Biomolecular modeling

Marcus Elstner and Tomáš Kubař

Theoretical Chemistry, TU Braunschweig

(Dated: December 10, 2010)

V. NON-BONDED INTERACTIONS

There are several reasons to take particular care of the non-bonded interactions:

- They are a key to understand the structure, function and in particular the efficiency of action of many proteins. It is the electrostatic and/or van der Waals interaction of the protein with the ligand that is responsible for the efficiency of a reaction, color of the chromophore etc. The solvent surrounding is co-responsible for the particular structure of nucleic acids, polypeptides and proteins (hydrophobic-hydrophilic residues).

- The non-bonded interactions are treated in MM by two-body potentials, and the computational effort of $\mathcal{O}(N^2)$ dominates the overall requirements for large molecular systems. Therefore, the non-bonded (above all, the long-range electrostatic) interactions represent a good target for optimizations.

- Solvation plays a crucial role in determining the structure and function of biomolecules. However, the necessary amount of water is often extremely large, becoming the main source of computational cost.[1] Therefore, there is a need to efficiently describe the solvation effects, which are of a predominantly electrostatic character (due to the large dipole moment of the water molecule).

A. Introduction to electrostatic interaction

The electrostatic interaction energy of two point charges q and Q separated by a distance r is given by Coulomb's law

$$E^{\mathrm{el}} = \frac{1}{4\pi\varepsilon_0} \cdot \frac{q \cdot Q}{r} \tag{V.1}$$

Of importance is the concept of *electrostatic potential* (ESP), induced at the point $\vec{r}$ by a point charge Q located at $\vec{r}_1$:

$$\Phi(\vec{r}) = \frac{1}{4\pi\varepsilon_0} \cdot \frac{Q}{|\vec{r} - \vec{r}_1|} \tag{V.2}$$

[1] Typically, the simulated molecular system consists from more than 90 % of water, so that more than 80 % of the computational time is spent by calculating the forces among the (uninteresting) water molecules around the (interesting) solute.

If we know the electrostatic potential at a point $\vec{r}$ in space, then we can obtain the total electrostatic energy of a charge q at this point:

$$E^{\mathrm{el}}(\vec{r}) = \Phi(\vec{r}) \cdot q \tag{V.3}$$

In this way, we can have an 'electrostatic potential energy surface' in analogy to mechanics. There, if we know the topography of the Alps, then we immediately know the potential energy of a person with a weight of 70 kg, at any point. In a similar way, if we know the electrostatic potential induced by the atoms of a protein, then we can readily obtain for instance the binding energy of a point charge (like a metal cation) at any place. The electrostatic potential induced by a number of point charges Q_i follows simply as a sum

$$\Phi(\vec{r}) = \frac{1}{4\pi\varepsilon_0} \sum_i \frac{Q_i}{|\vec{r} - \vec{r}_i|} \tag{V.4}$$

with the energy of a point charge q at $\vec{r}$ given by Eq. V.3.

In case of a continuous charge distribution, we have to consider the charge density $\rho = Q/V$, with $\rho(\vec{r})$ being the charge density at the point $\vec{r}$. Then, $Q_i = \rho(\vec{r}_i) \cdot V_i = \rho(\vec{r}_i) \cdot \Delta V$ is the charge in the i-th volume element V_i. Summing over all elements, one obtains

$$\Phi(\vec{r}) = \frac{1}{4\pi\varepsilon_0} \sum_i \frac{\rho(\vec{r}_i) \cdot \Delta V}{|\vec{r} - \vec{r}_i|} \tag{V.5}$$

If we make the volume elements infinitesimally small, this changes to (with $\mathrm{d}^3\vec{r} = \mathrm{d}V$)

$$\Phi(\vec{r}) = \frac{1}{4\pi\varepsilon_0} \int \frac{\rho(\vec{r}_1)}{|\vec{r} - \vec{r}_1|} \, \mathrm{d}^3\vec{r}_1 \tag{V.6}$$

