Biomolecular modeling

Marcus Elstner and Tomáš Kubař

Theoretical Chemical Biology, Karlsruhe Institute of Technology

(Dated: November 25, 2011)

Contents

VIII. ANALYSIS OF THE SIMULATION

A. Thermodynamic properties

As explained in detail earlier, we are able to obtain *time averages* of thermodynamic quantites from MD simulation. As long as the simulation is ergodic, these correspond to the *ensemble averages*, which are the values observed (in an experiment).

Some quantities may be evaluated directly, like the total (internal) energy:

$$U = \langle E \rangle_t \tag{VIII.1}$$

An interesting point is that the magnitude of *fluctuations* of certain quantities determines other thermodynamic properties of interest. So, the isochoric *heat capacity* is given by the variance of total energy:

$$C_V = \left(\frac{\partial U}{\partial T} \right)_V = \frac{\sigma_E^2}{k_\mathrm{B} T^2} = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_\mathrm{B} T^2} \tag{VIII.2}$$

Using this expression, we can obtain the heat capacity of the system in a very elegant way from a single NVT simulation at a given temperature.

B. Structural data

1. Single molecule immersed in solvent

In the area of biomolecular simulation, we usually deal with a single solute molecule (protein, DNA) immersed in solvent. The solute molecule is then the central object while our interest in the solvent is merely secondary. In such a case, we obviously wish to characterize the structure of the solute.

A common way to do so is to calculate the *average structure* of the molecule. The coordinates of every atom $\vec{r}_i$ are obtained as the arithmetic mean from the snapshots n saved along the MD trajectory:

$$\vec{r}_i = \frac{1}{N} \sum_{n=0}^{N} \vec{r}_i^{(n)} \tag{VIII.3}$$

This is a very clear and simple concept, which often yields a reasonable result. However, it may be problematic in some situations.

Imagine there are freely rotatable single bonds in the molecule, e.g. methyl groups in thymine (DNA) or in alifatic side chains (proteins). Then, be averaging of the coordinates, all three hydrogens of the methyl groups collapse to a single point, due to the free rotation of the group. This is just a minor issue; (not only) for this reason, the hydrogen atoms are usually excluded from the structure analysis, which is the restricted to the *heavy atoms* (C, N, O etc.).

A more serious issue would come up if the entire molecule rotated in the box from the initial orientation to another, in the course of the simulation. Then, the average structure of the molecule would be complete nonsense. To remedy this, the calculation of average structure usually involves the *fitting* of every snapshot to a reference structure[1] – the molecule (regarded as a rigid body) is translated and rotated so that its *RMS deviation* from the reference structure is minimized. Not until then are the coordinates taken to the sum in Eq. VIII.3.

The most unfavorable situation comes up if the molecule does not oscillate around a single structure. This may happen if the free energy surface (FES) features several available minima, which correspond to different structures. Then, the molecule will assume all these structures for certain periods of time, and the averaging of coordinates will most likely lead to an absurd structure, not corresponding to any of the minima on the FES. In such a case, it may be desirable to perform the averaging of structure on separate intervals of the trajectory, where the individual minima are being sampled.

The average structure of the molecule provides valuable information, however of an inherently static character. The development of the structure in time may be followed very simply by evaluating the *root mean square deviation* (RMSD)

$$\mathrm{RMSD}^2 = \frac{1}{N}\sum_{i=1}^{N}\left|\vec{r}_i(t) - \vec{r}_i^{\,\mathrm{ref}}\right|^2 \qquad \text{(VIII.4)}$$

of the structure in time t with respect to a reference structure; this may be the starting structure, the average structure or any other meaningful geometry of interest.[2]

Another similar quantity is the *root mean square fluctuation* (RMSF) of atomic positions,

[1] The starting structure may be taken as the reference.

[2] For instance, we may wish to compare the structure of a peptide to the idealized geometries of α-helix and β-sheet, or that of DNA to the idealized A-DNA and B-DNA.

or in other words the square of mean *variance* of atomic positions

$$\mathrm{RMSF}_i^2 = \left\langle |\vec{r}_i - \langle \vec{r}_i \rangle|^2 \right\rangle \tag{VIII.5}$$

for the atom i. This value tells us how vigorously the position of every individual atom fluctuates. RMSF may be converted to the so-called *B-factor*, which is an observable quantity in diffraction experiments (X-ray etc.):

$$B_i = \frac{8}{3}\pi^2 \cdot \mathrm{RMSF}_i^2 \tag{VIII.6}$$

Typically, the structure files deposited in the PDB contain these B-factors for all atoms. However, the comparison of B-factors obtained from a simulation with those from diffraction experiments may not be quite straightforward, as the simulation parameters and the experimental conditions may differ largely.

It is worth mentioning several further means of structure analysis which are used above all in the studies of proteins. It is possible to measure simply the distances of the individual amino-acid residues, represented for instance by their centers of mass or by the C^α atoms. This way, a *distance matrix* is constructed, which may be either time-dependent or averaged over the simulation. Distance matrices found their use in bioinformatics, and various tool have been developed for their analysis.

A classical means of analysis of protein structure is the *Ramachandran plot* – a two-dimensional histogram of dihedral angles ϕ and ψ along the protein backbone. Simulation programs usually contain tools to generate Ramachandran plots automatically.

2. Fluids

If we wish to describe the structure of a fluid (liquid or gas), for example pure argon or water, we will have to make use of another concept. Rather than one prominent molecule, we have many molecules in the system which are all equally important.

A useful way to describe the structure of such a system are the *radial distribution functions*. These describe how the molecular (or atomic) density varies as a function of the distance from one particular molecule (or atom). Consider a spherical shell of thickness δr at a distance r from a chosen atom; the volume of the shell is given by

$$\delta V \approx 4\pi r^2 \cdot \delta r \tag{VIII.7}$$

We count the number of molecules (atoms) present in this shell n, and divide this by δV to obtain a kind of 'local density' at the distance r. The *pair distribution function* $g(r)$ is then obtained by dividing by the ideal-gas distribution (which is the macroscopic density):

$$g(r) = \frac{n/\delta V}{\rho} \tag{VIII.8}$$

$g(r)$ is a dimensionless number which determines how likely it is to find a molecule (atom) in the distance of r from the reference particle, compared to the homogeneous distribution in the ideal gas.

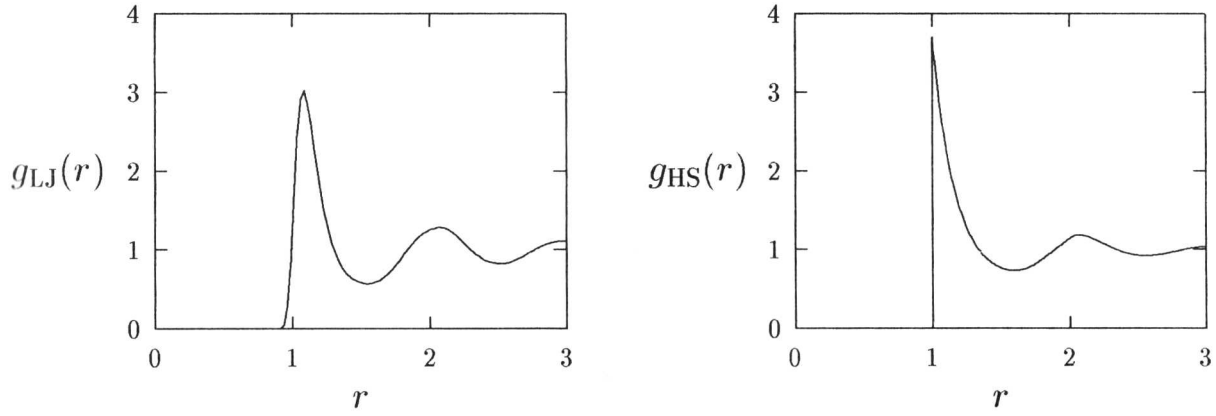


FIG. 1: Radial distribution function for a Lennard-Jones fluid near the triple point (left) and for a hard-sphere fluid (right). Reprinted from Nezbeda, Kolafa and Kotrla 1998.

A typical example of $g(r)$ for liquid water as well as hard spheres is shown in Fig. 1. The function vanishes on short distances, as the molecules cannot intersect. A high peak follows on roughly the van der Walls radius, where the interaction of molecules is favorable.[3] In other words, it is much more likely to find two molecules on this distance in a real liquid than in the ideal gas. On longer distances, several shallow minima and maxima are found, and $g(r)$ converges to unity at large distances – there, the probability of finding a particle is uniform, the same as in the ideal gas.

The importance of radial distribution functions consists not only in the information about the structure. If the pairwise additivity of forces is assumed, then thermodynamic properties can be calculated using $g(r)$ and the potential energy $u(r)$ and force $f(r)$ of a pair of particles.

[3] However, such a peak would be present in the case of hard spheres (which do not feature any attractive interaction) as well.

For example, the corrections to the ideal-gas values of total energy and pressure follow as

$$E - \frac{3}{2} N k_{\mathrm{B}} T \;=\; 2\pi N \rho \int_0^\infty r^2\, u(r)\, g(r)\, \mathrm{d}r \tag{VIII.9}$$

$$P - \rho\, k_{\mathrm{B}} T \;=\; -\frac{2\pi}{3} \rho^2 \int_0^\infty r^3 f(r)\, g(r)\, \mathrm{d}r \tag{VIII.10}$$

The Fourier transform of the pair distribution function is the *structure factor*, which may be measured in diffraction experiments (X-ray of neutron diffraction).

C. Monitoring the equilibration

Every simulation aimed at producing structural and/or thermodynamic data has to be performed in the *termodynamic equilibrium*. Therefore, the production simulation shall always by preceded by an *equilibration* run, in order to provide the system a possibility to achieve the equilibrium. The equilibration should proceed until the values of certain monitored properties become stable, i.e. until these does not exhibit a drift any more.

It is convenient to monitor the thermodynamic properties that are being evaluated and written out by the simulation program. These are usually the potential energy and the temperature; in case of NPT simulations, the pressure or the density should also be taken into account.

Apart from the thermodynamics, the structure of the system must be taken care of. Many simulations of the liquid state are being started from a configuration that exhibits some artifical regularity, like that of the crystal lattice.[4] This makes also the thermodynamics wrong, because the artifical regularity causes the entropy to be too small. Anyway, the equilibration must continue until such structural regularities are washed out. To guarantee this, we need appropriate quantities to characterize the regularity of the structure.

A measure of translational order/disorder was proposed by Verlet in the form of an *order parameter* λ

$$\lambda = \frac{\lambda_x + \lambda_y + \lambda_z}{3}, \qquad \lambda_x = \frac{1}{N} \sum_{i=1}^{N} \cos\left[\frac{4\pi x_i}{a}\right] \quad \text{etc.} \tag{VIII.11}$$

where a is the length of the edge of the unit cell. In the ideal crystal, λ assumes the value of one, while it drops to zero for a completely disordered structure. Thus, in an equilibration,

[4] Note that we usually fill the simulation box with water in the form of small and *identical* 'bricks'.

one should see λ to decrease to zero and then fluctuate around zero.

Another useful quantity may be the *mean squared displacement* (MSD) given by

$$\mathrm{MSD} = \frac{1}{N}\sum_{i=1}^{N} |\vec{r}_i(t) - \vec{r}_i(0)|^2 \tag{VIII.12}$$

which should increase gradually with time in a fluid with no specific molecular structure, whereas it would oscillate about a mean value for a solid.

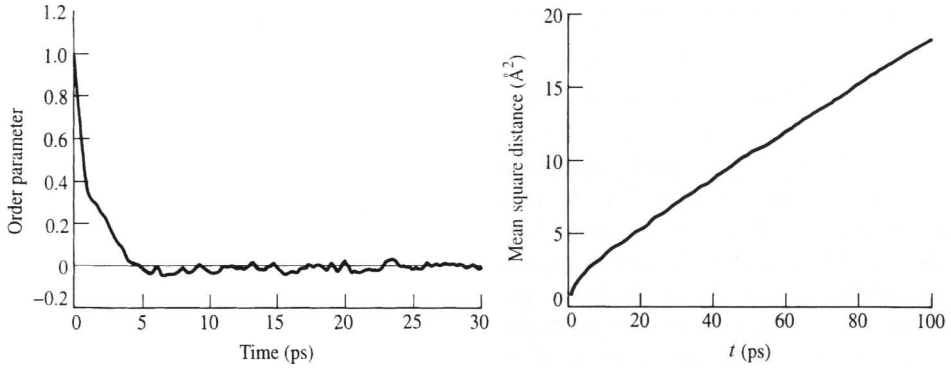


FIG. 2: The course of equilibration of liquid argon being followed by the Verlet order parameter (left) and the mean squared displacement (right). Reprinted from Leach: Molecular Modelling.

D. Time-dependent properties

1. *Correlation functions*

Suppose there are two physical quantities x and y, which may exhibit some *correlation*. This term indicates a relation of the quantities, opposed to *independence*. To quantify correlation, several kinds of *correlation functions* or *correlation coefficients* have been developed. Most common are the Pearson correlation coefficients, which describe the potential *linear* relationship between the quantities.

Typically, we consider two quantities fluctuating around their mean values $\langle x\rangle$ and $\langle y\rangle$. Then, it is of advantage to consider only the fluctuating part and introduce a correlation coefficient ρ_{xy}

$$\rho_{xy} = \frac{\langle (x - \langle x\rangle)\cdot(y - \langle y\rangle)\rangle}{\sqrt{\langle (x - \langle x\rangle)^2\rangle \cdot \langle (y - \langle y\rangle)^2\rangle}} = \frac{\mathrm{cov}(x,y)}{\sigma_x \cdot \sigma_y} \tag{VIII.13}$$

where $\mathrm{cov}(x,y)$ stands for the *covariance* of x and y, which is the generalization of variance.

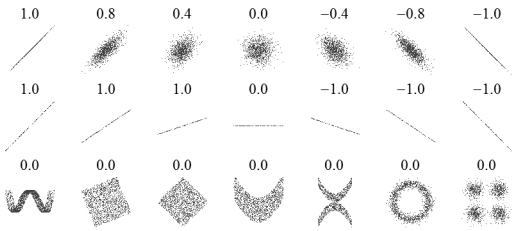


FIG. 3: Correlation of two quantities (on the x and y axes) and the coresponding correlation coefficients. Downloaded from WIKIPEDIA.

In an MD simulation, we obtain the values of various properties at specific times. It can happen at some point in time, that the value of a property x is correlated with the value of the same property at an earlier time point. This behavior may be described by the *autocorrelation function* (ACF) of this property

$$c_x(t) = \frac{\langle x(t) \cdot x(0) \rangle}{\langle x(0) \cdot x(0) \rangle} = \frac{\int x(t')\, x(t'+t)\, \mathrm{d}t'}{\int x^2(t')\, \mathrm{d}t'} \tag{VIII.14}$$

which denotes the correlation of the same property x at two time points separated by t, and the denominator $\langle x(0) \cdot x(0) \rangle$ normalizes c_x so that it takes values between -1 and 1.

2. Autocorrelation of velocity

The autocorrelation function indicates, to which extent the system retains a 'memory' of its previous values, or conversely, how quickly it takes for the system to 'forget' its previous state. A useful example is the *velocity autocorrelation function*, which tells us how closely the velocities of atoms at a time point t resemble those at a time 0. It is a good idea to average the ACF of velocity over all atoms i in the simulation:

$$c_v(t) = \frac{1}{N} \sum_{i=1}^{N} \frac{\langle \vec{v}_i(t) \cdot \vec{v}_i(0) \rangle}{\langle \vec{v}_i(0) \cdot \vec{v}_i(0) \rangle} \tag{VIII.15}$$

Typical ACF starts at the value of one in $t = 0$ and decreases afterwards. The time needed for the system to lose the autocorrelation of the quantity (velocity) whatsoever is often called *correlation time* or *relaxation time* τ_v:

$$\tau_v = \int_0^\infty c_v(t)\, \mathrm{d}t \tag{VIII.16}$$

There is a statistical issue related to the evaluation of properties of interest. In order to obtain correct average values of properties related to velocity (i.e. dynamical properties),

it is necessary to calculate the average of *uncorrelated* values. And now, the longer the relaxation time is, the fewer values can we take from the simulation of a certain length, to obtain correct averages. On the other hand, if the quantity (velocity) has short relaxation time, then it is possible to take many values for averaging.

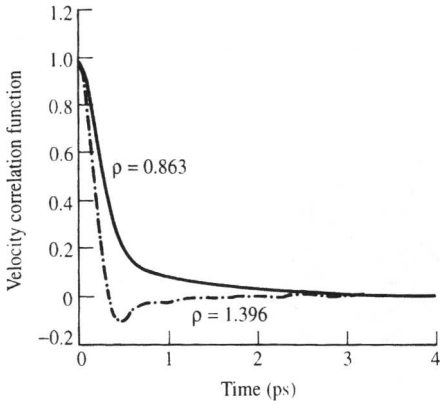


FIG. 4: Velocity autocorrelation functions for liquid argon (densities in g·cm^{-3}). Reprinted from Leach: Molecular Modelling.

Fig. 4 shows the velocity ACF from the simulations of a liquid at two different densities. At lower density, the ACF decreases gradually to zero. Unlike that, at higher density, the ACF comes faster to zero and even assumes negative values for a period of time. This means that the velocities point in the direction opposite to that at $t = 0$, which can be interpreted by the concept of a 'cage' structure of the liquid. The initial decay of ACF is slower than predicted by the kinetic theory, and this result together with its (slightly complex) explanation represents one of the most interesting achievements of early simulations.

There is a quite straightforward connection between the velocity ACF and the *transport properties* of the system. One of the Green–Kubo relations expresses the *self-diffusion coefficient* D by using the integral of the velocity ACF:[5]

$$D = \frac{1}{3}\int_0^\infty \langle \vec{v}_i(t) \cdot \vec{v}_i(0) \rangle_i \, \mathrm{d}t \qquad \text{(VIII.17)}$$

Diffusion coefficients are very interesting observable quantities, and it is an important point that we are able to obtain them from MD simulations. Interestingly, D may be obtained from another property easily accesible in the simulation – the mean squared displacement

[5] Recall Fick's laws of diffusion for flux J and concentration ϕ: $J = -D\frac{\partial \phi}{\partial x}$, $\frac{\partial \phi}{\partial t} = D\frac{\partial^2 \phi}{\partial x^2}$

(see Eq. VIII.12). The respective Einstein relation reads

$$D = \frac{1}{6} \lim_{t\to\infty} \frac{\left\langle |\vec{r}_i(t) - \vec{r}_i(0)|^2 \right\rangle_i}{t} \tag{VIII.18}$$

3. Autocorrelation of dipole moment

Velocity is an example of a property of a single atom. Contrary to that, there are quantities that need to be evaluated for the entire molecular system. Such a property of the system is the *total dipole moment*, which is the sum of the dipole moments of all individual molecules i in the system:

$$\vec{\mu}_{\text{tot}}(t) = \sum_{i=1}^{N} \vec{\mu}_i(t) \tag{VIII.19}$$

The ACF of total dipole moment is given as

$$c_\mu(t) = \frac{\langle \vec{\mu}_{\text{tot}}(t) \cdot \vec{\mu}_{\text{tot}}(0) \rangle}{\langle \vec{\mu}_{\text{tot}}(0) \cdot \vec{\mu}_{\text{tot}}(0) \rangle} \tag{VIII.20}$$

This quantity is very significant because it is related to the vibrational spectrum of the sample. Indeed, it is possible to obtain the infrared spectrum as the Fourier transform of the dipolar ACF. An example is presented in Fig. 5. Rather than sharp peaks at well-defined frequencies (as is the case of molecules in the gas phase), we see continuous bands, as the liquid absorbs at many frequencies in a broad interval. The frequencies correspond to the rate at which the total dipole moment is changing.

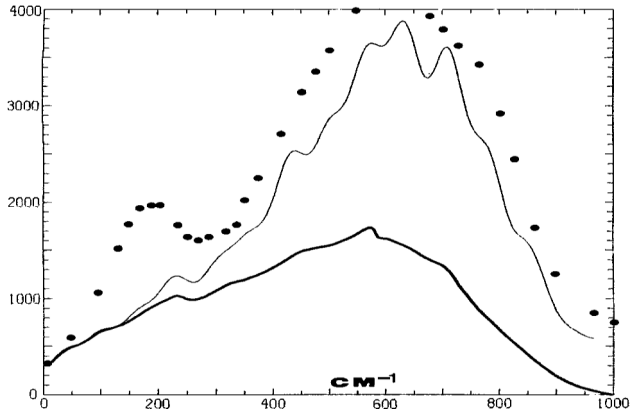


FIG. 5: Infrared spectra for liquid water. Black dots – experiment; thick curve – result from classical MD; thin curve – result with quantum corrections. B. Guillot, J. Phys. Chem. 1991.

4. Principal component analysis

It is possible to perform covariance analysis on the atomic coordinates in MD snapshots. This *principal component analysis* (PCA), also called *essential dynamics* uses the symmetric $3N$-dimensional covariance matrix C of the atomic coordinates $r_i \in \{x_i, y_i, z_i\}$:

$$C_{ij} = \langle (r_i - \langle r_i \rangle) \cdot (r_j - \langle r_j \rangle) \rangle_t \qquad \text{or} \tag{VIII.21}$$

$$C_{ij} = \left\langle \sqrt{m_i}(r_i - \langle r_i \rangle) \cdot \sqrt{m_j}(r_j - \langle r_j \rangle) \right\rangle_t \tag{VIII.22}$$

The latter definition is mass-weighted, with m_i being the masses of the respective atoms.

Standard diagonalization techniques can be used to obtain the eigenvalues and eigenvectors of this matrix. The eigenvectors the correspond to the principal or essential modes of motion of the system, an analogy of the normal modes; the respective eigenvalues may be expressed in terms of quasi-harmonic frequencies of these modes.

The first few eigenvectors with the largest eigenvalues (and thus the lowest frequencies of as little as 1 cm^{-1}) usually correspond to global, collective motions in which many atoms are involved. In the example of double-stranded DNA, the three weakest modes (see Fig. 6) are the same as would be expected for a simple rod made of a flexible material – two bending modes around axes perpendicular to the principal axis of the DNA, and a twisting mode.

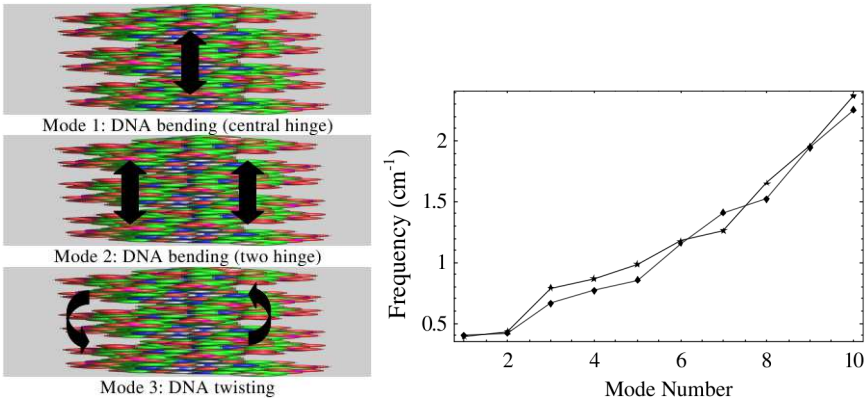


FIG. 6: First three principal modes of motion of double-stranded DNA (left) and their frequencies for two different sequences (right). Reprinted from S. A. Harris, J. Phys. Condens. Matter 2007.

Not only does this analysis give us an idea of what the modes of motion look like, it can also be used in thermodynamic calculations. The obtained vibrational frequencies may be used to evaluate *configurational entropy* of the molecule, which is otherwise hardly accessible.

