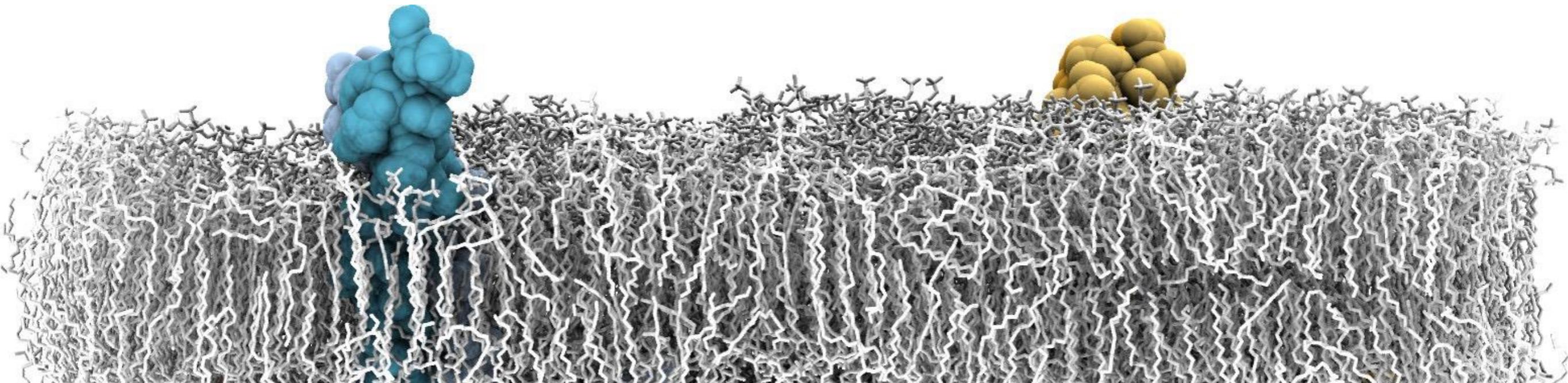


INTRODUCTION TO MOLECULAR DYNAMICS

Afroditi Maria Zaki

(slides based and adapted from a presentation by Matteo Aldeghi and Marc Dämgen)

*Wellcome Trust postgraduate course
16th Dec 2019*



OUTLINE

- **Introduction:** modelling reality
- **Molecular mechanics**
 - Assumptions and approximations
 - Force fields
 - Bonded terms
 - Non-bonded terms
- **Molecular dynamics**
 - Fundamentals:
 - $F=ma$
 - Finite difference method
 - Statistical mechanics: ensembles and ergodicity
 - MD practicalities:
 - Boundary conditions
 - Treatment of non-bonded energies: cutoffs, neighbour lists, and Ewald sum
 - Constraints
 - Typical workflow:
 - System preparation
 - Energy minimisation
 - Analysis

WHAT IS OUR GOAL?

- Description of the behaviour of molecules at the atomic level;
- Understand the properties of assemblies of molecules in terms of structure and interactions between them;
- Ultimately relate structure to dynamics to function.

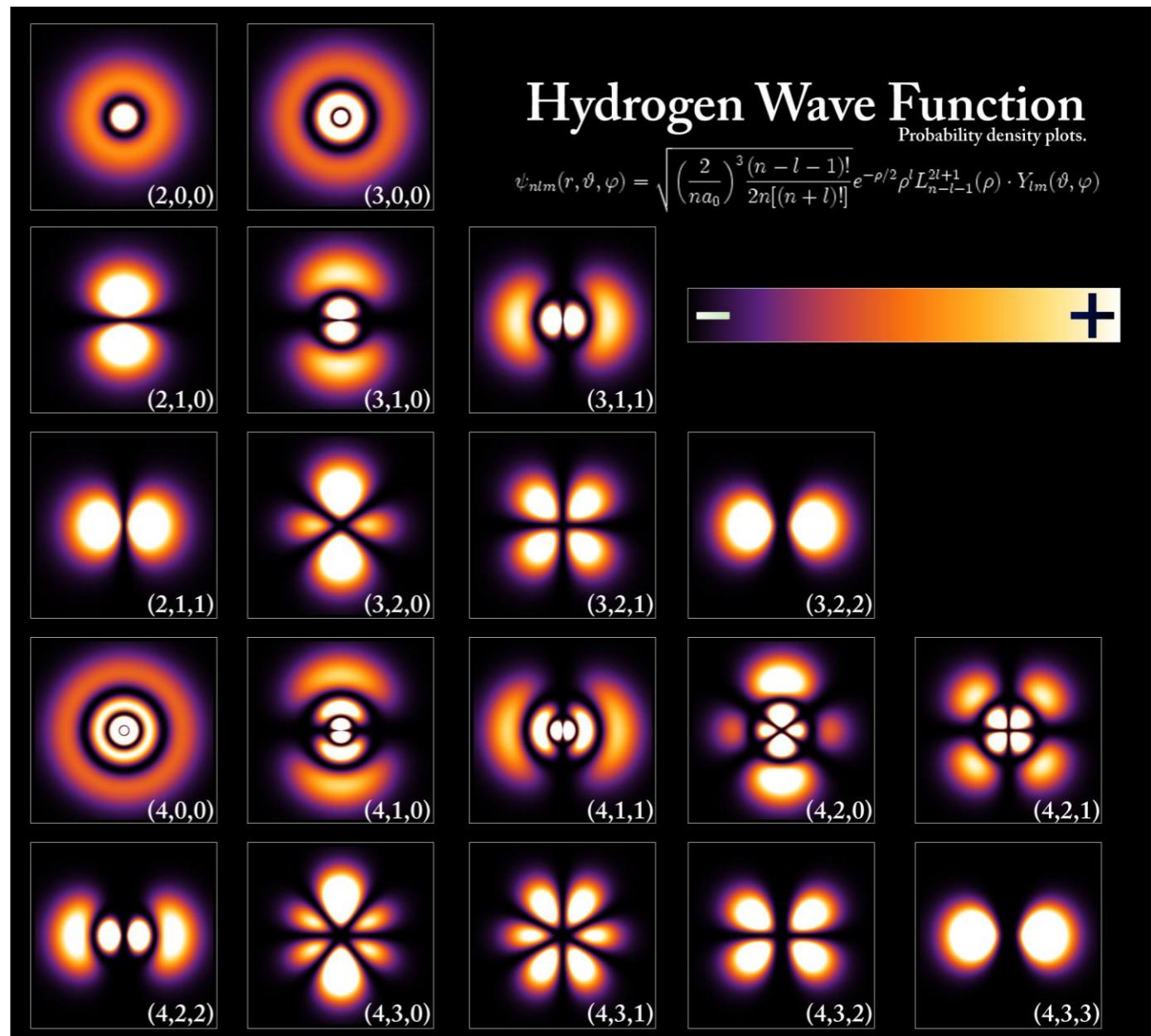
“...everything that living things do can be understood in terms of the jiggling and wiggling of atoms”

—R. Feynman

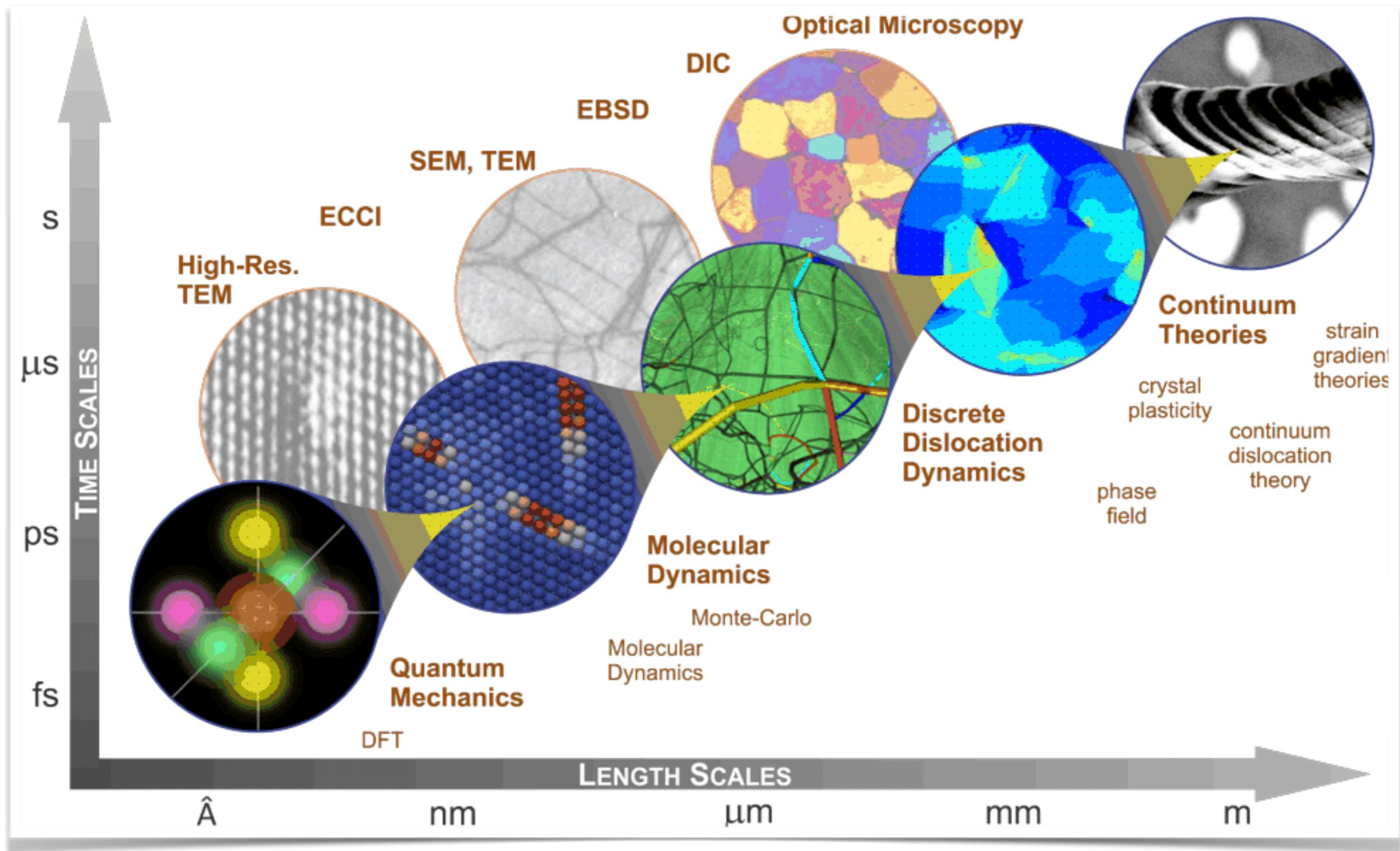
QUANTUM MECHANICS

$$i\hbar \frac{\partial}{\partial t} \Psi = H\Psi$$

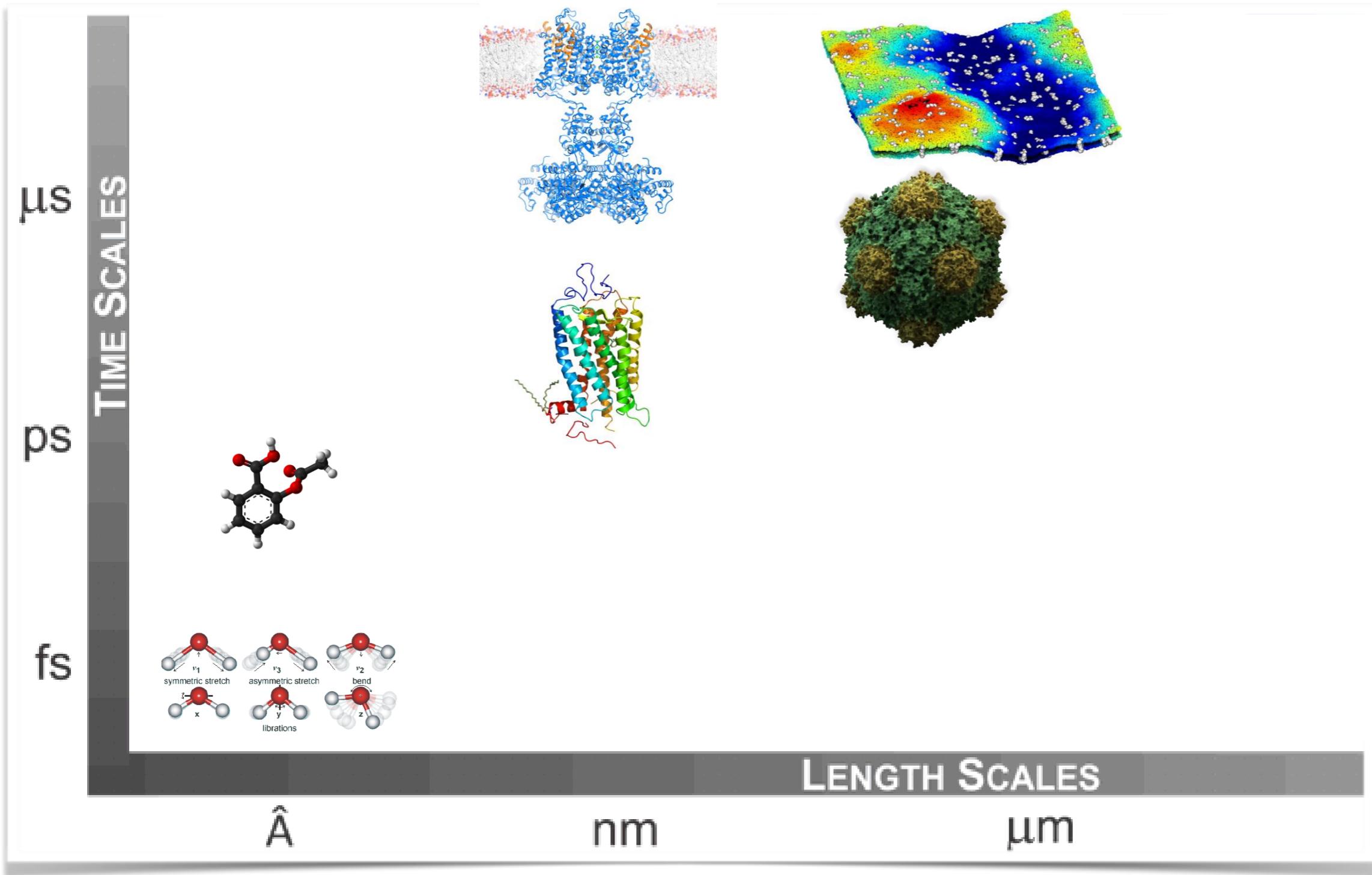
- Ideally one would use QM to study any molecular system;
- However, many problems and systems are too large to be studied by QM.