Finally, the electrostatic energy of a charge distribution $\rho(\vec{r})$ follows as

$$E = \frac{1}{2} \int \Phi(\vec{r}) \cdot \rho(\vec{r}) \, \mathrm{d}V = \frac{1}{8\pi\varepsilon_0} \iint \frac{\rho(\vec{r}_1) \cdot \rho(\vec{r})}{|\vec{r} - \vec{r}_1|} \, \mathrm{d}^3\vec{r} \, \mathrm{d}^3\vec{r}_1 \tag{V.7}$$

The main task is to get the electrostatic potential from a charge distribution. For that, one has to solve **Poisson's equation**

$$\nabla^2 \Phi(\vec{r}) = -\frac{\rho(\vec{r})}{\varepsilon} \tag{V.8}$$

(differential equation for Φ as a function of $\vec{r}$), or, if the permittivity ε is not constant,

$$\nabla \left(\varepsilon \nabla \Phi(\vec{r}) \right) = -\rho(\vec{r}) \tag{V.9}$$

As an example let us have a look at the ESP of a gaussian charge distribution. This distribution centered around the origin of coordinate system with a width σ is given as

$$\rho(r) = Q \cdot \frac{1}{\sigma^3 \sqrt{2\pi}^3} \cdot \exp\left[\frac{r^2}{2\sigma^2}\right] \tag{V.10}$$

The corresponding solution of the Poisson equation is

$$\Phi(r) = \frac{1}{4\pi\varepsilon} \cdot \frac{Q}{r} \cdot \mathrm{erf}\left[\frac{r}{\sqrt{2}\sigma}\right] \tag{V.11}$$

with erf being the error function. Here, if we move far enough from the center of the charge distribution (r is large), the error function converges to unity and the ESP is very near to that of a point charge placed in the origin (Eq. V.2). This is in accord with experience – a point charge and a well-localized charge distribution interact with distant charges in the same way. Actually, we need not go so far in order to see that – the error function assumes a value of 0.999 already at the distance of 2.4σ.

B. Periodic boundary conditions

The most frequent objective of MD simulations is to describe a molecular system in aqueous solution. The problem that we readily encounter is that we have to make the system as small as possible, in order to reduce the computational cost. The most straightforward way to do so is to consider only a single molecule of the solute (e.g. a protein or DNA species) with a smallest possible number of solvent (water) molecules. A typical size of such a system with several thousand water molecules is in the range of units of nanometer. Here, a serious issue occurs: while we are trying to describe the behavior of a molecule in *bulk solvent*, every point in such a small system is actually very close to the *surface*. The surface layer of a system has always properties very different from those of the bulk phase, and with such a setup, we would simulate something else that what we aim at.

An elegant way to avoid this problem is to implement the **periodic boundary conditions** (PBC). Here, the molecular system is placed in a box with a regular geometrical shape (the possibilities are listed below). This polyhedron is virtually replicated in all spatial directions, with *identical* positions (and velocities) of all particles, as shown in Fig. 1. This way, the system is made *infinite* – there is no surface in the system at all. The atoms in the vicinity of the wall of the *simulation cell* (like the black circle in Fig. 1) interact with the atoms in the neighboring replica.

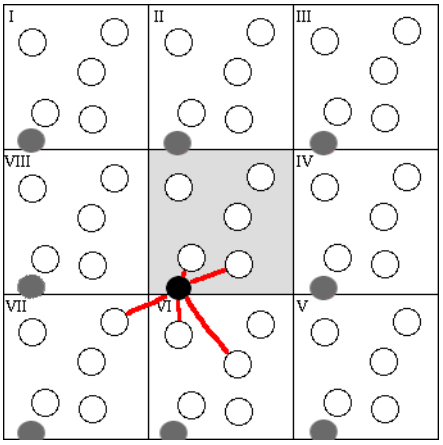


FIG. 1: Replication of the unit cell (grey) using periodic boundary conditions. Interactions of an atom (black) with the nearest images of all other atoms (red).