E. Appendix – Fourier transform

The Fourier transform (FT) is an operation that transforms one function of a real variable into another. In such applications as signal processing, the domain of the original function is typically time and is accordingly called the time domain. That of the new function is frequency, and so the FT is often called the 'frequency domain representation of the original function.' It describes *which frequencies* are present in the original function. In effect, the Fourier transform decomposes a function into oscillatory functions.[6]

FT of a function $f(x)$ in the domain of frequency ω is given by the expression

$$F(\omega) = \int_{-\infty}^{\infty} f(x) \cdot \exp\left[-\mathrm{i}\,\omega x\right]\,\mathrm{d}x \tag{VIII.23}$$

where the connection to oscillatory functions is evident by noting that

$$\exp\left[-\mathrm{i}\,\omega x\right] = \cos\left[\omega x\right] - \mathrm{i}\sin\left[\omega x\right] \tag{VIII.24}$$

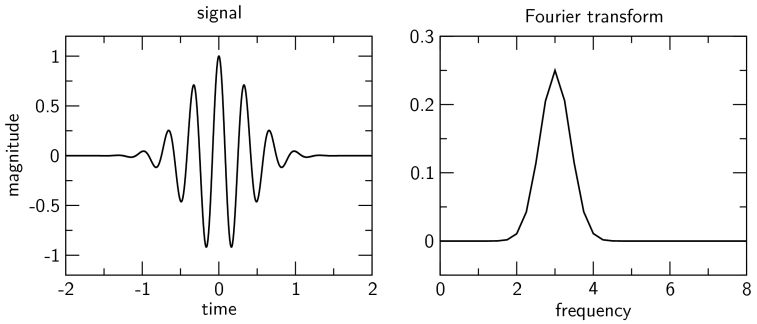


FIG. 7: Example of the Fourier transform (right) of a signal periodic on a time interval (left).

F. Exercises

- What does the radial distribution function of the ideal gas look like?
- What does the radial distribution function of an ideal crystal look like?

[6] The term Fourier transform refers both to the frequency domain representation of a function and to the process or formula that "transforms" one function into the other.

IX. FREE ENERGY SIMULATIONS

When searching for a physical quantity that is of most interest in chemistry, we could hardly find anything more appropriate than free energies – Helmholtz F or Gibbs G. Truly, these represent the holy grail of computational chemistry, both for their importance and because they are difficult to calculate.

These difficulties were hinted at in one of previous chapters. Recall that we can write

$$F = k_{\mathrm{B}}T \ln \iint \exp[\beta E(\vec{r},\vec{p})] \cdot \rho(\vec{r},\vec{p})\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p} + c \tag{IX.1}$$

The problem is that the large energy values (far from the minimum of energy) enter an exponential term, so that these high-energy regions may contribute significantly to the free energy F. So, in a simulation, if we have too few points in these high-energy regions of the phase space (*undersampling*), we may introduce sizeable errors in the calculated averages.

There are two fundamental approaches to overcome this difficulty: *free energy perturbation* and *thermodynamic integration*. Also, several computational tricks may be used for particular types of reactions, like *alchemical simulations* or *umbrella sampling*. An important observation is that it is not necessary to find the absolute value of the free energy. When considering a chemical reaction,[7] it is important to know merely the *free energy difference* (ΔF, ΔG) between the involved states (reactant A and product B).

A. Free energy perturbation (FEP)

For these states with energies $E_A(\vec{r},\vec{p})$ and $E_B(\vec{r},\vec{p})$, and partition functions Q_A and Q_B, free energy difference may be derived as

$$\begin{aligned}
\Delta F &= F_B - F_A = -k_{\mathrm{B}}T \ln \frac{Q_B}{Q_A} = -k_{\mathrm{B}}T \ln \frac{\iint \exp[-\beta E_B]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}}{\iint \exp[-\beta E_A]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}} \\
&= -k_{\mathrm{B}}T \ln \frac{\iint \exp[-\beta E_B]\ \exp[\beta E_A] \exp[-\beta E_A]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}}{\iint \exp[-\beta E_A]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}} \\
&= -k_{\mathrm{B}}T \ln \iint \exp[-\beta E_B] \exp[\beta E_A] \cdot \rho_A(\vec{r},\vec{p})\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p} \\
&= -k_{\mathrm{B}}T \ln \iint \exp[-\beta (E_B - E_A)] \cdot \rho_A(\vec{r},\vec{p})\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}
\end{aligned} \tag{IX.2}$$

[7] in a very general sense of a reaction that need not involve chemical bonds being created or broken – ligand binding a protein, passage of a molecule through membrane, or protein folding are reactions as well

The integral has the form of an average of a property S taken with the phase space density of state A

$$\langle S \rangle_A = \iint S(\vec{r}, \vec{p}) \cdot \rho_A(\vec{r}, \vec{p}) \, \mathrm{d}\vec{r} \, \mathrm{d}\vec{p} \tag{IX.3}$$

and so we can write equivalently

$$\begin{aligned} \Delta F(A \to B) &= -k_\mathrm{B} T \ln \langle \exp[-\beta(E_B - E_A)] \rangle_A \\ \Delta F(B \to A) &= -k_\mathrm{B} T \ln \langle \exp[-\beta(E_A - E_B)] \rangle_B \end{aligned} \tag{IX.4}$$

which is the free energy formula by Zwanzig (1954) and the essence of the FEP method. Thus, in principle, it is possible to perform a simulation of state A and obtain the free energy by averaging the exponential of the difference of energies of states B and A, or vice versa. Practically, we start an MD in state A to get the phase space density ρ_A, and then calculate the difference between the energies of states B and A along the trajectory.

- Free energy of deprotonation of an amino acid side chain in a protein. We would run the dynamics for the protonated species, and then evaluate the energy difference between protonated and unprotonated species to get the average of $\exp[-\beta(E_B - E_A)]$. This would only work if the conformations of the protein, and above all the configuration of water molecules, sampled along the MD were very similar with both forms. Usually, this is not the case.

- The ionization of a molecule. Again, we would perform a simulation of the neutral species and evaluate the energy differences. Alas, the configuration of water would be quite different here, too, leading to a very small overlap of phase space densities.

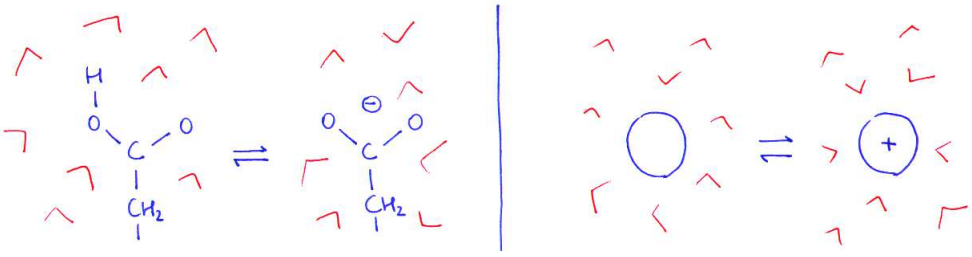


FIG. 8: Deprotonation of an amino acid (left) and ionization of a molecule (right), both in water.

Once again, let us emphasize the advantage of FEP over the direct evaluation of free energies. In the latter case, two simulations would be performed, one for each state A and

B, and the free energy difference would follow (using Eq. IX.1) as

$$\Delta F(A \to B) = k_\mathrm{B}T \ln \langle \exp[\beta E_B] \rangle_B - k_\mathrm{B}T \ln \langle \exp[\beta E_A] \rangle_A \tag{IX.5}$$

Here, note that the free energy difference is very small, of a few kcal/mol, while the total energies are very large, of hundreds or thousands kcal/mol, if the solvent or the like is included. So, we have to subtract two large numbers in order to get a small one. However, a small relative uncertainty (error) of the large values would be huge in comparison with the possibly small resulting free energy difference. Therefore, it is necessary to obtain these large values extremely accurate, which would mean the necessity to perform exceedingly long MD simulations – so long that we will never be able to afford it!

That is why we avoid performing individual simulations for the end states and rather evaluate the free energy difference directly in one simulation. Then, it is no longer necessary to sample the regions of the molecular system which do not change and are not in contact with the regions that are changing, as these do not contribute to the energy difference $E_B - E_A$. The region of phase space that has to be sampled thoroughly is much smaller, and the necessary simulation length may become feasible.

For the following, the concept of *overlap in phase space* or *overlap of phase space densities* is crucial. In a straightforward way, the more similar the states A and B are, the more similar are also the corresponding phase space densities, and they may exhibit an overlap, see Fig. 9. If the phase space densities for states A and B are similar (overlapping, Fig. 9 right), then

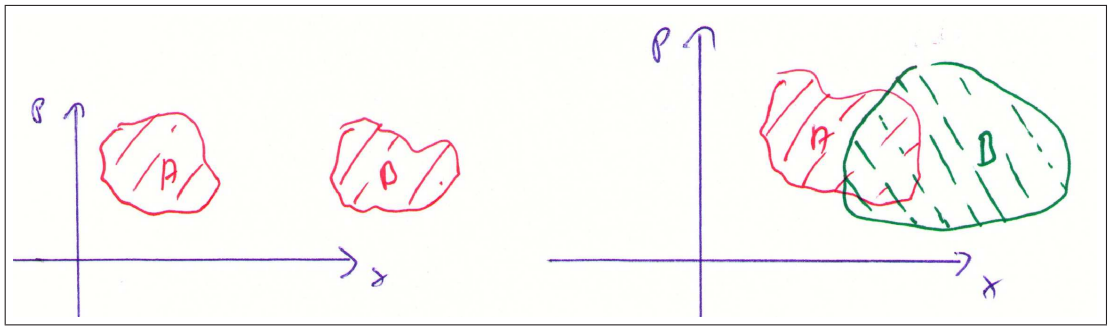


FIG. 9: Large (right) and no (left) overlap of phase space densities corresponding to two states.

the low-energy regions of state B may be sampled well even in the simulation of state A, and the free energy difference $\Delta F(A \to B)$ in Eq. IX.4 may converge. If this is not the case (like in Fig. 9 left), then the simulation of state A hardly comes to the region of phase space

where the state B has low energy; this region is undersampled, the averaging of the energy E_B is wrong, and the calculation will not converge. As a rule of thumb, this is the case if

$$|E_B - E_A| > k_\mathrm{B}T \tag{IX.6}$$

A way to overcome this problem is to insert an intermediate state (designated '1') which overlaps with both A and B, as in Fig. 10. The underlying idea is to make use of the fact

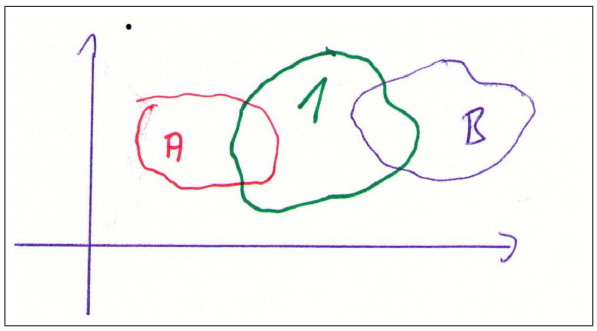


FIG. 10: Intermediate state '1' overlapping with state A and B

that free energy is a *state function*, and so

$$\Delta F(A \to B) = \Delta F(A \to 1) + \Delta F(1 \to B) \tag{IX.7}$$

Therefore, we can perform *two* MD simulations, one for each of the states A and 1, and evaluate free energies for the two subprocesses. These may be expected to converge better, and their sum gives the free energy of $A \to B$:

$$\begin{aligned} \Delta F &= -k_\mathrm{B}T \ln \left[\frac{Q_1}{Q_A} \cdot \frac{Q_B}{Q_1} \right] = \\ &= -k_\mathrm{B}T \ln \langle \exp[-\beta(E_1 - E_A)] \rangle_A - k_\mathrm{B}T \ln \langle \exp[-\beta(E_B - E_1)] \rangle_1 \end{aligned} \tag{IX.8}$$

Obviously, it is possible to insert more than one intermediate state between A and B, if these differ exceedingly. For N intermediate states $1, 2, \ldots, N$, we obtain

$$\begin{aligned} \Delta F = &- k_\mathrm{B}T \ln \left[\frac{Q_1}{Q_A} \cdot \frac{Q_2}{Q_1} \cdot \ldots \cdot \frac{Q_B}{Q_N} \right] = \\ = &- k_\mathrm{B}T \ln \langle \exp[-\beta(E_1 - E_A)] \rangle_A - k_\mathrm{B}T \ln \langle \exp[-\beta(E_2 - E_1)] \rangle_1 - \\ &- \ldots - k_\mathrm{B}T \ln \langle \exp[-\beta(E_B - E_N)] \rangle_N \end{aligned} \tag{IX.9}$$

and we have to perform $N + 1$ simulations, e.g. of states $A, 1, 2, \ldots, N$.

The description of this procedure may sound complicated, but it is implemented in the common simulation packages in a convenient way. Since we can change the chemical identities of the atoms or functional groups, this practice is often called *computational alchemy*. Typically, one introduces a parameter λ which 'converts' the force-field parameters (i.e. the Hamiltonian) from these of state A to those of state B:

$$E_\lambda = (1-\lambda)\cdot E_A + \lambda \cdot E_B \tag{IX.10}$$

- The (solvation) free energy difference of argon and xenon in aqueous solution. The two atoms differ only in the vdW parameters – the well depth ε and the radius σ. To transmutate the energy function from that of one species to the other, we interpolate:

$$\varepsilon_\lambda = (1-\lambda)\cdot \varepsilon_A + \lambda \cdot \varepsilon_B \tag{IX.11}$$

$$\sigma_\lambda = (1-\lambda)\cdot \sigma_A + \lambda \cdot \sigma_B \tag{IX.12}$$

  In the simulation, we start from $\lambda = 0$, i.e. an argon atom, and change it in subsequent steps to 1. For each step (called *window*), we perform an MD with the corresponding values of the vdW parameters, and calculate the relative free energies.

- A true chemical reaction like HCN $\rightarrow$ CNH. The situation is more complicated as we need the topologies of both molecules. Thus, a *dual-topology* simulation is performed: we have both molecules simultaneously in the simulation. These two molecules do *not* interact with each other, and we gradually switch off the interaction of one species with the solvent during the simulation while we switch on the other at the same time.

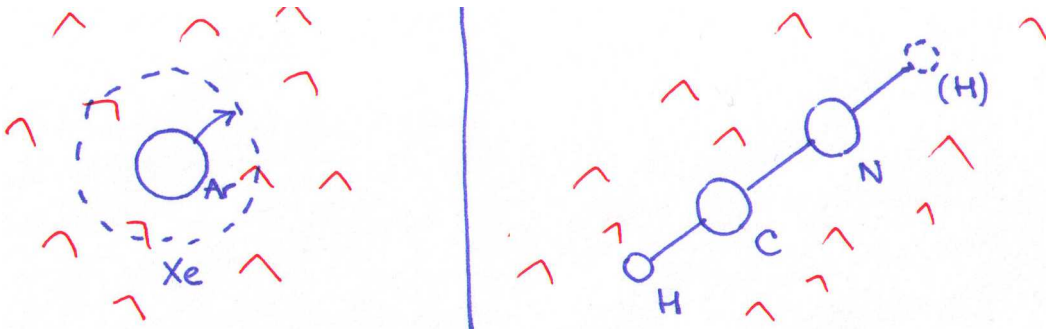


FIG. 11: Examples of 'alchemical' simulations.

B. Thermodynamic integration (TI)

In the last chapter, we have written the energy E as a function of the parameter λ. This means, that the free energy also becomes dependent on λ:

$$F = F(\lambda) \tag{IX.13}$$

with $F(0) = F(A)$ and $F(1) = F(B)$. Thus, we can write

$$\Delta F = F(B) - F(A) = \int_0^1 \frac{\partial F(\lambda)}{\partial \lambda} \mathrm{d}\lambda \tag{IX.14}$$

with

$$F(\lambda) = -k_\mathrm{B} T \ln Q(\lambda) \tag{IX.15}$$

The derivative of F rearranges to

$$\begin{aligned}
\frac{\partial F}{\partial \lambda}(\lambda) &= -k_\mathrm{B} T \frac{\partial \ln Q}{\partial \lambda}(\lambda) = -k_\mathrm{B} T \frac{1}{Q(\lambda)} \cdot \frac{\partial Q}{\partial \lambda}(\lambda) = -k_\mathrm{B} T \frac{1}{Q(\lambda)} \cdot \frac{\partial}{\partial \lambda} \iint \exp[-\beta E_\lambda] \, \mathrm{d}\vec{r} \, \mathrm{d}\vec{p} = \\
&= -k_\mathrm{B} T \frac{1}{Q(\lambda)} \cdot \iint (-\beta) \frac{\partial E_\lambda}{\partial \lambda} \exp[-\beta E_\lambda] \, \mathrm{d}\vec{r} \, \mathrm{d}\vec{p} = \\
&= -k_\mathrm{B} T \cdot (-\beta) \cdot \iint \frac{\partial E_\lambda}{\partial \lambda} \frac{\exp[-\beta E_\lambda]}{Q(\lambda)} \, \mathrm{d}\vec{r} \, \mathrm{d}\vec{p} \\
&= 1 \cdot \iint \frac{\partial E_\lambda}{\partial \lambda} \rho_\lambda(\vec{r}, \vec{p}) \, \mathrm{d}\vec{r} \, \mathrm{d}\vec{p} = \left\langle \frac{\partial E_\lambda}{\partial \lambda} \right\rangle_\lambda
\end{aligned} \tag{IX.16}$$

This is the essence of TI – the derivative of free energy F with respect to the coupling parameter λ is calculated as the average of derivative of total MM energy E, which can be directly evaluated in the simulation. Then, the free energy difference follows simply as

$$\Delta F = \int_0^1 \left\langle \frac{\partial E_\lambda}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{IX.17}$$

Practically, we perform a MD simulation for each chosen value of λ; it is usual to take equidistant values in the interval (0,1) like 0, 0.05,..., 0.95 and 1. Each of these simulations produces a value of $\left\langle \frac{\partial E}{\partial \lambda} \right\rangle_\lambda$, so that we obtain the derivative of free energy in discrete points for $\lambda \in (0, 1)$. This function is then integrated numerically, and the result is the desired free energy difference ΔF.

An example of the TI simulation is shown in Fig. 12. An atom of rare gas (neon) is dissolved in water; in course of the NPT simulation, the van der Walls parameters of the

neon atom are being gradually switched off by means of the λ parameter, so that the atom is effectively disappearing. The derivative of total energy with respect to λ is evaluated for several (21) values of λ ranging from 0 to 1. Eq. IX.17 is then used to obtain the (Gibbs) free energy difference of the two states: (i) a neon atom in water, and (ii) no neon atom in water, i.e. outside of the solution in vacuo. Thus, the calculated free energy difference corresponds directly to the solvation free energy, a quantity which is of considerable value in chemistry.

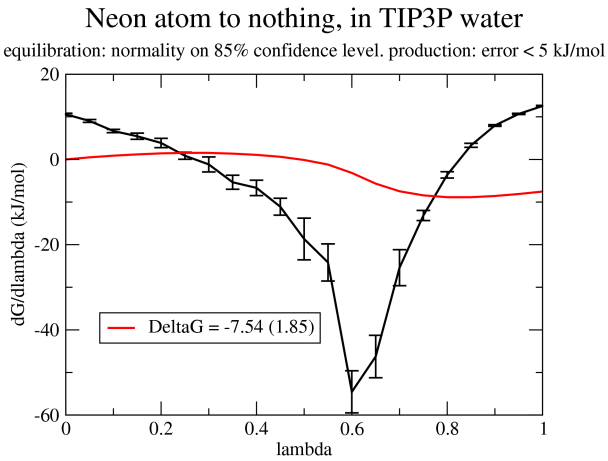


FIG. 12: TI simulation of a neon atom in water being disappeared. See text for explanation.

Finally, let us summarize the features of FEP and TI. Irrespective of the nature of the studied reaction, both FEP and TI require the introduction of a coupling parameter λ, which plays the role of the reaction coordinate with $\lambda = 0$ corresponding to the reactant and $\lambda = 1$ to the product. The fact that free energy is a state function guarantees the independence of the result on the chosen path between the reactant and the product, and so it does not matter if the reaction coordinate corresponds to an unphysical process like a change of chemical identity of one or more atoms (as is the case in the alchemical simulations).

The remaining open question regards the necessary number of windows in the simulation. We would like to have as few windows as possible, without compromising numerical precision of the calculation. In FEP, the assumption is that while simulating the state A, the low-energy regions of state B are sampled well. The closer the windows are, the better is this condition fulfilled. On the other hand, the free energy derivative is always evaluated for *one* λ-value with TI, and the problem present in FEP does not occur here. It is the numerical integration of the free energy derivative that brings on the numerical inaccuracy of TI.

C. Free energy from non-equilibrium simulations

A major disadvantage of the described methodology – TI using equilibrium simulations for discrete values of λ – is the very slow convergence of $\partial G/\partial\lambda$ once the alchemical change becomes large. So, it is often possible to describe the mutation of a single amino acid side chain in a protein provided the structure of the protein remains the same, but this should be considered a practical limit of the method.

To avoid this problem, the current development of free-energy methods makes use of *non-equilibrium simulations*. Here, the usual process of "equilibration" of the system for every of the selected values of λ followed by a "production phase" is not used; a non-equilibrium simulation consists of n MD steps, where the parameter λ starts at 0 and increases by $1/n$ in every MD step. This way, the simulation does not describe the system in equilibrium in any moment, as the external parameter λ is changing all the time. Whereas a single simulation of this kind is probably worthless, the remarkable equality by Jarzynski provides a link between an *ensemble* of such simulations and the desired free energy:

$$\exp[-\beta\Delta F] = \langle\exp[-\beta W]\rangle \tag{IX.18}$$

The true value of free energy ΔF is obtained as a special kind of ensemble average, for the ensemble of non-equilibrium TI simulations yielding "free energies" W. These values $W = \int_0^1 \partial E/\partial\lambda\,\mathrm{d}\lambda$ are no free energies whatsoever; instead, they may be called *(irreversible) work*. Since no convergence of any quantity is required within a single non-equilibrium simulation, these simulations may be very short – and this is the actual practice. However, the sampling problem persists because the largest statistical weight is carried by rarely occuring simulations (due to the unfavorable averaging in Eq. IX.18).

This sampling issue may be circumvented by *exponential work averaging* with gaussian approximation. An ensemble of simulations is performed for the 'forward' process $0 \to 1$ as well as for the 'reverse' process $1 \to 0$, and the obtained distributions of forward and backward irreversible work are approximated by gaussians with mean and standard deviation W_f, σ_f and W_r, σ_r, respectively. The free energy is calculated as an average of values

$$\begin{aligned} \Delta F_f &= W_f - \frac{1}{2}\beta\sigma_f^2 \\ \Delta F_r &= -W_r + \frac{1}{2}\beta\sigma_r^2 \end{aligned} \tag{IX.19}$$

A more general expression (than the Jarzynski equality) is the *Crooks fluctuation theorem* (CFS), according to which the distributions of forward and reverse work are related like

$$\frac{P_f(W)}{P_r(-W)} = \exp[\beta(W - \Delta F)] \tag{IX.20}$$

Then, once we have obtained well-converged distributions P_f and P_r, it is possible to apply *Bennett's acceptance ratio* for an equal number of forward and reverse simulation; the free energy follows from

$$\left\langle \frac{1}{1 + \exp[\beta(W - \Delta F)]} \right\rangle_f = \left\langle \frac{1}{1 + \exp[-\beta(W - \Delta F)]} \right\rangle_r \tag{IX.21}$$

It is possible to apply CFS more directly. A closer look at Eq. IX.20 reveals that the free energy corresponds to the value of work W for which the probabilities P_f and P_r are equal – to the intersection point of the distributions. To determine this point readily from the distributions may be difficult and a source of large errors if the overlap of the distributions is very small. Again, this issue can be solved by the assumption of normal distribution of the forward and reverse work, which was proven for a system with a large number of degrees of freedom. The procedure thus requires to perform a number of forward and reverse simulations sufficient to perform a good-quality gaussian fit to the resulting distributions of irreversible work. The free energy is calculated directly as the intersection points of these gaussian curves.

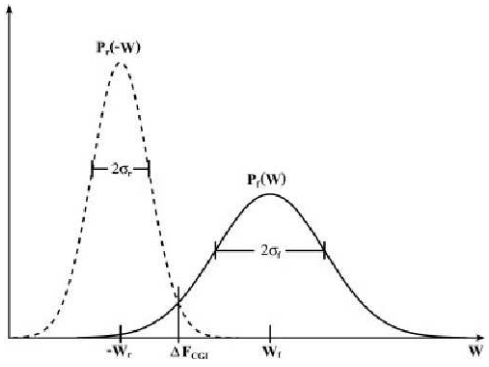


FIG. 13: The Crooks gaussian intersection (from Goette and Grubmüller 2009).

D. Thermodynamic cycles

Quite often, we are interested not in the absolute free energies and not even in the reaction free energies, but rather in the difference (Δ) of reaction free energies (ΔF) corresponding to two similar reactions. These may the be denoted as $\Delta\Delta F$ or $\Delta\Delta G$.