LIMITATION OF SCALES

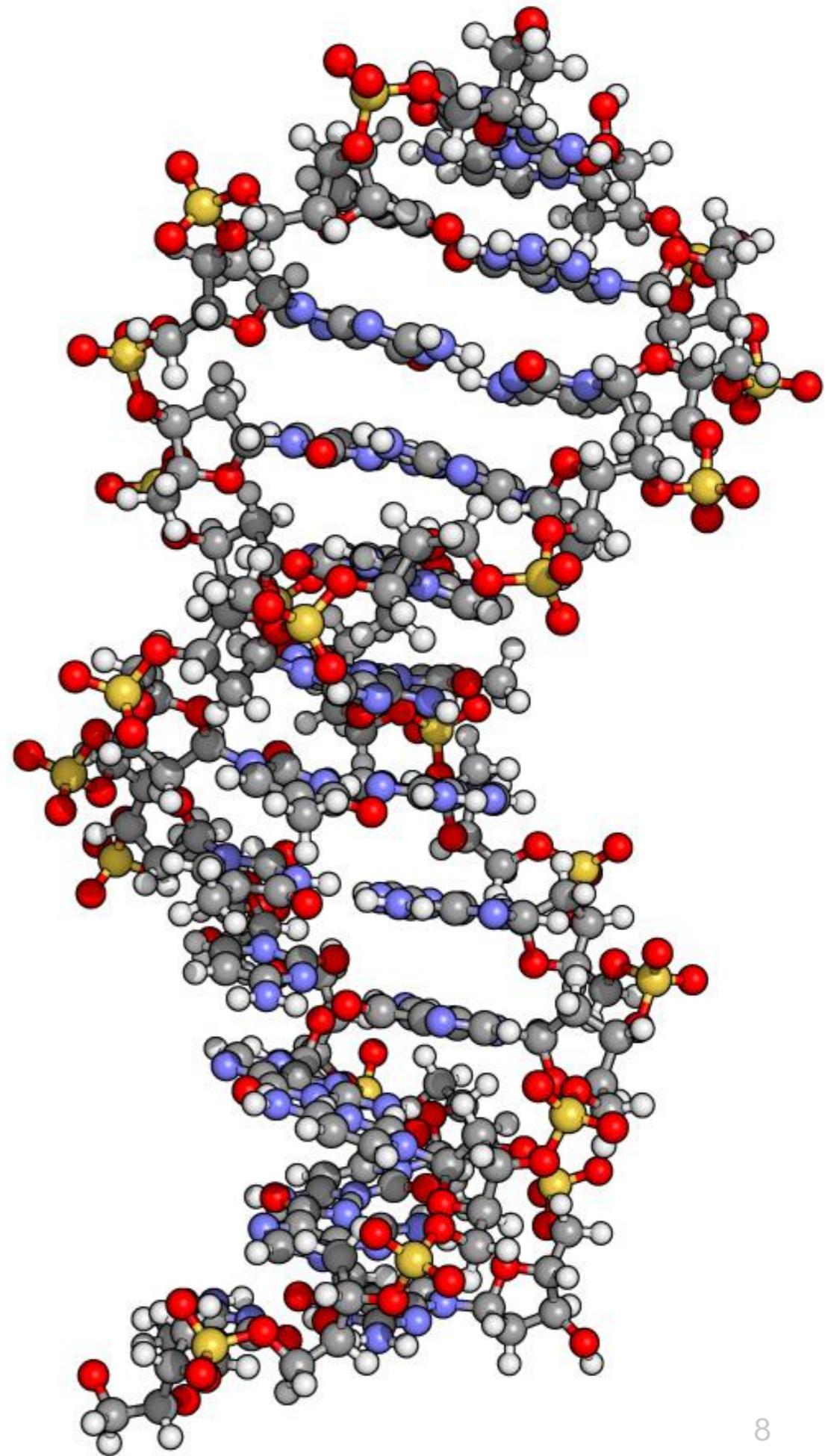


SCALES IN BIOLOGY



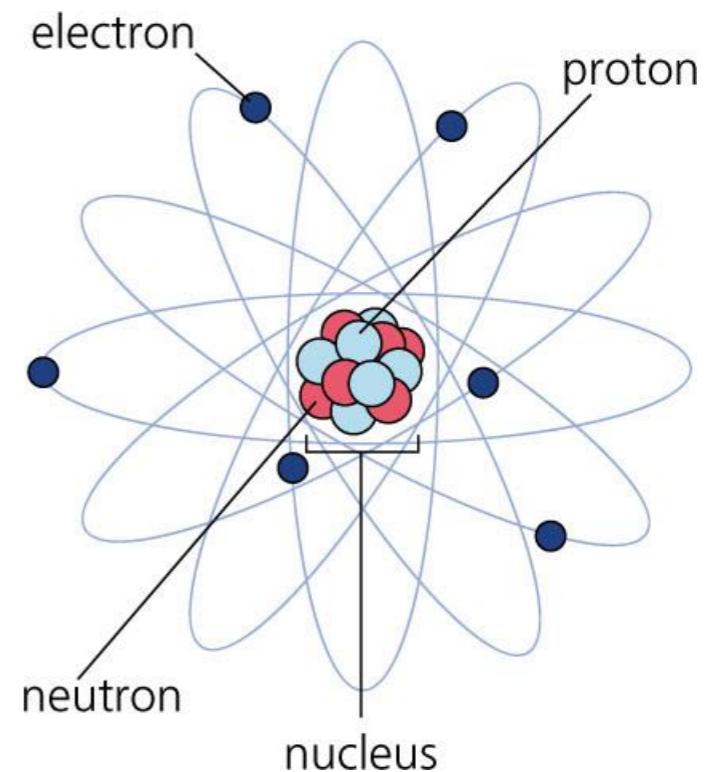
MOLECULAR MECHANICS

Empirical force field models



ASSUMPTIONS AND APPROXIMATIONS

- **Classical mechanics approximation:** we disregard the dynamics of electrons and each atom is represented as a single point particle;
- Atomic interactions can be described by **simple functions**, such as Hooke's law;
- **Transferability** of the parameters across similar chemical groups.



$$U(x) = \frac{1}{2}kx^2$$

FORCE FIELD

A set of energy functions and parameters that defines the potential energy of the system.

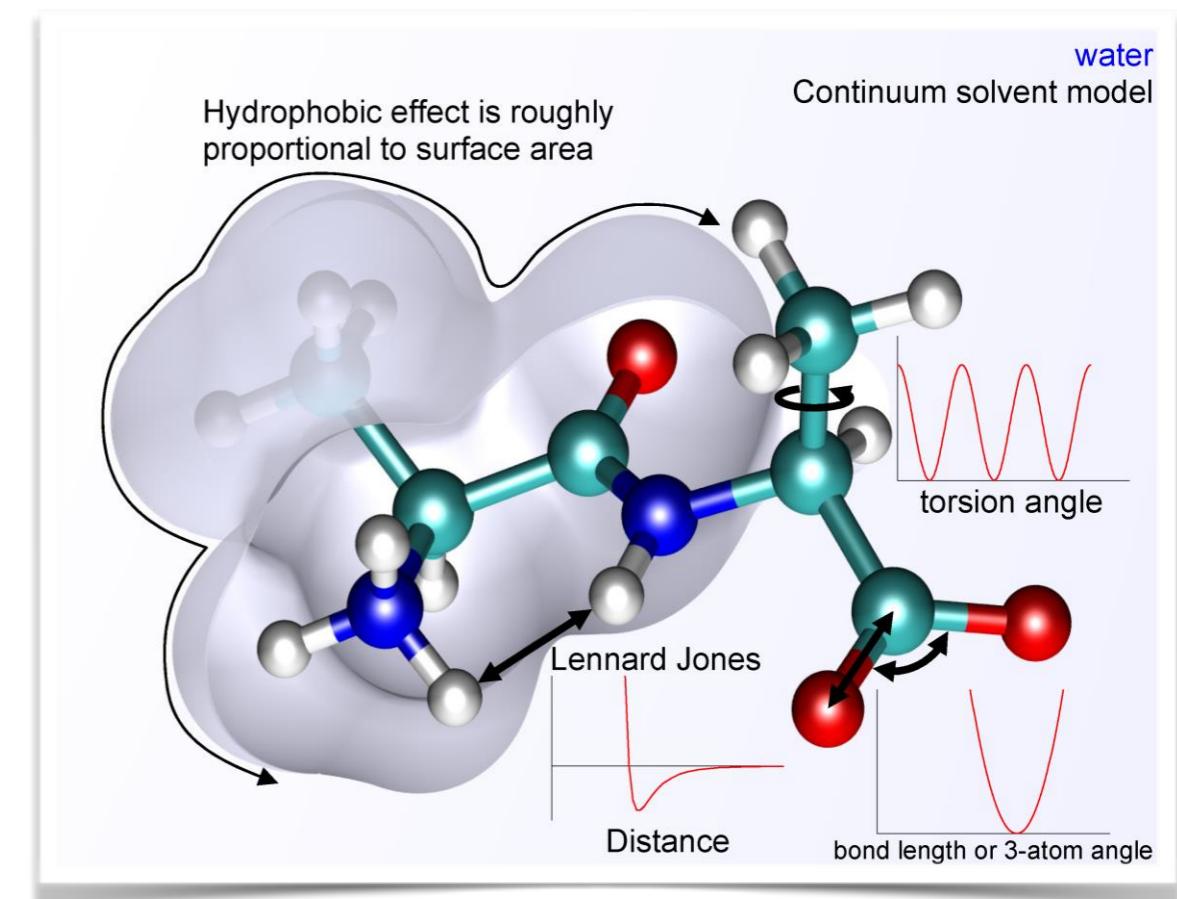
$$E_{\text{total}} = E_{\text{bonded}} + E_{\text{non-bonded}}$$

Bonded terms:

- Bonds
- Angles
- Dihedrals

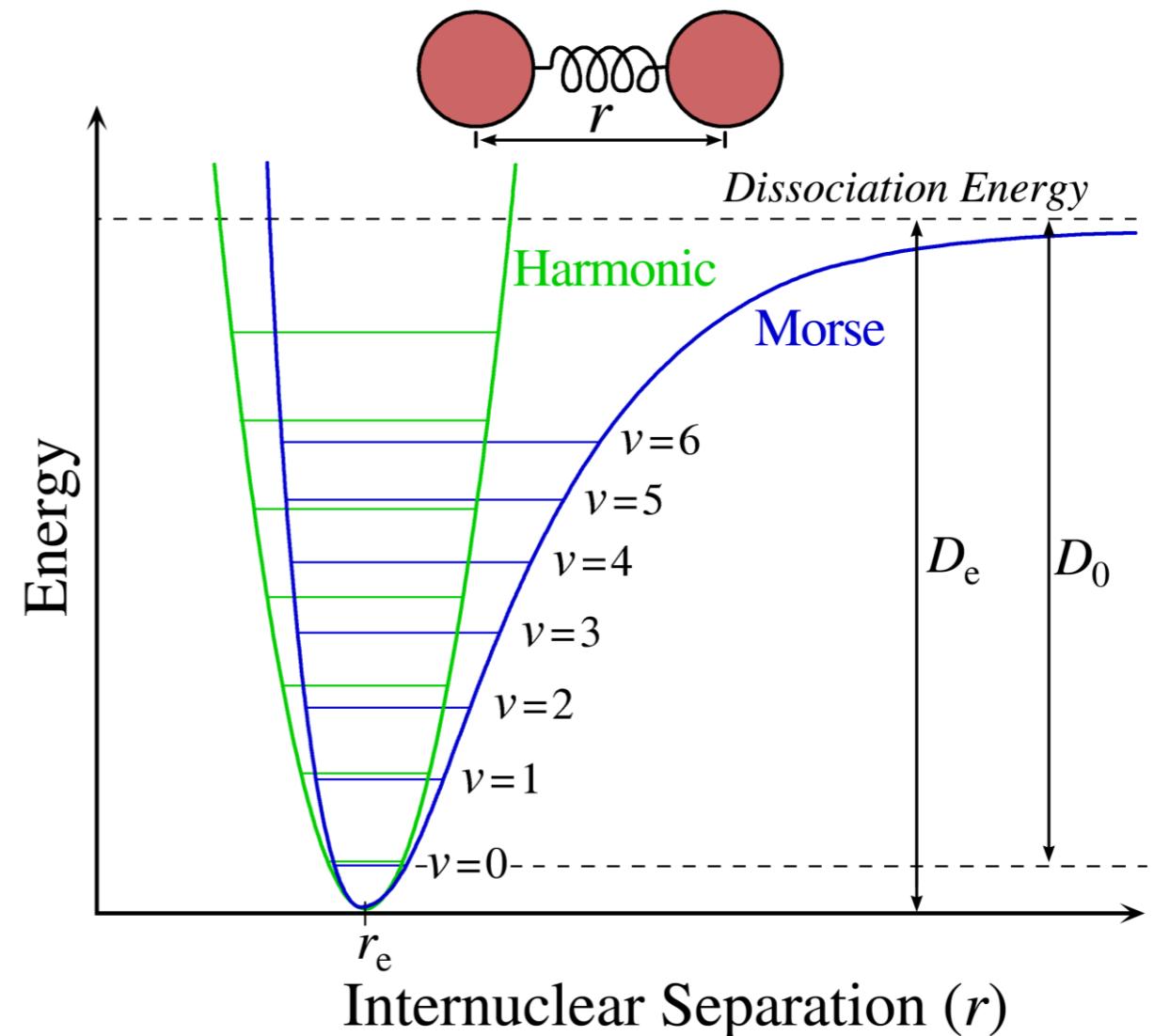
Non-bonded terms:

- Electrostatics
- Van der Waals



BONDED TERMS: BONDS

- **Morse potential** would be the closest to reality, but it is not efficient to compute;
- **Harmonic potential** is a reasonable approximation considering that during a simulation bond lengths do not differ much from the equilibrium value;



$$U_{Morse} = D_e(1 - e^{-a(r-r_e)})^2$$

$$U_{harmonic} = \frac{1}{2}k(r - r_e)^2$$

BONDED TERMS: ANGLES AND DIHEDRALS

- Angles are typically represented by a harmonic potential as well;
- Dihedrals are treated instead using a periodic function:

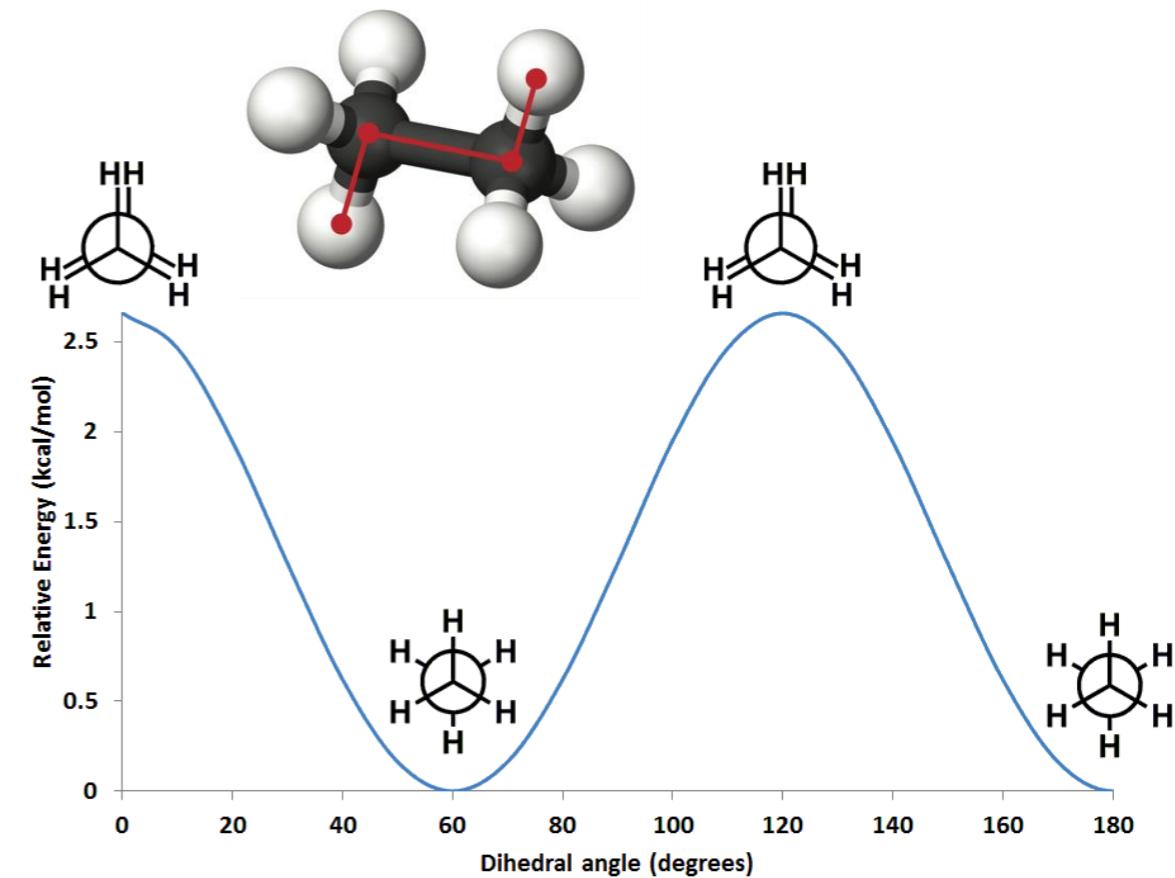
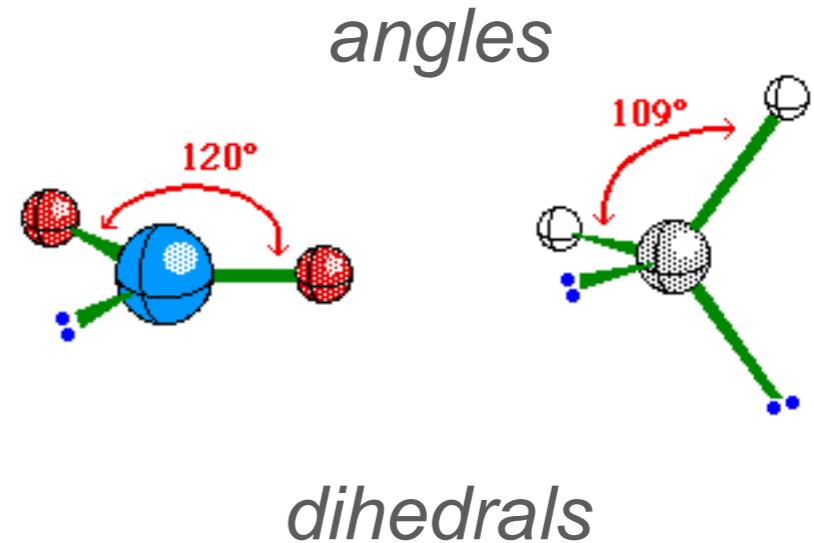
$$U(\theta) = \frac{1}{2}k(\theta - \theta_0)^2$$

