Università degli Studi di Bologna

Alma Mater Studiorum

Dipartimento di Scienze Farmaceutiche

Dottorato di Ricerca in Scienze Farmaceutiche
CHIM/08

Coordinatore: Chiar.mo Prof. Domenico Spinelli

Classical molecular dynamics studies of pharmaceutically relevant biological systems

Tesi di Dottorato presentata da

Matteo Masetti

Supervisor: **Chiar.mo Prof. Maurizio Recanatini**

Advisor: **Dr. Andrea Cavalli**

Doctor Philosophiæ

XIX Ciclo (2004-2006)

Classical molecular dynamics studies of pharmaceutically relevant biological systems

TOPICS:

Chapter 1

Introduction

A fundamental postulate in contemporary medicinal chemistry is that the effect of a drug in the human body is the consequence of the molecular recognition between a ligand (the drug) and a macromolecule (the target). The effect of binding can be either promotion or inhibition of signal transduction of some enzymatic activity or molecular transport. In this context, during last years, computational tools in medicinal chemistry have played a prominent role in the understanding of the molecular events lying at the basis of the therapeutic effects of drugs. In particular, computational chemistry tools allow the characterization of structure, dynamics and energetics of the above mentioned interactions. Moreover, the development of more accurate and reliable algorithms, the employ of more thoughtful protocols to apply them, and greatly increased computational power nowadays allow studies to be performed with the necessary reliability and accuracy. Even though there is no substitute for quantum mechanics when an explicit description of electronic features is demanded, classical mechanics based approaches can efficiently assist the study of pharmaceutically relevant systems, and these computationally relatively inexpensive methods are nowadays routinely used in modern rational drug design. However, since there is not an univocal strategy to solve a drug design-related problem even among the classical level of theory the appropriate computational method to be used will depend upon both the characteristics of the system itself and the available information. Accordingly, a number of approaches can be applied at different stages of the drug design process: at early stages the speed is usually required at the expense of an optimal physical description. Instead, at the end, namely during the lead-optimization stages, the emphasis unavoidably relies on the accuracy. Within this scenario, docking algorithms play a pivotal role in the first stage of the drug design, whereas for the second stage a wider spread of techniques are used. Nevertheless, molecular dynamics simulations represent one of the most used approaches, and they represents the main focus of the present thesis.

In particular, docking techniques are designed to find the correct conformation of a ligand at the binding site of a target protein, and have now been used with different success for decades. The idea behind this technique is only theoretically simple, since several entropic effects, which are hardly handled in the whole plethora of computational methods, could often take place. The mobility of both ligand and receptor, polarization effects acting on the small molecule arising from the protein environment, and their interactions with the neighboring water molecules, which would not be in principle neglected (but almost they are), further complicates the quantitative description of the process under investigation.

On the other hand, molecular dynamics (MD) simulations are valuable for understanding the dynamical behavior of proteins (or complexes) at different timescales, and they represent one of the most versatile and widely applied computational techniques for the study of biological systems. The first protein MD simulation was performed by the McCammon group in 1977, and consisted in the study of the bovine pancreatic trypsin inhibitor[1]. The system comprised 58 amino acidic residues, and the simulation was run in vacuum for a total time of 8.8 ps. Starting from this pioneering work, the relatively enormous computational power reached at the present, permits routinely the simulation of systems comprising 10^5 atoms for tens of nanoseconds. Hence, simulations of more realistic systems, including explicit water molecules, counterions, and even complete membrane-like environment are nowadays affordable, and observable properties can be monitored as they evolve in real time. Such a technological progress in mere computational power has been accompanied by methodological improvements: better force fields, improved treatment of long range electrostatics and boundary conditions, and better algorithms used in order to control temperature and pressure.

Finally, the close interplay between docking procedures and molecular dynamics techniques also needs to be accounted for. In particular, fast and relatively inexpensive docking protocols can be combined with accurate but more costly MD techniques to predict more reliable ligand-protein complexes. The strength of this combination relies in their complementary strength and weakness. Docking techniques are used to quickly investigate the huge conformational space of pharmaceutically relevant molecules, although the major drawback lies in the total or relative absence of protein flexibility, which actually prevents the treatment of important phenomena such as induced fit. On the other hand, MD simulations naturally take into account flexibility. Moreover, the effect of explicit water molecules can be directly treated. However, the main problems with equilibrium MD simulations are that they can be highly time consuming, and that they can easily get stuck in local free-energy minima, thus seriously limiting or slowing down the phase space sampling. From the above discussion, it is clear that the combination of the two techniques into a single protocol is a logical consequent approach to improve the drug design process. Actually, there are two main ways to link the two techniques: i) refinement, and ii) relaxing approaches. The first ones, consist in the dynamical study of the differential behavior of selected ligand-protein complexes previously obtained by means of a standard docking procedure (refinement). By analyzing the relative stability and/or conformational changes during the time course, important insight in respect to the correct binding mode can be achieved. Such an approach suffers from the fact that the ligand has to be properly parameterized in order to properly match with the force field chosen to treat the protein. Usually, protein force fields are highly specific in respect to peptides, hence it is not trivial that a suitable consistent parameterization for the ligand could be easily obtained. Anyway, until now, routinary methods still do not exists, therefore the ligand parameterization represents a crucial and at the same time quite tricky phase of the setup for molecular dynamics simulations of pharmaceutically relevant complexes. Since it could seriously affect the reliability of the resulting data, the ligand parameterization requires a rather high amount of experience in modeling. The second approach is less intuitive. It was originally introduced by McCammon and co-workers[2] and it attempts to take into account the possibility for a ligand to bind only a few relaxed conformations of the receptor (relaxation). Such an approach is usually (but not only) employed if a crystallographic structure of the protein is not available, and the docking of small molecules into the binding site of crude homology derived models would introduce serious artifacts as the protein is not properly relaxed in a suitable biological environment (water, lipid membrane). Hence, a long MD simulation of the *apo* form for the protein is firstly performed in order to extensively sample its phase space, then the docking of candidate ligands is carried out on a large ensemble of protein conformations. With such an approach the parameterization of the small molecule is usually avoided, even though a further dynamical equilibration of the ligand-protein complex would be in some case advisable.

In the present thesis, both equilibrium and non-equilibrium MD approaches applied to the study of biologically relevant systems will be discussed. As previously reported, although the main focus of the work is addressed to MD techniques, both the above mentioned strategies (refinement, relaxation) of linking MD with docking simulations will be covered, even if in different theoretical contexts (dynamics and metadynamics).

- The first study reported (chapter 3.1) focuses on the hERG potassium channel.
  Prolongation of the QT interval of the electrocardiogram is a typical effect of Class III antiarrhythmic drugs, achieved through blockade of potassium channels. It has been found that several classes of drugs used for non-cardiovascular therapies, may prolong the QT interval by means of a similar mechanism, and in particular by blocking the hERG potassium channel. The great interest in QT prolongation has arisen for several reasons. Among them, the most important is that drug-induced QT prolongation increases the likelihood of a polymorphous ventricular arrhythmia, called *torsades de pointes*, which in

turn may cause syncope and then degenerate into ventricular fibrillation and sudden death. Hence, the binding mode of drugs which show blocking effect against the hERG channel has become of great interest in the pharmaceutical community, and in particular within the computational area of research because the geometrical features of the binding site are still experimentally unknown. Such an increased interest has promoted the hERG channel to be one of the most popular anti-targets.

In this thesis, we believe that important insights about the binding mode of the most potent blocker so far known (Astemizole) will be given, even if an in depth understanding of a general binding mode has not been reached yet. With respect to this, further computational efforts are perhaps needed, and maybe it is reasonable to think that only a tight collaboration with experimentalists will at last completely unravel the hidden secrets of such a fascinating protein complex.

- The second study reported (chapter 3.2) relies on the proposal of a novel approach to be exploited in order to estimate the free-energy of binding of docked complexes. In particular, the approach is addressed to provide an univocal discrimination between poses which represent the outcome of docking protocols, hence avoiding the use of scoring functions, since their reliability has recently been criticized especially when entropic or solvation components of the free-energy of binding are not negligible. To do this we propose the combination of: i) the efficient search algorithm of (most) docking programs, ii) a robust geometrical cluster analysis program, and iii) metadynamics simulation performed by using a proper set of collective variables in order to accelerate the unbinding event and – at the same time – to reconstruct the free-energy of the investigated process.

  In this thesis, only a case study of the method will be explicitly discussed. In particular the docking/undocking of a typical ureidic inhibitor on the glycogen synthase kinase 3β protein will be taken into account. Moreover, accuracy and feasibility of the technique in a pharmaceutical perspective will be given.

References:

[1] McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature*. (1977); **267**: 585

[2] Lin, J. H.; Perryman, A. L.; Schames, J. R.; McCammon, J. A. *J Am Chem Soc*. (2002); **124**: 5632

Chapter 2

Theoretical Methods: from first principles to parameters.

In this section, the main theoretical background for classic mechanical simulations will be summarized, starting from the very basics of quantum mechanics to the classical one.

A short discussion on the different formalisms used in this thesis will introduce the chapter.

Except for metadynamics which is a recently developed method and hence references have to be explicitly accounted for, the following books were an inestimable source of information, needed to write the present chapter:

- Allen, M. P.; Tildesley, D. J. Computer simulation of liquids. *Clarendon Press* (1991).
- Foresman, J. B.; Frisch, A.; Exploring chemistry with electronic structure methods. *Gaussian Inc* (1996).
- Frenkel, D.; Smit, B.; Understanding molecular simulation. *Academic Press* (2002).
- Hinchliffe, A.; Modelling molecular structures. *John Wiley & Sons* (1995).
- Koch, W.; Holthausen, M. C.; A chemist's guide to density functional theory. *Wiley-VCH* (2001).
- Jensen, F.; Introduction to computational chemistry. *John Wiley & Sons* (2001).
- Leach, A. R; Molecular modelling. *Prentice Hall* (2001).

References for the theoretical background of metadynamics:

- Laio, A.; Parrinello, M. Escaping free-energy minima. *PNAS*. (2002), **99**: 12562 – 12566.
- Iannuzzi, M.; Laio, A.; Parrinello, M. Efficient exploration of reactive potential energy surfaces using Car-Parrinello molecular dynamics. *Phys Rev Lett*. (2003), **90**: 1 – 4.
- Micheletti, C.; Laio, A.; Parrinello, M. Reconstructing the density of states by history-dependent metadynamics. *Phys Rev Lett*. (2004), **92**: 1 – 4.
- Laio, A.; Fortea-Rodriguez, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. Assessing the accuracy of metadynamics. *J Phys Chem B*. (2005), **109**: 6714 – 6721.
- Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J Phys Chem B*. (2006), **110**: 3533 – 3539.

2.0 Different formulations of the laws of mechanics

Based on different principles, different equations of mechanics can be derived. Since the result has to be the same, the choice of a formalism instead of another is a mere matter of calculation convenience. Basically, three formalisms are reported:

- Newtonian mechanics;
- Lagrangian mechanics;
- Hamiltonian mechanics.

Newtonian mechanics represents the most intuitive one, since its equations directly derive from a “human-scale” experience. In this scheme, the fundamental equation rely on the Newton’s second law:

$$\mathbf{F} = m\ddot{\mathbf{r}} \qquad [2.1]$$

while, for the energy conservation principle, an isolated system is characterized by a total internal energy having the form:

$$\mathrm{E} = \mathrm{T} + \mathrm{V} \qquad [2.2]$$

where T and V stand for the kinetic and potential energy, respectively. Introducing the classical linear momentum as $\mathbf{p} = m\mathbf{v}$, the same equation becomes:

$$\mathrm{E} = \frac{\mathbf{p}^2}{2m} + \mathrm{V} \qquad [2.3]$$

from which is clear that the *trajectory* of a particle is univocally determined once the position and the momentum at a given time are known.

Conversely, non-Newtonian formulations are based on a variational principle, in particular the Hamilton’s principle. In the Lagrange’s formalism the mechanics of a system is described in terms of its generalized (not necessarily Cartesian) coordinates and velocities by means of the function:

$$\mathrm{L}(\mathbf{q}, \dot{\mathbf{q}}, t) \qquad [2.4]$$

The trajectory followed by a system over a time interval $\{t_0, t_1\}$ between an initial position $\mathbf{x}_0$ and a final position $\mathbf{x}_1$, is the one for which the *action*, that is the functional S, is an extremum (i.e. a stationary point, usually – but not necessarily – a minimum):

$$\mathrm{S} = \int_{t_0}^{t_1} \mathrm{L}(\mathbf{q}, \dot{\mathbf{q}}, t)\,\mathrm{d}t \qquad [2.5]$$

Being S stationary, it means that the action does not vary for infinitesimal deformations of the trajectory, or – in other words – the Hamilton’s principle can be rewritten as:

$$\delta\mathrm{S} = 0 \qquad [2.6]$$

Equivalently, considering the trajectory $\mathbf{q}(t)$ which for simplicity minimizes S (since not only minima are extrema), and considering a slight variation for which $\delta\mathbf{q}(t_1) = \delta\mathbf{q}(t_2)$, it follows:

$$\mathbf{q}'(t) = \mathbf{q}(t) + \delta\mathbf{q}(t) \qquad [2.7]$$

As a consequence, the variation $\delta\mathrm{S}$ is null, since the effective motion has now been replaced by an infinitesimally modified motion nearly close to the former. Hence, from equation 2.6 an infinitesimal variation of S can be written as:

$$\partial\,\mathrm{S} = \int_{t_1}^{t_2} \mathrm{L}(\mathbf{q}(t) + \partial\mathbf{q}(t), \dot{\mathbf{q}}(t) + \partial\dot{\mathbf{q}}(t), t)dt - \int_{t_1}^{t_2} \mathrm{L}(\mathbf{q}(t), \dot{\mathbf{q}}(t), t)dt = 0 \qquad [2.8]$$

leading by rearrangement to the Lagrange’s equation:

$$\frac{\partial\,\mathrm{L}}{\partial\mathbf{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\,\mathrm{L}}{\partial\dot{\mathbf{q}}} = 0 \qquad [2.9]$$

which corresponds to the Newton’s equations of motion in generalized coordinates.

From a computational point of view the trajectory is then calculated by minimizing the action S. Since it can be demonstrated that the Lagrangian function corresponds to the *difference* between the kinetic and the potential energy of the system:

$$\mathrm{L} = \mathrm{T} - \mathrm{V} \qquad [2.10]$$

the Lagrange’s equation can be rewritten as:

$$\frac{\partial\,\mathrm{L}}{\partial\mathbf{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\,\mathrm{L}}{\partial\dot{\mathbf{q}}} = \frac{\partial(T-V)}{\partial\mathbf{q}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial(T-V)}{\partial\dot{\mathbf{q}}} = 0 \qquad [2.11]$$

Switching to Cartesian coordinates, since kinetic energy is not a function of coordinates, it follows that the first term can be rewritten as:

$$\frac{\delta\mathrm{L}}{\delta\mathbf{q}} = -\frac{\delta\mathrm{V}}{\delta\mathbf{q}} = \mathbf{F} \qquad [2.12]$$

and since the potential energy is not a function of velocities, for the second term:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\,\mathrm{L}}{\partial\dot{\mathbf{q}}} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial T}{\partial\dot{\mathbf{q}}} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial\left(\frac{1}{2}m\dot{\mathbf{q}}^2\right)}{\partial\dot{\mathbf{q}}} = \frac{\mathrm{d}}{\mathrm{d}t}(m\dot{\mathbf{q}}) = m\ddot{\mathbf{q}} \qquad [2.13]$$

demonstrating that in Cartesian coordinates the Lagrangian equation correspond to the Newton’s second equation. Furthermore, it can be demonstrated that the internal energy of an isolated system is:

$$\mathrm{E} = \dot{\mathbf{q}}\frac{\delta\mathrm{L}}{\delta\dot{\mathbf{q}}} - \mathrm{L} \qquad [2.14]$$

Besides, Hamiltonian mechanics is a reformulation of Lagrangian formalism allowing the derivation of equation of motions in terms of generalized positions and their conjugated moment, thus providing a direct link with Newtonian mechanics. The Hamiltonian is defined as a Legendre transform of the Lagrangian:

$$\mathrm{H}(\mathbf{q},\mathbf{p},t)=\mathbf{p}\dot{\mathbf{q}}-\mathrm{L}(\mathbf{q},\dot{\mathbf{q}},t) \qquad [2.15]$$

where an infinitesimal variation of H can both be written as:

$$\mathrm{dH}(\mathbf{q},\mathbf{p},t)=\frac{\partial\,\mathrm{H}}{\partial\mathbf{p}}\mathrm{d}\,\mathbf{p}+\frac{\partial\,\mathrm{H}}{\partial\mathbf{q}}\mathrm{d}\,\mathbf{q}+\frac{\partial\,\mathrm{H}}{\partial t}\mathrm{d}\,t \qquad [2.16]$$

or as:

$$\begin{aligned}\mathrm{dH}(\mathbf{q},\mathbf{p},t)&=\mathrm{d}(\mathbf{p}\dot{\mathbf{q}})-\mathrm{dL}(\mathbf{q},\dot{\mathbf{q}})\\&=\mathbf{q}\,\mathrm{d}\,\mathbf{p}-\mathbf{p}\,\mathrm{d}\,\mathbf{q}-\frac{\partial\,\mathrm{L}}{\partial t}\mathrm{d}\,t\end{aligned} \qquad [2.17]$$

From which the Hamiltonian equations of motion directly follows:

$$\frac{\partial\,\mathrm{H}}{\partial\mathbf{p}}=\dot{\mathbf{q}} \qquad [2.18]$$

and:

$$\frac{\partial\,\mathrm{H}}{\partial\mathbf{q}}=-\dot{\mathbf{p}} \qquad [2.19]$$

which in Cartesian coordinates correspond to $\dot{\mathbf{r}}=\dot{\mathbf{r}}$ and $\mathbf{F}=m\ddot{\mathbf{r}}$ respectively, whereas the Hamiltonian function corresponds to the internal energy of an isolated system:

$$\begin{aligned}\mathrm{H}(\mathbf{r},\mathbf{p})&=\dot{\mathbf{r}}\mathbf{p}-\mathrm{L}(\mathbf{r},\dot{\mathbf{r}})\\&=m\dot{\mathbf{r}}^2-\frac{1}{2}m\dot{\mathbf{r}}^2+\mathrm{V}(\mathbf{r})\\&=\frac{1}{2m}\mathbf{p}^2+\mathrm{V}(\mathbf{r})\end{aligned} \qquad [2.20]$$

or in a simpler way:

$$\mathrm{H}=\mathrm{T}+\mathrm{V} \qquad [2.21]$$

As previously reported, different formulations of mechanics yield identical result, even if Hamiltonian equation are two *first-order* differential equations for 3*N coordinates* and 3*N conjugate momenta*, while Lagrangian equation is a single 3*N second-order* differential equation for the *coordinates only*. Hence, the choice of one formalism instead of another is primarily dictated by considerations of convenience. For instance, the Lagrangian formulation allows one to treat in a straightforward way holonomic constraints (namely dependent upon coordinates), such as in the continuous indirect version of metadynamics, as reported in chapter 2.3.1. Conversely, Hamiltonian formalism is helpful to handle certain systems, such as some aspects of quantum mechanics, in an easier way compared to other formulations of mechanics.

2.1 Quantum Mechanics: Basics Aspects of the Wave Function Theory (WFT)

The strict description of a system made of interacting particles can be handled by means of the notable Schrödinger equation, which is here reported in its non-relativistic, time-independent form and expressed in atomic units system (i.e.: in a compact form obtained omitting the fundamental physical constants, see Tab 2.1):

$$\hat{\mathrm{H}}\Psi_i(\mathbf{x}_{1,N},\mathbf{R}_{1,M}) = E_i\,\Psi_i(\mathbf{x}_{1,N},\mathbf{R}_{1,M}) \qquad [2.22]$$

where:

- $\Psi_i(\mathbf{x}_{1,N},\mathbf{R}_{1,M})$ stands for the wave function of the *i*-th state of the system, which is a function of the 3*N* spatial coordinates $\mathbf{r}_i$, and the *N* spin coordinates s_i (which are collectively termed $\mathbf{x}_i$), and the 3*M* spatial coordinates of the nuclei $\mathbf{R}_I$. The wave function contains all the information about the quantum system under investigation.
- E_i is the numerical value of the energy of the state described by Ψ_i.
- $\hat{\mathrm{H}}$ is the molecular Hamilton operator for a generic system consisting of *M* nuclei and *N* electrons in the absence of external magnetic or electric fields.

Basically, the Hamilton operator is the quantum mechanical analogue of the internal energy for the system, and in a mathematical formalism the Schrödinger equation is termed an *eigenequation*. Hence, the E_i allowed values of energy are the *eigenvalues* of the operator, whereas Ψ_i represent *eigenvectors*, or better *eigenfunctions* of the operator, and they both represent the solution for the eigenequation.

The total Hamilton operator can be written as a sum of the kinetic and potential energies of the nuclei and the electrons:

$$\hat{\mathrm{H}} = \left[-\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{A=1}^{M}\nabla_A^2\right] + \left[\sum_{i=1}^{N}\sum_{i>j}^{N}\frac{1}{\mathbf{r}_{ij}} + \sum_{A=1}^{M}\sum_{B>A}^{M}\frac{\mathrm{Z}_A\mathrm{Z}_B}{\mathbf{r}_{AB}} - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{\mathrm{Z}_A}{\mathbf{r}_{iA}}\right] \qquad [2.22]$$

where *A* and *B* run over the *M* nuclei, while *i* and *j* run over the *N* electrons of the system; the potential energy operator is the Coulomb potential, whereas the kinetic energy operator is accounted for by the Del-squared operator. The classic and non-relativistic kinetic energy for a particle can be written as:

$$\mathrm{K} = \frac{1}{2}m\mathbf{v}^2 = \frac{1}{2m}\mathbf{p}^2 \qquad [2.23]$$

Whereas in quantum mechanics the momentum is treated as an operator, and its formulation derives from the de Broglie's law $p = \hbar k$:

$$\hat{\mathbf{p}} \to -i\hbar \frac{\partial}{\partial \mathbf{r}} \qquad [2.24]$$

where *i* is the root square of -1. Such a formulation directly leads to the following expression for the kinetic energy operator:

$$\hat{\mathrm{K}} = \frac{1}{2}(-i\hbar\nabla)^2 \qquad [2.25]$$

From which is clear that the Del-squared operator is a second order differential operator defined as the divergence of the gradient for any given scalar or vector field (also called Laplace operator):

$$\Delta(\mathbf{r}) = \nabla(\mathbf{r}) \cdot \nabla(\mathbf{r}) = \nabla^2(\mathbf{r}) = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \qquad [2.26]$$

The total Hamilton operator can be rewritten in a more compact form as:

$$\hat{\mathrm{H}}_{\mathrm{tot}} = \hat{\mathrm{T}}_{\mathrm{e}} + \hat{\mathrm{T}}_{\mathrm{n}} + \hat{\mathrm{V}}_{\mathrm{ee}} + \hat{\mathrm{V}}_{\mathrm{nn}} + \hat{\mathrm{V}}_{\mathrm{ne}} \qquad [2.27]$$

and hence equation 2.22 can be rewritten as it follows:

$$\hat{\mathrm{H}}_{\mathrm{tot}} \Psi_{\mathrm{tot}}(\mathrm{x}, \mathrm{R}) = \mathrm{E}_{\mathrm{tot}} \Psi_{\mathrm{tot}}(\mathrm{x}, \mathrm{R}) \qquad [2.28]$$

The Schrödinger equation can be further simplified by taking advantage of the difference in masses between nuclei and electrons: as nuclei are much heavier than electrons, their velocities are much smaller, hence allowing one to neglect the kinetic energy term of the nuclei from the total Hamiltonian. This is the so called Born-Oppenheimer (BO) or adiabatic approximation, which states that the electronic wave function depends only on the position of the nuclei and not on their momenta, or – in other words – that the electronic wave function depends parametrically on the nuclear coordinates, and functionally on the $\mathbf{x}_{\mathrm{i}}$ spatial and spin coordinates of the *N* electrons of the system. The molecular Hamiltonian is then rewritten as:

$$\hat{\mathrm{H}}_{\mathrm{tot}} = \hat{\mathrm{H}}_{\mathrm{e}} + \hat{\mathrm{T}}_{\mathrm{n}} \qquad [2.29]$$

Where, the electronic Hamiltonian is defined to be:

$$\hat{\mathrm{H}}_{\mathrm{e}} = \hat{\mathrm{T}}_{\mathrm{e}} + \hat{\mathrm{V}}_{\mathrm{ee}} + \hat{\mathrm{V}}_{\mathrm{nn}} + \hat{\mathrm{V}}_{\mathrm{ne}} \qquad [2.30]$$

If nuclei do not have momenta, their kinetic energy is zero ($\hat{\mathrm{T}}_{\mathrm{n}} = 0$), while the potential energy due to nucleus-nucleus repulsion becomes a mere constant ($\hat{\mathrm{V}}_{\mathrm{nn}} = \mathrm{cost}$). Since electrons can be considered as moving in the field of fixed nuclei, the motion decoupling is handled by factorizing the total wave function:

$$\Psi_{\mathrm{tot}}(\mathrm{x}, \mathrm{R}) = \Psi_e(\mathrm{x}, \mathrm{R}) \cdot \Psi_n(\mathrm{R}) \qquad [2.31]$$

hence an electronic Schrödinger equation can be defined:

$$\mathrm{H}_e \Psi_e(\mathrm{x}, \mathrm{R}) = \mathrm{E}_e \Psi_e(\mathrm{x}, \mathrm{R}) \qquad [2.32]$$

where:

$$\mathrm{E_{tot}} = \mathrm{E_e} + \mathrm{E_n} \qquad [2.33]$$

Once the Born-Oppenheimer approximation holds, the problem is reduced to solve the electronic Schrödinger equation for a set of nuclear geometries, leading to the reconstruction of the Potential Energy (hyper-) Surface (PES). At the Hartree-Fock level of theory, the wave function is approximated as a single Slater determinant, and hence no electron correlation is taken into account, namely each electron feels the average field of the remaining electrons. In this framework, the electronic problem – electrons in the nuclear field – is tackled by means of the Self-Consistent Field approach (SCF) which is based upon the variational principle (which will not be discussed here), while the nuclear problem – nuclei in the electrons field – is solved using various minimization algorithms.

.

Tab. 2.1: **Atomic units**

Symbol (name)	**Quantity**	**Value in a.u.**	**Value in SI units**
m_e	Electron mass	1	$9.1094 \cdot 10^{-31}$ kg
e	Electron charge	1	$1.6022 \cdot 10^{-19}$ C
$\hbar$	$\frac{h}{2\pi}$, Atomic momentum unit	1	$1.0546 \cdot 10^{-34}$ Js
a_0 (Bohr radius)	$\frac{h^2}{4\pi^2 m_e e^2}$, Atomic distance unit	1	$5.2918 \cdot 10^{-11}$ m
E_h (Hartree)	$\frac{e^2}{a_0}$, Atomic energy unit	1	$4.3597 \cdot 10^{-18}$ J

2.2 The Potential Energy Function in Classical Mechanics: the Force Field

Unfortunately, most of the problems to deal with in molecular modeling of biological systems are too large to be solved by means of quantum mechanics, even if the computational power is constantly increasing. By tacking advantage of the BO approximation, we can consider the internal potential energy of a system in its ground state as a function of the solely nuclear coordinates. The electron contribution is not completely lost, but is implicitly taken in account by means of a parametric function of the nuclear coordinates. The parameters used are derived from experiments or from higher level of theory employing suitable fitting functions. Classical mechanics (or force field methods) are commonly used to perform calculations in systems containing a significant number of atoms, where the QM/MM average limiting threshold is nowadays about 100 nuclei (but it must be noted that in quantum mechanics the system size at a particular level of theory is more properly determined in terms of dimensions of the molecular orbital expansion). Obviously, classical mechanics cannot provide properties that explicitly depend upon the electronic distribution within a molecule. Moreover, as a consequence of the above reported considerations, molecules are associated to fixed topologies, that is no changes on bonded configurations are allowed. Such topologies must to be chosen a priori, and can not be changed during the course of simulation (for instance a protomeric form will not change its topology even if the environment would lead to a preferential stabilization of another tautomer, e. g. the ε- and δ- mono-protonated-neutral form of histidine).

As already stated, within the BO approximation, it is possible to express the Hamiltonian of a system as a function of the nuclear variables, as the electron motions have been averaged out. Since dependent upon parameters, the approach needs a full parameterization in each variable used for every kind of element's hybridization, that is the definition of the so called *atom types* table is demanded. The force field usually consists in a relatively simple expression of inter- and intra-molecular forces within the system of interest, which are modeled by means of bonded and non-bonded interactions, respectively. The basic functional form can be expressed as follows (by keeping in mind that more complex FF can include additional term, added for instance to explicitly model hydrogen bonding and so forth):

$$V(r^N) = \sum_{bonds} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{angles} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{torsions} \frac{V_n}{2}(1+\cos(n\omega-\gamma)) + \sum_{impropers} k_i(1-\cos 2\omega)$$

$$+ \sum_{vdW}\left(\varepsilon_{ij}\left[\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6}\right]\right) + \sum_{ele}\left(\frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}}\right) \qquad \text{[2.34]}$$

In the functional form of the force fields, two main groups of terms can be identified, since they attempt to model different kind of interactions:

BONDED INTERACTIONS: Interactions that depend upon connectivity. A harmonic function is used to quantify the energetic penalty associated with the deviation of bonds and angles away from their reference equilibrium values (denoted in equation 2.34 with the subscript 0), while a periodic function is used to describe the energy changes as dihedrals vary.

In general, the torsional energy in molecular mechanics is primarily used to correct the remaining energy terms. In other words, it represents the amount of energy that must be added or subtracted to the remaining terms to make the total energy agree with experiment or rigorous quantum mechanical calculation for a dihedral drive. It can be usually attributed either to electronic conjugation or hyper-conjugation effects.

NON-BONDED INTERACTIONS: Interactions that do not depend upon connectivity, and for this reason are modeled as a function of some inverse power of the distance.

Considering a system of N interacting particles, the potential energy can be divided into terms depending on the coordinates of individual atoms, pair, triplets, and so on:

$$\mathrm{V} = \sum_{i} \mathrm{v}_1(\mathbf{r}_i) + \sum_{i}\sum_{j>i} \mathrm{v}_2(\mathbf{r}_i, \mathbf{r}_j) + \sum_{i}\sum_{j>i}\sum_{k>j} \mathrm{v}_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \ldots \qquad [2.35]$$

where:

- the first term represent the effect of an external field on the system (usually it is not taken into account);
- the remaining terms represent the particle interactions, and – among them – the pair potential is the most important in magnitude.

To save computational cost, the non-bonded component of force fields is usually built up as a sum of *pair wise potentials* (two-body interactions), where interactions are in turns properly distinguished in two groups: electrostatic and van der Waals interactions.

The charge distribution is a continue three-dimensional function which arises from the local concentration or depletion of electronic density in a volume unit around the considered molecule. As any continue function, it is necessary to derive a discrete representation to be used in a numerical way. The easier and computationally fastest way to represent such a distribution is by means of a proper arrangement of fractional point charges, which are of course an abstraction since they are not an observable of the wave function. They are usually localized to nuclear centers, and thus they are often referred to as *partial atomic charges*. In this model the energy of the charge-charge interaction is then calculated by means of the Coulomb potential function.

Two main classes of charges derived from quantum mechanical calculations can be distinguished:

- **Mulliken charges** (often referred to as Coulson charges if the considered level of theory is semi-empirical)**:** they represent a set of charges based on the population analysis, namely a method used to artificially partition the electron density between nuclei so that each nucleus owns a (not necessarily integer) number of electrons. Given the wave function, this approach is rather trivial, but unfortunately it suffers from the fact that the charges derived are primarily dependent upon the way the atoms are bonded in the molecule, rather than to reproduce an inter-molecular electrostatic property.
- **ESP charges:** they represent a wide set of charges (Merz-Singh-Kollman, CHELP, CHELPG, RESP, to mention the most popular fitting schemes), which are explicitly derived to reproduce the electrostatic potential of the molecule. The electrostatic potential is an observable of the wave function, which can be expressed as follows:

  $$\varphi(\mathbf{r}) = \varphi_{\mathrm{N}}(\mathbf{r}) + \varphi_{\mathrm{e}}(\mathbf{r}) = \sum_{A=1}^{M} \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|} - \int \frac{\rho(\mathbf{r})}{|\mathbf{r} - \mathbf{r'}|} d\mathbf{r'} \qquad [2.36]$$

  A least-square fitting procedure is then used to derive the set of charges, which best reproduce the electrostatic potential: if the electrostatic potential at a point is φ_i^0 and if the value from the charge model is φ_i^{calc}, then the following deviation has to be minimized for each *i* grid nodes:

  $$\mathrm{d} = \sum_{i=1}^{N} \omega_i \left(\varphi_i^0 - \varphi_i^{calc}\right)^2 \qquad [2.37]$$

  where ω_i is a weighting function.

  Actually, the fitting schemes proposed are usually quite able to reproduce the main electrostatic properties of most of the molecules.

Within the multipolar expansion framework, an arbitrary charge distribution is represented as a decomposition in various electric moments. The most important component in determining the electrostatic potential is the first non-null electric moment, and usually – at least for systems of biological interest – the expansion is truncated at the quadrupole (namely to the third order):

- Monopole: it represents the net charge of the molecule, it is a scalar quantity, and it is expressed in Coulomb;
- Dipole: is the first derivative of the energy with respect to an external applied field. It represents a vector quantity, hence it is univocally defined by 3 elements. Nevertheless, if a proper reference system parallel to the line connecting the geometric centre of the charges having opposite sign is chosen, just the component in such a direction is needed to be

specified, being null the components along the remaining directions. For a discrete distribution of charges *i*, the dipole moment is defined as:

$$\boldsymbol{\mu} = \sum_i q_i \mathbf{r}_i = \begin{vmatrix} \sum_i q_i x_i \\ \sum_i q_i y_i \\ \sum_i q_i z_i \end{vmatrix} \qquad [2.38]$$

The dipole moment can be considered as a measure of the asimmetry in the charge distribution, it is measured in Debye ($\mathrm{C \cdot m}$).

- Quadrupole: is the second derivative of the energy with respect to an external applied field. It represents a rank 2 tensor quantity, and hence is univocally defined by 9 elements (3x3 matrix), even if – for symmetry reasons – independent elements are reduced to 6. For a discrete distribution of charges *i*, the traceless quadrupole moment (hence the matrix where the sum of the elements belonging to the principal diagonal is zero) is rigorously defined as:

$$\begin{aligned} \boldsymbol{\Theta} &= \frac{1}{2}\sum_{\alpha\beta} q_i \left(3\mathbf{r}_\alpha \mathbf{r}_\beta - r^2 \mathbf{1}\right) \\ &= \begin{vmatrix} \Theta_{xx} & \Theta_{xy} & \Theta_{xz} \\ \Theta_{yx} & \Theta_{yy} & \Theta_{yz} \\ \Theta_{zx} & \Theta_{zy} & \Theta_{xx} \end{vmatrix} \\ &= \frac{1}{2}\begin{vmatrix} \sum_i q_i \left(3x_i^2 - r_i^2\right) & 3\sum_i q_i x_i y_i & 3\sum_i q_i x_i z_i \\ 3\sum_i q_i y_i x_i & \sum_i q_i \left(3y_i^2 - r_i^2\right) & 3\sum_i q_i y_i z_i \\ 3\sum_i q_i z_i x_i & 3\sum_i q_i z_i y_i & \sum_i q_i \left(3z_i^2 - r_i^2\right) \end{vmatrix} \end{aligned} \qquad [2.39]$$

where $\mathbf{1}$ is the identity matrix, which corresponds to the Kronecker delta in a suffix formulation. Analogously to the dipole moment, it always exists a reference system (called principal axes) for which the tensor is diagonal, and is roughly determined by the symmetry of the charge distribution. The quadrupole moment can be thought as a measure of the deviation from the spherical geometry of the charge distribution, and it is expressed in $\mathrm{C \cdot m^2}$. From the formula definition it should be clear that for a spherical distribution of charges (but even for a tetrahedral one, such as methane, or octahedral, and so on), elements belonging to the principal diagonal of the tensor are null, whereas the presence of an element significantly greater than the others would imply an elongation of the spherical distribution along the considered axis. For a 2 fold symmetrical charge distribution, such as for the benzene molecule, it results:

$$\begin{vmatrix} \Theta_{xx} & 0 & 0 \\ 0 & \Theta_{yy} = \Theta_{xx} & 0 \\ 0 & 0 & \Theta_{zz} = -\left(\Theta_{xx} + \Theta_{yy}\right) \end{vmatrix} \qquad [2.40]$$

Atom-centered partial charges are straightforwardly implemented in force fields, but suffer from the strong drawback to be, trivially, localized at atomic positions. The problem becomes evident in some kind of molecules, such as benzene, where the interesting charge distribution is projected onto the π axis of the molecule, namely along an atom free axis. Hence, by using partial charges lying on the σ plane, no quadrupole moment of the benzene would be reproduced. This is particularly relevant when modelling in a classical way cation-π and π-π interactions, which are primarily driven by the electrostatic component arising form the π charge distribution of the aromatic fragment. Actually, some qualitative cation-π and π-π interactions can be reproduced by means of the six dipoles arising from the C-H bonds, hence the importance of an all-atoms (or at least extended-atoms) force field.