Consider as an example the binding of an inhibitor molecule I to an enzyme E, as shown in Fig. 14 left. Usually, we are interested in differences of binding free energies, for instance of an inhibitor I to two very similar enzymes E and E′:

$$\begin{aligned} \mathrm{E} + \mathrm{I} &\rightleftharpoons \mathrm{EI} \qquad \Delta G_1 \\ \mathrm{E}' + \mathrm{I} &\rightleftharpoons \mathrm{E}'\mathrm{I} \qquad \Delta G_2 \end{aligned} \tag{IX.22}$$

The binding of the inhibitor can induce large structural changes in the enzyme, and it would be very difficult (if not impossible) to describe this reaction in a simulation both correctly and efficiently at the same time. So, significant errors would seem to be inevitable. A way to solve this would be to simulate not the reaction of binding but rather the alchemical transmutation of enzyme E to E′. As we consider the enzymes to be very similar,[8] it is plausible to assume the structure of complexes EI and E′I to be similar as well. Then, the alchemical simulation may well be successful. As free energy is a state function, the sum of

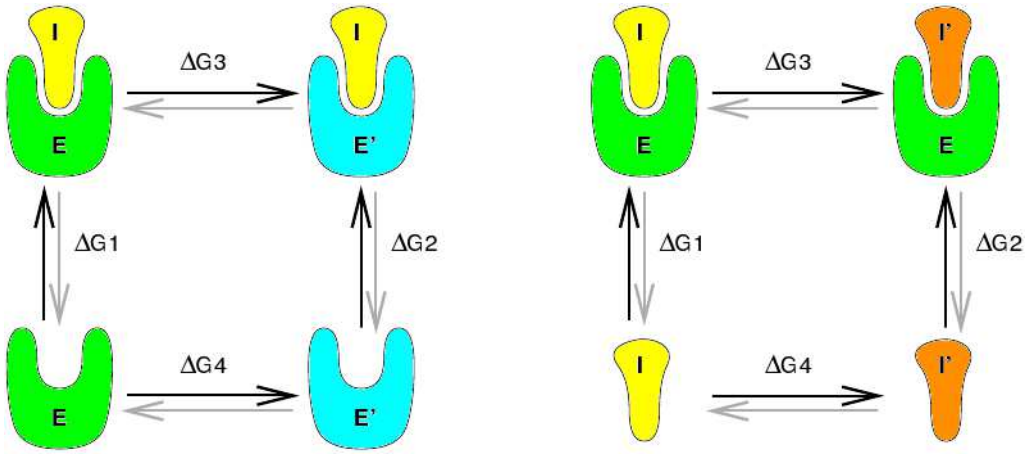


FIG. 14: Examples of the thermodynamic cycle.

free energies around a *thermodynamic cycle* vanishes (e.g. clockwise in Fig. 14 left):

$$\Delta F_1 + \Delta F_3 - \Delta F_2 - \Delta F_4 = 0 \tag{IX.23}$$

[8] Imagine E′ to be derived from E by a mutation of a single amino acid, e.g. leucine to valine.

The difference of binding free energies then follows to be equal the difference of free energies calculated in alchemical simulations:

$$\Delta\Delta F = \Delta F_1 - \Delta F_2 = \Delta F_3 - \Delta F_4 \tag{IX.24}$$

Similarly, it is possible to calculate the free energy difference of binding of two similar ligands to the same enzyme (Fig. 14 right), or the difference of solvation energy of two similar molecules. In the latter case, two alchemical simulations would be performed: one in vacuo and the other in solvent.

E. Potentials of mean force (PMF) and umbrella sampling

Sometimes, we wish to know not only the free energy difference of two states (the reactant and the product), but rather the free energy along the *reaction coordinate* q within a certain interval; the free energy is then a function of q while it is integrated over all other degrees of freedom. Such a free energy function $F(q)$ is called the *potential of mean force*. Examples of such a reaction coordinate q may be the distance between two particles if the dissociation of a complex is studied, the position of a proton for a reaction of proton transfer, or the dihedral angle when dealing with some conformational changes.

To separate the degree of freedom spanned by the reaction coordinate, we perform a coordinate transformation from $\vec{r} = (r_1, r_2, \ldots, r_{3N})$ to a set $(u_1, u_2, \ldots, u_{3N-1}, q)$, where the $(3N-1)$-dimensional vector $\vec{u}$ represents all remaining degrees of freedom, and we can write

$$\mathrm{d}\vec{r} = \mathrm{d}\vec{u} \cdot \mathrm{d}q \tag{IX.25}$$

Looking for the free energy at a certain value of q, all remaining degrees of freedom are averaged over (or 'integrated out'). One could think of performing an MD simulation and sampling *all* degrees of freedom except for q.

An example would be the free energy of formation of an ion pair in solution, as shown in Fig. 15. An MD simulation would be performed to calculate the free energy for every value of the reaction coordinate q.

The free energy is given by:

$$F = -k_\mathrm{B}T \ln \iint \exp[-\beta E(\vec{r}, \vec{p})]\, \mathrm{d}\vec{r}\, \mathrm{d}\vec{p} \tag{IX.26}$$

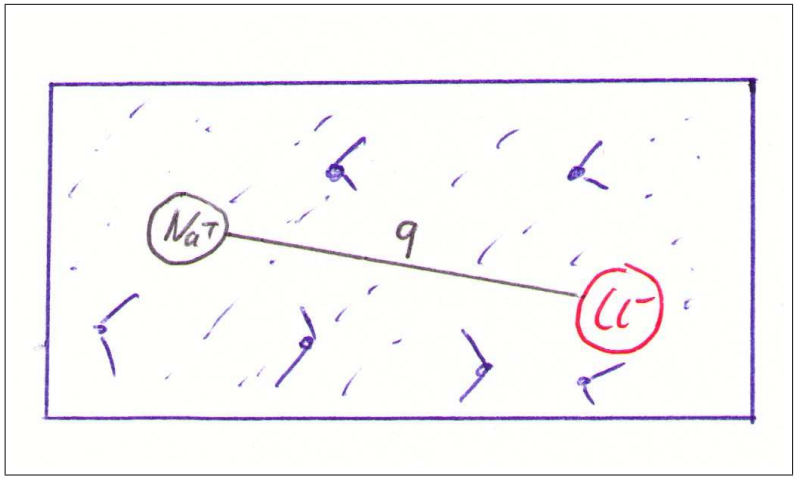


FIG. 15: Na^+ and Cl^- in water solution. The distance between the ions is the reaction coordinate q, and all other degrees of freedom (water) are represented by $\vec{u}$ and are free to vary.

If we wish to evaluate an expression for a coordinate q taking a certain value q_0, it is convenient to use the *Dirac delta function*,[9] $\delta(q - q_0)$. With that, we can write the free energy for the fixed reaction coordinate q_0 as

$$\begin{aligned} F(q_0) &= -k_\mathrm{B}T \ln \iint \delta(q - q_0) \exp[-\beta E(\vec{r}, \vec{p})] \, \mathrm{d}\vec{p} \, \mathrm{d}\vec{u} \, \mathrm{d}q \\ &= -k_\mathrm{B}T \ln \left[Q \cdot \iint \delta(q - q_0) \frac{\exp[-\beta E(\vec{r}, \vec{p})]}{Q} \, \mathrm{d}\vec{p} \, \mathrm{d}\vec{u} \, \mathrm{d}q \right] \\ &= -k_\mathrm{B}T \ln \left[Q \cdot \iint \delta(q - q_0) \cdot \rho(\vec{r}, \vec{p}) \, \mathrm{d}\vec{p} \, \mathrm{d}\vec{u} \, \mathrm{d}q \right] \\ &= -k_\mathrm{B}T \ln \left[Q \cdot \langle \delta(q - q_0) \rangle \right] \\ &= -k_\mathrm{B}T \ln Q - k_\mathrm{B}T \ln \langle \delta(q - q_0) \rangle \end{aligned} \tag{IX.27}$$

How to interpret this? $\rho(\vec{r}, \vec{p})$ is the probability, that the system is at the point $(\vec{r}, \vec{p})$. Then,

$$P(q_0) = \iint \delta(q - q_0) \cdot \rho(\vec{r}, \vec{p}) \, \mathrm{d}\vec{r} \, \mathrm{d}\vec{p} = \langle \delta(q - q_0) \rangle \tag{IX.28}$$

is the *probability* that the reaction coordinate q in the system takes the value of q_0 , because the integral proceeds over the whole phase space and the delta function 'cancels out' all points, where the reaction coordinate is *not equal* q_0! So, the integration collects all points in phase space, where the reaction coordinate has this specific value.

[9] This is a generalized function representing an infinitely sharp peak bounding unit area; $\delta(x)$ has the value of zero everywhere, except at $x = 0$ where its value is infinitely large in such a way that its integral is 1.

What would it work like in the example of the ion pair? We perform an MD simulation for the system, and then *count* how many times the reaction coordinate takes the specified value, in other words we calculate the probability $P(q_0)$ of finding the system at q_0.

Then, the free energy difference of two states A and B is:

$$\begin{aligned} F_B - F_A &= -k_\mathrm{B}T \ln Q - k_\mathrm{B}T \ln \langle \delta(q - q_B) \rangle - (-k_\mathrm{B}T \ln Q + k_\mathrm{B}T \ln \langle \delta(q - q_A) \rangle) \\ &= -k_\mathrm{B}T \ln \frac{\langle \delta(q - q_B) \rangle}{\langle \delta(q - q_A) \rangle} \\ &= -k_\mathrm{B}T \ln \frac{P(q_B)}{P(q_A)} \end{aligned} \tag{IX.29}$$

which is actually the known definition of the *equilibrium constant* $P(B)/P(A)$.

So, the task is clear: perform a MD, specify a coordinate, and then just count, how often the system is at special values of the reaction coordinate. The ratio of these numbers gives the free energy difference!

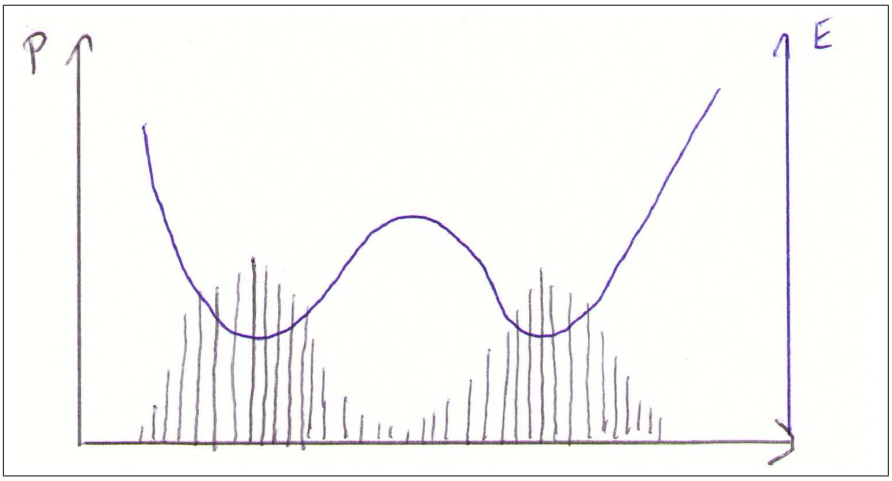


FIG. 16: Energy profile and probability distribution along the reaction coordinate. Note the undersampled region of the barrier.

This is very good, in principle. But, we also know the problem: If we there is a high barrier to be crossed along the reaction coordinate to come from A to B, a pure (*unbiased*) MD simulation will hardly make it,[10] and even if it does, the high-energy region (barrier) will be sampled quite poorly.

Then, a straightforward idea is to apply an *additional potential*, also called *biasing potential* in order to make the system spend a larger amount of time in that (those) region(s)

[10] In other words, the ergodicity of the simulation is hindered.

of phase space that would otherwise remain undersampled. This is the underlying principle of the *umbrella sampling*.[11] The additional potential shall depend only on the reaction coordinate: $V = V(q)$.[12] Then, what will the free energy look like in such a biased case? Let us start with the previously obtained expression:

$$\begin{aligned}
F(q_0) &= -k_\mathrm{B}T \ln \left[\frac{\iint \delta(q-q_0)\, \exp[-\beta E]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}}{\iint \exp[-\beta E]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}} \right] \\
&= -k_\mathrm{B}T \ln \left[\frac{\iint \delta(q-q_0) \exp[\beta V]\, \exp[-\beta(E+V)]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}}{\iint \exp[-\beta(E+V)]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}} \cdot \frac{\iint \exp[-\beta(E+V)]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}}{\iint \exp[-\beta E]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}} \right] \\
&= -k_\mathrm{B}T \ln \left[\langle \delta(q-q_0) \exp[\beta V] \rangle_{E+V} \frac{\iint \exp[-\beta(E+V)]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}}{\iint \exp[\beta V] \exp[-\beta(E+V)]\, \mathrm{d}\vec{r}\,\mathrm{d}\vec{p}} \right] \\
&= -k_\mathrm{B}T \ln \left[\langle \delta(q-q_0) \exp[\beta V] \rangle_{E+V} \frac{1}{\langle \exp[\beta V] \rangle_{E+V}} \right] \\
&= -k_\mathrm{B}T \ln \left[\exp[\beta V(q_0)]\, \langle \delta(q-q_0) \rangle_{E+V} \frac{1}{\langle \exp[\beta V] \rangle_{E+V}} \right] \\
&= -k_\mathrm{B}T \ln \langle \delta(q-q_0) \rangle_{E+V} - V(q_0) + k_\mathrm{B}T \ln \langle \exp[\beta V \rangle_{E+V} \\
&= -k_\mathrm{B}T \ln P^*(q_0) - V(q_0) + k_\mathrm{B}T \ln \langle \exp[\beta V] \rangle_{E+V}
\end{aligned} \tag{IX.30}$$

giving the free energy as function of reaction coordinate, or PMF in the form

$$F(q) = -k_\mathrm{B}T \ln P^*(q) - V(q) + K \tag{IX.31}$$

This result is very interesting: We have added an arbitrary potential $V(q)$ to our system. Now, we have to calculate the ensemble averages with the biased potential $E+V$ as indicated by $\langle\rangle_{E+V}$. We obtain the *biased probability* $P^*(q)$ of finding the system at the value of the reaction coordinate for the ensemble $E+V$, which can obviously be very different from that of the unbiased ensemble $P(q)$. Yet, we still get the right (unbiased) free energy $F(q)$, once we take the biased probability $P^*(q)$, subtract the biasing potential $V(q)$ at the value of the reaction coordinate and add the term K.

We can use this scheme efficiently, by way of moving the biasing (harmonic) potential along the reaction coordinate as shown in Fig. 17. In this case, we perform k simulations with the potentials V_k and get:

$$F(q) = -k_\mathrm{B}T \ln P^*(q) - V_k(q) + K_k \tag{IX.32}$$

[11] This should evoke the image of an interval of the reaction coordinate being covered by an umbrella.
[12] In such a case, $\langle \delta(q-q_0) \cdot \exp[\beta V] \rangle = \langle \delta(q-q_0) \cdot \exp[\beta V(q_0)] \rangle = \exp[\beta V(q_0)] \cdot \langle \delta(q-q_0) \rangle$ in the following.

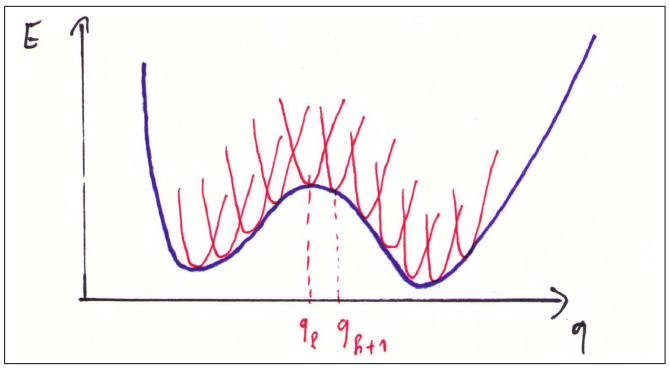


FIG. 17: Harmonic biasing potentials keep the system in the desired regions of reaction coordinate.

For each of these k simulations, we extract the probability $P^*(q)$ for every value of q and easily calculate $V^k(q)$. The curves of $-k_\mathrm{B}T \ln P^*(q) - V^k(q)$ for the simulations k and $k+1$ differ by a constant shift, which corresponds to the difference of K values, as shown in Fig. 18. The main task is to match the pieces together. One way is to fit the K_k in order

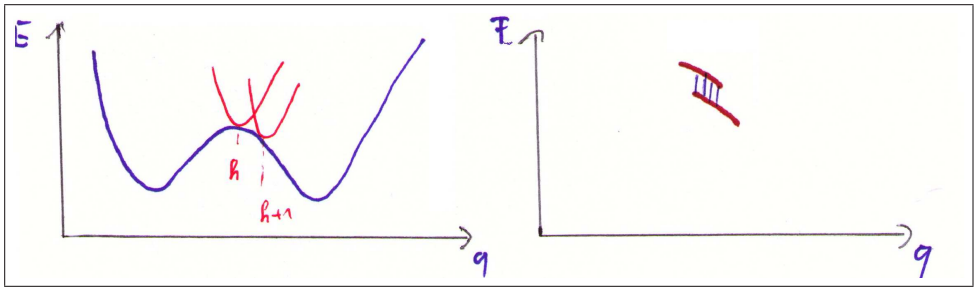


FIG. 18: The offset of free energy curves between two simulations k and $k+1$ is given by $K_k - K_{k+1}$

to get a smooth total $F(q)$ curve. This is possible if the pieces k and $k+1$ have sufficient 'overlap'.

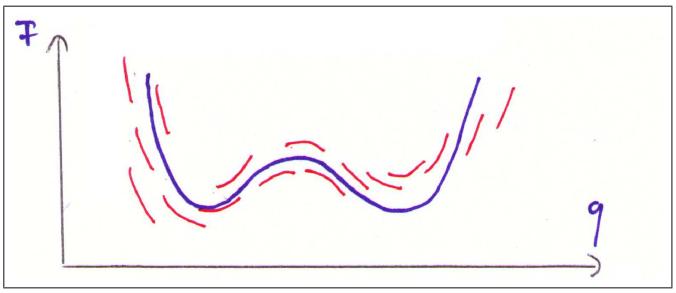


FIG. 19: Matching of histograms from different simulations

Another, quite involved method is the weighted histogram analysis method (WHAM). The starting point is the requirement of a perfect match, minimizing the total error. The

unbiased probabilities $P(x_j)$ of coordinate x falling into the bin j of the histogram and the shifts K_i are obtained by a self-consistent solution of a set of equations

$$\begin{aligned} P(x_j) &= \frac{\sum_{i=1}^{N} n_i(x_j)\ \exp[-\beta V_i(x_j)]}{\sum_{i=1}^{N} N_i\ \exp[-\beta(V_i(x_j) - K_i)]} \\ K_i &= -k\mathrm{T}\ \log \sum_{j}^{\text{bins}} P(x_j)\ \exp[-\beta V_i(x_j)] \end{aligned} \tag{IX.33}$$

(for a total of N simulations, i-th simulation contains N_i frames, $n_i(x_j)$ is the number of hits in bin j in simulation i). The WHAM procedure is included in a range of modern packages for MD simulations.

X. QM/MM

The standard force fields are designed to evaluate the energy of the system as fast as possible, and this requirement makes several quite crude approximations necessary. One of them is that the topology of the molecule remains the same in course of the simulation, meaning that the covalent bonds may be neither created nor broken in the simulation. Then, it is impossible to use such a force field to study the processes that would usually be designated as *chemical reactions*.

A. Empirical approaches to chemical reactions

In spite of the mentioned problems, it is not quite impossible to describe a chemical reaction with a force field. However, this may be done always for a single reaction, or a restricted class of reactions only, using approximations that are appropriate in the specific case; still, a generic force field applicable for *any* reaction is a mere illusion.

A possible way to describe a reaction would be as follows: An existing force field is used for all of the system, except the bonds that are being broken or created. The bonds involved in the reaction will then be re-parameterized, using probably a variant of Morse's potential or the like. Evidently, such an approach requires an *ad hoc* model of the molecule, and considerable effort is likely to be spent by the parameterization.

Also obvious are certain limitations of such an approach. The restrictions on the use of force field methods are more general than just that of the invariant connectivity of the molecules. Rather, it is the *electron density* that does not change at all. It is thus further impossible (or impracticable at the very least) to use a force field to describe a process involving *charge transfer*, in other words, the change of atomic charges. This fact poses another strong restraint on the classes of reactions that might be treated with molecular mechanics force fields. Other phenomena of interest that cannot be described with molecular mechanics, include *photochemical processes*, which involve electronic excitation of molecules.

B. The principle of hybrid QM/MM methods

Without loss of generality, we may assume that the changing electronic structure is localized in a small part of the studied molecular system. An example may be a reaction

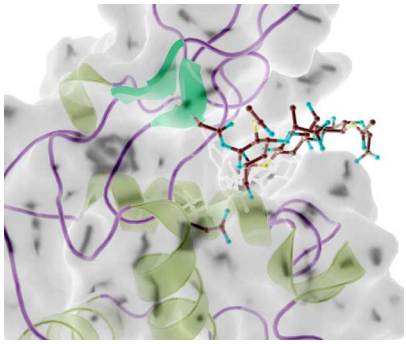


FIG. 20: Enzymatic reaction. The substrate in the binding pocket and the amino acids in contact shown as atoms; the rest of the enzyme shown as ribbons. Mulholland et al. (2008).

on a substrate which is catalyzed by an enzyme, see Fig. 20. Of the huge system, only the substrate and several atoms of the protein are involved in the reaction, while the rest of the protein and all the surrounding water and ions stay outside of the process. However, these seemingly inactive parts do interact with the substrate be means of non-bonded forces, and maintain the structure of the entire system.

So, the studied process is of quantum nature (a chemical reaction, but it may be some photochemistry as well) and thus, it must be described by a quantum chemical method. The overwhelming majority of the system (most of the enzyme and all of the water) is not directly involved in the process, but affects the reaction by way of non-bonded interactions; here, a description with an MM force field would be sufficient. It turns out to be a good idea to combine both approaches: The (small) region where the chemical reaction occurs will be described with a quantum-chemical method, while an MM force field will be used to deal with the (large) remaining part of the system, see Fig. 21. Obviously, the interaction

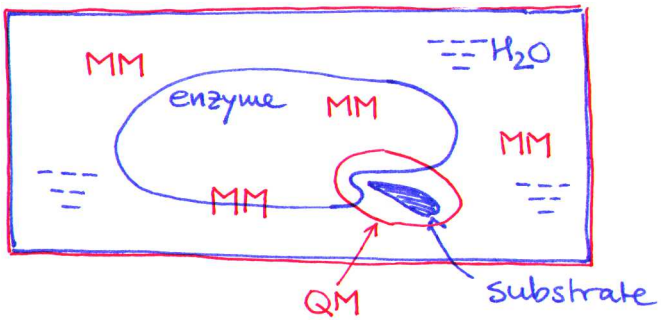


FIG. 21: QM/MM treatment of an enzymatic reaction.

of both subsystems must be taken into account correctly, as well, so that the total energy may by expressed in a simple fashion as

$$E_{\text{total}} = E_{\text{QM}} + E_{\text{MM}} + E_{\text{QM/MM}} \tag{X.1}$$

Quite a few hybrid schemes like that have been proposed so far, and they are usually called *quantum mechanics-molecular mechanics* (QM/MM) or *embedding*. These date back to a first study by Warshel and Levitt in 1976.

Both QM and MM calculations (yielding E_{QM} and E_{MM}) do not differ much from those performed on 'normal' systems not taking part in any QM/MM scheme. However, the key issue is how to treat the coupling of both parts of the system, to obtain $E_{\text{QM/MM}}$. This is the art of QM/MM calculations, and the rest of this section will deal with that topic.

C. Embedding schemes

The methods to couple the QM and MM systems differ in the excess of this coupling, or in that how large a part of this coupling is neglected. We will have a look at these methods in the order of increasing complexity (corresponding to the increasing completeness).