$$U(\omega) = \frac{1}{2}V_n[1 + \cos(n\omega - \gamma)]$$

n = multiplicity

V_n = barrier height

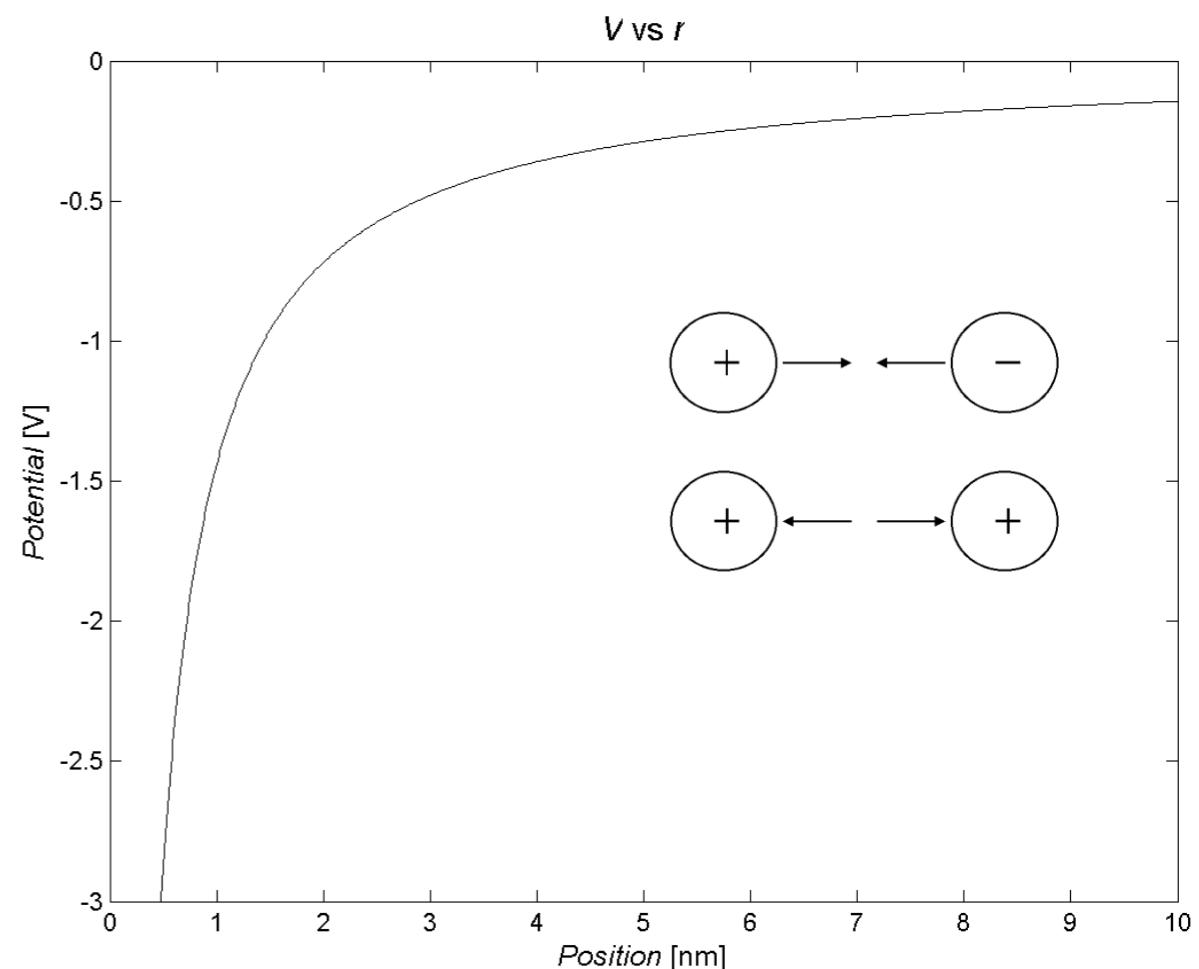
γ = phase (where the minima are)



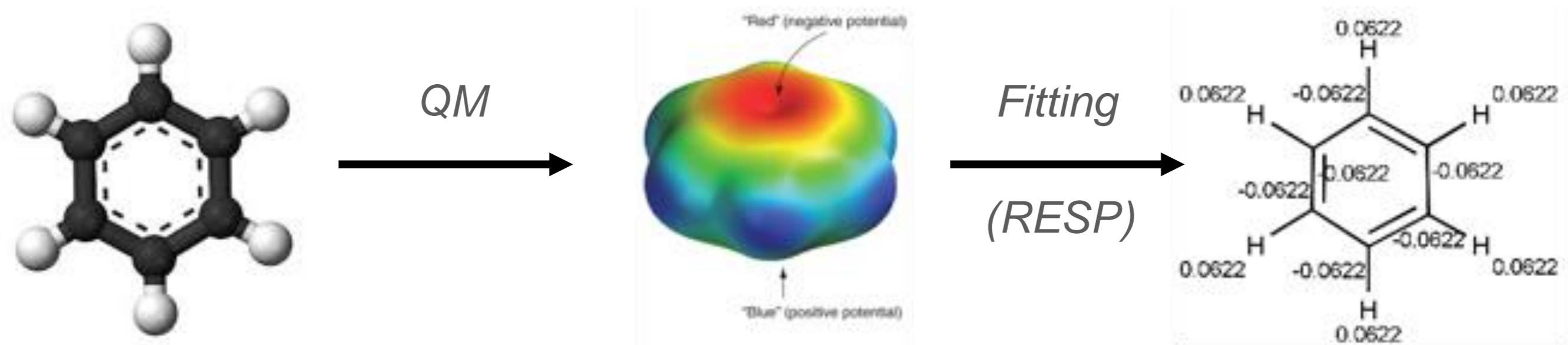
NON-BONDED TERMS: ELECTROSTATICS

- In molecular mechanics, each atom is assigned a **partial charge**;
- The interaction energy between any two partial charges (q_1 and q_2) on atoms i and j is described using **Coulomb's law**:

$$U_{electr} = \frac{q_1 q_2}{4\pi\epsilon_0 r_{ij}}$$



HOW DO WE GET THE PARTIAL CHARGES?

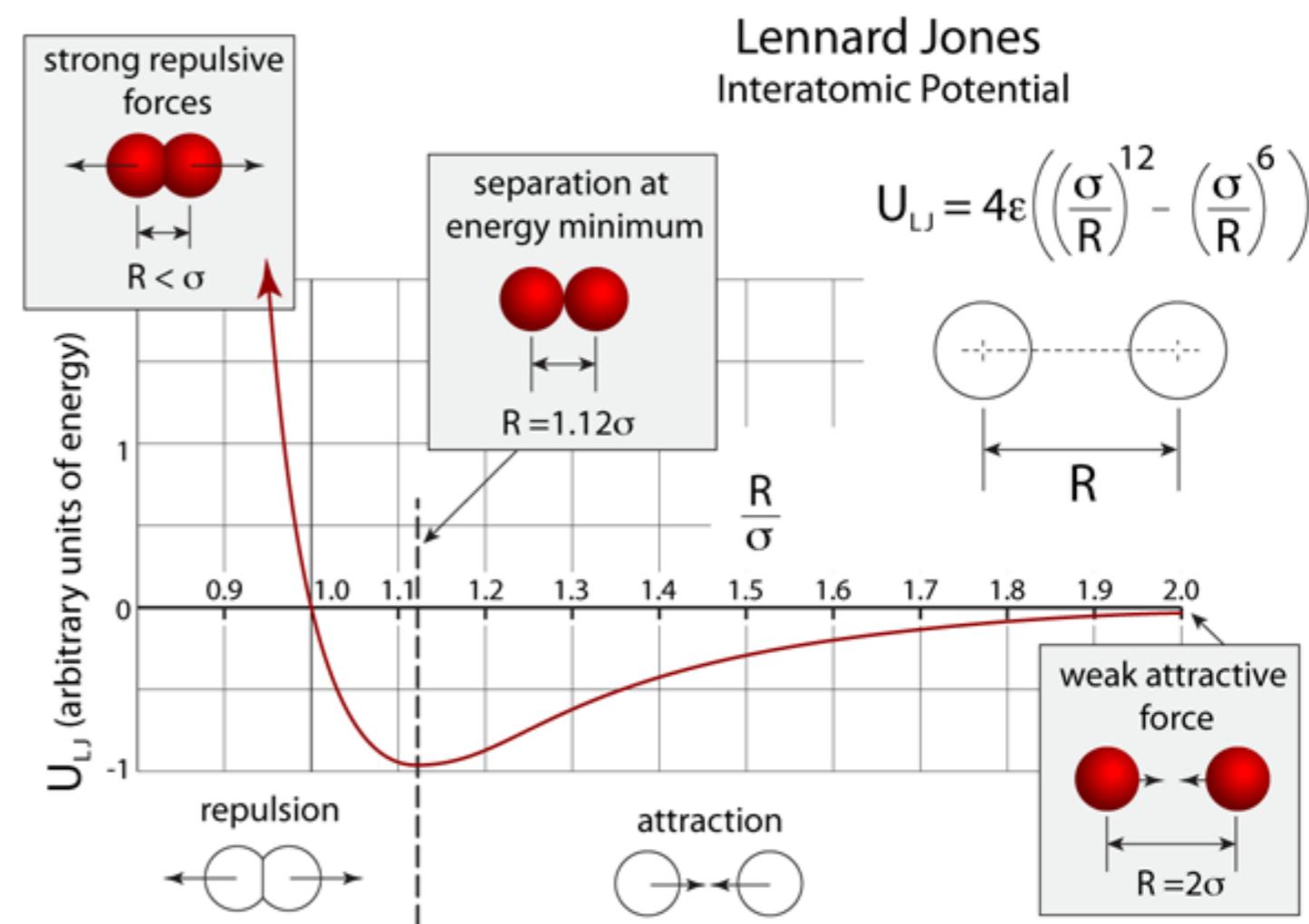


NON-BONDED TERMS: VAN DER WAALS

- Treated with the Lennard-Jones potential:

$$U_{LJ} = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right]$$

- Both attractive and repulsive;
- At short distance there is a steep repulsive potential, while at long distance the attractive energy approaches zero.



FORCE FIELD: SUMMARY

- Complete expression for potential energy (force field):

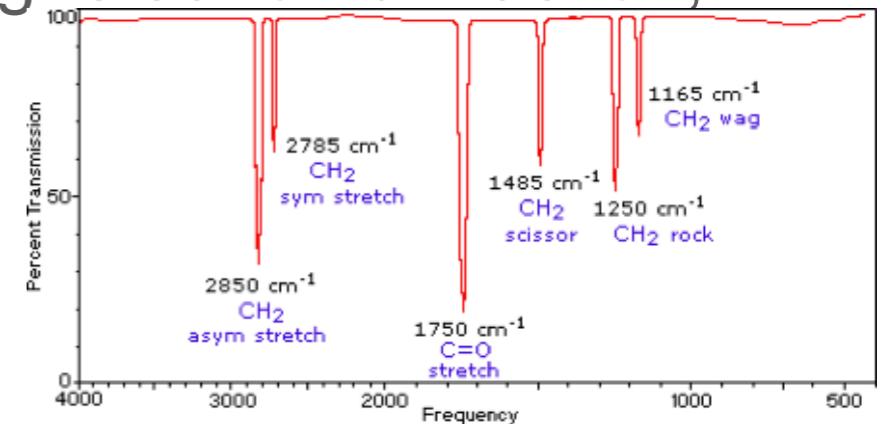
$$U(\mathbf{r}) = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \sum_{n=1}^{\infty} \frac{V_n}{2} [1 + \cos(n\varphi - \gamma)] \\ + \sum_{i < j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right] + \sum_{i < j} \frac{q_i q_j}{\epsilon r_{ij}}$$

- Parameters that need to be determined (force field parameterisation):

$$K_r, K_\theta, V_n, A_{ij}, B_{ij}, q_i$$

FORCE FIELD PARAMETRISATION

- Geometries often from structural studies (e.g. electron diffraction, microwave, crystallography);
- Force constants from spectroscopic data;
- QM data is often used too;
- Van der Waals parameters the hardest to derive; often iterative optimisation to match experimental data such as densities and heats of vaporisation;
- Partial charges can be obtained by fitting the QM electrostatic potential;
- Overall, different force fields are derived using different sets of data and procedures, so it is **inappropriate to mix and match parameters from different force fields.**



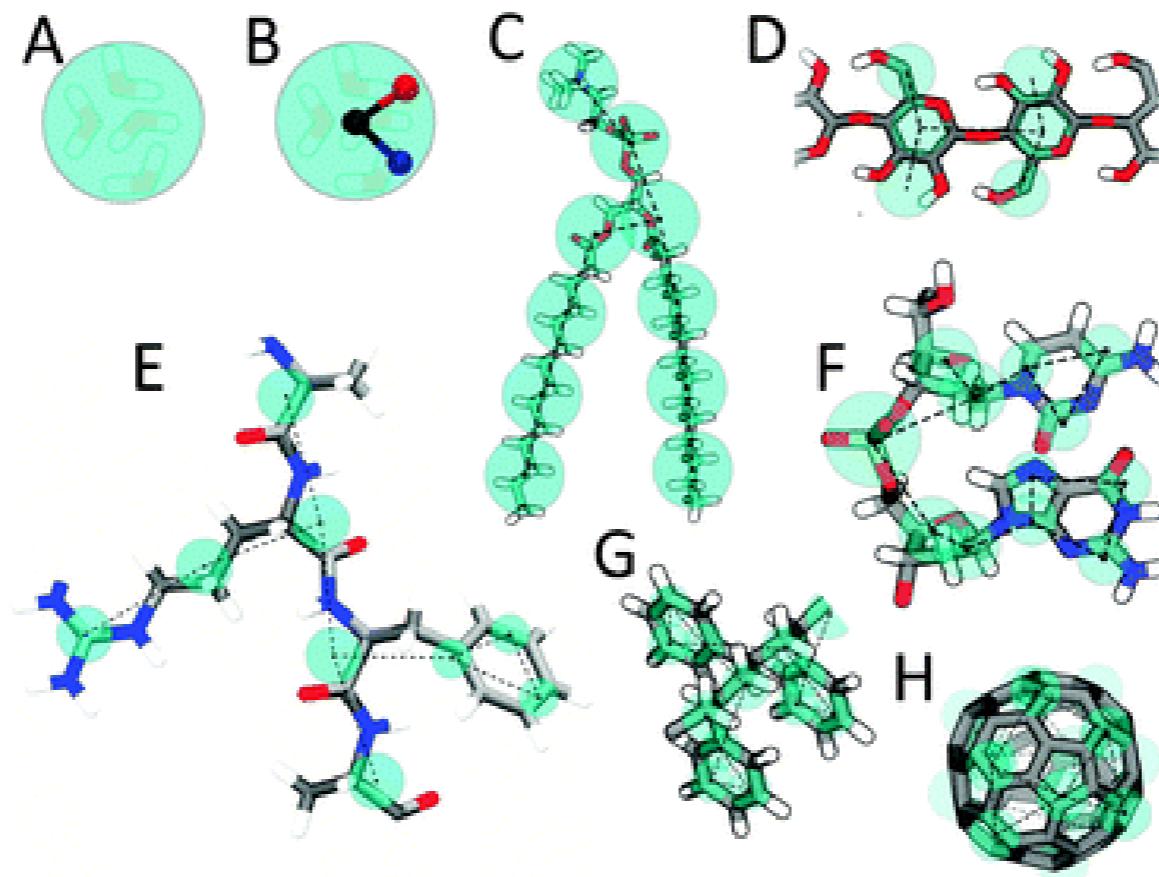
EXAMPLES OF DIFFERENT FORCE FIELDS

- AMBER: proteins, nucleic acids, lipids;
- GAFF: organic molecules (consistent with AMBER);
- CHARMM: proteins, nucleic acids, lipids;
- CGenFF: organic molecules (consistent with CHARMM);
- OPLS: proteins and organic molecules;
- GROMOS: proteins and nucleic acids;
- MMFF: organic molecules;

Constantly updated and revised.

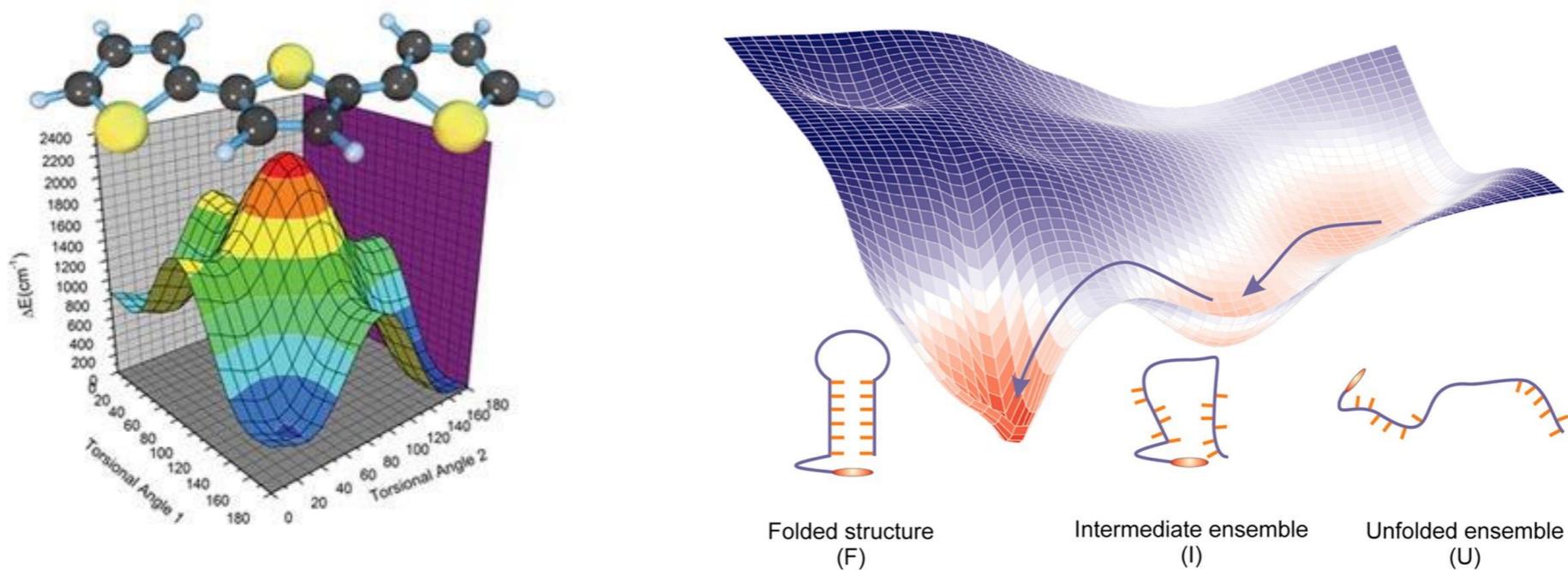
CORSE-GRAINED MODELS

- There are also more approximate (coarse-grained) models for organic and biological molecules, which allows the exploration of large systems (e.g. viruses, big membranes) and long time-scales at a lower computational cost;
- One particle represents more than one atom.