This method is not quite perfect as it introduces artificial periodicity in the system – all the replicas of the simulation cell look the same, making the thermodynamics of the system incorrect in principle.[2] However, this treatment is much better than simulating a too small system with artificial boundary with vacuum.

Practically, the implementation has the following features:

- Only the coordinates of the unit cell are recorded.
- If a particle leaves the box, then it enters the box from the other side.
- Carefull accounting of the interaction of the particles is necessary. The simplest approach is to make an atom interact only with the $N-1$ particles within the closest periodic image, i.e. with the nearest copy of every other particle (**minimum image convention**). This is to avoid the interaction of an atom with two different images of another atom, or even with another image of itself. If the box is cubic with boxsize L, then each atom can interact only with all atoms closer than $L/2$. Evidently, PBC have to be synchronized with the applied cut-offs, see below.

The unit cell may have a simple shape – cubic or orthorhombic, parallelepiped (specially, rhomboeder), or hexagonal prism; but also a more complicated like truncated octahedral

[2] For instance, the *entropy* of the entire system is obviously too small, because of its (wrong) translational symmetry. As a general rule, this is rarely a problem.

or rhombic dodecahedral. In the latter two cases, the corresponding PBC equations are quite complicated; the advantage of such shapes for the simulation of spherical objects (like globular proteins in solvent) is that there are no voluminous distant corners which increase the amount of solvent and thus the computational complexity (like in the case of cubic /orthorhombic box). Two-dimensional objects like phase interfaces are usually treated in a *slab* geometry.

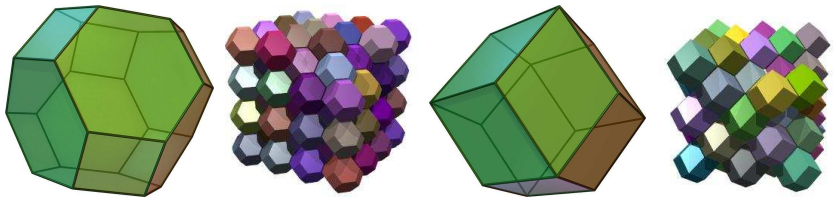


FIG. 2: Truncated octahedron (left) and rhombic dodecahedron (right).

C. Accelerating the calculation of non-bonded interactions – cut-off

As mentioned above, the evaluation of non-bonded terms becomes a bottleneck for large molecular systems, and in order to make simulations of extended systems possible, it is necessary to limit their computational cost.

The simplest and crudest approach is to neglect the interaction of atoms that are further apart than a certain distance r_c. This so-called *cut-off* is commonly used with the rapidly decaying $(1/r^6)$ Lennard-Jones interaction, which indeed nearly vanish already for moderate distances r_c, commonly around 10 Å. However, with the slowly decaying electrostatic interaction $(1/r)$, this would lead to a sudden jump (discontinuity) in the potential energy; even worse, this would be a disaster for the forces, which would become infinite at that point.

A better idea would be to *shift* the whole function by $V(r_c)$, so that there is no jump at r_c anymore. We would have

$$V^{\mathrm{sh}}(r) = \begin{cases} V(r) - V(r_c), & \text{for} \quad r \leq r_c, \\ 0, & \text{otherwise.} \end{cases} \tag{V.12}$$

However, the gradients (forces) are at r_c still not continuous! To eliminate this force jump,

it is possible to apply a *shift-force* potential ($V' \equiv \mathrm{d}V/\mathrm{d}r$):

$$V^{\mathrm{sf}}(r) = \begin{cases} V(r) - V(r_c) - V'(r_c) \cdot (r - r_c), & \text{for} \quad r \leq r_c, \\ 0, & \text{otherwise.} \end{cases} \tag{V.13}$$

The obvious drawback of this method is that the Coulomb energy is changed!