1. Unpolarized interactions (Mechanical embedding)

The simplest idea to account for the interactions between the QM and MM regions is to use a force field. In order to do that, atom types must be assigned to the QM atoms, because these determine the vdW parameters; further, their atomic charges must be evaluated – for instance, Mulliken charges may be used. It is then possible to calculate the QM/MM energy with the Coulomb law and the Lennard-Jones potential as

$$E_{\text{QM/MM}} = \sum_{i}^{\text{QM atoms}} \sum_{m}^{\text{MM atoms}} \left(\frac{q_i^{\text{Mull}} \cdot q_m}{r_{im}} + 4\varepsilon_{im} \left(\frac{\sigma_{im}^{12}}{r_{im}^{12}} - \frac{\sigma_{im}^{6}}{r_{im}^{6}} \right) \right) \tag{X.2}$$

where the Coulomb interaction may prove necessary to be scaled up for neutral QM zones, to account for the missing polarization of the wave function by the MM zone.

Certain specific combinations of force fields and quantum-chemical methods lead to very good results for specific classes of molecules and reactions; generally, care must be taken...

2. Polarized QM / unpolarized MM (Electronic embedding)

The clear deficiency of the mentioned model is that the QM system, or its wave function, is not affected by the MM system whatsoever. Actually, the wave function should be *polarized* by the environment (MM system), which is represented by point charges.

A step to improve the description is to include the electrostatic interaction with the MM charges in the QM Hamiltonian, whatever the QM method is – semiempirical, HF, DFT or a correlated method. The interaction of QM electrons with MM point charges moves from the $E_{\mathrm{QM/MM}}$ term (where it was described with a force field) to the quantum energy E_{QM}, and is described as an interaction of a charge density with point charges; then, it has the same form as the interaction with QM nuclei and brings on *little* increase of the computational cost. The interaction of QM nuclei with MM point charges may remain in $E_{\mathrm{QM/MM}}$.

Thus, the QM Hamiltonian changes to (schematically, may be method-dependent)

$$\hat{H}'_{\mathrm{QM}} = \hat{H}_{\mathrm{QM}} - \sum_{j}^{\text{QM electrons}} \sum_{m}^{\text{MM atoms}} \frac{q_m}{r_{jm}} \tag{X.3}$$

and the $E_{\mathrm{QM/MM}}$ term is considered excluding the electrostatic interaction of QM electrons with MM atoms, so that only the nuclear charges Z_i remain:

$$E'_{\mathrm{QM/MM}} = \sum_{i}^{\text{QM atoms}} \sum_{m}^{\text{MM atoms}} \left(\frac{Z_i \cdot q_m}{r_{im}} + 4\varepsilon_{im} \left(\frac{\sigma_{im}^{12}}{r_{im}^{12}} - \frac{\sigma_{im}^{6}}{r_{im}^{6}} \right) \right) \tag{X.4}$$

A QM/MM study would then run as follows:

1. The choice of specific QM and MM methods. Since the quantum-chemical calculation is used to describe the most interesting part of the system, as well as it is by far the most resource- and time-consuming component of the calculation, particular care must be taken with the selection of the QM method – the requirements regard both accuracy and computational efficiency, at the same time.

2. Determination of the Lennard-Jones parameters for the QM part of the system (for the calculation of $E_{\mathrm{QM/MM}}$). One can use either 'normal' parameters from a force field, or attempt to develop a special set of LJ parameters for the used QM method.

3. The simulation itself. Every step of the simulation involves one QM calculation, one MM calculation and a calculation of $E_{\mathrm{QM/MM}}$. The properties of interest (possibly but not necessarily of quantum character) are then evaluated as ensemble averages.

3. Fully polarized (Polarized embedding)

The QM/MM variant just described is already a very good approach with good changes for acceptable accuracy. The point at which it may be considered somewhat unbalanced is that whereas the QM system is being polarized by the MM charges, the MM molecules themselves cannot be polarized.

Should this fact be problematic in a study of a particular chemical process, it is possible to include this phenomenon in the QM/MM framework as well. However, in such a case, it is necessary to use a force field that makes it possible to account for the polarization of MM atoms or molecules.

Most of the standard force fields do not include polarization terms, mainly because of the extra computational effort. Every MM atom or molecule is assigned a polarizability α,[13] and an *induced dipole* at each polarizable center is then obtained as

$$\vec{\mu}^{\text{ind}} = \alpha \cdot \vec{E} \tag{X.5}$$

where $\vec{E}$ is the intensity of electric field induced by all the surrounding atomic point charges *and* all the induced dipoles. Because of that, the induced dipoles must be evaluated iteratively, until convergence (self-consistence) is reached. There are two possible issues with this procedure: (i) its iterative character makes the calculation an order of magnitude more time-consuming than a non-polarizable MM calculation of the same system, and (ii) the convergence of dipoles may be potentially problematic.

Within a QM/MM scheme involving a polarized MM method, the induced dipoles $\vec{\mu}^{\text{ind}}$ interact with the QM nuclei (i.e. some extra point charges) and with the QM electron density. Thus, the *entire* QM/MM calculation has to be performed iteratively until self-consistency is reached, and both the QM calculation and the MM treatment of induced charges proceed in a loop. This makes the computational cost rise dramatically.

To date, no conclusive answer has been given to the question if the completely polarized methodology brings a significant improvement if compared with the previously mentioned approach (polarized QM / unpolarized MM). Above all, the improvement would have to be necessary to justify the quite extreme computational cost.

[13] In principle, polarizability is a symmetrical tensor of rank 2. If isotropic polarizability is considered then α becomes a single value (scalar).

D. Covalent bonds across the boundary

All of the QM/MM schemes discussed so far involved purely non-bonded interaction between the QM and the MM subsystems. However, it may well turn out desirable or even necessary to divide the whole system in such a way that the QM and MM regions are connected with one or several covalent bonds. In such a case, a special treatment of the QM/MM boundary is necessary in order to perform a QM/MM calculation. Several possibilities are presented in the following.

1. Linear combination of molecular fragments

Imagine one wishes to simulate a large molecule, of which actually only a small part has to be treated quantum-chemically. The situation is more favorable if the interaction of the intended QM region with the rest of the molecule may be regarded as exclusively *steric*, i.e. if the electronic structure of the QM region is virtually unaffected by the rest of the molecule. This would be the case if this rest is composed of non-polar, i.e. alifatic or aromatic (even though large) groups(s), see Fig. 22 for an example.

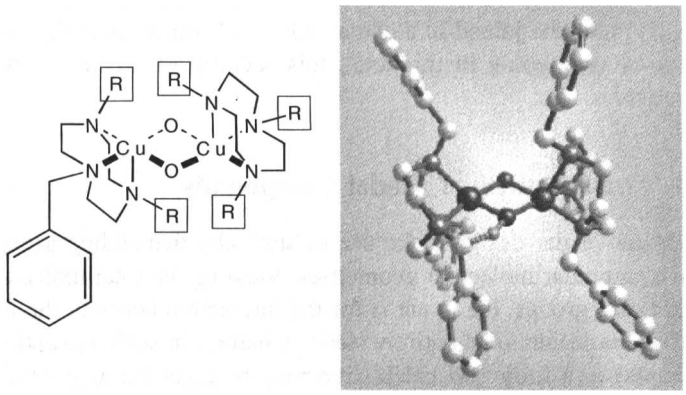


FIG. 22: A metallic complex with bulky non-polar functionalities. The five benzyl groups labeled with R are 'substituted' by H atoms in QM calculations. The remaining benzyl group is the donor in a hydrogen transfer reaction, and is thus included in the QM region. Reprinted from Cramer.

In such a case, the molecule may be regarded as a kind of a sum of the individual functional groups. The electronic structure of the QM system will be regarded as equal to the structure of a similar molecule where the bulky non-polar groups are replaced by *hydrogen atoms*. Then, the total energy may be expressed as the sum of energies of the

moleculare fragments (the QM-molecule 'capped' with hydrogens, and the bulky non-polar MM-molecules) like

$$\begin{aligned} E_{\text{total}} &= E_{\text{MM}}^{\text{large}} + \left(E_{\text{QM}}^{\text{small}} - E_{\text{MM}}^{\text{small}}\right) = \\ &= E_{\text{QM}}^{\text{small}} + \left(E_{\text{MM}}^{\text{large}} - E_{\text{MM}}^{\text{small}}\right) \end{aligned} \tag{X.6}$$

where the 'large' system is the entire molecule, and 'small' denotes the QM region. One can understand this approach so that the part of the MM energy corresponding to the 'small', interesting molecular fragment is substituted by corresponding QM energy (first line in Eq. X.6). The alternative way to think of Eq. X.6 (second line) is to concentrate on the 'small' fragment and its QM energy; the effect of the added non-polar groups is the added as a correction (in parentheses).

2. Link atoms

A more difficult situation arises if the intended MM region cannot be regarded as interacting only sterically, and there is for instance strong electrostatic interaction with the QM region. Typically, this is the case in proteins, where there are always polar and even charge amino-acid residues in the MM region, which polarize the electron density of the QM region (which is usually the binding site of a ligand or similar). What is missing in the approach presented in Section X C 2 is the description of covalent bonds crossing the QM/MM boundary.

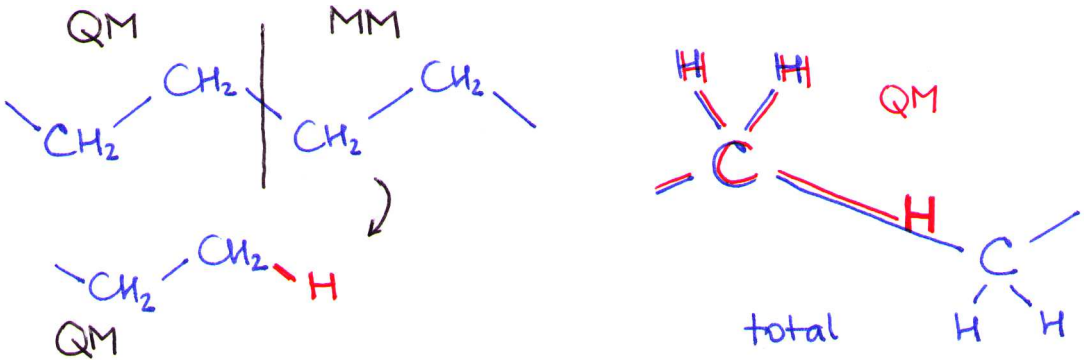


FIG. 23: Link atoms. Left: The QM/MM boundary cuts a bond between two sp^3-C atoms and a link hydrogen atom is used. Right: The link atom is placed on the C–C bond that has been cut; possible problem is that non-bonded interactions between the **H** and close MM atoms may diverge.

Link atoms are atoms that replace the covalently bonded MM system at the boundary, see Fig. 23. Usually, bonds between two sp^3-carbon atoms are chosen to be cut, and hydrogens are used as link atoms because of the similar electronegativity of carbon and hydrogen. It is thus desirable to define the QM region so that the bonds to be cut are as unpolar as possible, in order to minimize errors. The link hydrogen atom is then placed on the original C–C bond, in a typical C–H bonding distance from the QM carbon atom (Fig. 23 right).

The total energy may then be evaluated in the fashion called *additive coupling*. As the link atom is not really part of the QM region, the terms in expression for E_{QM} that involve the orbitals on the link atom are not evaluated. An obviously interesting point are the bonded interactions (bonds, angles and dihedral angles) involving the bond crossing the boundary. Their energy contributions are generally calculated with the force field. A possible exclusion are angles involving 2 QM atoms and 1 MM atom, and dihedrals with 3 QM atoms and 1 MM atom, which are omitted in some approaches.

Another important issue is that of MM atoms bearing point charges, that are situated very close to the QM system – typically, extremely close to a link atom. These would have unphysically large influence on the electronic structure of the QM system. Therefore, particular care must be taken of the charges of close MM atoms: These may be scaled down or even zeroed; alternatively, only their interaction with the QM atoms near the boundary may be scaled down or zeroed. A promising approach consists in the replacement of the close point charges by gaussian charge distributions – this maintains all charges while avoiding most of the trouble with too high interaction energies.

Alternatively, it is possible to apply the Eq. X.6 again, with the QM region ('small') now including the link atoms:

$$E_{\mathrm{total}} \;=\; E_{\mathrm{MM}}^{\mathrm{large}} + \left(E_{\mathrm{QM}}^{\mathrm{small+L}} - E_{\mathrm{MM}}^{\mathrm{small+L}}\right) \tag{X.7}$$

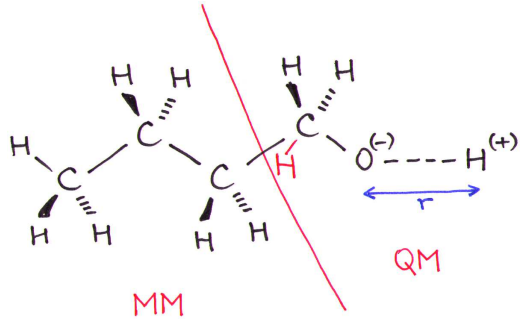
This is called the *subtractive coupling*.

The concept of link atoms is used very frequently in the studies of biomolecules, when quantum description is desired. The artifical separation of the molecular system in two brings on certain issues that need to be resolved in order to obtain correct results. Nevertheless, the development of QM/MM methodology has advanced considerably in the recent years, so that it is now considered to be a de facto standard tool in computational biophysics.

Deficiency of subtractive coupling – deprotonation of an alcohol

$$CH_3CH_2CH_2CH_2OH \rightarrow CH_3CH_2CH_2CH_2O^- + H^+$$

This reaction takes place in a small region of a large molecule. So, we can describe the 'interesting' region with QM – this will correspond to a *methanol* molecule. The rest of the molecule will be described as whole butanol with MM:



$$E^{\mathrm{QM/MM}}_{\mathrm{butanol}} = E^{\mathrm{QM}}_{\mathrm{methanol}} + \left(E^{\mathrm{MM}}_{\mathrm{butanol}} - E^{\mathrm{MM}}_{\mathrm{methanol}}\right)$$

We wish to evaluate the energy as a function of the O–H distance r. Let us assume the parameters for methanol and butanol to differ only in the force constant k^{OH} for the O–H bond. The remaining terms will give a constant independent of r:

$$\begin{aligned} E^{\mathrm{QM/MM}}_{\mathrm{butanol}}(r) &= E^{\mathrm{QM}}_{\mathrm{methanol}}(r) + \left(\frac{1}{2}k^{\mathrm{OH}}_{\mathrm{butanol}} \cdot (r - r_0)^2 - \frac{1}{2}k^{\mathrm{OH}}_{\mathrm{methanol}} \cdot (r - r_0)^2\right) + \text{const.} \\ &= E^{\mathrm{QM}}_{\mathrm{methanol}}(r) + \frac{1}{2}\left(k^{\mathrm{OH}}_{\mathrm{butanol}} - k^{\mathrm{OH}}_{\mathrm{methanol}}\right) \cdot (r - r_0)^2 + \text{const.} \end{aligned}$$

The MM energy remains in the form of a term proportional to r^2. For large r, the QM energy will be proportional to $-\frac{1}{r}$, due to Coulomb's law:

$$\lim_{r\to\infty} E^{\mathrm{QM}}_{\mathrm{butanol}}(r) = \lim -\frac{1}{r} = 0$$

The asymptotic behavior of total energy will look like

$$\lim_{r\to\infty} E^{\mathrm{QM/MM}}_{\mathrm{butanol}}(r) = \lim\left(-\frac{1}{r} + \frac{1}{2}k \cdot r^2\right) = \lim r^2 = \infty$$

So, the inequality of k^{OH} for methanol and butanol will make the total energy grow over all limits, for large distances r. The entire effort will go in vain.

Note that such an error will not arise in the *additive coupling* at all. Methanol will be calculated with QM, and something like propane with MM. This way, the parameters for the hydroxyl group *are not required* at all.

Excursion – the linking scheme

The atom charge in most force fields are designed in such a way that certain groups of atoms maintain neutral or integral charge. See an example of the CHARMM force field:

```
RESI SER          0.00
GROUP
ATOM N    NH1    -0.47  !      |
ATOM HN   H       0.31  !  HN-N
ATOM CA   CT1     0.07  !      |   HB1
ATOM HA   HB      0.09  !      |   |
GROUP                   !  HA-CA--CB--OG
ATOM CB   CT2     0.05  !      |   |     \
ATOM HB1  HA      0.09  !      |   HB2    HG1
ATOM HB2  HA      0.09  !   O=C
ATOM OG   OH1    -0.66  !      |
ATOM HG1  H       0.43
GROUP
ATOM C    C       0.51
ATOM O    O      -0.51
```

Now, let us take the side chain as the QM system. Then, the atom CA is very close to the link atom placed in between CA and CB, and so the point charge on CA would disturb the QM region drastically. Also, it is impossible to remove the CA atom simply, because the entire backbone would not be charge-neutral any more. There are a couple of possibilities to deal with this problem:

- A drastic but still often used linking scheme ('exgroup' in CHARMM) is to remove *all* charges of the group close to the QM region. In our example, these are CA, HA, N and HN. Obviously, we would lose the strong N–HN dipole within this scheme, which could lead to very inaccurate result.

- A better way ('div' in CHARMM) is to *divide* the charge of the so-called host atom (here CA) among the remaining atoms of the whole host group (composed of HA, N and HN). The resulting charges in this example would be: CA=0, HA=0.11, N=−0.44, HN=0.33.

The latter approach is limited by the requirement to have the QM/MM boundary between two individual charge groups. 'Cutting' of a covalent bond within a single charge group (e.g. between CB and OG in this residue) is only possible with the former approach (the modified charges would be CB=0, HB1=0, HB2=0).

3. Frozen orbitals

As mentioned, the introduction of link atoms may cause problems with non-bonded interactions, because of the possibly extremely short distance between these artificially added QM atoms and nearby MM atoms. Also, the representation of charge densities by MM point charges may result in inaccuracy and/or computational instability. A promising attempt to avoid this issue may be to introduce no new *atoms*, but rather treat the orbitals on the QM/MM boundary in a special way. The shape of these orbitals can be held constant during the simulation, hence the term *frozen orbitals*.

With this approach, we will have not two but rather *three* regions in the simulation: the QM and MM regions as before, and an *auxiliary region* on the QM/MM boundary on top of that. The atoms in this auxiliary region possess their normal nuclei and electron densities expressed using the basis of atomic orbitals. Then, the auxiliary region actually possesses a quantum character, but still its interaction with itself as well as with the MM system can be calculated classically.[14] Another only slight complication is the energy of interaction of the QM system with the auxiliary – this adds another term to the Hamiltonian which corresponds to the interaction of the wave function with the frozen charge density in the auxiliary region, which is only slightly more complex than the interaction with point charges.

In the simple case of a single covalent bond being cut by the QM/MM boundary, the auxiliary region may be very small, as seen in Fig. 24.

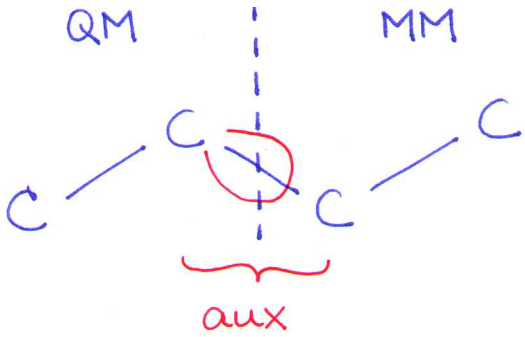


FIG. 24: Frozen orbital scheme. Red – the frozen sp^3 orbital for the bond crossing the boundary.

There are basically two approaches how to freeze the electron density. In the localized

[14] This may seem awkward but it is not so bad: interaction of (frozen) charge density with itself, and interaction of (frozen) charge denstity with a set of point charges; this is no issue in a calculation whatsoever.

SCF method (LSCF), every covalent bond crossing the QM/MM boundary is represented by a single frozen orbital – this is the (hybrid) atomic orbital localized on the QM atom before the boundary, which is calculated once at the beginning of the simulation and does not change shape afterwards any more. Some care must be taken of the occupation of the frozen orbitals in order to handle the density correctly, and this requires accurate accounting.

The generalized hybrid orbital approach (GHO) is different in that the QM/MM boundary does not cut any covalent bond, but rather the boundary passes through *an atom*. Here, the (hybrid) atomic orbitals on this particular atom which would belong to the MM region, are considered to be frozen. Their populations are calculated so that the resulting charge density corresponds to the point charge that this atom would have in an MM calculation. The remaining orbital on the atom, which points inwards the QM region, is frozen in shape but its occupation is free to vary in the QM calculation.

The approaches using frozen orbitals have received certain attention in the recent years and they are constantly being developed. However, the method of link atoms is clearly being applied more often, and has already been implemented in many popular simulation packages.

E. Advanced stuff and examples

1. QM/QM/MM

It is possible to improve the process of dividing the entire molecular system, so that there are *three* disjunctive regions: Then, one may be treated with an advanced, expensive QM method (correlated methods like CC, CAS...), another region surrounding the first one will be described with a faster QM method (like DFT or semiempirical), MM will be used for the rest of the system, probably including the solvent. This approach may be referred to as QM/QM/MM, and an example is the ONIOM scheme[15] implemented in GAUSSIAN.

2. Photochemistry–Retinal

Retinal is a polyene, covalently bound to a lysine side chain via a protonated Schiff base.

[15] a clear connotation with the layers of an onion

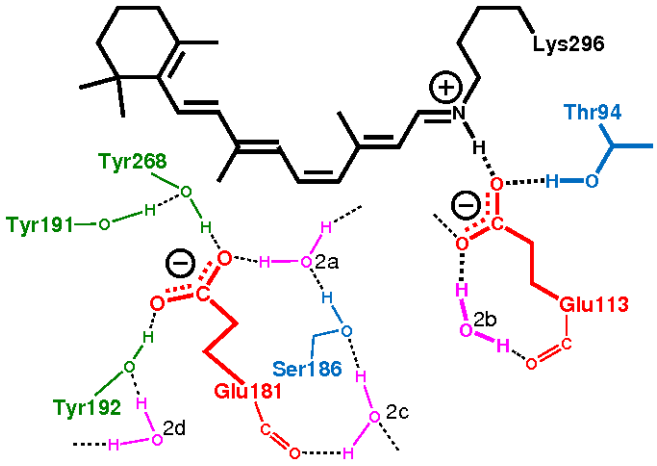


FIG. 25: Retinal (the chromophore) in rhodopsin, with the counterion Glu113 and charged Glu181.

It is not easy to choose a correct size of the QM region, and there are many possibilities:

	size	goodness, issue
	polyene (ring to NH)	**bad**, boundary cuts a polar bond
	retinal+CH_2	**bad**, link atom too close to the important region
QM1	retinal+sidechain to CB	**fair, but** no charge transfer to Glu113 possible
QM2	QM1+counterion	**better, but** no charge transfer to Wat
QM4	QM2+Wat2b+Thr94	**good, but** no polarization at Glu181
	QM4+Glu181	**very good, but**...

A highly correlated method (like CAS-PT2) is required to calculate the electronic spectrum of a chromophore in a protein (like the retinal within a rhodopsin). Also, it is crucial to describe the interaction of the electronic structure of retinal with the atoms of the protein. On the other hand, the vdW interaction with the protein is not so important, because the structure of the molecular system does not change during the photochemical process. Thus, a single-point calculation is sufficient. For the geometry optimization or MD simulation of the system, the complex electronic structure of retinal makes the use of a (modest) QM method necessary.

The calculation starts with a QM/MM calculation with approximative DFT, and the structure of the whole protein is optimized. The coordinates of QM atoms are written into a file, and so are the coordinates of all MM atoms together with the charges. These sets of data are fed to a highly correlated method, for the calculation of excitation energy.

XI. IMPLICIT SOLVENT AND COARSE GRAINING

A. Continuum electrostatic methods: Free energy of solvation

Up to now, we treated the molecules of interest either in the gas phase or immersed in an *explicit* solvent, meaning a solvent represented by individual atoms/molecules. Then, the difference of solvation free energy could be evaluate by such methods like the free energy perturbation or umbrella sampling (PMF).

Consider the example of a polypeptide in the α-helix and β-sheet conformations. The free energy difference of the two structures is given by

- the difference of *internal energies / enthalpies*
- the *entropic contributions* – above all the vibrational component of the configurational entropy
- the difference of free energies of *solvation*

The α-helix has a much larger dipole moment than the β-sheet, due to the peptide bonds pointing in the same direction with respect to the axis of the helix. Because of that, the α-helix is better solvated in a polar medium (like water). Therefore, the solvation (qunatified by the solvation free energy) plays a key role in the equilibrium between the conformations of the peptide in solution.

In this chapter, we will discuss how to describe the solvent with an *implicit treatment*. Since the number of solvent (water) molecules may easily become excessive for a large solute molecule, this approach may be preferred to an explicit representation of solvent, which would be too costly.