POTENTIAL ENERGY SURFACE

- The way the potential energy varies with the coordinates is referred to as the potential energy surface;
- Since the potential energy is defined by the force field, this will define the shape of the surface;
- We can use computer simulations such as MD to explore these complex multidimensional energy surfaces;



SOME CONSIDERATIONS

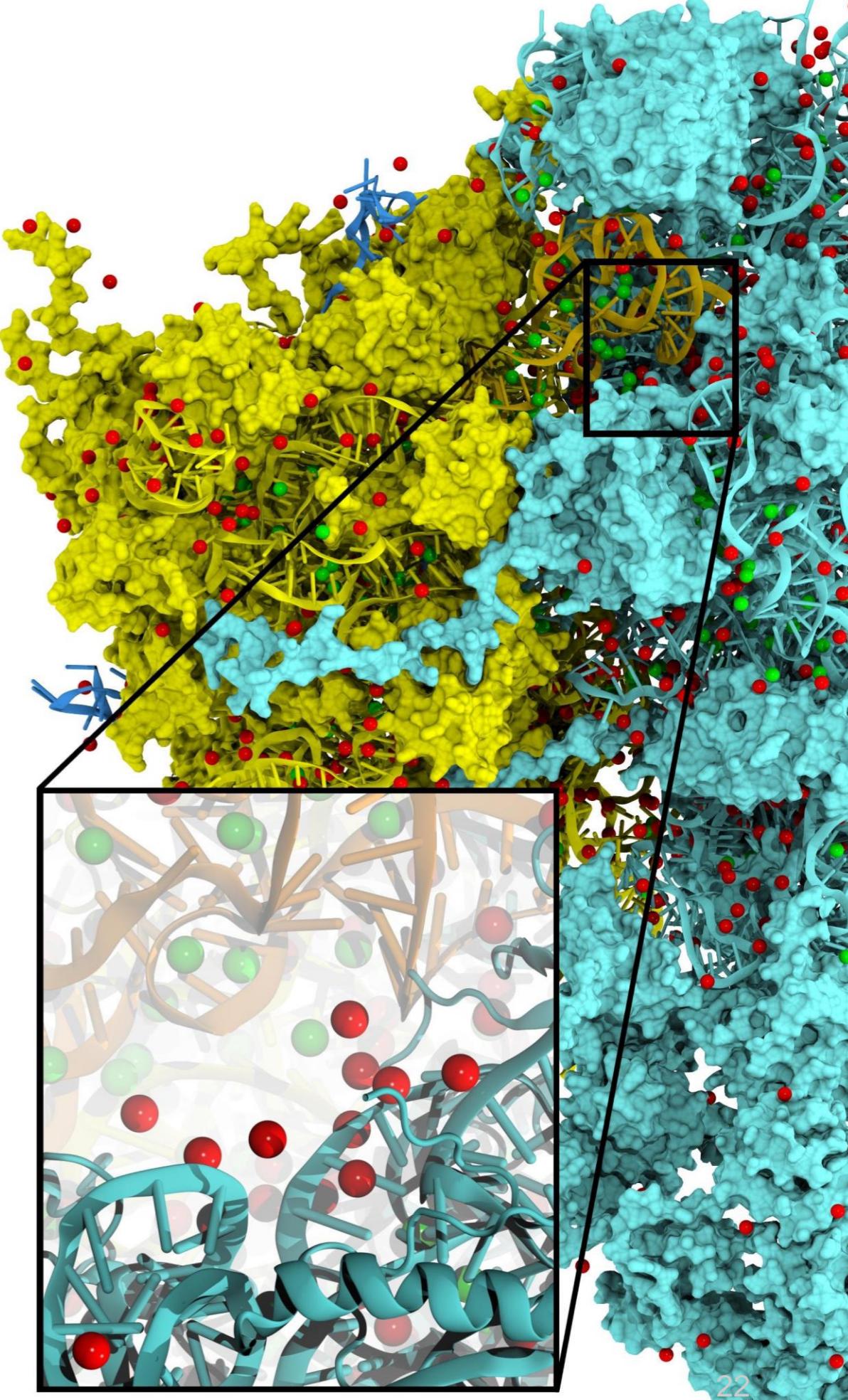
- Empirical since force fields are based on a number of assumptions and simplifications, and just parameterised to match experimental data; the functional forms used are not justified per se.
- There are a number of limitations:
 - lack of polarisation;
 - no bond-breaking;
 - limited transferability of some parameters;

“All models are wrong, some are useful.”

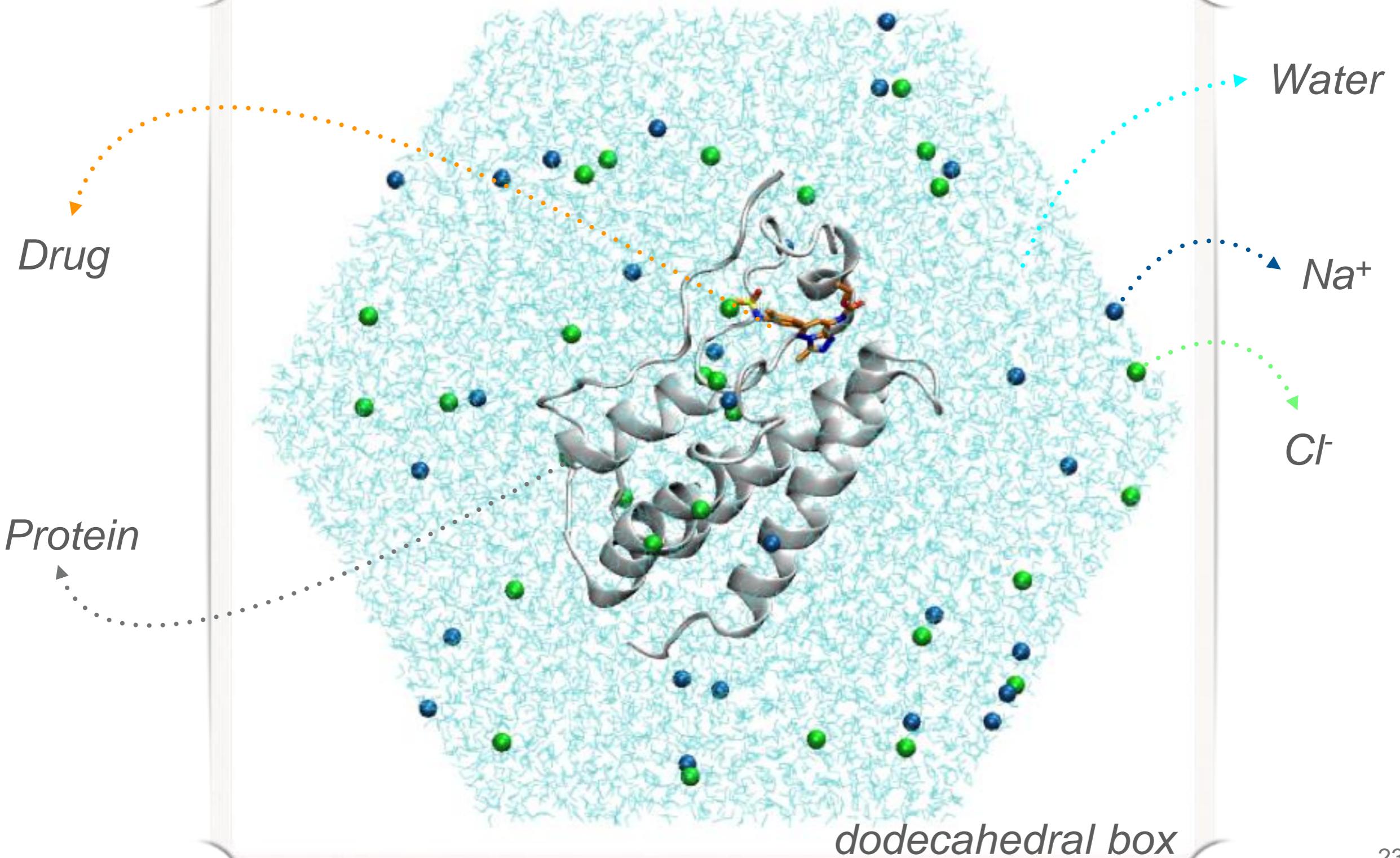
—G. Box

MOLECULAR DYNAMICS

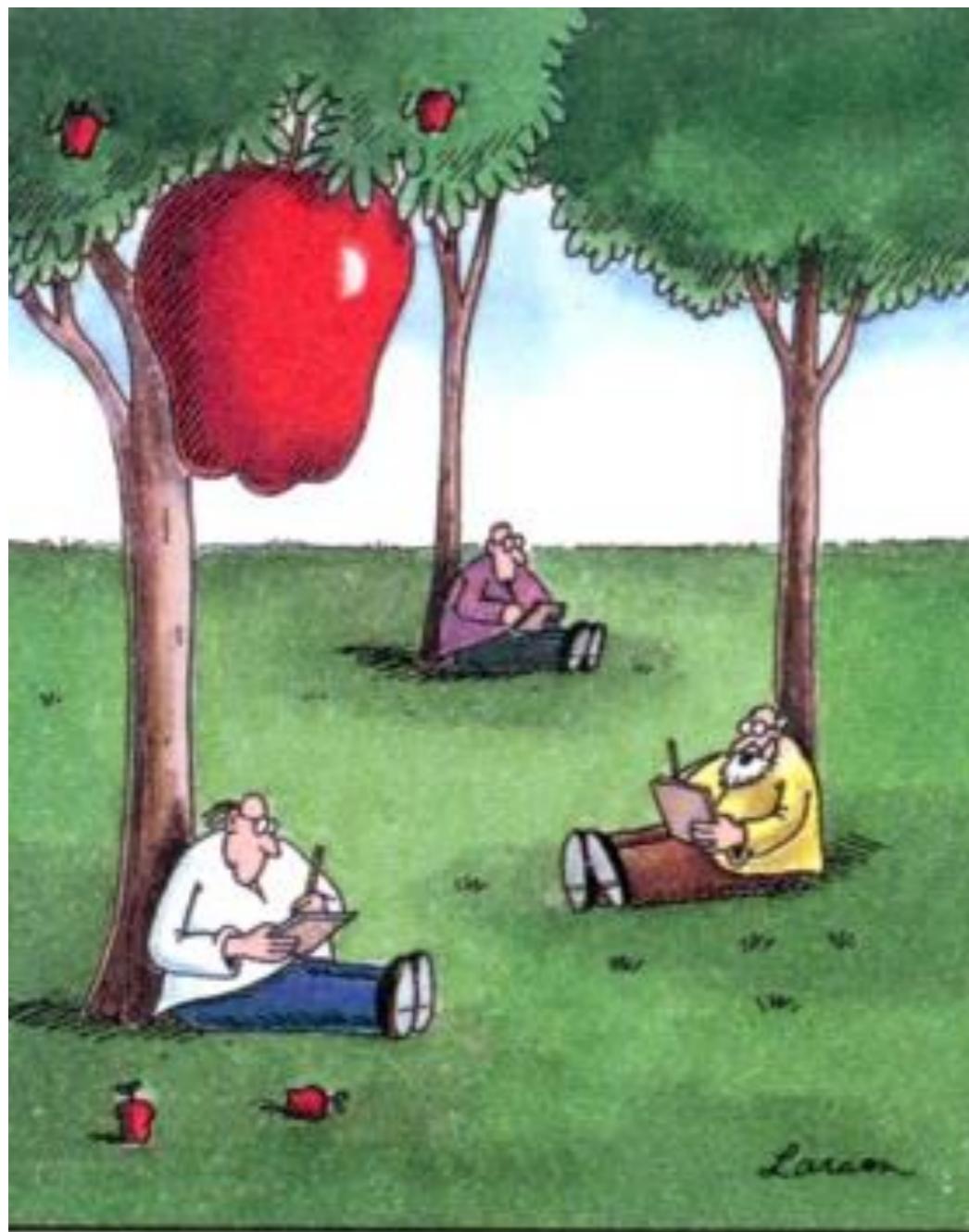
Fundamentals



AN EXAMPLE SYSTEM



MD IN A NUTSHELL



"Nothing yet... How about you, Newton?"

- We want to see how a molecular system evolves in time, based on the potential energy defined by the force field:

$$\vec{F} = m\vec{a}$$

$$-\nabla U = m\ddot{\vec{r}}$$

$$\frac{1}{m} \iint -\nabla U dt dt = \vec{r}$$

*potential energy
from the force field*

*the coordinates for all particles,
i.e. where the atoms are*

FINITE DIFFERENCE METHODS

- Due to the complexity of the systems, in general there is **no analytical solution** to the equations of motion, i.e. they must be solved numerically;
- **Integration is broken down into many time-steps**, separated by dt ;
- Different algorithms are available, e.g. Verlet, leap-frog, velocity Verlet, Beeman's;
- At each step, the total force on each particle is calculated and the acceleration determined; from the acceleration we obtain velocities and the new coordinates at $t+dt$;
- The force is assumed to be constant during each time step;
- time step is limited by the fastest motions in the system (in “vanilla” MD: vibrations of bonds limit time step to 1 - 2.5 fs).

Example: leap-frog algorithm

$$\vec{r}(t + dt) = \vec{r}(t) + \vec{v}(t + \frac{1}{2}dt)dt \quad \vec{v}(t + \frac{1}{2}dt) = \vec{v}(t - \frac{1}{2}dt) + \vec{a}(t)dt$$

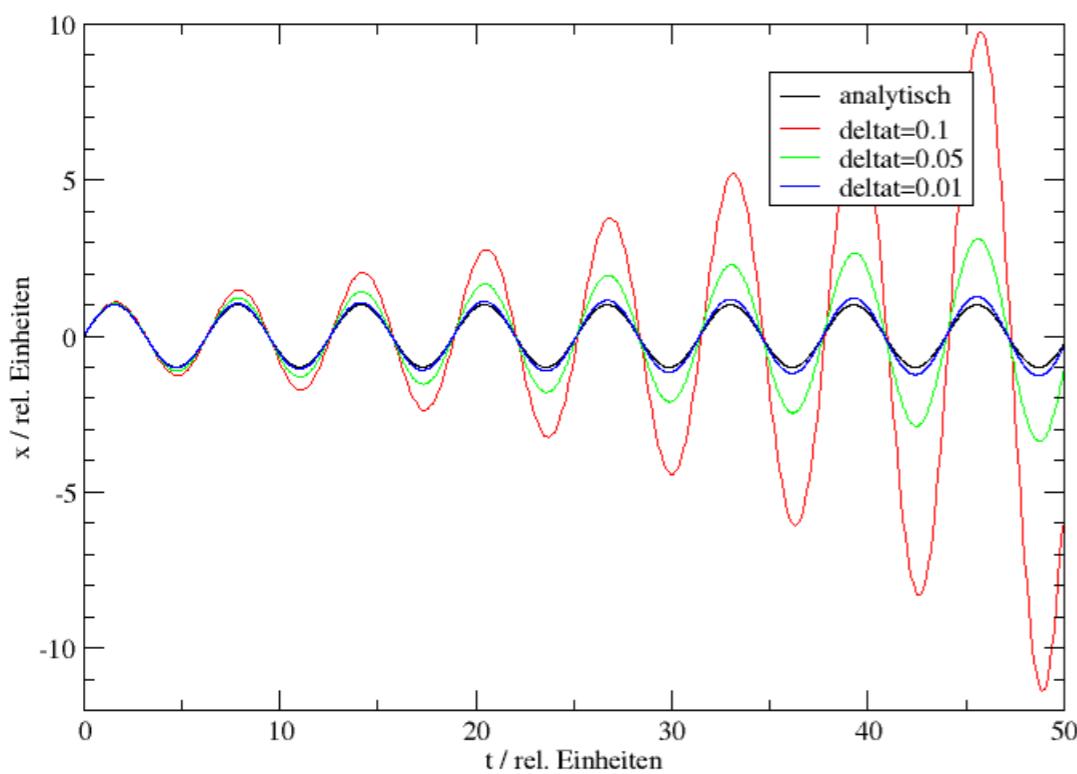
THE SIMPLEST MD: A HARMONIC OSCILLATOR (E.G. CO STRETCH)

