SCUOLA NORMALE SUPERIORE

PHD THESIS

Development, validation and application of accurate molecular force fields for complex soft matter systems

---

Author:
Gianluca DEL FRATE

Supervisor:
Prof. Vincenzo BARONE
Dr. Giordano MANCINI

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in

Methods and Models for Molecular Sciences

Contents

Preface

Up to now, the methodologies and results here discussed have led to the following publications:

- F. Fracchia, G. Del Frate[△], G. Mancini, W. Rocchia, V. Barone, "Force Field Parametrization of Metal Ions From Statistical Learning Technique", *Journal of Chemical Theory and Computation*, **2017**, Just accepted, DOI: 10.1021/acs.jctc.7b00779 ([△]co-first author)

- M. Macchiagodena, G. Del Frate*, G. Brancato, B. Chandramouli, G.Mancini, V. Barone, "Computational Study of DPAP Molecular Rotor in Various Environments: From Force Field Development to Molecular Dynamics Simulations and Spectroscopic Calculations", *Physical Chemistry Chemical Physics*, **2017**, Just accepted, DOI: 10.1039/C7CP04688J (*corresponding author)

- M. Macchiagodena, G. Mancini, M. Pagliai, G. Del Frate, V. Barone, "Fine-tuning of atomic point charges: Classical simulations of pyridine in different environments", *Chemical Physics Letters*, **2017**, 677, 120

- A. Salvadori, G. Del Frate, M. Pagliai, G. Mancini, V. Barone, "Immersive virtual reality in computational chemistry: Applications to the analysis of QM and MM data", *International Journal of Quantum Chemistry*, **2016**, 116, 1731

- G. Del Frate, F. Bellina, G. Mancini, G. Marianetti, P. Minei, A. Pucci, V. Barone, "Tuning of dye optical properties by environmental effects: a QM/MM and experimental study", *Physical Chemistry Chemical Physics*, **2016**, 18, 9724

Introduction

In the last decades, computational chemistry has established itself as an useful and powerful tool for designing molecular systems and forecast their properties over a wide range of space and time scales. *In silico* methods accelerate the design and discovery of novel materials with specific properties (e.g., mechanical, optical, pH responsive) as well as new chemicals with pharmacological activity. The proper theoretical framework for the description of molecular systems at the atomic level are the laws of quantum mechanics (QM). However, a satisfactory sampling of the phase space of large systems (up to a milion of atoms) is achievable only by employing classical methods, such as Molecular Dynamics (MD) and Monte Carlo (MC) simulations [1]. These techniques are based on molecular mechanics (MM) rules, which assume molecules being composed by a set of atoms modeled as spheres and linked together by springs. Electronic degrees of freedom are neglected, as entailed by the Born-Oppenheimer approximation. Atomic motions are described by means of the classical laws, and the potential energy surface (PES) is represented by a sum of analytical expressions with immediate physical meaning, aimed at describing bonded and non-bonded interactions between atoms. The integration the classical equation of motion is rather cheap, thus to allow to get structural and thermodynamic properties in molecular simulations at a computational cost which is significantly lower if compared to QM and hybrid QM/MM computations.

The complete set of functions and the corresponding parameters used in MM methods to describe both intra- and intermolecular interactions in a chemical system is named *force field* (FF). At present, a high number of FFs is available in computational chemistry, each to be used in a wide, yet specific chemical domain, from organic liquids (as the OPLS force field [2, 3]) to biomolecules (e.g., AMBER [4, 5] and GROMOS [6]) and minerals (e. g. INTERFACE [7], ClayFF [8]). In each of them, atoms are organized in "atom types", i.e. chemical elements with a specific hybridization and a specific chemical surrounding: identical atom types share the same FF parameters, and the same atoms (e.g., carbon) can be described by multiple atom types (e.g., in a sp^2 carbonyl group or in a sp alkyne). Transferability of both functional forms and parameters is an important feature, since it enables the employment of parameters developed on a small set of molecules to a wider range of chemical entities. The accuracy offered by transferable force fields however is limited by many factors (first of all, the similarity of the target molecule to those employed in the parameterization route and the thermodynamic conditions employed in the simulations during the parameterization procedure) and it may be also insufficient to guarantee the reproduction of several target properties of interest at the same level of the species included in the training set. Moreover, the invoked generality of the parameters may undermine the simulation of several chemical phenomena, such as spectroscopic data, the change of flexibility of organic dyes when going from liquid solutions to more obstructing environments, and bulk properties of liquids (as the static dielectric constant ϵ, density, structure, and many others).

With the growth of computer power and storage capacity, the role of classical simulation methods

has become ever more relevant in life and material science. Increasingly complex systems can be simulated by exploiting highly performing computing (HPC) facilities. In this scenario, there is wide margin for improve the currently available MM models and gather new insights from theoretical investigations with more detail and accuracy. Therefore, it is no coincidence that novel fitting methods and strategies, aimed at overcoming the drawbacks mentioned above, have been developed in the last few years: among them, innovative tools and software for the development of intramolecular FFs specifically tailored for one molecule by fitting QM optimized energies, gradients and Hessian matrix have recently attracted attention within the scientific community [9–13]. Parallel efforts are directed to the refinement or the computation *ex novo* of atomic charges [14], through the employment of novel schemes for electron density partitioning [15] or to the inclusion of virtual sites, which mimic somehow polarizability effects and correctly simulate hydrogen bonding patterns, within molecular topologies [16].

This thesis is devoted to improve classical simulations reliability through the application of novel strategies for classical FFs optimization. To this end, MM and QM state of the art approaches have been combined together and integrated in different protocols, devoted to the accurate parameterization and classical MD simulations of molecular systems, to overcome current performances offered by general parameters and to address new chemical problems of different kind. With this aim, the employed FFs have been developed from scratch or re-parameterized only in some parts depending on the circumstances, and showing how an accurate modeling, based on first principle computations and not relying on any empirical parameters, allow to compute with accuracy a series of thermodynamic, structural and spectroscopic properties of interest. In order to validate the optimized set of parameters, the computed MD trajectories have been extensively investigated: the reliability of the proposed parameters has been evaluated by comparing simulated results with available experimental data. In the cases of molecular probes force field optimization, the population of QM-predicted conformational energy minima along the trajectory has also been considered. Furthermore, MD has been opportunely integrated with QM methods (mainly routed in the density functional theory, DFT) to compute absorption spectra, which are not obtainable with standard force field-based procedures, in order to gain a further comparison with experiments.

The thesis is organized as follows. In the first part, the theory underlying the thesis work is briefly reported. The second part focuses on the applications arising from the protocols explained in the first part. The chemical investigations carried out in the course of this PhD are reported following, in ascending order, a sort of "degree of modification" of the used parameter set, from the first works (where a low level of modification of the FF is applied) until to the last one, where a totally new parameterization procedure is proposed.

Future perspectives concerning this research work are given in the last section. All the parameter sets developed in this thesis are given in the Appendice.

Part I

THEORY

This part is organized as follows. In Chapter 1, a summary on classical force fields and most common functional forms currently available in the literature is given. In Chapter 2 a general overview on MD technique (the main tool used in this thesis) is briefly reported. In this regard, MD simulations have been used to

1. validate the developed FFs;
2. investigate on structural, dynamic and spectroscopic properties of the considered chemical systems.

Chapter 3 presents two algorithms for force field development: *Joyce* [9], which performs a parameter optimization for the FF intramolecular potential, and *LRR-DE*, a new procedure for the generation of non-bonded models for metal ions, developed in the course of this PhD. The analysis of the MD trajectories have been performed by using a set of computational tools, which are outlined in Chapter 4.

Chapter 1

Molecular Mechanics force fields

The total energy E of a molecular system is given by the time-independent Schrödinger equation

$$\hat{H}\Psi(r,R) = E\Psi(r,R) \tag{1.1}$$

where $\hat{H}$ is the Hamiltonian operator which describes both potential and kinetic energy and Ψ is the wavefunction dependent on the electrons and the nuclei positions (r and R). Under the Born-Oppenheimer assumption, the motion of electrons can be separated from the motion of the nuclei, due to the large difference between elctron and nuclear masses: electrons moves faster and they readapt their positions instantly to nuclear motions. Therefore, the wave function can be factorized into an electronic and a nuclear part, and the electronic energy (E_{el}) can be computed for fixed nuclear positions. E_{el} defines the PES over all the possible nuclear coordinates: a chemical systems can be considered as a ball moving on the corresponding PES, whose sampling by means of dynamic simulations allow to get a lot of molecular properties.

At present, several approximation levels can be used in order to model the PES (from MM to DFT and coupled-cluster methods). In force field methods, it is calculated as a sum of pairwise potentials, dependent on the spatial coordinates of the nuclei. The specific decomposition depends upon the force field in use. One way of representing the potential energy of a molecule in vacuum is the following:

$$E_{intra} = E_{stretch} + E_{bend} + E_{Rtors} + E_{Ftors} + E_{nb} \tag{1.2}$$

Here, the intramolecular energy depends on five terms: the first four are due to bonded-atoms interactions, and single energy contributions derive from displacement from equilibrium values. $E_{stretch}$, E_{bend}, E_{Rtors} and E_{Ftors} correpond to the stretching, bending, rigid and flexible torsions potentials. The last (E_{nb}) term refers to non-bonded contributions: both 1-4 interactions (computed between end atoms involved in a dihedral angles, often scaled applying an empirical factor) and those between atoms separated by more than three bonds are included in such term. The first derivative of the potential E_{intra} computed on atom i corresponds to the atomic force on that atom.

Eq. 1.2 refers to a particular class of force field, i.e. the empirical non-reactive ones. Such kind of classical models have been used within this work: the corresponding terms, well adopted within widespread MD codes, are discussed in the next section. It has to be stressed that alternative ways of modeling the PES of a system, such as polarizable and reactive models, can be adopted depending on the chemical phenomenon that needs to be studied.

1.1 Force field terms

1.1.1 Bonded interactions

The most known approach to model chemical bonds in a molecule is to rely on Hooke's law formula, which assign quadratic penalties to the displacement from the reference bond length, according to

$$V(r) = \frac{1}{2} k^s_{ij} (b_{ij} - b^{eq}_{ij})^2 \tag{1.3}$$

where k^s_{ij} is the force constant associated to the ij bond, b_{ij} is the measured bond length and b^{eq}_{ij} is the bond length at the equilibrium. Anharmonic potentials too, with asymmetric energy profile and zero forces at infinite distance can also be used. Among them, the Morse potential [17], has the following form

$$V(r) = D_{ij}[1 - \exp(-\beta_{ij}(b_{ij} - b^{eq}_{ij}))]^2 \tag{1.4}$$

where D_{ij} and β_{ij} are the well depth and the corresponding steepness, respectively. The Morse potential is used to describe bond formation and bond breaking, since it considers the existence of unbound states: for this reason it is adopted in reactive force fields in modeling chemical reactivity.

The Hooke's law is used also to describe deviation of angles from their equilibrium values:

$$V(\theta) = \frac{1}{2} k^{\theta}_{ijk} (\theta_{ijk} - \theta^{eq}_{ijk})^2 \tag{1.5}$$

where k^{θ}_{ijk} is the force constant of angle $\widehat{\text{ijk}}$, θ_{ijk} and θ^{eq}_{ijk} are actual angle and corresponding reference value. Boradly speaking, values of k^{θ}_{ijk} are generally lower if compared to k^s_{ijk}, meaning that lower amounts of energy are required to deform angles and move them from their reference positions. GROMOS force field uses a cosine-based function for angle vibrations. A harmonic corrective term on the distance between i and k (i.e., the end atoms of the angle), in addition to Eq. 1.5, is used in the CHARMM force field [18] (known as the Urey-Bradley potential).

