CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

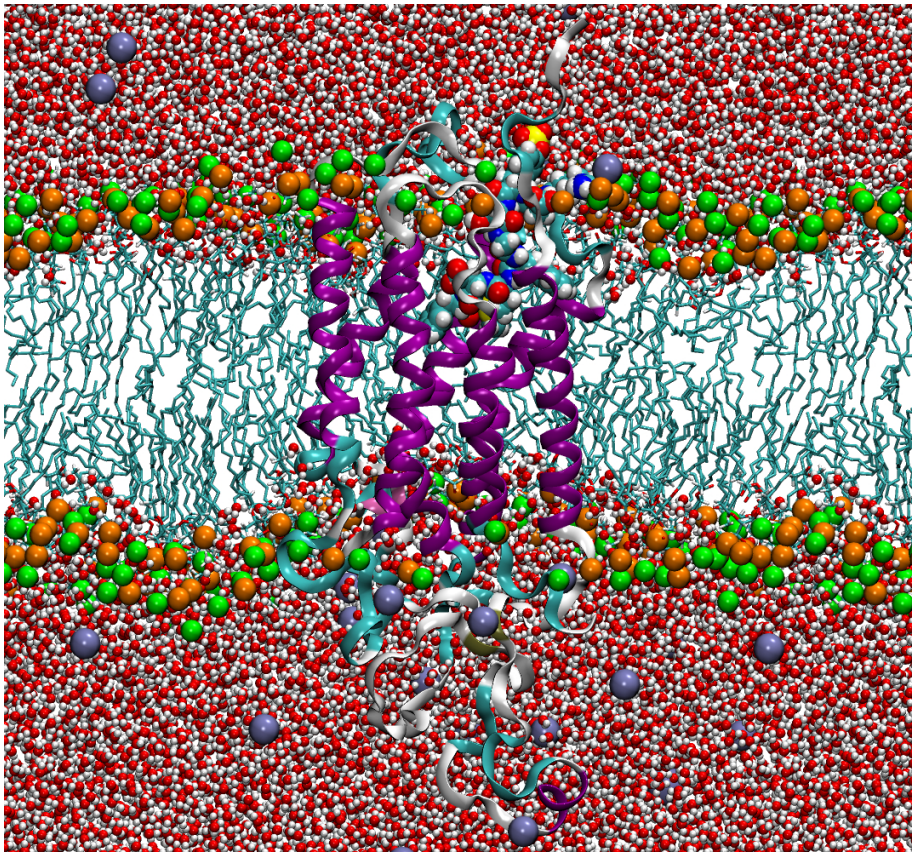
université HENRI POINCARÉ NANCY 1

Numerical methods for molecular dynamics simulations of biological systems

Christophe Chipot

Equipe de dynamique des assemblages membranaires, Unité mixte de recherche CNRS/UHP 7565, Université Henri Poincaré, B.P. 239, 54506 Vandœuvre–lès–Nancy, France

Christophe.Chipot@edam.uhp–nancy.fr



(Version: November 30, 2010)

Table of contents

1. Introduction

Knowledge at the atomic level of the structural and dynamic aspects of organized systems is of paramount importance to improve our understanding of the function of these complex molecular assemblies. In a number of instances, accessing the microscopic detail by means of conventional experimental techniques is not possible. Yet, the massive increase of computational resources ignited some ten years ago, associated with the development of efficient algorithms have enabled the study of supramolecular assemblies of increasing complexity using theoretical tools.

The main thrust of this course is to examine the facet of Theoretical Chemistry constituted by molecular mechanical statistical simulations. The goal of the latter is to access the atomic detail of condensed phases through *in silico* computational experiments. Several methods are currently available, among which molecular dynamics (MD), stochastic dynamics (SD) and its special cases — for example, Brownian dynamics or Langevin dynamics — or Monte Carlo (MC) simulations. These different theoretical approaches may be viewed as a bridge connecting macroscopic–level experimental observations to the microscopic world. In what follows, we shall focus on molecular dynamics [1].

1.1. Connecting the microscopic to the meso– and the macroscopic

Is it legitimate to use molecular simulations to model condensed phases? Rigorously, the complete study of a chemical system as complex as a molecular liquid would require solving the time–dependent Schrödinger equation for a large ensemble {electrons + nuclei}. Such an approach remains, however, totally illusory, in spite of the recent progresses made in the field of linear–scaling calculations, thereby constraining us to limit ourselves to a classical description of the system. Even in this framework, for obvious computational reasons, molecular simulations are generally restrained to a number of particles comprised between a few hundreds to several thousands.

In order to correlate the properties of the microscopic system to those of the macroscopic phase, it is pivotal to eliminate edge effects. In practice, use is made of periodic boundary conditions (PBC), which consists in replicating the finite ensemble of particles confined in a box, often orthorhombic, following the three directions of space (see Figure 1). It has been observed that a faithful, accurate reproduction of thermodynamic quantities from samples of reduced size warrants *a posteriori* the use of such an approach. The pseudo–infinite character of the system thus generated implies the necessity of a number of approximations for the treatment of intermolecular interactions [2]. In particular, the so–called "minimum image" approximation supposes that each particle i of the central cell interacts with the closest image of all other particles j.

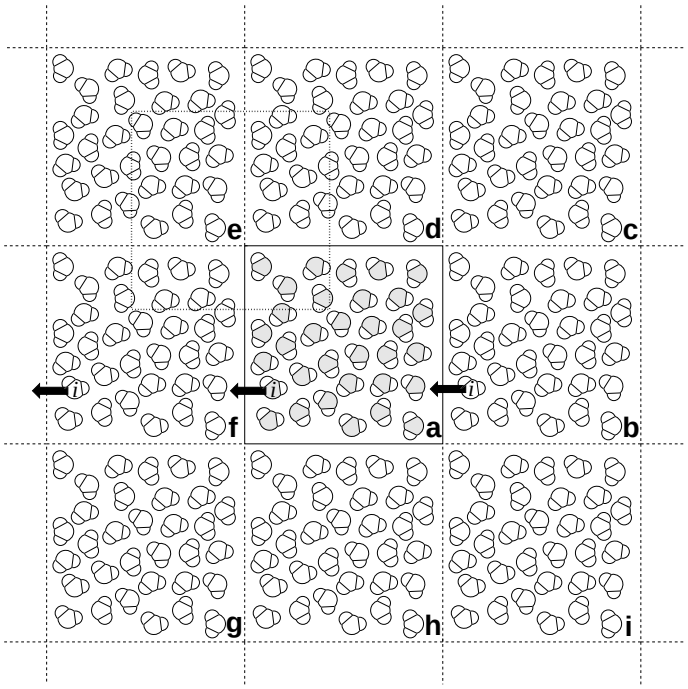


Figure 1: Two–dimensional view of a simulation cell replicated in the three directions of space. Employing periodic boundary conditions (PBC), when molecule i leaves the central box **a**, its images in the neighboring ghost boxes move in a similar fashion. The cell delineated by a dotted line, overlapping with cells **a**, **d**, **e** et **f**, symbolizes the so–called "minimum image" convention.

Furthermore, the introduction of a sphere of truncation, or cut–off sphere, can be employed to ignore interactions beyond an arbitrary distance, less or equal than half of the smallest dimension of the simulation cell (see Figure 2). Clearly, the validity of these approximations is conditioned by the range of the intermolecular interactions considered. Whereas short–range dispersion and repulsion interactions do not require any special care, the same cannot be said in the case of electrostatic interactions. The size of the system evolving roughly in r^3, it is generally accepted that $1/r^n$ interactions — where $n < 3$ — will not be handled correctly using a spherical truncation. To circumvent this difficulty, it is recommended to turn to a more adapted method, like those based on lattice sums, *e.g.* Ewald–Kornfeld [3] or Ladd [4], which consist in evaluating the interactions of a particle with all others located in the central box, as well as in all image cells. Adopting such an approach, however, increases considerably the global computational effort, but is rigorously indispensable for a correct description of long–range interactions.

1.2. How legitimate are molecular dynamics simulations?

The principle of MD, particularly simple, consists in generating trajectories for a finite ensemble of particles by integrating numerically the classical equations of motion. This approach, *a priori*

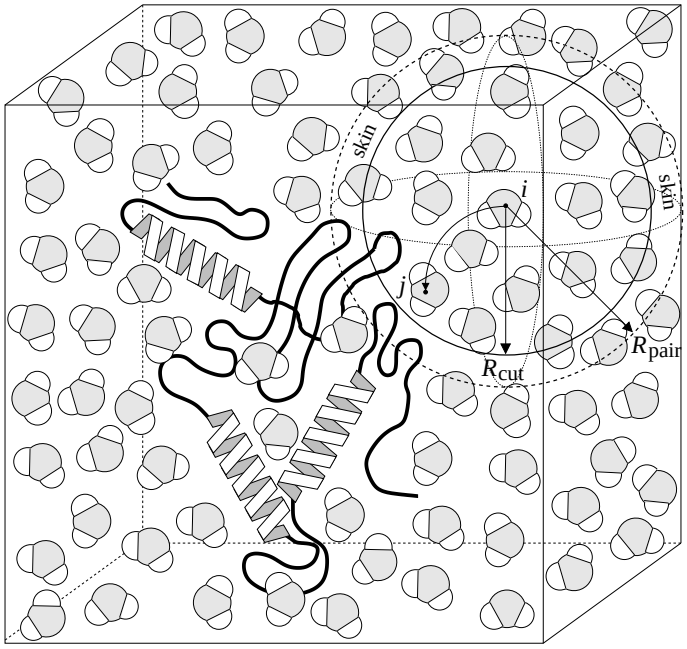


Figure 2: Use of a truncation sphere of radius $R_{\rm cut}$ to limit the computation of interactions of particle i with its neighbors in the "minimum image" convention. A sphere of radius $R_{\rm pair}$, greater than $R_{\rm cut}$, is employed to construct a neighbor list for i. This list of all pairs $\{i, j\}$ is updated periodically.

arguable, finds its justification in two remarkable facts: (i) on account of the Born–Oppenheimer approximation, the motion of the nuclei and that of the electrons can be dissociated, and, (ii) considering that, in most cases, the de Broglie wavelength is much shorter than typical intermolecular distances, quantum effects can be safely neglected. Trajectories determined by this approach are utilized to evaluate static and dynamic properties in the form of time averages, which, for *ergodic* systems, coincide with statistical averages:

$$\lim_{t\to\infty} \overline{\mathscr{A}}_t = \langle \mathscr{A} \rangle \tag{1}$$

Here, $\mathscr{A}$ denotes any observable property. $\overline{\mathscr{A}}_t$ represents its time average, and $\langle \mathscr{A} \rangle$, its statistical average. In practice, it is observed that the ergodicity hypothesis is verified, at least in the case of simple liquids.

2. The molecular dynamics equations

In classical MD [1, 2, 5], the trajectory for the various components of the system is generated by integrating the Newton equations of motion, which, for each particle i, write:

$$\begin{cases} m_i \dfrac{\mathrm{d}^2\mathbf{x}_i(t)}{\mathrm{d}t^2} = \mathbf{f}_i(t) \\ \\ \mathbf{f}_i(t) = -\dfrac{\partial \mathscr{V}(\mathbf{x})}{\partial \mathbf{x}_i(t)} \end{cases} \tag{2}$$

$\mathscr{V}(\mathbf{x})$ is the potential energy function of the N–particle system, which only depends upon the Cartesian coordinates $\{\mathbf{x}_i\}$. Equations (2) are integrated numerically using an infinitesimal time–step, δt, to guarantee the conservation of the total energy of the system — *viz.* typically 1–2 fs (see Figure 3).

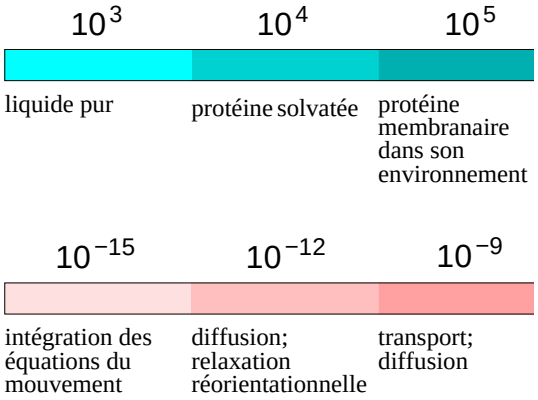


Figure 3: Size– (**a**) an time scales (**b**) accessible to MD simulations. The largest molecular system modeled to this date is the motion of transfer RNA in the ribosome, *i.e.* 2.64×10^6 atoms [6]. The longest simulation to this date remains the folding of the villin headpiece, a 47–residue fragment, over a period of 10^{-6} s [7].