More generally the main drawback of the pair wise potential representation of molecular electrostatic properties lies in the fact that fractional charges are of course *localized*, hence *static*. This means that:

- changes in charge distribution as a consequence of changes in conformation are not allowed (actually RESP charges are quite conformation-independent since atom equivalences are taken in account in the fitting function; moreover in principle they should be derived from a hyper-electrostatic potential calculated on the lowest energy conformations, and assigning a different fitting weight based upon their relative Boltzmann population at room temperature);
- polarization effects (changes in charge distribution in response to an external field) are by definition neglected.

Within the so called van der Waals potential energy term, interactions which determine deviation from the ideal gas behavior are accounted for, and in particular two contribution are involved. The mid-range attractive interactions are represented by the dispersive (or London’s) forces, which are due to the phenomenon known as *electron correlation*: in a pictorial view the favorable interactions arise as a consequence of a reciprocal induced temporary polarization in electronic charge distribution. Conversely, at shortest-range distances, the repulsive contribute is dominant, since the electron densities belonging to each interacting particle are about to overlap according to the Pauli’s principle. Therefore this kind of interactions are often referred to as *exchange forces*. Of course at even shorter distances, the repulsion between unshielded nuclei also takes place. The van der Waals interactions are usually taken into account by the 6-12 Lennard-Jones energy function, where the couple of numbers refers to the exponential dependence upon distance of the attractive and repulsive terms, respectively:

$$\mathrm{V}(r) = \varepsilon_{ij}\left[\left(\frac{r_{ij}^{0}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^{0}}{r_{ij}}\right)^{6}\right] \qquad [2.41]$$

where r_{ij}^{0} is the equilibrium distance between the particles, while ε_{ij} stands for the well depth, which defines the energy at the minimum of the function, namely the equilibrium distance (van der Waals radius). The collision diameter can be hence defined as: $\sigma_{ij} = \frac{r_{ij}^{0}}{2^{1/6}}$.

2.3 Statistical mechanics: the phase space sampling

The simulation of a microscopic system at a given temperature greater than zero Kelvin provides a set of configurations, each of them represents a distinct point in the so called *phase space*. Since the instantaneous mechanical state of such a system is usually specified in terms of positions and momenta of its constituting particles (i.e. the Hamiltonian mechanics formalism), a system containing *N* atoms will define a 6*N* hyper-dimensional configurational space (3*N* positions and 3*N* momenta), and the way the system moves through this phase space is determined by the mechanical laws at the given level of theory. If it was possible to visit all the points in phase space by means of a simulation, then the partition function (were ε_i is the energy of the i^{th} configuration) – defined as:

$$q = \sum_i e^{\left(\frac{-\varepsilon_i}{k_B T}\right)} \qquad [2.42]$$

– could be exactly and directly calculated (and hence the thermodynamic properties too). In practice, and in particular for systems of biological interest, the phase space is enormous, and the complete visitation is not achievable in a reasonable amount of time (the trajectory in phase space is hence said to be *non ergodic*). This also means that two simulations performed on the same system starting from two different point in phase space would give similar, but *not equal*, results.

Statistical mechanics can be seen as a useful tool for bridging information arising from simulations performed at a microscopic scale (atomic position, velocities) into macroscopic terms (pressure, internal energies, …) which are needed both to validate simulations and to predict structural or thermodinamical data (see Fig 2.1).

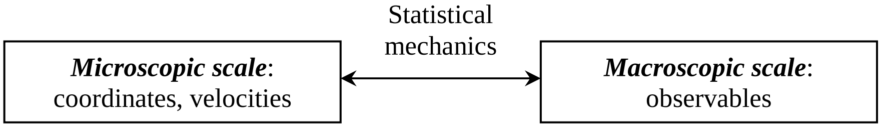

Statistical mechanics
Microscopic scale: coordinates, velocities
Macroscopic scale: observables

Fig. 2.1: Pictorial representation of the theoretical link provided by statistical mechanics.

For a particular configuration $\mathbf{\Gamma}$ in phase space, the instantaneous value of some property A (i.e. potential energy), can be written as a function $\mathrm{A}(\mathbf{\Gamma}(t))$, and hence:

$$\mathrm{A}_{obs} = \langle \mathrm{A} \rangle_{time} = \langle \mathrm{A}(\mathbf{\Gamma}(t)) \rangle = \lim_{t_{obs}\to\infty} \frac{1}{t_{obs}} \int_0^{t_{obs}} \mathrm{A}(\mathbf{\Gamma}(t))\,\mathrm{d}t \qquad [2.43]$$

Since we are dealing with computational techniques, and hence by performing discrete steps along the phase space, equation 2.43 should be rewritten as:

$$\mathrm{A}_{obs} = \langle \mathrm{A} \rangle_{time} = \langle \mathrm{A}(\mathbf{\Gamma}(t)) \rangle = \frac{1}{t_{obs}} \sum_{\tau=1}^{\tau_{obs}} \mathrm{A}(\mathbf{\Gamma}(t)) \qquad [2.44]$$

where τ identifies the large but finite number of time steps.

Because of the complexity of the time evolution of $\mathrm{A}(\mathbf{\Gamma}(t))$, Boltzmann and Gibbs suggested to replace the *time average* by the *ensemble average* concept. Within such a description, the usual single system evolving in time is thus replaced by a number of replicas that are simultaneously considered. In this context, the ensemble is defined as a *collection of points* $\mathbf{\Gamma}$ *in phase space*, which are distributed according to a probability density $\rho(\mathbf{\Gamma})$. The functional form of ρ is determined by the chosen fixed macroscopic variables, which in turn define the statistical ensemble itself: NVE, NVT or NPT, among the others (see table 2.2). Since $\rho_{ens}(\mathbf{\Gamma})$ represents an equilibrium ensemble, its time-dependence vanishes, and hence it follows that:

$$\mathrm{A}_{obs} = \langle \mathrm{A} \rangle_{ens} = \sum_{\mathbf{\Gamma}} \mathrm{A}(\mathbf{\Gamma})\rho(\mathbf{\Gamma}) \qquad [2.45]$$

In other words, in accordance with the ergodic hypothesis of the phase space, we assume that the time averages obtained from computations are equivalent to the ensemble averages. Finally, the ensemble average of the investigated property A is then calculated by integrating over all the obtained configuration of the system.

In table 2.2, the main four statistical ensembles used in computer simulations are reported.

Tab. 2.2: Definition of the main statistical ensembles used in computer simulations.

Statistical ensemble	Fixed thermodynamic variables	Equilibrium state
Microcanonic	NVE	Maximum: entropy (S)
Canonic	NVT	Minimum: Helmoltz free-energy (A)
Isotherm-Isobar	NPT	Minimum: Gibbs free-energy (G)
Grand canonic	μPT	Maximum: pressure • volume (PV)

Two main classes of sampling methods of the phase space are used: molecular dynamics (MD) and Monte Carlo (MC), by keeping in mind that many hybrid approaches between them have been described and (more or less) successfully applied. Here, we will focus on the former, nevertheless the basic differences between the two classes will be briefly summarized:

- MD is a *deterministic* method, where the dynamic behavior of the system is investigated evolving the Newton's second law equation in a proper statistical ensemble. The result is essentially a *trajectory*, where all discrete steps are correlated in time, and from which the thermodynamic averages of the ensemble can be directly calculated by means of numerical integration;

- MC is a *stochastic* method, where the configurations of the replica are randomly generated (hence the name), and their acceptance usually follows the Boltzmann distribution at the given temperature (the algorithm developed by Metropolis). The result is a *Markov chain*, that is a collection of configurations in which each of them depends only upon the previous state, and not upon any other state previously visited. In contrast of MD, the kinetic component of the total energy is missing, nevertheless by means of theoretical tricks the same information can be obtained.

The choice between the two methods of course depends upon the kind of problem, even if it must be stressed that for biological systems, MD have become much more popular than MC methods.

Because of the way in which the sampling is performed and directed along the phase space, both MD and MC techniques are known to be *equilibrium* methods, since they are intrinsically (MD) or naturally (MC) driven towards a Boltzmann distribution. Apart from the chosen method, providing a suitable formulation of the probability density, and hence of the partition function associated to the statistical ensemble of interest, the basic thermodynamic properties can be calculated as averages. The whole properties (thermodynamics and functions of average coordinates) that can be calculated from a simulation are here summarized:

1. ***structural properties***: such as atomic distribution functions,…
2. ***dynamical properties***: such as diffusion coefficient, auto-correlation functions,..
3. ***thermodynamic properties***: which can be further distinguished in:

- <u>mechanical properties</u> (internal energy, pressure, heat capacity), which are related to the *derivative* of the partition function;
- <u>thermal or entropic properties</u> (entropy, chemical potential, free-energy), which are related to the partition function *itself*;

Among the thermodynamic properties, the accuracy of the calculation of the mechanical ones is much better than that of the thermal ones. This could be demonstrated in a rigorous mathematical way by considering the dependence of the property under investigation upon the partition function, but it is also straightforwardly intuitive. As the trajectory (or the Markov chain) is not ergodic, these methods will preferentially sample the *low-energy* regions of the phase space, namely the most thermally populated configuration of the system. Whereas, in order to properly take into account entropic properties, (such as free-energy) also the *high-energy* states would have to be significantly visited.

The problem of the free-energy estimation. Denoting as X^N a configuration of the system in the phase space, the configurational partition function is defined as:

$$\mathrm{q}_N = \int \ldots \int e^{\left[-\beta \mathrm{E}\left(X^N\right)\right]} dX^N \qquad [2.46]$$

where $\mathrm{E}\left(X^N\right)$ is the energy of the N^{th} configuration, $\beta = \frac{1}{k_B T}$, whereas the integral is extended all over the space of the *N* configurations accessible to the system. The partition function provides an estimate of the number of states which are thermally accessible to the system at the given temperature. If $T = 0K$, $\beta \rightarrow \infty$, and hence $\mathrm{q}_N \rightarrow g_0$, namely the fundamental state is the only populated configuration of the system; whereas if $T = \infty K$, $\beta \rightarrow 0$, and hence $\mathrm{q}_N \rightarrow \infty$, which means that an infinite number of states will be populated. The probability to find the system in a particular configuration X^N at a given temperature is provided by the probability density function:

$$\rho\left(X^N\right) = \frac{e^{\left[-\beta \mathrm{E}\left(X^N\right)\right]}}{\mathrm{q}_N} \qquad [2.47]$$

From the probability density function the various thermodynamic properties can be calculated, for instance the value of the average internal energy will be:

$$\mathrm{U} = \int \ldots \int \mathrm{E}\left(X^N\right) \rho\left(X^N\right) dX^N = \left\langle \mathrm{E}\left(X^N\right) \right\rangle \qquad [2.48]$$

The free-energy is maybe the most important thermodynamic property of a system, and it is usually expressed as the Helmoltz function A which is appropriate for constant NVT statistical ensembles, or as the Gibbs function G which conversely is appropriate for constant NPT statistical ensembles. For instance, the Helmoltz free-energy is defined as:

$$\mathrm{A} = -k_B T \ln(\mathrm{Q}) \qquad [2.49]$$

where for distinguishable particles $\mathrm{Q} = \mathrm{q}_N$, whereas for indistinguishable particles $\mathrm{Q} = \frac{1}{N!}\mathrm{q}_N$, and it can be demonstrated that is equal to:

$$\mathrm{A} = k_B T \ln\left(\int \ldots \int e^{\left(\beta \mathrm{E}\left(X^N\right)\right)} \rho\left(X^N\right) dX^N\right) \qquad [2.50]$$

Since we are dealing with equilibrium methods, the sampling is intrinsically designed to be directed towards a Boltzmann-like distribution, namely it is naturally weighted to favor thermally populated states of the system. Hence, the estimation of some properties such as the free-energy by means of equilibrium methods is a difficult task, since it would require more substantial sampling over higer-energy configurations. An ergodic trajectory would of course visit all of the high-energy regions of the phase space as well, but in practice they will never be adequately sampled. Hence, many methods have been until now developed in order to encourage the system to explore regions of the phase space normally associated with a low-frequency of sampling.

In order to be quite clear, two main families of methods for free-energy estimation can be distinguished: i) *potentials of mean force*, which allow the study of the free-energy changes in respect to some inter- or intra-molecular coordinate and which are usually (but not only) implemented in MD codes, and ii) methods which allow the calculation of the free-energy difference between two states based on the fact that free-energy is a state function (thermodynamic perturbation methods, thermodynamic integration methods), which is a closely related albeit slightly different problem in respect to that discussed above. The latter family of methods is usually (but not only) implemented in MC code. Since we are focused on molecular dynamics sampling, here only the most important features of potentials of mean forces will be summarized.

The word potential of mean force is referred to the free-energy surface along a chosen coordinate. Various methods have been proposed for calculating potentials of mean force, but the most popular is the so called Umbrella Sampling. This method attempts to overcome the sampling problem by modifying the potential function so that also the unfavorable states are sufficiently sampled, and it can be implemented either in MD or in MC. The modification of the potential energy function can be written as the following perturbation:

$$\mathrm{V}'\left(\mathbf{r}^N\right) = \mathrm{V}\left(\mathbf{r}^N\right) + \mathrm{W}\left(\mathbf{r}^N\right) \qquad [2.51]$$

where $\mathrm{W}\left(\mathbf{r}^N\right)$ is the weighting function that usually takes a quadratic form:

$$\mathrm{W}\left(\mathbf{r}^N\right) = k_{\mathrm{W}}\left(\mathbf{r}^N - \mathbf{r}_o^N\right)^2 \qquad [2.52]$$

In such a way the sampling will be biased along some relevant *collective coordinate* (not necessary an intra- or inter-atomic coordinate) resulting on a non-Boltzmann distribution. The corresponding Boltzmann averages can be converted from the non-Boltzmann distribution through the equation:

$$\left\langle \mathrm{A} \right\rangle = \frac{\left\langle \mathrm{A}\left(\mathbf{r}^N\right) e^{\beta \mathrm{W}\left(\mathbf{r}^N\right)} \right\rangle_{\mathrm{W}}}{\left\langle e^{\beta \mathrm{W}\left(\mathbf{r}^N\right)} \right\rangle_{\mathrm{W}}} \qquad [2.53]$$

Usually, an Umbrella Sampling calculation is performed by splitting the run in a series of stages, each of them is characterized by a particular value for both $\mathrm{W}\left(\mathbf{r}^N\right)$ and $\mathbf{r}^N$. However, if the forcing potential is too large, the distribution is dominated by only few configurations namely the opposite problem seen in MD occurs: the non-Boltzmann distribution will be over-sampled, and the averages will take too long to converge.

The recently developed non-equilibrium sampling method, metadynamics, will be exhaustively discussed in chapter 2.3.2.

2.3.1 Classical molecular dynamics

Having already introduced the Hamiltonian formulation of mechanics in chapter 2.0 (equation 2.21), it is now possible to express the energy conservation principle for a system as a sum of kinetic and potential energy terms:

$$\mathrm{H}(\mathbf{q},\mathbf{p}) = \mathrm{T}(\mathbf{p}) + \mathrm{V}(\mathbf{q}) \qquad [2.54]$$

For Cartesian coordinates, equations of motion which govern the time-evolution of the system and all its properties become:

$$\dot{\mathbf{r}}_i = \frac{\mathbf{p}_i}{m} \qquad [2.55]$$

$$\dot{\mathbf{p}}_i = -\nabla_{r_i} \mathrm{V} = \mathbf{f}_i \qquad [2.56]$$

Where the potential V is accounted for by means of the potential energy function provided by the force field (equation 2.34).

For conservative systems (invariant potential function in time), the force acting on the *i*th particle is a function of the coordinates only. Since the potential energy function (which is independent of velocities and time) required for the force calculation is provided by the force field (chapter 2.2), initial velocities are solely required in order to evolve the system (as starting coordinates are obviously known).

It should be noticed that once the Hamiltonian is defined as above, we are intrinsically dealing with the microcanonical (NVE) ensemble. In such a framework, the system of interest moves in the phase space along a constant-energy hyper-surface.

As usual, as a consequence of a continuous potential, the motion of all particles is tightly correlated, giving rise to a many-body problem which can not be analytically solved. To overcome this, equations of motions are integrated using a finite difference method, where integration is performed on discrete time intervals δt. By doing so, two fundamental assumptions for classical MD are introduced:

1. forces are constant during each time step, and consequently:
2. collisions are elastics.

Basically, an MD code could be seen as follows:

- Starting velocities are initialized by random selecting from a Maxwell-Boltzmann distribution at the temperature of interest:

$$\mathrm{f}(v_{ix}) = \left(\sqrt{\frac{m_i}{2\pi k_B T}}\right) e^{\left(-\frac{m_i v_{ix}^2}{2k_B T}\right)} \qquad [2.57]$$

Since it represents a Gaussian distribution, it can be easily obtained from a random number generator.

- Since we are dealing with the NVE statistical ensemble, velocities are often rescaled so that the total momentum (linear and rotational) of the system is zero.

The discrete time course takes place in the next phase, namely for each integration time-step:

- The forces at time *t* are calculated by differentiating the potential energy function:

$$\mathbf{f}_{i,t} = -\frac{\partial \mathrm{V}(\mathbf{r}_{i,t})}{\partial \mathbf{r}_{i,t}} \qquad [2.58]$$

  The force on an atom may include contributions from the various terms in the force field, and represents the most time consuming part for a molecular dynamics code.

- The equations of motions are then integrated by means of a suitable algorithm. There are a lot (but not so many) algorithms to use in MD, each of them assumes that the positions and dynamic properties (velocities, accelerations, and so on) can be approximated by a Taylor series expansion. The Verlet algorithm, which is the simplest, represents a third-order truncation, and uses positions and accelerations (previously computed) at time t, and the positions from the previous step $(t-\delta t)$, to calculate the positions at time $(t+\delta t)$:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \dot{\mathbf{r}}(t)\delta t + \frac{1}{2}\ddot{\mathbf{r}}(t)\delta t + \frac{1}{3!}\dddot{\mathbf{r}}(t)\delta t \qquad [2.59]$$

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \dot{\mathbf{r}}(t)\delta t + \frac{1}{2}\ddot{\mathbf{r}}(t)\delta - \frac{1}{3!}\dddot{\mathbf{r}}(t)\delta t \qquad [2.60]$$

  where accelerations at the given time t, are obviously calculated by the forces at the same time step:

$$\ddot{\mathbf{r}}_{i,t} = \frac{\mathbf{f}_{i,t}}{m_i} \qquad [2.61]$$

  Adding together equations 2.59 and 2.60, gives:

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \ddot{\mathbf{r}}(t)\delta t \qquad [2.62]$$

  As it can be seen the Verlet algorithm is a third-order algorithm even though third derivatives do not appear in the above equation, since they have been cancelled out. Moreover, velocities do not explicitly appear as well, but they can be obviously obtained in a number of ways.

In general, MD integrators have to satisfy some requirements, such as:

1. to be time-reversible (even using an infinite numerical precision, algorithms would be time-reversible only in the limit of infinitively short time-steps);

2. to satisfying the conservation laws of energy and momentum;
3. to be accurate, usually third-order algorithms are used since second-order algorithm have a poor accuracy, while fourth-order are simply too expensive;
4. to be fast, namely to require just one force evaluation per time-step;
5. to be stable, namely to ensure a little amount of propagation errors. Usually stability decreases as the time-step increases.

Both accuracy and stability are quantified in terms of the divergence between the numerical and the analytical trajectory.

Choice of the time-step. The choice of the size of the time-step is primarily determined by a compromise between accuracy and speed of the calculation. The smaller the time-step, the less the numerical trajectory would diverge from the analytical one, while at the same time, the rate of phase space sampling decreases. Conversely, by using a large time step would lead to instability as a consequence of the increased probability for the atoms to cross their minimum energy separations during an integration step, and thus leading to an unrealistic gain in potential energy.

For these reasons, the chosen time-step should be small when compared to the mean time between collisions. As a safe rule of thumb, the time-step should be approximately one tenth the time of the shortest period of motion. Considering that the C-H bond stretching vibrates with a mean repeat period of 10 fs, the integration time step should not be greater than 1 fs:

$$\delta t = \frac{1}{10}\omega_{fastest} \qquad [2.63]$$

Because bond vibrations are usually of less interest than lower frequency modes such as torsions which correspond to major conformational changes, it would be advisable to increase the time-step constraining bond length involving hydrogen atoms. The most commonly used method for applying constraints in MD is the SHAKE procedure, and time-steps up to 2 fs are usually reached (or even more).

Boundaries. In MD simulations, the calculation of the Hamiltonian is limited to a microscopic, finite-size system, enclosed in the simulation cell. The link between the microscopic properties and the macroscopic ones is provided by statistical mechanics, as it has been summarized in the previous chapter. When dealing with biologically relevant systems, nowadays cubic or cuboid boxes are usually used as boundary geometry for the simulation cell.

The correct treatment of boundaries and boundary effects is crucial for a proper derivation of macroscopic properties. Actually, handling with a box creates six unwanted surfaces where hitting particles would reflect back into the interior of the cell, thus introducing artifacts that are relatively more important as the system sizes decrease. To reduce margin effects periodic boundary conditions are imposed, thus the unit cell is replicated in each dimension an infinite number of times. From a mathematical point of view this is defined as follows. In a cubic box of L length per side, for any observable A and for any integer n:

$$\mathrm{A}(\mathbf{r}) = \mathrm{A}(\mathbf{r} + nL) \qquad [2.64]$$

The computational implementation is that if a particle crosses a surface of the cell, it re-enters through the opposite wall with unchanged velocity. In such a way, surfaces are eliminated, simulating a quasi-infinite volume that would more closely represent the macroscopic system.

In this scheme, the potential energy due to the interaction of each non-bonded particles, would be expressed as:

$$V(\mathbf{r}_{ij}) = \sum_{i \neq j} V(\mathbf{r}_{ij}) + \sum_{n} \sum_{i \neq j} V(\left|\mathbf{r}_i - \mathbf{r}_j + nL\right|) \qquad [2.65]$$

where in the second term the summation runs over the n replicas of the system. In order to avoid an infinite summation, the *minimum image convention* is introduced, from which the pair-wise potential is truncated as:

$$\mathbf{r}_{ij} = \min\left\{\left|\mathbf{r}_i - \mathbf{r}_j + nL\right|\right\} \qquad [2.66]$$

Using this procedure, each particle interacts only with each of the n-1 other particles in the basic cell or in their nearest images. Actually, this is equivalent to set a cut-off on the potential energy function of the magnitude:

$$\mathbf{r}_{ij} \leq \frac{L}{2} \qquad [2.67]$$

In order to minimize finite-size effects, the value of L should be chosen large enough so that forces that would occur for distances greater than L/2 are negligibly small.

Nevertheless, elegant and rather complicated methods have been developed to treat the long-range electrostatics beyond the limits stated above (Particle Mesh Ewald, PME methods), and they will not be discussed here.

Actually, even periodic boundary conditions are not free from artifacts, in particular no fluctuations that have a wavelength greater then the length of the cell are obviously allowed. Such an order effect is particularly important when performing MD simulations of membrane systems.

Preservation of the conservation laws. Since the Hamiltonian is invariant either upon system translation and rotation, then the corresponding momentum should be conserved. Actually, dealing with a system confined in a cubic box none of these quantities would be conserved. Periodic boundary conditions allow preservation of translational invariance, hence total linear momentum is conserved, while the angular momentum is definitely not constant.

Simple thermodynamic averages: The kinetic, potential, and total internal energies may be easily calculated from the *phase equation*:

$$\mathrm{E} = \langle \mathrm{H} \rangle = \langle \mathrm{T} \rangle + \langle \mathrm{V} \rangle \qquad [2.68]$$

where brackets obviously denote ensemble averages.

Both temperature and pressure – which are observables in the microcanonical only, and in the microcanonical and canonical ensembles, respectively – may also be calculated as ensemble averages:

- The temperature is directly related to the kinetic energy by means of the equipartition theorem, which states that the internal energy of a system at thermal equilibrium will distribute among the quadratic degrees of freedom allowed to the particles of the system itself:

  $$\frac{1}{2} m v_i^2 = \frac{1}{2} k_B T \qquad [2.69]$$

  Since the system has three degrees of freedom per particle (in the absence of constraints), it follows:

  $$\mathrm{K} = \sum_{i=1}^{N} \frac{|p_i|^2}{2m_i} = 3\frac{k_B T}{2} N \qquad [2.70]$$

  where each degree of freedom contributes with a factor of $\frac{k_B T}{2}$.

- Conversely, pressure is calculated from the virial theorem of Clausius which concerns the connection between kinetic and potential energy ("virial" derives from the Latin word *vires* which means "forces"). By deriving the total kinetic energy of a system with respect to velocity ($\frac{\partial \mathrm{K}}{\partial v} = mv$), it can be obtained:

  $$\mathbf{v}\frac{\partial \mathrm{K}}{\partial v} = mv^2 = 2\mathrm{K} \qquad [2.71]$$

  Because of:

  $$\frac{d(\mathbf{p}\,\mathbf{r})}{dt} = \mathbf{p}\,\mathbf{v} + \mathbf{r}\,\dot{\mathbf{p}} \qquad [2.72]$$

  it can be easily derived:

  $$mv^2 = \mathbf{p}\,\mathbf{v} = \frac{d(\mathbf{p}\,\mathbf{r})}{dt} - \mathbf{r}\,\dot{\mathbf{p}} \qquad [2.73]$$

and thus:

$$2\mathrm{K} = \frac{d(\mathbf{p}\,\mathbf{r})}{dt} - \mathbf{r}\,\dot{\mathbf{p}} \qquad [2.74]$$

From the Eulero's formula, it can be demonstrated that in a time average:

$$\left\langle \frac{d(\mathbf{p}\,\mathbf{r})}{dt} \right\rangle = 0 \qquad [2.75]$$

hence, it follows:

$$\begin{aligned} \langle 2K \rangle &= \left\langle \frac{d(\mathbf{p}\,\mathbf{r})}{dt} \right\rangle - \langle \mathbf{r}\,\dot{\mathbf{p}} \rangle \\ &= \langle \mathbf{r}\,\dot{\mathbf{p}} \rangle \\ &= \langle \mathbf{r}\,\mathbf{F}^{tot} \rangle \end{aligned} \qquad [2.76]$$

where $\mathbf{F}^{tot}$ stands for the total force acting on the system, namely the vector sum of internal and external forces. Since the first derivative of the energy with respect to the displacement of a particle equals the force with changed sign, the virial theorem could be expressed as follows:

$$\langle 2\mathrm{K} \rangle = -\langle \mathrm{U} \rangle \qquad [2.77]$$

Hence, for a N particle system, equation 2.77 can be rewritten in an explicit way:

$$\left\langle 2\mathrm{K}^{tot} \right\rangle = \left\langle \sum_i^N 2\mathrm{K}_i \right\rangle = -\left(\left\langle \sum_i^N \mathbf{r}_i\,\mathbf{F}_i^{ext} \right\rangle + \left\langle \sum_i^N \mathbf{r}_i\,\mathbf{F}_i^{\mathrm{int}} \right\rangle \right) \qquad [2.78]$$

Since from the kinetic theory it results:

$$\mathrm{K}^{tot} = \frac{1}{2} Nm\left\langle v^2 \right\rangle = \frac{3}{2} Nk_B T \qquad [2.79]$$

equation 2.78 can be rewritten as:

$$\left\langle 2\mathrm{K}^{tot} \right\rangle = 3Nk_B T = \left\langle \sum_i^N 2\mathrm{K}_i \right\rangle = -\left(\left\langle \sum_i^N \mathbf{r}_i\,\mathbf{F}_i^{ext} \right\rangle + \left\langle \sum_i^N \mathbf{r}_i\,\mathbf{F}_i^{\mathrm{int}} \right\rangle \right) \qquad [2.80]$$

Again, from the kinetic theory, for an ideal gas the only forces are those arising from interactions between the gas and the container:

$$\mathrm{K}^{tot} = \frac{1}{2} Nm\left\langle v^2 \right\rangle = \frac{3}{2} Nk_B T = \frac{3}{2} PV \qquad [2.81]$$

and hence:

$$2\mathrm{K}^{tot} = 3PV \qquad [2.82]$$

For a system of particles, in the case of null internal forces:

$$\left\langle 2\mathrm{K}^{tot} \right\rangle = 3PV = -\left\langle \sum_{i}^{N} \mathbf{r}_i \, \mathbf{F}_i^{ext} \right\rangle \qquad [2.83]$$

Conversely, forces between particles in a real gas or in a liquid do affect the virial, and hence the pressure. In the light of these considerations, equation 2.80 will be rewritten as:

$$3Nk_BT = 3PV - \left\langle \sum_{i}^{N} \mathbf{r}_i \, \mathbf{F}_i^{int} \right\rangle \qquad [2.84]$$

which trivially equals to:

$$PV = Nk_BT + \frac{1}{3}\left\langle \sum_{i}^{N} \mathbf{r}_i \, \mathbf{F}_i^{int} \right\rangle \qquad [2.85]$$

Equation 2.85 represents an important result since it allows the calculation of the pressure for a system once the temperature and the internal forces are known. The second term in the right is often referred to as the "internal virial", W:

$$PV = Nk_BT + \left\langle \mathrm{W} \right\rangle \qquad [2.86]$$

Now, considering the shape of the simulation cell to be a cube or a cuboid of areas A_α that are perpendicular to the Cartesian axes α = *x, y, z,* the resulting forces F_β in the direction of the Cartesian axes β = *x, y, z* acting on each boundary area A_α, yield the elements $p_{\beta\alpha}$ of the pressure tensor:

$$p_{\beta\alpha} = \frac{\mathrm{F}_\beta}{A_\alpha} \qquad [2.87]$$

Alternatively, pressure can be calculated basing on thermodynamic relations. In the canonical ensemble, denoting as A the Helmoltz free-energy, the pressure results:

$$P = -\left(\frac{\partial \mathrm{A}}{\partial V}\right)_T \qquad [2.88]$$

Generally, the change in free-energy dA, which corresponds to the infinitesimal work δW performed at constant temperature, represents the result of the stress:

$$dV = A_\alpha d\beta \qquad [2.89]$$

Thus, the link between the thermodynamic and the mechanical definition of pressure can be expressed as:

$$p_{\beta\alpha} = -\left(\frac{\partial \mathrm{A}}{\partial V}\right)_T = -\frac{1}{A_\alpha}\left(\frac{\partial \mathrm{W}}{\partial \beta}\right)_T = \frac{\mathrm{F}_\beta}{A_\alpha} \qquad [2.90]$$

The usual scalar value for the pressure P can be calculated basing on the virial theorem as it has been previously demonstrated. In particular, *virial equation provides the pressure tensor*

$p_{\beta\alpha}$ in terms of the equilibrium ensemble average of the microscopic stress tensor $\sigma_{\beta\alpha}$. As a reminder, the rank-2 stress tensor is usually defined as:

$$dF_{\alpha} = \sum_{\alpha,\beta}^{3} \sigma_{\beta\alpha} dA_{\alpha} \qquad [2.91]$$

where the diagonal elements (σ_{xx}, σ_{yy}, σ_{zz}) represent the so called *normal stress*, whereas extra-diagonal elements represent the *shear stress*, and they are usually denoted as $\tau_{\beta\alpha}$. For an isotropic system, the shear stress elements are supposed to slightly fluctuate in time around the average value of zero. Providing a link among different formalism, pressure may be calculated from the trace of the stress tensor:

$$\begin{aligned} P &= \frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3} \\ &= \frac{1}{3V}\left(\left\langle \sum_{i}^{N} m_i \mathbf{v}_i \cdot \mathbf{v}_i \right\rangle + \left\langle \sum_{i}^{N} \mathbf{r}_i \, \mathbf{F}_i \right\rangle \right) \\ &= \frac{1}{3V}\left(\langle 2K \rangle + \left\langle \sum_{i}^{N} \mathbf{r}_i \, \mathbf{F}_i \right\rangle \right) \end{aligned} \qquad [2.92]$$

Sometimes, instead of the scalar value for the internal virial W, a tensor notation is preferred or advisable. The virial tensor is thus usually defined as:

$$\mathbf{\Xi} = -\frac{1}{2} \sum_{i \neq j} \mathbf{r}_{ij} \, \mathbf{F}_{ij} \qquad [2.93]$$

where $\mathbf{r}_{ij}$ denotes the distance between the couple of considered particles, and $\mathbf{F}_{ij}$ represents the force between the two centers of mass (and it is not the intra-particles force). Starting from the definition of the scalar pressure in terms of the trace of the stress tensor, it follows:

$$\begin{aligned} P &= \frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3} \\ &= \frac{1}{V}\left(\frac{2\mathrm{K}}{3} + \frac{1}{3} \sum_{i} \sum_{j} \mathbf{r}_{ij} \mathbf{f}_{ij} \right) \\ &= \frac{2}{3V}\left(\mathrm{K} + \frac{1}{2} \sum_{i} \sum_{j} \mathbf{r}_{ij} \mathbf{f}_{ij} \right) \\ &= \frac{2}{3V}\left(\mathrm{K} - \frac{1}{2} \sum_{i} \sum_{j} \mathbf{r}_{ij} \nabla_{ij} V(\mathbf{r}_{ij}) \right) \end{aligned} \qquad [2.94]$$

Hence finally, the pressure tensor results:

$$\mathbf{P} = \frac{2}{3V}\left(\mathrm{K} - \mathbf{\Xi} \right) \qquad [2.95]$$

Strictly speaking, all the above considerations about pressure are implicitly valid only in the microcanonical system, nevertheless they could be (more or less) easily modified to be extended to the other ensembles.

Thermostats: simulation of the canonical ensemble. The microcanonical ensemble represents an adiabatic system which does not correspond to the most common experimental conditions. Furthermore, both integration inaccuracies and the use of cut-offs introduce perturbations in the internal energy of the system, which usually are reflected by an increase in temperature. Hence, the need of a temperature control.

Briefly, algorithms for the temperature control can be summarized as follows:

- *Stochastic methods:* In stochastic collision methods, a particle is randomly chosen at fixed intervals and its velocity is reassigned by a random selection from the Maxwell-Boltzmann distribution (Anderson thermostat), which corresponds to a collision with an ideal heat-bath particle. The method generates a rigorous canonical ensemble, even if actually the system moves throughout the phase space on a constant-energy surface until the velocity of one molecule is changed as a consequence of the imaginary collision with the bath. When such an event occurs, the system jumps onto a different constant-energy surface, and the Hamiltonian motion restarts until the next collision. Thus, the sample is not strictly performed on a trajectory, but rather on a Markov chain. A too high collision rate will slow down the speed of the sampling, whereas a too low rate means that the canonical distribution of energies will only be sampled slowly.
- *Constraint methods:* Constraint methods simply rescale velocities by a proper factor in time. In the Berendsen implementation, the coupling with the ideal thermal bath (which owns the target temperature T_0) is reached introducing extra-terms in the equations of motions. In particular, forces are affected by means of both a *frictional* and a *stochastic* term, thus leading to a Langevin equation having the form:

$$m_i\dot{\mathbf{r}}_i = \mathbf{f}_i - m_i\gamma_i\mathbf{v}_i + R_i(t) \qquad [2.96]$$

where on the right the active (namely the effective) force, the forces rescaled by means of the frictional coefficient γ_i, and the stochastic term, appear respectively. In particular, the stochastic factor is built so as to provide a null average in time:

$$\langle R_i(t)R_i(t+\tau)\rangle = 2m_i\gamma_i k_B T_0 \delta(\tau)\delta_{ij} \qquad [2.97]$$

From the previous background, it can be demonstrated that the coupling with the ideal thermal bath is given by:

$$\left(\frac{dT}{dt}\right)_{bath} = \frac{(T_0 - T(t))}{\tau_T} \qquad [2.98]$$

If the coupling parameter τ_T is large, the coupling will be weak, whereas if τ_T is small, the coupling with the thermal bath will be strong. In the limit of $\tau_T = \delta t$ the algorithm becomes a classic rescaling method. Hence the new equations of motion become:

$$m_i\dot{\mathbf{r}}_i = \mathbf{f}_i - m_i\gamma_i\left(\frac{T_0}{T(t)} - 1\right)\mathbf{v}_i \qquad [2.99]$$

This method forces the system toward the desired temperature at a rate determined by τ_T, which has an optimal value of 0.4 for water. It does not provide a well defined statistical ensemble, nevertheless it is efficient and stable, thus it is generally performed on the equilibration phase of MD. It is usually known as isothermal MD.