- Equations of motion be solved analytically for harmonic oscillator:

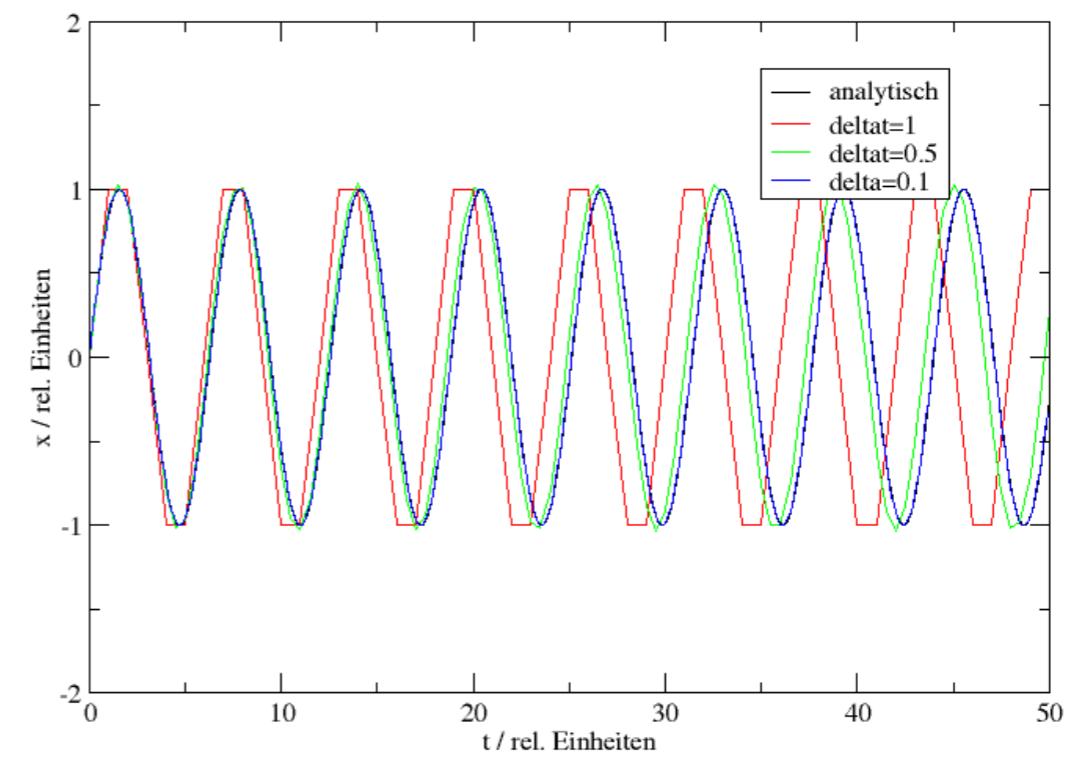
$$x(t) = \frac{v_0}{\omega} \sin(\omega(t - t_0)) + x_0 \cos(\omega(t - t_0)) \quad \omega = \sqrt{\frac{k}{m}}$$

- Numerical solution shows dependence on time step:

Euler

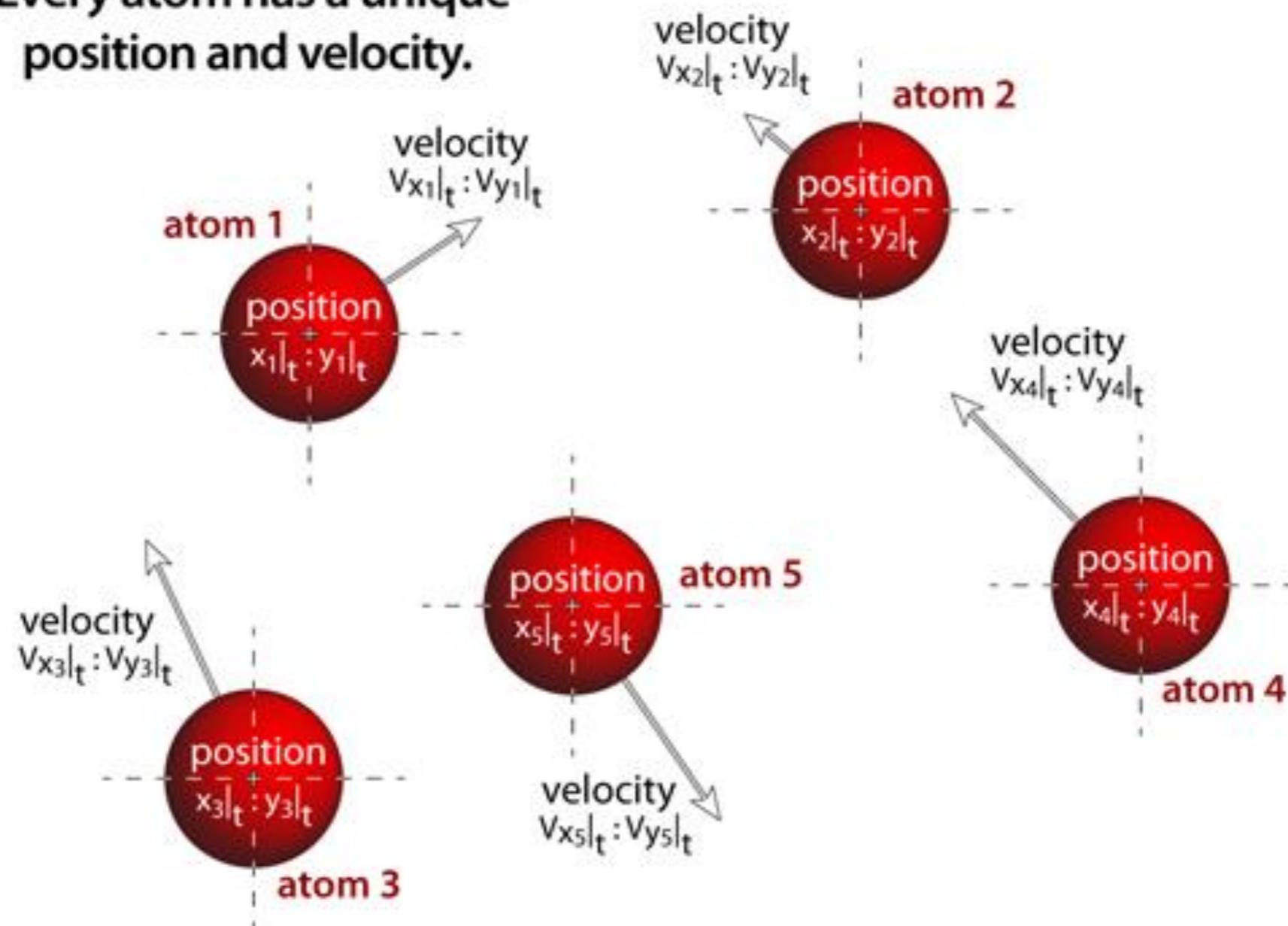


Verlet



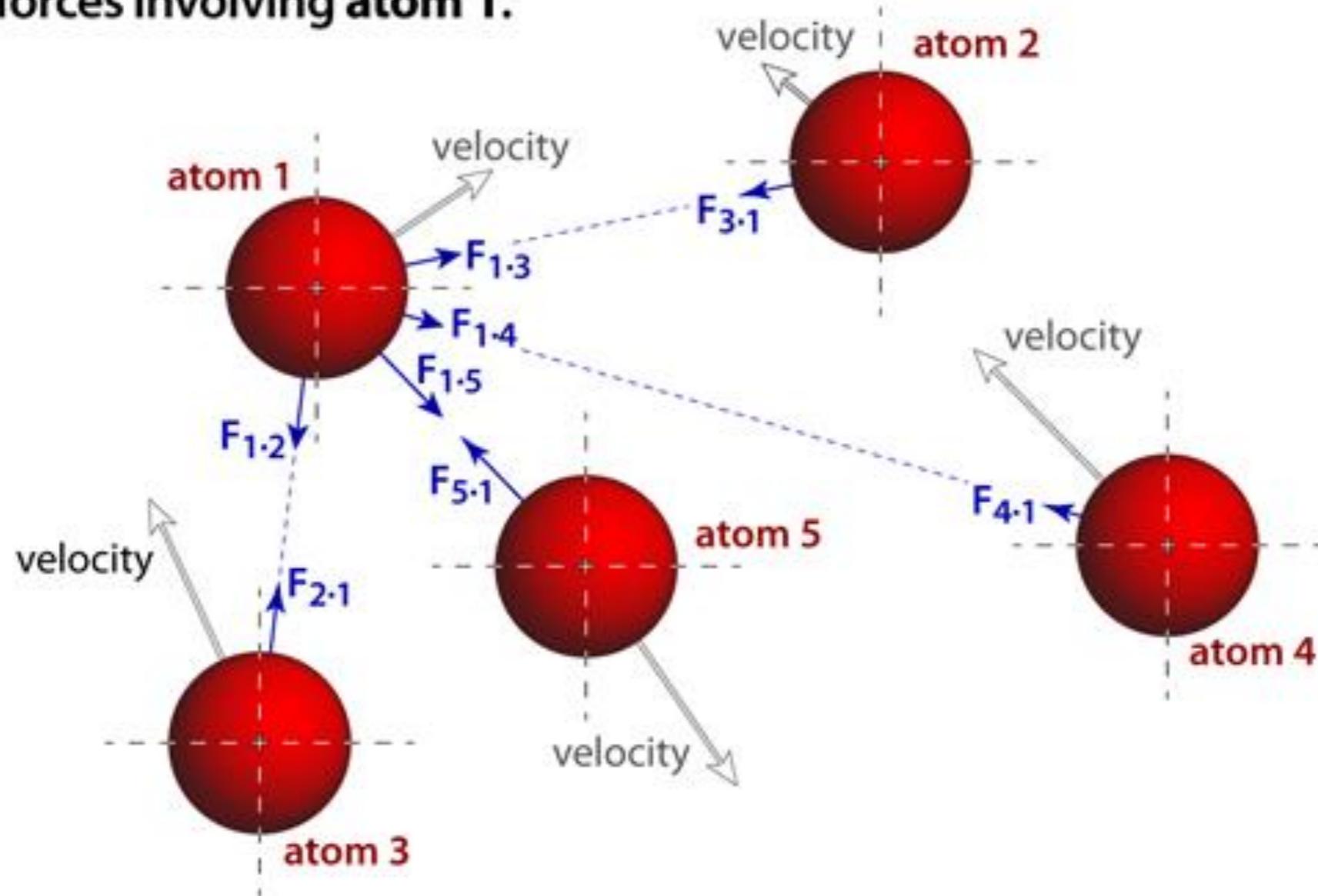
MD ILLUSTRATION

Every atom has a unique position and velocity.



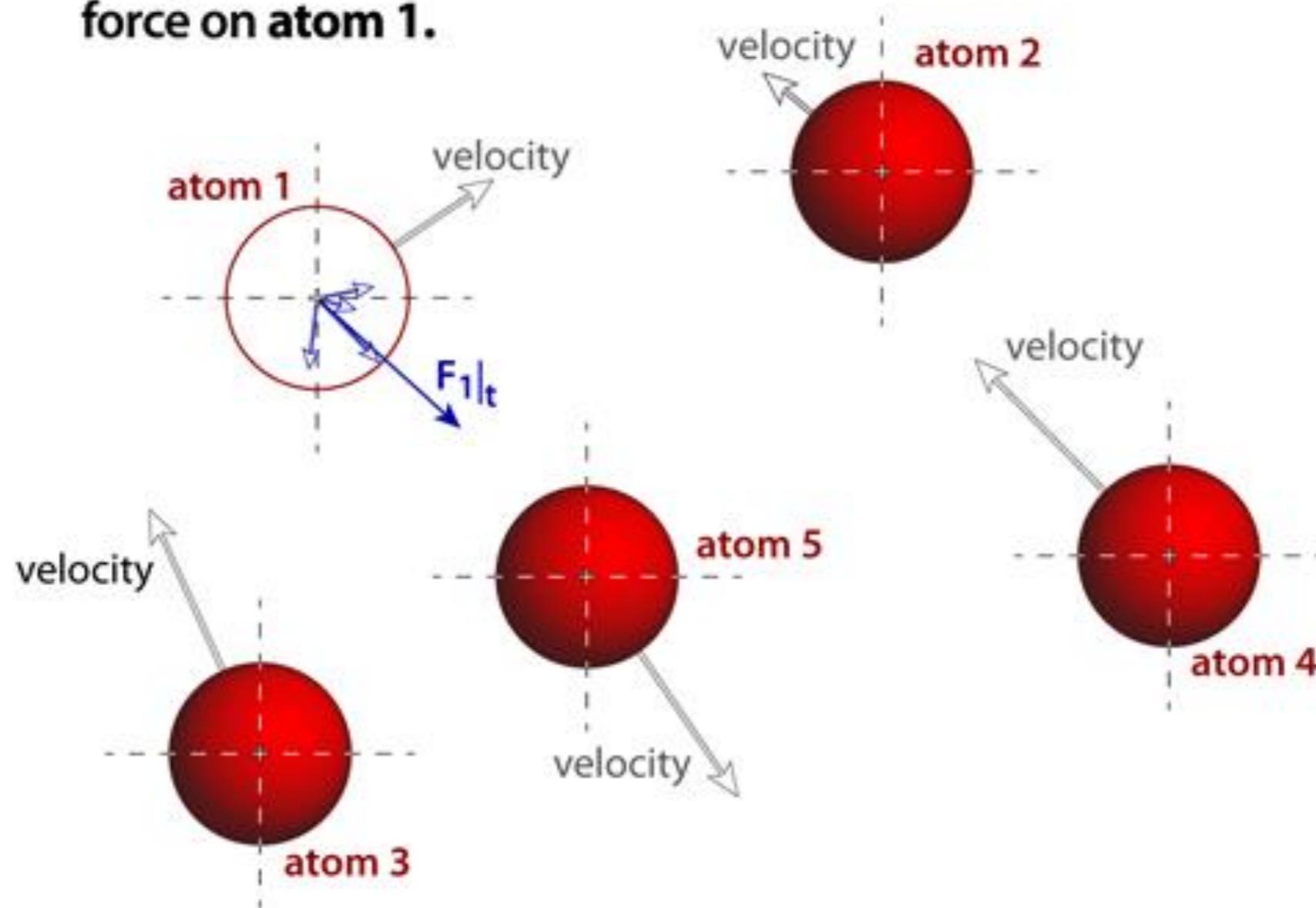
MD ILLUSTRATION

Calculate the interatomic forces involving atom 1.



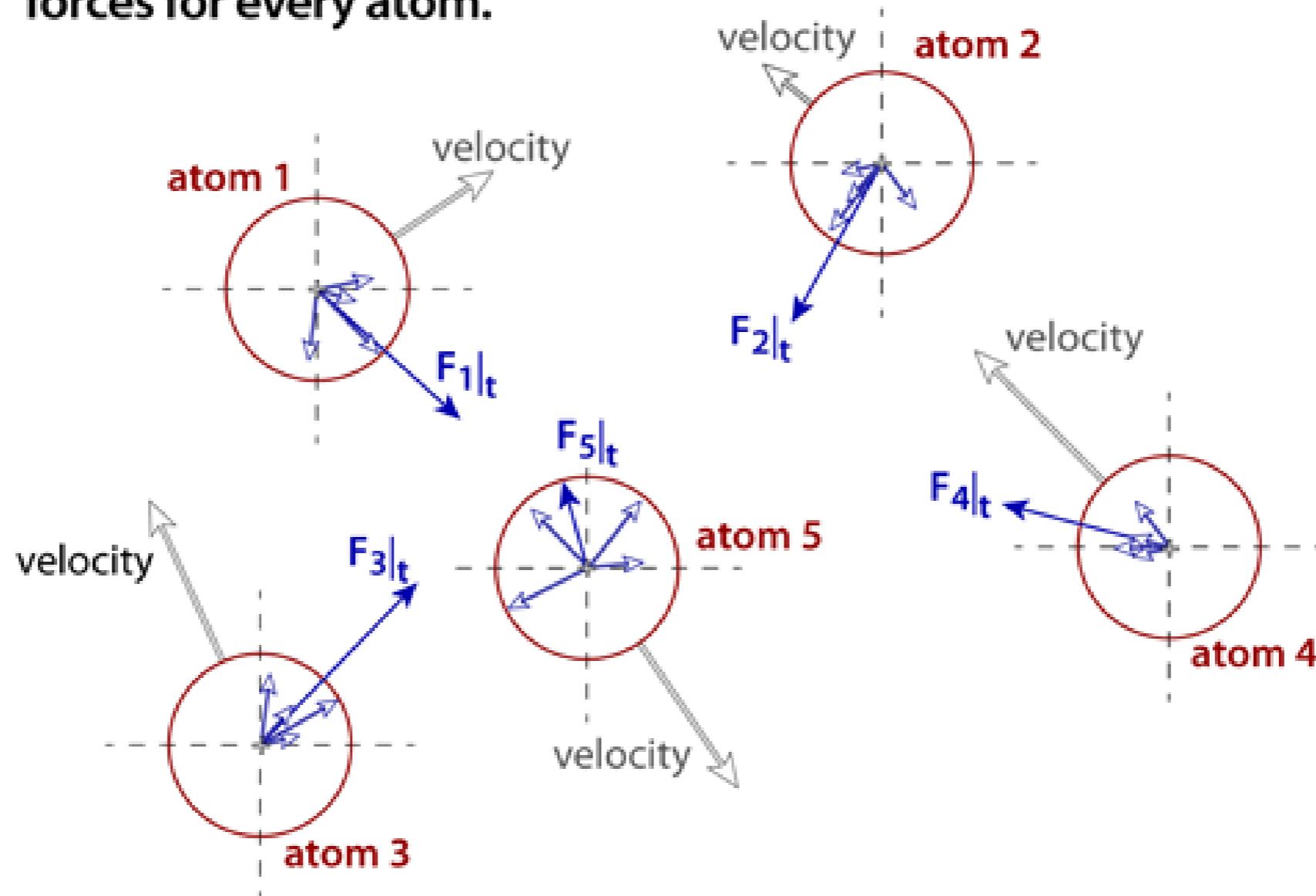
MD ILLUSTRATION

Compute the net force on atom 1.