2.1. Propagating the motion

Hoping to generate *exact* trajectories over long times is, however, illusory, considering that the Newton equations of motion are solved numerically, with a finite time–step. The exactness of the solution of equations (2) is, nevertheless, not as crucial as it would seem. What really matters in reality is the correct statistical behavior of the trajectory to ensure that the thermodynamic and dynamic properties of the system be reproduced with a sufficient accuracy. This pivotal condition is fulfilled only if the integrator employed to propagate the motion possesses the property of *symplecticity* [8–10]. A so–called *symplectic* propagator conserves the invariant metric of the phase space, $\mathbf{\Gamma}$. As a result, the error associated with this propagator is bound:

$$\lim_{n_{\mathrm{step}}\to\infty} \left(\frac{1}{n_{\mathrm{step}}}\right) \sum_{k=1}^{n_{\mathrm{step}}} \left| \frac{\mathscr{E}(k\delta t) - \mathscr{E}(0)}{\mathscr{E}(0)} \right| \leq \varepsilon_{\mathrm{MD}} \tag{3}$$

Here, $n_{\rm step}$ denotes the number of steps of the simulation, $\mathscr{E}(0) \equiv \mathscr{H}(\mathbf{x}, \mathbf{p}_x; 0)$, the initial total energy of the equilibrated system, and $\varepsilon_{\rm MD}$, the upper bound for energy conservation — *viz.* 10^{-4} constitutes an acceptable value. Assuming that the time–step is limited, integration of the equations of motion does not lead to an erratic growth of the error associated with the conservation of the total energy, which may affect significantly the statistical behavior of MD over long times. Interestingly enough, for a Hamiltonian system, the property of symplecticity implies that the Jacobian:

$$\mathbf{J}(\mathbf{\Gamma}_{\delta t}, \mathbf{\Gamma}_0) = \frac{\partial(\mathbf{\Gamma}^1_{\delta t}, \ldots, \mathbf{\Gamma}^N_{\delta t})}{\partial(\mathbf{\Gamma}^1_0, \ldots, \mathbf{\Gamma}^N_0)} \tag{4}$$

is unitary. $\mathbf{\Gamma}_0$ represents the initial vector of the N–dimension phase space, which contains all the position, $\mathbf{x}$, and momentum, $\mathbf{p}_x$, variables that describe the system.

As has been underlined previously, the long–range nature of charge-dipole interactions, *i.e.* $1/r^2$, and, *a fortiori*, charge–charge interactions, *i.e.* $1/r$, imposes the use of well–adapted algorithms for handling such contributions, which may increase appreciably the computational cost of the simulation. Rewriting equations (2) in a more formal fashion:

$$\mathbf{\Gamma}_t = \mathrm{e}^{i\mathscr{L}t}\, \mathbf{\Gamma}_0 \tag{5}$$

where $\mathscr{L}$ is the Liouville operator that generates distribution $\varrho(\mathbf{\Gamma}, t)$ for a given thermodynamic ensemble, following:

$$\frac{\partial \varrho(\mathbf{\Gamma}, t)}{\partial t} = -i\mathscr{L}\, \varrho(\mathbf{\Gamma}, t) \tag{6}$$

and applying the Trotter factorization:

$$\mathrm{e}^{i\mathscr{L}\Delta t} = \mathrm{e}^{i\mathscr{L}_1 \frac{\Delta t}{2}}\, \mathrm{e}^{i\mathscr{L}_2 \Delta t}\, \mathrm{e}^{i\mathscr{L}_1 \frac{\Delta t}{2}} + \mathscr{O}(\Delta t^3) \tag{7}$$

in which $i\mathscr{L} = i\mathscr{L}_1 + i\mathscr{L}_2$, deconvolution of the short– and long–range contributions becomes straightforward. Depending upon the nature of the interaction, different time–steps can, thus, be utilized to propagate the motion. Breaking down, for instance, the total Hamiltonian, $\mathscr{H}(\mathbf{x}, \mathbf{p}_x)$, that governs the system into kinetics, $\mathscr{T}(\mathbf{p}_x)$, valence, $\mathscr{V}_{\rm valence}(\mathbf{x})$, short–range, $\mathscr{V}_{\rm short}(\mathbf{x})$, and long–range, $\mathscr{V}_{\rm long}(\mathbf{x})$, contributions, it ensues that:

$$\mathrm{e}^{i\mathscr{H}(\mathbf{x},\mathbf{p}_x)\Delta t} = \mathrm{e}^{i\mathscr{V}_{\rm long}(\mathbf{x})\frac{\Delta t}{2}} \left\{ \mathrm{e}^{i\mathscr{V}_{\rm short}(\mathbf{x})\frac{\Delta t}{2n}} \left[\mathrm{e}^{i\mathscr{V}_{\rm valence}(\mathbf{x})\frac{\Delta t}{2pn}}\, \mathrm{e}^{i\mathscr{T}(\mathbf{p}_x)\frac{\Delta t}{pn}} \right. \right. \tag{8}$$

$$\times \quad \mathrm{e}^{i\mathscr{V}_{\mathrm{valence}}(\mathbf{x})\frac{\Delta t}{2pn}}\Big]^p \mathrm{e}^{i\mathscr{V}_{\mathrm{short}}(\mathbf{x})\frac{\Delta t}{2n}}\Big\}^n \mathrm{e}^{i\mathscr{V}_{\mathrm{long}}(\mathbf{x})\frac{\Delta t}{2}}$$

This partitioning of the various contributions of $\mathscr{H}(\mathbf{x},\mathbf{p}_x)$ constitutes the central idea of the so–called multiple time–step approaches, like the reversible reference system propagator algorithm (r–RESPA) [11]. It clearly highlights the use of distinct time–steps for updating these contributions, thereby reducing appreciably the computational effort of the statistical simulation.

2.2. The molecular dynamics propagators

Several approaches for integrating numerically the Newton equations of motion (2) are currently available. Among them, three will be detailed here. Unquestionably the simplest, the Verlet algorithm relies upon the knowledge of the triplet $\{\mathbf{x}_i(t),\mathbf{x}_i(t-\delta t),\mathbf{a}_i(t)\}$, where $\mathbf{a}_i(t) = \ddot{\mathbf{x}}_i(t) = \mathrm{d}^2\mathbf{x}_i(t)/\mathrm{d}t^2 = \mathbf{f}_i(t)/m_i$ denotes the acceleration of particle i [12]. Modifying the positions of the particles is achieved through a Taylor expansion of the position at $t-\delta t$ and at $t+\delta t$, leading to:

$$\mathbf{x}_i(t+\delta t) = 2\mathbf{x}_i(t) - \mathbf{x}_i(t-\delta t) + \mathbf{a}_i(t)\,\delta t^2 \tag{9}$$

which implies possible errors in $\mathscr{O}(\delta t^4)$. It is worth noting that the velocities, $\mathbf{v}_i(t) = \dot{\mathbf{x}}_i(t) = \mathrm{d}\mathbf{x}_i(t)/\mathrm{d}t$, do not appear explicitly in this scheme. They cancel out in the Taylor expansion of $\mathbf{x}_i(t+\delta t)$ and $\mathbf{x}_i(t-\delta t)$. Though unnecessary for describing the trajectory, their evaluation is an obligatory step for computing the kinetic energy, $\mathscr{T}(\mathbf{p})$, which depends upon the sole momentum variables, $\mathbf{p}$, and, consequently, the total energy of the system, $\mathscr{E} \equiv \mathscr{H}(\mathbf{x},\mathbf{p}_x)$, according to:

$$\mathbf{v}_i(t) = \frac{\mathbf{x}_i(t+\delta t) - \mathbf{x}_i(t-\delta t)}{2\,\delta t} \tag{10}$$

At each time–step, the associated error is in $\mathscr{O}(\delta t^2)$.

Exercise: From the Taylor expansion of position $\mathbf{x}$, en $t-\delta t$ et en $t+\delta t$, recover expression (9). Deduce from the latter that the associated error grows as $\mathscr{O}(\delta t^4)$.

The so–called *leap-frog* algorithm, derived from the preceding one, makes use of the $\{\mathbf{x}_i(t),\mathbf{v}_i(t-\delta t/2),\mathbf{a}_i(t)\}$ triplet. The origin of its name appears clearly in the writing of the algorithm:

$$\left\{\begin{array}{rcl} \mathbf{x}_i(t+\delta t) & = & \mathbf{x}_i(t) + \mathbf{v}_i(t+\dfrac{\delta t}{2})\,\delta t \\ \\ \mathbf{v}_i(t+\dfrac{\delta t}{2}) & = & \mathbf{v}_i(t-\dfrac{\delta t}{2}) + \mathbf{a}_i(t)\,\delta t \end{array}\right. \tag{11}$$

In practice, the first step is the computation of $\mathbf{v}_i(t + \delta t/2)$, from which $\mathbf{v}_i(t)$ is deduced, which is a requisite for evaluating the kinetic term, $\mathscr{T}(\mathbf{p})$, following:

$$\mathbf{v}_i(t) = \frac{\mathbf{v}_i(t + \frac{\delta t}{2}) + \mathbf{v}_i(t - \frac{\delta t}{2})}{2} \tag{12}$$

Last, the velocity form of the Verlet algorithm corrects the main deficiency of the standard Verlet or *leap-frog* scheme, namely the rigorous definition of the velocities — the associated error of which varies in $\mathscr{O}(\delta t^2)$. The explicit incorporation of the velocities in the Verlet algorithm may be expressed as:

$$\left\{ \begin{array}{rcl} \mathbf{x}_i(t + \delta t) & = & \mathbf{x}_i(t) + \mathbf{v}_i(t)\ \delta t + \dfrac{1}{2}\ \mathbf{a}_i(t)\ \delta t^2 \\ \\ \mathbf{v}_i(t + \delta t) & = & \mathbf{v}_i(t) + \dfrac{\mathbf{a}_i(t) + \mathbf{a}_i(t + \delta t)}{2}\ \delta t \end{array} \right. \tag{13}$$

This scheme involves the two following steps:

$$\mathbf{v}_i(t + \frac{\delta t}{2}) = \mathbf{v}_i(t) + \frac{1}{2}\ \mathbf{a}_i(t)\ \delta t \tag{14}$$

from which the thermodynamic forces, $\mathbf{f}_i$, and accelerations, $\mathbf{a}_i$, at time $t + \delta t$ can be evaluated. It ensues that:

$$\mathbf{v}_i(t + \delta t) = \mathbf{v}_i(t + \frac{\delta t}{2}) + \frac{1}{2}\ \mathbf{a}_i(t + \delta t)\ \delta t \tag{15}$$

The kinetic energy may then be deduced at time $t + \delta t$, while the potential energy, $\mathscr{V}(\mathbf{x})$, is computed in the force loops.

2.3. Constrained equations of motion

Under certain circumstances, it may be desirable to freeze a subset of degrees of freedom in the course of the MD simulation through the introduction of holonomic constraints. The latter allow hard degrees of freedom that correspond to high–frequency vibrations — *viz.* the vibration of covalent bonds involving hydrogen atoms — to be removed. By keeping through time these bond lengths at a nominal value, it is possible to increase the time step, δt, for integrating the equations of motion, without sacrificing the prerequisite of energy conservation.

Freezing certain degrees of freedom in the framework of classical MD is equivalent to solving constrained equations of motion. In the case of a fixed chemical bond length, the constraint writes:

$$\chi_{ij}(t) = |\mathbf{r}_j(t) - \mathbf{r}_i(t)|^2 - d_{ij}^2 \tag{16}$$

where d_{ij} stands for the equilibrium length of the chemical bond. It follows that, in addition to the force, $\mathbf{f}_i$, due to intra– and intermolecular interactions, a constraint force, $\mathbf{g}_i$, now appears in the equations of motion:

$$m_i \, \frac{\mathrm{d}^2\mathbf{r}_i(t)}{\mathrm{d}t^2} = \mathbf{f}_i + \mathbf{g}_i \tag{17}$$

This constraint force is defined by:

$$\mathbf{g}_i = -\sum_j \, \lambda_{ij}(t) \, \nabla_i \chi_{ij}(t) = -2 \sum_j \, \lambda_{ij}(t) \, \mathbf{r}_{ij}(t) \tag{18}$$

where $\lambda_{ij}(t)$ is the Lagrange multiplier associated to the constraint enforced along the chemical bond connecting atoms i and j. Combined with the Verlet integrator (9), the equations of motion now write:

$$\mathbf{r}_i(t+\delta t) = \mathbf{r}_i(t) + \delta t \, \mathbf{v}_i(t) + \frac{\delta t^2}{2m_i} \, [\mathbf{f}_i(t) + \mathbf{g}_i(t)] \tag{19}$$

These constrained equations of motion are solved in most cases following a Gauss–Seidel–like, iterative scheme — *i.e.* solving one by one the equations of a linear system — until each holonomic constraint is satisfied.