- *Extended system methods:* The extended system methods consider the thermal reservoir to be an integral part of the system: energy is allowed to dynamically flow from the system to the reservoir and back. Such a reservoir is represented by an additional degree of freedom *s*, which owns a conjugate momentum *p_s*, and a certain amount of thermal inertia *Q*, which in turn controls the rate of temperature fluctuations. Potential and kinetic energy for the reservoir are defined as:

$$V_s = (f + 1)k_B T \ln s \qquad [2.100]$$

$$\mathrm{K}_s = \frac{1}{2}Q\dot{s}^2 = \frac{p_s^2}{2Q} \qquad [2.101]$$

where f is the number of degree of freedom, that is 3*N*-3 if the total momentum is preserved, whereas *Q* can be considered as the fictitious mass associated to the additional degree of freedom . Thus, the Lagrangian of the extended system is:

$$\mathrm{L}_{ext} = \mathrm{K} + \mathrm{K}_s - \mathrm{V} - \mathrm{V}_s \qquad [2.102]$$

The above description allows both a conserved Hamiltonian and a microcanonical density function, which can be quite easily converted to the canonical one by means of theoretical tricks. The parameter *Q* controls the energy flow between the system and the reservoir: if *Q* is large then the rate of the energy exchange is low, and in the limit of infinite *Q* conventional MD is reached. Conversely, if *Q* is too small, the energy would oscillate, resulting in equilibration problems. The most popular extended system method is the Nosé-Hoover method.

Barostats: simulations of the isothermal-isobar ensemble. Usually experiments are performed under a condition of constant external pressure, hence the pressure control should be advisable in MD just as the temperature control is. A macroscopic system maintains constant pressure by changing its volume, hence simulations performed in the NPT ensemble allow the system to reach its proper bulk density as it is derived from the parameters of the force field. This is particularly important when only poor guesses of the system sizes are available such as in MD of membrane systems. Furthermore, from a biological point of view, the isotherm-isobar ensemble would induce (or at least would not inhibit as the other ensembles actually do) large conformational changes, thus allowing a more exhaustive sampling of the phase space.

The amount of volume fluctuation is related to the isothermal compressibility κ:

$$\kappa = -\frac{1}{V}\left(\frac{\partial V}{\partial P}\right)_T \qquad [2.103]$$

For instance, an easily compressible substance would have large values of κ, allowing larger volume fluctuations at a given pressure than a less compressible substance. Conversely, in a constant volume simulation a less compressible substance shows larger fluctuations in pressure, which is an observable of the simulation.

Again, the main algorithms for pressure control will be summarized in the following:

- *Constraint methods:* the Berendsen algorithm for pressure control, rescales coordinates and box vectors by means of a coupling with an ideal pressure bath. Analogously for the Berendsen thermostat, a pressure coupling factor can be defined as:

  $$\left(\frac{dP}{dt}\right)_{bath} = \left(\frac{P_0 - P(t)}{\tau_P}\right) \qquad [2.104]$$

  The rescaling is then accounted for in the equations of motions by affecting the equations of velocities:

  $$\dot{\mathbf{r}}_i = \mathbf{v}_i - \mu_{\alpha\beta}\mathbf{r}_i \qquad [2.105]$$

  where $\mu_{\alpha\beta}$ is a rescaling tensor defined as follows:

  $$\mu_{\alpha\beta} = \delta_{\alpha\beta} - \frac{\delta t}{3\tau_P}\kappa_{\alpha\beta}\left(P_{0,\alpha\beta} - P_{\alpha\beta}(t)\right)$$

  in which $\kappa_{\alpha\beta}$ is the isothermal compressibility of the system.

- *Extended system methods:* the extended system methods are obviously similar in their derivation to algorithms aimed at the thermal control. Here, the coupling mimics the action of a piston on a real system. Such a piston has a mass, and it is associated both to a kinetic and potential energy. Again the equations of motion follow an extended Lagrangian formalism. The most commonly used method is the Parrinello-Rahman, which was originally developed as an extension of the Andersen pressure coupling algorithm. Within the Parrinello-Rahman barostat, the box vectors as represented by the matrix $\mathbf{b}$ obey to the matrix equation of motion:

$$\frac{d\mathbf{b}^2}{dt^2} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}\left(\mathbf{P}-\mathbf{P}_0\right) \qquad [2.106]$$

where V is the volume of the box, and $\mathbf{W}$ represents the mass of the piston, namely a matrix parameter which determines the strength of the coupling, and the amount of deformation affordable from the box (hence it is linked to the isothermal compressibility, to the box size and to the pressure coupling constant). Besides, the matrices $\mathbf{P}$ and $\mathbf{P}_0$ are the current and the reference pressure tensors, respectively. The equation of motion are then changed in a similar way to the Nosé-Hoover temperature coupling. The Parrinello-Rahman barostat was specifically designed for studying solid-solid phase transitions, hence it intentionally allows relatively large box fluctuation which could lead to a disastrous results unless the pressure of the box is near the equilibrium value.

Apart from the obvious differences regarding the dissimilar nature of the coupled quantity, constraint methods and extended system methods for temperature and pressure maintain same characteristics. Weak coupling methods, such as the Berendsen, lead to a strong damped *exponential relaxation*, while extended system methods such as the Nosé-Hoover or the Parrinello-Rahman give rise to an *oscillatory relaxation* of the property of interest. Oscillations, usually lead to a slower relaxation, hence in extended system the coupling constant should be 4 or 5 time larger than that of constraint methods, i. e., relaxation times in different algorithm classes are not equivalent neither in the definition nor in the effect.

2.3.1 Non-equilibrium dynamics: Metadynamics

Metadynamics is a potentially powerful non-equilibrium method which allows at the same time the sampling and the reconstruction of the free-energy hyper-surface (FES) of rare events for complex many-body systems. To do this, two basic assumptions are made:

1. Reduced dimensionality of the search: it is assumed that the free-energy A associated to the process under investigation could be expressed as a function of few relevant collective coordinates:

   $$s_i(x); \quad i = 1, n \qquad [2.107]$$

   The exploration of the search is then guided by the forces:

   $$f_i = -\frac{\partial \mathrm{A}(s_i(x))}{\partial s_i(x)^t} \qquad [2.108]$$

   Such variables must obey to some requirements:

   - They must be function of the coordinates of the nuclei of the system;
   - They must be able to univocally discriminate between the two free-energy basins which connect the reaction coordinate of interest;
   - They must include all the slow relevant modes associated to the investigated event, which can not usually be sampled with equilibrium dynamics.
2. The evolution of the trajectory is biased by a history-dependent potential which disfavors the visitation of the space already investigated.

In its original *discontinuous* implementation, in order to efficiently estimate the previously reported forces, an ensemble of replicas of the system was introduced. Such replicas were allowed to evolve independently each other at the temperature T, each obeying the constraint that the collective variables have a pre-assigned value $s_i(x) = s_i(x)^t$. Since replicas are statistically independent, the estimate of thermodynamic variables, namely the forces on the constraints, is improved. Hence, averaging over time and over replicas, the derivative of the free-energy f_i is evaluated, and it is exploited to perform a steepest descent-like step along the direction of the gradient. Later, the code was further developed in order to facilitate its natural implementation in MD schemes by introducing a formulation were a continuous evolution of collective variables was allowed. In particular, this is achieved by using an extended Lagrangian where collective variables $s_i(x)$ are treated as additional dynamical variables of the system being coupled to fictitious auxiliary variables $\tilde{s}$ by means of harmonic potential restraints. Hence, in the *continuous indirect version* of the method the extended Lagrangian is defined as:

$$\mathrm{L} = \mathrm{L}_0 + \left\{ \sum_i \frac{1}{2} M_i \dot{\tilde{s}}_i^{\,2} - \sum_i \frac{1}{2} k_i [s_i(x) - \tilde{s}_i]^2 \right\} + \mathrm{V}(t, \mathbf{s}) \qquad [2.109]$$

where:

- L_0 is the usual Lagrangian which drives the microscopic system dynamics;
- The second term (in brackets) identifies the Lagrangian for the coarse-grained system, and it is in turn defined by the fictitious kinetic energy of the auxiliary variables, and by an harmonic potential which restrains the collective variables to the dynamic value of the auxiliary ones, thus providing a link between the microscopic and the coarse-grained system;
- The last term represents the history dependent biasing potential, which has the functional form of a summation of *i*-dimensional Gaussian functions deposed in time with the frequency τ_G, and centered along the trajectory of the collective variables (Figure 2.2):

$$\mathrm{V}(s_i(x), t) = w \sum_{t'<t} e^{\left(-\frac{[s_i(x) - s_i(x_G(t'))]^2}{2\delta s^2} \right)}; \quad t' = \tau_G, 2\tau_G, 3\tau_G, \ldots \qquad [2.110]$$

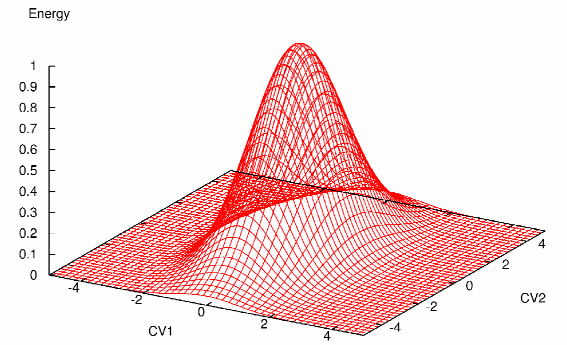


Fig. 2.2: Qualitative shape for a bidimensional Gaussian potential.

In equation 2.109, the masses M_i and the coupling constant k_i determine how fast auxiliary variables evolve in time with respect to the microscopic system. In particular, for a given value of k_i, if the masses M_i are large, then the motion of the auxiliary variables is slow, providing an adiabatic separation between the collective variables and the microscopic coordinates. In such a condition, the dynamics performed by the auxiliary variables is driven by forces arising form the harmonic potential:

$$f_i = k_i [s_i(x) - \tilde{s}_i] \qquad [2.111]$$

that are in turn an estimate of the derivative of the free-energy, which thus *does not have to be explicitly computed*. Hence, auxiliary variables are introduced in order to impose a set of dynamically evolving restraints acting over the collective variables, in a similar way to that seen for the discontinuous version of the method. In this picture, the kinetic energy of the auxiliary variables actually acts as a frictional term for the microscopic dynamics, which is exploited to better sample the local free-energy hyper-surface along the microscopic degrees of freedom, with the aim to obtain a better average estimate of the free-energy in the reduced dimensionality (hence mimicking the role of the replicas in the discontinuous version).

For all the above considerations, metadynamics can be defined as an *artificial dynamics performed in the space of few collective variables* (coarse-grained MD), *whose evolution in time is driven by a standard restrained microscopic dynamics* (hence the name, which refers to a "dynamics of a dynamics") *supplemented by an history dependent biasing potential*. At each meta-step (time has no longer its physical meaning) the evolution of the system is driven by the action of the contrasting generalized forces:

- Forces arising from the harmonic potential restraint, which represent a time average estimate of the thermodynamics forces and which would trap the system in the free-energy basin;
- Forces arising from the Gaussian potential which fills free-energy wells, and thus driving the system towards the nearest lowest saddle point.

If the Gaussian potentials have a suitable dimension (in terms of both height and scale) and if they are added sufficiently slowly, in the limit of a flattened and converged FES, the free-energy associated to the event of interest can be simply calculated from their sum changed of sign:

$$\mathrm{A}(s_i(x)) + \sum_{\tau} \mathrm{V}(s_i(x))_{\tau} = 0 \qquad [2.112]$$

Later on, it was discovered that the adiabatic separation is not a strict requirement for the success of the methodology, and that even if the masses are small, the history dependent potential still tends to flatten the underlying FES. Such a variant of the method, *continuous direct version*, were Gaussian potentials are directly added to the microscopic system, indubitably is faster in the sampling of the rare event, although the accuracy on the FES reconstruction is unavoidably reduced as a consequence of the loss of the coarse-grained forces averaging.

The efficiency of the FES exploration in metadynamics can be furthermore increased by taking advantage of the parallel facilities of current cluster machines. In particular, by defining a *walker* as a replica of the system which explores the FES in the space of the collective variables, in the *multiple walker* approach *n* walkers are each other independently evolved in time at the same temperature T, except for sharing the same Gaussian potential which is simultaneously deposed. The method which can be considered as an asynchronous parallelization, is straightforward and flexible, in the sense that walkers can be initialized and killed basing on the computational resources available at the moment, without significantly affect the error on the FES reconstruction.

Choice of collective variables. Of course, a crucial point for the success of the methodology is the proper choice of the collective variables for the reduced dimensionality in which the phase space of the rare event is projected. Usually, distances, angles and distribution functions are used, but sometimes *ad hoc* collective variables have to be designed in order to face a particular problem. In all cases, CVs should be representative of the (bio-)physical event under investigation. In this respect, a most dangerous phenomenon in metadynamics is the presence of hidden variables which can not be easily identified *a priori,* and thus the sampling is not accelerated along the wanted reaction coordinate. In a similar way, hidden variables could also lead to misleading or completely erroneous interpretations of the reconstructed FES.

Shape of the Gaussian potential. The analytic form of the added Gaussian functions, as long as their deposition time, have a not negligible effect on the accuracy of the reconstructed FES. The error in metadynamics is a measure of how different is the reconstructed free-energy $\mathrm{A}(s,t)$ from the real value $\mathrm{A}(s)$. It can be empirically demonstrated that the error is approximately proportional (independently of the dimensionality) to the square root of the basin size S, to the Gaussian width δs and to the Gaussian height w, while it is approximately proportional to the inverse square root of β and the diffusion coefficient D:

$$\bar{\varepsilon} \propto \sqrt{\frac{S\delta s}{\beta D}\frac{w}{\tau_G}} \qquad [2.113]$$

Again, in an empirical manner, it has been found that the optimal choice for the simulation parameters is:

- <u>Scaling factor δs</u>: it should be about $^1/_{10}$th of the basin size S. When the size of the basin can not be *a priori* estimated, a rule of thumb is to set δs lower than $^1/_3$ of the average fluctuation values obtained in a biasing potential free metadynamics;
- <u>Height w / Deposition time τ_G:</u> the error does not depend separately on w and τ_G, but only on the ratio w/τ_G. The value of approximately 0.001 kcal mol^{-1} fs^{-1} was found to be a good choice for a wide set of problems. Of course such a ratio can be obtained by using different combinations of w and τ_G, and in general for a given value of the ratio w/τ_G it is better to use a small deposition time (high frequency of deposition) along with small height. Intuitively, large values for w and τ_G will lead to significant discontinuities in the free-energy as a function of time, hence worsening the accuracy of the FES reconstruction.

Convergence of metadynamics. In spite of the huge reduction of dimensionality carried on the phase space sampled, convergence for pharmaceutically relevant systems remains an hard goal to pursue even when performing a multiple walker continuous direct metadynamics. For simpler systems, an useful rule of thumb is to consider the FES reconstruction converged immediately after a re-crossing event, which implicitly points out a flattening of the underlying free-energy. For the multiple walker approach, were trajectories of the single replicas are potentially constantly interchanging, the previously reported requisite is somewhat difficult to monitor. Besides, some root mean square distance function, calculated on subsequent free-energy surfaces, should be more informative.

2.4 Docking Simulations

Docking simulations represent a widely employed computational tool in pharmaceutical sciences, which attempts to predict a manifold of structures of intermolecular complex between at least two objects. Usually, but not necessarily, docking programs search along the degrees of freedom of a small molecule (the ligand) while the protein is treated as a rigid body. The result of such a search are configurations, namely conformations associated to a particular spatial orientation of the ligand at the binding site, which are usually referred to as solution *poses*.

A standard docking protocol consists of a step-wise process. First, a proper search algorithm predicts the various configurations of the ligand within the target binding site. In the second step, each docked pose is evaluated and ranked assessing the intermolecular interaction tightness throughout an estimation of the binding free-energy. Ideally, the correlation between the most favorable free-energy values and the best predicted poses should be very straight. The ability of a standard docking protocol to achieve its ultimate goal providing a reliable binding mode prediction, strongly depends on the accuracy of the scoring function used.

In the followings, the most common representation methods, conformational sampling algorithms and free-energy estimation methods will be briefly summarized.

Molecular representations. Docking techniques rely upon several receptor representation methods. The three basic methods are: atomic, surface or grid. The atomic representation is computationally expensive because of the complexity in evaluating pair wise atomic interactions by means of a suitable potential energy function. Besides, surface methods are based on the Connolly surface, which is defined as the van der Waals envelope surface accessible to a spherical probe. Such methods are mainly used in protein-protein docking, where a matching algorithm attempts to compute a rigid transformation that superimposes the protein surfaces mainly in terms of their complementarities. Finally, the grid approach stores physico-chemical features of the receptor on a regularly spaced grid. Within the assumption of a rigid receptor, the grid only needs to be computed once, hence saving computational time compared to the atomic approach. Basically, interactions of chemically diverse probes with the receptor are mapped in the grid and the protein-ligand affinity can be estimated by summing up the interaction energies for every probe corresponding to each ligand atom.

Ligand conformational sampling methods. The treatment of ligand flexibility can be summarized into three basic categories: systematic, stochastic and genetic algorithms. Systematic methods attempts to cover all the conformational degrees of freedom exploring each of them in a combinatorial way. Such methods provide an exhaustive search only in the limit of very rigid or simple molecules, otherwise a combinatorial explosion of the search dimensionality occurs, yielding the approach unfeasible. In order to avoid this, termination criteria are usually implemented which focus the sampling along regions of the conformational space that are more likely to lead an effective solution ("search and grow algorithms"). Conversely, stochastic approaches operate randomly selected changes along both the conformational internal and global (orientational/translational) degrees of freedom of the ligand, attempting to reach the global minimum for the molecule inside the binding site (Monte Carlo implementations). Within the Metropolis acceptance criterion, if the *i*-th solution of the conformational search bears an energy lower than the previous one (downhill move), it is always accepted. On the contrary, if the energy increases, a Boltzmann weighted probability function is then computed:

$$\begin{aligned} \rho(\Delta E) &= e^{(-\beta \Delta E)} \\ &= e^{\left(-\frac{1}{k_B T}\Delta E\right)} \end{aligned} \qquad [2.114]$$

where ΔE is the energy difference between the *i*-th and the (*i-1*)-th configurations. To accept an uphill move with the probability $\rho(\Delta E)$, a random number is uniformly generated in the range {0, 1}. If the random number is less than $\rho(\Delta E)$, then the uphill move is accepted, otherwise it is rejected. Here, an higher temperature can be usually applied to explore a wider potential energy surface. The search is then interrupted when the desired number of configuration is obtained. Genetic algorithms also implement a different amount of randomness, hence they should be formally classified between the stochastic ones. Nevertheless, compared with the properly called stochastic methods, they differ in the sense that they are based upon the principles of biological evolution and population dynamics, rather than on the laws of physics. Model parameters representing the degrees of freedom are encoded in data strings called "chromosomes". Such chromosomes are evaluated by a proper fitness function, and individuals whose chromosomes bear the largest fitness values have a better chance to reproduce and indeed to transmit their genetic inheritance to the next generation. Chromosomes are randomly varied by means of genetic-like operators, usually mutations and crossover, in order to increase density and prevent premature convergence. When applied to the docking problem, the genetic algorithm solution is a population of putative ligand conformations. For instance, in the software AutoDock 3.0.5, genetic algorithm represents an hybrid search technique that implements an adaptive genetic algorithm with a local search feature. The local searcher performs an energy minimization after the global sampling, hence the local changes occurred due to minimization are mapped back into the chromosomes. Since inheritance of acquired traits clearly contravenes the Mendelian genetic laws, in this sense the genetic algorithm is named “Lamarckian” after the discredited evolutionary theory of Lamarck. Similarly, another popular docking suite such as GOLD, employs a genetic algorithm whose most remarkable feature is the migration genetic operator. At the beginning, several subpopulations of chromosomes, called islands, are created instead of a large unique population. In order to preserve diversity, individuals are allowed to move among islands through the migration operator. Finally, only a fixed number of individuals can share the same place within an island. If there are more than a specified number of individuals in the same place, then the new individuals replaces the worst scoring member in the place, and not the worst individual in the overall population.

Scoring functions. The quantitative modeling of receptor-ligand interactions can be achieved by determining the equilibrium binding constant K_{eq}, which is in turn directly related to the Gibbs free-energy:

$$\begin{aligned} \Delta G &= -RT \ln K_{eq} \\ &= \Delta H - T\Delta S \end{aligned} \qquad [2.115]$$

The difficulties relying on the estimation of the free-energy by means of computational techniques have already been extensively covered in the paragraph 2.3, hence just an overview focused on the docking field will be given here. Docking simulations are at now usually performed in *vacuo*, although in principle implicit solvation models could be used as well. Nevertheless, in spite of their theoretical derivation, scoring functions are usually able to provide a proper assessment of the enthalpic contribution for the free-energy (a force field-like potential energy function), whereas the entropic contribution remains hard to estimate. The main entropic contributions to the stability of the receptor-ligand complex are provided by desolvation effects, and by the internal conformational degrees of freedom of the docked small molecule. Within the docking field, the need for a fast scoring method led to a number of different functions which bring various assumptions and approximations in the evaluation of modelled complexes. Widely employed approximations are:

- scoring functions assume that the free-energy of binding can be approximated using a single structure, which is a reasonable assumption since the lower is the energy of the configuration, the larger is its contribution to the partition function;
- the bound state for the complex is the only explicitly considered, whereas unbound components are implicitly accounted for;
- the free-energy is approximated by a linear combination of several terms, while several forces involved in the complex formation are non additive.

Empirical scoring functions provide an estimation of the binding free-energy by summing up interaction terms derived from structural parameters. The development of scoring functions is based on the idea that binding energies can be approximated by means of a sum of uncorrelated terms, which are derived by regression analysis from experimentally determined structures whose binding mode are known. Such kind of scoring functions are simple and intuitive, but their main drawback is that it is not clear whether they are able to predict the binding affinities for ligands whose structure is not covered among the training set.

Chapter 3

Applications.

In this section results of selected MD studies performed during the course of this thesis will be reported. In particular: i) standard equilibrium MD applied to the study of a pharmaceutically relevant system (the hERG channel), and ii) the proposal of a novel methodology aimed at the discrimination of the correct binding mode for docked complexes by means of a non-equilibrium MD (metadynamics), will be discussed.

3.1 Molecular dynamics simulations of membrane embedded-proteins

3.1.1 An overview of membrane simulations

Starting from the pioneering studies of van der Ploeg and Berendsen[1] who modeled for the first time a membrane bilayer mimetic (a system made of 16 decanoate molecules per leaflet in vacuo) more than twenty years ago, the increased computational power nowadays allows more realistic simulations in terms of: i) lipid models, ii) membrane dimensions, and iii) time scales of sampling. Nevertheless, the above reported ground-breaking work already showed the fundamental theoretical guidelines, which have been followed until now when performing MD simulations on lipid systems. In general, when dealing with simulations two key topics have to be taken into account:

1) simulation conditions (namely the choice of force-field, statistical ensemble and simulation parameters);
2) validation of the simulation against experimental data.

If experimental data are not well reproduced by the calculations, the simulation is devoid of any physically meaningful predictive power, hence one should change its simulation protocol in a iterative way until reaching a reasonable agreement with experiments. In lipid simulations, this task is somewhat more complicated than usual, because both the simulations conditions are not theoretically well established (that is a pretty amount of discussion have been arose in the scientific community) and the experimental reference data suffer from a high degree of uncertainty. The aim of this section is to provide some guide to better understand the literature from a pharmaceutical point of view.

3.1.1.1 Experimental data

Since MD simulations are based on models, results of the calculations have to be validated by experimental data. Over the last decades a variety of experimental techniques has been applied to membrane systems, such as diffraction methods and nuclear magnetic resonance techniques. However only a relatively small number of properties can be directly compared to simulation results[2, 9]. The main difficulties arise from the great differences both in dimensions and in timescales that theoretically and experimental techniques can intrinsically afford. Experimental structural properties of membranes are evaluated in ***macroscopic*** (usually) multilamellar patches over ***long time*** periods (hours), whereas in computational techniques the same properties are derived from a ***microscopic*** (or mesoscopic, if enough computational resources are available) system, usually studied in periodic boundary conditions, over extremely ***short time*** periods (tens or sometimes hundreds of nanoseconds)[3]. In other words, there is either a time and scale gap between theoretical and experimental studies which indeed complicate any comparison (Fig 3.1).

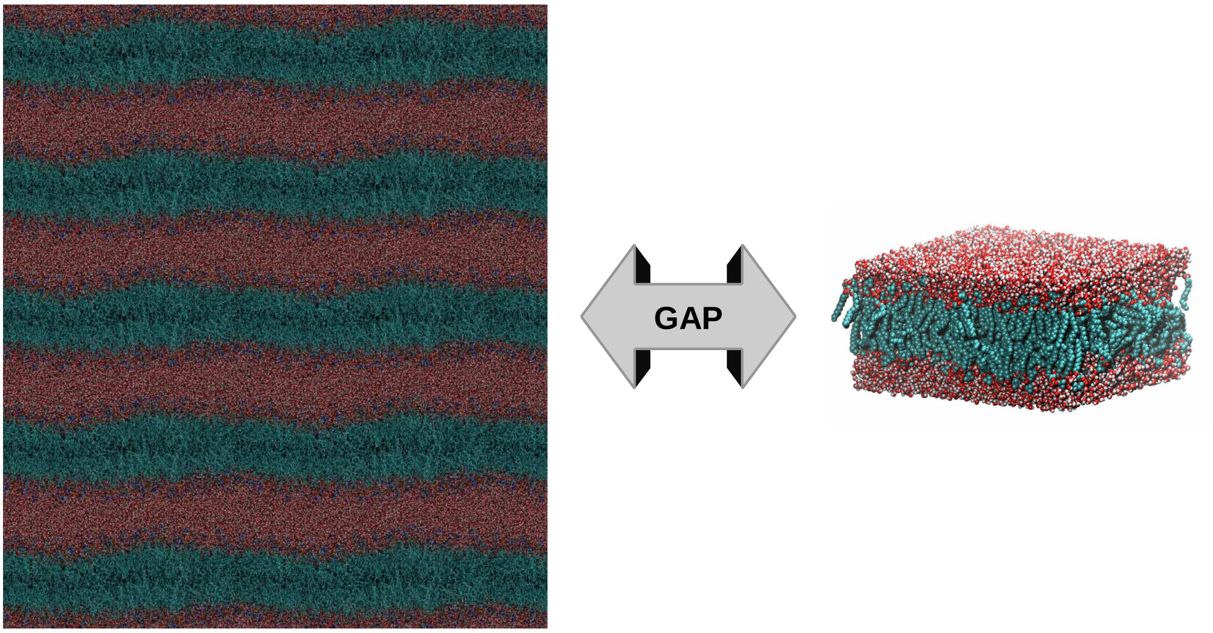


Fig 3.1: Pictorial view of the gap between the experiment (macroscopic regime) and the simulation (microscopic, sometimes mesoscopic regime).

In general, the experimental properties available to check the simulations can be distinguished in *structural*, *dynamical*, and *thermodynamic* quantities (see chapter 2.3). X-ray diffraction and nuclear magnetic resonance spectroscopy, which are the most used techniques in lipid bilayer studies[13], provide the structural class of quantities. Among them, the most widely used parameters (which are not necessarily the best ones to judge a simulation) are[2]:

- density profiles: electron and atom densities;
- cell parameters: area per lipid (A_0), lamellar repeat spacing (D);
- order parameters for the lipid chains (S_{CD}).

Conversely, neutron scattering experiments provide the main dynamical class of quantities, such as[13]:

- lateral diffusion coefficient;
- rotational diffusion coefficient.

As previously reported in simulation studies usually only structural parameters are reported, and since the simulation of a lipid bilayer environment is focused on the study of lipid-embedded proteins, only the biologically relevant liquid crystalline L_α-phase has to be considered.

Order Parameters. The structure of saturated lipid membranes has been in depth investigated by means of deuterium magnetic resonance, and in particular measuring the quadrupole splitting of selectively deuterated (in position 2, 3, 4, 5, 9, 10, 12, 14, 15 of the lipid acyl chains) non-sonicated bilayers of L_α-diplamitoylphosphatidylcholine[17]. The anisotropy of the CD bond direction with respect to the bilayer normal was quantified using the $-S_{CD}$ order parameter, according to the formula:

$$-S_{CD} = \left(\frac{1}{2}\left(3\left\langle\cos^2\theta\right\rangle - 1\right)\right)\Big/2 \qquad [3.1]$$

where θ denotes the temporary angle between the direction of a chain segment and the bilayer normal, while brackets indicate the average orientation. From a computational point of view the general order parameter tensor S_{ij} is defined as[1]:

$$S_{ij} = \frac{1}{2}\left\langle 3\cos\theta_i\cos\theta_j - \delta_{ij}\right\rangle \qquad [3.2]$$

in which θ_i is the angle between the *i*th molecular axis (*x*, *y* and *z*) and the bilayer normal, δ_{ij} is the Kronecker delta function (basically a suffix notation for the diagonal matrix with elements 1: δ_{ij} = 1 if $i = j$; δ_{ij} = 0 if $i \neq j$), whereas in this context brackets denote the ensemble average. The molecular axis for the *n*th CH_2 unit is defined as follows[1, 9]:

- *z*: vector from C_{n-1} to C_{n+1};
- *x*: H-H vector (or vector perpendicular to *z* and in the plane through C_{n-1}, C_n and C_{n+1} for a united-atoms representation);
- *y*: bisectrix of HC_nH angle (or vector perpendicular to *z* and *x* for a united-atoms representation).

S_{ij} represents a rank 2 tensor, and in particular – for reasons of symmetry – a 3x3 diagonal matrix, hence bearing a null trace, which leaves essentially two independent order parameters per unit[1]. Order parameters range between +1 (unit fully ordered parallel to the bilayer normal) and -1/2 (unit fully ordered perpendicular to the normal), while isotropic orientation corresponds to a value close to zero[1].

Experimental values for $-S_{CD}$ for the 2th to the 8th CH_2 unit (approximately half an acyl chain) are 0.20 ± 0.20, whereas proceeding along the end of the lipid tail the order parameter drops towards zero, indicating an increasing isotropy[3, 17]. The values reported are averaged over time and over the lipid chains (sn1 and sn2), even if quadrupole splitting indicates that they are not completely equivalent physically[17]. In the region of the constant order parameter, the same overall angular fluctuations for all the segment has to be involved, excluding the occurrence of isolated gauche conformations. Instead, it can be explained by means of fluctuating kink- or jog-like structures[17]. The gradual decrease of the order parameter which characterize the innermost region of the bilayer can be rationalized by an increasing probability of gauche conformations[17].

Cell Parameters. Other central structural quantities are represented by the primary lamellar repeat spacing D, and the average area per lipid molecule A_0[18, 19, 20]. From a computational point of view dealing with a finite simulation box, such quantities are definitely correlated. Experimentally, the repeat spacing is the easiest diffraction result that can be accurately obtained (D = 67.2 Å for the fully hydrated benchmark DPPC lipid at the temperature of 50 °C[19]). Conversely, data collected for the average area per lipid show a significant spread[18, 19], and in particular a huge degree of uncertainty affects the measure of the difference between the biologically relevant fluid phase area and the area of the gel phase[18] (Fig 3.2). It is informative to notice from Figure 3.2 that experimental data refer to different laboratories and to different experimental techniques as well. Moreover, it is apparently surprising that a considerable spread is also observed in NMR data, where A_0 is derived from sterical considerations from S_{CD} that is, as previously reported, a quite accurate measurement. This means that uncertainty is also (at least partially) due to different interpretation of the same measured quantities. Nevertheless, such a data scattering is not helpful to drive MD simulations. Most of the difficulty in obtaining good quantitative structural parameters for the biologically relevant, fully hydrated, fluid L_α phase is due to the intrinsic presence of fluctuations[19]. Based on the latter observation, Nagle and co-workers introduced a correction which provides an adjustment to the literature values of A_0[19]. A revision of the corrected values led the authors to propose 64 Å^2 as the best value for A_0[19].

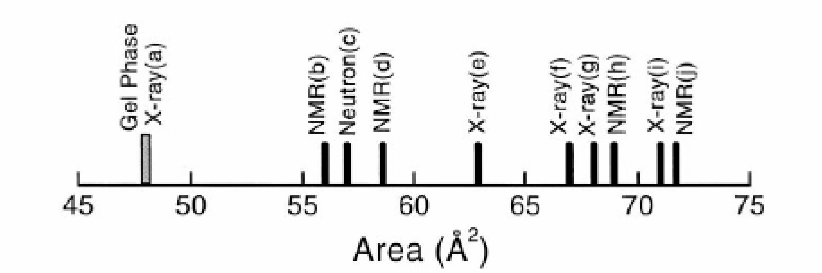


Fig 3.2: Summary of published areas for DPPC at 20 °C (gel phase, grey bar) and at 50 °C (fluid phase, black bars). The plot is taken from reference [19] where relative references can be found therein.

3.1.1.2 Simulation conditions

When planning a simulation, usually a compromise between time and length scales has to be chosen. The bigger is the length (per bilayer dimension) of the system, the shorter the sampling. In Table 3.1, some typical timescales of lipid relaxations are reported[2]. From Table 3.1 it should be clear that while MD is a powerful method to sample single lipid conformations, at the same time it would prevent any significant investigation far from the starting configuration, since the rotational and translational motion of lipids are quite slow to be significantly sampled[2]: hence, systematic drifts in any structural parameter are usually imputed to some artifact of the methodology.

Tab 3.1: Typical timescales of lipid relaxations.

Timescales	
Tens of picoseconds	Trans-gauche isomerization of dihedrals in the lipid tails.
Few hundreds of picoseconds	Trans-gauche isomerization of dihedrals close to the lipid head-group.
Few nanoseconds	Rotation around the lipid *z*-axis.

Tens of nanoseconds
(Average affordable simulation time)
	Lateral diffusion;
	Intra-leaflet lipid switch;
	Cooperative motions in phase transition

Minutes to hours	Rare events, e.g. inter-leaflet lipid flipping.

Furthermore, the computational cost of a "long" dynamics run is exponentially affected by the increasing of the length scale of the bilayer. On the other hand, when simulating a "big" bilayer system one should expect to reproduce properties that "small" patches simply cannot, such as long term wave fluctuations. This task is complicated by the fact that the simulation protocol for a relatively small membrane system should be different from that of a bigger one (not only in simulation parameters, but also in the proper statistical ensemble to choose, as some author suggest, see below). Nowadays, systems of few hundreds of lipids per leaflet are normally used in membrane simulations.

Lipid models. Biologically realistic lipid simulations should in principle take into account more than a single component, since considerable properties of cell membranes arise from a proper lipid mixture[15]. However, simulations of mixed bilayers remain challenging because of both the long timescales needed for relaxing the mixture of lipid components, and the large dimensions of the system that are required to reach a suitable balance among different components. Single component simulations are usually performed when dealing with lipid embedded proteins, and the choice of the lipid is generally made either in terms of the amount of experimental available data, and acquired experience as well. DPPC, POPC, DMPC are the most studied phospholipids, and among them – until now – DPPC is the preferred, since it represents the best experimentally characterized. Nevertheless, DPPC models for biological simulations suffer from the fact that acyl chains are completely saturated, thus the proper fluidity of the L_α-liquid crystalline phase cannot be reached unless an unrealistic temperature in simulations is used. Moreover, lipid models are usually available both in an united-atoms or in an all-atoms representations, although the united-atoms approximations is the most widely used since (at least for pure bilayer simulations) there is no need to an explicit representation of non-polar hydrogen atoms, hence computational time can be saved.