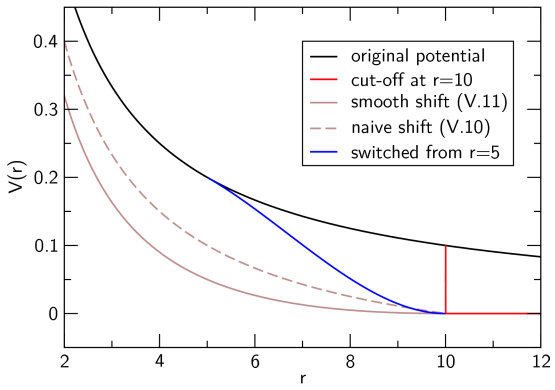


FIG. 3: Electrostatic interaction energy of two unit positive charges, evaluated using Coulomb's law and the various modifications.

A further alternative is the *switch potential*. Here, an additional potential is applied starting from a certain distance r_1, which brings the Coulomb interaction gradually to zero, as shown in Fig. 3. The drawback of this method is, that the forces are altered in the cut-off region.

Both methods can be applied to either the energy or the forces: when applied to the energy, the forces follow through differentiation, and vice versa, when applied to forces, the energy follows through integration.

Generally, the cut-off schemes can be based either only on atomic distances, or on functional groups. Usually, the latter is employed to assure charge conservation in the Coulomb interaction.

D. Accelerating the calculation of electrostatic interactions – Ewald summation

In many cases, above all if highly charged molecular systems (like DNA or some proteins) are simulated, the use of cut-offs is a bad approximation. For instance, the artificial forces if using a switching function may lead to the accumulation of ions in the regions of solution in the cut-off distance (measured from DNA). Therefore, it is desirable to abandon the minimum image convention and the cut-offs, and rather sum up the long-range Coulomb interaction between *all* the replicas of the simulation cell

Let us introduce a vector $\vec{n}$, which runs over all the replicas of the cell, denoting them uniquely:

- For $|\vec{n}| = 0$, we have $\vec{n} = (0, 0, 0)$ – the central unit cell.
- For $|\vec{n}| = L$, we have $\vec{n} = (0, 0, \pm L)$, $\vec{n} = (0, \pm L, 0)$, $\vec{n} = (\pm L, 0, 0)$ – the six neighboring unit cells.
- Further, we continue with $|\vec{n}| = \sqrt{2} \cdot L$ and the 12 cells neighboring over an edge, etc.

With this vector, we can write the sum of Coulomb interactions over all replicas as

$$E^{\text{Coul}} = \frac{1}{2} \sum_{i,j} \sum_{\text{images } \vec{n}} \frac{q_i \cdot q_j}{|\vec{r_{ij}} + \vec{n}|} \tag{V.14}$$

for indices i and j running over all atoms in the unit cell (r_{ij} is then their distance). This expression is an infinite sum which has special convergence problems. Such a sum decays like $1/|\vec{n}|$ and is a *conditionally convergent* series, meaning that it converges ($\sum_{i=1}^{\infty} a_i < \infty$) but does not converge absolutely ($\sum_{i=1}^{\infty} |a_i|$ cannot be summed up). The problem is that the convergence of such a sum is slow and, even worse, dependent on the order of summation. So, a conditionally convergent series may add up to *any* (!) value, as shown in this example:

$$\begin{aligned}
\text{I}: \quad S &= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \dots \\
\text{II}: \frac{1}{2}S &= \frac{1}{2} - \frac{1}{4} + \frac{1}{6} - \frac{1}{8} + \dots \\
\text{I} + \text{II}: \frac{3}{2}S &= 1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} + \frac{1}{11} - \frac{1}{6} + \dots \\
&= 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \dots = S \quad \text{(sic!)}
\end{aligned} \tag{V.15}$$

Therefore, we need a clever way to evaluate the potential resulting from the interaction of all images of all charges

$$\Phi(\vec{r_i}) = \sum_j \sum_{\text{images } |\vec{n}|} \frac{q_j}{|\vec{r_{ij}} + \vec{n}|} \tag{V.16}$$

in order to evaluate the Coulomb energy of the charges q_i in the unit cell.