Exercise: Establish the constrained equations of motion for a triatomic molecule, in which the bond lengths d_{21} and d_{23} are frozen and the valence angle, $\theta(1,2,3)$, may vary through the intramolecular potential.

3. The potential energy function

The potential energy function constitutes the corner stone of all molecular mechanics calculations, as it should reproduce the intra– and intermolecular interactions of the system as faithfully as possible. In principle, on account of the many–body character of the problem, this functional should write as an N–term sum:

$$\mathscr{V}(\mathbf{x}) = \sum_i v_1(\mathbf{x}_i) + \sum_i \sum_{j>i} v_2(\mathbf{x}_i, \mathbf{x}_j) + \sum_i \sum_{j>i} \sum_{k>j>i} v_3(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) + \dots \tag{20}$$

where $v_1(\mathbf{x}_i)$, $v_2(\mathbf{x}_i, \mathbf{x}_j)$, ...represent the intramolecular potential, the pair interaction potential, ... $\mathscr{V}(\mathbf{x})$ is, therefore, characteristic of an N–body problem, even though it might be argued that $v_2(\mathbf{r}_i, \mathbf{x}_j)$ is likely to constitute the prevailing term of the intermolecular contribution [2]. This point of view is at the origin of the so–called pairwise approximation, in which higher order effects are partially included in an effective potential:

$$\mathscr{V}(\mathbf{x}) \simeq \sum_i v_1(\mathbf{x}_i) + \sum_i \sum_{j>i} v_2^{\text{effective}}(x_{ij}) \tag{21}$$

This approximation is used in most commercial force fields, in particular those aimed at the study of macromolecular systems, for which the computational effort is intimately related to the complexity of $\mathscr{V}(\mathbf{x})$. Amongst the plethora of available potential energy functions, one of the most minimalist descriptions of the system is provided by the AMBER force field [13, 14]:

$$\begin{aligned}
\mathscr{V}(\mathbf{x}) &= \sum_{\text{bonds}} k_r\,(r - r_0)^2 + \sum_{\text{valence angles}} k_\theta\,(\theta - \theta_0)^2 \\
&+ \sum_{\text{torsions}} \sum_n \frac{V_n}{2}\,[1 + \cos(n\phi - \gamma)] \\
&+ \frac{1}{k_{\text{vdW}}^{\text{1-4}}} \sum_{\substack{i<j \\ \{i,j\}\in\text{1-4}}} \varepsilon_{ij} \left[\left(\frac{R_{ij}^*}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}^*}{r_{ij}}\right)^{6}\right] + \frac{1}{k_{\text{Coulomb}}^{\text{1-4}}} \sum_{\substack{i<j \\ \{i,j\}\in\text{1-4}}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1 r_{ij}} \\
&+ \sum_{\substack{i<j \\ \{i,j\}>\text{1-4}}} \varepsilon_{ij} \left[\left(\frac{R_{ij}^*}{r_{ij}}\right)^{12} - 2\left(\frac{R_{ij}^*}{r_{ij}}\right)^{6}\right] + \sum_{\substack{i<j \\ \{i,j\}>\text{1-4}}} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1 r_{ij}}
\end{aligned} \tag{22}$$

in which k_r and r_0 denote, respectively, the force constant of the chemical bond and its equilibrium length, k_θ and θ_0, the force constant of the valence angle and its equilibrium value, and $V_n/2$, n and γ, the torsional barrier, its periodicity and phase. ϵ_0 and ϵ_1 are, respectively, the vacuum and the relative dielectric permittivities. q_i is the partial charge borne by atom i. R_{ij}^* and ε_{ij} correspond to the van der Waals parameters for the pair of atoms $\{ij\}$, obtained from the Lorentz–Berthelot combination rules:

$$\begin{cases} \varepsilon_{ij} &= \sqrt{\varepsilon_i\,\varepsilon_j} \\ R_{ij}^* &= R_i^* + R_j^* \end{cases} \tag{23}$$

Considering that in the course of their parametrization, generally using sophisticated quantum–

chemical calculations, the torsional terms readily contain both an electrostatic and a van der Waals component, most commercial force fields distinguish between those interactions of atoms separated by exactly three chemical bonds (*i.e.* the so–called "1–4" terms), and all others, provided that $\{i, j\}$ are not separated by one or two chemical bonds (see Figure 4). van der Waals and Coulomb "1–4" contributions are usually weighted down by factors $1/k_{\mathrm{vdW}}^{1\text{-}4}$ and $1/k_{\mathrm{Coulomb}}^{1\text{-}4}$ that appear in expression (22).

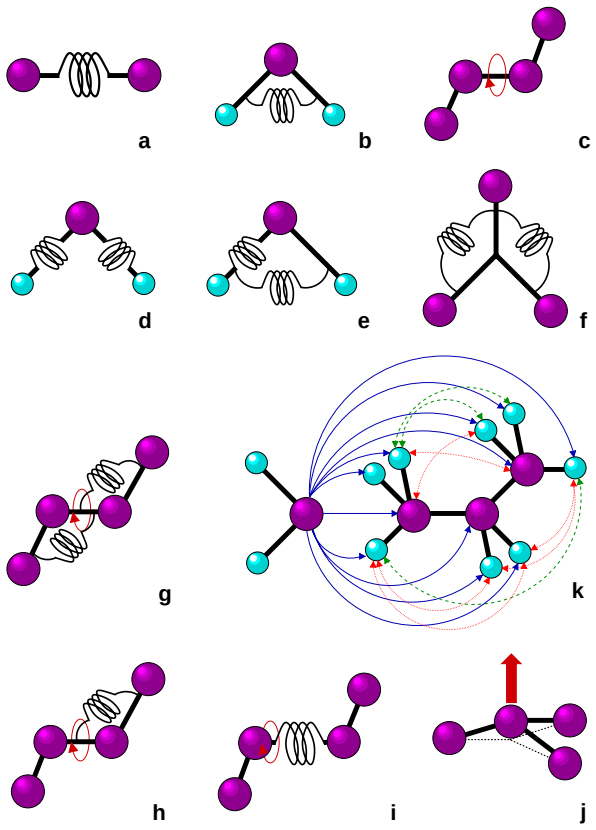


Figure 4: Illustration of the various terms included in an empirical potential energy function. Contributions **a-j** represent the valence force field, among which **d-j** are the so-called "crossed" terms. **j** is the out–of–plane term, guaranteeing that the central atom remains in the plane formed by its three neighbors to which it is bonded chemically. **k** characterizes the Coulomb and van der Waals interactions of atoms that are not bonded chemically: Intermolecular interactions (solid line), "1–4" intermolecular interactions (dotted line), and intramolecular interactions $>$ 1–4 (dashed line).

An important concept of macromolecular force fields is the underlying assumption of *transferability*, which, in a nutshell, presupposes that a vast variety of molecules can be described by a limited amount of parameters. To reach this goal, a restrained list of commonly found, yet chemically different atoms is sought. The bonded and non–bonded parameters involving these atoms are then determined, usually by means of quantum mechanical calculations on prototypical systems. The transferability hypothesis imposes that a particular type of atom can be described by the same set of parameters in chemically

distinct molecules and, hence, distinct chemical environments. For instance, in the AMBER force field, a CT sp_3 carbon atom can be found in n–butane, ethanol, or in the side chain of valine.

By and large, any force field consists of a very subtle equilibrium between its various contributions. The torsional component, which is crucial for describing conformational preferences, represents a small facet of this equilibrium. The choice of Lennard–Jones parameters, R_{ij}^* and ε_{ij}, and of the partial charges, q_i, is equally critical, because it modulates the accuracy of the computed thermodynamics and dynamics quantities. A potential energy function is a complex construction, the building blocks of which have been calibrated for the global reproduction of physical and chemical key–properties, yet without the necessity that these elements be physically or chemically meaningful. As a result, the modeler should avoid interchanging parameters between different force fields, for the latter correspond to distinct philosophies.

One of the fashionable approaches for developing models of point charges consists in fitting the latter to the electrostatic potential, the true fingerprint of the molecule. Limited to a monopole expansion of the electrostatic potential, the optimal set of N_{atoms} net atomic charges, $\{q_k\}$, is derived by minimizing the functional:

$$f(\{q_k\}) = \sum_{i=1}^{N_{\text{points}}} \left[V^{\text{reference}}(\mathbf{r}_i) - \sum_{j=1}^{N_{\text{atoms}}} \frac{q_j}{r_{ij}} \right]^2 \tag{24}$$

where $V^{\text{référence}}(\mathbf{r}_i)$ is the electrostatic potential evaluated at point $\mathbf{r}_i$ of a grid of N_{points} surrounding the molecule. $V^{\text{reference}}(\mathbf{r}_i)$ is obtained from quantum–chemical calculations, *i.e.* the expectation value $\langle \Psi \left| 1/\mathbf{r}_i \right| \Psi \rangle$, usually at a sophisticated level of theory.

Determination of Lennard–Jones parameters often turns out to be a far more daunting task. A possible route to reach this goal consists in fitting dispersion and repulsion contributions to the interaction energies obtained from a large number of high–level quantum chemical calculations, carried out with distinct configurations. This approach is generally applicable to chemical systems of reduced sizes, *e.g.* the formamide–water heterodimer, for which the atom–atom van der Waals interaction potential is sought. A more heuristic approach, alternative to numerous sophisticated quantum chemical computations, relies upon statistical simulations of condensed phases. Starting from a given set of Lennard–Jones parameters, $\{R_{ij}^*, \varepsilon_{ij}\}$, can we reproduce accurately fundamental thermodynamic quantities of a molecular liquid, like its density, ρ, its enthalpy of vaporization, ΔH_{vap}, and possibly dynamical properties like its self–diffusion coefficient, D.

3.1. Beyond the minimalist description

In a number of cases, the minimalist description imposed by the functional (22) may turn out to be inadequate. As is the case of most commercial force fields, this functional is multi–purpose, even though it was designed originally to study biopolymers, and specifically proteins and nucleic acids. The finer, deeper investigation of small organic molecules, often necessary in the emerging field of *de novo* drug design, requires a level of approximation that goes beyond the simplistic underlying hypotheses of expression (22). Among the latter, replacement of the harmonic term describing the stretch of chemical bonds (see Figure 4 **a**) by a dissociative Morse potential constitutes an interesting example:

$$\mathscr{V}(r) = D_0 \left[e^{-\alpha(r-r_0)} - 1\right]^2 \tag{25}$$

where D_0 stands precisely for the dissociation energy and r_0, the equilibrium bond length. The harmonic description may turn out to be unadapted as soon as anharmonicity effects can no longer be neglected. For both the stretch of chemical bonds and the deformation of valence angles, cubic and quartic corrections may be necessary:

$$\mathscr{V}(r) = k_r \left(r - r_0\right)^2 \left[1 - k'_r \left(r - r_0\right) + k''_r \left(r - r_0\right)^2\right] \tag{26}$$

Opening of a valence angle results in the shrinking of the associated chemical bonds — an effect clearly absent in equation (22). To correct this deficiency, the potential energy function can be enriched with the so–called *cross–terms*, like:

$$\mathscr{V}(r\text{-}\theta) = k_{r\theta} \left(r - r_0\right) \left(\theta - \theta_0\right) \tag{27}$$

In addition to the stretch–bend term, depicted in the above expression, it is possible to introduce other coupling terms in $\mathscr{V}(\mathbf{x})$, as shown in Figure 4, **e-i**. Aside from the heavier evaluation of the potential energy function at each time–step, δt, when integrating equations (2), inclusion of the cross–terms implies an additional effort for parameterizing these contributions.