System size. Because of increased computational power and algorithm advances, nowadays molecular dynamics simulations of lipids systems are about to reach the mesoscopic regime[14]. In such a domain, collective phenomena occur on length scales more than 10 nm and time scales of the order of 10 ns[14]. Spontaneous undulations, found in some MD simulations, are an example of such a collective effect. From the point of the simulation, it is important to recognize when the microscopic behavior turns into the mesoscopic one. Apparently, simulating too small systems (or systems with an applied surface tension, which actually stretches the membrane) would lead to an overestimated surface area per lipid due to the suppression of undulations. This behavior can be explained by the difference between the local area and the projected area for a system which undergoes collective fluctuations, such as a larger theoretical system or the experimental one[14].

Statistical Ensembles. During the years, basically 3 trends can be identified in the literature concerning the suggested proper statistical ensemble to use in MD simulations of bilayer systems. Keeping in mind that exceptions can always be found, for all the eighties and during the early nineties simulations dealing with fixed NVT thermodynamic variables were commonly performed. This was mainly due to the relevant computational demand of such calculations. Nevertheless, the cost to pay was that, unless a very good estimate of structural parameters such as the surface area per lipid was known for the lipid of interest as an initial condition, rough artifacts in lipid density could affect simulations, as it was later demonstrated by Tieleman and coworkers[3]. In 1995 Chiu et al.[4] for the first time introduced the need of the surface tension as an explicit fixed thermodynamic variable. By analogy with alkane/water interfaces and insoluble monolayers at the air/water interface, they suggested that there was a non zero surface tension at the lipid/water interface in bilayer phases that had to be accounted for in constant-pressure simulation of lipid bilayers[4]. Later on Feller, Pastor and collaborators supported this approach suggesting the presence of an applied external surface tension, actually stretching the membrane[5,6]. Briefly they observed that, while a bulk fluid can be described only by three thermodynamic variables (NVT, NPT and so on), for an interfacial system a fourth variable is required to take into account its inherent anisotropy. For a liquid/liquid interface made of two components each in a single phase, a generalization of the Gibbs phase rule (number of intensive variables equals to $c + p - 2$ with c number of components and p number of phases) for surface phases permits the specification of more than two variables for such a system, thus allowing the use of the statistical ensemble NP$_n\gamma$T (even if $\gamma = f(P_n, T)$)[5,6].

Considering a system made of two immiscible liquids forming a planar interface normal to the *z* direction with area *A*, the stress of such a system (as any other system) is defined as the measure of the distribution of the force per unit area, which is by definition a second-order tensor, namely a quantity that requires two array indices to be described, and thus is represented by a 3×3 square matrix. In three dimensions, the internal force *F* acting on the infinitesimal area *dA* can be resolved in three components: one normal to the plane, and the remaining parallel to the plane (Fig 3.3). In other words, the stress tensor σ_{ji} is defined as:

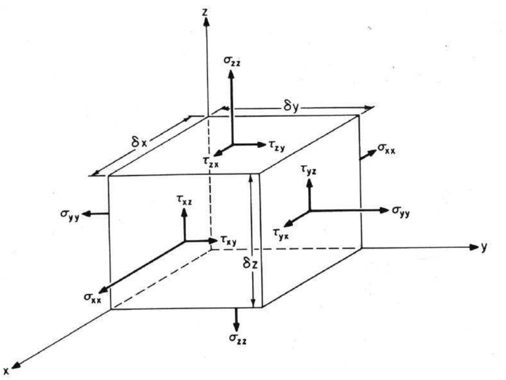


Fig 3.3: Geometrical representation of the nine components of the stress tensor for a cubical shape simulation cell.

$$dF_i = \sum_{j=x,y,z} \sigma_{ji} dA_j \qquad [3.3]$$

and hence:

$$\sigma_{ij} = \begin{bmatrix} \sigma_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \sigma_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \sigma_{zz} \end{bmatrix} \qquad [3.4]$$

where usually one can formally distinguish between σ, the *normal stress* for the normal force component, and τ, the *shear stress* for the parallel force components. If the system is momenta free (as MD simulation systems should be), the stress tensor is symmetric, and univocally defined by only six indices. Furthermore, for a system in hydrodynamic equilibrium the shear stress is null, and the intrinsically isotropic nature of such a condition leads to an invariant trace where each component corresponds to $\frac{1}{3}\sigma_{ij}$. The scalar pressure is hence defined as:

$$P_{iso} = \frac{\sigma_{xx} + \sigma_{yy} + \sigma_{zz}}{3} \qquad [3.5]$$

Handling with a semi-isotropic system, hereafter we will refer to the normal pressure and tangential pressure terms, respectively defined as follows:

$$P_n = \sigma_{zz} \qquad [3.6]$$

$$P_t = \frac{\sigma_{xx} + \sigma_{yy}}{2} \qquad [3.7]$$

While P_n is invariant, P_t strongly changes with respect to the *z* axis: in the bulk of the system it holds that $P_t = P_n = P$, while in proximity and in correspondence of the interface it becomes large and negative, as a consequence of the difference in density between the two components of the biphasic system. The surface tension is defined by the relation:

$$\delta W = \gamma \,\mathrm{d}\, A \qquad [3.8]$$

were δW is the work required to change the surface area by the amount $\mathrm{d}\, A$. Alternatively, one may calculate the work required to change the shape of a bilayer slice at constant volume against the normal pressure P_n and the lateral pressure $P_t(z)$. If this is set equal to $\gamma \,\mathrm{d}\, A$, one obtains the integral:

$$\gamma = \int_{-\infty}^{+\infty} [P_n - P_t(z)]dz \qquad [3.9]$$

In computer simulations, the surface tension can be coupled either implicitly, by specifying an anisotropic stress tensor ($P_t < P_n$) thus leading to the ensemble (not rigorously theoretically defined) $\mathrm{NP_nP_tT}$, or explicitly by the ensemble $\mathrm{NP_n\gamma T}$ (to date just the most popular ensembles). Nevertheless, such an approach showed both experimental and theoretical drawbacks. For instance, it is questionable the way to derive the surface tension for a bilayer, which is experimentally not (yet?) measurable[2, 10]. Chiu et al.[4] *estimated* the surface tension for a *monolayer* water/lipid interface, and then they *assumed* the surface tension of the *bilayer* to be twice the former:

$$(\gamma_{monolayer\ \ phase\ change})_{\exp erimental} = (\gamma_{water\,/\,air})_{known} - (\gamma_{water\,/\,lipid})_{unknown} \qquad [3.10]$$

$$\gamma_{bilayer} = 2 \times \gamma_{water\,/\,lipid} \qquad [3.11]$$

Applying such an approach they proposed for a DPPC lipid bilayer in the $\mathrm{L_\alpha}$-liquid crystalline phase a surface tension of about 56 dynes/cm. Even though they demonstrated that a liquid crystalline phase emerged from simulations starting from a $\mathrm{L_\beta}$-gel phase, the methodology seems to switch the problem from the NVT ensemble, where a good guess of the dimensions of the simulation box is needed, to another ensemble were another experimental quantity is still just approximately known (furthermore, it is not even clear if it exists, see below). Apart from this, in 1996 Klein and colleagues claimed for the theoretical weakness of the procedure, reminding that the surface tension is thermodynamically defined as the derivative of the free-energy with respect to the area at constant temperature and volume:

$$\gamma = \left(\frac{\partial F}{\partial A}\right)_{\mathrm{T,V}} \qquad [3.12]$$

and asserting that at the equilibrium condition, membrane systems adjust their area such that the free-energy is a minimum, thus the surface tension vanishes[7, 8]. For this reason they performed simulations of both liquid and gel phase of DPPC preserving its behavior by using isotropic NPT macroscopic boundary conditions in a fully flexible simulation box, namely where each dimension is allowed to independently scale in respect to the others[7, 8]. A slight debate occurred in the biophysics community, involving authors who both joined the match or basically acted as referees[10]. For instance, in one of the last publications of his life, Jähnig stressed that, although the surface tension of a bilayer is not directly accessible, indirect measurement support the idea that it is zero[11]. Again Feller and Pastor answered to the criticism arguing that a non-zero surface tension should be used to reproduce the correct surface area per lipid value as a consequence of the limited sizes of the simulation box, where long wavelength fluctuations are absent since prevented by periodic boundary conditions, and then the surface area per lipid would shrink[12]. This controversial question went on for a while, even if Tieleman and co-workers already stated that – in practice – the choice of the NPT or $NP_n\gamma T$ ensemble makes a very little difference[9, 2]. Nowadays, the most popular software of MD actually allows the use of every ensemble henceafter discussed, along as diverse kind of box scaling, since it has been clarified that once the principal structural parameters are in good agreement with experiment, the simulation conditions are not so strictly important.

Simulation parameters. Among the whole set of simulation parameters, it has been recognized that a major role is determined by the truncation of short-range electrostatics and van der Waals interactions[16]. It is not surprising that treatment of electrostatic interactions has a strong impact both in structural and in dynamic properties of the bilayer, since phopsholipids are highly charged molecules[16]. Concerning the long-range electrostatics, the particle-mesh Ewald technique has been increasingly used in lipid bilayer simulations[16]. As long as the PME method is used, the short-range electrostatic truncation seems to be not very important, since reasonable values of area per lipid can usually be reproduced[16]. In particular, artificial order in the bilayer plane may arise in simulation performed without an explicit treatment of long-range electrostatics, which in turn implies that the lipid bilayer no longer owns a truly fluid-like state[16]. However, even PME and related techniques are not free from potential artifacts. In details, such artifacts are related to the periodicity of the system, as periodic boundary conditions are used to eliminate finite size effects[16].

The cut-off radius for the Lennard-Jones energy function (r_{vdW}) is another parameter which has a non-negligible impact on simulations, although its effect is not *a priori* as intuitive as that due to the electrostatic truncation. Actually, a systematic study performed by Patra et al.[16] revealed an inverse relationship between r_{vdW} and the calculated area per molecule. This trend can be rationalized considering that an increase in the cut-off radius actually would increase the attractive interactions between acyl chains, thus reducing A_0[16].

3.1.2 The hERG Potassium Channel

The hERG potassium channel (also known as $K_V11.1$ according to the IUPHAR nomenclature[21], gene: KCNH2) is a human voltage-gated homo-tetrameric protein composed by the radial assembly of transmembrane (TM) spanning α-subunits (S1-S6 segments).

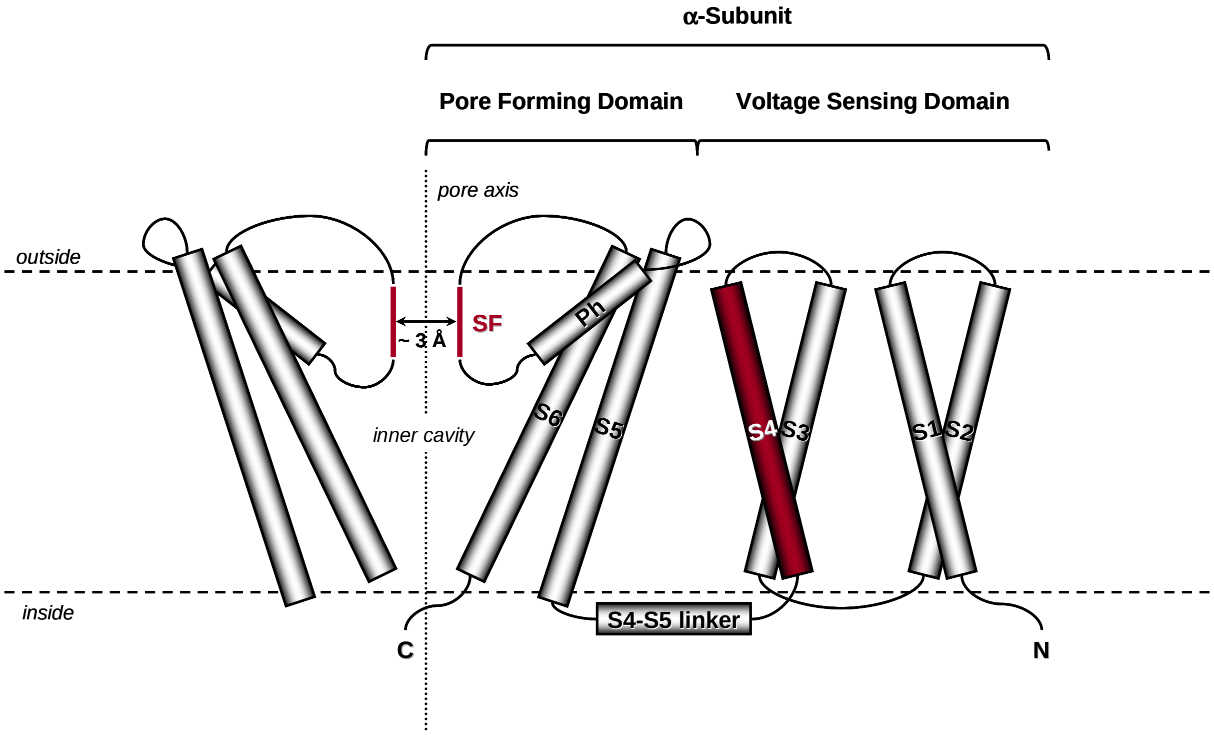


Fig. 3.4: Main topological features for the transmembrane portion of a "standard" potassium channel belonging to the K_v family, based upon the latest experimental insight provided by the crystallographic structures solved by the MacKinnon group.

In the K_v family members, two main TM domains can be distinguished: i) the voltage sensing domain formed by four α-helices (S1-S4, where S4 represents the voltage-sensing element), and ii) the pore forming domain (S5-S6). Comparing the amino acid sequences, the pore of these membrane proteins are well conserved among all K^+ channels, since they need to perform the same physiologic function, namely, the capability to preserve both high ion selectivity and rate conduction[41, 42, 46]. In details, all K^+ channels display a "signature sequence" (residues TVGYG) in the so-called P-loop segment that is located between the pore helix (Ph) and S6. In hERG, the signature sequence shows a slight different amino acids composition (SVGFG), which however does not seem to affect the ion permeation and rate. This sequence forms the so called selectivity filter (SF) that structurally corresponds to the narrowest part of the pore (about 3 Å of diameter) where the ion crossing discrimination occurs[43 – 45, 55]. Conversely, the widest portion of the pore (usually referred to as the cavity) is formed by part of S6 helices (one for each subunit) located underneath the SF. The gating mechanism affects the cavity width (in terms of diameter and dimension), and is usually determined by a concerted conformational modifications in the S6 helix at the PVP motif, which is highly conserved throughout most of the voltage-gated potassium channels, and it is supposed to provide a kink in S6 that allows the opening and the closure (deactivation) of the cavity. In Figure 3.4 the main topological transmembrane features of a "standard" K_v potassium channel are schematically reported.

Belonging to the K_v10-K_v12 sub-family (according to the IUPHAR nomenclature[21], better known as eag sub-family), the hERG channel owns some individual sequence features, likely responsible for a unique gating mechanism. In fact, the hERG kinetics is characterized by a slow activation and deactivation, but a rapid inactivation and recovery from the inactive state[39, 40]. In particular, hERG lacks the PVP motif in the S6 helix, responsible for the cavity enlargement. In eag sub-family, a similar structural function seems to be played by a conserved glycine hinge (Gly648 of hERG) in analogy with prokaryotic potassium channels[46]. Moreover, hERG is provided of a large (~ 43 residues including a putative α-helix[47, 48]) extracellular S5-Ph linker, which is supposed to be responsible for a rather unique fast inactivation mechanism[27, 47].

hERG is mainly expressed in the heart muscle tissue[22] and is responsible for the rapidly activating component of the delayed inward rectifier currents (I_{kr}), which play a major role in the modulation of the repolarization phase (phase III) of the myocyte action potential[23]. Impairments in hERG functionality, hence alterations in the I_{kr} currents, are commonly referred to as the second form of the Long QT Syndrome (LQT2), as they are clinically associated to a broad widening of the QT interval in the electrocardiogram recording[25, 26 and references therein]. Inherited mutations in the hERG channel expressing gene[24, 27] or drug induced block[25] cause the congenital or acquired LQT2 syndrome, respectively. In particular, a number of structurally diverse drugs belonging to different pharmaceutical classes (e.g., antihistamines, antidepressants, antipsychotic, gastrointestinal prokinetics, etc.)[28], has been reported to block the channel, leading to a consequent significant prolongation of the ventricular repolarization. This occurrence, along with several complementary causes, potentially gives rise to an occasionally lethal polymorph ventricular tachi-arrhythmia, named *torsades de pointes*[28]. Withdrawal of some common drugs that showed this remarkable side effect (such as Astemizole, Sertindole, Thioridazine, Terfenadine, Grepafloxacyn, Cisapride, etc.)[52],has strongly focused the attention on the hERG channel as a pharmaceutical anti-target[54]. At this respect, a detailed molecular understanding of the interactions lying at the basis of the channel functioning and block would be of major interest for the rational drug design. For this purpose during the last years a large effort has been done in order to characterize the physicochemical features required for a high affinity binding[30 – 33].

Among the in-silico techniques, both ligand-[34, 35] and target-based[36, 37, 38, 56] approaches have been undertaken. Until now, the latter approach has been strongly hampered by the lack of any crystallographic data about the TM domain of the hERG channel, where experimental studies suggest to be located the drug binding site[55, 53]. Nevertheless, the pioneering work done by the MacKinnon's group has provided a fairly large set of K^+ channel crystal structures useful as templates for comparative modeling work. Recently, Reynolds and co-workers, starting by the crystal structures of KcsA (closed state) and MthK (open state), modeled the hERG channel in a multiple state representation in order to account for the protein flexibility[36]. These models were then used to properly assess the binding affinity for a set of docked ligands. In a second paper, Åqvist and co-workers docked a series of Sertindole analogues (antipsychotic compounds bearing a potent hERG blocker activity) at the channel cavity in an open state homology model built using the crystal structure of the voltage-gated KvAP as a template. The lowest energy docked poses belonging to the most populated clusters were then selected for molecular dynamics (MD) refinement[37]. Pearlstein and co-workers, based on the same template structure, docked a series of well known blockers[38]. A multiple docking configuration was found to be achieved by different compounds, suggesting a more flexible description of the drug binding[38]. Finally, Choe et al. proposed a further hypothesis of drugs binding to hERG using an homology model of the channel again based on the KvAP template[56].

In the present work, we modeled both the closed and open states of the hERG channel using the crystallographic structures of KcsA[49] and KvAP[50, 51] as templates, respectively. Since alanine scanning mutagenesis experiments showed that the binding site of most drugs is located inside the cavity[29, 53], only the channel portion ranging from S5 and S6 was built. To assess the reliability of the models, MD simulations in explicit membrane environment was then carried out, and a careful analysis of the pore volume in the putative drug binding site was also undertaken. Furthermore, we probed the suitability of including MD simulations in the docking of a ligand to the channel cavity. In fact, we found that snapshots from the MD trajectory, but not the starting conformation could provide reasonable docking complexes, thus showing that our simulation protocol was able to take into account an induced fit-like effect (actually, a proper thermal protein relaxation) needed for the drug binding of the drug to the channel.

3.1.2.1 Sequence alignment and homology modeling

Comparative models of hERG channel in a closed and open state were built starting from the crystallographic coordinates of KcsA (PDB entry: 1K4C[49]) and KvAP (PDB entry: 1ORQ[50, 51]).

The extent of the TM helices of the hERG channel was assessed by means of the PHDhtm server[57, 58], as it was able to overall reproduce both the KcsA and KvAP (segments ranging from S5 to S6, i.e. the modeled portion) transmembrane topology (data not shown). To date, only one mammalian K_v potassium channel has been solved, i.e. the K_v1.2, which owns the previously introduced PVP motif[59, 60].

The multiple sequence alignment was performed with T-Coffee[61]. In order to take care on the structural consistency between the two templates, local manual adjustments of the alignment were required. In particular, the alignment of the innermost helices, namely those ranging from Ph to S6, was quite straightforward and useful to address the manual refinement of the remaining part of the alignment. The overall percent of identity calculated against the whole 98 residues per subunit, was 17.3 and 28.6% for the KcsA/hERG and the KvAP/hERG pair, respectively (Fig. 3.5). It should be noted that for both the alignments, the identity significantly increases reaching the modeling reliability threshold of 30% on the S6 helix, namely the binding site for most drugs.

The closed and open state models of the hERG channel (hereafter referred to as $hERG_C$ and $hERG_O$, respectively) were therefore built by comparative modeling using Modeller 7v7[62]. As Figure 3.4 points out, the homology model procedure was restricted at the segments spanning from S5 through S6, while the extracellular S5-Ph linker was not included in the models (see further). Besides, the alignment clearly showed some critical features of the hERG channel pore moiety. In most voltage-dependent potassium channels, the SF is formed by the sequence TVGYG[41, 42], whilst in hERG, the sequence is replaced by SVGFG. These mutations could lead to a different stability of the different SFs. Moreover, the conserved putative glycine gating hinge, located at position 648 in the target sequence, is shown in Figure 3.4. Finally, the two aromatic amino acids (Tyr652 and Phe656) crucial for binding hERG blockers are also displayed. Although the overall sequence identities were quite low, that is 17.3% and 28.6% for the KcsA/hERG and KvAP/hERG pairs, respectively, they raised up to more than 30% in the S6 helix, which is the region of the channel mostly involved in the drug binding (Figure 3.4). Since the models were mainly aimed at studying ligand-channel interaction mode, this made us confident enough on the accuracy of the proposed theoretical models.

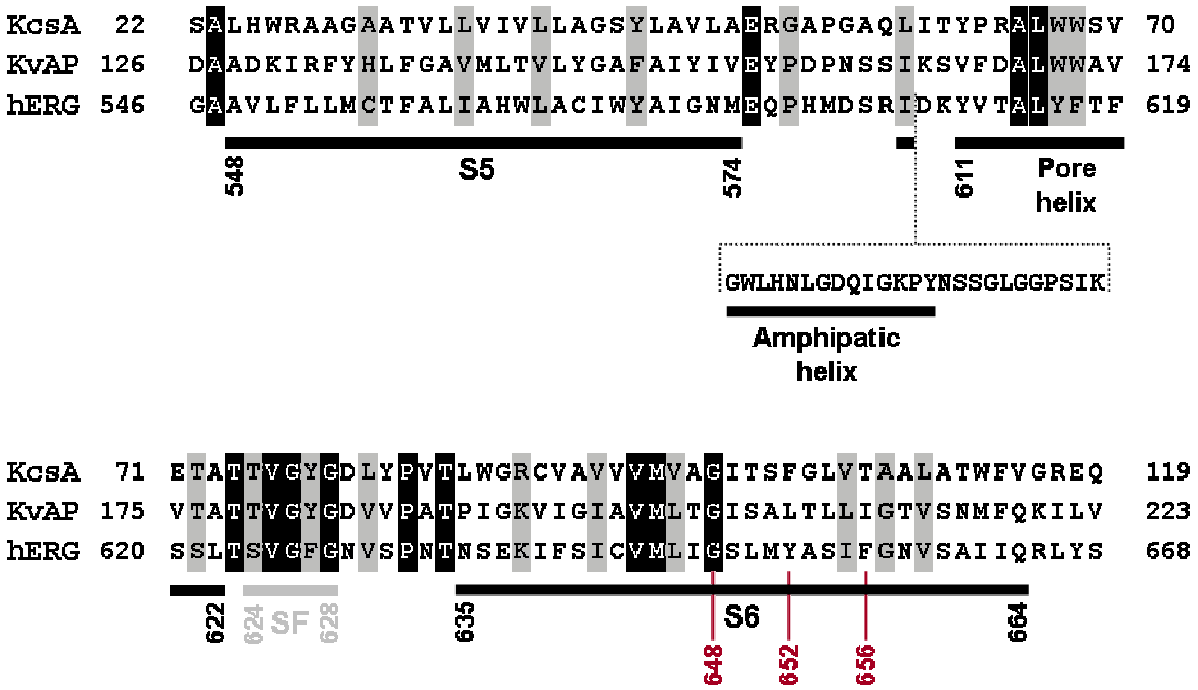

KcsA 22 SALHWRAAGAATVLLVIVLLAGSYLAVLAERGAPGAQLITYPRALWWSV 70
KvAP 126 DAADKIRFYHLFGAVMLTVLYGAFAIYIVEYPDPNSSIKSVFDALWWAV 174
hERG 546 GAAVLFLLMCTFALIAHWLACIWYAIGNMEQPHMDSRIDKYVTALYFTF 619
548
S5
574
611
Pore helix
GWLHNLGDQIGKPYNSSGLGGPSIK
Amphipatic helix
KcsA 71 ETATTVGYGDLYPVTLWGRCVAVVVMVAGITSFGLVTAALATWFVGREQ 119
KvAP 175 VTATTVGYGDVVPATPIGKVIGIAVMLTGISALTLLIGTVSNMFQKILV 223
hERG 620 SSLTSVGFGNVSPNTNSEKIFSICVMLIGSLMYASIFGNVSAIIQRLYS 668
622
624
SF
628
635
S6
664
648
652
656

Fig. 3.4: Sequence multiple alignment between the modeled portion of the hERG channel and the chosen templates (KcsA and KvAP). A common putative secondary structure is highlighted underneath the hERG sequence, and helices from S5 to S6 are labeled. Conserved and homologous amino acids throughout all the sequences are highlighted in black and grey, respectively, while the key residues in the S6 helix of hERG are explicitly shown in red. The S5-P linker bearing the putative amphipatic helix (not modeled) is separately shown for clarity.

For each gating state a set of 10 models was generated by imposing fourfold rotational symmetry as well as *ad hoc* orientational restraints acting on the χ dihedrals angles of the amino acidic side chains of the SF, in order to preserve its overall geometry as regards to the respective template. The couple of models candidate as a starting structure for the MD simulation was then selected mainly in terms of the stereochemical parameters, which were evaluated by means of Procheck v3.3 validation tool[63]. The most satisfactory models are reported in figure 3.6.

The selected closed and open models showed an overall G-factor of 0.01 and -0.03, respectively, as a main consequence of a high amount of residues having a combination of ϕ-ψ dihedral angles lying either in the core or in the allowed regions of the Ramachandran plot. In particular, while the closed state model lacks of residues located outside the allowed areas of the plot (95% core and 4.5% allowed), the open state showed only a residue per subunit (Arg582, an amino acid belonging to the loop which replaces the S5-Ph linker) in a generally allowed region (93.3% core, 5.6% allowed, and 1.1% generally allowed).

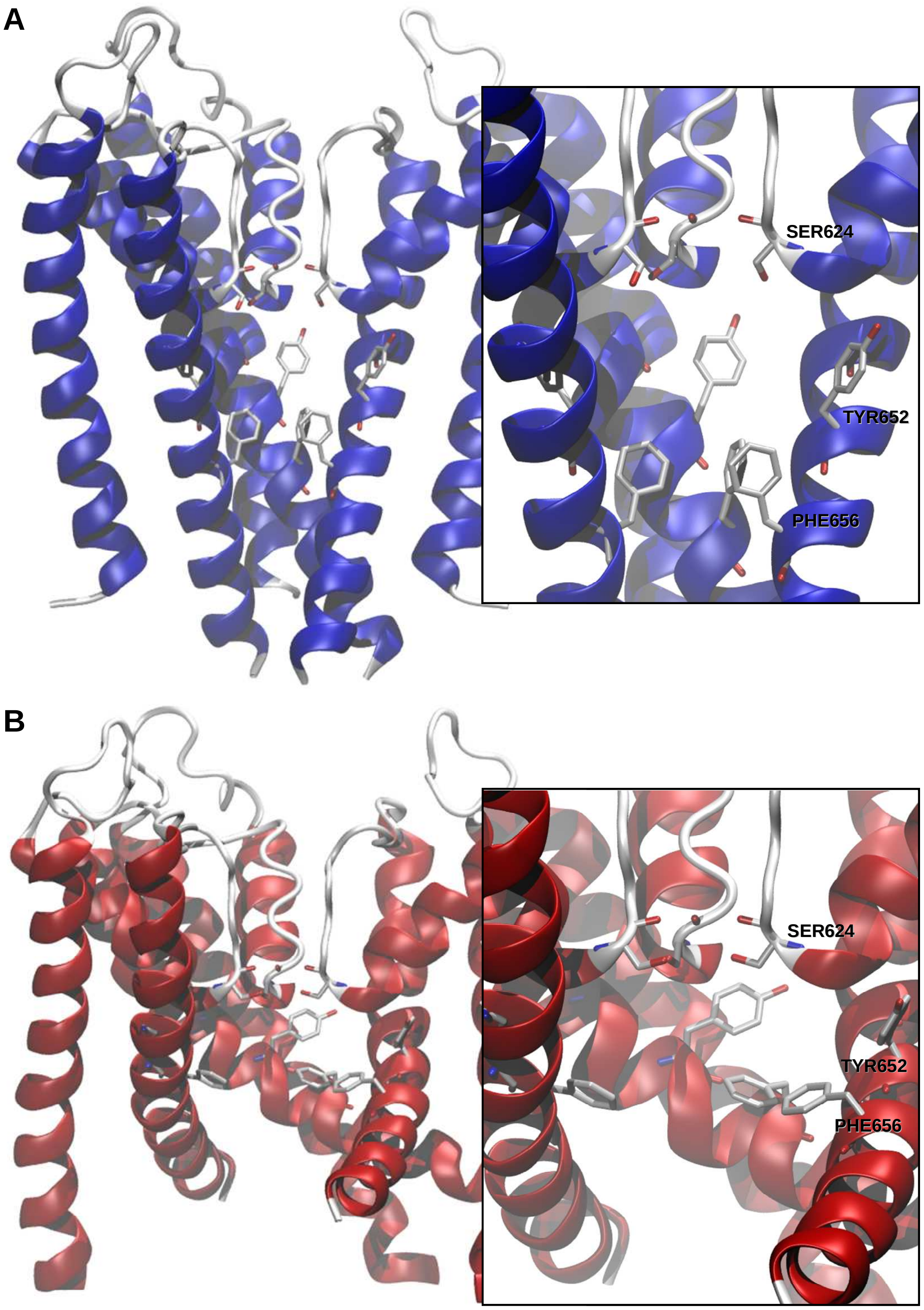


Fig. 3.6: Cartoons representation of the selected homology models for the hERG channel in the closed (hERG$_C$, blue) and open (hERG$_O$, red) state. For each model a focus on the binding site is shown, where the key amminoacidic residues Ser624, Tyr652 and Phe656 are explicitly displayed as well. For clarity just three out of four subunits are shown (namely the chains A and C in foreground, and D in background).

As expected, the overall folding of the two models closely reproduced the one of their respective templates. This was assessed by evaluating the RMSD calculated over the Cα atoms after superposition of the template/target pair carried out on the same set of atoms, which was 0.85 and 1.12 Å for the closed and the open states pairs, respectively. The most striking difference between the modeled proteins and the templates was the presence of Phe656 lying in the middle of the cavity, which considerably reduced the pore radius, especially in the closed state model. Notably, this is also one of the two residues (the other is Tyr652) univocally required for the binding of all kind of drugs to hERG (see Introduction).

As regards the difference between the two gating states, visualization of the pore lining reveals the main difference that is the opening at the intracellular mouth of the channel. In particular, as figure 3.6 points out, the portion ranging from the Ph to the SF was highly similar in the open and closed models, whereas in the open state model the glycine hinge (Gly648) induced a kink in the C-terminus portion of the S6 helix of about 37° with respect to the closed one, which significantly enlarged the channel cavity.

3.1.2.2 Molecular dynamics

In order to carry out a proper simulation of a transmembrane protein, both suitable lipid bilayer environment and simulation protocol needed to be set up.

All the simulations were performed using GROMACS 3.1.4[64, 65] implemented with the native GROMOS-87 extended-atoms force field[66], and running on a local LINUX cluster employing an openMosix ® architecture.

Membrane set up. A pure di-palmitoyl-phosphatidyl-choline (DPPC, 16:0) lipid model in a united atom approximation was used. The overall system was built starting from a smaller pre-equilibrated membrane[67]. The original system was then replicated by means of symmetry operations until reaching the wanted dimensions. The final model comprises 256 lipid molecules per leaflet in a water:lipid ratio of about 29:1 as shown in figure 3.7. Such a condition of full hydration allowed us to simulate the Lα-liquid crystalline phase, namely the biologically relevant thermodynamic phase, at the temperature of 325 K and at the total pressure of 1 bar[9, 67, 68].

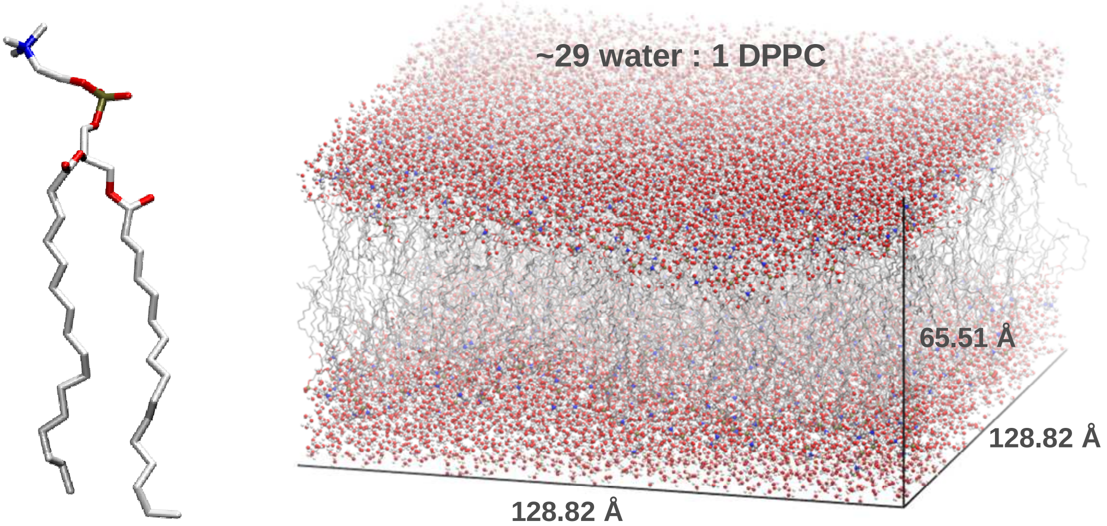


Fig. 3.7: Representations of the di-palmytoil-phosphatidyl-choline in a united-atom approximation (*left*), and the whole system used (*right*). Geometrical features of the simulation cells are also explicitly reported.

A 5 ns MD simulation was run in the NPT statistical ensemble using semi-isotropic pressure coupling, as it has been already clarified to be the most suitable theoretical approach to be used when dealing with lipid bilayer systems[7 – 11, 66 – 68]. Actually, using the above mentioned parameters, the extension of the box in the *z* direction, which is normal to the bilayer plane, was permitted to vary independently of both the *x* and *y* sides, thus allowing the adjustment of the surface area per lipid along the simulation run. Lipid parameters were those calculated by Jähnig et al.[69] used together with the consistent *ab initio* derived point charges[4], whilst SPC water molecules[70] were used as they well reproduce the proper solubility of interfacial systems[9]. van der Waals interactions and short range electrostatics were explicitly handled with a twin-range cut off scheme by updating the neighbor list every 20 time steps. In particular for the Lennard-Jones and the Coulomb functions cut offs of 10.0 and 9.0 Å were respectively used, while for the internal radius of the twin-range scheme the value of 9.0 Å was applied. Conversely, long range electrostatics were taken into account by means of the particle mesh Ewald method by using a Fourier spacing of 1 Å, interpolated by fourth-order B-spline, and by setting the direct sum tolerance to 10^{-5} [71]. The Berendsen algorithm for pressure coupling[21-72] was used with a time constant of 2 ps, whereas the temperature was independently coupled to the lipids and to the solvent with the Berendsen algorithm for temperature[72] with a time constant of 0.5 ps. All bonds involving hydrogen atoms were constrained with the LINCS algorithm[73], while the water geometry was kept fixed with the SETTLE algorithm[74]. The integration timestep was set to 2 fs.