$$E^{\text{Coul}} = \frac{1}{2} \sum_i q_i \cdot \Phi(\vec{r_i}) \tag{V.17}$$

The idea of the Ewald methods is to convert the difficult, slowly convergent series to the sum of two series, which both converge much more rapidly, like

$$\sum \frac{1}{r} = \sum \frac{f(r)}{r} + \sum \frac{1 - f(r)}{r} \tag{V.18}$$

where $\sum 1/r$ represents the difficult series that we have to deal with. Whereas the terms on the right-hand side look more complicated, they actually exhibit a much more rapid convergence than $\sum 1/r$ in our case, and such an awkwardly looking 'decomposition' is the preferred way to go.

Since the summing over point charges leads to convergence problems with conditionally convergent sums, the Ewald method uses rather *normal distributions* of charge of the same magnitude:

$$q_j \;\rightarrow\; q_j \cdot \left(\frac{\alpha}{\sqrt{\pi}}\right)^3 \exp\left[-\alpha^2 \cdot |\vec{r_j}|^2\right] \tag{V.19}$$

To get the electrostatic potential induced by this smeared charge distribution, Poisson's equation (Eq. V.8) has to be solved. This leads to the potential being represented by the so-called *error function*:[3]

$$\Phi(\vec{r}) = q_j \cdot \frac{\text{erf}\left[\alpha \cdot r\right]}{r} \tag{V.20}$$

[3] The error function is defined as the definite integral of the normal distribution

$$\text{erf}[x] = \frac{2}{\sqrt{\pi}} \int_0^x \exp[-t^2]\,\mathrm{d}t$$

and the complementary error function as

$$\text{erfc}[x] = 1 - \text{erf}[x]$$

and, in the special case of $\vec{r} = \vec{o}$:

$$\Phi(\vec{o}) = q_j \cdot \frac{2\alpha}{\sqrt{\pi}} \tag{V.21}$$

If we sum up the potentials given by Eq. V.20 for all charges, we obtain

$$\Phi(\vec{r_i}) = \sum_j \sum_{\text{images}\,|\vec{n}|} q_j \cdot \frac{\text{erf}\,[\alpha \cdot |\vec{r_{ij}} + \vec{n}|]}{|\vec{r_{ij}} + \vec{n}|} \tag{V.22}$$

This has to be compared with the potential induced by the point charges (Eq. V.16). The difference between Eq. V.16 and Eq. V.22 is given by the complementary error function. The genuine potential induced by the point charges can then be expressed as

$$\Phi(\vec{r_i}) = \sum_j \sum_{\text{images}\,|\vec{n}|} q_j \cdot \frac{\text{erfc}\,[\alpha \cdot |\vec{r_{ij}} + \vec{n}|]}{|\vec{r_{ij}} + \vec{n}|} \tag{V.23}$$

$$+ \sum_j \sum_{\text{images}\,|\vec{n}|} q_j \cdot \frac{\text{erf}\,[\alpha \cdot |\vec{r_{ij}} + \vec{n}|]}{|\vec{r_{ij}} + \vec{n}|} \tag{V.24}$$

The first term V.23 called the *real-space contribution* is shown graphically in Fig. 4 (top). From a certain (quite small) distance, the point charges and the gaussian charge distributions (with opposite signs) cancel each other. This distance depends on the gaussian width α – a small gaussian width would lead to a rapid convergence.