In flexible molecules, dihedral angles affect molecular conformation. Two different kind of dihedral angles can be defined: rigid torsions, modeled through standard harmonic potentials, as the ones described by equation 1.3 and 1.5, and flexible torsions. Improper dihedral angles, often included in force fields in order to keep a group of atoms fixed at a particular geometry (e.g., planar aromatic rings), are included in the first category. In the other one, dihedral angles are modeled using periodic functions (cosine series expansion). A typical form is the following:

$$E^F_{tors} = k^{\phi}_{ijkl}(1 + cos(n\phi - \gamma)) \tag{1.6}$$

Here, k^{ϕ}_{ijkl} is the force constant which governs the flexible torsion defined by the i, j, k and l atoms. γ is the phase factor, and it determines where the function passes thorugh a minimum. The multiplicity value instead defines the number of minima along a complete scan of the $ijkl$ dihedral angle. Magnitudes of k^{ϕ}_{ijkl} are notably lower if compared to bonds and angles force constants: for some systems, significant deviations of dihedral angles from their equilibrium states can be easily observed in MD simulations of few hundreds of picoseconds.

The parameterization of the bonded interactions is usually performed by means of spectroscopic experiments: also quantum mechanics calculations on the isolated molecule can be used.

1.1.2 Non-bonded interactions

Non-bonded interactions are essentially of van der Waals (vdW) and electrostatic nature. Usually modelled as a function of the inverse of the distance beetween two non-bonded atoms, force field non-bonded part is important for chemical structure determination: in fact, both van der Waals and electrostatic interactions have a key role in determining the global structure of large molecules, e.g. protein folding. Furthermore, they contribute to intermolecular interactions and macroscopic properties. vdW interactions, in particular, have been shown to dominate the heats of vaporization of nonpolar organic molecules and have effects on several other condesed-phase properties, such as density and molecular volumes.

Depending on the force field in use, 1-4 interactions (i.e., those between non-bonded atoms which are separated by three bonds) can be scaled down by using empirical factors. Just for example, OPLS uses a 0.5 scale factor for both electrostatic and vdW interactions; AMBER applies 0.8333 for electrostatics, while it scales vdW in equal measure as OPLS. Scale factors for non-bonded interactions are not considered in GROMOS force field.

van der Waals interactions

The most common way to think about vdW forces is to consider two non-bonded atoms. In a classical way, such atoms could be seen as two spheres with an atomic radius, r_{vdW}: at infinte distances there are no attractive forces between them. As the two atoms approach one another, dispersive forces beetween them arise, mainly due to the correlated fluctuations of the electron clouds. This results in the formation of two dipoles. Such interactions are commonly named *London forces* and they have an inverse sixth power dependence on the distance between the two considered atoms [19]. The most popular function used to describe Van der Waals interactions is the Lennard-Jones potential [20] (Eq. 1.7).

$$V(r) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.7}$$

In such equation, r_{ij} indicates the distance beetween the considered atoms i and j, σ_{ij} is the separation at which the energy value is zero and ϵ_{ij} is the minimum (the well depth). The well depth parameter is often expressed in units of temperature, as ϵ/k_b, where k_b is the Boltzmann's constant. Often, the same law is expressed in function of the distance at which the energy value matches the well depth, r_m. In this case, the LJ function is

$$V(r) = \epsilon_{ij}\left[\left(\frac{r_m}{r_{ij}}\right)^{12} - 2\left(\frac{r_m}{r_{ij}}\right)^{6}\right] \tag{1.8}$$

It is easy to demonstrate that the value of r_m correspond to $2^{1/6}\sigma$, since, at this separation, the first derivative of the energy with respect to the distance r is equal to zero.

The following more simplified formulation

$$V(r) = \frac{A_{ij}}{r_{ij}}^{12} - \frac{B_{ij}}{r_{ij}}^{6} \tag{1.9}$$

is commonly used in molecular simulation packages. Here, $A_{ij} = 4\epsilon_{ij}\sigma_{ij}^{12}$ and $B_{ij} = 4\epsilon_{ij}\sigma_{ij}^{6}$.

In order to describe vdW behavior, the potential function must contain both an attractive and a repulsive contribution: as said before, it is theoretically demonstrated that the attractive term has a dependence on the inverse of the sixth power of the distance. As regarding the repulsive contribution, there is no theoretical justification to the r^{-12} dependence. The Buckingham potential instead has a more reliable (from a theoretical point of view) repulsion term [21]:

$$V(r) = Ab_{ij}\exp(-Bb_{ij}r_{ij}) - \frac{B_{ij}}{r_{ij}^6} \tag{1.10}$$

where Ab_{ij}, Bb_{ij} are constants specific for atoms i and j in the Buckingham potential. B_{ij} is the same as in Eq. 1.9. Another functional form with an exponential dependence is the already shown Morse potential (Eq. 1.4). In spite of their agreement with quantum mechanics theories, exponential forms for the VdW potential are not so popular and this fact is mainly due to their excessive computational cost. It has to be mentioned that some mathematical operations, like square roots and exponential functions are computationally more expensive if compared to more simple multiplications and additions. Force field calculations use atomic cartesian coordinates as variables in the energy expression and so the determination, for example, of a distance between two atoms involves the use of a square root. This value is directly employed in the functional of the exponential expression, while it compares raised to even powers inside the LJ potential. Hence in the latter case the exact value of the distance is not needed. Furthermore, it has to be considered that the main differences between the functional forms above mentioned regard only the repulsive contribution to the whole potential, which is less important during non-bonded energies calculations since it arises under the r_m value.

Polyatomic systems are made up by different atom types, and if a cutoff value has not been set, the vdW energy ordinarily has to be computed for all the atom pairs. The application of the potential described above, however, requires the knowledge *a priori* of the σ and the ϵ parameters. Nevertheless, these factors are usually reported as single values for each *individual* atom type. Currently it is not feasible to calculate individual parameters for all the atom-pair interactions of interest (i.e., the different σ_{ij} and ϵ_{ij} values). Thus, a way to combine single atom type parameters is necessary in order to calculate long-range interactions between different atom types. Moreover, good results for the vdW interactions between unlike atoms will be attainable only if appropriate combination rules are used. Current molecular mechanics force fields use either the geometric mean to express the well depth ϵ_{ij}, while different solutions are adopted for the vdW minimum distance r_m between atom i and atom j. The geometric mean used by OPLS and GROMOS (generally named as the OPLS combination rule) tends to underestimates the experimental values, specially when the considered atoms differ substantially. In a similar way, but with a lesser degree, performs the arithmetic mean rule, used by AMBER and CHARMM (Lorentz-Berthelot combination rule [22, 23]). Anyway, the minimum energy distance tends to fall closer to that of the larger atom in the considered pair. Broadly speaking, both OPLS and Lorentz-Berthelot rules consider parameters of dissimilar atoms as almost an additive. Other, more specific rules, as the Fender-Halsey [24] and the Waldman-Hagler [25], treat the interaction between substantially different atoms to be significantly weakened [26].

As mentioned above, the determination of good-quality energy parameters in the evaluation of vdW interactions is a very important step during force fields parameterization, because of the key role

that such factors play on several macromolecular properties. Different experimental techniques, as beam scattering [27] and cromatography [28], have been used many decades ago in order to estimate potential parameters. Moreover, σ and ϵ have been determined for polar molecules like NH_3 and water by viscosity measures and diffusion studies [29] correlating transport properties with the LJ potential. The analysis of crystal packing and X-ray structures is an alternative experimental practice used to calculate energy parameters at the atomistic level. In fact, since crystalline structures are determined by the balance between repulsive and attractive forces between molecules inside a crystal, theoretical studies on these systems could shed light on intermolecular forces potential [30, 31].

Taking into account the r_m value, it is well known that it could be approximated as the sum of the atomic vdW radii (r_{vdW}) of the two interacting atoms. However, the selection of a set of vdW radii appears unfortunately arbitrary, due to the variability of contact distances among the different X-ray data as the atoms experience different environment and are more or less compressed. In the case of enough crystal structures and available zero-point density data, one could pick for that value of r_{vdW} which yields the correct packing density ρ_O at 0°K [32].

Looking for a periodical equation for the vdW radii, some semiempirical rules were developed. Correlations of these parameter with electron density values, with the de Broglie wavelenght λ_b of the outermost electrons (that assumded $r_{vdW} = \lambda_b/2$) and with the covalent radius r_c were made [33]. Pauling, for example, proposed $r_{vdW} = r_c + 0.8$Å for VA-VII elements [34]. One of the most easy way to test the validity of experimentally determined potential parameters may be the *second virial coefficient* ($B(T)$) of a gas composed by the considered atoms [35]. Values of $B(T)$ are computed as a function of r_{vdW} and ϵ, which are modified until B(T) coincides with the experimental datum at the considered temperature. Since the curves of r_{vdW} versus ϵ are temperature dependent, values of LJ energy parameters can be determined from an overlap region, where every curves met together. In this region the parameters have to be found in order to have the best fit with the experimental curve of B(T).

Nowadays most used methods in the determination of the ϵ and σ could be denoted as 'fitting methods'. Thanks to the increase of computational performances, it is usual to perform molecular simulations where non bonded parameters are modified in a iterative way, in order to reproduce experimental data [36]. In fact, although it is well accepted that parameters assigned to one atom in a molecule are quite well transferable to describe the same atom in a different molecule, often some sort of optimization is needed. This is particularly required if several experimental properties have to be contemporaneously reproduced. These procedures start with an initial guess of the force field parameters, which are then used to perform several computer simulations. By these, different properties, like vapor pressure and caloric properties are computed and a cost function (usually gradient-based) calculates the deviation of the predicted quantities from the experiments. Then, from these results, a new set of parameter is achieved and used in a next ensemble of simulations, leading to a less deviation from the experimental observables, and so on. This process continues, unless a minimum is reached [37]. OPLS and GROMOS non bonded parameters for alkanes were developed in such manner, using heat of vaporization, C-C radial distribution functions and heat of vaporization as target properties [38, 39]. Recent procedures have also highlighted the dependence of observables to a single LJ parameter: e.g. surface tension γ at the liquid vapor interface could be modulated varying ϵ, while density is closely related to the σ factor. This could be exploited by developing systematic procedures where specific parameters could be modified leaving almost unchanged the others [40]. Unfortunately, other

parameterization studies have shown that very different values of the LJ parameters can well reproduce condensed phase properties. This problem, known as the *parameter correlation problem* [41] indicates that often more information beyond the macroscopic observables is needed. The reproduction of *ab initio* data on interactions between rare gas atoms and model compounds could represent a solution to this issue [42]. LJ parameters are selected in order to reproduce interaction energies between gas atoms and model compounds computed at the QM level. Only in a second step, such values are modulate in order to yeld condensed-phase properties in good agreement with experiment. However, due to the required computational cost to determine accurate parameters able to describe vdW interactions from quantum mechanics calculations, fitting procedures remain widely used.