Several energy contributions have to be considered for the process of solvation:

- A *cavity* (Fig. 26) in the solvent bas to be formed against the outside pressure. This also involves the necessary rearrangement of the solvent molecules at the surface of the cavity. The energy contribution ΔG_{cav} accounts for i.a. the decrease of entropy and the loss of solvent–solvent interactions.
- The van der Waals (ΔG_{vdW}) and electrostatic (ΔG_{ele}) interaction of the solute molecule with the solvent (Fig. 27).

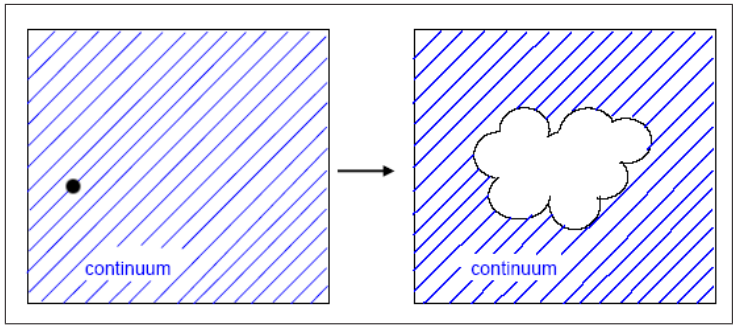


FIG. 26: Formation of the cavity

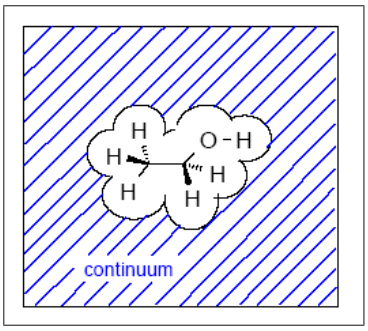


FIG. 27: Electrostatic and vdW interactions upon inserting the molecule into the cavity.

Then, the total solvation energy is

$$\Delta G_{\text{solv}} = \Delta G_{\text{cav}} + \Delta G_{\text{vdW}} + \Delta G_{\text{ele}} \tag{XI.1}$$

An important concept is that of the *solvent accessible surface area* (SASA): We consider the molecule to be a solid body exposed to the solvent, and SASA is then the surface area of this body. In a reasonable approximation, the terms ΔG_{cav} and ΔG_{vdW} are taken to be proportional to SASA. Since arbitrary surfaces are difficult to involve in calculations, it is convenient to obtain the total surface of the molecule from the surfaces of the individual atoms of the molecule, S_i.[16] A practical approach is then to write

$$\Delta G_{\text{cav}} + \Delta G_{\text{vdW}} = \sum_i c_i \cdot S_i \tag{XI.2}$$

Since this term does not contain the electrostatic contribution (which will be discussed in the following) it is appropriate to parameterize it with respect to the solvation energies of hydrocarbons.

[16] In principle, it is possible to determine the SASA of the molecule as a whole. This may be done by rolling an imaginary ball of a certain diameter (typically 2.8 Å to mimic H_2O) on the molecular surface.

When does it work?

- will probably work if the electrostatic effect of the surrounding solvent is dominant. An example is the shielding of solvent-exposed charged side chains of proteins.
- will not succeed if some kind of specific interaction between the solute and the solvent plays a role, such as hydrogen bonding. An example may be the dynamics of small peptides dissolved in water; a tripeptide can form either an intramolecular hydrogen bond (with a seven-membered ring) or hydrogen bonds with the solvent. This fine balance is difficult to describe only via general electrostatic interactions.

In the following, we will discuss several models to evaluate the term ΔG_{ele}. As discussed in the chapter on non-bonded interactions, the electrostatic energy of a point charge q located at $\vec{r}$ is

$$E_{\mathrm{ele}} = q \cdot \Phi(\vec{r}) \tag{XI.3}$$

where $\Phi(\vec{r})$ is the electrostatic potential (ESP) induced by the charge distribution in the rest of the system. To obtain the solvation energy, we have to calculate the electrostatic potential of the protein in vacuo $\Phi_{\mathrm{vac}}(\vec{r})$, and in solution, $\Phi_{\mathrm{solv}}(\vec{r})$. The solvation energy then follows:

$$\Delta E_{\mathrm{ele}} = q \cdot \Phi_{\mathrm{solv}}(\vec{r}) - q \cdot \Phi_{\mathrm{vac}}(\vec{r}) \tag{XI.4}$$

With the defintion of the *reaction field*

$$\Phi_{\mathrm{rf}}(\vec{r}) = \Phi_{\mathrm{solv}}(\vec{r}) - \Phi_{\mathrm{vac}}(\vec{r}) \tag{XI.5}$$

the solvation energy follows as

$$\Delta E_{\mathrm{ele}} = q \cdot \Phi_{\mathrm{rf}}(\vec{r}) \tag{XI.6}$$

As we learned up to now, the potential energy E is related to one point on the potential energy surface. Moving to the free energy, we have to average over all the solvent configurations to include the entropy contributions.[17] Thus, it would be af advantage if we are able to determine the reaction field in such a way that it includes these contributions, so that we effectively obtain the free energy:

$$\Delta G_{\mathrm{ele}} = q \cdot \Phi_{\mathrm{rf}}(\vec{r}) \tag{XI.7}$$

[17] And, in an NPT ensemble, the enthalpy includes the PV term in addition ($PV = Nk_{\mathrm{B}}T$ for ideal gas).

1. *Continuum electrostatic methods: the Born and Onsager models*

Born (1920) determinded analytically the work needed to bring a charge q from vacuo into a spherical cavity of radius a formed in a solvent with a *dielectric constant* ε (Fig. 28 left) as

$$\Delta G_{\text{ele}} = -\frac{q^2}{2a}\left(1 - \frac{1}{\varepsilon}\right) \tag{XI.8}$$

The dielectric constant takes values of 1 for vacuo (thus $\Delta G_{\text{ele}} = 0$), 80 for water and between 2 and 20 for protein environment.

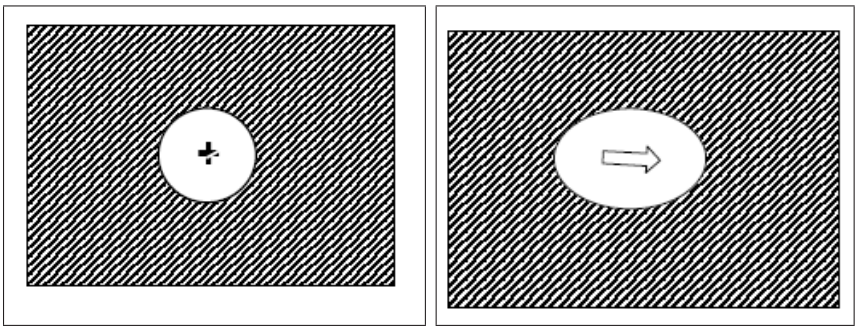


FIG. 28: Solvation of a point charge (left) and a point dipole (right).

Onsager and Kirkwood (1930) developed a model for a dipole in a cavity (Fig. 28 right). The dipole moment of a molecule μ induces charges at the surface of the cavity – the molecular dipole is an "action" which induces a "reaction" of the solvent, hence the electrostatic potential is called the *reaction field*, which was derived as

$$\Phi_{\text{rf}} = \frac{2(\varepsilon - 1)}{2\varepsilon + 1} \cdot \frac{1}{a^3} \cdot \mu \tag{XI.9}$$

$$\Delta G_{\text{ele}} = -\frac{1}{2}\Phi_{\text{rf}} \cdot \mu \tag{XI.10}$$

These simple models are implemented in many standard quantum chemistry programs as well as simulation packages, in order to calculate solvation energies. Of course, even for small molecules, the point charge or dipole approximation in combination with a spherical or ellipsoidal surface is quite unrealistic. Therefore, the *polarizable continuum model* (PCM) extends these schemes to arbitrary surfaces constructed with the use of vdW radii of the atoms. An alternative approach are the *conductor-like screening models* (COSMO), which derive the polarization of the dielectric (insulating) solvent from a scaled-conductor approximation.

2. Continuum electrostatic methods: Poisson–Boltzmann equation (PBE)

For large molecules, other approximations were developed, starting from the Poisson equation

$$\nabla\varepsilon\nabla\Phi = -4\pi\rho \tag{XI.11}$$

This is a partial differential equation. Given are the charge distribution ρ and the dielectric constant ε, and we wish to solve the equation for Φ.

One way to solve it is to discretize the problem on a three-dimensional grid. Here, we have the charge distribution and the (non-constant) dielectric constant distributed on the grid, and the potential Φ is calculated on every grid point iteratively (Fig. 29), using finite differences for the second derivative.

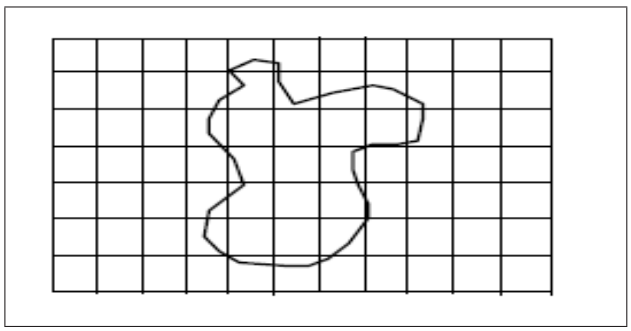


FIG. 29: Representation of Φ, ρ and ε on a grid.

Very often, we want small *ions* to be part of the solvent. In certain situations, ions are very important, like in the simulations of DNA, where counterions are necessary to compensate for the charge on the phosphate groups. If we do an MD with explicit solvent, the include the ions as particles, naturally. But, how can we accomplish this within a continuum representation of the solvent?

If we know the electrostatic potential in the system, the energy of an ion is

$$E_i(r) = q_i \cdot \Phi(r) \tag{XI.12}$$

and with the Boltzmann distribution, the density at that point is

$$n_i(r) = n_i^0 \cdot \exp\left[-\frac{q_i \cdot \Phi(r)}{k_\mathrm{B}T}\right] \tag{XI.13}$$

with n_i^0 being the number density in bulk solution, or *concentration*. Therefore, anions concentrate in regions with positive Φ whereas cations in regions with negative Φ. Multiplying with the ion charges, we obtain the charge distribution of the ions:

$$\rho_{\text{ions}} = \sum_i q_i \cdot n_i^0 \cdot \exp\left[-\frac{q_i \cdot \Phi(r)}{k_\text{B}T}\right] \tag{XI.14}$$

Now, if we have two kinds of ions with opposite charges (e.g. Na^+ and Cl^- with $q = \pm 1$) in the solution, we will have terms like

$$1 \cdot \exp[-1 \cdot \Phi(r)/k_\text{B}T] - 1 \cdot \exp[1 \cdot \Phi(r)/k_\text{B}T] \tag{XI.15}$$

which may be combined by noting the definition of hyperbolic functions:

$$\exp[x] - \exp[-x] = 2\sinh[x] \tag{XI.16}$$

Then, adding the charge distribution due to the ions, to the Poisson equation, we obtain the *Poisson–Boltzmann equation*:

$$\nabla\varepsilon\nabla\Phi = -4\pi\rho + \sum_i q_i \cdot n_i^0 \cdot \sinh\left[\frac{q_i \cdot \Phi(r)}{k_\text{B}T}\right] \tag{XI.17}$$

This equation is usually written in the form

$$\nabla\varepsilon\nabla\Phi = -4\pi\rho + \varepsilon \cdot \kappa^2 \cdot \frac{k_\text{B}T}{q} \cdot \sinh\left[\frac{q \cdot \Phi(r)}{k_\text{B}T}\right] \tag{XI.18}$$

with the Debye–Hückel parameter

$$\kappa^2 = \frac{8\pi q^2 I}{\varepsilon \cdot k_\text{B}T} \tag{XI.19}$$

(ionic strength $I = \frac{1}{2}\sum_i c_i z_i^2$, c_i – concentration, z_i charge of ion i).

At low ionic strength, the difficult differential equation may be simplified by truncating the Taylor expansion of sinh, which yield the *linearized PBE* of the form

$$\nabla\varepsilon\nabla\Phi = -4\pi\rho + \varepsilon \cdot \kappa^2 \cdot \Phi(r) \tag{XI.20}$$

The PBE describes two effects of solvation: First, the charge distribution on the protein polarizes the dielectric outside (the "solvent"). This leads to a *screening* of the effect of the *solvent-exposed charges* of the protein atoms. The "solvent molecules" will arrange around the charge, and dipoles will be induced, which will compensate for the charge largely.

Effectively, the charges pointing into the solvent will be nearly canceled. The second effect is that the *solvent ions* will be distributed so that the overall charge distribution will become more uniform. For instance, if a negative charge points into the solvent, a positive ion will be located close to it, effectively reducing the magnitude of the electrostatic field. These two points usually become important when (photo-)chemical reactions in proteins are to be described. The solvent around a protein should always be taken into account.

When calculating solvation energies, we have to determine the reaction field. For this, we perform one PBE calculation 'in vacuo' ($\varepsilon = 1$) and one for the solution ($\varepsilon = 80$)

$$\Phi_{\mathrm{rf}} = \Phi_{\varepsilon=80} - \Phi_{\varepsilon=1} \tag{XI.21}$$

yielding the solvation free energy as

$$G_{\mathrm{elec}} = \frac{1}{2} \sum_i q_i \Phi_{\mathrm{rf}} \tag{XI.22}$$

The computational cost of the solution of PBE becomes excessive if PBE has to be solved several million times during a MD simulation (remember, it has to be done in *every* MD step). Therefore, approximations have been developed.

3. The generalized Born (GB) model

A simple idea is to use the Born equation XI.8 for the atomic charges of the biomolecule, to calculate the solvation energy of the charges:

$$\Delta G^1_{\mathrm{ele}} = -\left(1 - \frac{1}{\varepsilon}\right) \sum_i \frac{q_i^2}{2a_i} \tag{XI.23}$$

What changes upon solvation as well, is the interaction of the individual charges. The interaction energy in a medium with $\varepsilon > 1$ may be expanded as

$$\begin{aligned} E_{\mathrm{ele}} &= \frac{1}{2} \sum_{i \neq j} \frac{1}{\varepsilon} \frac{q_i \cdot q_j}{r_{ij}} = \\ &= \frac{1}{2} \sum_{i \neq j} \frac{q_i \cdot q_j}{r_{ij}} - \frac{1}{2} \left(1 - \frac{1}{\varepsilon}\right) \sum_{i \neq j} \frac{q_i \cdot q_j}{r_{ij}} \end{aligned} \tag{XI.24}$$

where the red term is the usual Coulomb interaction in vacuo, and the blue one corresponds to the reaction field contribution – the contribution due to solvation:

$$\Delta G^2_{\mathrm{ele}} = -\frac{1}{2} \left(1 - \frac{1}{\varepsilon}\right) \sum_{i \neq j} \frac{q_i \cdot q_j}{r_{ij}} \tag{XI.25}$$

The solvation contribution to the free energy then follows as the sum of the terms $\Delta G^1_{\text{ele}} + \Delta G^2_{\text{ele}}$:

$$\Delta G_{\text{ele}} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\left(\sum_i \frac{q_i^2}{a_i} + \sum_{i \neq j} \frac{q_i \cdot q_j}{r_{ij}}\right) \tag{XI.26}$$

This formula describes the interaction of charges that are located in spherical cavities with radii a_i. For charged bodies of generalized shapes, the derivation is only valid if the distance between the charges is large ($r_{ij} \gg a_i + a_j$). In other words, Eq. XI.26 can be considered valid for the interaction of the charges q_i and q_j in one of two limiting cases:

$$E = \begin{cases} \dfrac{q_i^2}{a_i}, & \text{if } i = j \text{ ('self-interaction, i.e. solvation energy)} \\ \dfrac{q_i \cdot q_j}{r_{ij}}, & \text{if } i \neq j \text{ and } r_{ij} \to \infty \end{cases} \tag{XI.27}$$

Therefore, the interaction of two charges with finite radii becomes the interaction of point charges once the distance is large. But, what is the interaction energy for intermediate distances of, say, 2 Å to 10 Å? What we need here is an expression that interpolates between the limiting cases. It can be easily shown that a function of the form

$$f(r_{ij}) = \sqrt{r_{ij}^2 + a_i a_j \exp\left[-\frac{r_{ij}^2}{4 a_i a_j}\right]} \tag{XI.28}$$

exhibits this behavior: The exponent vanishes for large r, so that $f \approx r$; and the exponent approaches unity for small r, giving $f \approx \sqrt{a_i a_j}$ or a_i. With this function, we can write

$$\Delta G_{\text{ele}} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right) \cdot \sum_{i,j} \frac{q_i \cdot q_j}{f(r_{ij})} \tag{XI.29}$$

for the free energy of solvation of charges within the Born approximation, involving

1. the solvation energy of every charge due to the Born formula

2. the change of the Coulomb interaction energy of the charges, due to solvation

Unfortunately, there is a fundamental problem with this equation. The Born equation was derived for a charged particle with radius a, *in contact* with the solvent. But, many charges will be deeply buried inside the protein, and will not 'feel' much of the solvent! Therefore, if we use the same value of a_i for all charges, the solvation energy of some charges will be grossly overestimated.

A solution would be to build an empirical model: The solvation energy of charge q_i

$$\Delta G^1_{\mathrm{ele},i} = -\frac{1}{2}\left(1 - \frac{1}{\varepsilon}\right)\frac{q_i^2}{a_i} \tag{XI.30}$$

depends on a_i. Then, if we wish to scale down this energy for a charge inside the protein, we can use a *larger* value of a_i than for the same charge located at the surface of the protein. What needs to be done is to determine a_i for every charge. In principle, this could be done by performing PBE calculations for every charge, which would yield $\Delta G^1_{\mathrm{ele},i}$ and also the a_i. Alas, this is too costly, and doing PBE calculations is exactly what we wanted to avoid. Therefore, we need an approximation to calculate the radii a_i.

4. A simple approximation to the Generalized Born (GB) model

The work necessary to transfer a charge distribution ρ into a polarizable medium is

$$\Delta G = \frac{1}{2}\int \rho \cdot \Phi \, \mathrm{d}V \tag{XI.31}$$

Now, consider a charge q_i inside a protein surrounded by water (ε_W). It can be shown that the energy of this charge can be written as

$$\Delta G^i_{\mathrm{ele}} = -\frac{1}{8\pi}\left(1 - \frac{1}{\varepsilon_W}\right)\int_{\mathrm{ext}} \frac{q_i^2}{r^4}\,\mathrm{d}V \tag{XI.32}$$

where the integration proceeds over the 'exterior' of the protein, i.e. over the whole space outside the protein (Fig. 30).

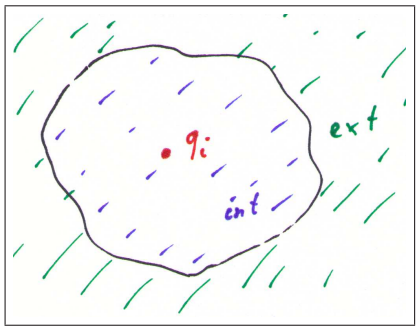


FIG. 30: Integration over 'int' or 'ext'

Comparing with the Born formula (Eq. XI.8), we find

$$\frac{1}{a_i} = \frac{1}{4\pi}\int_{\mathrm{ext}} \frac{1}{r^4}\,\mathrm{d}V \tag{XI.33}$$

with r being the distance from the charge to the 'boundary' of the protein. This a_i will vary depending on the location of the charge – it will be larger for charges buried inside of the protein! The integral over the outside of the protein can be transformed into an integral over the 'interior' of the protein, using the van der Waals radius α_i of atom i:

$$\frac{1}{a_i} = \frac{1}{\alpha_i} - \frac{1}{4\pi} \int_{\text{int}, r>\alpha_i} \frac{1}{r^4} \, \mathrm{d}V \tag{XI.34}$$

A possible approximation of this is to fill the space inside with spheres, and approximate thereby the volume of the protein molecule by the volume of the individual spheres:

$$\frac{1}{a_i} = \frac{1}{\alpha_i} - \sum_{j \neq i} \frac{1}{4\pi} \int_{\text{sphere } j} \frac{1}{r^4} \, \mathrm{d}V \tag{XI.35}$$

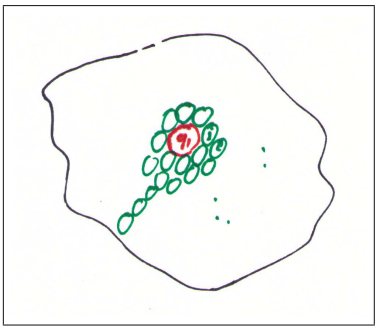


FIG. 31: The simple GB model

However, this approximation turns out to be insufficient. Instead, one can try to model the space to be integrated over with an empirical formula: the model has to represent the space 'int' in Eq. XI.34. Every atom has a volume V_j, and since $\Delta G_i \propto a_i^{-1}$, the volumes of all other atoms reduce the solvation energy of atom i, i.e. they increase a_i by

$$\frac{1}{a_i} = \frac{1}{\alpha_i} - \frac{V_j}{r_{ij}^4} \tag{XI.36}$$

where r_{ij} is the distance between the charge i and the atom j, which reduces its solvation energy. The model has the following terms:

$$\begin{aligned} a_i^{-1} \;=\; & \frac{1}{\lambda \cdot R_{\text{vdW},i}} - P_1 \frac{1}{R_{\text{vdW},i}^2} - \sum_j^{\text{bond}} \frac{P_2 V_j}{r_{ij}^4} - \sum_j^{\text{angle}} \frac{P_3 V_j}{r_{ij}^4} \\ & - \sum_j^{\text{nonbond}} \frac{P_4 V_j}{r_{ij}^4} \cdot \text{CCF}(P_5, r_{ij}) \end{aligned} \tag{XI.37}$$

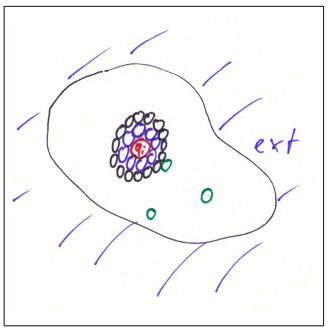


FIG. 32: The empirical GB model. blue – the binding neighbors, black – the angle, green – the nonbond 'atoms'.

The Born radius of atom i in solution is $\lambda \cdot R_{\mathrm{vdW},i}$, and then reduced due to a quadratic term, the sum over the bonded, neighbors (bonded, angles) and the all non-bonded interactions. For the latter, the function CCF is unity when the atoms do not have overlap, but reduced when they overlap. The parameters $\lambda, P_1, \ldots, P_5$ are fitted to reproduce the PBE results for the solvation energies of atoms in peptides and proteins. This model works (in contrast to the simple analytical one discussed above Fig. 31) due to the empirical fitting of parameters.

5. Practical example – MM-PBSA

The implicit solvent models are used to evaluate the solvation energy and force acting upon the atoms of solute in an MD simulation, but this not the only possible application. The considerable interest in free energies of binding of ligands to biomolecules, or even in the absolute free energies of molecules in solution led to the development of *post-processing* approaches to evaluate free energies. Here, a normal simulation (no matter if with an implicit or explicit solvent) is run, and the necessary components of the free energies of interest are evaluated by an analysis of the trajectory obtained.

The MM total energy of the system is evaluated without cutoff to yield the internal energy. The electrostatic contribution to the solvation free energy is evaluated with some of the methods described in this chapter, whereas the non-polar contribution is determined with SASA-dependent terms. Finally, the configurational entropy can be estimated with a normal-mode analysis. The total free energy is approximated by the sum of these terms.

This approach is undoubtedly very approximative and the various methods used are of very different character. Yet, results of very good quality may still be obtained.

B. United-atom force fields and coarse-grained models

In the studies of biomolecules, a proper and efficient treatment of the solvent is the key to the feasibility of the entire model. However, it may well happen that there are other components in the system that contain a large number of atoms – an example may be the lipid in the studies of transmembrane proteins. Even worse, the biomolecule itself may be exceedingly large – a very large protein or a long nucleic acid species. In such cases, it is necessary to modify the description of the biomolecule, and to design a simplified molecular model.