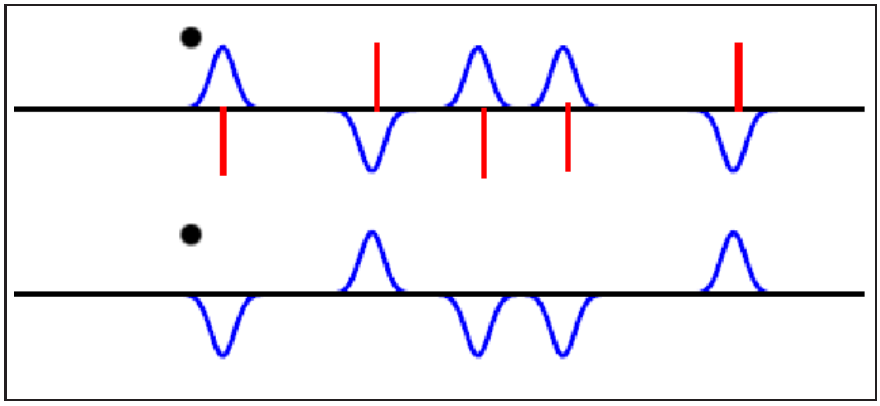


FIG. 4: Top: Real-space contribution to the Ewald sum consists of the original point charges (red) and gaussian charge distributions (blue) of the same magnitude but opposite sign. Bottom: Reciprocal-space contribution.

On the other hand, the second term V.24 called the *reciprocal-space contribution* is best evaluated in the form ($\vec{k}$ – the reciprocal lattice vector of periodic images)

$$E^{\text{rec}} = \frac{1}{2V\varepsilon_0} \cdot \sum_{\vec{k} \neq \vec{o}} \frac{1}{\vec{k}^2} \cdot \exp\left[-\frac{|\vec{k}|^2}{4\alpha^2}\right] \cdot \left| \sum_j q_j \cdot \exp[-\text{i} \cdot \vec{k} \cdot \vec{r_j}] \right|^2 \tag{V.25}$$

The usually applied *Fourier transform* techniques[4] need a large gaussian width α for fast convergence, therefore the value of α is a necessary compromise between the requirements for the real- and reciprocal-space calculations. All in all, both mentioned contributions exhibit quite favorable convergence behavior, making the evaluation of the electrostatic potential due to all periodic images feasible.

After calculating these two terms (Fig. 4), yet another one has to be taken into account: Since we have broadened charge distributions, they do interact with themselves, as shown in Fig. 5. This interaction has been brought about by the substitution of point charges by gaussian charge distributions, and thus it must be subtracted from the final result.

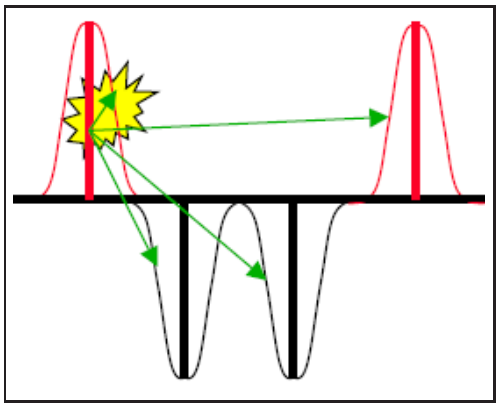


FIG. 5: Interaction of the charge with the gaussian distribution

The potential of a broadened gaussian is given by Eq. V.21, which leads to Coulomb energy of

$$E^{\text{self}} = \sum_j q_j \cdot \Phi(\vec{o}) = \sum_j q_j \cdot q_j \cdot \frac{2\alpha}{\sqrt{\pi}} \tag{V.26}$$

At the end, we have three energy contributions: one from the 'real-space' evaluation of Φ^{real} in Eq. V.23, which gives

$$E^{\text{real}} = \frac{1}{2} \sum_j q_j \cdot \Phi^{\text{real}}(\vec{r_j}) \tag{V.27}$$

one from the 'reciproal-space' evaluation of Φ^{rec} in Eq. V.25 and the 'self-energy' in Eq. V.26, so that

$$E^{\text{Ewald}} = E^{\text{real}} + E^{\text{rec}} - E^{\text{self}} \tag{V.28}$$

[4] A popular implementation is the FFTW (Fastest Fourier Transform in the West), with a favorable computational cost scaling as $\mathcal{O}(N \cdot \ln N)$.

E. Explicit solvent models – water

The most simulations of biomolecules are performed with water as the solvent, to mimic the physiological or *in vitro* experimental conditions. If a not too concentrated solution is to be simulated, then the necessary amount of the solvent is quite large, often many thousand molecules.