The membrane system was equilibrated in the NPT statistical ensemble by using a semi-isotropic pressure coupling for a total time of 5 ns, with the aim to provide a suitable lipid environment where the homology-built channels had to be inserted. Usually, the accuracy of such a simulation is assessed by monitoring a series of structural quantities and by comparing their values to the experimental ones[9]. The Surface Area per Lipid (A_0), namely the most commonly reported parameter, is shown in Figure 3.8 plotted against the simulation time. In the graphic, the experimental value corrected and revised from Nagle and co-workers,[19] is also reported. As it can be noticed from the plot, the calculated A_0 compared to the experimental value was over-estimated during the simulation time of about 1.5 units. Noteworthy, a typical periodic behavior of fluid systems could also be detected, in other words no systematic drifts in the above considered parameter were found all along the whole 5 ns of MD simulation. Nevertheless, the A_0 parameter by itself is a proper measure neither of the force-field nor of the methodology, as reported by Marrink et al.[67]. Therefore, it is useful to compare the calculated $-S_{CD}$ parameter in the light of the above reported behavior of A_0. Thus, the second quantity taken into account during the analysis was the deuterium chain order parameter ($-S_{CD}$) calculated on the carbon atoms belonging to the lipid acyl chains. In Figure 3.9, plots of the latter quantity averaged over time intervals of 100 ps sampled along the MD trajectory are shown, together with the experimental value[17].

The trend of this parameter quantifies the average anisotropy of the lipid tails relative to the normal of the bilayer plane (*z* axis). Considering the experimental values, starting from the sp^2 carbon acyl chain (atom number 1 in Figure 3.9) and proceeding along the lipid tail, the parameter shows a plateau at about 0.20 units for half a chain, and then it decreases for the remaining carbon atoms, reflecting an increased anisotropy of the lipid chains. What is informative is that the experimental trend was somewhat reproduced in the calculation, and it is almost kept along the trajectory, as the average time interval plots show (Figure 3.8).

These findings demonstrated that the chosen simulation parameters along with the selected force field were able to avoid serious artifacts such as a forthcoming geliphication. Accordingly, a suitable lipid environment for the following protein simulations was achieved, as the Lα-crystalline phase of the membrane was kept throughout all simulation time.

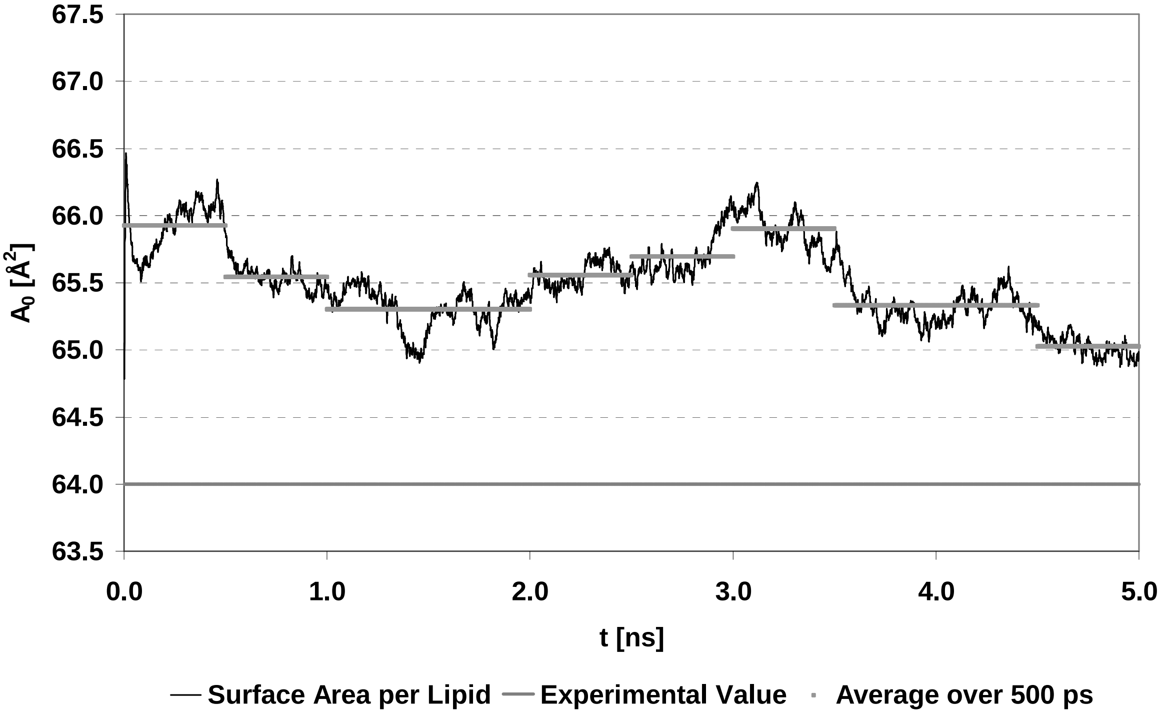


Fig. 3.8: Surface Area per Lipid (A_0) plotted against simulation time. The calculated and the experimental value are shown in black and in grey, respectively.

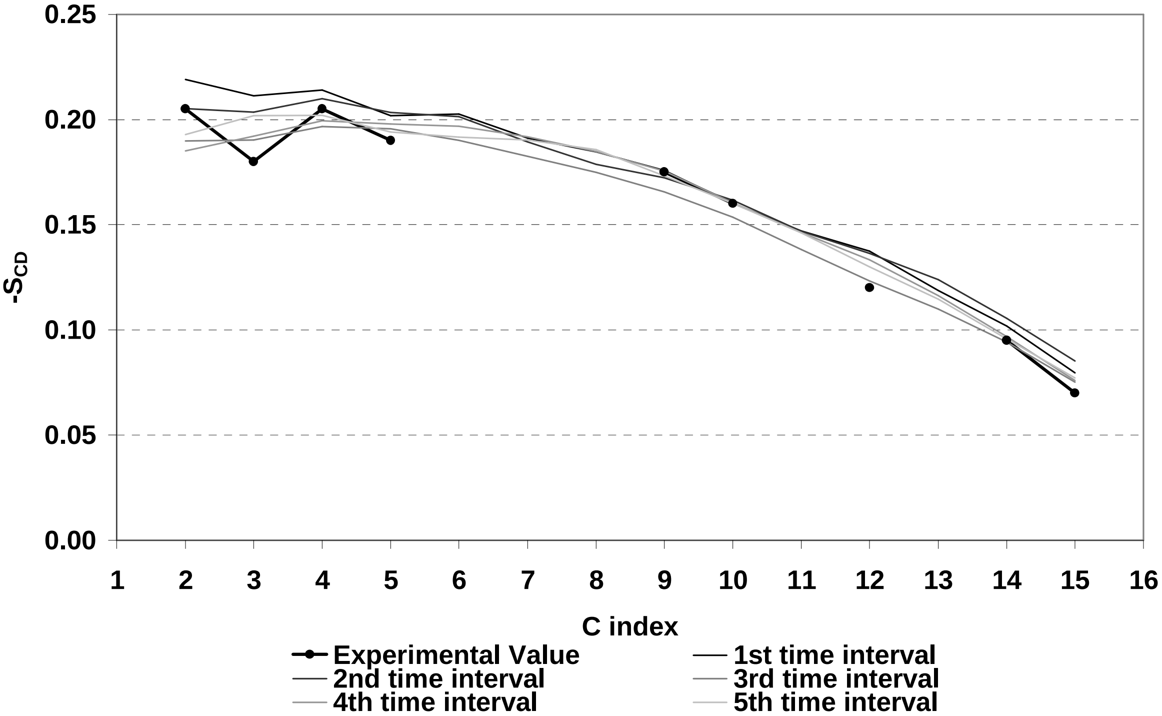


Fig. 3.9: Deuterium chain order parameter ($-S_{CD}$). The calculated and the experimental value are shown. Each time interval is defined by a MD fraction of 100 ps of duration, regularly sampled along the trajectory. The calculated value was computed averaging over both the lipid tails for all the lipids of both the membrane leaflets, and over the time.

A snapshot taken at 2 ns of the membrane equilibration was used as input configuration for the following step of protein set up. To deal with the dimensions of the channels along the pore axis, and in order to reduce artifacts due to the periodic boundary conditions, two additional layers of water molecules (each of 10 Å of thickness) were added both in the extra- and in the intra-cellular sides of the membrane. Water was once more briefly thermalized by means of an additional 50 ps of MD using the same previously reported equilibration protocol.

Protein set up. The couple of candidate models (hERG$_C$ and hERG$_O$) was introduced in the geometric centre of the equilibrated membrane environment adapting the Sansom et al. two-stage protocol[75]. First, lipids overlapping a cylindrical volume equivalent to that of the protein, which was assessed by means of a solvent excluded surface (SES) model, were removed. In particular, in order to take into account the anisotropic shape of the channel along the *z* axis (especially for the closed model), the protocol was reiterated for both layers of the membrane, by calculating for each step the partial protein volume embedded in the considered leaflet. The second phase of the procedure consisted in the close-contact minimization of the protein-lipid interface. This was achieved by a series of short NVT ensemble MD runs, by exerting an incremental radial force originating from the vertices of the protein surface, projected in the *xy* plane and acting on the atoms of the left lipid molecules. By means of an empirical approach, the following three-step optimized protocol was found to properly work for both the models.

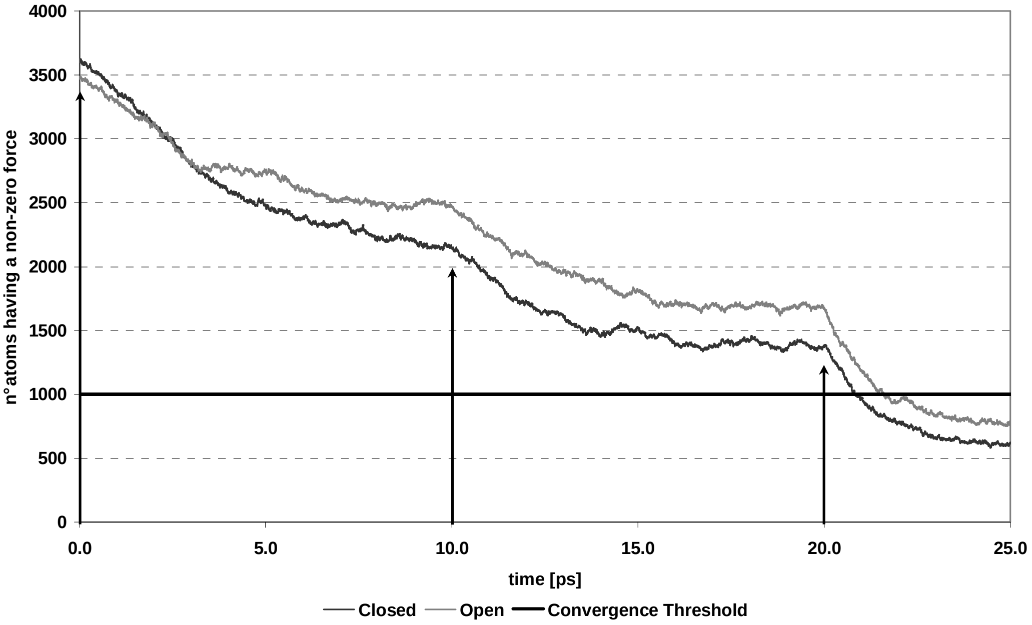


Fig. 3.10: Three-step relaxation procedure of the membrane environment in respect to the solvent accessible surface of the implicit channel models. Each step is shown by means of a vertical arrow. The convergence threshold was set to be lower than 1000 atoms experiencing a null force.

The magnitude of the radial force was of 5 (step 1), 10 (step 2) and 30 kJ mol^{-1} Å^{-1} (step 3), and it was exerted for 10, 10 and 5 ps, respectively, whilst the overall shape of the membrane was kept by applying *z* restraints on the DPPC headgroups (10 kcal mol^{-1} Å^{-2}). A steady state was then reached for each step, namely a condition in which the number of solvent (meant waters and lipids) atoms having a non zero-force remained approximately constant, and the convergence was achieved when this value was lesser than the arbitrary threshold of 1000 atoms (Figure 3.10). A pictorial representation of the process is given in Figure 3.11.

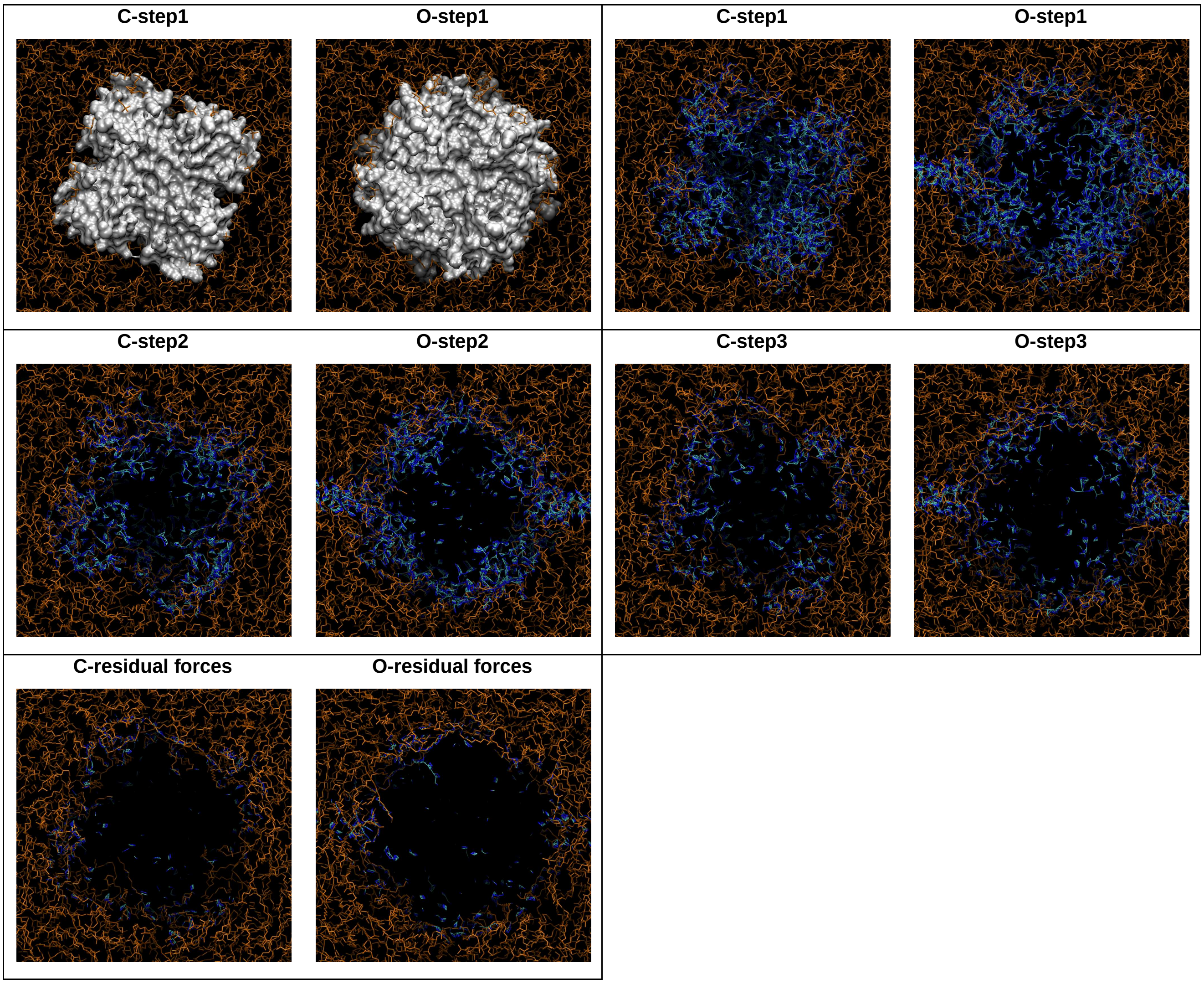


Fig. 3.11: Pictorial representation of the three-step relaxation procedure of the membrane environment. Snapshots referring to each temporary steady state are taken from the top of the extracellular side, and both for the closed (*C*) and open (*O*) state models of the channel. Lipids are shown in orange, whilst atoms which experience a direct non-null force are coloured in blue. The predicted displacement of such atoms after the chosen time step is shown in cyan. For the step1 the implicit SES model of the channels is shown as well. The residual forces of the last couple of snapshots refers to the same magnitude of step3.

The correct positioning of the protein with respect of the bilayer is crucial for the success of the above reported procedure, and this is usually achieved by maximizing the contacts between aromatic sidechains and lipid headgroups[75, 76]. Unfortunately, in our opinion, the hERG primary sequence ranging from S5 to S6 does not allow the previous unbiased approach, therefore we preferred to first insert the well studied KcsA template according to Carloni et al.[76], and then to coherently superimpose the models by minimizing the RMSD function calculated against the SF Cα atoms: RMSD KcsA/hERG$_C$ = 0.13 Å; RMSD KcsA/hERG$_O$ = 0.11 Å. The straightforward positioning of aromatic residues both in S5 and S6 helices for the KcsA channel is shown in Figure 3.12.

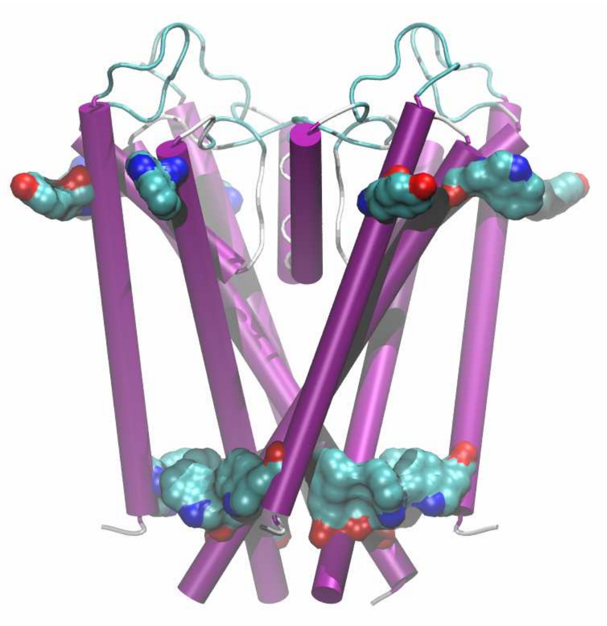


Fig. 3.12: Aromatic residues located at the extrema of the transmembrane spanning helices in KcsA which were exploited to drive the *z* positioning of the hERG models into the membrane environment.

According to the nomenclature for the selectivity filter occupation proposed by Åqvist et al.[77] the 10101 configuration was chosen as input structure, which corresponds to three potassium ions located in the S_0, S_2 and S_4 crystallographic sites, separated by two water molecules lying in the S_1 and S_3 sites. This configuration is consistent with the single file ion motion model in which the K^+ ions alternate with water molecules in the selectivity filter[78, 79]. Moreover, an additional potassium ion was placed inside the pore, namely at the so called S_{cav} site, which is supposed to be important to stabilize the ions in the selectivity filter in the closed state of the channel[80].

Technically, filter water and ions were positioned by superposing the backbone of the templates filter onto the target models, and transferring across the respective ion coordinates. For both the S_1 and S_3 sites an oxygen atom took place of the potassium coordinates. Besides, since the KvAP crystal structure lacks of the K^+ ions both in S_0 and (obviously, being a channel in an open state) in S_{cav}, for the hERG$_O$ model these ions were transferred from the superposed KcsA crystal structure. Ions and water molecules involved in the coordination will be henceforth referred to as K_0, W_1, K_2, W_3, K_4 and K_{cav} proceeding from the extra- to the intra-cellular side, respectively (Figure 3.13).

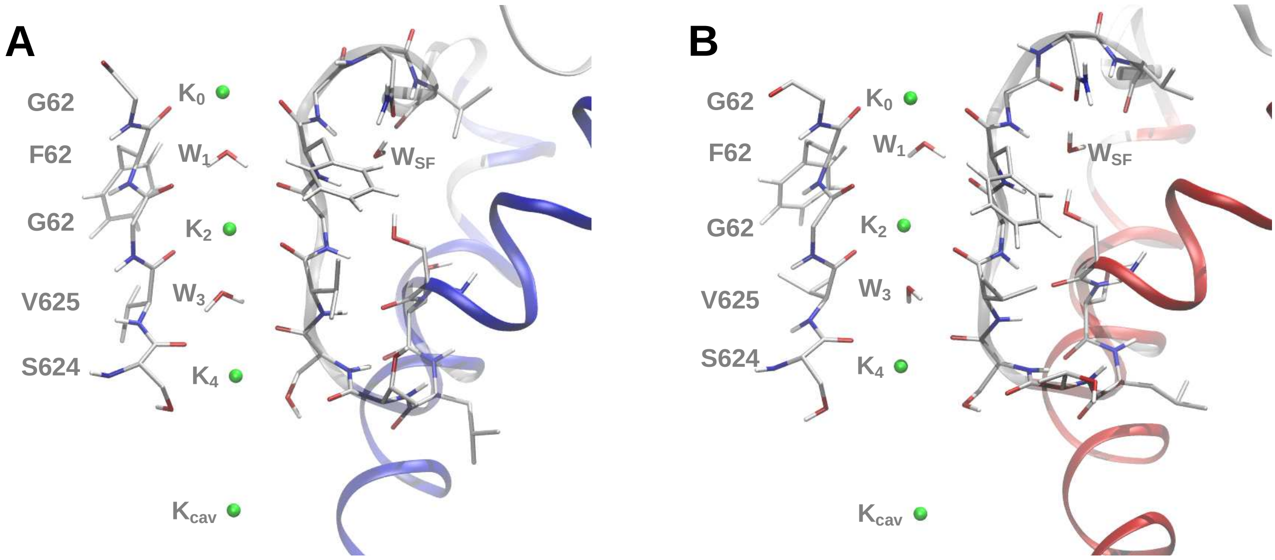


Fig. 3.13: Starting configuration for the occupation of the SF in $hERG_C$ (*left*) and in $hERG_O$ (*right*) models. The snapshot is referred to the equilibrated systems, bearing each channel already inserted in the membrane.

For both channels the electro-neutrality was reached by adding 4 Na^+ counter ions, since each modeled subunit had a null total charge. The cavity of the open state of the hERG channel was solvated by means of the specific GROMACS tool, whereas for the closed model a more accurate procedure was needed. In particular, with the VOIDOO software[81] the shape and the accessible volume of the pore were assessed, then by means of FLOOD[81] the cavity was filled by water molecules. By taking into account the van der Waals volume of K_{cav} a total of 22 (out of 25) water molecules were added in the cavity. An additional crystallographic (in the KcsA solved at high resolution, i.e.: 1K4C[49]) water molecule located “behind” the SF as regards to the pore axis, was found to be fundamental to preserve the conformation of the upper moiety of the filter along the production run[82 – 84]. Such a supplementary water molecule, hereafter reported to as W_{SF} (Figure 3.13) seems to provide important interactions with the filter and the remainder of the protein.

Simulation parameters slightly varied compared with those used for the membrane equilibration. Specifically, the Nosé-Hoover[85, 86] algorithm with a coupling constant of 0.5 ps and the Parrinello-Rahman[87] algorithm with a coupling constant of 0.5 ps, were used as pressure and temperature coupling methods, respectively. These algorithms were preferred because of their more accurate statistical ensemble, while in the membrane simulation the Berendsen algorithms were chosen as they give rise to lesser oscillations in the structural parameters used to evaluate the quality of the simulations itself (that is, the surface area per lipid and the deuterium chain order parameter, see the results).

Short-range non-bonded interactions were taken into account by means of a twin-range cutoff scheme by updating the neighbor list every 20 steps, and by using a cut-off of 9.0, 10.0, and 12.0 Å for the internal radius, the Lennard-Jones function, and the Coulomb function, respectively. Conversely, long-range electrostatic was handled in the same way as reported for the membrane alone simulation.

The system was thermalized with the following equilibration protocol. Firstly the whole system was energy minimized by using the Steepest Descent algorithm starting from the solvent (water and lipids) while the protein was restrained, and then by releasing the protein structure. Then, the solvent was equilibrated during 50 ps in the NVT statistical ensemble MD run at the temperature of 100 K, half time by freezing both the DPPC molecules and the entire protein, and the remaining time by constraining only the protein atoms. After that, the whole solvent equilibration was further extended for 50 ps, by increasing the simulation temperature up to 325 K. The purpose of this stage of the equilibration was mainly to relax the lipid molecules as regards to the protein geometry. To this end, according to Roux et al. the temperature was intentionally set above that of the gel-liquid phase transition of DPPC (315 K), in order to confer a higher fluidity to the lipid molecules[78], whereas in the following of the simulation the temperature was set at the standard value of 300 K. A short (5 and 10 ps for $hERG_C$ and $hERG_O$, respectively) qualitative steered dynamics, carried out pulling K_0 in the *z* direction towards K_2, was needed to properly arrange the former ion in its coordination site, namely S_0. It must be noted that this is the only selectivity filter site which is half composed by four carbonyl oxygen atoms, whilst the remaining ligands are provided by extra-cellular water molecules[78, 76]. The NVT equilibration protocol was continued up to 200 ps of duration, in which the constraints acting on the proteins atoms were smoothly released (20 and 15 ps for $hERG_C$ and $hERG_O$, respectively, by keeping K^+ ions and the main chains frozen, 25 ps by keeping K^+ ions and the backbone frozen, and 50 ps by keeping K^+ ions and Cα atoms frozen), whereas position and distance restraints took place in the selectivity filter moiety. Finally, we switched to the NPT statistical ensemble, and the constraints left in the previous stage were converted in strong positional restraints. Again, the restraints (either of positional and distance kind) were gradually decreased along the last 300 ps of equilibration by using a similar approach to the previous one.

This extensive constrained and restrained equilibration for a total of 500 ps, was found to be necessary in order to deliver the selectivity filter in a reasonable geometry to the following production run. This behavior has already been attributed to the significant electrostatic forces arising from the series of carbonyls oxygen atoms which points together towards the pore axis[78]. With respect to the classical treatment of the five coordination sites, as suggested by Roux et al.[88], the partial charges on the carbonyl atoms were re-parameterized according to Tieleman et al. (namely by assigning the values of 0.60 and -0.60 to the point charges for the carbon and the oxygen, respectively)[89], in order to reproduce the correct interaction energy with potassium. The configurations of $hERG_C$ and $hERG_O$ at the end of the equilibration runs are showed in figure 3.14.

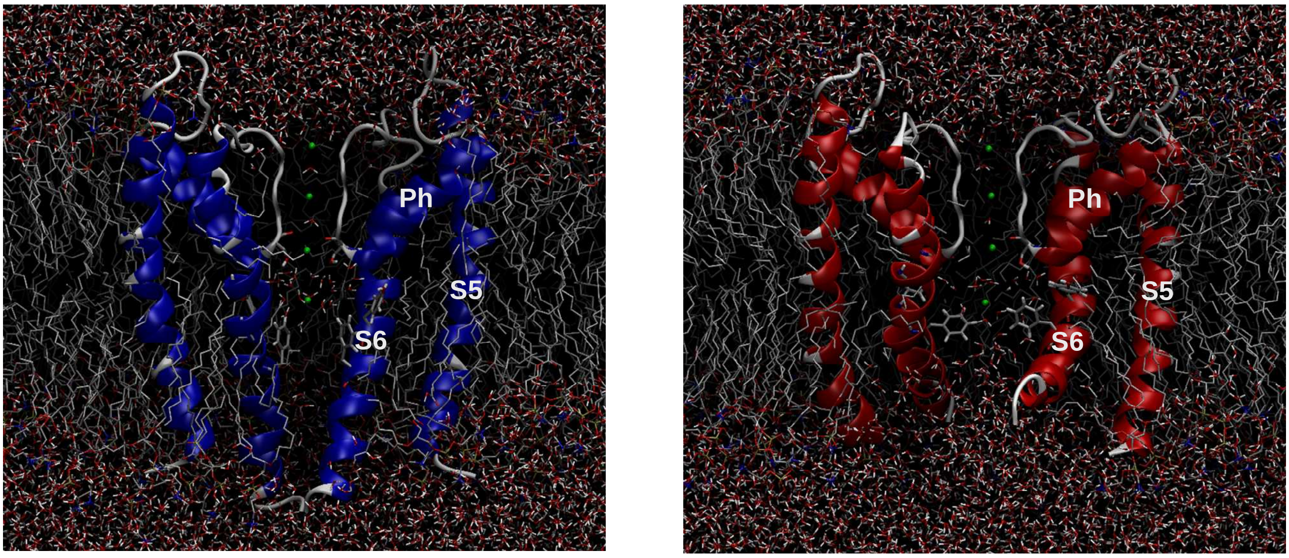


Fig. 3.14: Pictorial view of the channel models (closed to the *left*, and open to the *right*) at the end of the equilibration run.

The MD production run was performed in the NPT statistical ensemble for a total time of 5 ns. The system was completely free except for residual weak distance restraints of 5 kcal mol^{-1} $Å^{-2}$ to strengthen the H-bond network around the additional W_{SF} water.

The trajectory analysis was mainly carried out with GROMACS modules, while in order to coherently monitor the pore volume for both the closed and the open state of the models, an in house code was purposely developed (see further).

The dynamic behavior of the system was found to be very sensitive both to the starting configuration of the proteins and to the equilibration protocol. Many attempts were especially done in order to preserve the fold of the SF during the production run, still using the lowest amount of imposed restraints. This variability of the system was probably due to the series of four carbonyl oxygens pointing together towards the axis of the channel pore, as previously reported[78]. This led us to an optimized protocol showing the following system setup: 10101 configuration for the SF, an additional potassium ion located in the middle of the cavity, and a supplementary water molecule (W_{SF}) per subunit placed “behind” the residues forming the SF (Fig. 3.13). The H-bond network involving this W_{SF} and the neighbor residues was properly reproduced by applying some appropriate weak harmonic restraints. It must be noticed that the pivotal role of such water molecules was already reported for dynamics simulations of both crystal structures and homology models[82 – 84].

The stability of the system during the 5 ns of MD production run was assessed by evaluating the RMSD calculated as the difference between the position of each Cα atom in the outcome of the equilibration run and in every sampled conformation. Both the models reached their thermal stability after about 3.5 ns, getting in the last nanosecond to an average RMSD of 2.3 and 2.2 Å for the closed and open channels, respectively (Fig. 3.15).

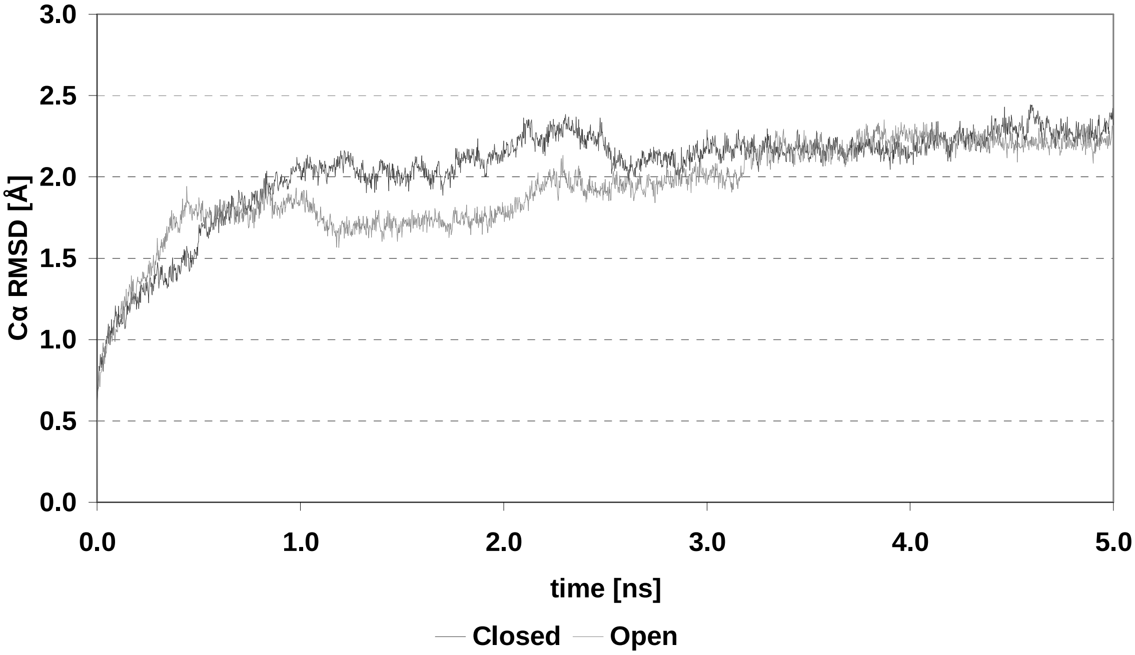


Fig. 3.15: Cα RMSD versus time for hERG_C (*dark*) and hERG_O (*light*).

Besides to the RMSD, the root mean squared fluctuation (RMSF) calculated over the Cα positions of each residue was also analyzed (Fig. 3.16). As expected, the RMSF value was small in the middle of the α-helices, ranging from about 0.4 to 1.5 Å, whilst a significant increase was observed at the C- and N-termini of each chain and also in the loop region. In particular, the greater flexibility showed by the extra-cellular loop (i.e., the truncated S5-Ph linker) can be due to the percentage of identity between the templates and the hERG channel lower in the truncated S5-Ph linker than in the helices and in the SF moieties.

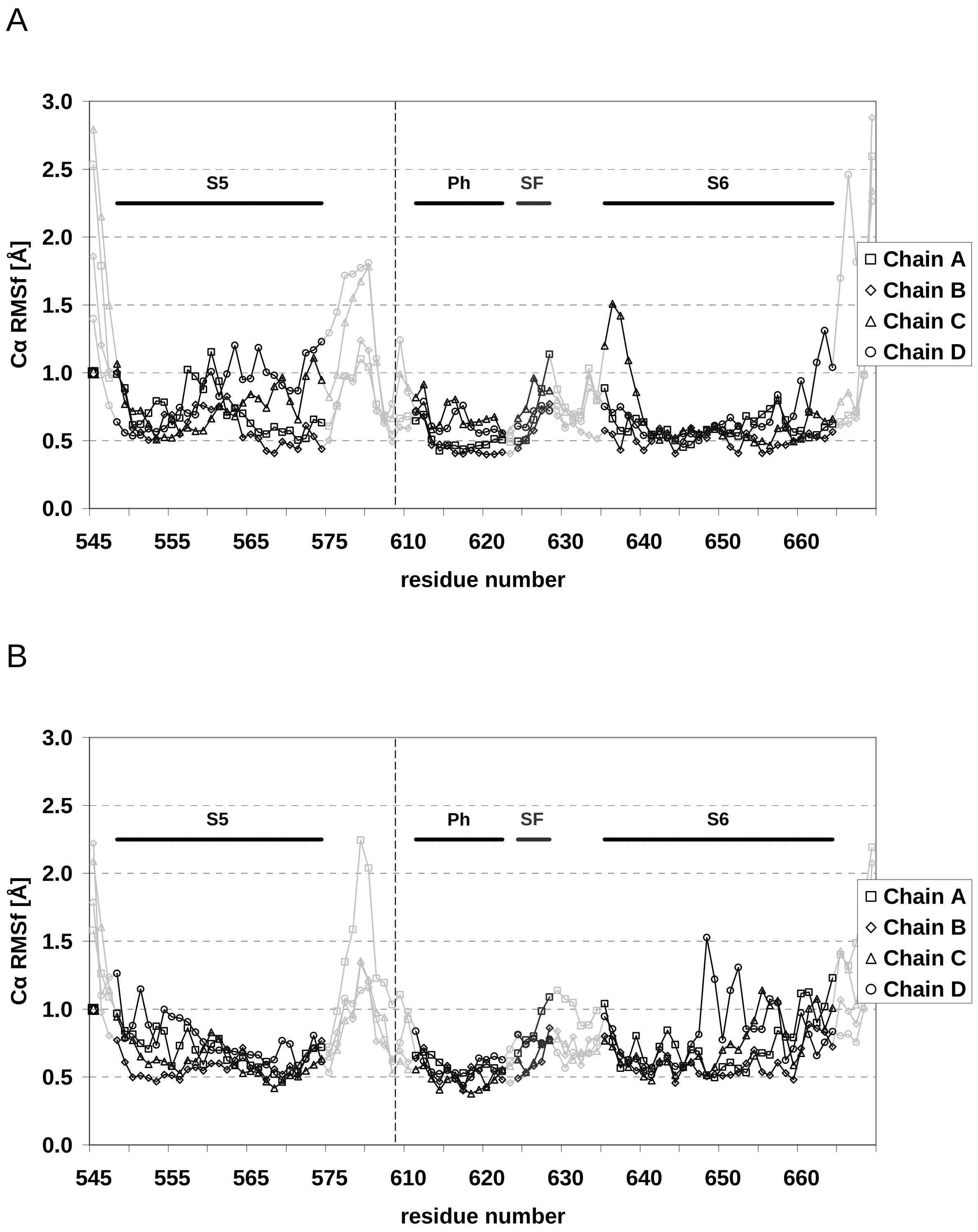


Fig. 3.16: Structural fluctuations as a function of residue position for hERG$_C$ (*A*) and hERG$_O$ (*B*) channel models. Data are averaged over all the 5 ns of MD. Helices are shown in black, while the SF is shown in dark grey. The vertical dashed lines indicates the gap in the models, which lack of the long S5-P linker (see Methods for details).