Early force fields (like Weiner 1984 and others) already used a similar idea. Within the *united-atom* force fields, each hydrogen atom was considered not individually, but rather condensed to the heavy atom to which it was connected. This way, the number of atoms was reduced considerably if compared with the *all-atom* force fields, which earned popularity in the 1990's. It is necessary to mention here that this approach works very well for non-polar C–H bonds, so that it is a very good approximation to consider a methyl group constituting one united atom. On the other hand, the substitution of a polar O–H group by a single particle is obviously a very crude approximation which will not work unless there are further correction terms in the force field. The united-atom force fields found their use in the modern computational chemistry e.g. in studies involving lipids, where each methylene group constitutes a united atom, cf. Fig. 33.

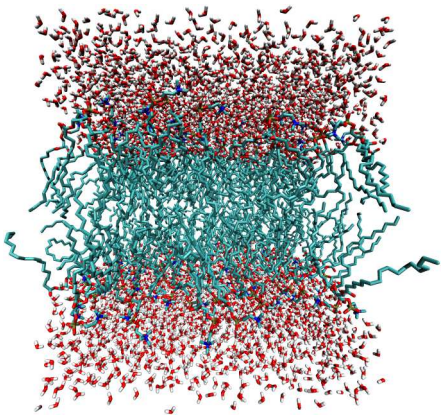


FIG. 33: A snapshot from the simulation of a lipid bilayer in water. The lipid (DOPC) is described with a united-atom force field – every CH_2 group is represented by a united atom. Downloaded from the website of R. Böckmann.

An advanced and sophisticated approach to cut the computational expense of simulations is the *coarse graining* (CG) of the problem. Quite naturally, a way to accelerate the evaluation of interactions is to reduce the number of particles involved. As it may not be always possible to reduce the number of atoms, an alternative idea is to consider particles composed of *several* atoms, so-called *beads*. Then, the number of inter-particle interactions will decrease, and in spite of the possibly more complex form of these interactions, the computational expense may be largely reduced as well. The necessary parameters of the force field are often obtained by fitting to all-atom force fields.

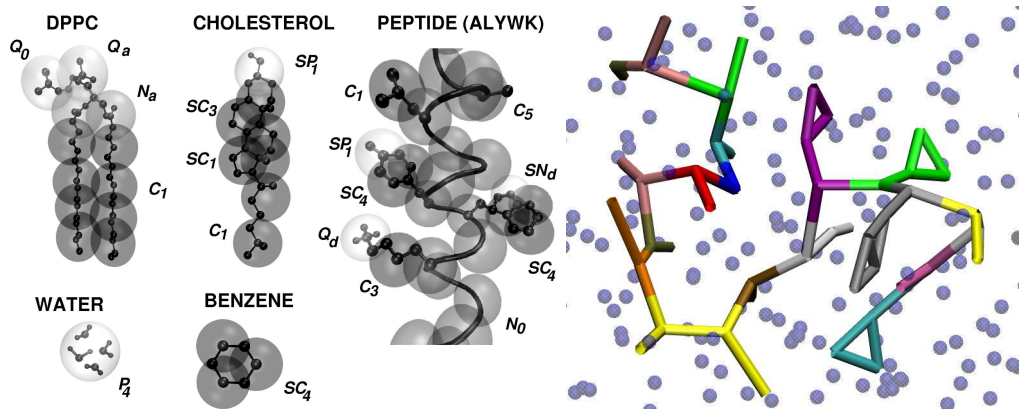


FIG. 34: Left: The CG force field MARTINI – mapping of beads onto molecular fragments. Right: A solvated peptide with MARTINI. Downloaded from the MARTINI website.

Every bead usually represents several atoms, and a molecule is composed of several beads, refer to Fig. 34 for the MARTINI force field. Such CG force fields are particularly useful for simulations of large-scale conformational transitions, involving either exceedingly large molecular systems or excessive time scales, or both. Another example is the VAMM force field for proteins, where every amino acid is represented by a single bead at C-α, see Fig. 35.

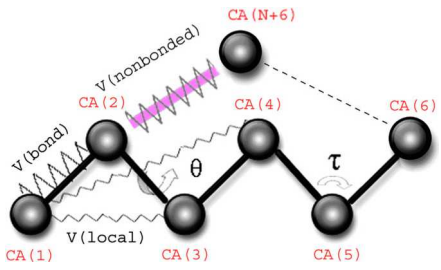


FIG. 35: The CG force field VAMM. Reprinted from Korkut & Hendrickson 2009.

XII. ENHANCING THE SAMPLING

At room temperatures, normal nanosecond length MD simulations have difficulty overcoming barriers to conformational transitions and may only sample conformations in the neighborhood of the initial structure.

A. Molecular dynamics as a way to the global minimum

Quotation from "A molecular dynamics primer" by Furio Ercolessi, University of Udine, Italy (`www.fisica.uniud.it/~ercolessi`).

Molecular dynamics may also be used as an optimization tool. Let us suppose that a set of N particles has many possible equilibrium configurations – this is truly the case with large (bio)molecules. The energy of these configurations is in general different, and one of them will be the lowest; each of the configurations, however, corresponds to a local minimum of the energy and is separated from every other by an energy barrier.

Finding the most energetically favorable structure – i.e. the global minimum of the energy function – within an approach based on traditional minimization techniques (steepest-descents, conjugate gradients, etc.) is tricky as these methods do not normally overcome energy barriers at all and tend to fall into the nearest local minimum. Therefore, one would have to try out several (many) different starting points, corresponding to different "attraction basins" in the energy landscape, and relax each of them to the bottom of the basin. The optimal structure would then be the one with the lowest energy, provided we were lucky enough to select it in the list of candidates.

1. Simulated annealing

Temperature in an MD (or Monte Carlo) simulation is the key to overcome the barriers: States with energy E are visited with a probability of $\exp[-E/k_BT]$. If T is sufficiently large, then the system will "see" the simultaneous existence of many different minima, still spending more time in the deeper ones. By decreasing T slowly to zero, there is a good chance that the system will pick up the deepest minimum and stay trapped there. This consideration is the principle of simulated annealing: The (molecular) system is equilibrated at a certain temperature and then (slowly) cooled down to $T = 0$. While this procedure does

not guarantee that the true global minimum will be reached, it often does so. And, since no a priori assumptions are made about the optimal structure, it often yields structures that would have been difficult to foresee by intuition alone.

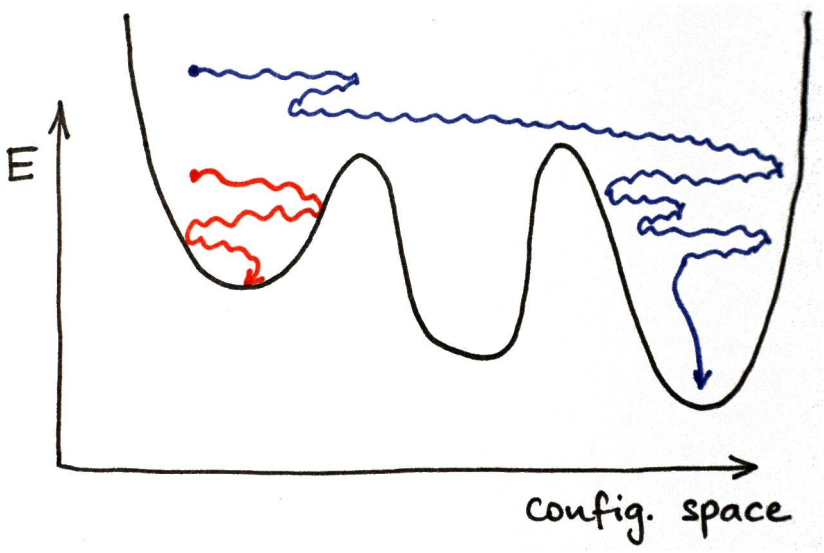


FIG. 36: Simulated annealing.

This method is often used to optimize the structure of molecular systems, but its validity is much more general: Given an objective function $Z(\alpha_1, \ldots, \alpha_N)$ depending on N parameters, one can regard each of these parameters as a degree of freedom, assign it a "mass", and let the system evolve with a molecular dynamics or Monte Carlo algorithm to perform simulated annealing. One of the early applications of this method can be found in a famous paper discussing an application to the problem of the traveling salesman (Kirkpatrick et al., Science 1983).

2. MD quenching

There is yet another possibility to make use of molecular dynamics not only to obtain the minima of the energy, but even to approximate their relative free energies (or equilibrium constants). An MD/quenching simulation consists of a usual MD trajectory, which is a basis for subsequent minimizations: In regular intervals, the structure from the simulation is subject to energy-minimization. In principle, we avoid the need to select starting structures for our minimizations – instead, we let the MD simulation take care of that.

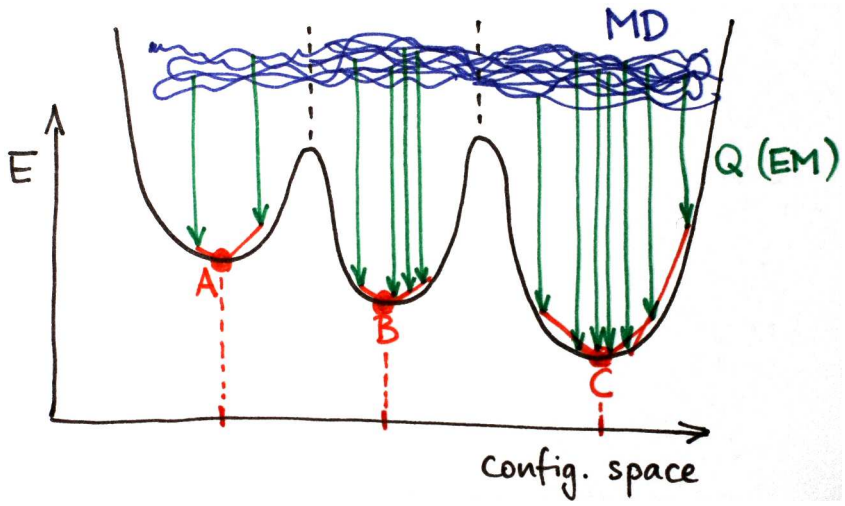


FIG. 37: MD quenching.

The obtained (possibly many) minimized structures can be processed e.g. by a cluster analysis to determine the set of unique optimal structures, their total energies and number of hits. For a small molecular system, we would observe few unique structures, each occuring many times; for larger systems, the number of unique structures would grow rapidly.

A potentially appealing feature of MD/quenching is the possibility to estimate the relative free energies of the observed structures. If the MD simulation subject to post-processing is long enough (i.e. if sufficient sampling of the configuration space is guaranteed) then the ratio of their occurence (number of hits, n_i) determines the equilibrium constant K, and thus the free energy ΔG:

$$\begin{aligned} K &= \frac{n_2}{n_1} \\ \Delta G &= -k_B T \log K = k_B T \log \frac{n_2}{n_1} \end{aligned} \tag{XII.1}$$

It is important to note that we consider whole regions of configuration space (as in Fig. X) rather than points to be individual structures. Therefore, we obtain no curves of free energy as a function of coordinate(s) but rather single values of free energy differences for certain pairs of "structures". There is an interesting, neraly philosophical question connected to this – is there something like "free energy surface" at all? Or, like obviously is the case with quenching, is it only meaningful to ask for discrete values of free energy differences?

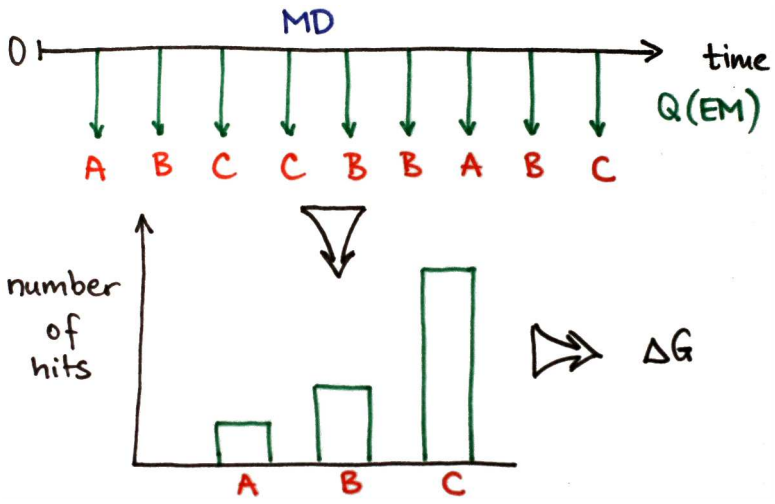


FIG. 38: MD quenching2.

B. Replica-exchange MD

Replica-exchange molecular dynamics (REMD, a.k.a. parallel tempering) is a method to accelerate the sampling of configuration space, which can be applied even if the configurations of interest are separated by high barriers. With REMD, several (identical) copies, or replicas of the molecular system of interest are simulated at the same time, with different temperatures. The essence of the method is that the coordinates together with velocities of the replicas may be switched (exchanged) between two temperatures. In practice, the probability of the replica exchange between temperatures $T_1 < T_2$ is determined in (regular) time intervals from the instantaneous potential energies U_1 and U_2 in the corresponding simulations as

$$P(1 \leftrightarrow 2) = \begin{cases} 1 & \text{if } U_2 < U_1, \\ \exp\left[\left(\frac{1}{k_B T_1} - \frac{1}{k_B T_2}\right) \cdot (U_1 - U_2)\right] & \text{otherwise.} \end{cases} \tag{XII.2}$$

Then, if $P(1 \leftrightarrow 2)$ is larger than a random number, the replicas in simulations at temperatures T_1 and T_2 are exchanged.

When using REMD, there usually one replica is simulated at the temperature of interest (often $T_1 = 300$ K) and several other replicas at higher temperatures ($T_1 < T_2 < T_3 < \ldots$). After, say, 1000 steps of MD, replica exchanges $1 \leftrightarrow 2$, $3 \leftrightarrow 4$ etc. are attempted, and after next 1000 steps the same is done for $2 \leftrightarrow 3$, $4 \leftrightarrow 5$ etc. so that only the replicas at "neighboring" temperatures can be exchanged. With such setup, the advantages of the

simulations at high temperatures – fast sampling and frequent crossing of energy barriers – combine with the correct sampling at all temperatures, above all at the (lowest) temperature of interest. Although the computational cost of REMD simulations is increased (because many simulations are running simultaneously), this additional investition of resources pays off with extremely accelerated sampling. Moreover, the simulations running at different temperatures are completely independent of each other between the points of attempted exchange, making this problem trivially (*embarassingly*) parallelizable. The first application of REMD was for a truly biophysical problem – folding of a protein (Sugita & Okamoto, Chem. Phys. Lett. 1999).

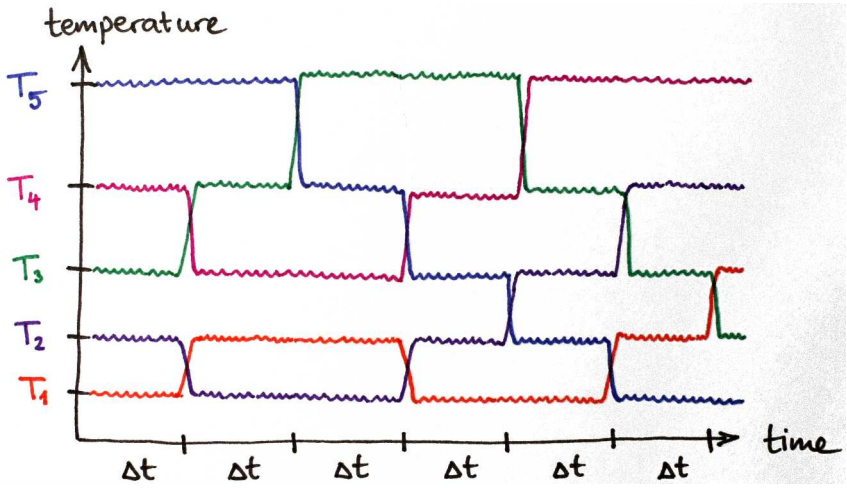


FIG. 39: Replica-exchange MD.

An important point with REMD is a suitable choice of temperatures T_i. This depends on (i) how frequent exchanges we wish (average probability $P(1 \leftrightarrow 2)$), (ii) the size of the system (the number of degrees of freedom N_{dof}) and (iii) the number of temperatures/simulations. For protein/water systems with all bond lengths constrained to their respective equilibrium values (so that $N_{\mathrm{dof}} \approx 2N$, N – number of atoms), the average probability is related to the difference of temperatures $T_2 - T_1 = \varepsilon T_1$ as

$$\overline{P(1 \leftrightarrow 2)} \approx \exp\left[-2\varepsilon^2 N\right] \tag{XII.3}$$

Using this relation, we can design the set of temperatures to suit our needs.

The REMD method can be likened to "super simulated annealing" without a need to restart. The systems at high temperatures can feed new local optimizers to the systems at

low temperatures, allowing tunneling between metastable states and improving convergence to a global optimum.

1. Replica-exchange umbrella sampling

There is an interesting application of the replica-exchange idea concerning biasing potentials rather than thermodynamic parameters (Okamoto et al., J. Chem. Phys. 2000). With the replica-exchange umbrella sampling approach (REUS), several copies of the molecular system are simulated with different biasing potentials – these are the separate umbrella-sampling simulations as presented in a previous chapter. As with the previously described REMD, an exchange of replicas with 'neighboring' umbrellas is attempted in regular intervals. Obiously, the criterion for the acceptance of a replica exchange has to be modified, and may read for instance

$$\Delta = \frac{1}{kT_1}\left(U_1(q_2) - U_1(q_1)\right) - \frac{1}{kT_2}\left(U_2(q_1) - U_2(q_2)\right) \tag{XII.4}$$

$$P(1 \leftrightarrow 2) = \begin{cases} 1 & \text{if } \Delta \leq 0, \\ \exp\left[-\Delta\right] & \text{otherwise.} \end{cases} \tag{XII.5}$$

where U_i is potential energy calculated with the energy function (including bias – umbrella) from simulation i, and q_i are the coordinates of all atoms from simulation i. With this setup, improved sampling of the configuration space and thus increased efficiency of the simulation may be expected.

It is even possible to do *multidimensional replica exchange* simulations, where the molecular system is replicated with multiple different simulation parameters – for instance, various temperatures *and* various biasing potentials.

C. Methods using biasing potentials

Using quotations by Helmut Grubmüller
(`www.mpibpc.mpg.de/home/grubmueller/projects/MethodAdvancements/ConformationalDynamics`

The energy landscapes occuring in large (bio)molecular systems feature a multitude of almost iso-energetic minima, which are separated from each other by energy barriers of various heights. Each of these minima corresponds to one particular structure ('conformational

substate'); neighboring minima correspond to similar structures. Structural transitions are barrier crossings, and the transition rate is determined by the height of the barrier.

Since in conventional MD simulations only nanosecond time scales can be covered, only the smallest barriers are overcome in simulations, and the observed structural changes are small. The larger barriers are traversed more rarely (however the transition process itself may well be fast), and thus are not observed in MD simulations.

Several approaches to remedy this drawback by way of modifying the potential energy surface of the molecular system have been proposed.

1. Conformational flooding

(Grubmüller, Phys. Rev. E 1995)

A method called 'conformational flooding' accelerates conformational transitions in MD simulations by several orders of magnitude and thereby actually can bring slow conformational transitions into the scope of simulations. From the ensemble generated by the (unbiased = normal) MD simulation, a localized artificial 'flooding potential' V_{fl} of certain (variable) strength can be constructed, meeting two requirements: (i) V_{fl} shall affect only the initial conformation and vanish everywhere outside of this region of conformational space, and (ii) it shall be well-behaved (smooth) and 'flood' the entire initial potential-energy well. A multivariate (n-dimensional) Gaussian function exhibits such a behavior:

$$V_{\mathrm{fl}} = E_{\mathrm{fl}} \cdot \exp\left[-\frac{E_{\mathrm{fl}}}{2k_{\mathrm{B}}T} \cdot \sum_{i=1}^{n} q_i^2 \lambda_i\right] \tag{XII.6}$$

where E_{fl} is the strength of the flooding potential. Here, the first n essential dynamic modes with eigenvalues λ_i will be flooded, with q_i being the coordinates along these modes.

This potential is included within subsequent 'flooding' (biased) simulations and rises the minimum of the initial conformation. Thereby, the barrier height is reduced, and the transitions are accelerated (following the theory of transition states). It is important to note that this is achieved solely by modifying the energy landscape within the minimum where the dynamics is already known and thus uninteresting; the barriers and all the other minima – which we are interested in – are not modified at all. The bottom-line is that 'conformational flooding' is expected to induce unbiased transitions, i.e. those which would be observed without the flooding potential, too, on a much longer time scale.

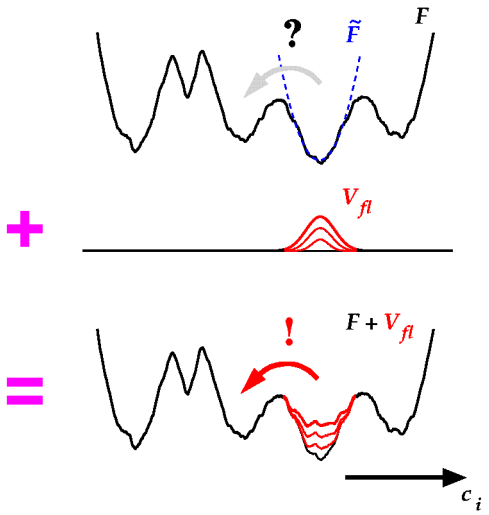


FIG. 40: Sketch of the conformational flooding (from the website of H. Grubmüller).

2. *Metadynamics*

Using quotation by Alessandro Laio (`people.sissa.it/~laio/Research/Res_metadynamics.php`)

The method is aimed at reconstructing the multidimensional free energy of complex systems (Laio & Parrinello, Proc. Natl. Acad. Sci. USA 2002). It is based on an artificial dynamics (metadynamics) performed in the space defined by a few collective variables S, which are assumed to provide a coarse-grained description of the system. The dynamics is biased by a history-dependent potential constructed as a sum of Gaussians centered along the trajectory of the collective variables. A new Gaussian is added at every time interval t_G, and the biasing potential at time t is given by

$$V_G(S(x),t) = \sum_{t'=t_G,2t_G,3t_G,\ldots} w \cdot \exp\left[\frac{(S(x)-s_{t'})^2}{2\cdot\delta s^2}\right] \tag{XII.7}$$

where w and δs are the height and the width of the Gaussians, and $s_t = S(x(t))$ is the value of the collective variable at time t. In the course of time, this potential is filling the minima on the free energy surface, i.e. the biased energy surface (sum of the Gaussians and the free energy) as a function of the collective variable(s) S is becoming constant. So, the MD protocol exhibits a kind of memory via the changing potential-energy function – a concept that was introduced earlier under the name "local elevation" (Huber et al., J. Comp. Aided Molec. Design 1994).

This approach can be exploited to explore new reaction pathways and accelerate rare

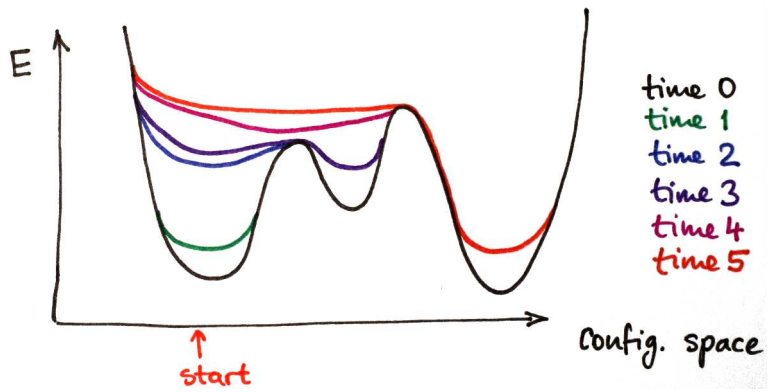


FIG. 41: Metadynamics.

events, and also to estimate the free energies efficiently. The features of metadynamics:

- The system escapes a local free energy minimum through the lowest free-energy saddle point.
- The dynamics continues, and all the free-energy profile is filled with Gaussians. At the end, the sum of the Gaussians provides the negative of the free energy. This latter statement is correct if the dynamics along S is much slower than the dynamics along the remaining (transversal) degrees of freedom.

The crucial point of the method is to identify the variables that are of interest and that are difficult to sample, since the stable minima in the space spanned by these variables are separated by barriers that cannot be cleared in the available simulation time. These variables $S(x)$ are functions of the coordinates of the system; practical applications allow the definition of up to three such variables, and the choice depend on the process being studied. We can think for instance of the principal modes of motion obtained with principal component analysis (covariance analysis, essential dynamics). However, the choice of S may be far from trivial.