For instance, in a typical simulation of a DNA oligomer with ten base pairs (see Fig. 6), the dimensions of the PBC box are $3.9 \times 4.1 \times 5.6$ nm, and there are 630 atoms in the DNA molecules, 8346 atoms of water and 18 sodium counterions. The macroscopic concentration of DNA in this 'sample' reaches an astonishingly large value of 18 mmol/L![5] At the same time, 86 % of all *pair interactions* are those where each of the partner atoms belongs to a water molecule,[6] and the most remaining interactions involve one atom of water. This is a huge portion, and the smallest possible at the same time, as we have the minimal number of water molecules.

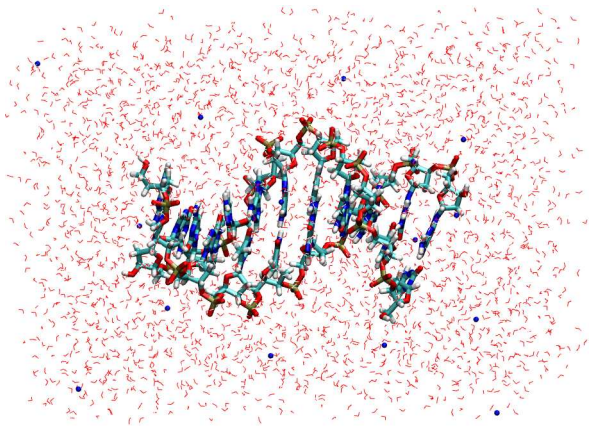


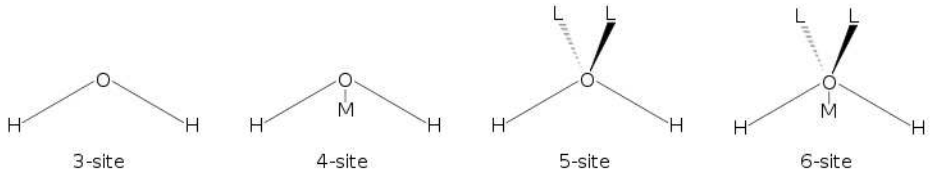
FIG. 6: Typical setup of the simulation of a DNA oligomer.

We can see that the most interactions involve water, and that is why it is necessary to turn our attention to the description of water in the simulations. The model of water must be made simple enough in order to reduce the computational complexity, but at the same time it is necessary not to compromise the accuracy of the description.

[5] Due to the commonly accepted criteria, such a box is the smallest possible. Thus, the amount of water is also the smallest possible, and the concentration the highest possible.

[6] There are 8346 water atoms, that is roughly 8346^2 interactions water–water, and 8994 atoms altogether, corresponding to 8994^2 pair interactions. The ratio of these figures is 0.86.

Many simple water models have been developed so far. They are usually *rigid*, so that the bond lengths as well as angles remain constant during the simulation. A molecule is composed of at least three sites (corresponding to atoms in this case), but possibly also as many as six sites – three atoms and optional dummy particles corresponding to a 'center' of electron density, or to the lone electron pairs on the oxygen atom.



The most frequently used atomic model of water is the TIP3P (very similar is the SPC). A TIP3P molecule consists of three atoms connected by three rigid bonds. A charge is placed on every atom (-0.834 on the O and $+0.417$ on the Hs), while *only* the oxygen atom possesses non-zero Lennard-Jones parameters.[7]

If the negative charge is placed on a dummy atom M rather than on the oxygen, then the electric field around the molecule is described better. This idea is implemented e.g. in the TIP4P model.

A further improvement may be achieved if two dummy particles L bearing negative charge are placed near the oxygen atom, to mimic the lone electron pairs. Consequently, such a five-site model (like TIP5P) describes the directionality of hydrogen bonding and derived effects (radial distribution function, temperature of highest density) better than less sophisticated models.

[7] This makes it possible to additionally optimize the algorithm for the calculation of energy and forces.