Electrostatic interactions

In MM, electrostatic potential in a molecule is computed as a sum of pairwise interactions between partial charges located at the centre of the nuclei. In classical force field, this issue is achieved by applying the Coulomb law

$$V(r) = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.11}$$

where q_i and q_j are partial charges located on atom i and j, respectively. Within this formalism, charges are fixed, i.e. they do not depend on molecular conformation and they are not influenced by alterations in the local enviroments which they perceive during a MD simulation. Polarizable force fields in principle should be more accurate than fixed charge force fields, but for many systems of interest, current fixed-charged models may provide results that are comparably reasonable in aqueous solution to polarizable contemporaries [43]. Despite several limitations, which derive mainly from the isotropic distribution around the nuclei, point-charges model using Coulomb equation is still largely used: forces due to electrostatic interactions (i.e., the first derivative of Eq. 1.11) are easy to compute and directly act on the nuclei, so to be really useful in MM simulations.

Since their importance in chemistry, great attention have been paied in developing robust methods to compute partial charges during the years. These efforts are usually classified in four main categories, depending on the fitting scheme adoptetd and/or the target property of the fitting procedure itself.

1. *Class I charges* are not dependent on QM quantities and they are computed by using intuitive approaches, usually based on electronegativity concepts. As an example, the method proposed by Marsili and Gasteiger in 1980 [44] can be cited: it is a iterative process based on atoms electronegativity, where atomic charge quantity Q is transferred from low- to high-electronegative atom types at each step k. The electron charge transferred depends on the electronegativity difference between the atoms and it is strictly modulated by k, since $Q \propto f^k$, where f is a dumping factor of value 0.5. Another model belonging to this first category is the *QEq* (*charge equilibration model*) by Rappè and Goddard [45]. The *QEq* model requires as input data the ionization potentials (IP), electron affinity and atomic radii of the different atom types of a molecule. Charges are computed taking into account shielded electrostatic interactions between all charges: point charges therefore depend on molecular geometry, and they are computed during MD simulations. QEq is used within the Universal Force Field (UFF) [46], so to be extended to all the elements of the periodic table. Class I charge models are known to be fast, and they are used for chemoinformatic purposes [47].

2. *Class II* charge models include population analysis schemes, such as Mulliken [48], Hirshfeld [49] and Lowdin [50] population analysis. Atomic charges are obtained from the partitioning of the electron density (obtained at the QM level) into atomic populations following orbital-based processes.

3. *Class III charges* are derived in order to reproduce physical observables computed from the wave function. Here, only the molecular electrostatic potential (MEP) as a property to be fitted is considered. The ESP (again, from "electrostatic potential") procedure [51] is a lest-squares algorithm which assign the best atomic charge set able to reproduce the MEP, computed at QM level, around the molecule. Different points, located on a cubic grid encompassing the vdW surface of the molecule, are considered for the evaluation of the MEP. The restrained-ESP (RESP) method [52] tries to overcome some of the typical problems of the ESP procedure, such as the conformational dependence of computed charges as well as their high absolute values. RESP charges are currently adopted in AMBER.

4. *Class IV charges* are developed using either Class II or Class III charges as precursor set transformed in a new one by a semi-empirical mapping, which is optimized so that the new set of charges better reproduce a physical observable (e.g., molecular dipole moment or the MEP computed at a high level of theory) than the precursor one. By using experimental quantities, Class IV charges are designed to adjust for systematic errors that occur systematically for a given level of electronic structure [53]. Class IV charges deriving approaches are commonly defined as Charge Models (CM). Last CM, CM5 [54], uses Hirshfeld population analysis as input charges, which are less sensitive to basis set size as well as the choice of the basis. Such model has been parametrized on a large training set of more than six hundred of instances, using the data for 26 elements.

1.2 A survey of existing force fields

In the context of classical simulations, force fields encode and predict the chemical traits of the simulated chemical system. Therefore, the choice of the force field to be used is of paramount importance for any computational chemistry investigation. As already anticipated in the Introduction, FF parameters are obtained in order to reproduce experimental and/or high level QM data for a selected set of similar molecular target. This practice allows for the transfer of the optimized parameters to chemical entities with similar features to the one included in the original target set. A common way of FF classification is based on the degree of transferability of the corresponding parameter set [55]. Thus, in the first group, FFs aimed at large transferability can be included, such as the ones designed in order to cover the whole periodic table (e.g. UFF). A second large group is made up by FFs well focused on a specific class of chemical systems (proteins, lipid, organic molecules). A third group includes high-quality force fields, properly derived in order to accurately reproduce a range of molecular properties, from conformational structure to vibrational frequencies. Force fields which belong to this category (e.g., COMPASS [55]) feature complex yet flexible functional forms and off-diagonal cross-coupling terms. At last, a series of algorithm and procedures devoted to the *ad hoc* parameterization of FFs specifically

tailored for one molecules have to be mentioned. These tools are commonly employed when spectroscopic accuracy has to be achieved, thus that specificity for the investigated system is largely preferred over transferability of the employed parameters.

In the following, a general but not exhaustive overview on currently available force fields is provided.

1.2.1 Biomolecules

From the beginning of the 80's the modeling of proteins and macromolecules of biological interest has been placed at the heart of computational chemistry. Current force fields for biological macromolecules look very similar to each other in their functional forms and often in the parameters: they rely on fixed point-charges, they mostly employ the standard LJ potential (Eq. 1.9) for modeling attractive and dispersive interactions, and they use harmonic potentials for stretching and bending (Eq. 1.3 and 1.5) and Eq. 1.6 for flexible dihedral angles.

The first AMBER (Assisted Model Building with Energy Refinement) force field was released in 1984 [56]. Significant advances were done by Cornell *et al.* [57] with the Amber94 force field: bonded parameters were determined to reproduce structural and vibrational frequency data on small molecular fragments that make up proteins and nucleic acids. Atomic charges were computed using a 6-31* basis set with RESP fitting, while vdW parameters were iteratively varied during Monte Carlo simulations until bulk properties were reproduced. Later versions have been focused on the improvement of amino-acid side-chain and protein backbone dihedral angles description using NMR measurement as reference data [5, 58]. In 2004, an extension of Amber aimed at including organic molecules have been developed (that is, the General Amber Force Field, GAFF [59]).

The development of OPLS (Optimized Potentials for Liquid Simulations) was aimed initially at the modeling of liquids. Original OPLS used a united-atoms paradigm: sites for non-bonded interactions were placed on all non-hydrogen atoms and on hydrogens attached to heteroatoms or carbons in aromatic rings [60]. Later, OPLS moved to an all-atoms representation. Force field non-bonded parameters were derived in order to reproduce several experimental properties [2], such as densities and heat of vaporization, while stretching and bending paramaters have been adopted from AMBER. The recent OPLS3 version [3] correctly deal with proteins as well as accurately predict protein-ligand binding affinities.

CHARMM (Chemistry at HARvard Molecular Mechanics) force field was released together with its simulation package in the early 1980s [18]. As in OPLS, also the initial CHARMM used a united-atoms representation. Parameters were developed and tested mainly on gas-phase simulations, but this parameterization also used more sophisticated fits to quantum mechanics calculations, typically including hydrogen bonded complexes between water and different molecular fragments [61]. Several improvements have been made during the years to reliably describe lipids [62]. The Charmm general force field (CGenFF) [63] should be seen as an extension of the chemical space covered by the standard CHARMM to include organic molecules such as drug-like compounds.

GROMOS (GROningen MOlecular Simulation) non bonded interaction parameters were obtained at the beginnning from crystallographic data and atomic polarizabilities, and adjusted such that experimental distances and interaction energies of individual pairs were reproduced for minimum energy configurations [64]. Since then, the parameters have been deeply improved by exploiting the increase

in computational power. Latest versions have been developed using free enthalpies of hydration and apolar solvation for a range of compounds as target data, since the importance of such property in many processes of biological interest [65].

1.2.2 Water models

Classical water models are parametrized in order to reproduce experimental data (radial distribution function, diffusion coefficient, density, dielectric constant and so on). TIP3P [66] and SPC [67] are among the simple water models commonly used in biomolecular simulations: they use three sites (each on one atom) which are kept at a fixed geometry, and they slightly differ on both atomic charges and LJ parameters. The TIP4P model adds one more site (without mass) along the bisector of the $\widehat{\mathrm{HOH}}$ angle at a 0.15 Å distance from the oxygen atom. TIP5P instead uses two dummy atoms negatively charged, which represent the two lone pairs of the oxygen atom and leading to a tetrahedral geometry. A more complex form than the one of the models already mentioned is used in the Toukan and Rahman model [68]: here, the structure is assumed to be flexible and the O-H bond is described anharmonic. New versions of the TIP3P and TIP4P models have been recently developed [69] which exhibit a better reproduction of several experimental properties (the dielectric constant, in particular).

1.2.3 Minerals

ClayFF [8] and INTERFACE [7] are two force fields specially derived to model inorganic compounds and their interfaces with fluids, using the same energy expression as the one of the biomolecular FFs. PMMCS [70] instead employs a three-terms interatomic potential, made by the Coulomb and Morse potentials together with the repulsive contribution C/r^{12} term of the LJ potential.

1.2.4 Reactive force fields

In this category reactive models, which explicitly take into account bond formation and breaking in order to model chemical reactions, have to be included. ReaxFF [71], developed in 2001, used a bond-order potential, which provide a general relationship between bond distance and bond energy that leads to proper dissociation of bonds to separate atoms. The terms of the potential have been designed in order to go to zero as atoms dissociate. Other contributions in the force field take into account over- and undercoordination penalties and conjugation effects on the global energies. ReaxFF have been developed initially on hydrocarbon compounds, aimed at reproduce heats of formation, bond lengths and angles data available in the literature. Experimental data of both non-reactive and reactive behavior were fairly reproduced. The same model has been applied also to other chemical systems such as metal ions in water and metal-catalyzed reactions [72, 73].

1.2.5 Polarizable models

Polarization refers to the ability of charge distribution to rearrange itself as a consequence of a surrounding electrostatic field. Respect to standard, fixed charges models, polarizable force fields offer an improvement in functional forms by including many-body effects.

A way to model polarization is to consider variable point charges: the combination of the fluctuating charge QEq model [45] based on electronegativity to model electrostatics with standard OPLS has been applied to small peptides to predict energetics of different configurations [74]. AMOEBA [75, 76] replace the partial charges model with contributions of both permanent and induced multipoles. The polarization is achieved through a mutual induction scheme which requires an induced dipole to polarize all the other sites, until all the induced dipoles at each site reach convergence. The computational cost offered by such models is higher with respect to standard non-polarizable models, although less expensive if compared to hybrid QM/MM approaches. A cheaper way is offered by the Drude oscillator model [77], which uses an additional particle to be attached to each polarizable site to account for polarizabiltiy, thus preserving the classical particle-particle electrostatic interaction of non-polarizable force fields.The polarizable version of CHARMM is based on the Drude oscillator [78].

1.2.6 Algorithms for high-quality FFs

As already stressed in this document, general transferable force fields may fail in the reproduction of chemical properties which are of paramount importance. Indeed, only a small part of the chemical space is covered by current parameterizations. Even if universal force fields as the UFF are aimed to deal with a molecular system of every kind, they are insufficient for many purposes. In order to fill this gap, many routines have been developed to accurate parametrize one chemical system, in a particular electronic configuration, by exploiting specific reference data and minimizing a specific objective function.

The force-matching approach [79] has been applied to parametrize non-polarizable force fields (of the GROMOS form) by computing atomic forces at the QM level during QM/MM MD simulations [80]. The method optimizes all the interaction parameters, except for the vdW ones, which are retained from pre-existing sets of parameters. The atomic charges are derived at first by reproducing the electrostatic potential and forces experienced by the MM part of the system; then, the contribution given by the computed charges and the used LJ parameters is subtracted by the atomic forces in the QM region: the remaining part is used for the fitting of the bonded part of the optimizing force field.