The accurate description of the dihedral angles also constitutes a particularly crucial aspect in the development of potential energy functions. The behavior of the torsional potential, $\mathscr{V}(\phi)$, is often exceedingly complex to be mimicked appropriately by a single term of the Fourier series in equation (22) The case of phospholipids represents an excellent illustration of this difficulty. The key to a faithful reproduction of the order parameters, S_{CD}, of the aliphatic chains resides, indeed, in the

correct description of the dihedral angles in the course of the statistical simulation. Only a potential that includes several terms is able to account for the subtle *trans–gauche* equilibrium along the lipid chains. The Ryckaert and Bellemans potential [15],

$$\mathscr{V}(\phi) = \sum_{i=1}^{6} a_i \, \cos^i \phi \tag{28}$$

in which the coefficients a_i have been optimized on the basis of the internal rotation of n–butane, is often employed for simulating phospholipid bilayers.

3.2. Circumventing the pitfalls of the pairwise additive approximation

Treatment of the electrostatics is without any doubt the most problematic in many respects. First, owing to the fact that the monopole approximation in equation (24) is not necessarily sufficient for a faithful description of any molecule. It may, thus, be desirable to include permanent dipoles to the simple model of net atomic charges. More crucial are the possible induction effects, obviously absent in the pairwise additive approximation (21). To circumvent this difficulty, the number of solutions available remains limited. The less expensive approach, still widely utilized in numerical simulations of macromolecular systems, consists in inflating artificially the point charges, so that the latter reproduce a permanent dipole moment characteristic of the condensed phase rather than the low–pressure gaseous phase. In this representation, polarization effects are taken into account in an *average* sense. Explicit introduction of polarizability contributions in the force field [16] constitutes the more rigorous, and at the same time the more costly, solution. In this case, the total electrostatic term writes:

$$\mathscr{V}_{\text{élec}}(\mathbf{x}) = \frac{1}{2} \sum_i q_i \, V_i \tag{29}$$

Here, the electrostatic potential, V_i, arises from two distinct sources — on the one hand, the ensemble of point charges of the system; on the other hand, the induced multipole moments at site i. Limiting ourselves to the sole induced dipole moment, $\boldsymbol{\mu}_i$, which is linearly related to the electric field, $\mathbf{E}_i$, created at this point by all other polarizable sites $j \neq i$, that is, $\boldsymbol{\mu}_i = \alpha_i \, \mathbf{E}_i$, the potential may be expressed as:

$$V_i = \sum_{j \neq i} \left[\frac{q_j}{4\pi\epsilon_0 r_{ij}} + \frac{\mathbf{r}_{ij} \cdot \boldsymbol{\mu}_j}{4\pi\epsilon_0 r_{ij}^3} \right] \tag{30}$$

Even when using as a starting point at time $t+\delta t$, the induced moments, $\boldsymbol{\mu}_i$, obtained at time t, convergence of the latter in a MD simulation of a polarizable liquid increases the computational effort by a factor of 2 at the most, with respect to a simulation featuring an additive pairwise approximation [17].

3.3. Coarse–graining the problem

At the antipodes of all–atom simulations that explicitly take induction effects into account, far more rudimentary approaches have been devised to push back the usual limitations of classical MD, in terms of both size– and time–scales — see Figure 3. Coarse–grain descriptions provide a convincing answer to these stringent requirements by reducing significantly the number of particles of the system and allow longer time steps to be employed — *e.g.* typically δt = 40 fs, by eliminating the hardest degrees of freedom.

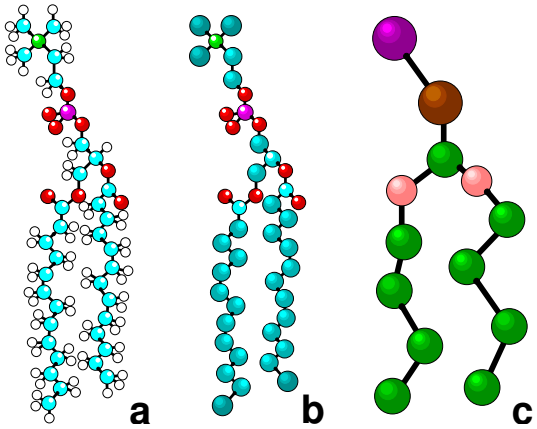


Figure 5: All–atom (**a**) *vs.* united–atom (**b**) *vs.* coarse–grain (**c**) description of a lipid unit. Whereas in a united–atom model, methylene, –CH_2–, and methyl, –CH_3, groups are treated as van der Waals spheres, in a coarse–grain description, one single sphere encompasses several methylene and methyl groups.

In the particular case of Figure 5, the ratio between the number of atoms in an all–atom and a united–atom or a coarse–grain description is 2 or 9, respectively. Difference in the interaction potentials utilized in coarse–grain and all– or united–atom models is often appreciable. The electrostatic term is generally Coulombic, but may involve the interaction of point dipoles. van der Waals interactions are described by modified Lennard–Jones potentials — *e.g.* 6–9, or Gay–Berne potentials [18]. It is worth noting that in spite of their very rudimentary nature, coarse–grain models can reproduce a number of physical properties of the system at a semi–quantitative level. Such is the case, for instance, of atomic density profiles in lipid bilayers [19, 20].

4. Exploring thermodynamic ensembles

In the course of the simultaneous integration of Newton's equation of motion, the total energy of the system is conserved. If the volume is preserved, the generated ensemble is microcanonical, *i.e.* $(N, V, \mathscr{E})$. This type of simulation is often referred to Newtonian MD. This simplistic situation is, however, not always satisfactory and adapted to the system examined, and it may be desirable to perform MD simulations, where the temperature and the pressure are considered as independent, rather than derived quantitities — see Figure 6.

Ensemble:	Free energy:	Conserved Hamiltonian:
$(N, V, \mathscr{E})$		$\mathscr{H}(\mathbf{x}, \mathbf{p}_x) = U$
(N, V, T)	$A = U - TS$	$\mathscr{H}(\mathbf{x}, \mathbf{p}_x) \simeq \sum_i \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + \mathscr{V}(\mathbf{x})$
$(N, P_\perp, \mathscr{A}, T)$	$A = U - TS + P_\perp V$	$\mathscr{H}(\mathbf{x}, \mathbf{p}_x) \simeq \sum_i \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + \mathscr{V}(\mathbf{x}) + P_\perp V$
$(N, P_\parallel, \mathscr{A}, T)$	$A = U - TS + P_\parallel V$	$\mathscr{H}(\mathbf{x}, \mathbf{p}_x) \simeq \sum_i \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + \mathscr{V}(\mathbf{x}) + P_\parallel V$
(N, V, γ, T)	$A = U - TS - \gamma\mathscr{A}$	$\mathscr{H}(\mathbf{x}, \mathbf{p}_x) \simeq \sum_i \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + \mathscr{V}(\mathbf{x}) - \gamma\mathscr{A}$
$(N, P_\perp, \gamma, T)$	$A = U - TS + P_\perp V - \gamma\mathscr{A}$	$\mathscr{H}(\mathbf{x}, \mathbf{p}_x) \simeq \sum_i \frac{1}{2m_i} (p_{xi}^2 + p_{yi}^2 + p_{zi}^2) + \mathscr{V}(\mathbf{x}) + P_\perp V - \gamma\mathscr{A}$

Figure 6: Examples of thermodynamic ensembles accessible to MD simulations. U denotes the internal energy of the system, S, its entropy and A, its Helmholtz free energy. $P_\perp$ stands for the normal pressure and $P_\parallel$, the lateral pressure applied to the simulation cell.

4.1. Constant temperature molecular dynamics

A number of schemes, more or less sophisticated, for carrying out isothermal MD simulations have been proposed. Perhaps the simplest consists in rescaling periodically the velocities by a factor $\sqrt{T/T_{\mathscr{T}}}$, where $T_{\mathscr{T}}$ denotes the instantaneous kinetic temperature — *viz.* $2\mathscr{T}(\mathbf{p}_x)/3Nk_B$, where k_B is the Boltzmann constant — and T, the desired temperature. Application of this corrective factor does not yield, however, a Newtonian MD. Newtonian mechanics implies that both the energy

and the momentum be conserved. Constant kinetic temperature MD requires solving the constrained equations of motion [2]:

$$\begin{cases} \dot{\mathbf{x}}_i = \dfrac{\mathbf{p}_{x,i}}{m_i} \\ \\ \dot{\mathbf{p}}_{x,i} = \mathbf{f}_i - \xi(\mathbf{x};\mathbf{p}_x)\ \mathbf{p}_x \end{cases} \tag{31}$$

in which $\xi(\mathbf{x};\mathbf{p}_x)$ may be related to a friction term that guarantees $\dot{T}_{\mathscr{T}} = 0$. This constraint is chosen in such a way that perturbation of the Newtonian trajectory is minimal:

$$\xi(\mathbf{x};\mathbf{p}_x) = \frac{\sum\limits_i \mathbf{p}_{x,i} \cdot \mathbf{f}_i}{\sum\limits_i |\mathbf{p}_{x,i}|^2} \tag{32}$$

Note: Replacing $\xi(\mathbf{x};\mathbf{p}_x)\ \mathbf{p}_x$ by $\xi(\mathbf{x};\mathbf{p}_x)\ \mathbf{p}_{x,i}$ would correspond to a rigorous thermostat, yielding the expected canonical distribution [2]. A second approach, more rigorous, consists in introducing in the equations of motion an additional degree of freedom, s. The velocity of particle i can, thus, be written $\mathbf{v}_i = s\ \dot{\mathbf{x}}_i = \mathbf{p}_{x,i}/m_i s$. Potential and kinetic terms are associated with the degree of freedom s, which can be related to the thermostat of the system:

$$\begin{cases} \mathscr{V}_s = \dfrac{1}{\beta}\ (f+1)\ \ln s \\ \\ \mathscr{T}_s = \dfrac{1}{2}\ \mathscr{Q}\ \dot{s}^2 \end{cases} \tag{33}$$

where $\beta \equiv 1/k_B T$, $\mathscr{Q}$ is the parameter of thermal inertia that regulates the fluctuations of temperature, and f is the number of degrees of freedom in the system — $3N - 3$ if the total momentum, $\mathbf{p}$, is a constant. Such an approach is known as extended Lagrangian, owing to the fact that the latter can be expressed as $\mathscr{L}_s(\mathbf{r};\mathbf{p}) = \mathscr{T}(\mathbf{p}_x) + \mathscr{T}_s(\mathbf{p}_x) - \mathscr{V}(\mathbf{x}) - \mathscr{V}_s(\mathbf{x})$. The equations of motion may then be restated as:

$$\begin{cases} \ddot{\mathbf{x}}_i = \dfrac{\mathbf{f}_i}{m_i s^2} - 2\ \dfrac{\dot{s}\dot{\mathbf{x}}}{s} \\ \\ \mathscr{Q}\ddot{s} = \sum\limits_i\ m_i \dot{x}_i^2 s - \dfrac{f+1}{\beta s} \end{cases} \tag{34}$$

This formalism, devised by Nosé [21], has been revisited by Hoover [22, 23], who suppressed the time–dependent parameter s. In the constrained equations of motion (31), the friction term is now given by a first–order differential equation:

$$\dot{\xi} = \frac{f}{\mathscr{Q}}\, k_B\, (T_{\mathscr{T}} - T) \tag{35}$$

The conserved quantity, here, is the total Hamiltonian, *i.e.* that of the chemical system plus the thermostat, $\mathscr{H}_s(\mathbf{x}; \mathbf{p}_x) = \mathscr{T}(\mathbf{p}_x) + \mathscr{T}_s(\mathbf{p}_x) + \mathscr{V}(\mathbf{x}) + \mathscr{V}_s(\mathbf{x})$.

The last approach, referred to as *weak coupling* [24], consists in letting the instantaneous kinetic temperature, $T_{\mathscr{T}}(t)$, "relax" towards the reference temperature, T, following:

$$\frac{\mathrm{d}T_{\mathscr{T}}(t)}{\mathrm{d}t} = \frac{T - T_{\mathscr{T}}(t)}{\tau_T} \tag{36}$$

where τ_T represents precisely the relaxation time associated with the fluctuations of the temperature. The kinetic energy is modified by a quantity $\Delta\mathscr{T}$, defined as:

$$\Delta\mathscr{T} = \frac{1}{2}\, (\chi^2 - 1)\, N\, k_B\, T_{\mathscr{T}}(t) \tag{37}$$

during a time–step, δt, by rescaling the velocities by a factor χ:

$$\chi = \left[1 + \frac{\delta t}{\tau_T} \left(\frac{T}{T_{\mathscr{T}}(t)} - 1\right)\right]^{1/2} \tag{38}$$

This aperiodic coupling to a "heat reservoir", by means of a first–order process, does not lead to oscillating responses to temperature changes. Yet, as has been demonstrated, neither does it yield the correct canonical distribution, in sharp contrast with the approach of Nosé and Hoover.