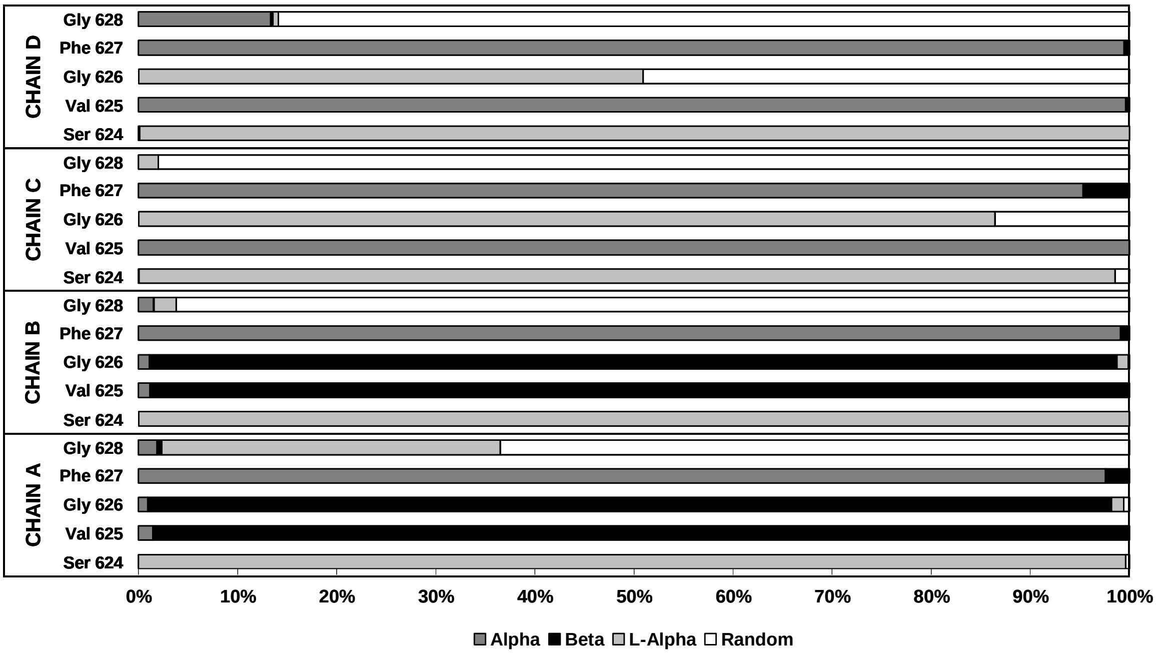
Since the pore is a crucial feature for a channel protein, a more in-depth examination of the dynamic behavior of that region was undertaken, and it is here summarized for both the SF and the cavity.

Selectivity Filter. Trial MD simulations performed both on the open and closed states of the hERG channel, showed that starting from the crystallographic ions coordinates led to very unstable systems. In particular, K_0 (followed by W_1) left its coordination site from the extracellular side at the very beginning of the equilibration, leading to an early SF unfolding during the subsequent production run. This prompted us to perform short and qualitative steered MD simulations, by pulling K_0 towards K_3 and bringing the former ion at a stable position. At the end of the steered dynamics, K_0 turned out to be properly stabilized by the four carbonyl oxygens belonging to Gly628 (see Figure 3.11) and by water molecules in the upper side of the coordination site (half a coordination shell). This behavior was already observed in MD simulation of the crystal structure of KcsA by Girardet et al.[76], and earlier by Roux et al.[79].

As many other MD studies performed on potassium channels pointed out, the SF underwent moderate changes in conformation during the simulations, supporting a flexible picture of the permeation mode. Furthermore, the intrinsic flexibility of the SF region has been proposed as a main feature for determining the ion selectivity in potassium channels[45, 93]. According to Sansom et al.[82] as an unspecific measure of the mobility of the SF, we monitored the value of the φ/ψ dihedral angle combination for the involved residues (namely, those ranging from Ser624 to Gly628, for each subunit) as a percentage over the simulation time (Fig. 3.17)

As it can be seen from the plot, the innermost residue, namely Ser624, is the most stable amino acid, as it spent about all of its time in the left-handed α-helix conformation. Val625 is also a quite stable residue, even though it adopted different conformations as regards both the different subunits and the gating state. Apparently, the three outermost amino acids of the SF were the most flexible ones, and among them the less stable turned out to be Gly626. It has been suggested that the SF could play a role in the unusual rapid inactivation kinetic of the hERG channel[27]. In particular, it is likely that a phenylalanine in the SF of hERG in place of a tyrosine (namely, the most conserved amino acid at the same position[46] for most K_v channels), could weaken the whole filter, as a consequence of a decrease of stabilizing H-bond interactions[99]. These results could corroborate this hypothesis, even though further studies should be undertaken.

A



B

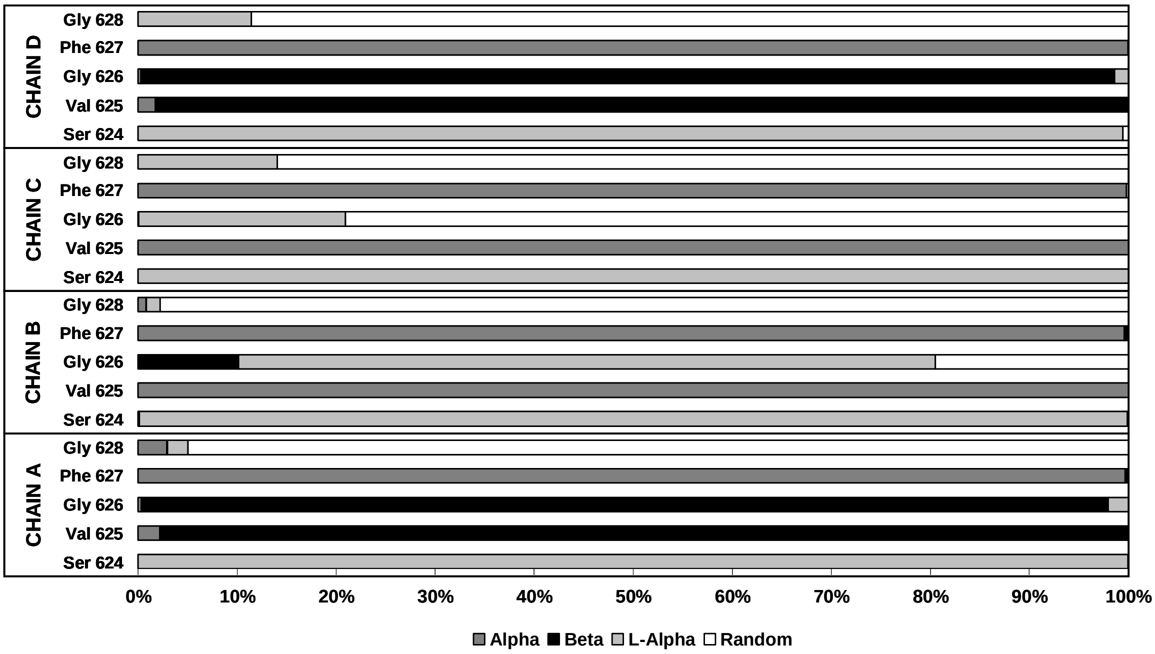


Fig. 3.17: Summary of the conformations adopted by the residues belonging to the SF for hERG$_C$ (*A*) and hERG$_O$ (*B*) models of the hERG channel, plotted as a percentage over the simulation time.

As often reported in the literature, some major changes in the conformation of residues in the SF occur during MD simulations[78, 83, 94, 95]. Here, the carbonyl flipping of Gly626 (in analogy to the glycine at the same position in simulation of other channels) was early observed in the simulations of both the closed and the open channel models. In particular, such a conformational change happened in the A and B chain in hERG$_C$, and only in the chain A in hERG$_O$, as the Val625 ψ and Gly626 φ angles plotted versus time in Figure 3.17 point out. This conformational change was responsible for a relevant motion of the SF lumen. The rotation around a backbone dihedral angle brought the carbonyl oxygen of Gly626 away from the SF lumen, while the hydrogen of the amide group established hydrogen bond interaction with a water molecule inside the SF (W$_3$). However, in contrast to studies where a temporary isomerization was observed, in the present hERG dynamics the above reported behavior seemed to be irreversible, that is, the affected amino acid did not flip back during the run. This is particularly clear for the hERG$_C$ simulation, where the SF early reached a non-conductive (or defunct) conformation. This could also be inferred from the trajectories of the potassium ions and water molecules located in the SF (Figure 3.19). Again, in the closed state model a non-physical outward switch in the occupational configuration was early observed (from 10101 to 01010 at about 700-900 ps), along with the exit of K$_0$ in the extracellular environment. Such a behavior has already been reported in the literature for closed state simulation of crystal structures[78, 96] and homology models as well[97], even when dealing with different force fields. It is therefore questionable that the reported behavior could be due to the quality of the homology models of the hERG channel. Actually, the carbonyl flipping has recently also been interpreted in terms of a sort of intrinsic gating of the SF[83, 98]. We do not intend to speculate on this aspect of the hERG channel as it is out of the scope of the present work. Conversely, in the hERG$_O$ model a greater stability of the SF configuration was paradoxically observed. While K$_{cav}$ left the channel cavity in the intracellular side in the first nanosecond of molecular dynamics as expected (Fig. 3.19 B), the remaining potassium ions were kept in their coordination sites for all the early stages of the simulation. Eventually, in our opinion, it is likely that the classical level of theory is not accurate enough to describe events happening in the SF moiety, where coordination effects having a non-neglecting quantum mechanical component, play a key role. These effects should be properly described by means of quantum chemical calculations that however cannot reasonably be carried out on homology models. In spite of this, it is generally accepted that the conformation of the SF does not affect the other regions of the channels, so that the analysis of the cavity can be both conceptually and practically separated from that of the SF.

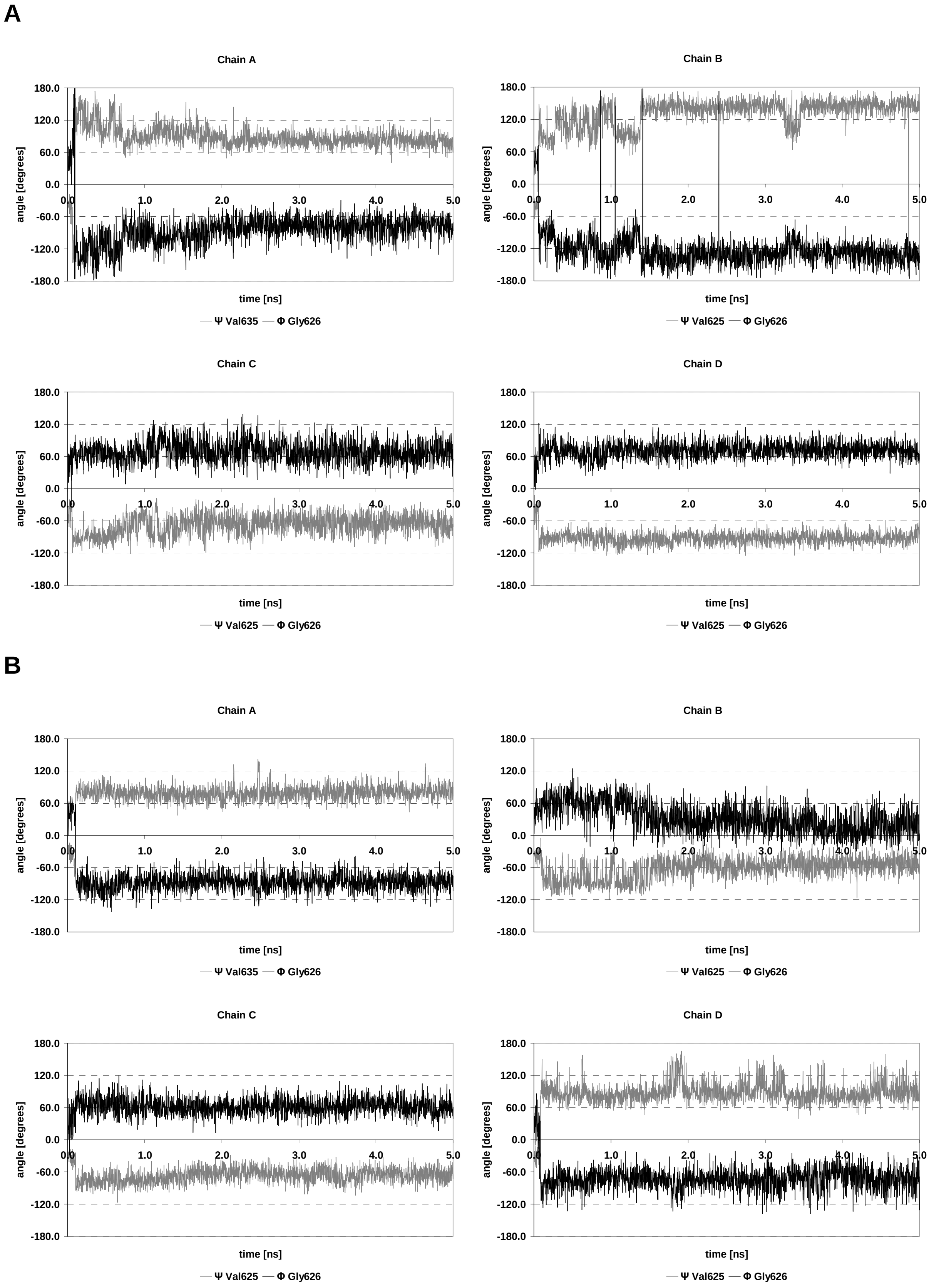


Fig. 3.18: Backbone torsion angles versus time for the closed (A) and open (B) models.

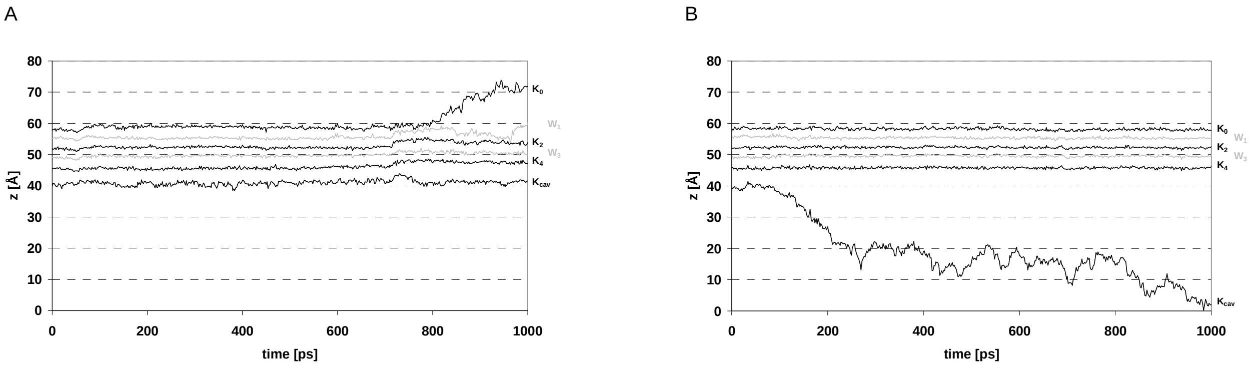


Fig. 3.19: Trajectories projected onto the z axis of K^+ ions (*black*) and water molecules (*gray*) belonging to the SF for the closed (*A*) and the open (*B*) channel models.

Cavity. In order to properly study the change in shape experienced by the portion of the cavity of the channels which define the drug binding site, a suitable in house code was developed to analyze the trajectory.

Dynamical accessible volume: The program performs a numeric calculation of the accessible dynamical volume of a user defined region.

The algorithm could be briefly described as follows. The accessible volume is calculated by building a three dimensional Cartesian grid (having an adequate resolution) inside a (*internal*) cylinder whose shape (height and radius) is determined by the active site definitions provided by the user.

In particular, the height and the radius of the *internal* cylinder are calculated from the C_α coordinates of the selected amino acids, so that the potential volume of the region in study is allowed to dynamically change its shape along the MD trajectory. The program is interfaced with a shell script which extracts the snapshots from the trajectory at a given frequency as pdb files, and converts the protein from the GROMOS-87 extended-atoms representation to an AMBER all-atoms description[90]. This is directly made by supporting the pdb2pqr driver[91], and it turned out to be

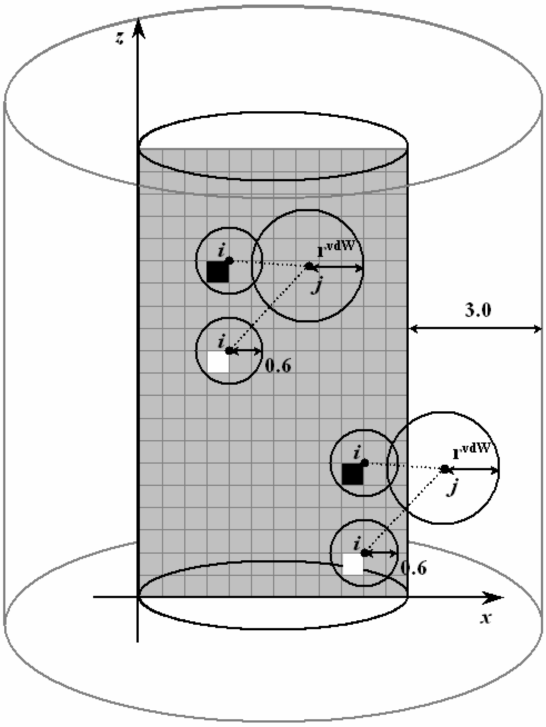


Fig. 3.20: Pictorial representation of the code used for the numerical calculation of the dynamical accessible volume of the binding sites. For the sake of clarity, only a *zx* projection of the three-dimensional Cartesian grid is shown.

a fundamental trick in order to properly assess the accessible volume of the monitored region. Thus, for every snapshot, the distance between each *i* grid node and all the *j* protein atoms located inside the potential cylinder are calculated. Each volume element whose vertices correspond to the *i* node is thus judged to be empty if the following condition occurs:

$$d_{ij} \geq (r_j^{vdW} + 0.6) \qquad [3.13]$$

where r_j^{vdW} and 0.6 stands for the *j*-esim atom and for the smallest van der Waals radius according to the AMBER parm99 parameterization[90] of the hydrogen element, respectively. All the atoms located outside the *internal* cylinder but whose van der Waals sphere protrudes inside of it, where taken into account by extending the above procedure to an additional layer of 3 Å per dimension (*external* cylinder which is coaxial to the *internal* one). This value was chosen because it is bigger than the largest van der Waals radius of any atom type normally occurring in proteins. In this way, summing over all the empty volume elements, we were able to calculate the maximum volume accessible from the smallest atom in nature (i.e. the hydrogen), in other words the volume assessed was always over-estimated.

As it can be easily recognized, provided the previous algorithm, in order to accurately assess the accessible volume for any given cavity two main requirements have to be fulfilled,:

1. within the numerical approximation, the algorithm must be able to reasonably reproduce the analytical value of volume for any given cylinder having dimensions comparable to those of the investigated region;
2. the axis of the cylinder must always be coherently oriented in respect to the protein, whose absolute orientation (in a general way) is allowed to change along the trajectory.

Actually, the first condition is not problematic since the resolution of the grid can be chosen arbitrarily high in order to satisfy the requisite, without any significant increase in computational time. Conversely, the second condition is somewhat more tricky, and it should be addressed in a rigorously way by dealing with the Euler angles for a rigid body and relative matrix transformation, or – even better – by means of the quaternion parameters for generalized coordinates. Here, the problem was solved in a simpler manner exploiting polar coordinates, and the fact that the cylinder shape bears a rotational axis which actually reduces one degree of freedom. Hence, for each snapshot, the protein is translated to the axis origin, which corresponds to the center of the highest face of the

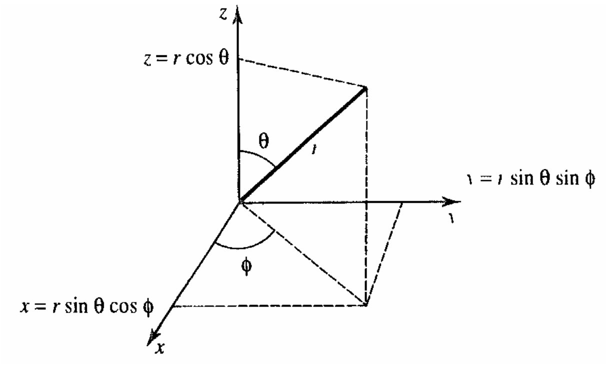


Fig. 3.21: Relationships between polar and Cartesian coordinates.

cylinder, and it is rotated along θ and ϕ in such a way that the axis of the cavity (which in our case fortuitously corresponds to the axis of the channel too) corresponds to the *z* Cartesian axis.

By taking advantage of the previously reported code, we first monitored the volume of the whole cavity of both channels. However, for such an investigation, to monitor the volume of the whole inner cavity was not informative for hERG$_O$, as slight variations in the drug binding pocket could roughly be identified. Therefore, we focused on the putative binding site for the blockers, which is delimited in the upper side from the selectivity filter (Thr624 ring) and downwards from the Phe656 ring, extended of an arbitrary value in order to avoid boundary effects. Hence, our internal cylinder definitions were the following:

- *upper side*: <Cα:Ser624> ring;
- *lower side*: <Cα:Phe656> ring;
- *radial side*: max{<Cα:Tyr652> ring, <Cα:Phe656> ring}.

Where brackets denote spatial average calculated over the four subunits, separately for each Cartesian dimension. Such conditions are schematically summarized in figure 3.22.

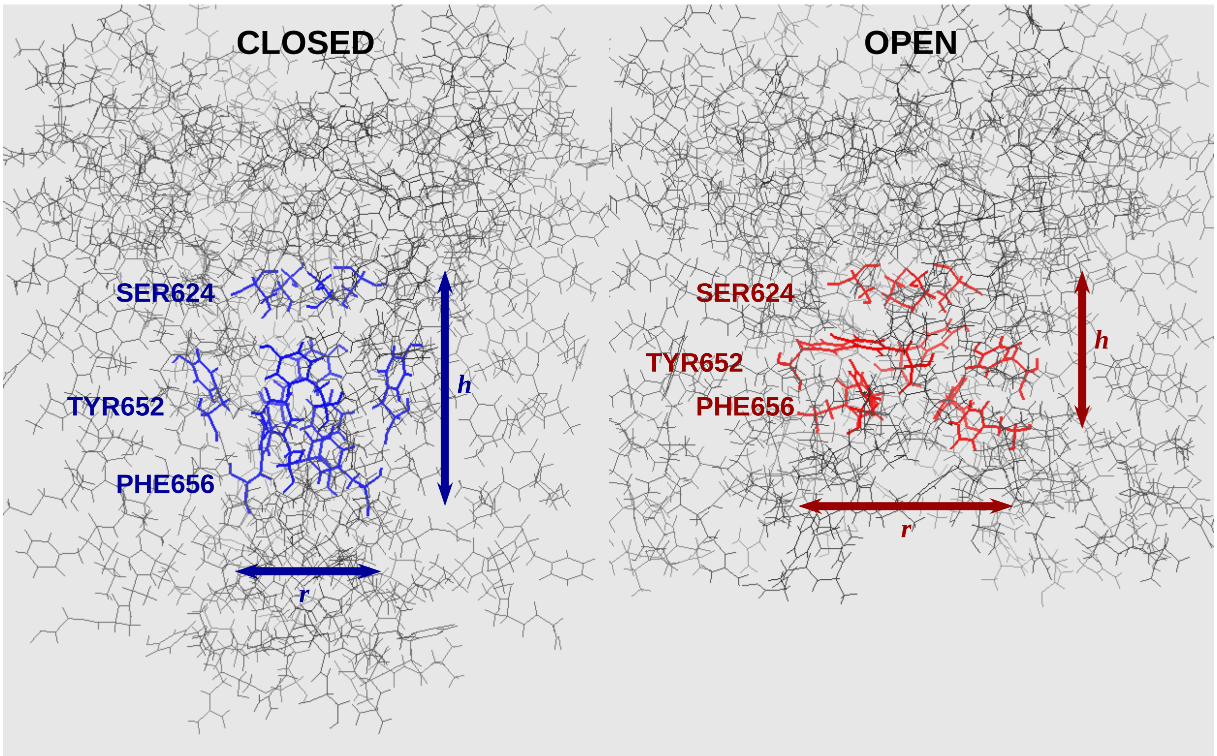


Fig. 3.22: Schematic representation of the calculation of both the radius and the height for the internal cylinder, where the Cartesian grid will be built.

The volume of the active site monitored along the whole simulation time is shown in Figure 3.23, whereas in Figure 3.25 a pictorial view both of the internal cylinder and the accessible volume are shown.

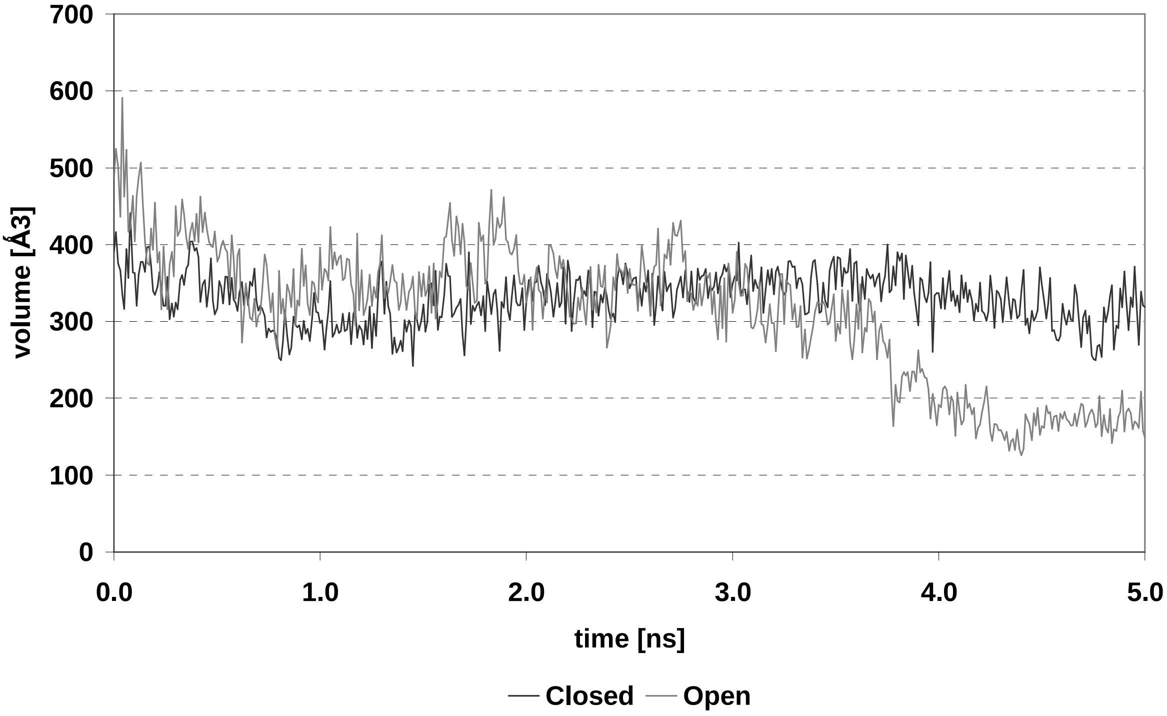


Fig. 3.23: Volume of the active site moiety for hERG$_C$ (dark grey) and hERG$_O$ (light grey) channel models versus time.

As expected, at the beginning of the production run the active site of the open model of the hERG channel is considerably larger than that of the closed one (of about 100 Å^3). Apparently, during the MD, while the volume of the active site of hERG$_C$ was approximately constant, the volume of hERG$_O$ quickly decreased reaching the value of the closed state model after about 2 ns. In the last nanosecond of MD simulations, the putative drug binding site of the open model irreversibly got smaller than the one of the closed state. This behavior could be easily explained considering the different shape of the cavity in the two models, and the type of amino acids which face the pore in the binding site. For the closed model, a mainly constant volume was not surprising. Actually, since the water molecules (along with K$_{cav}$) trapped inside the channel were not allowed to leave the pore from the intracellular side, they exerted an essentially constant pressure on the surrounding residues. In contrast, in hERG$_O$ water molecules were free to diffuse in the intracellular water bulk. Actually, the active site volume remained close to its original value, until K$_{cav}$ was present in the cavity. As soon as this potassium ion and its first coordination shell left the protein environment (which happened approximately after 200 ps of production run, see Figure 3.17), the volume of the binding site suddenly decreased (compare to Figure 3.23). Moreover, the partial desolvation of the cavity allowed four Phe626 to “hydrophobically collapse” giving rise to a further contraction of the cavity. To better show this phenomenon we monitored the distance between the geometric centers of the benzene moiety of the side chain of each couple of Phe626 residue (Figure 3.25).

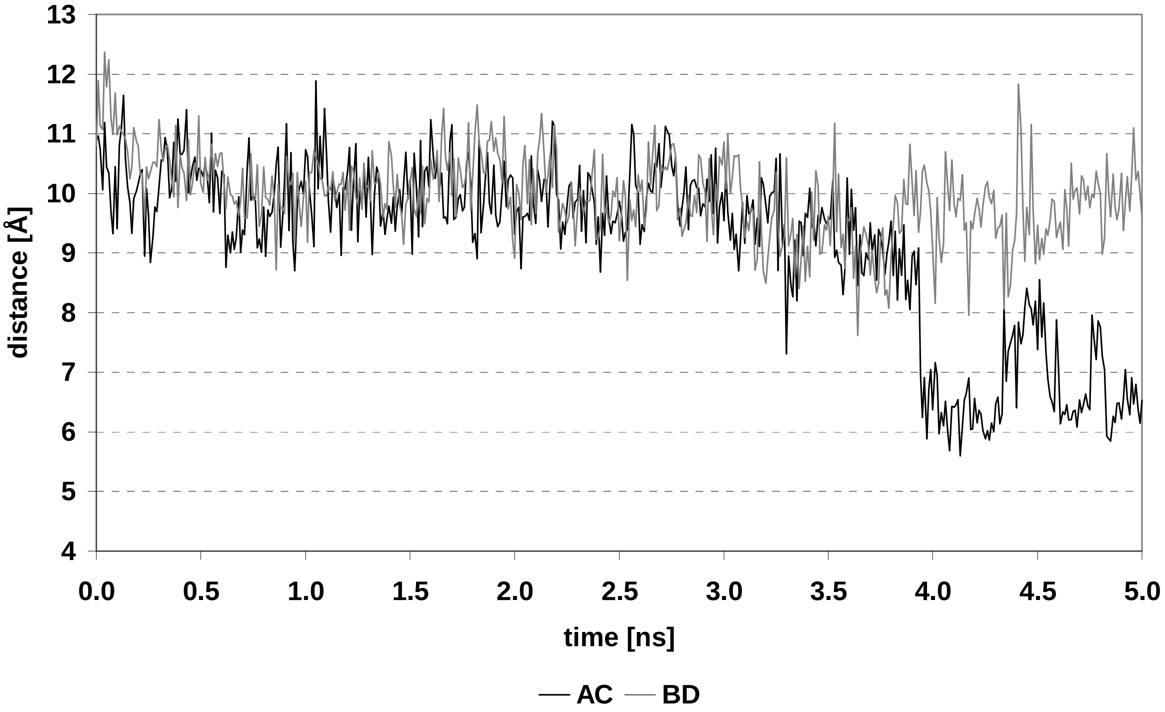


Fig. 3.24: Distance between the geometric centre of aromatic ring of the amino acid Phe626 belonging to two opposite chains: AC (*black*) and BD (*grey*).

From the plot it is clear that the distance between the geometric centers of the amino acids belonging to the opposite chains AC decreases abruptly at about 3900 ps, excluding the cross interaction between the subunits BD (light gray plot in figure 3.22). It is quite relevant to notice from the plot of figure 3.22, that while hERG$_O$ Phe656 side chains initially relaxed as a consequence of the thermalization of the homology-built channel, at ~ 4 ns of MD, when the system showed a quite stable dynamic behavior in terms of overall Cα RMSD (figure 3.15), the hERG$_O$ cavity started to suddenly collapse leading to the above-reported “closure”.

Since the drug binding site is supposed to be mainly located in the channel cavity, we believe that the above-described dynamic behavior could strongly influence the outcome of docking simulations that intend to describe the binding mode of drugs to hERG. In the light of these considerations, we subsequently investigated the role of the side chains cavity dynamics (induced fit effects) in the docking simulations.

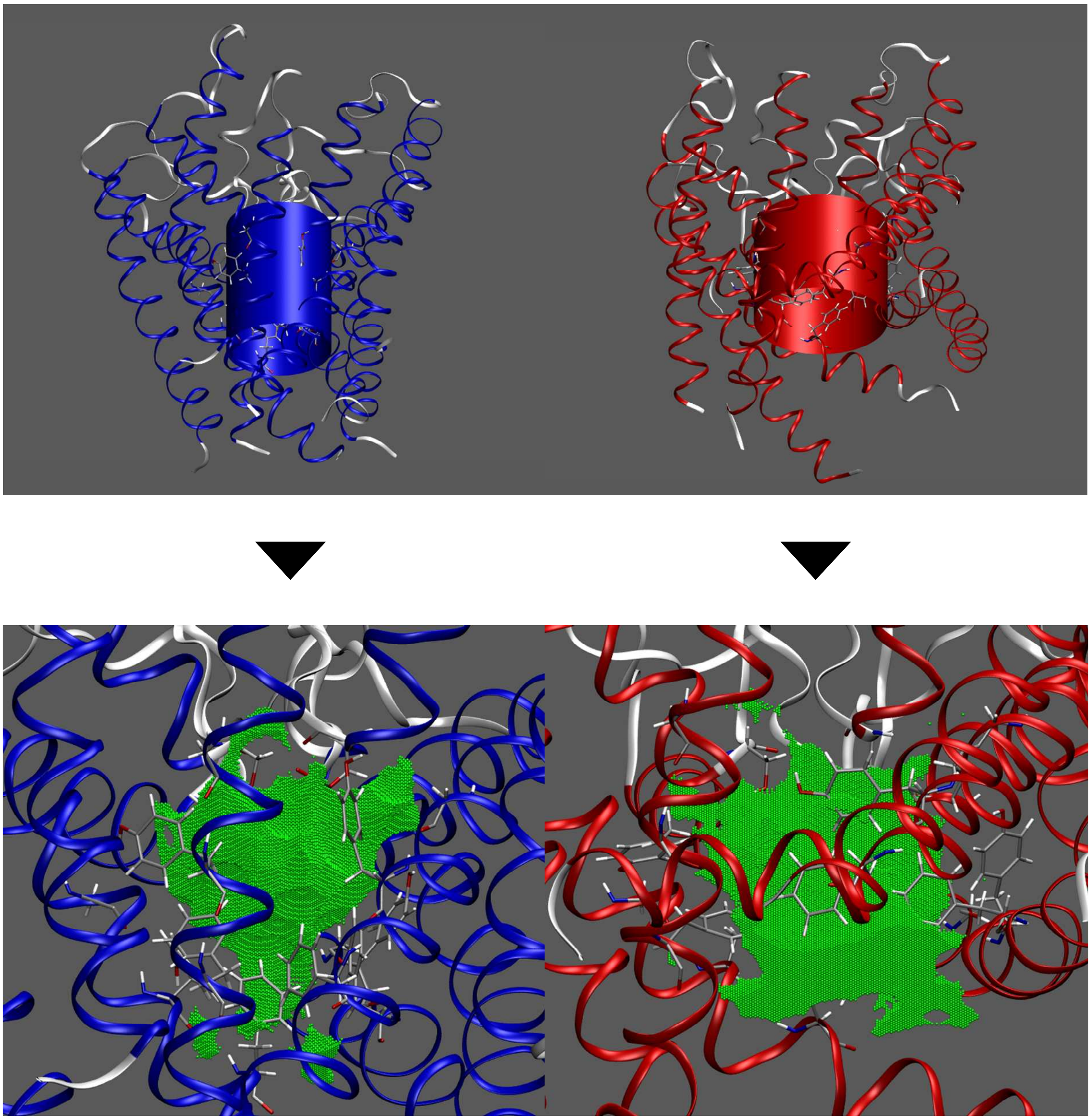


Fig. 3.25: Pictorial view of the construction of the internal cylinder (*up*) and the calculated accessible volume (*down*) for the binding site for both the closed (*left*) and open (*red*) channel models, at the beginning of the production run.