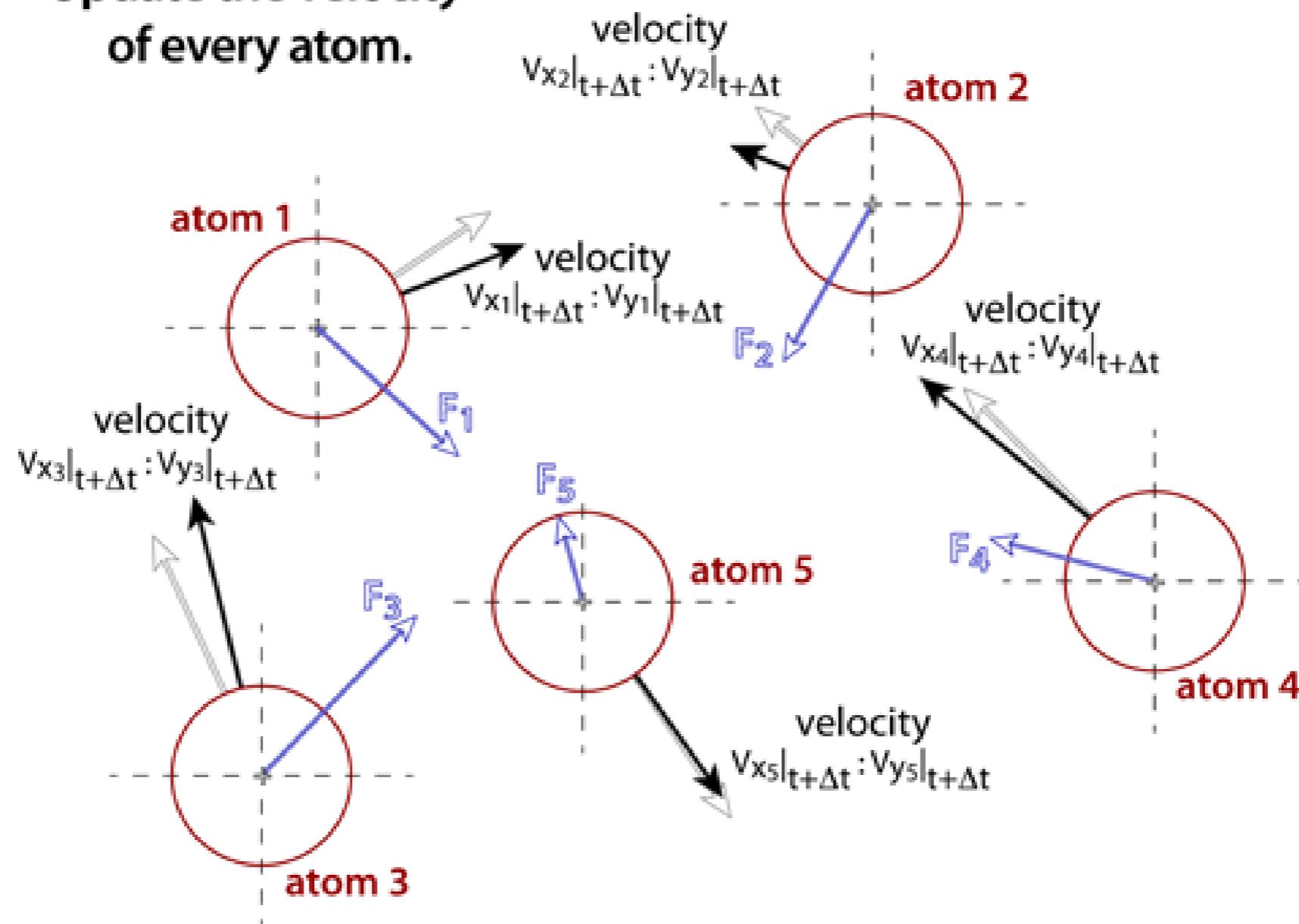
MD ILLUSTRATION

Compute the interatomic
forces for every atom.



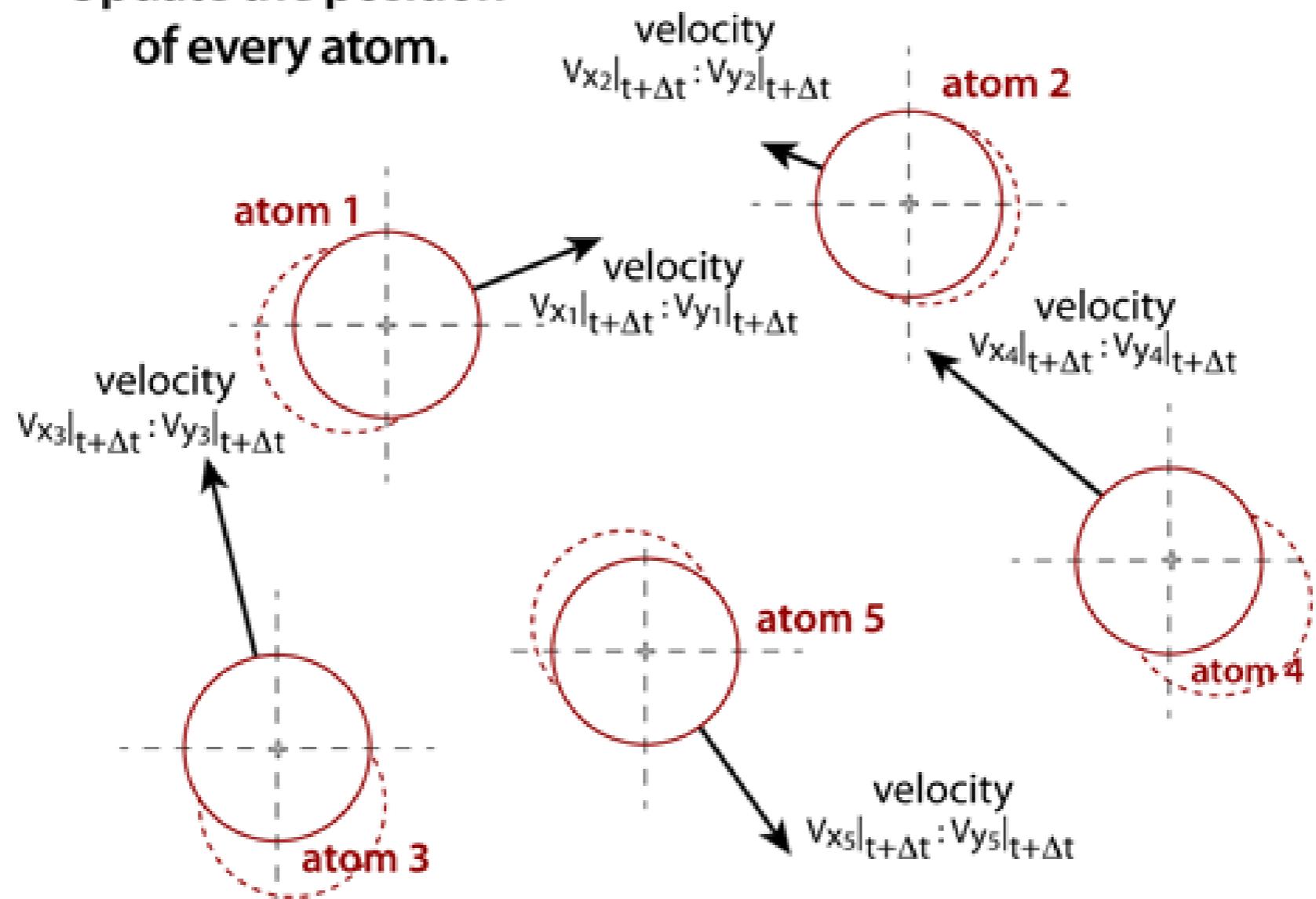
MD ILLUSTRATION

Update the velocity
of every atom.



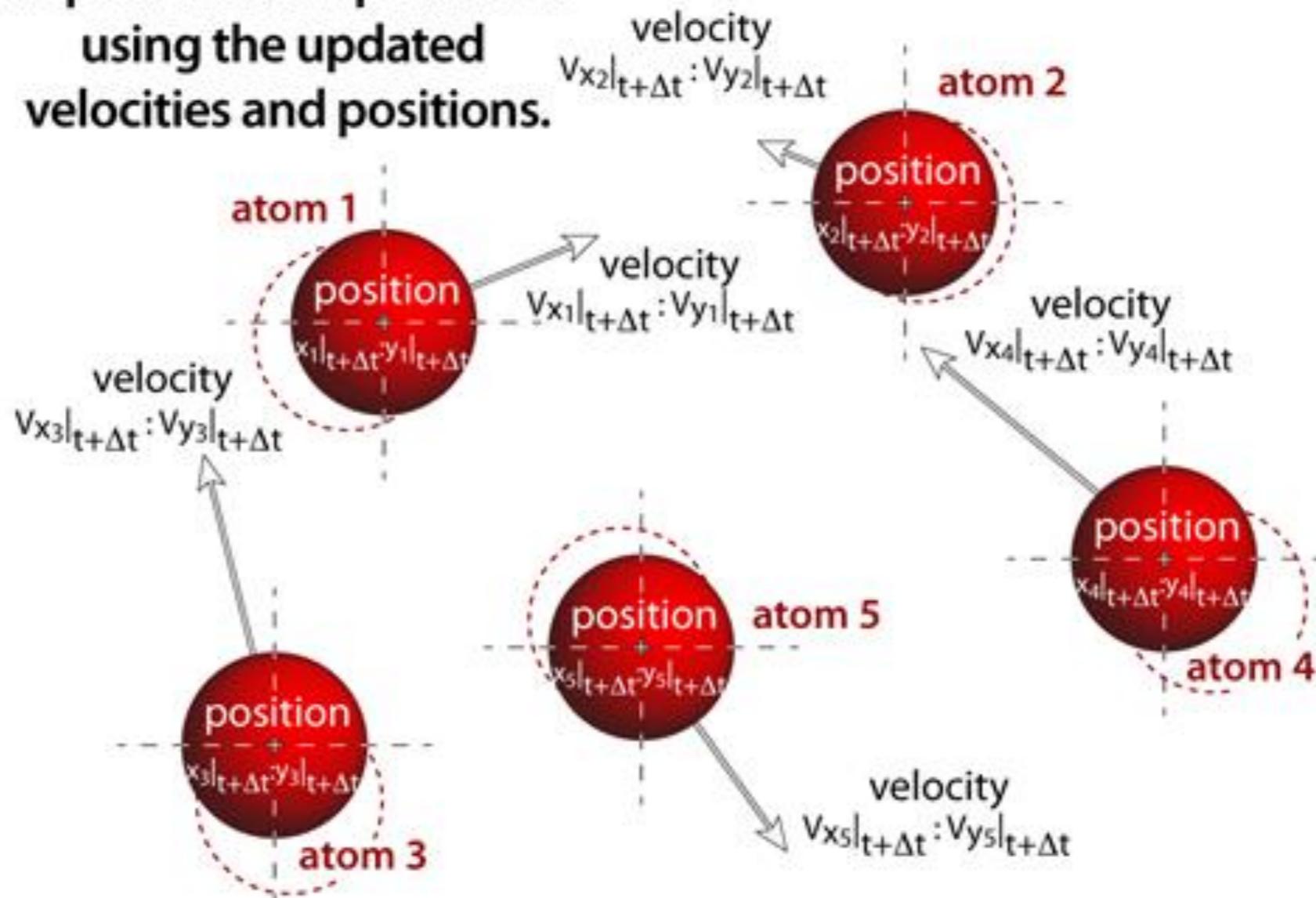
MD ILLUSTRATION

Update the position
of every atom.



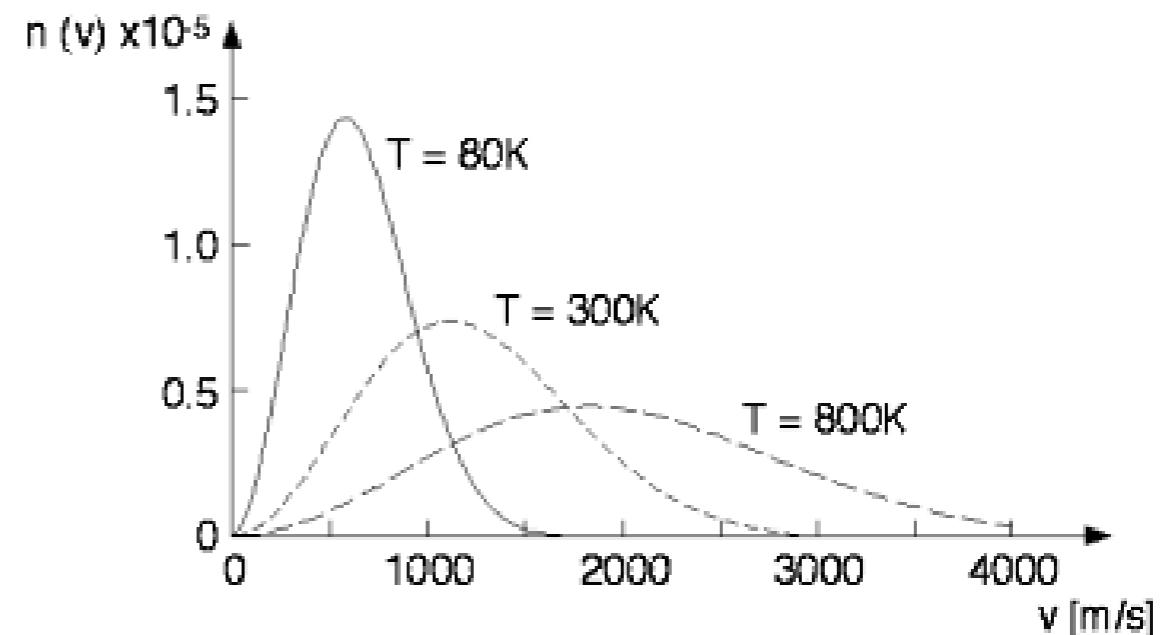
MD ILLUSTRATION

Repeat the computations
using the updated
velocities and positions.