4.2. Constant pressure molecular dynamics

Here again, several of alternative methods, more or less sophisticated, can be adopted to maintain the simulation cell at a constant pressure in the course of time. It might, indeed, be desirable, in a number of instances, to generate trajectories in the isobaric–isothermal, (N, P, T), thermodynamic ensemble.

Just like for keeping the temperature at a constant value, the extended Lagrangian formalism can be applied to pressure. Initially devised by Andersen [25] this approach implies that the system be coupled to an external variable, V, characterizing the volume of the simulation box. This coupling symbolizes the action a piston would exert on the system, to which a kinetic and a potential term is associated:

$$\begin{cases} \mathscr{V}_V = \dfrac{1}{2}\, m_P\, \dot{V}^2 \\ \\ \mathscr{T}_V = P\, V \end{cases} \tag{39}$$

where m_P can be seen as the mass of the piston, and P denotes the desired pressure. Scaling of the positional variables, $\mathbf{r}$, and the velocities, $\mathbf{v}$, in the form $\mathbf{s} = \mathbf{r}/V^{1/3}$ and $\dot{\mathbf{s}} = \mathbf{v}/V^{1/3}$, one can rewrite the kinetic and potential energies as: $\mathscr{V}(\mathbf{r}) \equiv \mathscr{V}(V^{1/3}\mathbf{s})$ et $\mathscr{T}(\mathbf{p}_x) = \dfrac{1}{2}\, m\, V^{2/3} \sum\limits_i \dot{s}_i^2$. It follows that from the Lagrangian, $\mathscr{L}_V(\mathbf{x}; \mathbf{p}_x) = \mathscr{T}(\mathbf{p}_x) + \mathscr{T}_V(\mathbf{p}_x) - \mathscr{V}(\mathbf{x}) - \mathscr{V}_V(\mathbf{x})$, one can establish the new equations of motion:

$$\begin{cases} \ddot{\mathbf{s}}_i = \dfrac{\mathbf{f}_i}{m_i V^{1/3}} - \dfrac{2}{3}\, \dfrac{\dot{\mathbf{s}}_i \dot{V}}{V} \\ \\ \ddot{V} = \dfrac{P_{\mathscr{P}} - P}{m_P} \end{cases} \tag{40}$$

where the force, $\mathbf{f}_i$, and the instantaneous pressure, $P_{\mathscr{P}}$ — derived from the virial, $P_{\mathscr{P}} = \dfrac{1}{V}\, (N/\beta - \dfrac{1}{2} \sum\limits_i \mathbf{x}_i \cdot \mathbf{f}_i)$ — are evaluated from the unscaled Cartesian coordinates and momenta. Here, the conserved quantity throughout the MD simulation is the Hamiltonian of the extended system, $\mathscr{H}_V(\mathbf{x}; \mathbf{p}_x) = \mathscr{T}(\mathbf{p}_x) + \mathscr{T}_V(\mathbf{p}_x) + \mathscr{V}(\mathbf{x}) + \mathscr{V}_V(\mathbf{x})$, that is its enthalpy, to which is added a kinetic contribution of $\frac{1}{2\beta}$ arising from the fluctuations of the volume of the simulation cell. It should be underlined that, formally, this algorithm generates an isobaric–isoenthalpic distribution, (N, P, H). Its coupling to a thermostat, like the one governed by equation (34), yields the true isobaric–isothermal distribution.

It has been observed that the above scheme would lead to oscillations of $P_{\mathscr{P}}$, depending upon the mass of the piston, m_P. Feller *et al.* have devised an alternative that suppresses this undesirable effect by damping out the degree of freedom of the piston through a Langevin equation. Restating equation (40), it follows that:

$$\begin{cases} \ddot{\mathbf{s}}_i = \dfrac{\mathbf{f}_i}{m_i V^{1/3}} - \dfrac{2}{3}\, \dfrac{\dot{\mathbf{s}}_i \dot{V}}{V} \\ \\ \ddot{V} = \dfrac{P_{\mathscr{P}} - P}{m_P} - \gamma \dot{V} + R(t) \end{cases} \tag{41}$$

where γ is the collision frequency and $R(t)$, a random force taken from a Gaussian distribution of zero mean. Interestingly enough, $R(t)$ satisfies the fluctuation–dissipation relationship, *i.e.*

$\langle R(t_1)R(t_2)\rangle = \dfrac{2}{\beta m_P}\,\kappa(t_1 - t_2)$, where $\kappa(t)$ stands for the damping factor.

Another approached, proposed by Berendsen *et al.* [24] is an extension of the *weak coupling* algorithm described previously to constant–pressure simulations. Just like for the constant–temperature scheme, the equations of motion are modified in response to the relaxation of the instantaneous pressure, $P_{\mathscr{P}}(t)$, towards its reference value, P, according to:

$$\frac{\mathrm{d}P_{\mathscr{P}}(t)}{\mathrm{d}t} = \frac{P - P_{\mathscr{P}}(t)}{\tau_P} \tag{42}$$

Here, τ_P is the relaxation time associated to the fluctuations of the pressure. By rescaling the atomic coordinates of the system and the size of the periodic cell by a factor ς, the total volume is modified by $\Delta V = (\varsigma^3 - 1)\,V$, leading naturally to a variation of the pressure, which can be expressed as:

$$\Delta P = \frac{\Delta V}{\beta_{\mathscr{T}}\,V} \tag{43}$$

where $\beta_{\mathscr{T}}$ denotes the isothermal compressibility. Solving equations (42) and (43) for a given value of ς, it follows that:

$$\varsigma = \left[1 - \beta_{\mathscr{T}}\,\delta t\,\frac{P - P_{\mathscr{P}}(t)}{\tau_P}\right]^{1/3} \tag{44}$$

For reasons similar to the case of constant temperature, this algorithm does not yield a well–defined thermodynamic ensemble.

5. Handling electrostatic interactions

One particularly critical aspects of MD simulations lies in the appropriate treatment of electrostatic interactions. For obvious cost–effectiveness reasons, spherical truncation, which has been mentioned previously, remains widely utilized, especially when significant time–scales, beyond the nanosecond range, are being explored — such is the case of the simulation of a fully hydrated phospholipid bilayer. If the long–range nature of $1/r^3$ dipole–dipole interactions is sufficiently limited to guarantee a satisfactory reproduction of structural properties and statistical ensemble averages of the system, it still remains that the influence of these interactions on complex physical and chemical processes, like protein folding, would deserve a detailed analysis. The presence of ionic species is clearly more problematic, considering that the use of a spherical cut–off induces numerous artefacts that evidently distort the results of the simulation. Besides, it should be underlined that, limiting ourselves to dipolar

species, employing a spherical truncation causes singularities in the derivatives of the potential energy at the cut–off boundary. The deleterious effects of this spherical truncation can be reduced by adding a so–called switching function, [2] that replaces the abrupt Heavyside–type behavior by a smooth decrease in the neighborhood of the spherical boundary. This method does not circumvent, however, the difficulties arising from the presence of ions in the system. A more rigorous approach for handling charge–charge and charge–dipoles, seemingly inexpensive, relies upon the Debye–Hückel theory and the solution of the linearized Poisson–Boltzmann equation [26, 27]. As illustrated in Figure 8, this alternative referred to as *generalized reaction field* does not eliminate completely the artefacts when $r \rightarrow R_{\text{cut-off}}$, the radius of the sphere of truncation. Possibly the most rigorous approach is that proposed by Ewald. Admitting that the Coulomb sum,

$$\mathscr{V}_{\text{Coulomb}}(\mathbf{x}) = \sum_{i<j} \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1 r_{ij}} \tag{45}$$

over the simulation cell and its image neighbors, does not converge formally, the central idea of the method, known as *Ewald sum*, consists in breaking down expression (45) into two sums evaluated, respectively, in the direct and in the reciprocal spaces.

$$\sum_{\mathbf{n}} \frac{1}{|\mathbf{n}|}\mathscr{F}(\mathbf{n}) + \sum_{\mathbf{m}} \frac{1}{|\mathbf{m}|}[1 - \mathscr{F}(\mathbf{n})] \tag{46}$$

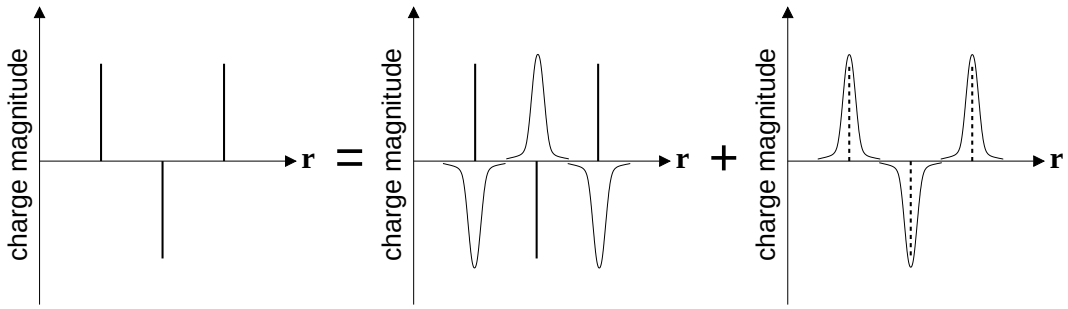


Figure 7: Components of an Ewald sum in a one–dimensional system of point charges. In the direct space, each charge is surrounded by a Gaussian charge distribution, $\varrho_i(\mathbf{x})$, of equal magnitude but opposite sign. This contribution is counter–balanced in the reciprocal space by a Gaussian distribution, $\varrho_j(\mathbf{x})$, of opposite sign.

By surrounding each point charge of the system by a Gaussian charge distribution:

$$\varrho_i(\mathbf{r}) = q_i\alpha^3 \frac{\exp(-\alpha^2 r^2)}{\sqrt{\pi^3}} \tag{47}$$

where α is a positive parameter characterizing the width of the Gaussian distribution, the first summation converges rapidly when $\mathbf{n} \longrightarrow \infty$, because $\mathscr{F}(\mathbf{n})$ decreases rapidly. The direct–space contribu-

tion is a short–range one — see Figure (7). The second summation evaluated in the reciprocal space uses a Fourier transform to solve the Poisson equation, *i.e.* $\nabla^2 V_i(\mathbf{r}) = -4\pi\varrho_i(\mathbf{r})$. The transform decreases rapidly, and the sum converges equally fast [28].

Implementation of lattice sums according to the scheme devised by Ewald in an MD program that uses a macomolecular force field, like the one described in equation (22), may be summarized as follows:

$$\begin{aligned}
\mathscr{V}_{\text{Ewald}}(\mathbf{r}) &= \frac{1}{2V\epsilon_0}\sum_{\mathbf{k}\neq\mathbf{0}}\frac{\exp(-k^2/4\alpha^2)}{k^2}\left[\sum_j q_j\exp(-i\mathbf{k}\cdot\mathbf{r}_j)\right]\left[\sum_j q_j\exp(i\mathbf{k}\cdot\mathbf{r}_j)\right] \\
&+ \frac{1}{4\pi\epsilon_0}\sum_i\sum_{j>i}\frac{q_iq_j}{r_{ij}}\,\text{erfc}(\alpha r_{ij}) - \frac{\alpha}{4\pi^{3/2}\epsilon_0}\sum_i q_i^2 \\
&+ \frac{1}{4\pi\epsilon_0}\sum_i\sum_{j>i}^{j\ \text{lié à}\ i}\frac{q_iq_j}{r_{ij}}
\end{aligned} \tag{48}$$

The first term corresponds to the reciprocal–space summation over **k**–vectors. The second term is the direct–space summation; $\text{erfc}(x)$ is the complementary error function, *i.e.* $1-\text{erf}(x)$, that explains the short–range nature of this rapidly converging term. The third and fourth contributions are corrections, owing to the fact that the reciprocal–space summation is carried over *all* atomic pairs $\{i,j\}$, that necessarily include self, 1–2 and 1–3 terms.