The metadynamics method may be also classified as a variant of the adaptive umbrella sampling approach.

D. Locally enhanced sampling

Quotation from the Amber website, by David A. Case (`www.ambermd.org`).

Locally enhanced sampling (LES) is a mean-field technique which allows selective application of additional computational effort to a portion of the system, increasing the sampling of the region of interest (Elber & Karplus, 1990). The enhanced sampling is achieved by replacing the region(s) of interest with multiple copies. These copies *do not interact* with each other, and interact with the rest of the system in an *average way*. This average is an average force or energy from all of the individual copy contributions, not one force or energy from an average conformation of the copies.[18] A key feature is that the energy function is modified such that the energy is identical to that of the original system when all LES copies have the same coordinates.

During the simulation, the copies are free to move apart and explore different regions of conformational space, thereby increasing the statistical sampling. This means that one can obtain multiple trajectories for the region of interest while carrying out only a single simulation. If the LES region is a small part of the system (such as a peptide in solution, or a loop in a protein), then the additional computational effort from the added LES particles will be a small percentage of the total number of atoms, and the multiple trajectories will be obtained with a *small additional computational effort*.

Perhaps the most useful feature of the LES method is that it has been shown that the barriers to conformational transitions in a LES system are reduced as compared to the original system, resulting in more frequent conformational changes (Roitberg & Elber, 1991). This can be rationalized with a simple model: Imagine a protein side chain that has been replaced with 2 copies. At finite temperatures, these copies will have different conformations. Now consider the interaction of another part of the system with this region – previously, steric conflicts or other unfavorable interactions may have created high barriers. Now, however, the rest of the system sees each of these 2 copies with a scaling factor of $\frac{1}{2}$. Whereas one copy is in an unfavorable conformation, the other may not be, and the effective barrier with a distribution of copies is lower than with a single copy (as in normal MD).

Another way to consider the LES copies is that they represent an intermediate state between a normal simulation where each point in time represents a single structure, and a purely continuum model where the probability distribution of regions of interest are repre-

[18] Note the difference! The forces from all copies are calculated and their average is then taken. *No average structure* or the like is calculated.

sented by a continuous function. The atoms outside a LES region interact with that region as if it were (in the limit of many copies) a continuum, with a probability scaling given to all interactions. Therefore, the most unfavorable interactions are reduced in magnitude as compared to the original system.

Another major advantage of LES over alternate methods to reduce barriers or improve sampling is that it is compatible with current state-of-the-art simulation techniques such as molecular dynamics in explicit aqueous solvation (problems for techniques such as Monte Carlo or genetic algorithms) and the particle–mesh Ewald technique for accurate treatment of long-range electrostatic interactions. Higher temperatures can increase rates of barrier crossing, but one is then faced with issues related to solvent behavior at higher temperatures, maintaining proper densities and pressures, stability of the molecule of interest at the elevated temperature, and so on. LES gives more direct control over which regions should be enhanced, and also provides other benefits such as improvement in statistical sampling discussed above.

XIII. OTHER GENERATORS OF CONFIGURATIONS

A. MD simulation of hard bodies

The first MD simulation of a system in the condensed phase used the model of hard spheres (Alder & Wainwright, J. Chem. Phys. 1957). Representing a first step from the ideal gas model towards realistic molecules, this model has been a valuable tool above all in statistical thermodynamics, deriving e.g. equations of state and virial expansions.

1. The hard-sphere potential

The potential is a pairwise one. The potential energy of a system of two hard spheres with radius R equals zero for distances larger than the diameter of the spheres and rising above all bounds (infinity) for shorter distances when the spheres overlap:

$$V(r) = \begin{cases} 0 & \text{if } r > 2R \\ +\infty & \text{otherwise} \end{cases} \tag{XIII.1}$$

The potential is *discontinuous* and thus not differentiable, and this is different from the potentials typically used in biomolecular simulation.

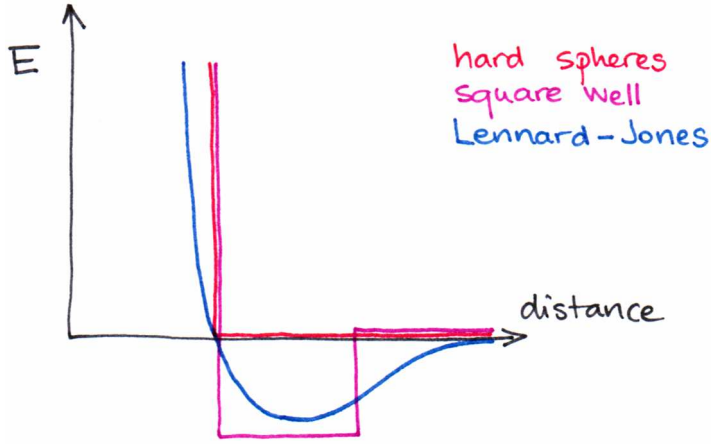


FIG. 42: The potentials of hard spheres, square well and Lennard-Jones.

If we wished to proceed further towards realistic description, however preserving the simplicity of the interaction model, we would probably opt for the so-called *square well model*, which features a region of negative potential energy (corresponding to attraction)

starting at the contact distance $2R$. Clearly, such an approximation goes in the direction of the Lennard-Jones potential, which describes the behavior of nonpolar fluid very well.[19]

Hard-convex-body potential is another extension used in statistical thermodynamics. Still, the potential energy function is discontious – zero if the bodies do not intersect and infinity if they do. The enhancement is represented by the shape of the bodies, which is not spherical anymore but rather ellipsoidal or the like. Such a shape may better describe diatomic molecules for instance.

2. Simulation protocol

As stated in the previous chapters a few times, the integration of Newton's equations motion requires the used (pair) potential to be continuous and possibly smooth (i.e. with continuous first derivative). If this is not the case, then the atoms will experience sudden 'jumps' in forces, leading to unstable simulations and wrong sampling of the configuration space, see Fig. 43.

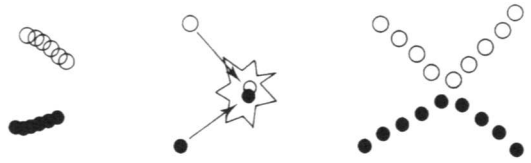


FIG. 43: If we attempted to simulate hard spheres with an integrator, we would see an explosion caused by a sudden occurence of an overlap of atoms, much the same as in the case of a simulation with continuous potential and a way too large time step (center). However, with hard spheres, an arbitrarily short time step would be still too long.

The situation with the hard-body potential is even worse, as there is an infinitely high jump of potential energy at the edge of the body (particle). What would a simulation of hard spheres with (say) the Verlet integrator look like? There are no forces in any initial configuration, and so the spheres move with their initial velocities until, all of a sudden, two spheres start to overlap. At that very moment, the energy and the forces are infinite, and the simulation crashes.

[19] This is probably the reason why physical chemists like argon so much. The simple LJ potential describes argon extremely accurately.

The simulation protocol for a system of particles interacting with a hard-body potential has to be adjusted to the discontinous character of this potential. The spheres (bodies) move along straight lines between collisions, which are perfectly elastic and instantaneous. A simulation of such a system proceeds as follows:

1. Identify the next pair of spheres (bodies) to collide, and calculate when this collision will occur.

2. Calculate the positions of all spheres at the collision time, using the principle of conservation of linear momentum and kinetic energy.

3. Determine the new velocities of the two spheres after collision.

4. Repeat from start.

Obviously, no further approximations are involved in this protocol, and a simulation will be exact within the model of hard spheres. (This is different with continuous potentials, where approximations have to be made, usually via a stepwise integration of the equations of motion.)

The potential energy is constant (zero) throughout the simulation. Thus, the conservation of total energy forces the conservation of kinetic energy, meaning that in any simulation with hard spheres, the temperature is actually constant.

B. Monte Carlo approach

In many (if not most) of the applications of molecular dynamics, the main objective is not to study how the molecular system evolves in time, but rather to generate as many configurations of the system of possible in order to sample the configuration space and estimate some thermodynamic quantites. MD is not the only possibility to do this, and we are actually free to design a method to generate the needed configurations as long as these sample the correct (e.g. canonical ensemble).

Another possibility are the Monte Carlo methods (MC). Actually, an MC technique was the first technique used to perform a computer simulation of a molecular system. The not-too-chemically sounding name comes from the crucial role that random numbers play in the MC algorithm.

1. *Monte Carlo integration*

As mentioned above, one of the major goals if molecular simulations is to calculate the thermodynamic properties. Formally, this is done by the integration over the entire configuration space. Now then, how could we use a method based on randomness to integrate a function?

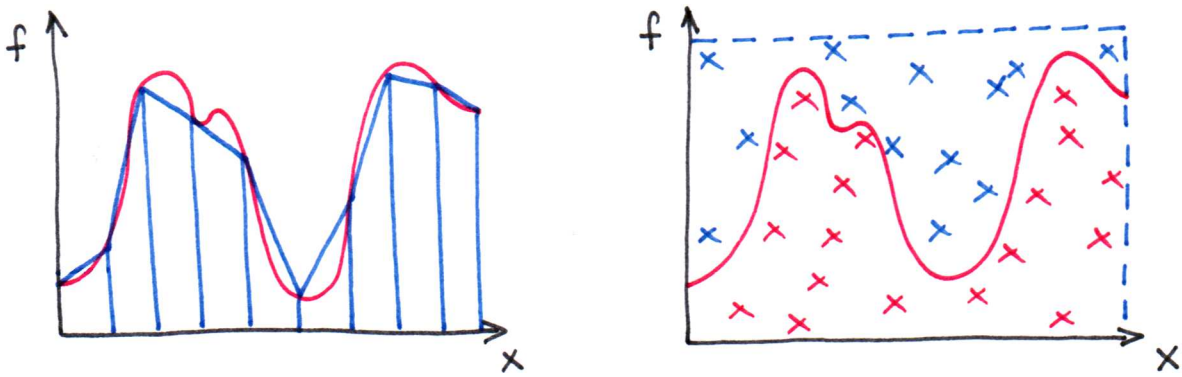


FIG. 44: Integration with the trapezoidal rule (left) and with Monte Carlo (right).

An example is shown in Fig. 44 – the task is to estimate the area under the curve, or to integrate the function. This could be done by the application of the trapezium rule. However, this method (as well as all the other commonly used ones) comes into trouble if we have to integrate a function of many variables, as is always the case in the studies of molecular systems. Here, we can make use of an alternative idea: Generate N randomly placed points within a rectangle, and count how many points (n) lie under the curve. Then, the ratio n/N approximates the ratio of area under the curve to the area of the rectangle.[20]

Importantly, it is straightforward to extend this idea to a problem in many dimensions – and we can make use of this in studies of molecular systems. Conveniently, the integration will be made even more straightforward if we are able to generate the configurations with the right probability, i.e. sampling the correct thermodynamic (e.g. canonical) ensemble. Such *importance sampling* will make it possible to average the thermodynamics quantity trivially over the ensemble of generated configurations.

[20] Apply the Monte Carlo idea to calculate π as follows: Generate pairs of random number between 0 and 1 (x, y). Count the pairs for which $x^2 + y^2 < 1$, i.e. the point (x, y) lies within the circle centered at (0,0) with a radius of 1. The ratio of this number to the total number of pairs approximates the value of $\pi/4$.

2. Metropolis' method

A typical MC simulation of a molecular system generates a sequence of configurations in an iterative way – in every iteration, one configuration is produced. Usually, a new configuration is constructed from the current one by randomly shifting a single randomly chosen atom (or, in general, particle). In practice, the new set of Cartesian coordinates is calculated with random numbers $\xi \in (0, 1)$ as

$$\begin{aligned} x_{\text{new}} &= x + (2\xi - 1) \cdot \delta r \\ y_{\text{new}} &= y + (2\xi - 1) \cdot \delta r \\ z_{\text{new}} &= z + (2\xi - 1) \cdot \delta r \end{aligned} \tag{XIII.2}$$

where δr is the maximum allowed displacement.

Then, a test is performed to inspect if this configuration shall be accepted or not. To do this, potential energy of the entire molecular system is calculated. The calculation can be optimized by realizing that only a small part of the system (a single particle) has moved since the previous iteration. Consequently, only a small part of the usually considered pair interactions changes.

The acceptance probability of the trial configuration is obtained from the current potential energy U and that of the trial configuration U_{new} as

$$P = \begin{cases} 1 & \text{if } U_{\text{new}} < U \\ \exp\left[-\frac{U_{\text{new}} - U}{k_{\text{B}} T}\right] & \text{otherwise} \end{cases} \tag{XIII.3}$$

For $P < 1$, a (pseudo)random number is drawn from the interval (0,1). The trial configuration is accepted if P is larger than this random number. If it is not the case, the trial configuration is discarded and a new one is generated by modifying the coordinates of the current configuration.

The percentage of accepted configurations (among all the generated) is governed by the maximum allowed displacement δr, which is an adjustable parameter. It is usually chosen so that $\frac{1}{3}$ to $\frac{1}{2}$ of all configurations are accepted. Such acceptance ratio was shown to lead to the most efficient sampling of the configuration space. If δr is too small, then most configurations are accepted though, but the configurations are very similar and the sampling is slow. On the other hand, if δr is too large, then too many trial configurations are rejected. Often,

δr is adjusted in the course of the simulation in order to reach a certain target acceptance ratio.

There are some modifications possible to the described recipe. Instead of selecting the atom to move randomly, it is possible to move the atoms sequentially, in a preset order. This way, one less random number per iteration has to be obtained. Alternatively, several atoms can be moved at once, instead of a single atom. With an appropriate maximum allowed displacement, this procedure may sample the configuration space very efficiently.

3. Intermezzo: generators of pseudorandom numbers

A Monte Carlo algorithm requires several random numbers to be obtained in every iteration, and since many steps have to be performed in a typical simulation (where many may mean millions or so), it is necessary to have a reliable and efficient source of random numbers. It would be most convenient to be able to 'calculate' random numbers in some way. This is actually a paradoxical requirement: computers are intrinsically deterministic devices, which are designed to deliver results that are determined by the input.

However, there are ways to generate sequences of 'pseudorandom' numbers, which are actually not random in the true meaning of the word. Still, they are independent enough of each other and have the right statistical properties, which makes them useful for MC.

Most commonly used are the *linear congruential generators*, which produce sequences of pseudorandom numbers. A following number in the sequence ξ_{i+1} is obtained by taking the previous number ξ_i, multiplying by a constant (a), adding another constant (b) and taking the remainder when dividing by yet another constant (m). Obviously, an initial value ('seed') has to be given to the generator (the system time on the computer is often used). If 'real' values are requested rather than integers, the obtained number is divided by the modulus m (to get to the interval (0,1)).

$$\begin{aligned} \xi_0 &= \text{seed} \\ \xi_{i+1} &= (a \cdot \xi_i + b) \bmod m \end{aligned} \tag{XIII.4}$$

Here, it is essential to choose 'good' values of a, b and m. If they are chosen carefully, then the generator will produce all possible values $0, \ldots, m-1$ and the sequence does not start to repeat itself until m numbers have been generated. If they are not, the sequence starts

to repeat much earlier, and there is not much randomness in there at all. A disadvantage of these generators is that the generated points in an N-dimensional space are not distributed uniformly in the space but rather lie on at most $\sqrt[N]{m}$ $(N-1)$-dimensional planes (i.e. on straight lines if we have a 2D space). If the generator is poor, the number of these planes is much smaller than $\sqrt[N]{m}$.[21]

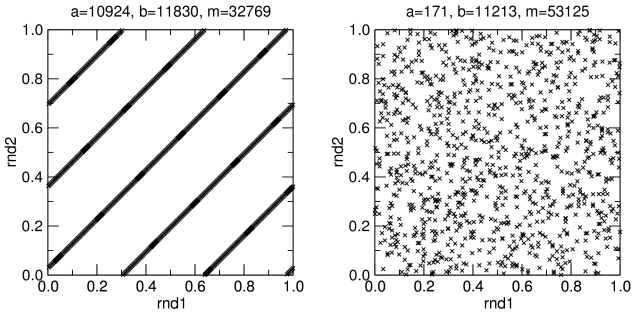


FIG. 45: A bad and a good generator of pseudorandom numbers. Each point (rnd1,rnd2) is a pair of consecutive numbers in the generated sequence.

In spite of the mentioned disadvantages, linear congruential generators are often used in MC simulations because of their extreme simplicity and thus computational efficiency. The classes of pseudorandom number generators of higher quality include the linear feedback shift register generators (LFSR) or Mersenne twister (MT). LFSR uses several bits from the current number to generate a new sequence of bits constituting a newly generated number, and it does not suffer from the cumulation of the generated numbers on hyperplanes.
MT is the current state of the art among generators and outperforms the previously mentioned e.g. by an extremely long period of $2^{19937}-1$ and no cumulation of numbers on hyperplanes in spaces with up to 623 dimensions. In a modified form, it is even suitable for cryptographic applications.

Alternative generators – from WIKIPEDIA: In Unix-like operating systems (with Linux being the first), `/dev/random` (or `/dev/urandom`) is a special file that serves as a random number generator or as a pseudorandom number generator. It allows access to environmental noise collected from device drivers and other sources.

[21] An example of such generator is RANDU: ξ_0 is odd and $\xi_{i+1} = 65539 \cdot \xi_i \bmod 2^{31}$. All generated values are odd, the period is only 2^{29} and the points $(\xi_i, \xi_{i+1}, \xi_{i+2})$ cumulate on as few as 15 planes in space.

4. Monte Carlo simulation of molecules

The easiest implementation of MC is for systems of monoatomic molecules, because it is only necessary to deal with the translational degrees of freedom. In polyatomic molecules, the implementation is more complex, and the situation is most difficult if there is much conformational flexibility in the molecules. Then, the internal degrees of freedom have to be free to vary, but this may often lead to an overlap of atoms accompanied by energy growing steeply. The ratio of acceptance of configurations would be extremely low.

It is still quite easy to simulate rigid molecules with MC. Apart from their position in space, their orientation has to be varied. This is accomplished by a rotation along one of the Cartesian axes (x, y or z) by a randomly chosen angle. There is some trigonometry to do to obtain the position of the molecule in the trial configuration.

5. Monte Carlo simulation of polymers

Many approximative models of polymers have been developed that are suitable for MC simulation. A class of convenient representations of polymers is that of a chain of monomer units, which are elementary particles (without further internal structure).

Lattice models are very simple and thus useful for very efficient studies of polymers. Here, monomer units connected with a bond can occupy neighboring lattice points in a cubic or tetrahedral lattice (Fig. 46). The used expressions for potential energy are usually very

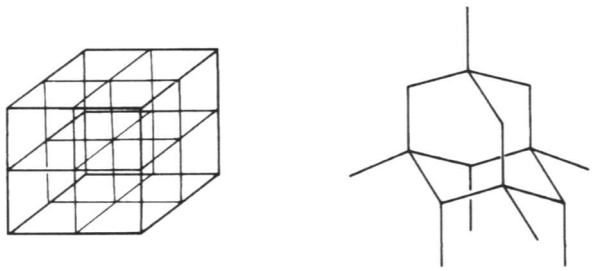


FIG. 46: Monte Carlo of a polymer – cubic (left) and diamond-like (right) lattices.

simple, which is partially forced by the simple structure of the model but also required to evaluate the energy rapidly and to sample the configuration space efficiently. An example of a more realistic and thus more complex lattice model is the 'bond fluctuation' model, where the lattice is finer-grained with respect to the bond length and the bonds between

the particles (which actually stand for several covalent bonds each) are not constrained to lie on the lattice edges (Fig. 47).

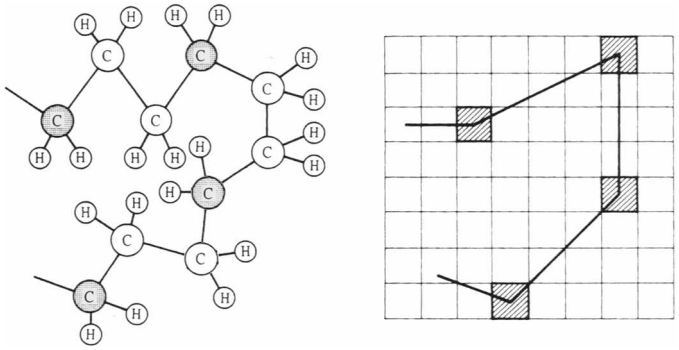


FIG. 47: Monte Carlo of a polymer – the bond fluctuation model. A single 'effective' bond in the model (right) consist of three covalent bonds along the chain of the real polymer molecule (left).

The simplest type of simulation of such a polmer chain is a *random walk*. Here, the chain grows in a random direction until the desired length is achieved. The first implementation does not consider the excluded volume of the previous segments, and the chain is free to cross itself. It is possible to evaluate various structural properties with this model, by averaging the results over many 'simulations.' For instance, the end-to-end distance R_n and the radius of gyration s_n are obtained for a chain composed of n bonds with length l as

$$\begin{aligned} \left\langle R_n^2 \right\rangle_0 &= n \cdot l^2 \\ \left\langle s^2 \right\rangle_0 &= \left\langle R_n^2 \right\rangle / 6 \end{aligned} \qquad \text{(XIII.5)}$$

While the missing description of excluded volume may seem to be a serious flaw at the first sight, this may not be always the case. In the so-called theta state, the effects of excluded volume and attractive interactions within the polymer and between the polymer and the solvent exactly cancel (also, the second virial coefficient vanishes), and the expressions derived with the simple random walk are actually valid. (The calculated parameters are often designated with the subscript '0').

The excluded volume can be taken into account by not allowing the chain to extend to the already occupied lattice points – *self-avoiding walk* (Fig. 48). This model was used to generate *all* possible configurations of a polymer of given length, in order to evaluate the partition function leading to all thermodynamic properties. The 'potential energy' may be calculated with a reasonable model of interaction of the nearby monomer units. Also,

it is possible to consider copolymers consisting of two different types of monomer units. Extreme attention has been paid to the structural properties again; an example result is the end-to-end distance in a limit of large number of elements of

$$\left\langle R_n^2 \right\rangle \approx n^{1.18} \cdot l^2 \tag{XIII.6}$$

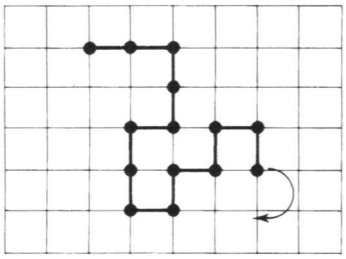


FIG. 48: Monte Carlo of a polymer – self-avoiding walk.

While it need to be difficult to generate a configuration of the polymer chain, it can be nearly impossible to modify this configuration e.g. with a MC engine, especially for densely packed polymers. A widely used algorithm in MC simulations, which is not limited to lattice models, is the *slithering snake* model. Here, one end of the polymer chain is randomly chosen as the head, and an attempt is made to connect a new monomer unit to it. If the attempt is successful, one monomer is removed from the other end. The whole procedure is then repeated.

A natural way to improve the lattice models is to leave the lattice. The simplest of such 'continuous' polymer models consists of a string of connected beads (particles), which are freely connected and interacting with each other with a sperically symmetric potential (like Lennard-Jones). Note that the beads do not generally correspond to monomer units and so the links are not the chemical bonds between monomers. The links may be either of fixed length or free to vary with a harmonic potential.

The most unrealistic property of such a model is continuous variation of link angles. The *freely rotating chain model* improves this behavior by holding the link angles fixed while allowing free rotation about the links (i.e. continuous variation of 'dihedral angles'). Obviously, this will affect the overall structure of the polymer chain compared to the freely connected one; the characteristic ratio

$$C_n = \frac{\langle R_n^2 \rangle}{n \cdot l^2} \tag{XIII.7}$$

indicating the extension of the chain will converge to the value of

$$C_\infty = \frac{1-\cos\theta}{1+\cos\theta} \tag{XIII.8}$$

with bond angle θ. For instance, $C_\infty \approx 2$ for a tetrahedral bond angle of 109°.