3.1.2.3 Docking simulations

Docking simulations were carried out in order to identify the most suited hERG cavity conformation for studying channel-drug complexes. To this end the most potent channel blocker so far known, Astemizole (IC_{50} = 0.9 nM on I_{Kr} current of human embryonic kidney, HEK, cells[100]), was exploited as a molecular probe. The geometric center calculated between the four Tyr652 and the four Phe656 was used as coordinates for the binding site origin, whereas active site radius was set equal to 15 Å. As suggested by the GOLD authors[92], genetic algorithm default parameters were set: population size was 100, selection pressure was 1.1, number of operation was 10^5, number of islands was 5, niche size was 2, migrate was 10, mutate was 95, and crossover was 95. In order to better model the binding site features, the GOLD scoring function was used.

Astemizole was then docked at both hERG channel models (i.e., open and closed). At the $hERG_C$ we were unable to find a reliable binding pose within the channel cavity. This was not a surprising result in the light of the smaller accessible volume of the channel cavity (Figure 3.21). Actually, $hERG_C$ model showed a rather crowded cavity lumen because of the orientation of both S6 backbone atoms and Tyr652 and Phe656 side chains. This is in line with a similar result reported by Rajamani et al. for the docking of ligands to a closed hERG model (based on the KcsA template) [49], and is also consistent with experimental data that point to the hERG inactive state as the conformation suitable for the drug binding.

Docking simulations were then carried out with $hERG_O$ either using the crude homology model ($hERG_O^0$), or snapshots from MD runs. While Astemizole could also bind in somehow at $hERG_O^0$, we could identify a reasonable binding mode only when using snapshots from MD simulations. As previously in-depth described, the most reasonable docking pose for Astemizole was that obtained with the MD snapshot at 150 ps (average GOLD Score: 62.26). In Figure 3.27, a docking complex between the MD snapshot at 150 ps and Astemizole is reported. In particular, the benzimidazole ring favorably interacts both with Tyr652 and Phe656 of the same subunit (D) by means of a parallel-displaced and a parallel-stacked π-π interaction, respectively. A parallel-displaced π-π interaction is also involved between the p-fluorobenzene ring of Astemizole and the Tyr652 residue of the next right-hand side subunit (C), and the possibility of a hydrogen bond between the ligand fluorine atom and Ser624 of subunit (B) can be identified. Finally, the positively charged nitrogen belonging to the piperidine ring clearly interacts by means of a cation-π interaction with Phe656 of the hidden subunit in front of the observation view (A). Many furhter van der Waals contacts between the molecule and the protein can not be ruled out in the energy stabilization of the complex. Conversely, the p-methoxybenzene ring does not seem to specifically interact with the protein, and it remains rather exposed to the solvent. The binding mode thus described is in line with the current hypothesis on the determinants of hERG-drug binding based on extensive mutagenesis and assessment of the sensitivity of mutant channels to the action of drugs, which, among others, proposes as critical residues for the ligand binding Tyr652 and Phe656 (through cation-π and hydrophobic interactions, respectively), and Ser624 as well.

The Astemizole-$hERG_O^0$ binary complex provided a model in sharp contrast with the above mentioned hypothesis based on experimental evidence, in particular with regard to the interactions with Tyr652 and Phe656. Actually, what seemed to cause an unfeasible pose of the ligand was the position of the aromatic side chains of the deemed critical residues, in particular the tyrosines 562. We reasoned that in this situation, the channel cavity was not in a proper rotamer configuration to allow a reasonable prediction of a physically meaningful drug binding mode. Therefore, we employed MD simulations to sample the conformational space within the cavity, and used MD snapshots as random conformations for the docking experiments. Only in this way, we were able to sort out binding site configurations where the drug could interact with the side chains of the two key amino acids in a way that agreed well with the experiments regarding the involvement of such residues. Notably, we obtained this result in an unbiased manner (i.e., without any "manual" intervention on the position of the side chains), but simply by probing different snapshots with the ligand. The drug-channel complex configurations detected in this way (Figure 15) might then be considered as taking into account the induced-fit caused by the ligand binding to the protein.[85, 86]

In all the complexes of Figure 15, the side chains of Tyr652 and Phe656 take part in some way in the binding of the drug, even if a frank cation-π interaction is seldom detected. In this regard, it is remarkable that, Farid et al.[30] following a careful docking procedure of several ligands on an open hERG model could not obtain any evidence of cation-π interactions between the hERG cavity and several drug molecules. However, this should not rule out the involvement of Tyr652 in the ligand binding, but, as the authors thoughtfully proposed, might indicate an indirect effect of the mutation on drug binding linked to a change in the cavity structure. This reasoning then leads to an interpretation of the role of both Tyr652 and Phe656 not so strictly bound to a specific physicochemical effect invariant for all kind of ligands. The same authors state: "The specific configuration of such residues in the open channel creates opportunities for multiple simultaneous ring stacking and hydrophobic interactions that can be achieved in multiple binding conformations/orientations". We agree, and think that only a proper exploration (through MD sampling) of the "specific configurations" of the Tyr652 and Phe656 side chains may allow the identification of binding opportunities for the ligands.

As a final corollary observation on the usefulness of carrying out MD simulations before performing the docking procedure, we note that in this way there is also the possibility to identify the proper simulation interval wherein to perform the docking experiments. In fact, as shown in Figure 15, where the average GOLD score is plotted versus the simulation time, only $hERG_O$ conformations in the first 3.5 ns of MD were suitable for docking studies. For instance, that time coincided with the simulation time at which the collapse of the inner cavity occurred (see Fig. 12), and, correspondingly, the average GOLD score dropped to a value of -290 (Fig. 15), and docking complexes could not be identified any longer.

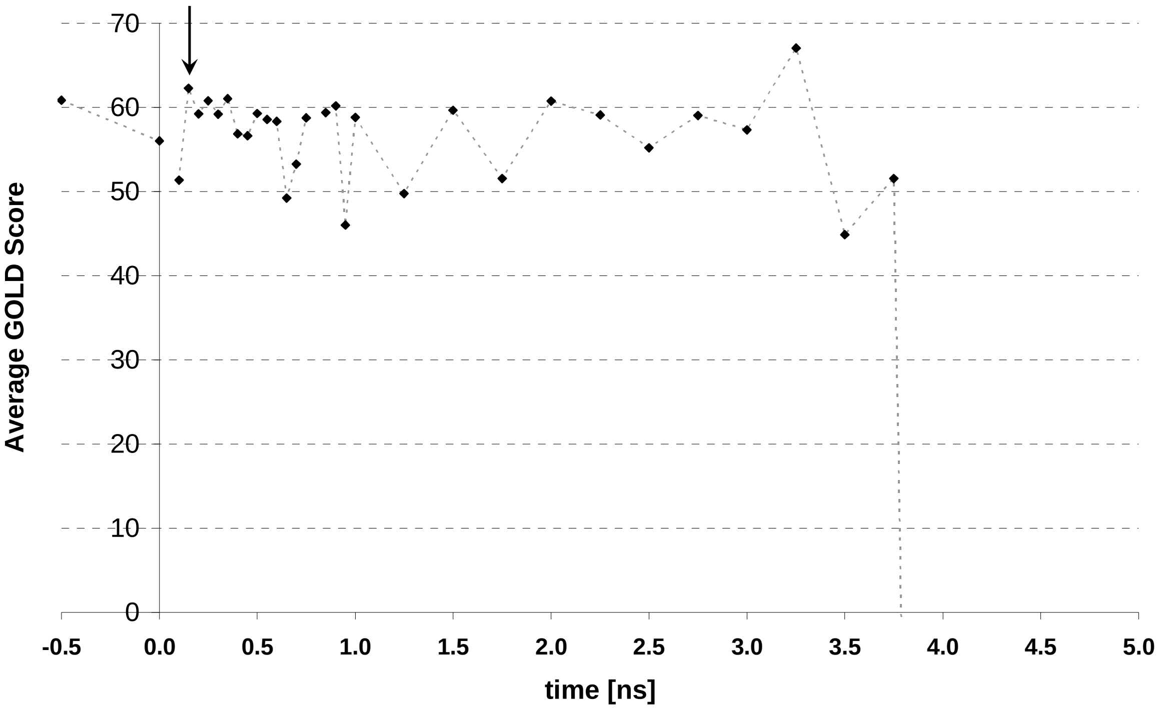


Fig. 3.26: Average GOLD Score plotted against simulation time. The sampling frequency is 50 ps^{-1} during the first nanosecond of dynamics, and 250 ps^{-1} for the remaining trajectory. Only a meaningful portion of the *y* axis is shown, namely just positive values for the GOLD scoring function are plotted. The dashed line is discontinuous in two points, namely at 50 and 800 ps, where the optimization procedure was not able to locate meaningful minima. The result of the docking procedure performed on the crude output of the homology modeling (hERG$_O$^{0}) is arbitrarily plotted at the fictitious time of -0.50 ps, for comparison. The arrow highlights the selected channel conformation.

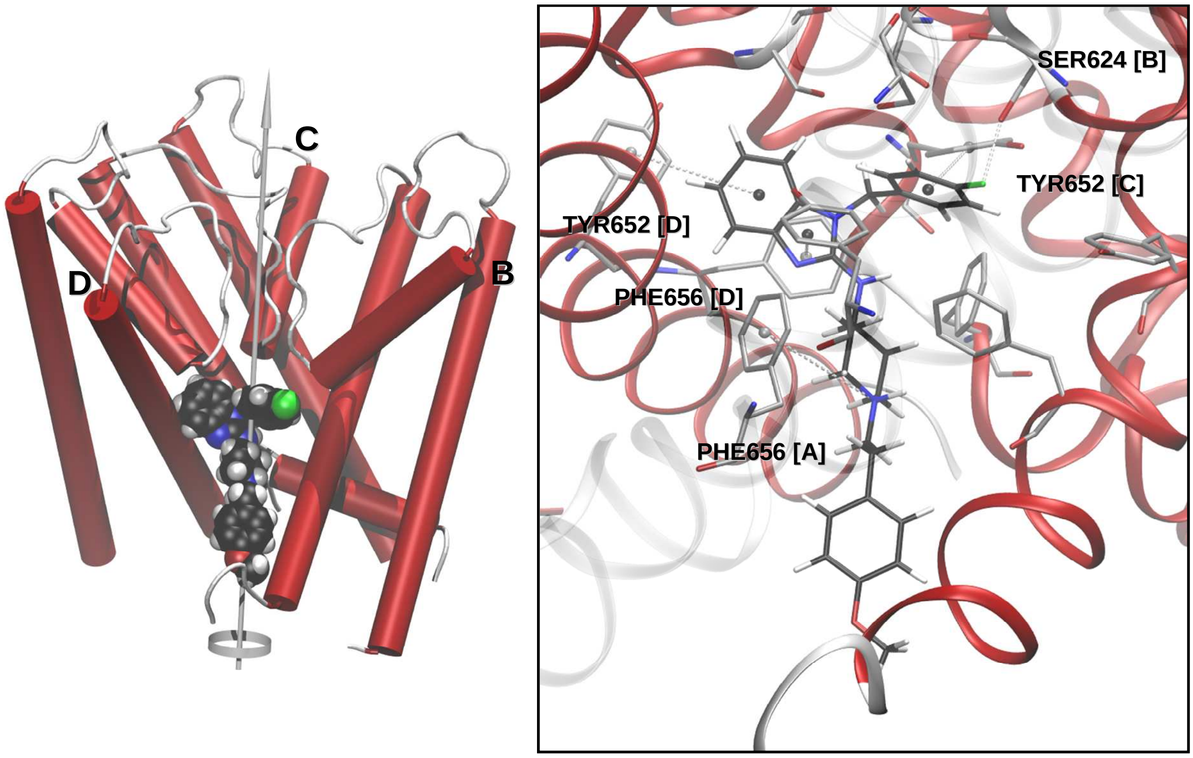


Fig. 3.27: A solution of the docking of astemizole in the hERG$_O$ configuration corresponding to the snapshot of 150 ps. Details of the interactions are reported in the text. For clarity, just three out of four subunits are explicitly shown.

3.2 Metadynamics as a post-docking tool

3.2.1 The protocol

It has been demonstrated, and nowadays it is widely accepted, that docking methods provide a collection of putative binding modes, rather than supply a single univocal solution[102, 103 and references therein]. The above reported behavior is sometimes referred to as the so called *docking problem*.

Typical docking programs commonly employ a two phase approach:

1. Posing: configurations of the ligand within the binding site are generated by means of diverse searching techniques;
2. Scoring: the poses identified in the previous step are ranked in terms of the outcome (score) of a suitable – usually energetic – potential function.

Of course the two phases are strictly connected, since the score associated to a particular pose will bias the search towards the lowest local minima, basing upon the rules of the chosen search algorithm. Anyway, it is well known that scoring functions bear a limited reliability, especially when entropic effects (primarily salvation/desolvation phenomena) play a non negligible role in stabilizing the protein-ligand complex. Moreover, it has been demonstrated that even if the experimental configuration can be generally reproduced (when effects as induced fit and the presence of structural waters do not take place), the corresponding pose is seldom found among the top score ranked[102]. However, in order to partially overcome such a problem a geometrical cluster analysis approach has already been suggested[102, 103]. Actually, it has been demonstrated that the poses which at the best reproduce the experimental configuration, judged in terms of their RMSD calculated on the heavy atoms, are very often located within the most populated clusters[102], [103]. Nevertheless, a more robust approach would be advisable in order to unambiguously discriminate the correct binding mode among different equally populated clusters.

Metadynamics is a non-equilibrium sampling technique which is both able to accelerate rare events and to reconstruct the free-energy changes associated to the selected reaction coordinate, provided a proper set of few collective variables is available which coarsely describe the investigated event (see chapter 2.3.1). Such a method is general, and it has been successfully applied in several different fields, ranging from chemistry and material science, to crystal structure prediction and biophysics. Furthermore, the potentiality of metadynamics as a flexible-docking procedure has been investigated for simple ligands as well[104, 105]. In particular, starting the simulation with the ligand in solution, not only the reproduction of the crystallographic binding mode was achieved, but the reconstructed FES also provided insights in respect to the energetically preferred binding path[104,

[105]. Nonetheless, the main drawback of such simulations relied in their computational demand, determined by the huge complexity and by the high amount of degrees of freedom which affected the microscopic dynamics when applied to pharmaceutically relevant systems. The basic idea of this study is to exploit the straightforward ability of the posing algorithms of docking programs, combined to a robust geometrical cluster analysis, to quickly provide a crude set of binding modes, which act as a multiple guess for the subsequent slower but more accurate metadynamics simulation. In such a way, the ligand in solution as a starting point for a metadynamics run is no longer needed, and the sampling would be fastened since only the un-docking event would be actually investigated. Summarizing, we propose metadynamics as a methodological post-docking tool to be used in place of scoring functions.

Technically, the protocol consists in ideally splitting the study of the protein-ligand complex into two stages as it is schematically summarized in Figure 3.28. As it can be seen, only the posing algorithm of the docking program is explicitly exploited. Moreover, as it serves only to provide binding mode guesses, the choice of the docking program is not of a pivotal importance for the success of the methodology. Instead, the robustness of the selected cluster analysis method is crucial, since it will judge and select poses, which will be studied by means of the following metadynamics simulation.

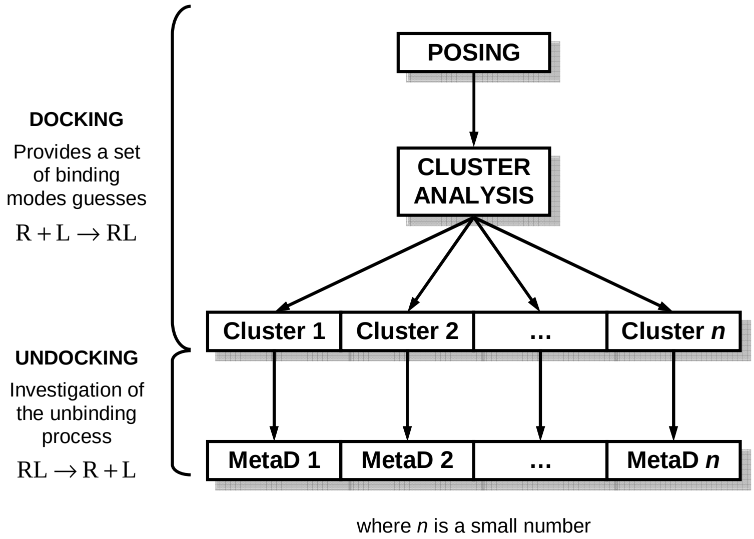


Fig. 3.28: Schematic overview for the proposed approach.

The proposed methodological protocol has been tested on a limited set of crystallographic complexes and artificially docked (non-crystallographic) complexes as well (Table 3.2) with the aim to investigate accuracy, efficiency and the potential usefulness in the pharmaceutical computational area. On the basis of the previously reported considerations, for each system the docked poses were generated by means of the AutoDock 3.0.5 software[106], which is nowadays one of the most theoretically crudest docking program, since it considers the receptor as a rigid body, and uses both for the protein and the ligand an united-atom description. The search was then performed by means of the Lamarckian genetic algorithm combined with default parameters. For the protein Kollman-UA charges were used, whereas for the ligand RESP charges were derived from the *ab initio* electrostatic potential calculated at the HF/6-31G(d)//HF/6-31G(d) level of theory. In order to carry out the most unbiased general approach, any structural water was systematically removed, and for each complex a total number of 100 poses were generated. Conversely, since the clustering procedure is of primary importance for the success of the methodology, the program AClAP 1.0[103] was used, given that it has been demonstrated to provide a robust and reproducible (in respect to different input sorting) clustering without any subjective involvement. In particular, the hierarchic-agglomerative clustering method was used along with the average-linkage rule. Besides, the partition of the poses which at best fulfills the highest intra-cluster homogeneity joined together with the lowest number of clusters (cutting rule), was assessed by means of the KGS penalty function. Moreover, a preliminary clusterability assessment was performed with the Hopkins's H*-test, and the *a posteriori* cluster significance was estimated in terms of relative population assessed by the Chauvenet criterion. Finally, atomic equivalences were imposed for symmetry reasons when necessary. The continuous direct version of metadynamics, as it is at the moment implemented in the software for molecular dynamics simulations ORAC 4.0, was used along with the asynchronous parallelization method of the multiple walkers. Both the parm94 and parm99 versions of the AMBER force filed were used.

The above reported protocol was applied to the dataset summarized in Table 3.2. Such a dataset was opportunely split in two classes: set A, which would represent somehow "standard" complexes, while in set B a "difficult" case is taken into account.

Tab. 3.2: Dataset used to test the procedure. PDB codes marked with a * are those at the moment converged.

set A	**set B**
1Q5K* 3ERT* 1AGW Mixed system: 1Q5K protein + non-crystallographic ligand (pirazole [1.5-b] pyridazinic scaffold) [abandoned]	1HVR

The choice of collective variables was driven by the following requirements:

1. they should be general, in order to provide a wide extension of applicability of the protocol;
2. they should be small in number, in order to improve the efficiency of the search.

The original idea was to test the feasibility of the use of a couple of collective variables, no matter how the complexity or the shape of the binding site and/or the flexibility of the ligand. The selected variables were:

- the modulus of the distance calculated between the center of mass of selected amino acids located close to the binding site, and the centre of mass of the ligand itself;
- the angle vector between the centre of mass of selected amino acids located close to the binding site and the major inertia axis of the ligand.

However, it was early clear that such an approach would be only able to describe simpler systems, namely 1Q5K, 3ERT and 1AGW. In order to properly treat an inter-conversion in a bond having a partial character of double bond of the ligand belonging to the mixed system (Table 3.2), a third collective variable sampling along such a dihedral angle had to be added. Analogously, it has been realized that no realistic undocking would be possible for 1HVR, unless a proper collective variable describing correlated slow motions of the protein was taken into account. For this reason a third variable describing the motion of the C_α of the protein projected along the direction of the slowest principal mode was added. In order to identify the direction of such a motion, a principal component analysis calculated on the C_α positions (essential dynamics analysis) was performed on a trajectory of a preliminary NVT dynamics achieved by means of the GROMACS 3.2 software[65, 65]. In particular the preliminary dynamics was carried out until a satisfactory decrease below the threshold value of 0.1 of the cosine content in the projection over time of the displacement for the eigenvector associated to the first eigenvalue, was obtained. The convergence was reached in 5 ns of production run, and the dynamics was then stopped at 6 ns of duration.

For all the complexes belonging to the set A, metadynamics was performed on the representative poses (namely those closer to the geometric centre of the respective cluster) of the two most populated clusters, once having checked whether the crystallographic one was present within them. Conversely for 1HVR just a single pose was studied as a consequence of a univocal solution of the docking program. Obviously, before starting metadynamics, all the complexes must reach both thermal and pressure equilibria by means of a short equilibration performed in the NVT and NPT statistical ensembles, respectively.

Among all the simulations performed on the dataset, the most elegant and informative solution was achieved for the complex 1Q5K, and thus it will be in depth discussed in the following paragraph.

3.1.2 1Q5K: a case study

It must be stressed that the training set was primarily chosen based on the interest of each system in terms of the application of the proposed methodology, rather than for the pharmaceutical or biological importance of the respective target *per se*.

The PDB code 1Q5K corresponds to the protein glycogen synthase kinase 3β complexed with the nanomolar inhibitor AR-A014418 (K_i = 38 nM)[107]. In particular, GSK3 represents a serine/threonine kinase which is involved in the phosphorylation of tau protein, and its activity in adult brain has been directly linked to several of the key neuropathological mechanisms lying at the basis of the Alzheimer's disease[107]. The ligand, N-(4-methoxybenzyl)-N'-(5-nitro-1,3-thiazol-2-yl)urea, is an ATP-competitive inhibitor, which possesses a high selectivity in respect to CDK2 and CDK5 kinases (IC50 > 100 μM), and its principal effect relies in the inhibition of neural degeneration mediated by deposition of β-amyloid peptide[107].

The protocol introduced in the previous paragraph was applied in order to obtain a total number of 100 of poses. The outcome of the docking procedure was optimally partitioned by the cluster analysis program into 20 clusters (the KGS penalty function is plotted in Figure 3.29), and the most populated are reported in Table 3.3.

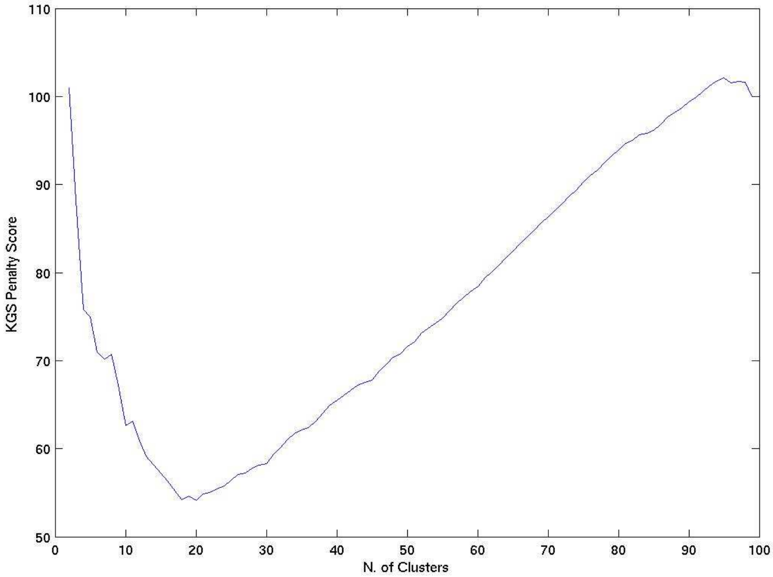


Fig. 3.29: KGS penalty function. The minimum corresponding to the value of about 20 clusters, which points out the best partition of the 100 input poses, is clearry visible.

Tab. 3.3: Results of the cluster analysis performed on the poses generated by the docking routine. The first line refers to the Hopkins test, revealing a border line clusterability of the elements, whereas columns are reported as they appear in the outcome of the program. In particular, from left to right it is shown the number of cluster, its cardinality (namely the population), the number of the representative pose, the distance of such a pose with respect to geometric center of the cluster, the RMSD of the representative pose with respect to the crystallographic configuration and the significance of the cluster, respectively.

H* = 0.62

MOST POPULATED CLUSTERS

CLUSTER #	CARD	REPR	CEN.DI	RMSD	CHAU
5	9	36	2.123	8.135	no
11	10	57	4.849	8.790	no
14	19	44	11.196	3.354	no
15	27	7	13.600	1.350	yes

As it can be inferred from Table 3.3 in the column *RMSD*, the configuration which more closely reproduces the crystallographic one is pose number 27, belonging to the cluster number 15, which is at the same time the most populated one (27 elements), and the only significantly populated according to the Chauvenet criterion. In this particular case, the combined approach posing/cluster analysis turned out to be clear enough to discriminate the real binding mode, conversely pose 27 was not present in the top ranked results of the scoring function (data not shown). Nevertheless, as previously stated, a further confirm is always advisable, hence the original aim of the work was to study by means of metadynamics the undocking process for all the poses reported in Table 3.3. Actually, during the standard dynamics equilibration, the pose number 44 – representative of the cluster 14 – almost converged in a

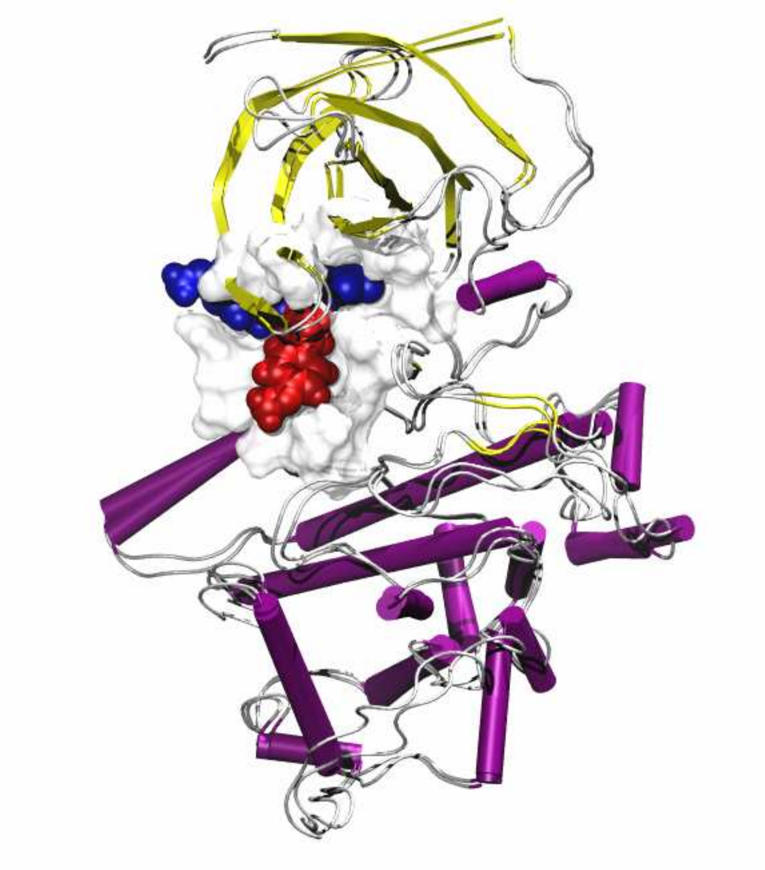


Fig. 3.30: Superposition of the complexes arising from pose 7 (blue nd 36 (red), as they appear at the end of the equilibration MD run.

configuration similar to the crystal structure, whereas pose number 57 – representative to the cluster 11 – assumed a particularly unfavorable orientation leading to an unrealistic binding mode, hence it was discarded by visual inspection in order to save computational resources. Hence, only pose number 7 and pose number 36 were finally studied by means of metadynamics.

Metadynamics was performed until reaching a reasonable convergence on the reconstructed FES, by using the collective variables of the distance and the angle vector (described in the previous paragraph) and by keeping restrained the Cα atoms of amino acids ranging from 105 to 350 (by using an harmonic potential having a force constant of 100 kcal mol^{-1} Å^{-2}), in order to reduce at the most the degrees of freedom for the microscopic dynamics, while allowing the protein conformational changes needed for the undocking of the ligand from the binding site. The most important parameters used for the simulation were the followings:

▪ n° walkers:		range 3 – 6;
▪ Gaussian height:		ω = 0.4 kJ/mol;
▪ Deposition time:		τ = 1000 fs × replica;
▪ Gaussian width:	CV distance:	δs = 0.4 Å;
	CV angle:	δs = 0.13 rad.

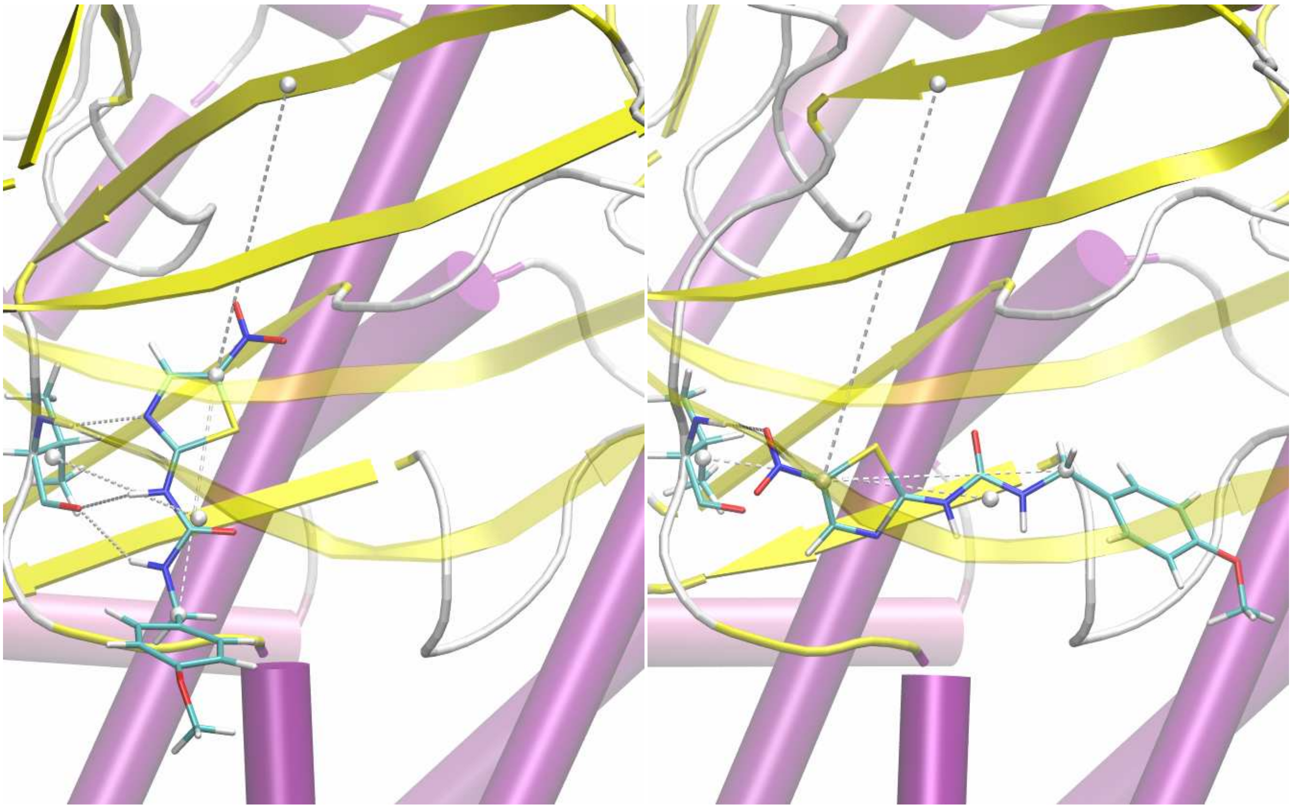


Fig. 3.31: Binding modes for the pose number 7 (*left*) and the pose number 36 (*right*) at the end of the MD equilibration run. Hydrogen bonds between the ligand and the key amino acid residues are explicitly shown as long as the collective variables used in the Metadynamics simulation.

The starting coordinates in the space of the collective variables were (6.13 Å, 0.67 rad) and (8.85 Å, 1.82 rad) for the pose number 7 and 36, respectively. A focus on the binding mode for both the complexes is reported in figure 3.31. As it can be seen, the ligand in pose 7 favourably interacts with both the carbonyl and the amide of the key residue (Val101) by means of three hydrogen bonds, whereas in pose 36, just only a hydrogen bond involving the nitro group and the amide of the same key residue is involved.

The evolution of the sampling along the collective variables can be analyzed by monitoring the evolution of the reconstructed FES plotted along the time course as Figure 3.32 points out, as it should be stressed that in metadynamics time has no longer its physical meaning.

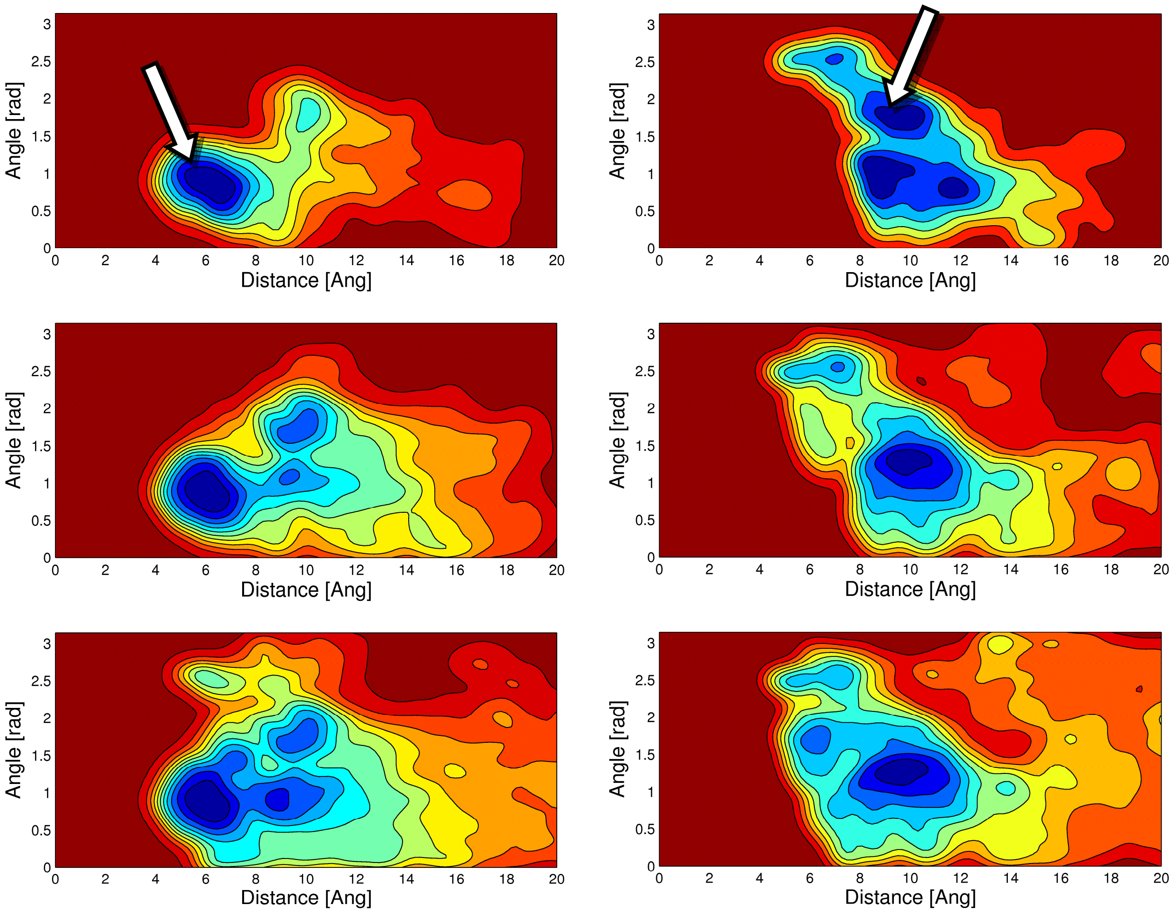


Fig. 3.32: Evolution of the reconstructed FES for the undocking process of pose number 7 (left), and pose number 36 (right). The free energy surface is shown as iso-contours calculated with a step size of 2 kcal/mol. The white arrows indicates the starting configuration in the space of the collective variables.