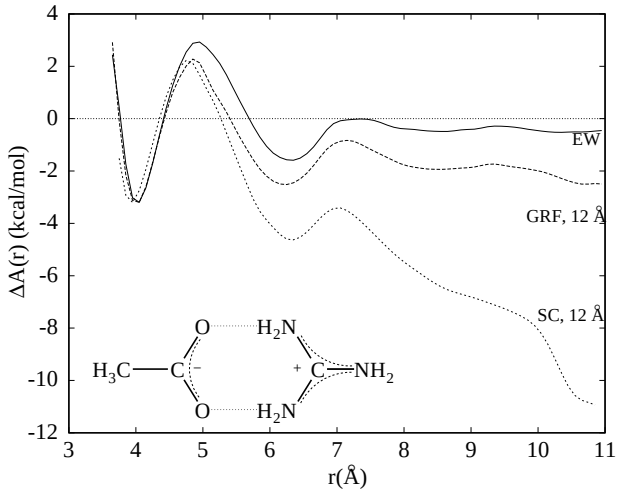


Figure 8: Free energy profile delineating the association of a guanidinium cation with an acetate anion in the $\mathscr{C}_{2v}$ geometry, in an aqueous environment. [29] The solid line curve corresponds to an Ewald sum (EW) simulation — α = 0.3 Å^{-1}; The dashed line curve, to a generalized reaction field (GRF) simulation — $R_{\text{cut-off}}$ = 12 Å; The dotted line curve, to simulation with a spherical truncation (SC) of intermolecular interactions — $R_{\text{cut-off}}$ = 12 Å.

Formally, the computational effort involved in a classical Ewald lattice sum is $\mathscr{O}(N^2)$, where N is the total number of particles of the system. As has been shown by Perram *et al.* [30], on the one hand, and Fincham [31], on the other hand, this cost can be reduced to $\mathscr{O}(N^{3/2})$ by choosing judiciously the width of the Gaussian distribution, α, the number of **k**–vectors, and the truncation of all pair interactions in the direct space. It is generally advised that the CPU time invested in the direct– and in the reciprocal–space sums be well balanced to reach the desired $N^{3/2}$–scaling.

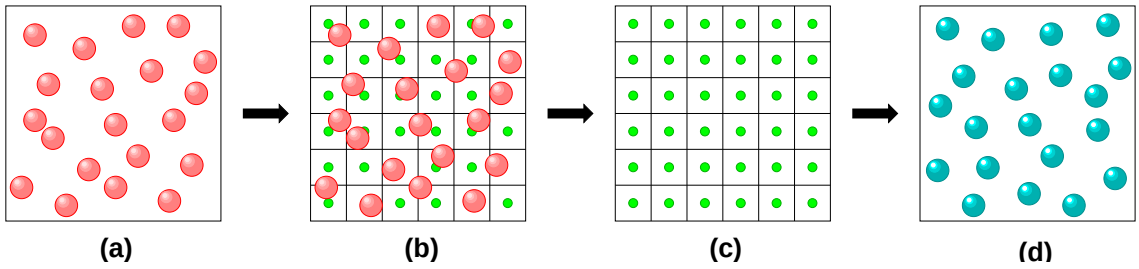


Figure 9: A *particle–mesh* scheme on a two–dimensional lattice. **(a)** A system of charged particles. **(b)** The charges are interpolated on a two–dimensional grid. **(c)** Employing a fast Fourier transform (FFT) approach, the potential and the forces are evaluated on each point of the grid. **(d)** Forces are interpolated back towards the particles, the position of which is subsequently updated.

Less expensive alternatives to the standard Ewald sum rely upon an evaluation of the reciprocal–space term using a fast Fourier transform (FFT). A three–dimensional grid filling the Cartesian space in which the MD simulation is carried out is constructed. The point charges borne by the particles of the system are interpolated over this grid, and the corresponding charge distribution, $\varrho(\mathbf{r})$, is computed. Employing an FFT technique, the transform of the charge distribution, $\hat{\varrho}(\mathbf{k})$, is determined in the basis of the reciprocal–space **k**–vectors. Next, the long–range contribution of the electrostatic potential is evaluated according to $\hat{V}_{\text{long}}(\mathbf{k}) = \hat{\mathscr{G}}(\mathbf{k})\hat{\varrho}(\mathbf{k})$, where $\hat{\mathscr{G}}(\mathbf{k})$ is the so–called influence function, defined by $\hat{\mathscr{G}}(\mathbf{k}) = \hat{\lambda}(\mathbf{k})/\epsilon_0 k^2$, in which $\lambda(\mathbf{r})$ is a distribution that depends upon the sole geometrical characteristics of the simulation cell. The $V_{\text{long}}(\mathbf{r})$ contribution is estimated at the various points of the three–dimensional grid by means of an inverse transform. Electrostatic forces are then determined by numerical derivation of the potential. Finally, the electric field and the potential are interpolated back from the grid towards the position of the particles — see Figure (9). This scheme constitutes the central idea of all *particle–mesh* algorithms [32,33], *e.g. particle–mesh Ewald* (PME) or *particle–particle–particle–mesh* [34] (P^3M), the CPU cost of which is formally $\mathscr{O}(N \ln N)$.

6. Accessing properties of the system from the trajectory

Among the static properties that can be extracted from the trajectories generated in the course of an MD simulation, structural quantities are particularly interesting, because they account for the local

order in the molecular system. Radial distribution functions (RDF), $g(r)$, indicate the probability to find a pair of atoms separated by a distance r, with respect to the probability expected for a completely random distribution with the same density [2, 35, 36], The definition of $g(r)$ requires the integration of the configurational distribution function over all atomic positions, except that of the two tagged particles:

$$g(\mathbf{x}_1;\mathbf{x}_2) = \frac{N(N-1)}{\rho^2 \int \mathrm{e}^{-\beta\mathscr{V}(\mathbf{x}_1,\ldots,\mathbf{x}_N)}\,\mathrm{d}\mathbf{x}_1\ldots\mathrm{d}\mathbf{x}_N} \int \mathrm{e}^{-\beta\mathscr{V}(\mathbf{x}_1,\ldots,\mathbf{x}_N)}\,\mathrm{d}\mathbf{x}_3\ldots\mathrm{d}\mathbf{x}_N \tag{49}$$

where ρ denotes the density of the liquid. For a system in which all atoms were identical, the RDF would reduce to a simple statistical average over pairs of atoms:

$$g(r) = \frac{V}{N^2}\left\langle \sum_i \sum_{j\neq i} \delta(\mathbf{r}-\mathbf{r}_{ij}) \right\rangle \tag{50}$$

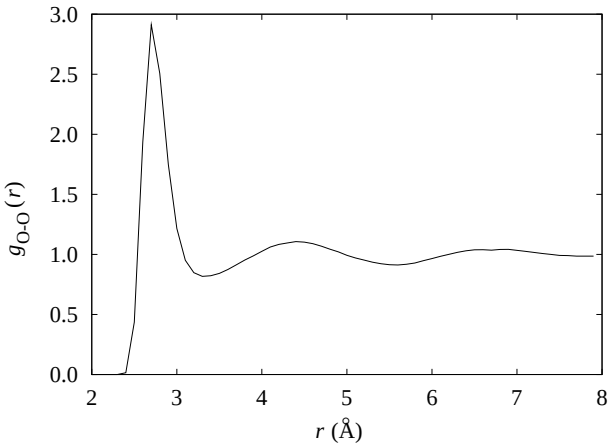


Figure 10: Oxygen–oxygen radial distribution function (RDF) of the TIP4P water model, obtained from a molecular dynamics simulation at 300 K.

For a given pair of atoms, $\{i, j\}$, the RDF can be readily evaluated through the following relationship:

$$g_{ij}(r) = \frac{\langle n_j(r+\delta r)\rangle}{4\pi\rho_j \int r^2 \mathrm{d}r} \tag{51}$$

where $\langle n_j(r+\delta r)\rangle$ corresponds to the average number of sites j, for which the distance to i is comprised between r and $r+\delta r$; ρ_j is a mean density of sites j in the sample of interest.

A second quantity, just as relevant to appreciate the order in a molecular liquid is the distance–dependent Kirkwood factor, $G_K(R)$ [37]. This quantity characterizes the correlation between the

dipole moment, $\boldsymbol{\mu}_i$, borne by a given molecule i and that borne by the neighboring molecules j contained in a sphere of radius R and centered at site i:

$$G_K(R) = \frac{\left\langle \sum_{i,j;r_{ij}<R} \boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_j \right\rangle}{N\mu^2} \tag{52}$$

The R–dependent Kirkwood factor is particularly sensitive to the treatment of electrostatic interactions in the molecular simulation. Using a spherical truncation causes several artefacts that deteriorate the dipole–dipole correlation near the edge of the sphere.

MD, as hinted by its name, may also supply valuable information about dynamical properties of the system investigated. Among the latter, time–correlation functions are particularly informative about the relaxation times for the various degrees of freedom in the system [1, 2]. The correlation between two observable quantities, $\mathscr{B}$ and $\mathscr{B}'$, can be expressed by:

$$c_{\mathscr{B}\mathscr{B}'} = \frac{\langle \delta\mathscr{B}\ \delta\mathscr{B}' \rangle}{\sigma(\mathscr{B})\ \sigma(\mathscr{B}')} \tag{53}$$

where $\delta\mathscr{B} = \mathscr{B} - \langle\mathscr{B}\rangle$, with $\langle\mathscr{B}\rangle$, is the statistical ensemble average of quantity $\mathscr{B}$; $\sigma(\mathscr{B}) = \sqrt{\langle\mathscr{B}^2\rangle - \langle\mathscr{B}\rangle^2}$. $c_{\mathscr{B}\mathscr{B}'}$ varies between 0 and 1, 0 corresponds to an absence of correlation. This formulation can be generalized in the case where $\mathscr{B}$ are $\mathscr{B}'$ are evaluated at distinct times. It follows that the correlation function now writes:

$$c_{\mathscr{B}\mathscr{B}'}(t) = \frac{\langle \delta\mathscr{B}(t)\ \delta\mathscr{B}'(0) \rangle}{\sigma(\mathscr{B})\ \sigma(\mathscr{B}')} \tag{54}$$

The time average at the numerator is carried over all time origins. In the event $\mathscr{B}' \equiv \mathscr{B}$, the calculated function is called an *auto–correlation function*, defined by:

$$c_{\mathscr{B}\mathscr{B}}(t) = \frac{\langle \delta\mathscr{B}(t)\ \delta\mathscr{B}(0) \rangle}{\langle \delta\mathscr{B}(0)\ \delta\mathscr{B}(0) \rangle} \tag{55}$$

Integration of the correlation function between $t = 0$ and $t = \infty$ yields the correlation time. As an illustration, let us consider the reorientational correlation time of benzene in water. The simulation time is necessarily longer than for a neat liquid, owing to the fact the statistical average is estimated from a single molecule.

The dynamical quantity evaluated here is the auto–correlation coefficient $c_{\hat{\mathbf{u}}\hat{\mathbf{u}}}(t) = \langle \hat{\mathbf{u}}(t)\hat{\mathbf{u}}(0) \rangle$, where $\hat{\mathbf{u}}$ is the unit vector borne by the C_6–axis or the C_2'–axis of benzene. Looking at the inertia tensor of

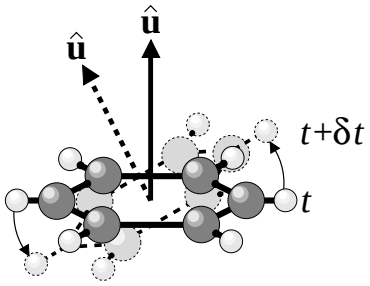


Figure 11: Reorientation of a benzene molecule as a function of time. $\hat{\mathbf{u}}$ is the unit vector borne by the C_6–axis of the molecule.

the molecule, it can be anticipated that the rotational motion about C_2' will be decorrelated faster than that about C_6.