The *rotational isomeric state model* (RIS) by Flory (1969) improves the description by allowing every link to adopt only one of a defined set of rotational states (i.e. dihedral angles). These states usually correspond to minima of potential energy, for instance the trans, gauche(+) and gauche(-) conformations for a polyalkane chain. An elegant feature of the model is that it uses various matrices to described conformation-dependent properties. RIS is the best known one of the 'approximative' ways to describe polymer chains. It can be conveniently combined with MC simulation to estimate a wide range of properties. In such a simulation, conformations of the chain are generated with probability distributions corresponding to their statistical weights, which are a component of the RIS model (in a matrix form). With u_{ab} being the statistical weight of dihedral state b following a link in the dihedral state a, the matrix of statistical weights for an example of polyalkane chain may look like this:

$$U \equiv \begin{pmatrix} u_{tt} & u_{tg^+} & u_{tg^-} \\ u_{g^+t} & u_{g^+g^+} & u_{g^+g^-} \\ u_{g^-t} & u_{g^-g^+} & u_{g^-g^-} \end{pmatrix} = \begin{pmatrix} 1.00 & 0.54 & 0.54 \\ 1.00 & 0.54 & 0.05 \\ 1.00 & 0.05 & 0.54 \end{pmatrix} \tag{XIII.9}$$

Starting on one end of the chain, a conformation is generated by calculating the dihedral angles sequentially, until the whole chain is done. The probability of each dihedral angle is determined by the a priori probabilities of the dihedral states and on the state of the previous dihedral angle; a Monte Carlo engine is then used to select one of the values.

In a typical study, a large number of such chain will be grown, and the properties of interested will be calculated for each of them and averaged. The RIS-MC approach can be used to estimate properties like pair correlation functions (for atoms within the polymer chain), scattering functions and the force–elongation profiles.

Black-and-white figures were reprinted from Leach, Molecular Modelling.

XIV. STRUCTURE OF PROTEINS AND DRUG DESIGN

A. Basic principles of protein structure

The structure of protein molecules is not at all regular but rather far more complex. However, there are structural patterns that occur frequently. These secondary structure elements include alpha-helix and beta-strand as well as some more rarely occurring kinds of helices and several kinds of loops and turns, which exhibit certain structural patterns in spite of their generally less regular composition. These elementary structures are held together by means of hydrogen bonds. Tertiary structure is the relative orientation of secondary structural patterns, like e.g. beta barrel. Quaternary structure constitutes of the way the individual subunits of the protein – separated molecules – combine to form the native, active state of a multi-subunit protein.

The structure of a polypeptide chain can be characterized by the dihedral angles along the backbone. Ignoring the usually planar configuration on the amide bond, there are two dihedral angles per amino acid: ϕ (along the N–C$^\alpha$ bond) and ψ (along C$^\alpha$–C). The Ramachandran plot (1963) is a way to record this structure in a two-dimensional diagram (Fig. 49). In a structural analysis of a protein, any amino acids lying outside of the common regions in the Ramachandran plot would be paid special attention.

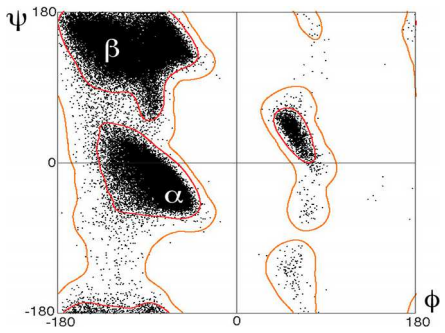


FIG. 49: Ramachandran plot obtained by an analysis of a protein databank.

Quite independently of the secondary structure elements, it is a general rule that the surface of soluble (globular) proteins is formed by polar and charged amino acids, whereas non-polar AAs (Trp, Phe, Leu, Ile, Val) tend to cumulate in the interior of the protein. This observation is said to be the consequence of the hydrophobic effect, which is one of the most important factors driving the stability of a protein. As a phenomenon, it still not

completely resolved, yet it is generally explained with entropic considerations. When the protein is folding, the free surface of the (bulky) non-polar AA side chains is decreasing. Thus, some of the water molecules that had previously formed a kind of cage around these AAs are being freed to leave to the bulk water (Fig. 50), bringing on an increase of entropy. This contribution is believed to dominate the entire free energy of the process of creation of the native structure of the protein – the folding of the protein.

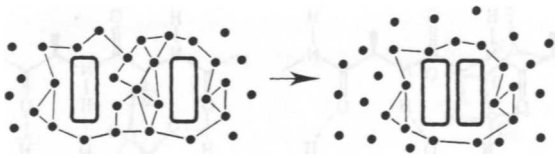


FIG. 50: The hydrophobic effect.

Another large group is the trans-membrane proteins. Typically, non-polar (hydrophobic) AA side chains are located on the surface of the protein in the membrane-spanning region, in order to match the hydrophobic character of the environment in the interior of the lipid membrane. On the other hand, charged and polar residues will be found in the parts of the protein exposed to the aqueous solution. The resolution of structure of membrane proteins is generally a very hard problem due to the extreme difficulties with crystallization of such proteins.

B. Comparative/homology modeling

Comparative modeling is a method to obtain a reasonable model of protein structure. The 3D structure is built on the basis of comparison of the sequence to that of (a) certain other (homologous) protein(s). Here, we understand the phenomenon of 'homology' as structural similarity in general, although homologous proteins are defined as such that have a common evolutionary origin. The fundamental underlying idea is that the 3D structure of proteins with similar sequence is similar. Expressed more strongly: even though the AA sequence of homo-logous proteins differs, sometimes by a seemingly large margin, their 3D structures may still be nearly identical. Obviously, this need not necessarily be the case, yet still it works often.

The procedure of creating a homology model is as follows:

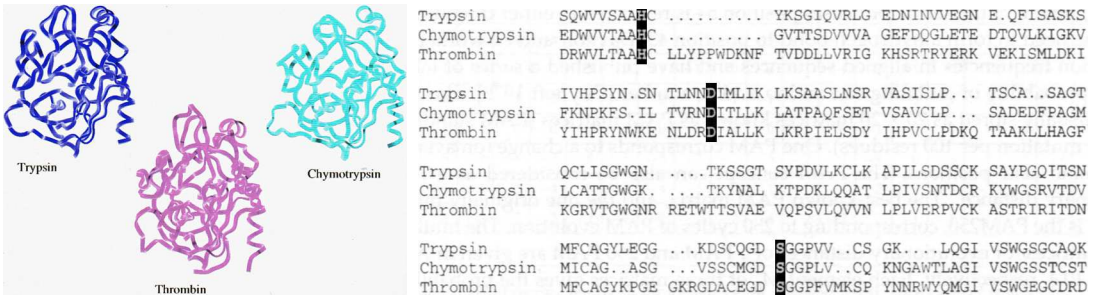


FIG. 51: The 3D structure and the AA sequence of three homologous proteins.

1. Identify a template – i.e. a protein that we consider homologous to the protein that we want to determine the structure of. There may be more than one template.

2. Produce the alignment of the sequences. Literally, the two (or more) sequences are to be laid next to each other so that their 'match' is as good as possible.

3. Identify which regions are structurally conserved between/among the sequences, and the regions of variable structure.

4. Create a model (coordinates) of the conserved region – 'core' – for the unknown structure, based on the known structure of the core of the template protein(s).

5. Generate the structure of the variable region(s) in the unknown structure. These are often fragments with no regular secondary structure, like various loops.

6. Handle the AA side chains.

7. We are done. The structure should be verified and possibly further refined with e.g. molecular mechanics.

1. Identification of the template

The basic assumption of the entire procedure is the existence of a suitable template – a protein that we expect to be structurally very similar to the unknown one. Having only the AA sequence as input, we have to rely on some kind of comparison of the sequence with a database of proteins with known 3D structure. Thus, we will take one or more proteins with certain sequence similarity with the unknown.

Also, it may be of interest to look for a possible function of an uncharacterized protein, for which only the sequence is known (for instance derived from a DNA sequence). In such a case, we would look for fragments of sequences that are strongly conserved in certain protein families – these are typically AA side chains binding a cofactor or catalytic sites.

2. Alignment of the sequences

The procedure of aligning the sequences along each other in order to obtain a best-possible match may look simple at the first sight though, but actually it is a crucial and highly non-trivial step in the development of a structural model. Several algorithms are available for alignment, and the choice of the algorithm is one of the tasks that need to be performed, together with the choice of the scoring method and the potential application of gap penalties.

The many algorithms are generally based on the so-called dynamic programming algorithm (Needleman & Wunsch, 1970). The available possibilities are FASTA, Smith-Waterman and BLASTP (which does not handle gaps). In the words of Leach, FASTA works like this (see Fig. 52):

A locate regions of identity

B scan these regions using a scoring matrix and save the best ones

C optimally join initial regions to give a single alignment

D recalculate an optimized alignment centered around the highest scoring initial region

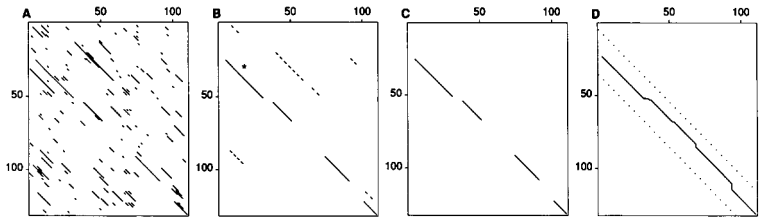


FIG. 52: Creating an alignment with FASTA.

3. Scoring of the alignment

A scoring method is used to characterize the quality of alignment of the specific sequences with a single number. In general, if an AA is identical in the aligned sequences, the contribution to the score is high, while it may be smaller if the AAs are chemically similar but not identical (conservative substitution), and it should be unfavorable if there are very different AAs aligned. There are several possibilities to perform the scoring:

- Identity – only identical AAs have favorable score.
- Genetic code – the score is given by the number of nucleobases in DNA/RNA that are needed to be changed to change one of the aligned AAs to the other.
- Chemical similarity – not only identical AAs in the aligned sequences will score favorably, but it is still OK (i.e. the score is favorable) if physico-chemically 'similar' AAs are aligned. That is, if Glu is in one sequence and Asp in the other, or two different non-polar aliphatic AAs, etc.
- Observed substitutions – this is based on the analysis of protein databases and the frequency of mutations of AAs in the alignment of sequences in these databases.

The schemes based on observed substitutions are considered to be the best choice to score alignments of sequences. An early approach is the 'Percentage of Acceptable point Mutations' (PAM, Dayhoff 1978) which give the probability of mutation of an AA to another within a certain interval of evolutionary time. Varying this time, the scoring method would find either short runs of highly conserved AAs or longer parts of the sequence with weaker similarity. Another approach is to base the scoring matrices on alignments of 3D structures rather than sequences alone; JO matrices are an example (Johnson & Overington 1993). These matrices have the potential to render the similarities of 3D structures of different sequences more sensitively – even if the sequences are formally less similar than required with other approaches. Still, there is no ultimate scoring approach that performs best for all possible alignment problems, and the selection of the scoring matrix remains non-trivial. Further, one has to decide if a global alignment shall be made (with the whole length of the sequences) or rather a local alignment, with just some fragments of the sequences; in such case, the template(s) need not be of the same length as the unknown protein.

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
Arg	R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
Asn	N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
Asp	D	5	3	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
Cys	C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Gln	Q	3	5	5	6	1	10	7	3	8	2	3	5	3	1	4	3	3	1	2	2
Glu	E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
Gly	G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
His	H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
Ile	I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
Leu	L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
Lys	K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
Met	M	1	1	1	1	0	1	1	1	1	2	3	2	7	2	1	1	1	1	1	2
Phe	F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
Pro	P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	19	6	5	1	2	4
Ser	S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
Thr	T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
Trp	W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Tyr	Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
Val	V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

TABLE I: Mutation probability matrix for 250 PAM. Each ij element give the probability (in %) that the amino acid in column j will have mutated to that in row i by the end of the period of 250 PAM. (Based on Dayhoff 1978 and Leach.)

A usual component of alignment algorithms is the *gap penalty*. It is allowed that the alignment of the sequences is discontinuous though, i.e. one or several amino acids is (are) missing in one of the sequences (or, looking from the other side, there is (are) (an) extra amino acid(s) in the other sequence(s)), but such a situation is penalized by an unfavorable contribution to the score of such an alignment. The simplest possibility is to assign each of such *indels* (insertions/deletions) a constant negative contribution. It is more common to have a penalty of the form $u + v \cdot k$ for a gap of length k AAs, with the opening penalty u being larger than the extension penalty v. Even more sophisticated are gap penalty schemes that take into account if the gap lies within a secondary structure element (alpha helix, beta sheet) or even within an active center of the protein – in such cases, the penalty would be larger than if the gap is e.g. in solvent-exposed loops or other regions on the protein surface.

4. Structurally conserved/variable regions

As soon as the alignment is available, it has to be determined which regions of the sequence will have the same 3D structure in the unknown protein as in the template(s) – conserved regions (CR) – and the variable regions (VR), which will require special treatment in the design of the 3D structure. This is more feasible if more than one template is available. CRs are usually secondary-structure elements (alpha helices, beta sheets) and sites of binding of cofactors or substrates. CRs of such character can be recognized even if only one template is used.

If more than one template is used, then the templates are first aligned with each other. The CRs are identified in the group of templates, and the alignment of the unknown protein is performed after that.

5. Create the 3D structural model

The most straightforward part here is to generate the coordinates of the main-chain atoms in the CRs – this is done simply by using the structure of the template(s). As for the side chains, the situation is still easy if the AAs are identical in the unknown protein, or if they are at least similar. If the difference of individual AAs in the CRs is more significant, then a kind of systematic approach may be used to obtain a reasonable conformation of the side chain – for instance, rotamer libraries may be used to generate the possible (most favorable) conformations of the side chain, from which the most appropriate for the specific case may be chosen.

Obviously, it is more difficult to create a good structural model for the VRs, which often correspond e.g. to the solvent-exposed loops on the surface of the protein. In those favorable cases where the sequence of the unknown protein is very similar to that in (one of) the template(s), then the VR from the template may be copied. Is this not the case, the particular sequence of AAs in the possible loop together with an additional couple of AAs on both ends of the loop may be sought among all available proteins (and not only the templates). It is quite likely here that the perfect match would not be achieved and considerable effort in application of rotamer libraries would be necessary to find a good model for the structure of the VR.

Databases of structure from comparative modeling – ModBase, SwissModel Repository. Automated web-based comparative modeling – SwissModel via the ExPASy web server, What If via the EMBL servers.

7. Evaluation and refinement of the generated structure

The structure of protein molecules on atomic level has been the focus of research of a huge number of experts in the recent decades, and a vast amount of knowledge has been accumulated on this topic. Thus, the fundamental principles of protein structure are known and quite well defined, providing us with the criteria that may be used to assess if the generated 3D structure of the unknown protein can be considered reasonable. The criteria may include:

- Conformation of the main chain in expected regions of the Ramachandran plot
- Planar peptide bonds
- Conformation of the side chains in accordance with those previously observed (rotamer library)
- Polar groups should be hydrogen bonded to a suitable partner if they are buried in the interior of the protein
- There should be a reasonable match between the hydrophilic and hydrophobic side chains (and possibly H-bonding between polar side chains and the backbone)
- No unfavorable atom–atom contacts
- No empty space (hole) in the interior of the structure. (That would be an extremely unfavorable situation.)

There are programs available to perform such an analysis – Procheck, 3D-Profiler. The output of the analysis may be not only a single determinant describing the quality of the overall 3D structure, but it can even tell which parts of the structure have been modeled probably correctly and which are suspicious or unlikely, based on the empiric criteria mentioned above.

As soon as such a simple analysis of the 3D structure has been performed and any revealed problems have been resolved, the prepared structure may be subject to further refinement. This would include energy minimization with molecular mechanics, and probably also molecular dynamics. It may be of advantage to apply constraints to the coordinates of the CRs at least at the start of the refinement, while the VRs are free to move. These constraints would be (gradually) decreased/removed during the process. Also, it is advisable to consider the solvent in these calculations (implicit/explicit, maybe PBC), and even crystallographic water molecules in the CRs of the templates can be introduced.

C. Molecular modeling in the drug design

One of the most exquisite applications of molecular modeling in the broadest sense is to construct new chemical compounds interacting in a defined way with natural materials – usually proteins but also nucleic acids, carbohydrates etc. A typical example of a task in the 'drug design' is to find a potent inhibitor of an enzyme, which does not interact harmfully with other substances in the organism. This example immediately illustrates the difficulties in drug design – mentioning just the most important requirements: the drug (medicine, for instance) has to be a potent inhibitor of the given enzyme, but it must not interact with other enzymes (which might be lethal), it must not decompose too early (before reaching the desired destination), and its metabolites must not be (too) toxic. To find a substance that meets all of these criteria is a truly hard business, and an expensive one – it is no exception that the costs to develop and test a single drug reach several hundred million euros. Although the purely experimental methodologies in this area have improved largely in the recent 20 years, involving approaches like the high-throughput screening, the exceptional amount of time and money needed to invest on the experimental side make this field an ideal target of computational approaches.

1. Molecular docking

In this chapter, we will concentrate on a typical pharmacological problem – to find a (small) molecule (ligand, guest, key) that would bind to a protein (receptor, host, lock) as strongly and specifically as possible. Thus, it is necessary (1) to generate the structure of a

complex of a known receptor (protein) and an up to this point unknown compound, and (2) to evaluate this structure. A good news is that the binding site – binding pocket – is usually known, as it is often the active or allosteric place of the protein that is to be inhibited.

Otherwise, there is bad news. The problem has many degrees of freedom – translation and rotation of the ligand as well as its internal flexibility; the relaxation of protein structure may be often neglected (although not always). A single molecule (or a small number of molecules) can be docked manually, once the binding mode of a similar molecule is known. It should be noted that even such a straightforward approach may fail, as even similar molecules may sometimes bind in different ways (and with different strength).

There is a sequence of tasks to accomplish, fairly similar to that in the search for the structure of a protein, indeed:

1. Take the compounds to test from somewhere – database of compounds, construction from a database of moieties,...

2. For a selected compound, place the molecule in the binding site in the most favorable way – orientation and conformation (if applicable – nearly always).

3. Evaluate the strength of the orientation. Accurate determination of binding free energy is impossible, and so some kind of scoring is desired.

Various levels of approximation may be employed when searching for a molecule that would fit a binding pocket. The simplest approach is to process a database of molecules and consider each of them as rigid body; this would be attempted to fit into a rigid binding pocket in the protein. This is the essence of the action of the Dock program, which first creates a 'negative image' of the binding pocket as a unification of several spheres, and then looks which molecule(s) would fit this shape best.

A natural expansion of this approach is to consider the flexibility of the ligand in some way. To do so, it is possible to apply any means of exploring the configuration space of the molecule – be it Monte Carlo, sometimes in conjunction with simulated annealing, simple minimization of molecular dynamics. A simple (and thus robust) force field would be used with any of these generators of configurations. Alternatively, a quite efficient approach is the incremental construction of the ligand. Here, the ligand is partitioned into chemically reasonable fragments; the first fragment is docked into the binding site in a usual way, and

the other fragments are 'grown' consecutively. This provides a natural possibility to account for the conformational flexibility of the molecule, regarding the relative orientation of the individual fragments.

We can already see the problem of docking very well – not at all surprisingly, it is all about sampling. There is no way to try to do MD for every candidate molecule, because (1) MD takes much longer than we can afford having to process a lot of molecules, and (2) MD could work probably only for quite rigid molecules and a binding pocket which does not constrain the movement of the ligand, which is usually not the case. If our goal is to dock a single, specific molecule, we can afford a particularly thorough search that would probably involve MD, possibly with a king of biasing potentials. However, if we have to dock and assess many candidate ligands, simpler approaches have to be chosen. The current state of the art is to consider the flexibility of the ligands, while ways to describe the conformational flexibility of the protein (on the level of side chains) are under development.

2. Scoring functions for docking

If we have a plenty of molecules to evaluate, we need an extraordinaly efficient way to quantify the strength of the binding in order (1) to find the right binding mode of each ligand, and (2) to compare the strength of binding of various ligands. So, the quantity of interest here is the binding free energy. We know many methods to evaluate free energies, but the problem is that these procedures are many orders of magnitude slower than required for docking. What we need here is a simple additive function to approximate the binding free energy, which would give a result rapidly, in a single step. Such *scoring function* would have the form

$$\Delta G_{\text{bind}} = \Delta G_{\text{solvent}} + \Delta G_{\text{conf}} + \Delta G_{\text{int}} + \Delta G_{\text{rot}} + \Delta G_{\text{t/r}} + \Delta G_{\text{vib}} \qquad \text{(XIV.1)}$$

$\Delta G_{\text{solvent}}$ covers the change of hydration effects during the binding reaction – the different hydration of the isolated ligand and protein and that of the complex. ΔG_{conf} describes the change of conformation of the ligand upon binding – the 'deformation energy'; the shape of the binding pocket may constrain the ligand to another conformation than what is favored with a free ligand, and this costs energy. (The conformation of the protein is usually ignored, or considered unchanged.) ΔG_{int} – the 'interaction energy' – a favorable contribution to free

energy stemming from the specific interactions between the ligand and the protein. ΔG_{rot} is the loss of entropy ($\Delta G = -T \cdot \Delta S$) brought about by the frozen rotations around single bonds within both the bound ligand and the protein. It is possible to approximate this contribution as $+RT \log 3 = 0.7$ kcal/mol per rotatable bond with three equienergetic states (trans and 2x gauche). $\Delta G_{\mathrm{t/r}}$ is the loss of translational and rotational entropy upon association of two molecules, which is approximately constant for all ligands of similar size, therefore it need not be considered when comparing the ligands. ΔG_{vib} should describe the change of vibrational modes, which is difficult to estimate and is often ignored.

As a result, there is a kind of force field for the free energy of binding. The problem is that in spite of its approximative character, this expression may be still computationally too costly to evaluate for a huge number of ligands that is usually to be processed. For this reason, the many ways proposed so far to estimate the contributions to the free energy are usually very simple, looking over-simplified in comparison with molecular-mechanics force fields. An illustrative example of such a simplified approach is the following equation:

$$\begin{aligned}\Delta G = \Delta G_0 \; + \; \Delta G_{\mathrm{Hbond}} \cdot \sum_{\mathrm{Hbonds}} f(R,\alpha) + \Delta G_{\mathrm{ionpair}} \cdot \sum_{\mathrm{ionpairs}} f'(R,\alpha) \\ + \; \Delta G_{\mathrm{lipo}} \cdot A_{\mathrm{lipo}} + \Delta G_{\mathrm{rot}} \cdot N_{\mathrm{rot}}\end{aligned} \tag{XIV.2}$$

where ΔG_0 – a constant term; $\Delta G_{\mathrm{Hbond}}$ corresponds to an ideal hydrogen bond, and $f(R,\alpha)$ is a penalty function for a realistic hydrogen bond of length R and angle α; analogic quantities ($\Delta G_{\mathrm{ionpair}}$ and $f'(R,\alpha)$) apply for ionic contacts. ΔG_{lipo} is the contribution from hydrophobic interaction, considered as proportional to the area of the non-polar surface of the molecule A_{lipo}; N_{rot} is the number of rotatable bonds in the ligand that are being frozen upon binding, contributing ΔG_{rot}. (Böhm, 1994).

A number of similar functions followed this study. These involved for instance the partitioning of the surface areas of both the proteins and the ligand into polar and non-polar regions, and assigning different parameters to the interactions of different kinds of regions (polar-polar, polar-nonpolar, nonpolar-nonpolar). Also, statistical techniques were used to derive the scoring function and the parameters. The problem is that these functions only describe well those ligands that bind tightly to the protein. Modestly binding ligands, which are of increasing interest in docking studies, are more poorly described by such functions. A possibility to resolve this issue is the ‘consensus scoring’ – combining results from several scoring functions, which was shown to perform better than any single scoring functions.

In all studies aimed at the strength of binding expressed in terms of the binding free energy, it must be constantly borne in mind that a change (error) of binding free energy of 1.4 kcal/mol corresponds to a ten-fold in/decrease of equilibrium constant of binding. In other words, as little as 4.2 kcal/mol of binding free energy lies between a micro- and a nanomolar inhibitor of a protein, which is figuratively infinite difference. Therefore, the requirements on the accuracy of the scoring function are actually quite big.

3. De novo design of ligands

While it is very often useful to search a database of molecules for a suitable ligand, there is still a chance to miss the 'ideal' ligand simply because no such compound has been included in the database. To avoid such a failure, it may be a better choice to construct the ligand 'from scratch' – without relying on the content of a database. There are two basic types of the de novo design: In the 'outside–in' approach, the binding site is first analyzed and tightly-binding ligand fragments are proposed. Then, they are connected together, possibly using a database of molecular linkers, providing a molecular skeleton of the ligand, which may be converted to an actual molecule. The 'inside–out' approach constitutes of 'growing' the ligand in the binding pocket, driven by a search algorithm with a scoring function.