INITIAL VELOCITIES

- What are the velocities at $t=0$? We need velocities at each time-step to determine the new atom positions!
- Solution is to randomly **assign velocities from the Maxwell-Boltzmann distribution** at a given temperature T .



$$p(v_{ix}) = \left(\frac{m_i}{2\pi k_B T} \right)^{1/2} \cdot \exp \left[- \frac{m_i v_{ix}^2}{2k_B T} \right]$$

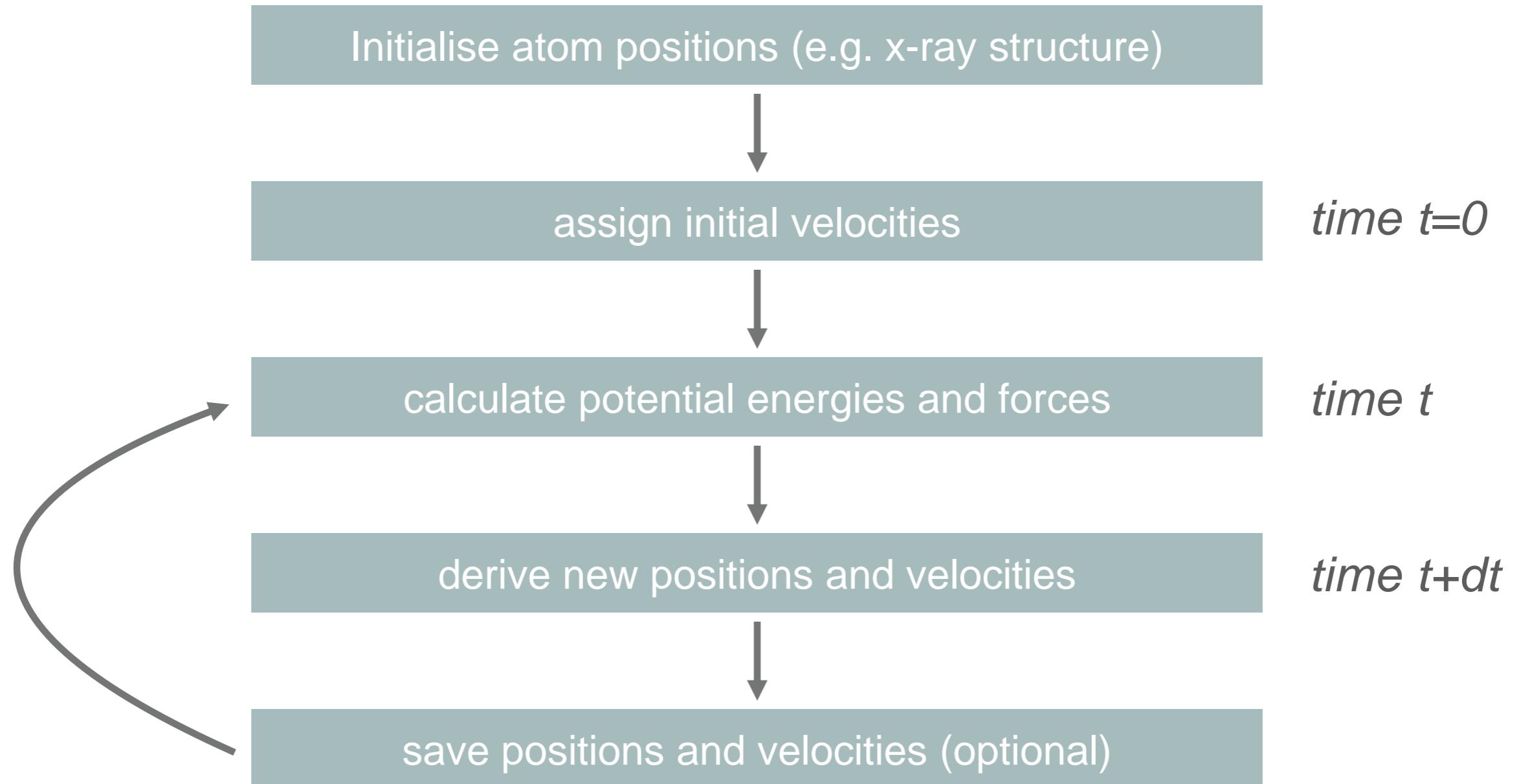
$$T = \frac{1}{3N} \sum_{i=1}^N \frac{|p_i|}{2m_i}$$

N = number of particles

p = momentum

u_{ix} = velocity of particle i in x direction

SCHEMATIC OF MD ALGORITHM



STATISTICAL MECHANICS: ENSEMBLES

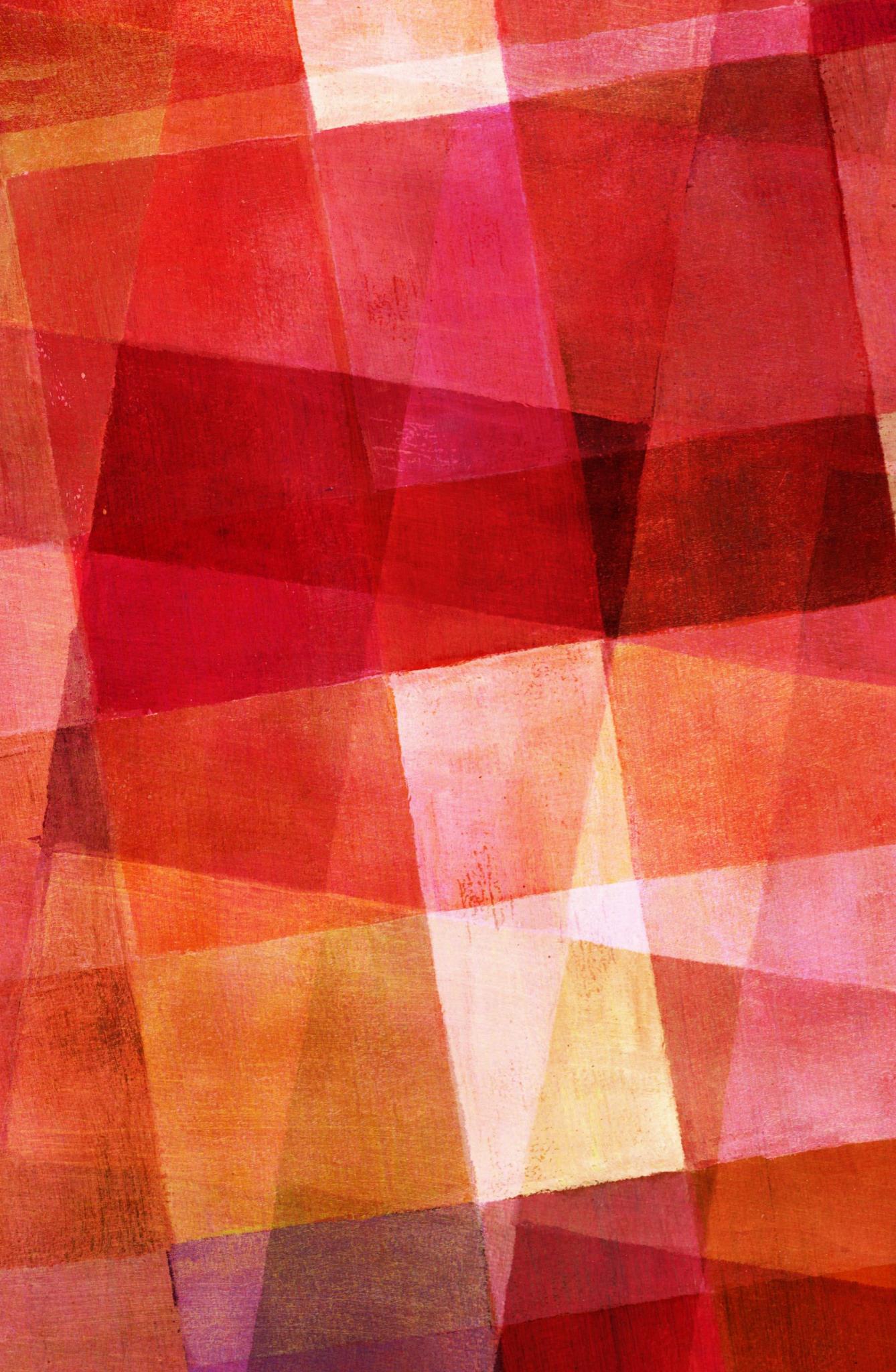
- An ensemble is the collection of all microscopic states accessible to a particular system;
- To mimic experimental conditions, one might want to keep the temperature, volume, or pressure constant during the simulation;
- The typical ensembles used in MD are:
 - **NVE: microcanonical ensemble**
 - **NVT: canonical ensemble**
 - **NPT: isothermal-isobaric ensemble**
- The NVE ensemble will maximise the entropy, while the NVT and NPT ensembles will tend to the minimum Helmholtz and Gibbs free energies respectively.

N = number of particles
 V = volume
 T = temperature
 P = pressure
 E = energy

ERGODIC HYPOTHESIS

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time}$$

- With MD we can calculate time averages, but experimental observables are assumed to be ensemble averages;
- Basically, the ergodic hypothesis suggests that over a long time period the microstates sampled by MD will match the microstates of the statistical ensemble; so we can take the average of a property over the time of the MD trajectory as the ensemble average;
- It follows that appropriate sampling of the phase space is needed to ensure the ergodicity of the system.



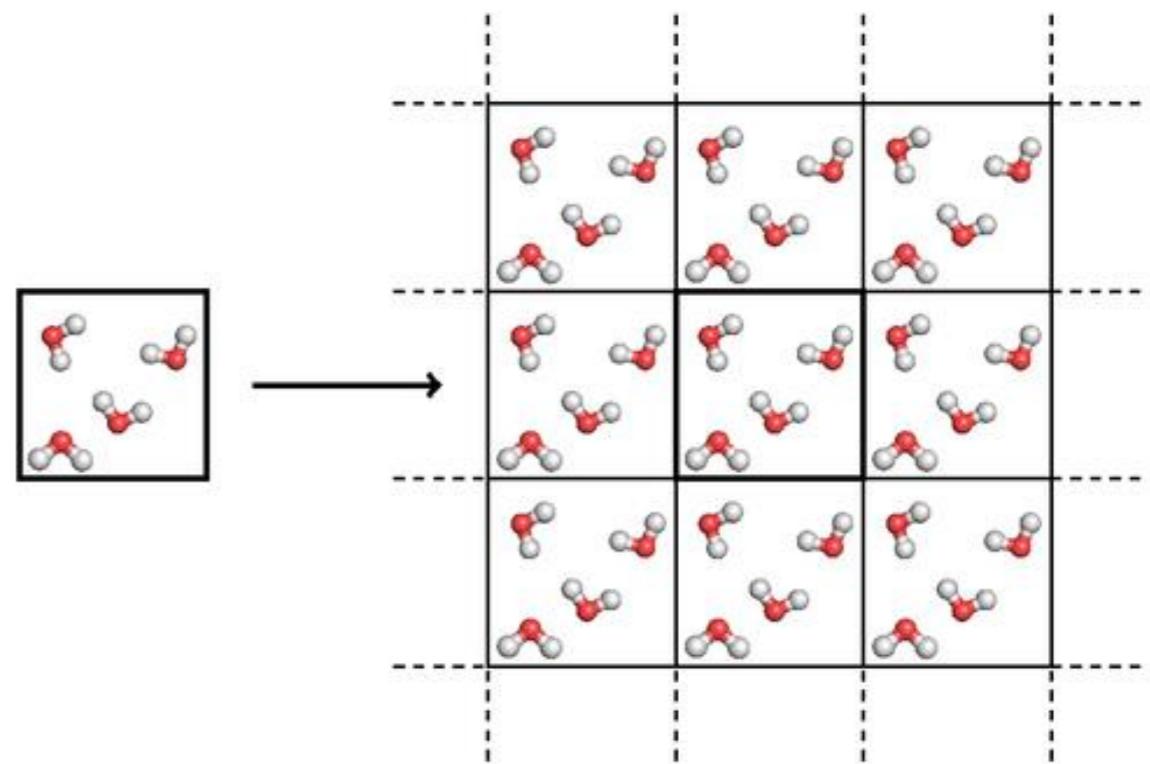
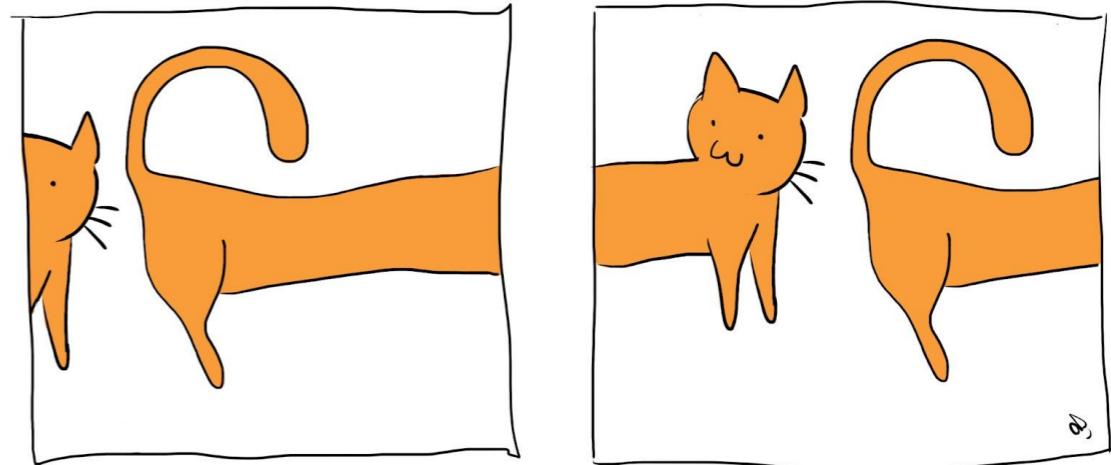
MD PRACTICALITIES

- Boundary conditions
- Truncating the potential
- Neighbour lists
- Particle Mesh Ewald
- Constraints

BOUNDARY CONDITIONS

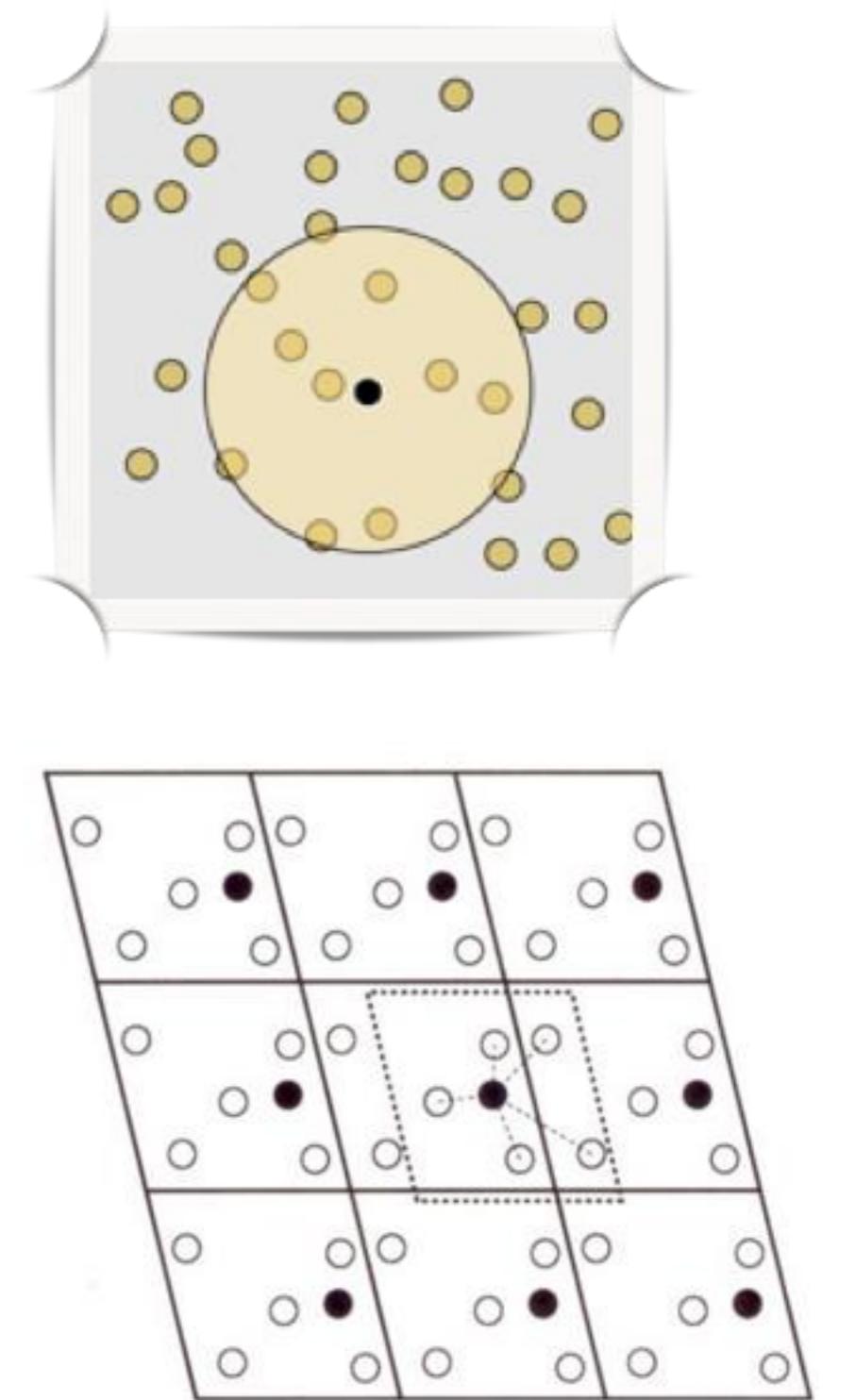
- We are simulating a very small box, but want to extract macroscopic properties of the system: what happens when a molecule approaches the boundary of the box? What force does it experience?
- **Periodic boundary conditions** allows us to simulate a small number of particles in such a way that these particles will experience forces as if they were in bulk;

PERIODIC BOUNDARY CONDITIONS



LENNARD-JONES CUTOFF

- The most **time consuming** part of MD is the calculation of the **non-bonded energies**, since each particle interacts with all other particles in the system: N^2 problem;
- The Lennard-Jones potential however falls off pretty rapidly with the distance (r^{-6} dependence);
- An option is thus to simply use a cut-off and the LJ interactions will be calculated only for particles within the cut-off; outside the cutoff the LJ energy is set to zero;
- The **minimum image convention** is applied (each atom sees only one copy of any other atom) and the cut-off must be smaller than half the size of the cubic box.



NEIGHBOUR LISTS

- How do we know whether an atom is within the cutoff or not?
We need to calculate the distances between all atoms at each time-step, but this defeats the point of using a cutoff for computational efficiency;
- However, when simulating fluids, an atom's neighbours (within the cutoff) do not change much over 10-20 time steps;
- A neighbour list **stores all atoms that are within the cutoff** (and the ones slightly further away) and it is **updated at regular intervals**, e.g. 10 every time steps;
- Only the distances and energies of the atoms in the list are calculated at each time-step.

