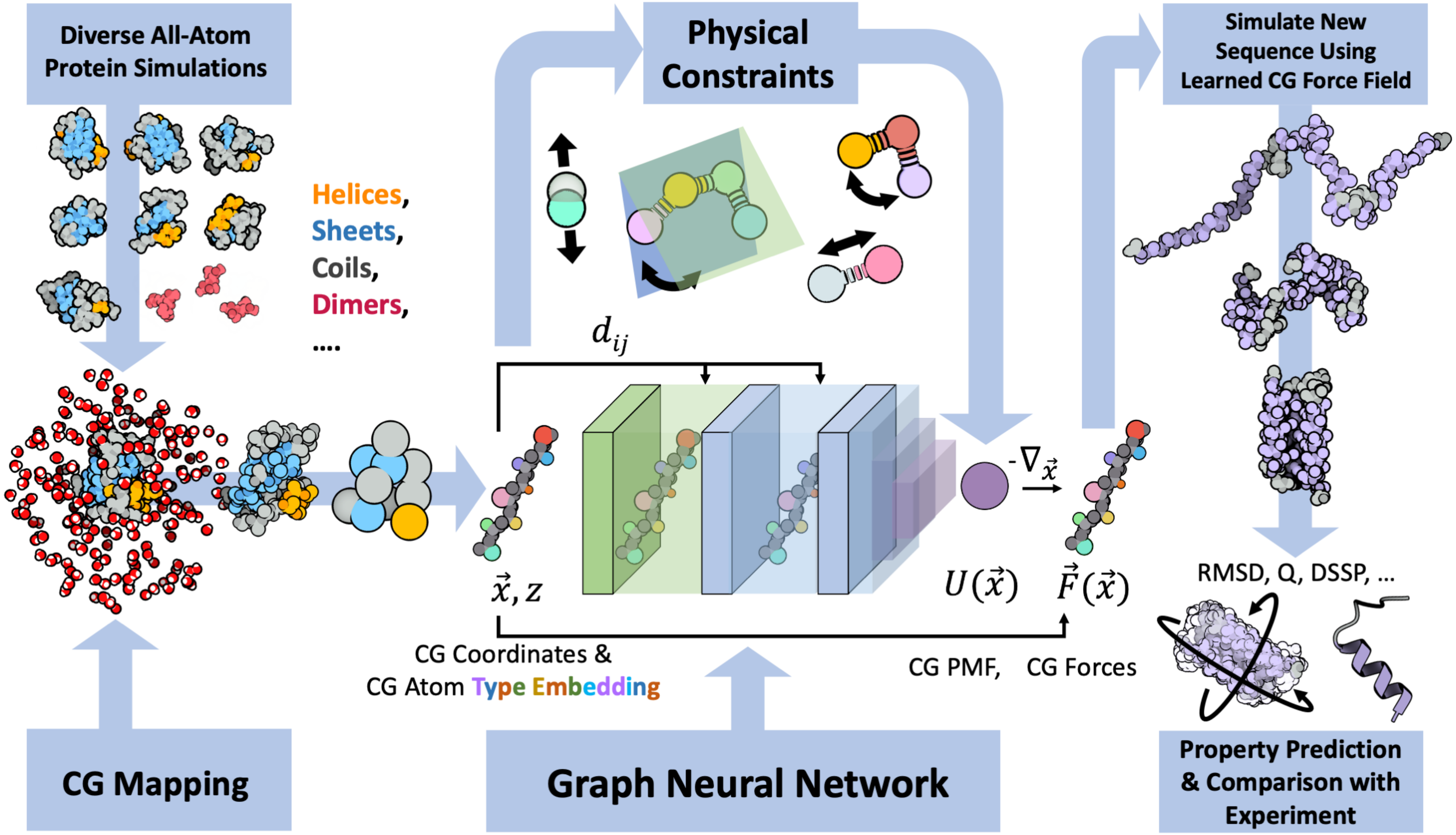
Supra-coarse-grained
- interaction sites comprising many atoms, protein parts or proteins
- implicit solvent
- stochastic dynamics

Continuum
- materials as a continuous mass
- implicit solvent
- continuum mechanics