Other tools use the Hessian matrix computed at the QM level as a reference quantity in order to optimize the bonded part of the force field [10, 81]. The improvement of bonded description is the main goal of the Paramfit tool [12] by fitting high level energy and forces. GAAMP [12] optimizes both standard as well as Drude polarizable atomic models, using existing parameters as initial data; then, electrostatics is parametrized by using ESP and the QM-level interactions between water and the considered molecule as target data, and torsion potentials are refined to match QM energies of different conformers.

Efficient methods able to optimize new, more demanding functional forms (as the Class 3 FFs) have been recently developed. ForceBalance [13] gives high freedom to the user in the choice of the potential form and reference data. To this end, this tool is able to optimize both linear and non-linear parameters (as the exponential parts of the Buckingham and the Morse potentials). Moreover, the objective function is regularized in order to i) prevent from the overfitting to the target data, and ii) to avoid large and unphysical values of the optimized parameters.

Chapter 2

Molecular Dynamics

Molecular Dynamics (MD) simulations solve Newton's equation of motion (EOM) for a system of N interacting particles [82]:

$$m_i \frac{\partial^2 \vec{r}_i}{\partial t^2} = \vec{F}_i \quad i = 1, 2, \ldots, N \tag{2.1}$$

where $\vec{r}_i$ are the coordinates of particle i, t is the time, and $\vec{F}_i$ is the sum of the forces acting over particle i:

$$\vec{F}_i = -\frac{\partial V}{\partial \vec{r}_i} \tag{2.2}$$

where the potential energy $V(\vec{r}_i)$ is a function of the types of atoms in the system, of their relative distance and parameters. EOM is solved at discrete time intervals, saving the coordinates (and in some cases the velocities too) of all atoms in the system as a function of time (i.e. storing a *trajectory* of the simulated system). Following the ergodic hypothesis, the average of a observable over the simulated time corresponds to the average over the statistical ensemble in use. Hence, macroscopic properties can be extracted performing time averages of the saved coordinates. Table 2.1 shows the global MD algorithm.

The use of classical mechanics at normal temperatures is usually a good approximation. However very light atoms, most notably hydrogen atoms, show quantum mechanical behavior in certain situations such as tunneling phenomena or hydrogen bonding formation. Moreover in MD simulations covalent bonds and bond angles are usually approximated as classical oscillators. A classical approximation can give a reasonable description of quantum oscillator as long as the resonance energy, $h\nu$ is small enough compared to $k_B T$. At room temperature this value is about 200 cm^{-1}, i.e. very low compared to the energies of common covalent bonds such as the covalent C-C bond stretching. This means that, when performing simulations, both a correction to the energy of classical oscillators or *constrains* to the atoms can be applied. Constraints are used very often in classical simulations because this allows to increase the integration time step (neglecting the higher oscillations) and thus to perform longer simulations at a reasonable time (*vide infra*).

2.1 Integration of the equations of motion

The equation of motion is integrated using finite difference methods. The information on the state of the system at time t are used to calculate the forces at time $t + \delta t$ and then to predict the new positions at time $t + \delta t$ (δt is the integration step). If the new position of a particle at time $t + \delta t$, $\vec{r}(t + \delta t)$, is

1. Input and initial conditions
Initial coordinates, $\vec{r}$, of all atoms in the system
Initial velocities $\vec{v}$ of all atoms in the system
Calculation of the potential energy V as a function of $\vec{r}$ and $\vec{v}$
$\Downarrow$
2. Force calculation
The force exerted on each atom is
$\vec{F}_i = -\frac{\partial V}{\partial \vec{r}_i}$
calculated taking into account non bonded interactions between atom couples
$\vec{F}_i = \sum_j \vec{F}_{ij}$
bonded interactions (that can depend from 2 to 4 atoms)
geometrical constraints and/or external forces
The kinetic and potential energy and the pressure tensor are calculated
$\Downarrow$
3. Position and velocity update
Atomic motion is performed solving with numerical methods
Newton's equations of motion
$\frac{\vec{F}_i}{m_i} = \frac{d^2\vec{r}_i}{dt^2}$
or
$\frac{d\vec{r}_i}{dt} = \vec{v}_i; \vec{a}_i = \frac{\vec{F}_i}{m_i}$
Coordinates, velocities, energies... are saved in the trajectory

steps 2,3,4 are repeated up to the established number of time steps

TABLE 2.1: An overview of the MD algorithm.

expanded in Taylor series:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \vec{v}(t)\delta t + \vec{a}(t)\frac{\delta t^2}{2} + O(\delta t^3) + \cdots \tag{2.3}$$

where $\vec{r}$ and $\vec{a}$ are the velocity and acceleration at time t. In the precedent equation, $\vec{a} = \frac{\vec{F}}{2m}$ ($\vec{F}$ is the force acting on the particle at time t), so that an analogous expansion of $\vec{r}$ for negative times can be written:

$$\vec{r}(t+\delta t) = \vec{r}(t) + \vec{v}(t)\delta t + \frac{\vec{F}(t)}{2m}\delta t^2 + O(\delta t^3) + \cdots \tag{2.4}$$

$$\vec{r}(t-\delta t) = \vec{r}(t) - \vec{v}(t)\delta t + \frac{\vec{F}(t)}{2m}\delta t^2 - O(\delta t^3) + \cdots \tag{2.5}$$

Summing side by side these two equations:

$$\vec{r}(t+\delta t) = 2\vec{r}(t) - \vec{r}(t-\delta t) + \frac{\vec{F}(t)}{m}\delta t^2 + O(\delta t^4) \tag{2.6}$$

The forces are calculated once per cycle, and a trajectory obtained with this algorithm is time reversible. The velocity is no present in the algorithm and can be approximated as

$$\vec{v}(t) = \frac{\vec{r}(t+\delta t) - \vec{r}(t-\delta t)}{2\delta t} \tag{2.7}$$

This algorithm for calculating the new positions is known as the Verlet algorithm [83]. Another time reversible method used to integrate the equations of motions is known as the *leap-frog* algorithm [84]. It uses the positions at time t and the velocities at time $t - \frac{\delta t}{2}$. Positions are calculated subtracting to the series expansion of $\vec{r}(t+\delta t)$ that of $\vec{r}(t)$ at time t:

$$\vec{v}(t+\frac{\delta t}{2}) = \vec{r}(t-\frac{\delta t}{2}) + \frac{\vec{F}}{m}\delta t + O(\delta t^4) \tag{2.8}$$

$$\vec{r}(t+\delta t) = \vec{r}(t) + \vec{v}(t+\frac{\delta t}{2})\delta t + O(\delta t^3) \tag{2.9}$$

Each cycle of integration requires the calculation of $\vec{r}(t+\delta t)$, $\vec{v}(t+\frac{\delta t}{2})$, and of the acceleration $\frac{\vec{F}}{2m}$. Velocities at time t, needed to calculate kinetic energy (and consequently other properties such as temperature and pressure) are obtained by:

$$\vec{v}(t) = \frac{\vec{v}(t+\frac{\delta t}{2}) + \vec{v}(t-\frac{\delta t}{2})}{2} \tag{2.10}$$

The *leap-frog* algorithm has the advantage of providing a direct method to calculate the velocity which provides a better precision and a way of direct control of the temperature of the system (by calculating the kinetic energy).

2.2 Boundary Conditions

The common way to minimize edge effects in the simulated (finite) system in MD is to adopt periodic boundary conditions (PBC) when surface effects are not of interest. In a system made up of 1000 atoms arranged in a $10 \times 10 \times 10$ cube, nearly half the atoms are on the outer faces, and these will have a large effect on the measured properties. Even for a system consisting of 10^6 atoms, the surface atoms amount to 6% of the total, which is a relevant part of the whole assembly. Surrounding the simulation box with replicas of itself takes care of this problem. Provided the potential range is not too long, the *minimum image convention* assures that each atom interacts with the nearest atom or image in the periodic array. In the course of the simulation, if an atom leaves the basic simulation box, attention can be switched to the incoming image.

Calculations exploiting PBC are rather expensive, since they require a space-filling box once it is replicated across each dimension. Moreover, it is worth noting that the imposed artificial periodicity can lead to artifacts when considering properties which are influenced by long-range correlations. Among them, effects on the counter ion distribution, conformational equilibria and energetic bias in the simulation of charged systems can be included. Special attention must be paid to the case where the potential range is not short: for example for charged and dipolar systems. Usually long range forces are not calculated beyond a certain cutoff distance to save computational time. Outside the cutoff

range the forces are calculated using lattice sum methods (like the Ewald sum method, see next section) or by a simple truncation. The use of the minimum image convention implies that the cutoff radius cannot exceed half the box side (or half the shortest box vector for a non cubic box). To further reduce the simulation time a list of neighbors is often used: this is a list of all the atoms in the cutoff range of atoms i. Because in liquid systems atoms are continuously entering or leaving the cutoff sphere the neighbor list is compiled using a greater range and updated every few steps.

To avoid PBC-induced spurious effects, non-periodic boundary conditions (NPBC) can be employed. The hard part of these models is related to the proper description of edge effects introduced by the presence of an artificial confinement of the system. To this end, restraints to the sphere boundary atoms or a proper modeling of the interactions between the simulated system and the wall of the cavity via elastic collisions have been evaluated in the literature [85–88].

2.3 Long range electrostatics

Suppose a system of N positively and negatively charged particles, located in a cube with side L. The system is periodic (i.e., PBC are used) and as a whole is electrically neutral, i.e. $\sum_i q_i = 0$ The Coulomb potential here can be written as:

$$V_{Coul} = \frac{1}{2}\sum_{i=1}^{N} q_i \Phi(\vec{r}_i) \tag{2.11}$$

$$\Phi(\vec{r}_i) = \sum_{j,\mathbf{n}}{}' \frac{q_j}{|\vec{r}_{ij} - \mathbf{n}L|} \tag{2.12}$$

where Φ is the electrostatic potential at the position of ion i and the prime on the summation indicates that the sum is over all periodic images n and over all particles j, except $j = i$ if $n = 0$. Equation 2.11 cannot be used to compute the electrostatic energy, because the sum is only conditionally convergent. Now every particle charge q_i are assumed to be surrounded by a diffuse charge distribution (say a Gaussian distribution) of the opposite sign, such that the total charge of this cloud exactly cancels q_i. In that case the electrostatic potential due to particle i is due exclusively to the fraction of q_i that is not screened. At large distances, this fraction rapidly goes to 0, and the contribution to the electrostatic potential at a point n due to a set of screened charges can be easily computed by direct summation. However, a correction for the screening charge cloud on every particle must be added. This is equal to adding a smooth charge density. There are three contributions to the electrostatic potential: i) the one due to the point charge q_i, ii) the one due to the (Gaussian) screening charge cloud (with charge q_i), and iii) the one due to the compensating charge cloud with charge q_i. If the Coulomb self interactions are not excluded (correcting for them afterwards) the compensating charge distribution is not only a smoothly varying function, but it is also periodic. Such a function can be represented by a (rapidly converging) Fourier series, and thus can be easily evaluated in a numerical implementation. The single slowly-converging sum of equation 2.11 has been converted into two quickly-converging terms: the Fourier series for the screening and compensating charge clouds and the direct sum of the screened charges. The use of a fixed cutoff range makes the Fourier part of the Ewald summation scale as $\mathcal{O}(N^2)$ making the technique inefficient for large systems. To improve computational efficiency it is possible to apply discrete Fast Fourier Transform methods, distributing the charges on a mesh. This

significantly decreases the computational cost. Widely used mesh based approaches are the PPPM (Particle-Particle Particle-Mesh) and PME (Particle Mesh Ewald) techniques [89, 90].