PARTICLE MESH EWALD (PME)

- Electrostatic interactions decay with r^{-1} , therefore the error introduced by the use of a cutoff would be large;
- The most popular way to deal with this is to use Ewald summation, which is a **method for computing long-range interaction in periodic systems**;
- The electrostatic interaction is divided into a short-range and a long-range contribution, where the latter is calculated using a Fourier transform;
- Overall it is accurate and fast to converge, therefore quite popular.

CONSTRAINTS (NOT RESTRAINTS)

- The time-step used for integrating the equation of motion needs to be longer than the highest frequency motion in the system;
- But the more time-steps the more demanding the calculation, and high-frequency motion (e.g. bonds vibration) might not be of interest;
- Constraints allow to ‘fix’ **specific degrees of freedom** (e.g. vibration of bonds) so to be able to **increase the time-step** without prejudicing accuracy; typically 2 fs time-step with constrained bonds to hydrogens;
- Having the bonds constrained means they are forced to adopt a specific value throughout the simulation;
- Note that “restraints” instead introduce an additional energy term to the potential, biasing the simulation, but they still allow deviations from the desired values.



TYPICAL WORKFLOW

- MD software
- System preparation
- Energy minimisation
- Equilibration and production
- Analysis and errors

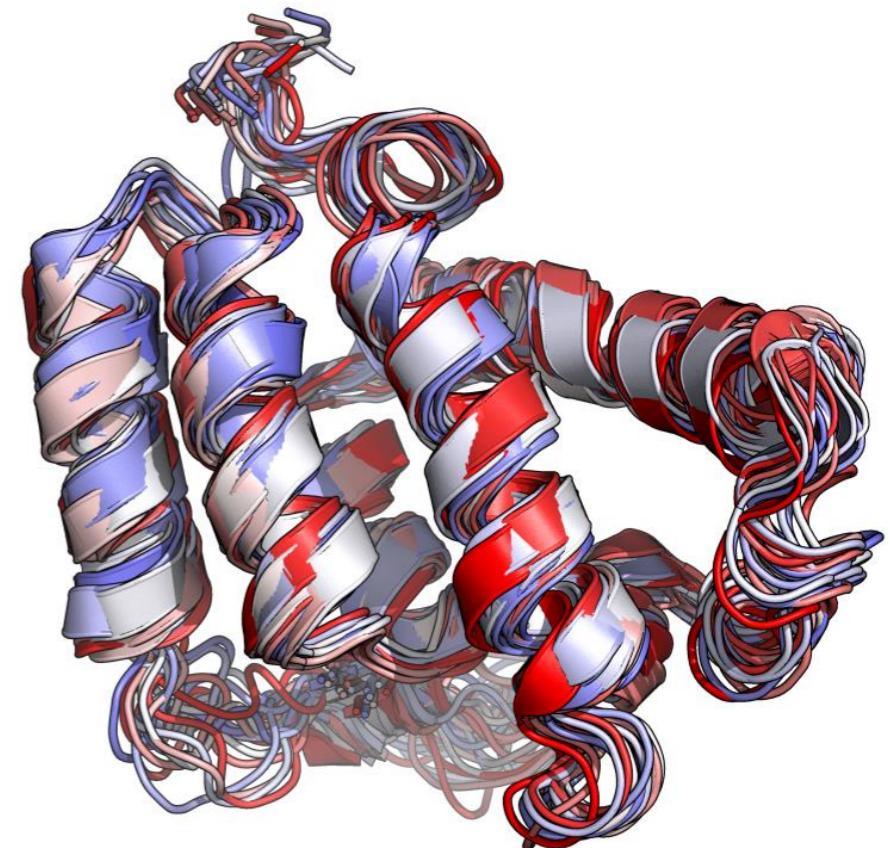
MD SOFTWARE

Just to name the most popular ones...

- Gromacs (Groningen Machine for Chemistry Simulations)
- AMBER (Assisted Model Building with Energy Refinement)
- CHARMM (Chemistry at HARvard Molecular Mechanics)
- NAMD (Not just Another MD programme)

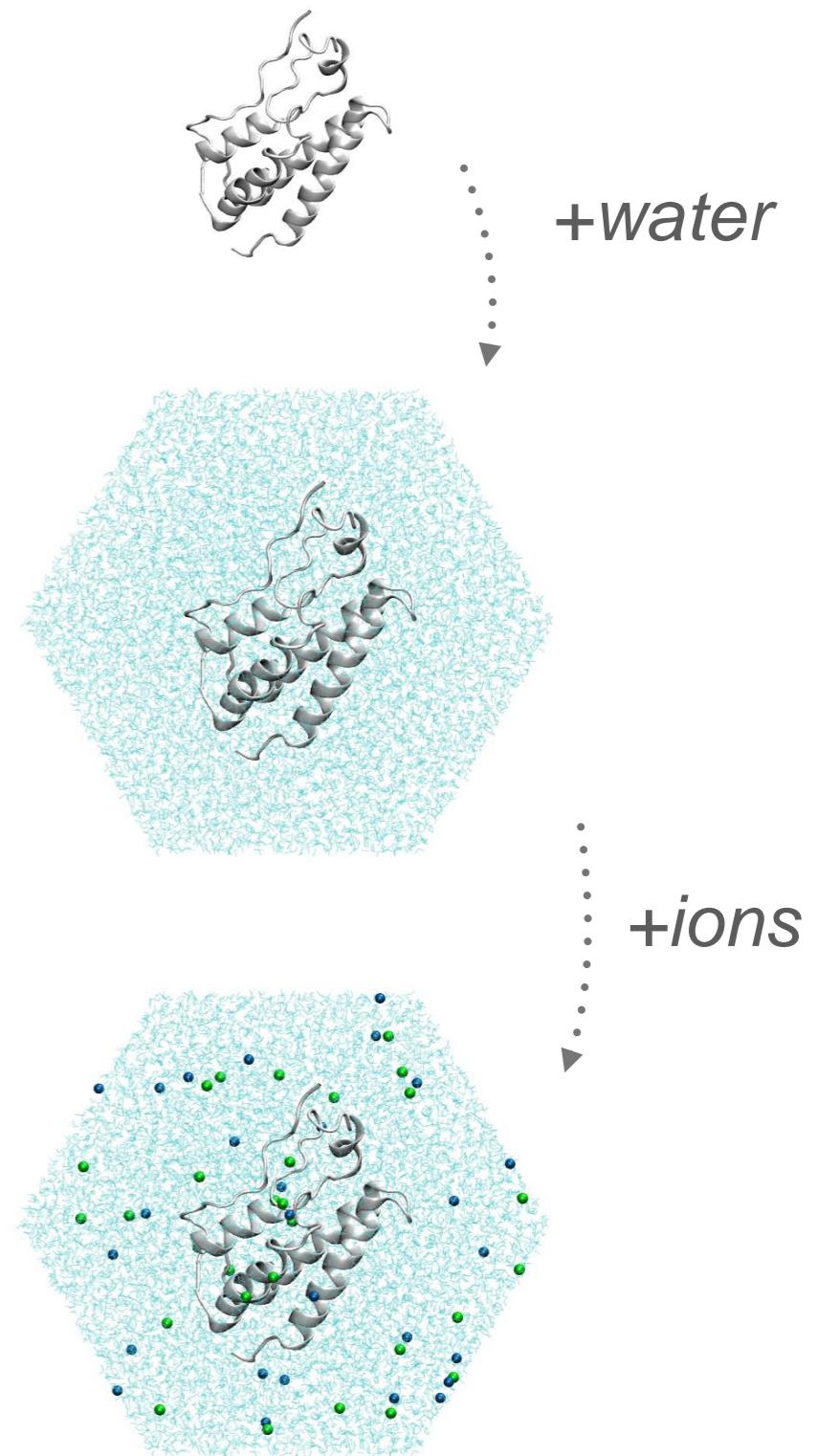
SYSTEM PREPARATION

- Get a protein **structure** from x-ray or NMR (or homology model);
- Fix structure; if there are **missing atoms/residues** they will need to be modelled (e.g. WHATIF server);
- Add hydrogens; **protonations states** for certain residues such as histidines, or small molecules, might be tricky and pka calculations needed;
- Pick the force field to use;



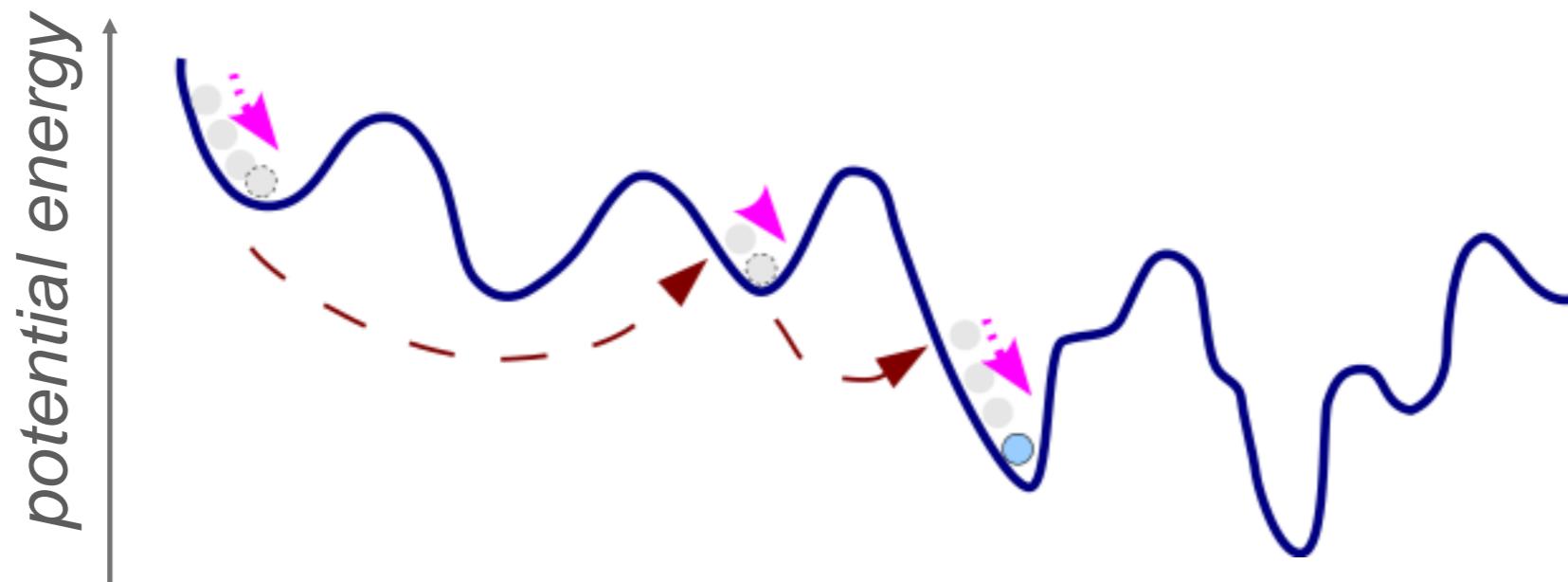
SYSTEM PREPARATION

- Choose the **box type**, e.g. cubic, octahedron, dodecahedron;
- Add other molecular species if needed, e.g. drugs, co-solvents, lipid bilayer;
- Add **solvent**, typically water; choose from a number of existing models (e.g. SPC, TIP3P, TIP4P);
- Add a concentration of **ions** (e.g. NaCl) and make the box neutral*.



*condition dictated by PME

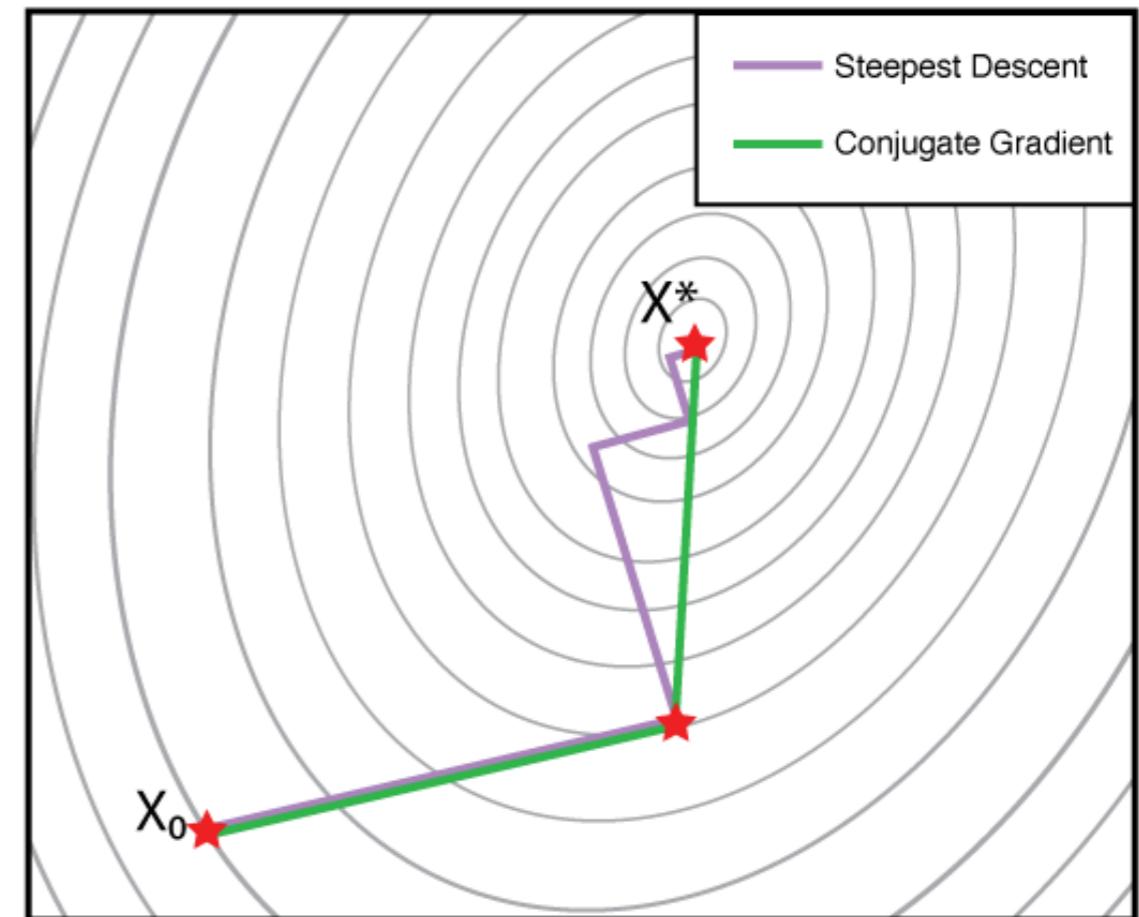
MINIMISATION



- ▶ Optimise the structure in order to **resolve potential clashes**;
- ▶ The system goes to a minimum of the potential energy surface so that it is in a stable state; this is a good starting point for a simulation so that we avoid large forces on some atoms that might blow-up the system;
- ▶ You need to decide when to stop it (tolerance), otherwise the minimisation will keep going on and on;

MINIMISATION TECHNIQUES

- Steepest descent:
 - Follows gradient downhill;
 - Poor convergence near minima (oscillatory behaviour);
- Conjugate gradient:
 - Follows gradient downhill but takes into account the direction and gradient of the previous steps;
- Newton-Raphson:
 - Uses first and second derivatives to find a minimum;
 - More computationally expensive but also more accurate



EQUILIBRATION AND PRODUCTION MD

- After minimisation, depending on your system you might want to run a short MD in order to equilibrate things like pressure, temperature, distribution of water molecules around the protein etc.;
- You can then start the MD simulation you will use for analysis.

ANALYSIS AND ERRORS

- First, simple thing to do is to **look at the trajectory**; is there anything obvious problem? e.g. is the protein unfolding;
- Test if the simulation is at **equilibrium**: is there a drift in the C-alpha RMSD, or energy;
- Check if distribution of thermodynamic properties is Gaussian;
- When taking averages it is important to estimate the uncertainty too; the block method helps to get an **error estimate**;
 - Stratified systematic/random sampling: divide the simulation in blocks and pick one value from each block;
 - Coarse-graining: use the blocks to obtain an average of averages;
- Time series: take into account **autocorrelation** for the definition of the blocks;

FURTHER READING

- M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation*, 2010
- H.J.C. Berendsen, *Simulating the Physical World*, 2007
- D. Frenkel & B. Smit, *Understanding Molecular Simulation*, 2nd edition, 2002
- M.P. Allen & D.J. Tildesley, *Computer Simulation of Liquids*, 2nd edition, 2017
- A.R. Leach, *Molecular Modelling: Principles and Applications*, 2nd edition, 2001
- W.F. van Gunsteren & A E. Mark(1998) *Validation of molecular dynamics simulation*. J. Chem. Phys. 108, 6109-6116
- M. Karplus & J.A. McCammon (2002) *Molecular Dynamics Simulations of Biomolecules*. Nat. Struct. Biol. 9,646-52