New directions

universal and computationally efficient machine-learned CG model for proteins



Diverse All-Atom Protein Simulations

Physical Constraints

Simulate New Sequence Using Learned CG Force Field

CG Mapping

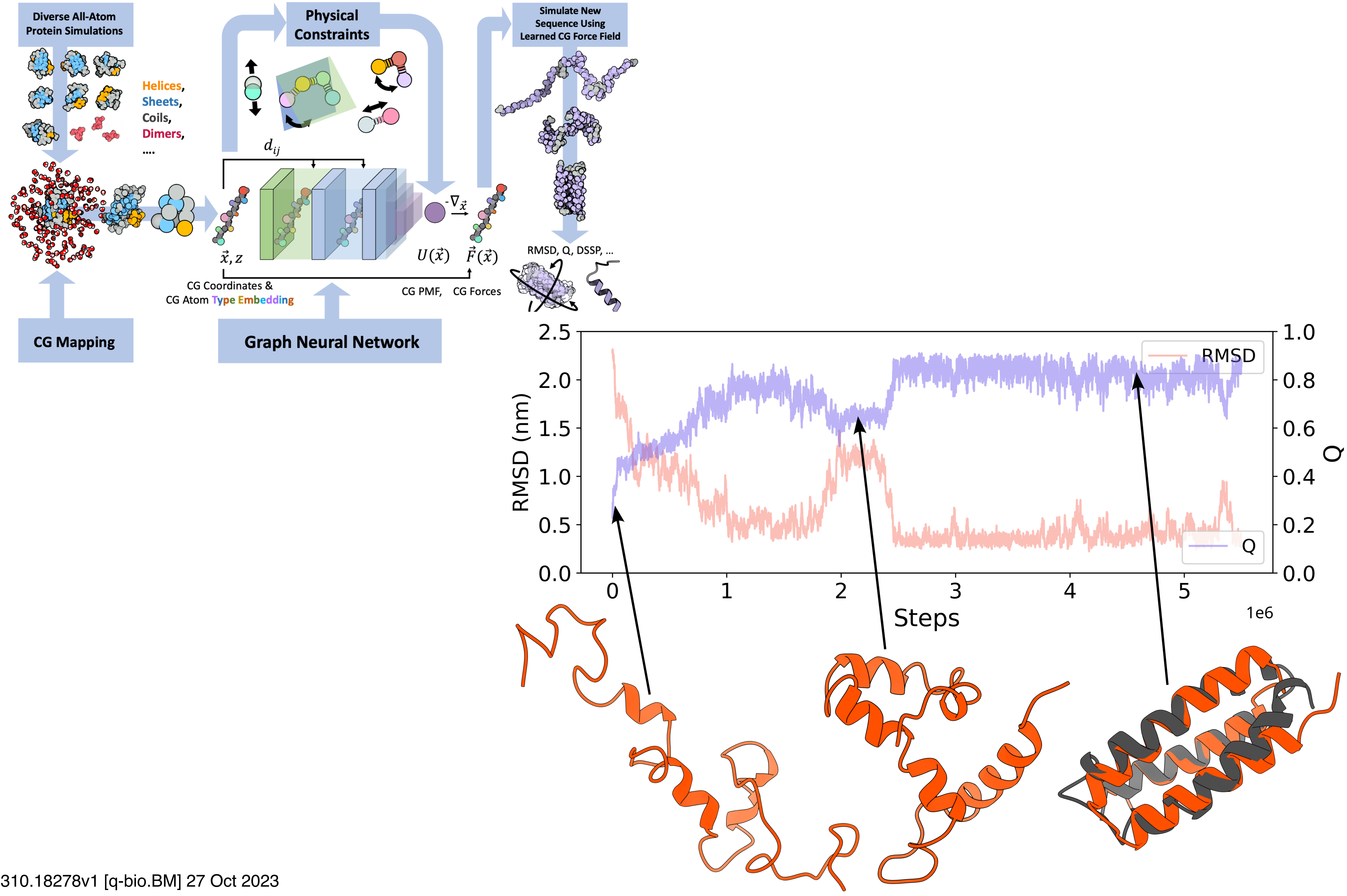
Graph Neural Network

Property Prediction & Comparison with Experiment

arXiv:2310.18278v1 [q-bio.BM] 27 Oct 2023

New directions

universal and computationally efficient machine-learned CG model for proteins



arXiv:2310.18278v1 [q-bio.BM] 27 Oct 2023