From the knowledge of the correlation functions for the derivatives, $\dot{\mathscr{B}}$, of the observable, $\mathscr{B}$, rather than the observable itself, numerical simulations give access to transport coefficients. Equilibrium MD offers the possibility to estimate diffusion coefficients by means of the following integral [2]:

$$D = \frac{1}{3} \int_0^\infty \langle \dot{\mathbf{r}}_i(t) \cdot \dot{\mathbf{r}}_i(0) \rangle \; dt \tag{56}$$

Here, $\dot{\mathbf{r}}_i(t) \equiv \mathbf{v}_i(t)$ is the velocity of the center of mass of the molecule. At sufficiently long times, D may be obtained using the Einstein relationship:

$$D = \frac{1}{3} \frac{\left\langle |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \right\rangle}{2t} \tag{57}$$

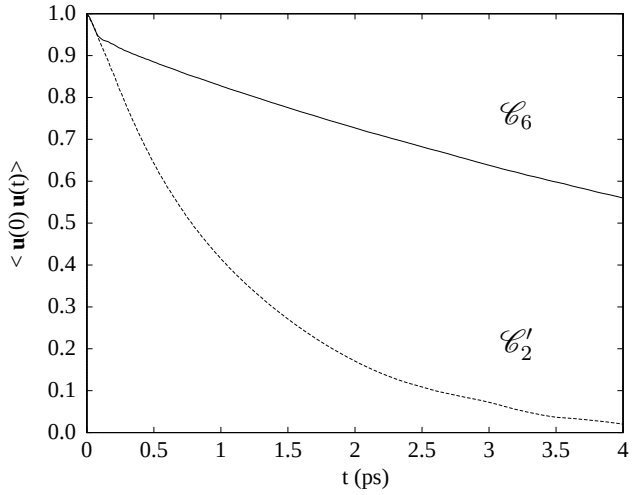


Figure 12: Reorientational auto–correlation functions of benzene in liquid water, at 300 K. Integration of the profiles yields the correlation times $\tau(C_6)$ = 2.4 ps and $\tau(C_2')$ = 0.5 ps.

7. Molecular dynamics and free energy calculations

We have seen that MD could provide the modeler with valuable information on the structural properties of the system, but also on its dynamics. We will now show that from an ensemble of configurations generated by MD, it is possible to access key thermodynamic quantities, like the free energy, utilized to predict the propensity of chemical species to associate or react. In many respects, the grounds for free energy calculations were laid several years ago by Kirkwood [38], Zwanzig [39], Bennett [40] and Valleau [41], but had to wait for the availability of significant computational power to be applied to molecular systems of chemical and biological relevance. In 1954, Zwanzig put forth a strategy, This approach, referred to as free energy perturbation (FEP), based on first principles of statistical mechanics, which allows free energy differences, ΔA, between two thermodynamic states, a and b, to be determined:

$$\Delta A = -\frac{1}{\beta} \ln \left\langle \exp \left\{ -\beta \left[\mathscr{H}_b(\mathbf{x}, \mathbf{p}_x) - \mathscr{H}_a(\mathbf{x}, \mathbf{p}_x) \right] \right\} \right\rangle_a \tag{58}$$

Here, $\mathscr{H}_a(\mathbf{x}, \mathbf{p}_x)$ is the Hamiltonian of the N–particle system in thermodynamic state a. $\langle \cdots \rangle_a$ denotes an ensemble average over configurations representative of this state. In most cases, states a and b are sufficiently disparate to prevent the brute application of equation (58). In practice, the reaction pathway between the reference and the target states is broken down into non–physical intermediates that are connected by means of a general extent parameter, λ, often called coupling parameter. Formally, equation (58) may be restated as a continuous integral over λ, thus leading to an alternative approach referred to as thermodynamic integration (TI) [38]:

$$\Delta A = \int_0^1 \left\langle \frac{\partial \mathscr{H}(\mathbf{x}, \mathbf{p}_x; \lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda \tag{59}$$

where λ connects state a, *i.e.* $\lambda = 0$, to state b, *i.e.* $\lambda = 1$.

Exercise: Starting for the original expression of the Helmholtz free energy, $A = -1/\beta \ln Q$, where Q stands for the canonical partition function, show that the free energy difference between two states a and b can be expressed as the ensemble average of equation (58).

The *umbrella sampling* method (US) [41] constitutes yet another route towards free energy differences that can be compared directly to experimental measurements. In this approach, sampling along an ordering parameter, ξ — *e.g.* possibly a true reaction coordinate — is restrained to a finite region of the configurational space by means of properly chosen external biasing potentials, $\mathscr{V}_{\mathrm{ext}}(\xi)$. Torrie and Valleau demonstrated that the unbiased statistical ensemble average of some quantity $\mathscr{B}$ may be

recovered from a biased ensemble:

$$\langle \mathscr{B} \rangle = \frac{\langle \mathscr{B} \ \exp\left[-\beta \mathscr{V}_{\mathrm{ext}}\right] \rangle_{\mathrm{biais}}}{\langle \exp\left[-\beta \mathscr{V}_{\mathrm{ext}}(\xi)\right] \rangle_{\mathrm{biais}}} \tag{60}$$

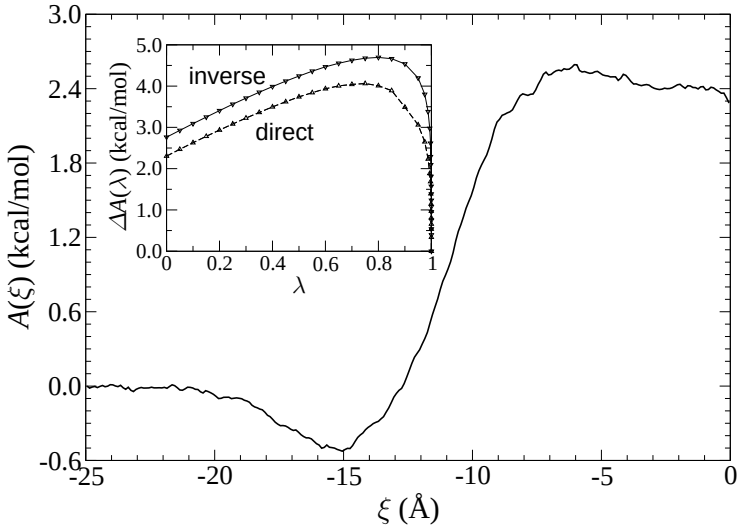


Figure 13: Estimation of the hydration free energy of methane through its transfer from the gas phase to the aqueous medium, and by annihilation (direct simulation) and creation (reverse simulation) in bulk water (inset). In the first case, one methane molecule diffuses freely along the normal to the water–air interface, using the ABF approach. In the second case, the water–methane interactions are progressively canceled or created as a function of λ. These two methodologies provide quantitatively comparable results, *viz.* 2.4$\pm$0.4 kcal/mol [42], in good agreement with the experimental estimate of 2.0 kcal/mol [43].

It follows that the free energy profile along ξ may be determined directly from the probability, $\mathscr{P}_{\mathrm{bias}}(\xi)$, to find the system over the range of values taken by the ordering parameter, computed in the biased configurational space:

$$A(\xi) = -\frac{1}{\beta} \ \ln \mathscr{P}_{\mathrm{bias}}(\xi) - \mathscr{V}_{\mathrm{ext}}(\xi) + A_0 \tag{61}$$

whereA_0 is a constant. One of the limitations of the US method lies in the necessity to define, without much it *a priori* knowledge, the shape of the external biasing potentials, $\mathscr{V}_{\mathrm{ext}}(\xi)$, which may turn out to be a daunting task in the case of qualitatively new problems. An appealing solution to this conundrum is provided by a recently proposed method called adaptive biasing force (ABF) [44], which relies on the integration of the average force exerted along ξ and obtained from unconstrained MD [42]. In the course of the simulation, a biasing force is rapidly estimated so that, when applied to the system, it supplies a Hamiltonian bereft of any residual average force acting along the ordering parameter.

As a result, all accessible values of ξ are sampled with a uniform probability, thereby improving significantly the accuracy of the computed free energies and the efficiency of the calculation.

As may be seen in Figure 13, coincidence of the estimates of the hydration free energy of methane from two distinct approaches — *viz.* ABF et FEP, suggests, irrespective of the intrinsic quality of the force field, that, by and large, the modeler has tamed the free energy methodology. This example further underlines the key role played by free energy calculations as a tangible link between theory and experiment.

Evidently, the present section does not pretend to be exhaustive. The reader interested by the topic of free energy calculations is invited to refer to specialized monographs — *e.g.* reference [45], for further detail.

8. Molecular dynamics and parallelism

Among the intrinsic limitations of MD, the most constraining one is without a doubt its computational cost when exploring complex molecular systems over significant time scales. Complexity increases with the number of atoms, which, in the case of biological systems, easily exceeds 10,000. In such assemblies, attaining simulation times compatible with the physical reality of the modeled phenomenon — which embraces the nano– to the microsecond time scale — evidently represents a numerical challenge. The difficulty to model these phenomena essentially stems from the minute time steps utilized to integrate the equations of motion, which is incommensurate with the duration of the proposed MD simulations. The example of the modeling of biological membranes, in which the collective motions of the lipid chains relax on a time scale that extends from 10^{-11} to 10^{-6} second, illustrates rather well this difficulty. This time scale should be compared to the usual 2×10^{-12} second time step that guarantees an appropriate conservation of the total energy of the system. The emergence of coarse–grain models, emancipated from the hardest degrees of freedom in the system, allow longer time steps to be used — *viz.* on the order of 40×10^{-12} second, thereby opening the way to the exploration of complex phenomena over microsecond time scales. Yet, on account of their very rudimentary nature, these models only supply a qualitative picture of the investigated processes. An atomistic description of the molecular systems and its inherent limitations are the price to pay to access a quantitative view of the targeted phenomena.

Over the past fifteen years, the development of parallel architectures has increased considerably the available computational resources, thereby pushing back the usual limitations of MD simulations. In particular, multiple–instruction, multiple–data (MIMD) massively parallel machines have made

possible the sampling of phase space over time scales hitherto never attained for complex molecular assemblies. As novel architectures would appear, so would alternative strategies for parallelizing numerical simulations, MD being recognized to represent an inherently parallelizable problem. Decomposition of the computational task in terms of subset of atoms constitutes undoubtedly the simplest scheme among all — see Figure 14. The ensemble of N particles of the system is spread out over the available n_{proc} processors of the machine. An array **a**, defined in the central memory, contains all the relevant information on the position of the particles. This information is necessarily shared by the n_{proc} processors for the computation of the forces involving atoms belonging to distinct subsets.

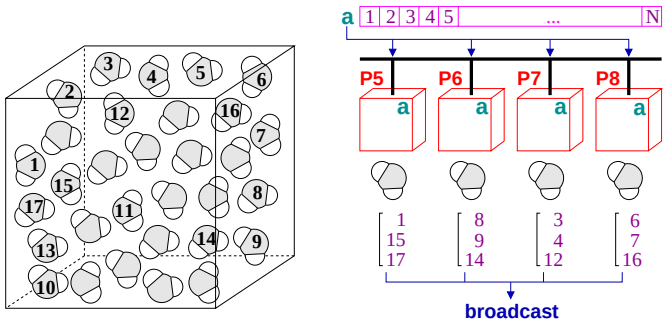


Figure 14: Decomposition of the computational task in terms of packets of particles. The task carried out in the main force loop is broken down over the available processors, each one of them handling a subset of atoms reasonably close in Cartesian space.

From the onset, it may be observed that this paradigm is limited to systems of rather moderate size and requires a significant memory. Parallelization of the computational task through breaking down the evaluation of the forces over the available n_{proc} processors is evidently more efficient in terms of memory usage — see Figure 15. In this approach, pair interactions are computed by blocks spread out over the various processors of the machine. One noteworthy advantage of this scheme lies in the possibility to "recycle" a scalar MD code by parallelizing the computationally expensive routines. Usage of compilation directives, *e.g.* OPENMP, in the framework of shared memory architectures is generally well suited for the force–decomposition strategy [46]. The difficulty to implement it with an appropriate efficiency is rooted in the possibility that atom i interacts simultaneously with atom j and atom k, the corresponding contributions being evaluated in distinct blocks. In this particular instance, update of the force, $\mathbf{f}_i$, exerted on atom i, from contributions due to atoms j and k, cannot be achieved concomitantly on distinct *threads*.

One may rapidly see in the force–decomposition strategy an asymptotical limit of the performances as a function of n_{proc}. This solution is an attractive one insofar as it applies to systems of moderate size, on parallel architectures with a limited number of processors — *viz.* typically $n_{\mathrm{proc}} <$

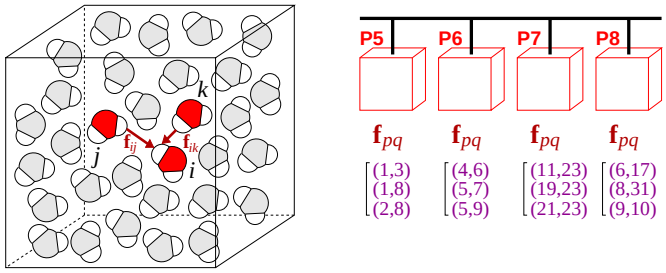


Figure 15: Parallelizing of the computational task by means of a force decomposition. The central loop for the calculation of the forces is spread over the different processors available and the contributions to the force exerted on the current atom, i, are evaluated concomitantly.