2.4 Geometric constraints

In order to gain computational time, the time step can be increased by freezing the fastest molecular motions, such as the intramolecular vibrations and rotations. This can be achieved imposing a set of geometrical constraints that keep the constrained distances and angles in a threshold from a given value. One of the most common methods used to restrain molecular motion is the SHAKE algorithm [91]. At each simulation step, after the integration of motion have been calculated, SHAKE transforms the set of non constrained coordinates $\{r'\}$ in the set of constrained coordinates $\{r''\}$; this is done by the Lagrange multipliers method. If σ_k is the generic equation of a constrain, then

$$\sigma_k(\vec{r}_1,\ldots,\vec{r}_K) = 0 \qquad k = 1,\ldots,K \tag{2.13}$$

$$\vec{f}_i = -\frac{\partial}{\partial \vec{r}_i}\big(V + \sum_{k=1}^{K} \lambda_k \sigma_k\big) \tag{2.14}$$

$$\vec{g}_i = -\sum_{k=1}^{K} \lambda_k \frac{\partial \sigma_k}{\partial \vec{r}_i} \tag{2.15}$$

where λ_k is the Lagrange multiplier, $\vec{f}_i$ is the generalized force and $\vec{g}_i$ is the force associated to a constrain. The use of SHAKE involves the resolution of a system of K second order equations , neglecting the quadratic terms λ_k^2. Water molecules often make up to the 80% of atoms in a simulation box and, for this reason, a non iterative version of SHAKE (called SETTLE [92]) has been optimized to be used with water.

Another method used to impose constraints is the LINCS [93] algorithm. It is a non iterative method (it takes two steps only) based on the resolution of matrix equations. LINCS is faster and more stable than SHAKE but can be applied only on bond lengths or isolated bond angles, such as the $\widehat{\mathrm{HOH}}$ angle in a water molecule, and thus can be used only for small molecules.

2.5 Starting conditions

To start a MD simulation an initial set of coordinates and velocities of all the atoms in the system is needed. Furthermore the set of molecular bonds and angles of all the molecules present (i.e. the topology of the system) may be specified. To avoid the superposition of atoms a structure from a precedent simulation or from a crystalline structure is used. The initial velocities, if not available, are generated using a Maxwellian distribution at the chosen temperature of simulation:

$$p(\vec{v}_i) = \sqrt{\frac{m_i}{2\pi k_B T}}\, exp\Big(-\frac{m_i v_i^2}{2k_B T}\Big)$$

$p(\vec{v_i})$ is the probability for particle i to have velocity $\vec{v}_i$. The integration time step choice is based on the rate of the fastest process that takes place in the simulation: the integration algorithms are based on the assumption that the average velocity in the time interval $t(t+\delta t)$ is equal to the instantaneous

velocity in $t + \frac{\delta t}{2}$. The time step is usually set at 0.1 times the relaxation time of the fastest process in the system, e.g. if molecular vibrations and rotations are considered a time step of one or few femtoseconds is used (the relaxation times being in the range of $10^{-11} - 10^{-14} s$). At the same time the time step is chosen as long as possible to minimize the computational cost and perform a simulation with a greater number of atoms and/or a longer simulation time. Table 2.2 shows the complete configuration update algorithm including the application of geometric constraints.

Update algorithm
Given: atomic positions, $\vec{r}$, at time t, atomic velocities, $\vec{v}$, at time $t - \frac{\Delta t}{2}$, accelerations $\frac{\vec{F}}{m}$ at time t, (constraints are neglected) total kinetic energies and virial. $\Downarrow$ **1.** Compute scaling factors μ e λ (see sec. 2.6) $\Downarrow$ **2.** Velocities are updated and scaled as λ: $\vec{v}' = \lambda(\vec{v} + a\Delta t)$ $\Downarrow$ **3.** Non constrained positions are computed: $\vec{r}' = \vec{r} + \vec{v}'\Delta t$ $\Downarrow$ **4.** Constrains are applied (with the SHAKE or LINCS algorithm); new coordinates $\vec{r}''$ $\Downarrow$ **5.** Velocities are corrected for constraints $\vec{v} = (\vec{r}'' - \vec{r})/(\Delta t)$ $\Downarrow$ **6.** Atomic coordinates and box dimensions are scaled: $\vec{r} = \mu\vec{r}''$;$\mathbf{b} = \mu\mathbf{b}$

TABLE 2.2: configuration update algorithm.

2.6 Temperature and Pressure control

In MD simulations the integration of the equation of motion keeps the total energy (i.e. the simulation is performed in a microcanonical ensemble). This type of simulation is not well suited to be compared with experimental data that is usually obtained at constant temperature and pressure. Moreover, due to the approximations used to perform simulations, such as the cutoff long range interactions, there is a need to monitor the temperature and pressure of the system. The temperature T of a system with N_{df} degrees of freedom is a function of the kinetic energy of the atoms:

$$E_{kin}(t) = \sum_{i=1}^{N} \frac{1}{2} m_i \vec{v}_i^2(t) = \frac{1}{2} N_{gl} k_B T(t) \tag{2.16}$$

where k_B is the Boltzmann constant, and $T(t)$ is the time dependent temperature. The thermal capacity per degree of freedom allows to relate the total kinetic energy and the temperature: if (C_V^{df}) is the

thermal capacity per degree of freedom, then

$$\Delta E_{kin}(t) = N_{df} C_V^{df} \Delta T(t) \tag{2.17}$$

The control of the temperature is performed by a weak coupling with an external thermal bath at the temperature T_0. If the temperature drifts from T_0 it is slowly corrected using the following equation (τ is a time constant)

$$\Delta E_{kin}(t) = N_{df} C_V^{df} \Delta T(t) \tag{2.18}$$

This method is known as the Berendsen's thermostat [94] and has the advantage of being able to change the strength of the coupling between the system and the external bath (usually a short τ, about the integration time step, is used during the equilibration phase, while a longer one is used during the sampling). The heat flow to or from the system takes place changing the λ parameter:

$$\Delta E_{kin}(t) = [\lambda^2 - 1]\frac{1}{2} N_{df} k_B T(t) \tag{2.19}$$

$$\lambda = \left[1 + \frac{\Delta t}{\tau_t}\left(\frac{T_0}{T_t} - 1\right)\right]^{-1/2} \tag{2.20}$$

where $\tau_t = \frac{\tau k_B}{2C_V}$; τ_t is different from τ (for aqueous solutions the ratio is usually $\tau/\tau_t = 3$). The same relaxation method may be applied to monitor the pressure (which can be calculated fomr the virial) in the simulation box, using the isothermal compressibility to correlate the changes pressure and volume at time t.

The Berendsen bath method is very efficient for relaxing a system; however it does not generate a canonical surface. To resolve this problem methods that allow to keep the temperature and the pressure constant while generating a canonical surface have been developed, such as the velocity-rescale method [95] for the temperature and the Parrinello - Rahman scheme [96] for pressure. This methods make use of an extended Hamiltonian; the coupling with an external bath is achieved adding a friction term to atomic velocities with a friction constant that is function of the current difference between the actual and target parameter being monitored (i.e. pressure or temperature). It is noteworthy to observe that with an extended Hamiltonian method an oscillating relaxation, which is very different from the damped exponential relaxation (that is, the result of the Berendsen scheme and methods of the first type), is obtained, and more time steps to reach equilibrium are needed.

Chapter 3

Force field development

The following sections focus on two procedures made up of several algorithms devoted to the *ad hoc* parameterization of chemical systems which have been used within this thesis work: the Joyce and the LRR-DE procedures. These two methods are complementary: Joyce is in fact devoted to the parameterization of the intramolecular part of a force field, while LRR-DE optimizes at present the non-bonded part of a pair-wise model.

3.1 Joyce

Joyce [81] is a force field parameterization scheme which performs bonding parameters optimization using QM data as reference. Recently, the program has been provided of a user-friendly GUI, called Ulysses [9].

3.1.1 An overview

The intramolecular potential in Joyce has the general form

$$\begin{aligned} E_{intra} = & \sum_{\mu \in bonds} \frac{1}{2} k^{s}_{\mu} (b_{\mu} - b^{eq}_{\mu})^2 + \\ & + \sum_{\mu \in angles} \frac{1}{2} k^{\theta}_{\mu} (\theta_{\mu} - \theta^{eq}_{\mu})^2 \\ & + \sum_{\mu \in R_{tors}} \frac{1}{2} k^{\phi}_{\mu} (\phi_{\mu} - \phi^{eq}_{\mu})^2 \\ & + \sum_{\mu \in F_{tors}} \sum_{j}^{N^{\mu}_{cos}} k^{\delta}_{j\mu} (1 + cos(n^{\mu}_{j} \delta_{\mu} - \gamma^{\mu}_{j})) \\ & + \sum_{i,j \in atoms} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{i,j \in atoms} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \end{aligned} \tag{3.1}$$

The meaning of the single terms have been already illustrated in Chapter 1; i, j run over atoms, while R_{tors} and F_{tors} indicate stiff and flexible torsions, respectively.

Given a molecule of interest, QM geometry optimization is performed initially. Such step is usually performed at the DFT level. The program use the energy, gradients and Hessian matrix (i.e. energy second derivatives with respect to the nuclear displacements) computed on the located global minimum for the force field optimization. The other input needed by the program is a selection of internal coordinates (ICs) which consist in all bond stretches, angle bendings and dihedral torsions that can

be obtained from a given connectivity criteria referred to the reference conformation (here, the global minimum). The chosen RICS will be used during the parameters fitting procedure. By default, Joyce creates by itself a complete set of internal coordinates of the molecule. The input topology file may contain non-bonded parameters (atomic charges and LJ - σ and ϵ - parameters), altough these values are not optimized. Non-bonded parameters can be easily transferred from literature, or, as in the case of point charges, re-computed at the QM level, using for instance one of the procedure outlined in section 1.1.2.

QM input data obtained from the minimum energy conformation are enough to correctly optimize the harmonic term of Eq. 1.2, thus to describe with some accuracy the molecular system close to the minimum. Concerning flexible dihedral, which are modeled as a sum of cosine functions (see Eq. 1.6) , harmonic approximation may be insufficient, and additional data, as the one computed along a whole dihedral angle scan, are required. Therefore, when flexible dihedral angles are present, energy scan along such coordinates are performed: the torsion is varied from -180° to 180°, spaced by fixed intervals, and the whole geometry is relaxed while keeping freezed at the chosen value the considered dihedral angle.

The Joyce merit function has the following form

$$I^{intra} = \sum_{g=0}^{N_{geom}} \left[W_g \left[U_g - E_g^{intra} \right]^2 \right] + \sum_{K}^{3N-6} W' \left[G_K - \left(\frac{\partial E^{intra}}{\partial Q_K} \right) \right]_{g=0}^2 + \\ + \sum_{K \leq L}^{3N-6} \frac{2W''_{KL}}{(3N-6)(3N-5)} \left[H_{KL} - \left(\frac{\partial^2 E^{intra}}{\partial Q_K \partial Q_L} \right) \right]_{g=0}^2 \tag{3.2}$$

Here, K, L run over the normal coordinates, N_{geom} is the number of sampled conformations; U_g is the energy difference between the energy of the $g_t h$ conformation and the one computed on the global minimum ($g = 0$). G_K is the energy gradient with respect to the normal coordinate K, while H_{KL} is the Hessian matrix with respect to K and L. Both G_K and H_{KL} are evaluated at $g = 0$. The constants W , W' and W'' weight the several terms at each geometry and can be chosen in order to drive the results depending on the circumstances. The energy, gradient and Hessian terms are normalized in order to account for the different number of terms and to make the weights independent from the number of atoms in the molecule.