As it can be noticed from Figure 3.32, the reconstructed FES is striking similar in a distance ranging from 10 to 20 Å, suggesting a similar undocking pathway for both the poses. This result would be reasonable only in the assumption of an occurring interchange between the poses. In order to corroborate such hypothesis, a more in depth investigation and characterization of the main free-energy basin was undertaken (Figure 3.33). Thus, representative configurations were sampled along the metadynamics trajectory within an energy window of 1 kcal from the local minimum of each basin highlighted in Figure 3.33. The collection of such configurations is reported in Figure 3.34.

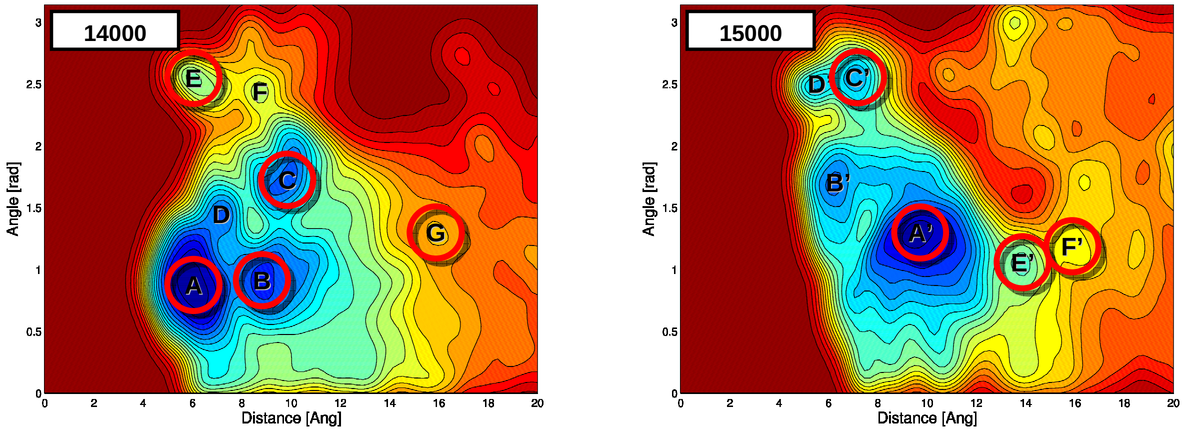


Fig 3.33: Reconstructed FES for the undocking process of pose 7 (left) and pose 36 (right) at the end of metadynamics, that is after the deposition of 14000 and 15000 Gaussian functions. Iso-contours are calculated by means of a step size of 1 kcal/mol; the main informative basins are highlighted with a red circle.

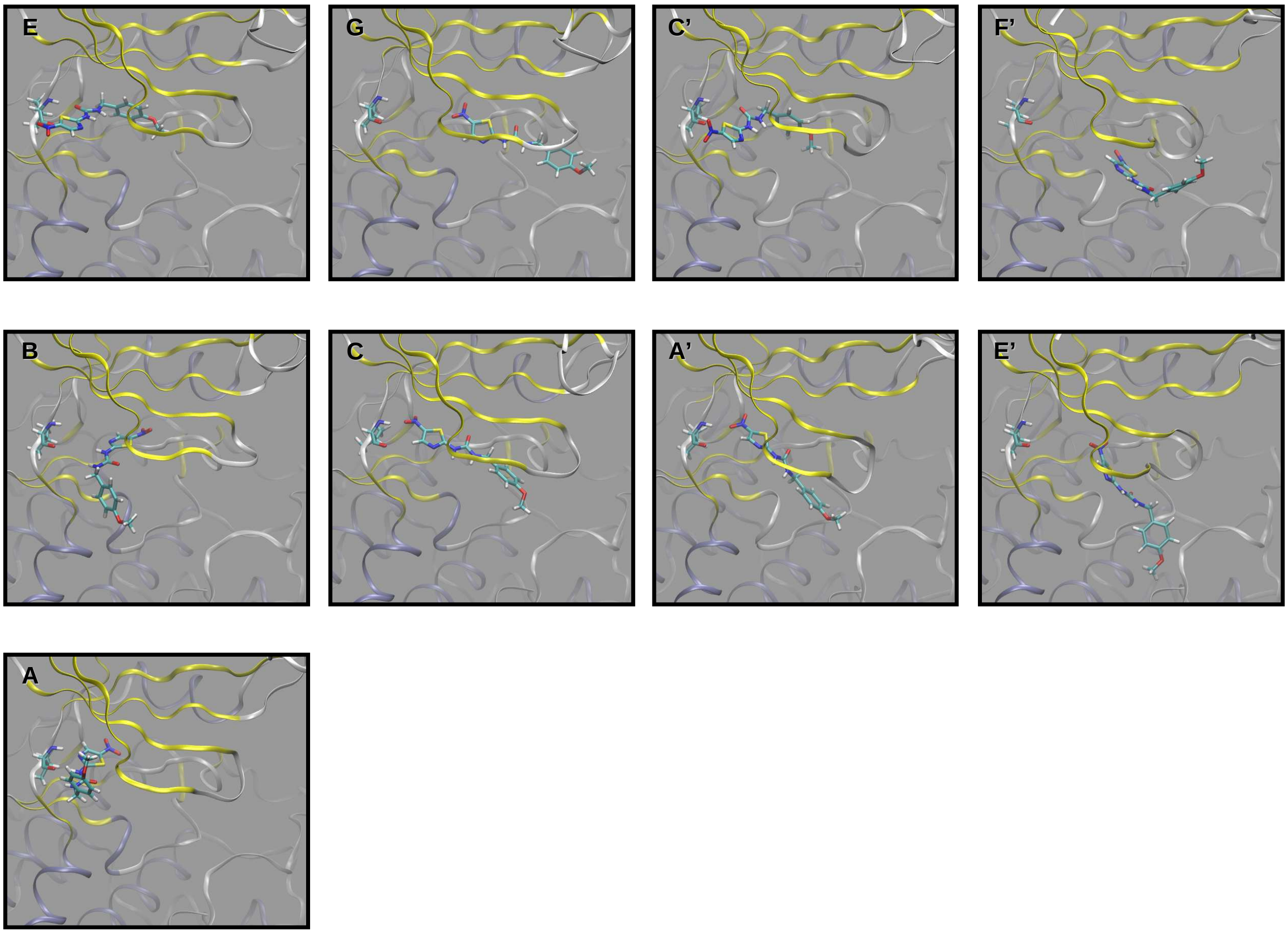


Fig. 3.34: Representative configurations for each highlighted basin reported in Figure 3.33.

The sampling will be analyzed first discussing the most important (and well characterized) configurations belonging to the undocking pathway for pose 7 (denoted with capital letters), and then configurations for pose 36 (denoted with primed capital letters plus a prime) will be taken into account as well.

Pose 7 – Undocking pathway.

A. It represents the deepest basin, and the representative configuration corresponds to the crystallographic binding mode;

B. It represents a metastable configuration: the ligand is already significantly separated from the key aminoacid of about 3 Å, by keeping almost the same angular orientation compared to the crystallographic binding mode;

C. This is an extremely important configuration, since it closely reproduces the starting configuration of the metadynamics simulation for pose number 36 (A'), hence supporting the above assumption about the interchanging undocking path. As it can be seen the centre of mass of the ligand is almost at the same distance for basin B, but the small molecule notably changed the angular orientation of its principal inertia axis, by pointing the amino group moiety directly towards the key amino acid;

E. It represents a non effective basin as regards the undocking pathway. Nevertheless, it is important to notice the metadynamics capability to find energetically meaningful orientations for the ligand. Interestingly enough, it should be noticed that basin E can be roughly considered as a reflection of basin A along the angle vector collective variable;

G. Within the approximation of a (relatively) quick converging FES, it represents the basin that has been considered corresponding to the undocked configuration for the ligand. Here the ligand is separated from the key amino acid of about 10 Å, while keeping a similar angular orientation compared with that of the configuration representative for basin C.

Pose 36 – Undocking pathway.

A'. It represents the deepest basin for the undocking process of pose 36, and the presented configuration corresponds to the starting one. As it can be noticed from the plot in Figure 3.33, basin A' is quite large and this is reflected into a non-univocal representative pose, hence in Figure 3.34, the most recurring configuration of the ligand is actually shown. Such a behaviour could be imputed either to inaccuracies or to non-convergence of the reconstructed FES;

C’. It represents a basin homologous to basin E for the undocking of pose 7. Again, metadynamics was able to widely explore the FES and to identify energetically meaningful configurations which would be *a priori* hardly foreseen;

E’. It represents a configuration closely related to the undocked one, which is represented by basin F’. A similar microscopic configuration can be also appreciated along the metadynamics trajectory for pose number 7, even if a real basin can not be actually identified. Once more, a missing convergence along the reconstructed FES could be imputed in order to explain such a discrepancy;

G’. Such a configuration represents the basin which is supposed to provide a description of the undocked ligand. As it can be seen from figure 3.34, the representative configuration is quite similar to that for the homologous basin G for the undocking process of pose 7.

Once basins have been qualitatively identified and characterized, the next step relies on the quantitative comparison between the reconstructed FES between the investigated poses. In order to facilitate such an evaluation, the free-energy was projected along the collective variable of the distance, since this is the most informative coordinate to characterize the investigated event among the chosen CVs. The plots are reported in Figure 3.35.

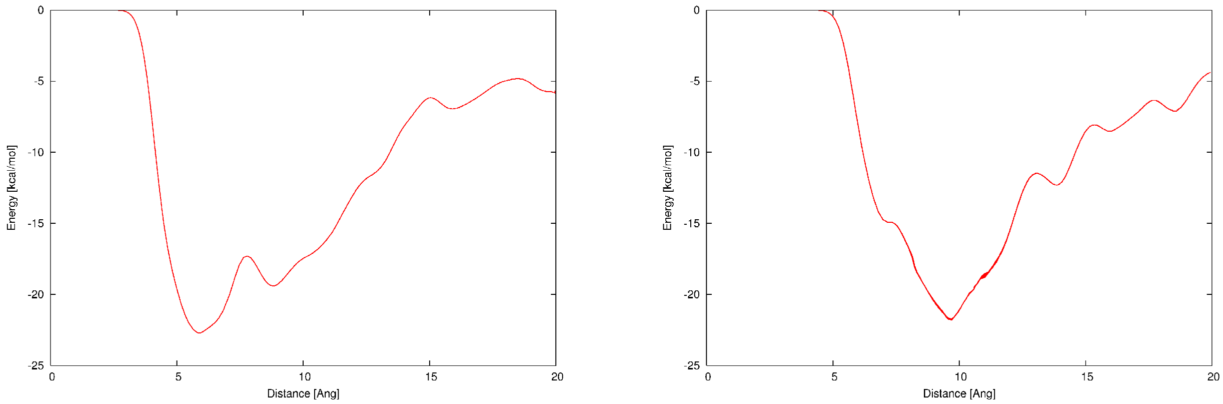


Fig. 3.35: Comparison between the reconstructed FES projected along the collective variable of the distance, for pose number 7 (*left*) and number 36 (*right*)..

The reconstruction of the FES led to a value of 16 and 13 kcal/mol for the undocking process A → G and A’ → F’ for pose 7 and 36, respectively. As it can be noticed, metadynamics was able to unambiguously discriminate the configuration which at best reproduces the crystallographic binding mode, providing some insight as regards to the dynamics of the event investigated as well. It is worth mentioning that the calculated values bear the same order of magnitude of the experimental datum, even if a significative difference can be appreciated between the absolute values. Such a dissimilarity can be imputed either to an insufficient convergence of sampling or to inaccuracies arising from the use of a force field which, actually, plays a similar role than that of the scoring functions in standard docking procedures.

Finally, some considerations on the behavior of metadynamics in such a particular case study are also matter of interest. As it has been previously discussed, during the undocking process for the complex arising from pose number 7 (namely, the "correct" pose with respect to the crystallographic binding mode), the starting configuration for the undocking simulation of pose number 36 (namely, the "wrong" pose) has been found in an unbiased manner. It is important to stress that such a configuration is associated for both the metadynamics simulations to a deep basin (in an ideally converged metadynamics the relative well compared to the respective undocked basin should be perfectly equal), hence it represents an energetically meaningful configuration along the space of collective variables. Such a configuration provide an important link between the two simulations, pointing out an interchange (7 → 36) between the starting poses. In other words, the so called "wrong" pose was not an artifact provided by the posing algorithm: it actually represents a possible binding mode, even if it is not the most favorable in terms of interactions, and hence in terms of energetics. Furthermore, it is possible to speculate that such "wrong" configuration provides the first match between the ligand and the key amino acid of the target during the dynamical docking process. Then, the complex would optimize its favorable contacts leading to the definitive "correct" configuration.

On the other hand, given the possibility of such an interchange between the two original configurations, the reader could wonder why the "correct" pose (7) can not be found during the metadynamics simulation of the "wrong" one (36). In fact, the 36 → 7 interchange should be observed as well, hence its absence from the simulation can be imputed to an artifact occurring in the metadynamics run. Actually, by analyzing the trajectories of the various walkers for the metadynamics simulation performed on the complex starting from the pose number 36, many configurations bearing an orientation very similar to that of pose number 7 can be visually identified. Such configurations differ from pose number 7 for the presence of some (ranging from 1 to 2) water molecules "trapped" between the ligand and the key amino acid which did not allow to achieve the configuration reproducing the crystallographic binding mode. In the light of this result, two consideration can be made. First, in the limit of a very long metadynamics simulation run such water molecules would be ejected from the binding site, and the “correct” binding mode should be reached as well. The second consideration directly follows from the previous one: even if the undocking event is accelerated via the used collective variables, the finding of alternative binding configurations is not. Hence, for such an event (which is closely related to the undocking, but actually slightly differs from it) a collective variable is hidden, thus the event is not accelerated along that direction path. Work is still in progress in order to avoid this annoying behaviour, nevertheless two strategies (not necessarily excluding each other) appear to be promising. The most obvious is to take into account a proper third collective variable, such as the water coordination number for the ligand. The second is to improve the microscopic phase space sampling by means of a parallel tempering procedure, namely a method in which different replicas of the systems – each in a different thermodynamic state (usually, but not necessarily such states differ in temperature) – are simultaneously handled in order to further improve the speed of the sampling and the crossing of high energetic barriers.

3.3 Bibliography

[1] van der Ploeg, P.; Berendsen, H. J. C. Molecular dynamics simulations of a bilayer membrane. *J Chem Phys* (1982), **76**: 3271.

[2] Tieleman, D. P.; Marrink, S. J.; Berendsen, H. J. C. A computer perspective of membranes: molecular dynamics studies of lipid bilayer systems. *Biochim et Biophys Acta* (1997), **1331**: 235.

[3] Benz, R. W.; Castro-Román, F.; Tobias, D. J.; White, S. H. Experimental validation of molecular dynamics simulations of lipid bilayers: a new approach. *Biophys J* (2005), **88**: 805.

[4] Chiu, S.-W.; Clark, M.; Balaji, V.; Subramaniam, S.; Scott, H. L.; Jakobsson, E. Incorporation of surface tension into molecular dynamics simulation of an interface: a fluid phase lipid bilayer memnbrane. *Biophys J* (1995), **69**: 1230.

[5] Zhang, Y.; Feller, S. E. Brooks, B. R.; Pastor, R. W. Computer simulation of liquid/liquid interfaces. I. Theory and application to octane/water. *J Chem Phys* (1995), **103**: 10252.

[6] Feller, S. E.; Zhang, Y., Pastor, R. Computer simulation of liquid/liquid interfaces. II. Surface tension-area dependence of a bilayer and monolayer. *J Chem Phys* (1995), **103**: 10267.

[7] Tu, K; Tobias, D. J.; Klein, M. L. Constant pressure and temperature molecular dynamics of a fully hydrated liquid crystal phase dipalmitoylphosphatidylcholine bilayer. *Biophys J* (1995), **69**: 2558.

[8] Tu, K.; Tobias, D. J.; Blasie, J. K.; Klein, M. Molecular dynamics investigation of the structure of a fully hydrated gel-phase dipalmitoylphosphatidilcholine bilayer. *Biophys J* (1996), **70**: 595.

[9] Tieleman, D. P.; H. J. C. Berendsen, Molecular dynamics simulations of a fully hydrated dipalmitoylphosphatidylcholine bilayer with different macroscopic boundary conditions and parameters. *J Chem Phys* (1996), **105**: 4871.

[10] Roux, B.; Commentary: Surface tension of biomembranes. *Biophys J* (1996), **71**: 1346.

[11] Jähnig, F. What is the surface tension of a lipid bilayer membrane? *Biophys J* (1996), **71**: 1348.

[12] Feller, S. E.; Pastor, R. W. On simulating lipid bilayers with applied surface tension: periodic boundary conditions and undulations. *Biophys J* (1996), **71**: 1350.

[13] Essmann, U.; Berkowitz, M. L.; Dynamical properties of phospholipid bilayers from computer simulation. *Biophys J* (1999), **76**: 2081.

[14] Marrink, S. J.; Mark, A. E. Effect of undulations on surface tension in simulated bilayers. *J Phys Chem B* (2001), **105**: 6122.

[15] Forrest, L. R.; Sansom, M. S. P.; Membrane simulations: bigger and better? *Curr Opin Struct Biol* (2000), **10**: 174.

[16] Patra, M.; Karttunen, M.; Hyvönen, M. T.; Falck, E.; Lindqvist, P.; Vattulainen, I. Molecular dynamics simulations of lipid bilayers: major artifacts due to truncating electrostatic interactions. *Biophys J* (2003), **84**: 3636.

[17] Seelig, A.; Seelig, J. The dynamic of fatty acyl chains in a phospholipid bilayer measured by deuterium magnetic resonance. *Biochem* (1974); **13**: 4839.

[18] Nagle, J. F.; Tristram-Nagle, S. Lipid bilayer structure. *Curr Op Struct Biol* (2000); **10**: 474.

[19] Nagle, J. F.; Tristram-Nagle, S. Structure of lipid bilayers. *Biochim et Biophys Acta* (2000); **1469**: 159.

[20] Tristram-Nagle, S.; Nagle, J. F. Lipid bilayers: thermodynamics, structure, fluctuations, and interactions. *Chem Phys Lip* (2004); **127**: 3.

[21] Gutman, G. A.; (et al); Wymore, R. A. International union of pharmacology. XLI. Compendium of voltage-gated ion channels: potassium channels. Pharmacological Reviews. (2003); 55: 583 – 586.

[22] M.C. Trudeau, J.W. Warmke, B. Ganetzky, G.A. Robertson. *Science* **269** (1995) 92.

[23] Sanguinetti, M. C.; Jiang, C.; Curran, M. E.; Keating, M. T.; A mechanicistic link between an inherited and an acquired cardiac arrhythmia: HERG encodes the I_{Kr} potassium channel. *Cell*. (1995); **81**: 299 – 307

[24] Vandenberg, J. I.; Walker, B. D.; Campbell, T. J.; HERG K^+ channels: friend and foe. *Trends Pharmacol Sci*. (2001); **22**: 240 – 246.

[25] Recanatini, M.; Poluzzi, E.; Masetti, M.; Cavalli, A.; De Ponti, F.; QT prolongation through hERG channel: current knowledge and strategies for the early prediction durino drug developement. *Med Res Rev*. (2005); **25**: 133 – 166.

[26] Sanguinetti, M. C.; Tristani-Firouzi, M. hERG potassium channel and cardiac arrhythmia. *Nature*. (2006) **440**: 463 – 469.

[27] Vandenberg, J.; Torres, A. M.; Campbell, T. C.; Kuchel, P. W.; The HERG K^+ channel: progress in understanding the molecular basis of its unusual gating kinetics. *Eur Biophys J*. (2004); **33**: 89 – 97.

[28] De Ponti, F.; Poluzzi, E.; Montanaro, N.; Organising evidence on QT and occurrence of torsades de pointes with non-antiarrhythmic: a call for consensus. *Eur J Clin Pharmacol*. (2001); **57**: 185 – 209.

[29] Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C.; A structural basis for drug-induced long QT syndrome. *PNAS*. (2000); **97**: 12329 – 12333.

[30] Mitcheson, J. S.; Chen, J.; Sanguinetti, M. C.; Trapping of a methanesulfonanilide by closure of the HERG potassium channel activation gate. *J Gen Physiol*. (2000); **115**: 229 – 239.

[31] Chen, J.; Mitcheson, J. S.; Tristani-Firouzi, M.; Lin, M.; Sanguinetti, M. C.; The S4-S5 linker couples voltage sensing and activation of pacemaker channels. *PNAS*. (2001); **98**: 11277 – 11282.

[32] Kamiya, K.; Mitcheson, J. S.; Yasui, K.; Kodama, I.; Sanguinetti, M. C.; Open channel block of HERG K^+ channels by vesnarinone. *Mol Pharmacol*. (2001); **60**: 244 – 253.

[33] Perry, M.; de Groot, M. J.; Helliwell, R.; Leishman, D.; Tristani-Firouzi, M.; Sanguinetti, M. C.; Mitcheson, J. Structural determinants of HERG channel block by clofilium and ibutilide. *Mol Pharmacol*. (2004); **66**: 240 – 249.

[34] Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A.; Three-dimensional quantitative structure-activity relationship for inhibition of human Ether-a-go-go-related gene potassium channel. *JPET*. (2002); **301**: 427 – 434.

[35] Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M.; Toward a pharmacophore for drugs inducing the long QT syndrome: insights from a CoMFA study of HERG K(+) channel blockers. *J Med Chem*. (2002); **45**: 3844 – 3853.

[36] Rajamani, R.; Tounge, B. A.; Li, J.; Reynolds, J. L.; A two-state homology model of the hERG K^+ channel: application to ligand binding. *Bioorg & Med Chem Lett*. (2005); **15**: 1737 – 1741.

[37] Österberg, F.; Åqvist, J.; Exploring blocker binding to a homology model of the open hERG K^+ channel using docking and molecular dynamics methods. *FEBS Lett*. (2005); **579**: 2939 – 2944.

[38] Farid, R.; Day, T.; Friesner, R. A.; Pearlstein, R.; New insights about HERG blockade obtained from protein modeling, potential energy mapping, and docking studies. *Bioorganic & Medicinal Chemistry* (2006); **14**: 3160 – 3173.

[39] Wang, S.; Liu, S.; Morales, M. J.; Strauss, H. C.; Rasmusson, R. L.; A quantitative analysis of the activation and deactivation kinetics of HERG expressed in Xenopus oocytes. *J Physiol*. (1997); **502**: 45 – 60.

[40] Zhou, Z.; Gonq, Q.; Ye, B.; Fan, Z.; Makielski, J. C.; Robertson, G. A.; January, C. T.; Properties of HERG channels stably expressed in HEK 293 cells studied at physiological temperature. *Biophys J.* (1998); **74**: 230 – 241.

[41] Heginbotham, L.; Abramson, T.; MacKinnon, R.; A functional connection between the pores of distantly related ion channel as revealed by mutant K^+ channels. *Science*. (1992); **258**: 1152 - 1155.

[42] Doyle, D. A.; Morais Cabral, J.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.: Chait, S. L.; MacKinnon, R.; The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science*. (1998); **280**: 69 – 77.

[43] Morais-Cabral, J. H.; Zhou, Y. F.; MacKinnon, R; Energetic optimization of ion conduction rate by the K^+ selectivity filter. *Nature*. (2001); **414**: 37 – 42.

[44] Zhou, Y.; Morais-Cabral, J. H.; Kaufman, A.; MacKinnon, R.; Chemistry of ion coordination and hydration revealed by a K^+ channel-Fab complex at 2.0 Å resolution. *Nature*. (2001); **414**: 43 – 48.

[45] Noskov, S. Y.; Bernèche, S.; Roux, B.; Control of ion selectivity in potassium channels by electrostatic and dynamic properties of carbonyl ligands. *Nature*. (2004); **43**: 830 – 834.

[46] Shealy, R. T.; Murphy, A. D.; Ramarathnam, R.; Jakobsson, E.; Subramaniam, S. Sequence-function analysis of the K^+-selective family of ion channels using a comprehensive alignment and the KcsA channel structure. *Biophys J.* (2003); 84: 2929 – 2942.

[47] Pardo-Lopez, L.; Zhang, M.; Liu, J.; Jiang, M.; Posani, L. D.; Tseng, G. N.; Mapping the binding site of a human ether-a-go-go-related gene-specific peptide toxin (ErgTx) to the channel’s outer vestibule. *J Biol Chem.* (2002); **277**: 16403 – 16411.

[48] Zhang, M.; Korolkova, Y.; Jiang, M.; Grishin, E. V.; Tseng, G. N.; BeKm-1 is a HERG-specific toxin that shares the structure with ChTx but the mechanism of action with ErgTx1. *Biophys J.* (2003) **84**: 3022 – 3036.

[49] Roux, B.; MacKinnon, R. The cavity and pore helices in the KcsA K^+ channel: electrostatic stabilization of monovalent cations. *Science*. (1999) **285**: 100 – 102.

[50] Jiang, Y. ; Lee, A. ; Chen, J. ; Ruta, V. ; Cadene, M. ; Chait, B. T. ; MacKinnon, R. *Nature*. (2003) ; **423** : 33 – 41.

[51] Jiang, Y. ; Ruta, V.; Chen, J.; Lee, A.; MacKinnon, R. The principle of gating charge movement in a voltage-dependent K+ cannel. *Nature*. (2003); **423**: 42 – 48.

[52] Fernandez, D.; Ghanta, A.; Kauffman, G. W.; Sanguinetti, M. C.; Physicochemical features of the hERG channel drug binding site. *J Biol Chem*. (2004); **279**: 10120 – 10127.

[53] Mitcheson, J.S.; Perry, M. D. Molecular determinants of high affinity drug binding to HERG channels. *Current Opinion In Drug Discovery & Development*. (2003); **5**: 667 – 674.

[54] Fernandez, D.; Ghanta, A.; Kauffman, G. W.; Sanguinetti, M. C. Physicochemical features of the hERG channel drug binding site. *J Biol Chem*. (2004); **279**: 10120 – 10127.

[55] Miller, C. An overview of the potassium channel family. *Genome Biology*. (2000); **4**: 1 – 5.

[56] Choe, H.; Nah, K. H.; Lee, S. N.; Lee, H. S.; Lee, H. S.; Jo, S. H.; Leem, C. H.; Jang, Y. J.; A novel hypothesis for the binding mode of HERG channel blockers. *Biochem Biophys Res Comm*. (2006); **344**: 72 – 78.

[57] Rost, B.; Casadio, R.; Fariselli, P.; Sander, C.; Prediction of helical transmembrane segments at 95% accuracy. *Protein Sci*. (1995); **4**: 521 – 533.

[58] Rost, B. Fariselli P.; Casadio, R.; Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci*. (1996); **5**: 1704 – 1718.

[59] Long, S. B.; Campbell, E. B.; MacKinnon, R.; Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. *Science*. (2005); **309**: 897 – 903.

[60] Long, S. B.; Campbell, E. B.; MacKinnon, R.; Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science*. (2005); **309**: 903 – 908.

[61] Notredame, C.; Higgins, D. G.; Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol*. (2000); **302**: 205 – 217.

[62] Sali, A.; Blundell, T. L.; Comparative protein modeling by satisfaction of spatial restraints. *J. Mol. Biol*. (1993); **234**: 779 – 815.

[63] Laskowski RA, McArthur MW, Moos DS, Thornton JMJ. PROCHECK: a program to check the stereochemical quality of protein structures. Appl Cryst 1993;**26**:283-291.

[64] Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comp Phys Comm*. (1995), **91**: 43 – 56.

[65] Lindahl, E.; Hess, B.; van der Spoel; D.; GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model*. (2001); **7**:306 – 317.

[66] van Gunsteren, W. F., and H. J. C. Berendsen. 1987. GROMOS Manual. BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningen, The Netherlands.

[67] Anézo, C.; de Vries, A. H.; Höltje, H. D.; Tieleman, D. P.; Marrink, S. J.; Methodological issues in lipid bilayer simulations. *J Phys Chem B*. (2003); **107**: 9424 – 9433.

[68] Nagle, J. F.; Zhang, R.; Tristram-Nagle, S.; Sun, W.; Petrache, H. I.; Suter, R. M.; X-ray structure determination of fully hydrated L_α phase dipalitoylphosphatidylcholine bilayers. *Biophys J*. (1996); **70**: 1419 – 1431.

[69] Berger, O.; Edholm, O.; Jähnig. F.; Molecular dynamics simulations of a fluid bilayer of dipalmitoylphosphatidylcholine at full hydration, constant pressure, and constant temperature. *Biophys J*. (1997); **72**: 2002 – 2013.

[70] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J.; Interaction models for water in relation to protein hydration. Intermolecular Forces. (1981); Reidel, Dordrecht, The Netherlands. B. Pullman (ed) pp. 331 – 342.

[71] Essman, U. L.; Perera, L.; Berkowitz, M. L.; Darden, H. L. T.; Pedersene, L. G.; A smooth particle mesh Ewald method. *J Chem Phys*. (1995); **103**: 8577 – 8592.

[72] Berendsen, H.; Postma, J.; van Gunsteren, W.; DiNola, A.; Haak, J.; Molecular dynamics with coupling to an external bath. *J Chem Phys*. (1984); **81**: 3684 – 3690.

[73] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M.; LINCS: a linear constraint solver for molecular simulations. *J Comput Chem*. (1997); **18**: 1463 – 1472.

[74] Miyamoto, S.; Kollman, P. A.; Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem*. (1992); **13**: 952 - 962.

[75] Faraldo-Gómez, J. D.; Smith, G. R.; Sansom, M. S. P.; Setting up and optimization of membrane protein simulations. *Eur Biophys J*. (2002); **31**: 217 – 227.

[76] Compoint, M.; Carloni, P.; Ramseyer, C.; Girardet, C.; Molecular dynamics study of the KcsA channel at 2.0-Å resolution: stability and concerted motions within the pore. *Biochim et Biophys Acta*. (2004); **1661**: 26 – 39.

[77] Åqvist, J.; Luzhkov, V. Ion permeation mechanism of the potassium channel. *Nature*. (2000); **404**: 881 – 884.

[78] Bernèche, S.; Roux, B.; Molecular dynamics of the KcsA K^+ channel in a bilayer membrane. *Biophys J*. (2000); **78**: 2900 – 2917.

[79] Bernèche, S.; Roux, B.; Energetics of ion conduction through the K^+ channel. *Nature*. (2001); **414**: 73 – 77.

[80] Guidoni, L.; Torre, V.; Carloni, P.; Water and potassium dynamics inside the KcsA K^+ channel. *FEBS*. (2000); **477**: 37 – 42.

[81] Kleywegt, G. J.; Jones, T. A.; Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Cryst* (1994); **D50**: 178.

[82] Capener, C. E.; Proks, P.; Aschroft, F. M.; Sansom, M. S. P.; Filter flexibility in a mammalian K channel: models and simulations of Kir6.2 mutants. *Biophys J*. (2003); **84**: 2345 – 2356.

[83] Domene, C.; Grottesi, A.; Sansom, M. S. P.; Filter flexibility and distortion in a bacterial inward rectifier K^+ channel: simulation studies of KirBac1.1. Biophys J. (2004); 87: 256 – 267.

[84] Domene, C.; Sansom, M. S. P.; Potassium channel, ions, and water: simulation studies based on the high resolution X-ray structure of KcsA. *Biophs J.* (2003); **85**: 2787 – 2800.

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

[86] Hoover, W. G. Canonical dynamics: equilibrium phase-space distributions. *Phys Rev A.* (1985); **31**: 1695 – 1697.

[87] Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys* (1981); **52**: 7182 – 7190.

[88] Bernèche, S.; Roux, B.; On the potential functions used in molecular dynamics simulations of ion channels. *Biophys J.* (2002); **82**: 1681 – 1684.

[89] Monticelli, L.; Robertson, K. M.; MacCallum, J. L.; Tieleman, D. P.; Computer simulation of the KvAP voltage-gated potassium channel: steered molecular dynamics of the voltage sensor. *FEBS.* (2004); **564**: 325 – 332.

[90] Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham III, T. E.; Wang, J. Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L. Singh, U. C.; Weiner, P. K.; Kollman, P. A. AMBER 7, University of California, San Francisco (2002).

[91] http://pdb2pqr.sourceforge.net/

[92] Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R.; Development and validation of genetic algorithm for flexible docking. *J Mol Biol*. (1997); **67**: 727 – 748.

[93] Allen, T. W.; Andersen, O. S.; Roux, B.; On the importance of atomic fluctuations, protein flexibility, and solvent in ion permeation. J. Gen. Physiol. (2004); **124**: 679 – 690.

[94] Holoyake, J.; Domene, C.; Bright, J. N.; Sansom, M. S. P.; KcsA closed and open: modelling and simulation studies. *Eur Biophys J.* (2004); **33**: 238 – 246.

[95] Bernèche, S.; Roux, B.; A gate in the selectivity filter of potassium channels. *Structure*. (2005); **13**: 591 – 600.

[96] Shrivastava, I. H.; Sansom, M. S. P.; Simulations of ion permeation through a potassium channel: molecular dynamics of KcsA in a phospholipid bilayer. *Biophys J.* (2000); **78**: 557 – 570.

[97] Capener, C. E.; Shrivastava, I. H.; Ranatunga, K. M.; Forrest, L. R.; Smith, G. R.; Sansom, M. S. P.; Homology modeling and molecular dynamics simulations studies of an inward rectifier potassium channel. *Biophys J.* (2000); **78**: 2929 – 2942.

[98] Bernèche, S.; Roux, B.; A gate in the selectivity filter of potassium channels. *Structure*. (2005); **13**: 591 – 600.

[99] Fan, J. S.; Jiang, M.; Dun, W.; McDonald, T. V.; Tseng, G. N. Effects of outer mouth mutations on hERG channel function: a comparison with similar mutations in the Shaker channel. *Biophys J.* (1999); **76**: 3128 – 3140.

[100] Zhou, Z.; Vorperian, V. R.; Gong, Q.; Zhang, S.; January, C. T. Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole, *J Cardiovasc Electrophysiol* (1999), **10**: 836-843.

[101] Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem*. (2006); **49**: 534 – 553.

[102] Bottegoni, G.; Cavalli, A.; Recanatini, M. A comparative study on the application of hierarchical-agglomerative clustering approaches to organze outputs of reiterated docking runs. *J Chem Inf Model* (2006); **46**: 58 – 65.

[103] Bottegoni, G.; Rocchia, W.; Recanatini, M.; Cavalli, A. AClAP, aoutonomous hierarchical agglomerative cluster analysis based protocol to partition conformational datasets. *Bioinformatics*. (2006); **22**: 58 – 65.

[104] Gervasio, F. L.; Laio, A.; Parrinello, M. Flexible docking in solution using metadynamics. *J Am Chem Soc* (2005); **127**: 2600 – 2607.

[105] Branduardi, D.; Gervasio, F. L.; Cavalli, A.; Recanatini, M.; Parrinello, M. The role of the peripheral anionic site and cation-p interactions in the ligand penetration of the human AChE gorge. *J Am Chem Soc* (2005); **127**: 9147 – 9155.

[106] Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A.J. Automated docking using a Lamarckian genetic algorithm and empirical binding free energy function. *J Comput Chem*. (1998); **19**: 1639 – 1662.

[107] Bhat, R.; Xue, Y.; Berg, S.; Hellberg, S.; Ormö, M.; Nilsson, Y.; Radesäter, A.-C.; Jerning, E.; Markgren, P.-O.; Borgegård, T.; Nyölf, M.; Giménez-Cassina, A.; Hernández, F.; Lucas, J. J.; Díaz-Nido, J.; Avila, J. Structural insights and biological effects of glycogen synthase kinase 3-specific inhibitor AR-A014418. *J Biol Chem*. (2003); **46**: 45937 – 45945.