16. A more promising alternative, and without any contest a far more effective one, is the so–called domain–decomposition paradigm illustrated in Figure 16. The basic idea of this strategy consists in breaking the simulation cell into n_{cell} subcells that are handled in the most favorable circumstances quasi–independently by the n_{proc} processors of the architecture. The complete independence is never possible, even in the simplistic example of atomic fluids, because communication, even a minimalist one, is necessary to update the forces exerted onto the various particles of the system. The situation may become rapidly intricate when the same molecule — *e.g.* a protein, spans several contiguous subcells. This, among other aspects, explains why the development of a domain–decomposition MD code may be only achieved *ex nihillo*, following a philosophy distinct from that of the atom– and force–decomposition paradigms examined previously, and hence precluding the possibility of "recycling" an old scalar code.

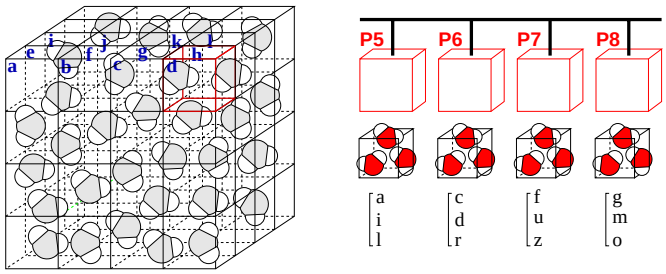


Figure 16: Decomposition into spatial domains. The simulation cell is broken down into several subcells, which are handled by the various processors of the parallel architecture. In order to balance the work load across the different processors, the main cell is usually broken down into a number of subcells greater than the number of available processors.

This paradigm has been the object of several improvements over the years, the detail of which is beyond the scope of this chapter. Among these improvements, load–balancing represents a notewor-

thy step to increase the performances of the parallel MD code, through spreading the $n_{\rm cell} \gg n_{\rm proc}$ subcells over the available processors, and shuffle the former until the computational effort evens out across these processors. This elegant strategy has been implemented in a variety of MD codes, like NAMD [47], today one of the most widely used for modeling complex biological systems.

9. Conclusion

The few illustrations reported here provide a glimpse into the amazing field of investigation accessible to numerical simulations for tackling problems encountered on a daily basis by the modeler. This field of investigation increases progressively as the price/performance ratio of modern computers decreases. Furthermore, access to massively parallel architectures, still somewhat uneven depending upon the country, opens the way to complex and realistic applications amenable to MD simulations. Returning to Figure 3, *in silico* studies of macromolecular systems of physical, chemical and biological interests, over significant time scales, are glaring examples of the progresses made in recent years not only on the hardware front, but also on the software one. It is apparent from this figure that simulation of assemblies of *ca.* 10^5 atomes over several nanoseconds has reached the level of routine modeling. In this range of size and time scales, the study of the structural fluctuations of a membrane protein in its natural lipid environment represents a concrete example. Combined to free energy calculations that emancipate the modeler from the limitations of Boltzmann sampling, MD constitutes an appropriate tool for modeling the slow events that occur in such intricate molecular systems, like the assisted transport of small molecules through a dedicated membrane protein [48].

Yet, these progresses suggest a number of comments. First, the quality of the computations carried out is strongly conditioned by the chosen strategy of the simulation. As a function of the problem tackled, a number of approximations should be banished, and it becomes necessary to turn to more rigorous methods, that are also often more expensive. This is, for instance, the case of the treatment of electrostatic interactions, which, in its simplest form, is limited to an additive pairwise approximation and a spherical truncation of the long–range forces, but, for a more sophisticated description, can include lattice sums and possibly models of distributed polarizabilities. If the methodology is nowadays well established and robust, simulating everything using MD is not yet possible, on account of the numerical integration of the Newton equations of motion by means of infinitesimal time–steps of *ca.* 1 to 2 fs, which should be compared to the time scales explored, together with the size of the system examined. This may constitute an insuperable obstacle to the routine use of such large–scale computations. Moreover, if it is tempting to turn systematically to numerical simulations, as molecular modeling tends to proliferate in an increasingly large number of domains of physics, chemistry

and biology, in the vast majority of cases, MD codes should not be considered as *black boxes*. Only a careful examination of the results obtained, a detailed analysis of the trajectories generated and a thorough verification of the derived thermodynamic properties will confirm the correctness and the physical and chemical meaning of the calculations performed [49].

Acknowledgments

I gratefully acknowledge my colleagues, François Dehez, Mounir Tarek, Jérôme Delhommelle and Michael A. Wilson for having accepted to proof–read this manuscript, and for their insightful comments and helpful suggestions to improve its quality and readability.

References

[1] Frenkel, D.; Smit, B., *Understanding molecular simulations: From algorithms to applications*, Academic Press: San Diego, 1996.

[2] Allen, M. P.; Tildesley, D. J., *Computer Simulation of Liquids*, Clarendon Press: Oxford, 1987.

[3] Ewald, P., Die Berechnung optischer und elektrostatischer Gitterpotentiale, *Ann. Phys.* **1921**, *64*, 253–287.

[4] Ladd, A. J. C., Long–range dipolar interactions in computer simulations of polar liquids, *Mol. Phys.* **1978**, *36*, 463–474.

[5] van Gunsteren, W. F.; Berendsen, H. J. C., Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry, *Angew. Chem. Int. Ed. Engl.* **1990**, *29*, 992–1023.

[6] Sanbonmatsu, K. Y.; Simpson, J.; Tung, C. S., Simulating movement of tRNA into the ribosome during decoding, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15854–15859.

[7] Duan, Y.; Kollman, P. A., Pathways to a protein folding intermediate observed in a 1–microsecond simulation in aqueous solution, *Science* **1998**, *282*, 740–744.

[8] Tuckerman, M. E.; Martyna, G. J., Understanding modern molecular dynamics: Techniques and applications, *J. Phys. Chem. B* **2000**, *104*, 159–178.

[9] Leimkuhler, B.; Reich, S., *Simulating Hamiltonian dynamics*, Cambridge University Press, 2005.

[10] Hairer, E.; Lubich, C.; Wanner, G. Geometric numerical integration. Structure–preserving algorithms for ordinary differential equations. in *Springer series in computational mathematics*, vol. 31. Springer–Verlag, Berlin, [2nd. edition].

[11] Martyna, G. J.; Tuckerman, M. E.; Tobias, D. J.; Klein, M. L., Explicit reversible integrators for extended systems dynamics, *Mol. Phys.* **1996**, *87*, 1117–1128.

[12] Verlet, L., Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard–Jones molecules, *Phys. Rev.* **1967**, *159*, 98–103.

[13] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz Jr., K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. C.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids, and organic molecules, *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

[14] Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. The development/application of a "minimalist" force field using a combination of ab initio and experimental data. in *Computer simulation of biomolecular systems: Theoretical and experimental applications*, Van Gunsteren, W. F.; Weiner, P. K., Eds. Escom, The Netherlands, 1997, pp. 83–96.

[15] Ryckaert, J.; Bellemans, A., Molecular dynamics of liquid alkanes, *Chem. Soc. Faraday Discuss.* **1978**, *66*, 95–106.

[16] Chipot, C.; Ángyán, J. G., Continuing challenges in the parametrization of intermolecular force fields. Towards an accurate description of electrostatic and induction terms, *New J. Chem.* **2005**, *29*, 411–420.

[17] Wang, W.; Skeel, R. D., Fast evaluation of polarizable forces, *J. Chem. Phys.* **2005**, *123*, 164107.

[18] Gay, J. G.; Berne, B. J., Modification of the overlap potential to mimic a linear site–site potential, *J. Chem. Phys.* **1981**, *74*, 3316–3319.

[19] Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L., A coarse grain model for phospholipid simulations, *J. Phys. Chem. B.* **2001**, *105*, 4464–4470.

[20] Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L., Coarse grain models and the computer simulation of soft materials, *J. Phys.: Condens. Matter* **2004**, *16*, R481–R512.

[21] Nosé, S., A molecular dynamics method for simulations in the canonical ensemble, *Mol. Phys.* **1984**, *52*, 255–268.

[22] Hoover, W. G., Nonequilibrium molecular dynamics, *A. Rev. Phys. Chem.* **1983**, *34*, 103–127.

[23] Hoover, W. G., Canonical dynamics: Equilibrium phase–space distributions, *Phys. Rev.* **1985**, *A31*, 1695–1697.

[24] Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath, *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[25] Andersen, H. C., Molecular dynamics simulations at constant pressure and/or temperature, *J. Chem. Phys.* **1980**, *72*, 2384–2393.

[26] Barker, J. A., Reaction field, screening, and long–range interactions in simulations of ionic and dipolar systems, *Mol. Phys.* **1994**, *83*, 1057–1064.

[27] Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F., A generalized reaction field method for molecular dynamics simulations, *J. Chem. Phys.* **1995**, *102*, 5451–5459.

[28] Toukmaji, A. Y.; Board Jr., J. A., Ewald summation techniques in perspective: A survey, *Comput. Phys. Comm.* **1996**, *95*, 73–92.

[29] Rozanska, X.; Chipot, C., Modeling ion–ion interaction in proteins: A molecular dynamics free energy calculation of the guanidinium–acetate association, *J. Chem. Phys.* **2000**, *112*, 9691–9694.

[30] Perram, J. W.; Petersen, H. G.; de Leeuw, S. W., An algorithm for the simulation of condensed matter which grows as the 3/2 power of the number of particles, *Mol. Phys.* **1988**, *65*, 875–889.

[31] Fincham, D., Optimisation of the Ewald sum for large systems, *Mol. Sim.* **1994**, *13*, 1–9.

[32] Darden, T. A.; York, D. M.; Pedersen, L. G., Particle mesh Ewald: An $NlogN$ method for ewald sums in large systems, *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[33] Essman, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. G., A smooth particle mesh Ewald method, *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[34] Hockney, R. W.; Eastwood, J. W., *Computer simulation using particles*, IOP Publishing Ltd.: Bristol, England, 1988.

[35] McQuarrie, D. A., *Statistical mechanics*, Harper and Row: New York, 1976.

[36] Chandler, D., *Introduction to modern statistical mechanics*, Oxford University Press, 1987.

[37] Neumann, M.; Steinhauser, O., The influence of boundary conditions used in machine simulations on the structure of polar systems, *Mol. Phys.* **1980**, *39*, 437–54.

[38] Kirkwood, J. G., Statistical mechanics of fluid mixtures, *J. Chem. Phys.* **1935**, *3*, 300–313.

[39] Zwanzig, R. W., High–temperature equation of state by a perturbation method. I. Nonpolar gases, *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[40] Bennett, C. H., Efficient estimation of free energy differences from Monte Carlo data, *J. Comp. Phys.* **1976**, *22*, 245–268.

[41] Torrie, G. M.; Valleau, J. P., Nonphysical sampling distributions in Monte Carlo free energy estimation: Umbrella sampling, *J. Comput. Phys.* **1977**, *23*, 187–199.

[42] Hénin, J.; Chipot, C., Overcoming free energy barriers using unconstrained molecular dynamics simulations, *J. Chem. Phys.* **2004**, *121*, 2904–2914.

[43] Ben-Naim, A.; Marcus, Y., Solvation thermodynamics of nonionic solutes, *J. Chem. Phys.* **1984**, *81*, 2016–2027.

[44] Darve, E.; Pohorille, A., Calculating free energies using average force, *J. Chem. Phys.* **2001**, *115*, 9169–9183.

[45] Chipot, C. Calculating free energy differences from perturbation theory. in *Free energy calculations. Theory and applications in chemistry and biology*, Chipot, C.; Pohorille, A., Eds. Springer Verlag, Berlin–Heidelberg, 2006.

[46] Couturier, R.; Chipot, C., Parallel molecular dynamics using OPENMP on a shared memory machine, *Comp. Phys. Comm.* **2000**, *124*, 49–59.

[47] Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, L.; Schulten, K., Scalable molecular dynamics with NAMD, *J. Comput. Chem.* **2005**, *26*, 1781–1802.

[48] Tajkhorshid, E.; Nollert, P.; Jensen, M. Ø.; Miercke, L. J. W.; O'Connell, J.; Stroud, R. M.; Schulten, K., Control of the selectivity of the aquaporin water channel family by global orientational tuning, *Science* **2002**, *296*, 525–530.

[49] van Gunsteren, W. F.; Mark, A. E., Validation of molecular dynamics simulation, *J. Chem. Phys.* **1998**, *108*, 6109–6116.