The minimization of Eq. 3.2 leads to a linear problem, so that k_{ij}^s, k_{ijk}^θ and k_{ijkl}^ϕ are analytically derived. Equilibrium values instead are simply measured on the minimum geometry. The first term is evaluated only if flexible dihedral angles are intended to be parametrized: in such case, N_{geom} corresponds to the number of scanned geometries submitted to partial QM optimization. Such process is evaluated under the Frozen Internal Rotation Approximation (FIRA), which assumes that no relevant geometry rearrangements are experienced by the molecule during the scan, except for the scanned dihedral itself.

3.1.2 Current compatibilities

The code reads QM input files obtained through the Gaussian software [97]: in particular, energies, gradients and Hessian matrix are read from a formatted *.fchk* file. The definition of all the ICs setting up the force field are retrieved from a GROMACS [98] *.top* topology file. As main output file, a topology

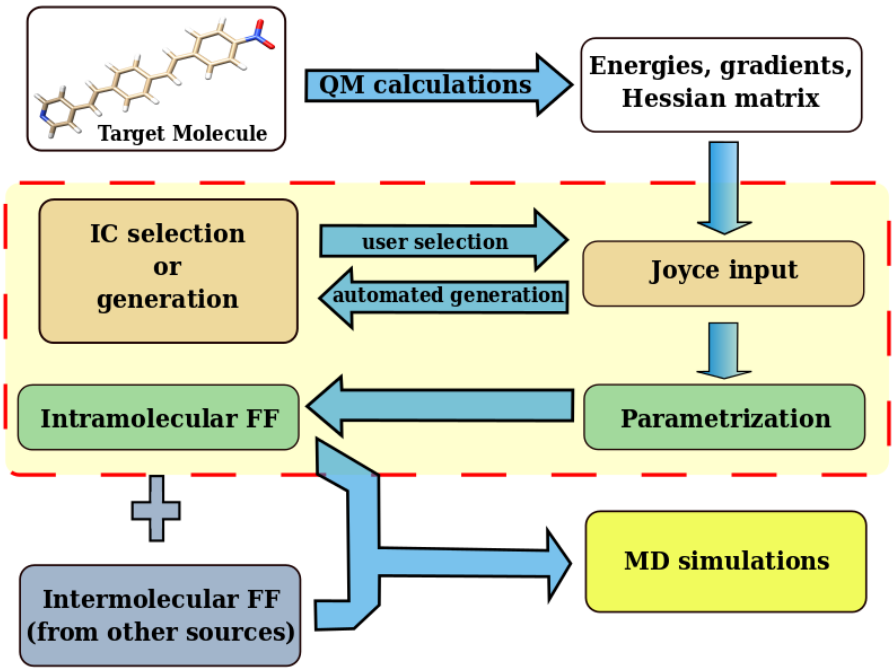


FIGURE 3.1: Flowchart of the Joyce protocol.

file of the molecule under examination with the intramolecular optimized parameters is written again in the GROMACS topology format.

3.1.3 Parametrization protocol and validation

In this thesis, Joyce has been used for the parameterization of force field specifically tailored for single molecules (Chapters 7 and 8). The following protocol has been applied:

1. Global minimum of the investigated molecule is located by means of DFT computations. Energy, gradients and Hessian matrix are computed on the obtained geometry. Moreover, CM5 point charge are computed. Environment effects are taken into account by means of the Polarizable Continuum Model (PCM) [99]. The inclusion of polarization effects due to the surrounding environment allows for a better description for the species to be parametrized within the environments where they are intended to be simulated. However, the trasferibility of the developed force field in other solvents, especially in those with significant different dielectric constants from the one used in the continuum model, is reduced. The choice of the functional, as well as the basis set, depends on the nature of the system. B3LYP exchange-correlation functional has been mainly used, because of its good compromise between accuracy and computational cost. Such step is performed using the Gaussian software.

2. In the case of molecules with flexible torsions, which are known to affect molecular conformations, energy scans around each of them are performed. This step can be not applied to CH_3 groups and similar, since their influence on the global structure determination is poor. Dihedral angles are varied in steps of 30° (or less), and optimizing the obtained structure while keeping

Molecules	GAFF [59]	Seminario [101]	Parafreq [100]	Joyce
H-peroxide	-	-	65.6	49.8
biphenyl	130.0	80.7	62.2	62.9
bicyclo[2.2.2]octane	73.5	129.4	68.6	52.9
nitro-pyridine	156.5	128.7	78.7	63.2
toluensulfonic acid	158.8	108.4	46.7	41.6

TABLE 3.1: Frequency related standard deviations (cm^{-1}) with respect to QM reference values obtained from MM vibrational frequencies computed with different force fields. All the reported values are taken from Ref. [100].

freezed the considered soft variable. The employed level of theory is the same as the one used in the previous step.

3. The Joyce program is used to generate and accurately select a suitable set of internal coordinates to be parametrized. In particular the definition of bond, angles, torsions and improper dihedral angles is performed at this step. A redundant set of coordinates (RICS) can also be employed.

4. Only harmonic terms are parametrized in first instance. Dependencies are applied on identical coordinates (i.e., those coordinates defined by the same atom types or which are equal for symmetry reasons), in order to describe such variables with the same set of parameters.

5. A second parameterization step is performed. Here, only flexible dihedral angels are parametrized. The number of cosine functions to be used for each torsion is established by the user, as well as the corresponding multiplicity value: these choices have to be made in order to reproduce the related reference QM energy scan. Identical dihedral angles within the same molecule are refined only once, and the obtained parameters are transferred to the others. The harmonic part terms are constrained to the values computed during step 4.

6. The obtained force field is validated by inspecting at first the square root of I^{intra} (Eq. 3.2), minimized by Joyce during the parameterization. Then, the normal mode wavenumbers, computed at the force field level, are compared to the QM ones, by computing the standard deviation. In the case of flexible dihedral parameterization, the force field energy scan is compared to the single-point DFT energies to assess the quality of the fit.

The Joyce algorithm has been validated in Ref. [9] by looking at the predicted vibrational frequencies and modes on a set of selected systems. In the target set both rigid (hydrogen peroxide, biphenyl, bicyclo[2.2.2]octane) as well as flexible (nitro-pyridine and toluensulfonic acid) molecules have been included, thus providing a heterogenous dataset. Moreover, this set was investigated in a previous work with available force fields [100], allowing therefore for a direct comparison with previous tests. The molecules have been parametrized following the procedure previously itemized, with few differences: in particular, no environment descriptions were provided during the QM computations, so to increase transferability of the developed models among multiple solvents.

Through the inspection of Table 3.1 it is undeniable that the Joyce force fields are able to reproduce rather accurately the vibrational behavior of the target molecules, giving frequencies in agreement or even better than the ones provided by standard force fields and other parameterization methods. In order to properly describe the flexibility of the floppy molecules included in the set, corresponding

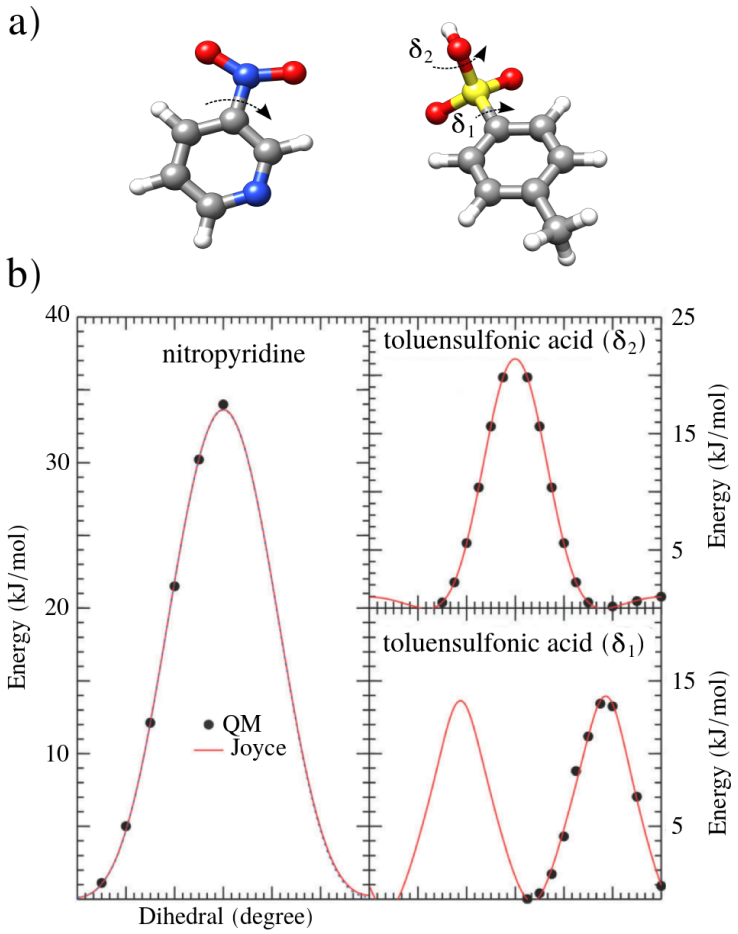


FIGURE 3.2: a) Structure of nitropiridine and toluensulfonic acid, with flexible dihedral angles highlighted. b) Torsional profiles comparison between Joyce and QM counterparts for the two flexible molecules. For toluensulfonic acid, dihedrals δ_1 and δ_2 are those indicated in the top panel.

soft dihedral angles have been parametrized according to the workflow discussed above. Only two molecules are considered in this regard: nitropiridine and toluensulfonic acid (corresponding structures are shown in 3.2a). The accordance for these molecules with the reference QM data is shown in Figure 3.2b: the computed energy profiles correctly account for both conformation minima and barriers along the investigated torsions.

3.2 LRR-DE

The linear ridge regression-differential evolution (LRR-DE) is a procedure able to optimize the non-bonded part of a FF without altering the functional form and the parameters of the FF of the other atoms of the system. Thanks to such approach, the obtained models can be integrated into consolidated MM packages and FFs. The method exploits a combination of machine learning techniques, that in the recent years are increasingly finding applications in computational chemistry [102–104]. Inspired by recent works [105], LRR-DE combines the linear ridge regression technique with differential evolution [106, 107], a metaheuristic optimization algorithm which is effective in the exploration of high dimensionality search spaces. In particular, LRR-DE uses linear ridge regression to optimize

the linear parameters of a tunable model and differential evolution to optimize the non-linear parameters, minimizing the leave-one-out cross-validation error (*vide infra*). In the most general form of the methodology *ab initio* forces and energies of sampled configurations are used as reference output, leading to a multi-objective optimization problem. Some of the features which characterize the proposed method, such as a regularized multi-objective cost function, aimed to prevent from overfitting, and the ability of optimize either linear and non linear parameters, are already implemented in the ForceBalance tool [13] already mentioned in Section 1.2.6. However, significant novelties can be outlined. The combination of algebraic techniques and metaheuristics employed in LRR-DE, using the LOOCV error as criterion to optimize the non-linear parameters, enforces the protection from the overfitting with respect to the training set data and increases the efficiency in finding the global minimum in the parameters space. Moreover, the weights which tune the contribution of the single objective functions are predetermined in ForceBalance. In contrast, the proposed protocol introduces the optimization of the weights so as to allow for error checking in order to obtain the most balanced compromise solution. An high level flowchart of the algorithm is shown in Figure 3.3. In the following, the main features and capabilities are discussed in some detail. The algorithm has been validated using the parameterization of the non-bonded models of metal ions (Zn^{2+}, Ni^{2+}, Mg^{2+}, Ca^{2+} and Na^{+}) in water as test case. The whole validation study is shown in Chapter 9.

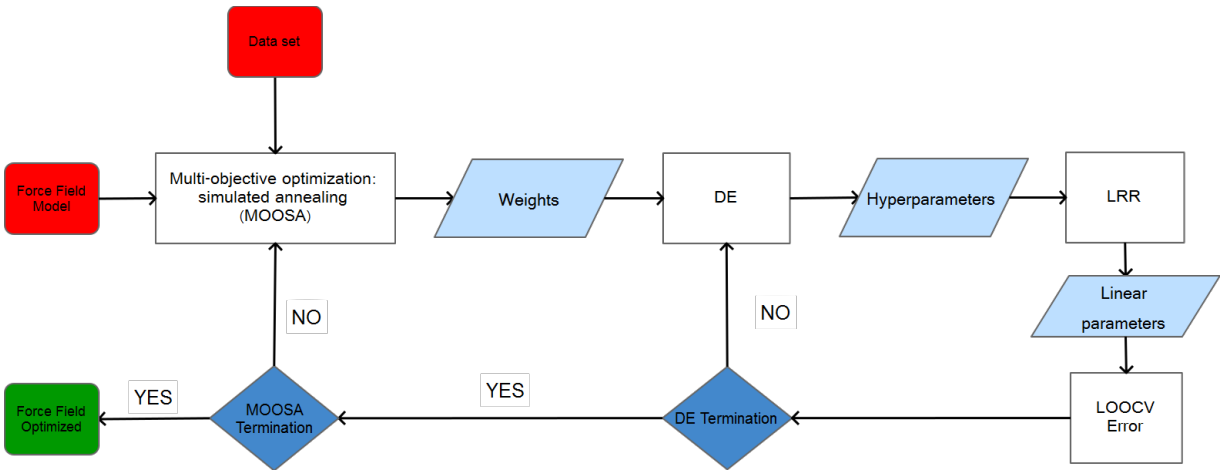


FIGURE 3.3: High level flowchart of the proposed algorithm.

3.2.1 The Linear Ridge Regression Differential Evolution procedure

Given a data set $\{\mathbf{x}_l, y_l\}$, where $\mathbf{x_l}$ is the l-th input vector and y_l the corresponding output value, an interpolative general model can be built as linear combination of the functions $\varphi(\mathbf{x},\theta)$, called *predictors* or *descriptors* in the language of the statistical learning:

$$y_{est} = \sum_{j}^{N_{functions}} C_j \varphi_j(\mathbf{x},\boldsymbol{\theta}_j) \tag{3.3}$$

where $\{\mathbf{C}\}$ and $\{\boldsymbol{\theta}\}$ are the linear and non-linear parameters of the model respectively. In the linear ridge regression technique [108–112], the optimal linear parameters are obtained minimizing the regularized cost function

$$J = \frac{1}{2M}\sum_{l}^{M}\left(y_l - \sum_{j}^{N_{functions}} C_j\varphi_j(\mathbf{x}_l,\boldsymbol{\theta}_j)\right)^2 + \lambda \sum_{j}^{N_{functions}} C_j^2 \tag{3.4}$$

where M is the size of the data set and λ is the regularization parameter. The introduction of the regularization term prevents the overfitting of the model penalizing high values of the linear parameters. In order to evaluate properly the regularization term, all the predictors are scaled with respect to the relative standard deviations

$$\widetilde{\varphi}_j(\mathbf{x}_l,\boldsymbol{\theta}_j) = \frac{\varphi_j(\mathbf{x}_l,\boldsymbol{\theta}_j)}{\sqrt{\frac{1}{M}\sum_{l}^{M}\left(\varphi_j(\mathbf{x}_l,\boldsymbol{\theta}_j) - \overline{\varphi_j}(\mathbf{x}_l,\boldsymbol{\theta}_j)\right)^2}} \tag{3.5}$$

The minimization of the cost function (eq. 3.4), in the scaled form, can be performed analytically solving the system of linear equations

$$\frac{\partial J}{\partial \widetilde{C_j}} = \sum_{l}^{M}\left(y_l - \sum_{j}^{N_{functions}} \widetilde{C}_j\widetilde{\varphi}_j(\mathbf{x}_l,\boldsymbol{\theta}_j)\right)\widetilde{\varphi}_j(\mathbf{x}_l,\boldsymbol{\theta}_j) + 2M\lambda\widetilde{C_j} = 0 \tag{3.6}$$

and the solutions are given by the normal equation

$$\widetilde{\mathbf{C}} = \left(\mathbf{H}^T\mathbf{H} + 2M\lambda\mathbf{I}\right)^{-1}\left(\mathbf{H}^T\mathbf{y}\right) \tag{3.7}$$

where $\mathbf{H}$ is the $M \times N_{functions}$ matrix of the scaled descriptors, $\mathbf{I}$ is the $N_{functions} \times N_{functions}$ identity matrix, and $\mathbf{y}$ is the vector of the output.

The evaluation of the equation 3.7 can be performed if the values of λ and $\{\theta\}$ parameters have been previously established, therefore they are considered as hyperparameters. In order to obtain the optimal values of the hyperparameters, the criterion here employed is the minimization of the cross-validation error.

Cross-validation

Cross-validation (CV) [112] is a resampling method applied in statistical learning for the model assessment and model selection. In order to estimate the accuracy of a regression model on observations not included in the training set, a test set of instances should be available. However, this is usually not the case. CV overcomes this obstacle executing multiple fittings of subsets of the training set and evaluating the errors on the remaining data. In the k-fold CV, the data set is randomly split into k equally sized subsets. Each of these subsets is used in turn as a test set, while the remaining $k-1$ are used for the training. Therefore, k models are built, each one provides a validation error averaging the deviations of the predictions with respects to the data point of the corresponding test set. The cross-validation error is computed as the mean of the k validation errors. An illustrative scheme of the cross validation technique is shown in Figure 3.4.

When k is equal to the number of the instances of the data set, the case is called *leave-one-out* cross-validation (LOOCV). LOOCV provides an approximated unbiased prediction of the expected test error, because the training sets of the subsets are almost identical to the general training set. In statistical learning, the minimization of the LOOCV error is a standard criterion to optimize the hyperparameters of the model. The LOOCV error is computed as

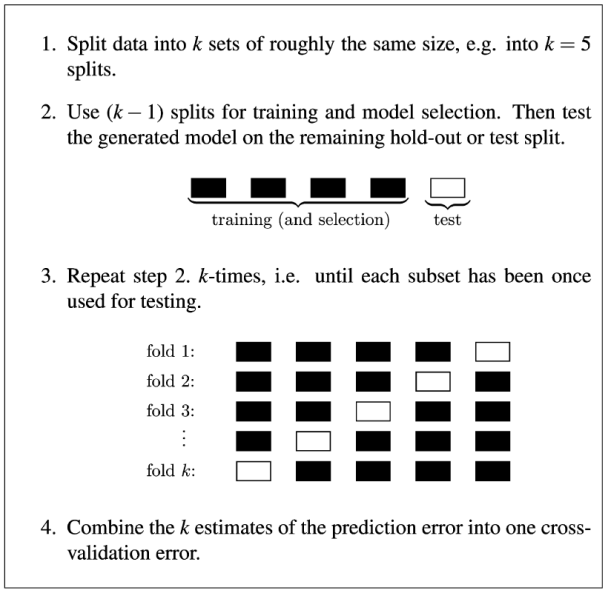


FIGURE 3.4: Cross-validation scheme. Excerpt from the paper of Hansen et al. [102]

$$LOOCV_{error}(\lambda,\{\theta\}) = \frac{1}{M}\sum_{l}^{M}\left(y_l - y_{est}^{(-l)}(\mathbf{x}_l,\lambda,\{\boldsymbol{\theta}\})\right)^2 \tag{3.8}$$

here $y_{est}^{(-l)}(\mathbf{x_i},\lambda,\{\theta\})$ is the prediction for the l-th instance, using the model trained with all the data except the l-th instance. The equation 3.8 represents the mean squared error (MSE); the mean absolute error (MAE) can be equally used, nevertheless the MSE is more sensitive to the outliers, therefore it is a better choice to reduce the occurrence of large errors of the model. Calculating this estimate can be computationally demanding because it requires to repeat the resolution of equation 3.7 M times. However, for the linear ridge regression method the following relationship holds [113]

$$LOOCV_{error}(\lambda,\{\theta\}) = \frac{1}{M}\sum_{l}^{M}\left(\frac{y_l - y_{est}(\mathbf{x}_l,\lambda,\{\boldsymbol{\theta}\})}{1-h_l}\right)^2 \tag{3.9}$$

where $y_{est}(\mathbf{x_l},\lambda,\{\theta\})$ is the prediction of the model trained with the complete data set for the l-th instance, and h_i is the leverage defined as

$$h_l = \frac{1}{M} + \frac{(\boldsymbol{\varphi}_l - \bar{\boldsymbol{\varphi}})^2}{\sum_{l'}^{M}(\boldsymbol{\varphi} i_{l'} - \bar{\boldsymbol{\varphi}})^2} \tag{3.10}$$

The formula 3.9 reduces of a factor M the computational cost of the estimate of $LOOCV_{error}$, nevertheless an efficient method is necessary to sample the hyperparameters space: each evaluation of $LOOCV_{error}$ in fact involves the calculation of the elements of the **H** matrix (unless $\{\theta\} = \emptyset$) and the solution of the normal equation (3.7).

Optimization of the hyperparameters using differential evolution

The minimization of $LOOCV_{error}$ (eq. 3.9) with respect to the hyperparameters λ and $\{\theta\}$ is a non convex optimization, therefore a metaheuristic algorithm is necessary to search for the global minimum of

Initialize population $\{\boldsymbol{\rho}_i,\ i = 1...NP\}$

$G \leftarrow 0$
while not converged **and** $G < G_{max}$ **do**
  for $i = 1...NP$ **do**
    randomly select $\boldsymbol{\rho}_a$, $\boldsymbol{\rho}_b$, $\boldsymbol{\rho}_c$ ($a \neq b \neq c$) from population
    draw random integer j_{rand} between 1 and D
    for $j = 1...D$ **do**
      if rand[0,1] < CR **or** $j = j_{rand}$ **then**
        $\mathbf{u}[j] \leftarrow \boldsymbol{\rho}_a[j] + F \cdot (\boldsymbol{\rho}_b[j]$ - $\boldsymbol{\rho}_c[j])$
      else
        $\mathbf{u}[j] \leftarrow \boldsymbol{\rho}_i[j]$
      end if
    end for
    if $f(\mathbf{u}) < f(\boldsymbol{\rho}_i)$ **then**
      $\boldsymbol{\rho}_i \leftarrow \mathbf{u}$
    end if
  end for
  $G \leftarrow$ G + 1
end while

TABLE 3.2: Pseudocode of DE/rand/1/bin. "←" is the assignment operator.

the objective function. To accomplish this task, the procedure here presented exploits the evolutionary algorithm known as differential evolution (DE) in its basic version DE/rand/1/bin. DE has been proved to be very competitive in benchmarks tests [114] and in real world applications [115] compared to other global optimization algorithms. Moreover it offers the great advantage of providing stable performances varying the parameters on which it depends. This technique has been already applied to the optimization of hyperparameters for a support vector machine classifier [116] providing better results than grid search and particle swarm optimization algorithms. Recently DE has been identified as convenient global optimizer also in computational chemistry [117, 118].

DE is a population-based derivative-free algorithm that consists of three main steps: mutation, crossover, and selection. After a set $\{\boldsymbol{\rho}_i\}$ of NP trial solution vectors defined in the domain of the objective function is randomly initialized, the three steps of the algorithm proceed iteratively on each vector $\boldsymbol{\rho}_i$ (called target vector) of the population until a tolerance criterion is satisfied. In the mutation step, a *donor* vector is created through the differential mutation operation

$$\boldsymbol{\nu}_j = \boldsymbol{\rho}_a + F(\boldsymbol{\rho}_b - \boldsymbol{\rho}_c) \tag{3.11}$$

where F is a parameter of the algorithm called *differential weight* and the indices a, b, c are chosen randomly with the condition $a \neq b \neq c \neq i$. This implies that the size of the population must be larger than four units.

In the crossover step, the donor vector exchanges its components with $\mathbf{p_i}$. According to the binomial scheme, the crossover is performed following the rule

$$u_{i,j} = \begin{cases} v_{i,j} & \text{if } U(0,1)_{i,j} \leq CR \quad \text{or} \quad j = j_{rand} \\ \rho_{i,j} & \text{otherwise} \end{cases} \tag{3.12}$$

where $U(0,1)_{i,j}$ is a random number selected from an uniform distribution in the range $[0,1]$ and CR is a parameter of the algorithm called *crossover rate*; j_{rand} is an integer random number between 1 and D (being D the dimension of the vector). In all the applications conducted in this work, F and CR have been set to 0.7 and 0.85, respectively, as result from calibrations on some test cases.

In the selection step, the objective function (f, in Table 3.2) is evaluated in $\mathbf{u}$. If the new vector yields a lower or equal value than $\boldsymbol{\rho}_i$, it will replace the target vector in the next generation.

When the termination condition is satisfied the best solution provides the optimal hyperparameters.

Properties of LRR-DE

The LRR-DE procedure is a method capable of reproducing data by optimizing the parameters, both linear and non-linear, of a model chosen by the user. The application of the regularization and cross-validation protects the optimization from overfitting. The DE algorithm guarantees high efficiency in the search of the optimal hyperparameters. These features make LRR-DE suitable to optimize the parameters of physical models with respect to experimental or *ab initio* data.

As a simple illustrative example, the LRR-DE method is applied to the fitting of the potential energy curve of the $Zn^{2+}\cdots H_2O$ interaction, calculated at the MP2/aug-cc-pVTZ level, as function of the only variable d, the interatomic distance between the zinc ion and the oxygen atom. A training set of 16 points is employed to build models of increasing complexity. The results are collected in Table A.14. In Figure 3.5 the graphical representations of two cases are shown.

The simplest considered model, 12-3, includes a repulsive d^{-12} term analogous to that of the Lennard-Jones potential and an attractive d^{-3} term to account for the charge-dipole interaction. The performance of this model is poor, as can be seen by observing Figure 3.5 (a). The reduction of the error is drastic if the d^{-12} term is substituted by an exponential repulsion. Even better results can be obtained employing a buffered d^{-12} term to describe the repulsion. Both the exponential and buffered d^{-12} terms have a non-linear parameter. Further terms, even without an immediate physical interpretation, can be added to the models to reduce the errors. For instance, in the 12b-3-G and 12b-3b-G models a Gaussian function is included, resulting in a significant performance improvement. This

Model	Linear Parameters	Non-linear parameters	MSE (LOOCV)	MSE (test)
12-3	2	0	719.040	768.565
Exp-3	2	1	3.530	3.526
12b-3	2	1	0.525	0.434
12b-3-G	3	3	0.079	0.068
12b-3b-G	3	4	0.001	0.002

TABLE 3.3: Mean squared errors (MSE), in $(\text{kcal/mol})^2$, for five models optimized to reproduce the MP2/aug-cc-pVTZ potential energy curve of the $Zn^{2+}\cdots H_2O$ interaction. The test errors are calculated with respect to 100 points not included in the training set. The analytical expressions of the models are shown in Table 3.4.

Model	Analytical expression
12-3	$\frac{C_1}{d^{12}} + \frac{C_2}{d^3}$
Exp-3	$C_1 \exp(-\theta d) + \frac{C_2}{d^3}$
12b-3	$\frac{C_1}{(d-\theta)^{12}} + \frac{C_2}{d^3}$
12b-3-G	$\frac{C_1}{(d-\theta_1)^{12}} + \frac{C_2}{d^3} + C_3 \exp\left(-\frac{(d-\theta_2)^2}{2\theta_3^2}\right)$
12b-3b-G	$\frac{C_1}{(d-\theta_1)^{12}} + \frac{C_2}{(d-\theta_2)^3} + C_3 \exp\left(-\frac{(d-\theta_3)^2}{2\theta_4^2}\right)$

TABLE 3.4: Analytical expressions of the models tested in the fitting of the potential energy curve of the $Zn^{2+} \cdots H_2O$

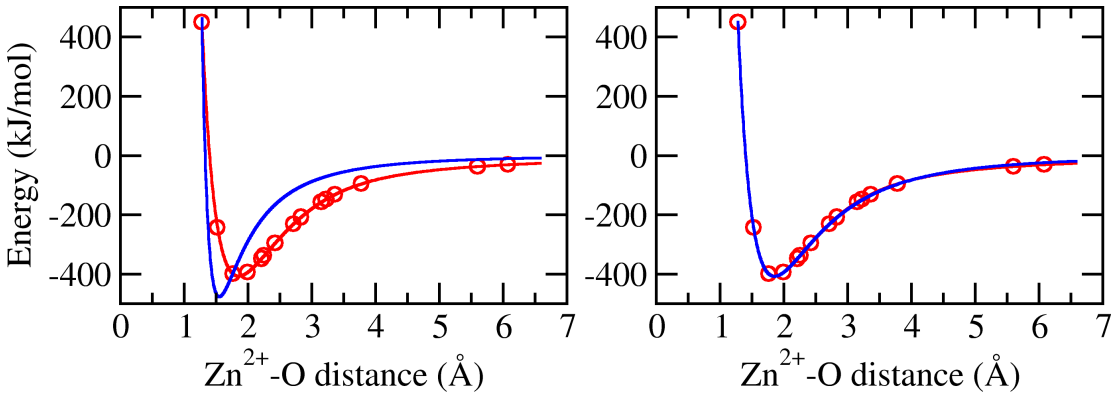


FIGURE 3.5: Graphical representation of the potential energy curves of the models 12-3 (blue line, left) and 12b-3 (blue lines, right), compared with the target data (red line), namely the MP2/aug-cc-pVTZ energies (red lines) for $Zn^{2+} \cdots H_2O$ interaction. The blue circles are the points included in the training set.

simple univariate example highlights the crucial role of the descriptor selection in the outcome of the fitting. In general, the choice of the functional form of the model can be made evaluating the performances in the reproduction of the quantities of reference in relation to the particular operational needs. It is worth noting that LRR-DE does not use constraints in the optimization of the linear coefficients. Therefore the sign of each term, which indicates if it describes a repulsion or an attraction, emerges spontaneously from the optimization and it is not imposed by the user. However, the j-th linear parameter can be fixed to a constant value, K, performing the fitting of the other parameters with respect to the output subtracted of the contribution of the j-th descriptor $(y_l - K\varphi_j(\boldsymbol{x}_l))$. This possibility has been exploited in the validation tests to generate force fields with the electrostatic component defined by the formal charge of the ions. Tighter control can be exerted on the non-linear parameters by defining the lower and upper limits of the search domain.

3.2.2 Single-objective application of the LRR-DE procedure: the force-matching approach

The application of the LRR-DE procedure to the generation of non-bonded force fields can be performed using as target output one or more types of reference quantities, calculated with *ab initio* methods or obtained by the experiments. In this section the single-objective case is illustrated, in which only one type of reference data is used. As stated in Section 3.2, the parameterization of metal ions will be used as validation set of the method; therefore, as case study, the single objective mode is here applied to this kind of systems, using the *ab initio* forces computed on the metal ion as reference. This approach recalls the force-matching method [79, 119, 120] with the important differences that here the

cost function is regularized and the hyperparamaters are tuned to minimize the cross-validation error.

Assuming that the potential of the metal ion is the result of the sum of pairwise potentials with respect to all the other atoms,

$$V_M(\mathbf{R}_l) = \sum_{i}^{N_{atoms}-1} V_{M-i}(\{\mathbf{C}\},\{\boldsymbol{\theta}\},\mathbf{R_l}) \tag{3.13}$$

where $\mathbf{R_l}$ is the l-th configuration of the system, if V_{M-i} is expressed as a linear combination of functions v

$$V_{M-i}(\mathbf{R}_l) = \sum_{j}^{N_{functions}} C_j v_j(\{\boldsymbol{\theta}\},\mathbf{R}_l) \tag{3.14}$$

the k-th component of the molecular mechanics model of force exerted on the metal ($F^{MM}_{M,k}(\mathbf{R_l})$) as a result of interactions with all other atoms is given by

$$F^{MM}_{M,k}(\mathbf{R}_l) = -\frac{\partial\left(\sum_i^{N_{atoms}-1}\sum_j^{N_{functions}} C_j D_{ij} v_j(\{\boldsymbol{\theta}\},\mathbf{R}_l)\right)}{\partial k} \tag{3.15}$$

$$F^{MM}_{M,k}(\mathbf{R}_l) = -\sum_{j}^{N_{functions}} C_j \sum_{i}^{N_{atoms}-1} D_{ij}\frac{\partial v_j(\{\boldsymbol{\theta}\},\mathbf{R}_l)}{\partial k} \tag{3.16}$$

D_{ij} is a characteristic parameter of the i-th atom, assuming that the combination rule for the j-th function is multiplicative.

The model of the force field corresponds to the equation 3.3 if the following identity is set

$$\varphi_j(\mathbf{x},\boldsymbol{\theta}_j) = -D_{ij}\frac{\partial v_j(\{\boldsymbol{\theta}\},\mathbf{R}_l)}{\partial k} \tag{3.17}$$

then the LRR-DE procedure can be applied if the target output are set equal to the *ab initio* forces, $F^{QM}_{M,k}$, and the values of the descriptors are assigned to the elements of the **H** matrix, according to the equation 3.5.

3.2.3 Generalization to the multi-objective fitting

The multi-objective optimization of the parameters exploits simultaneously different types of reference output, for example the *ab initio* forces on the metal ion, the forces on the nearest neighbor atoms from the metal ion, the contribution to the total energy due to the force field to optimize, different levels of the theory for the calculations, and systems of different composition. The simplest way to approach a multi-objective optimization problem is the reduction to a single-objective one building a weighted cost function. In this case the equation 3.4 becomes

$$J = \sum_{b}^{N_b} w_b \frac{1}{2M_b}\sum_{l}^{M_b}\left(y_{l,b} - \sum_{j}^{N_{functions}} \widetilde{C}_j\widetilde{\varphi}_{j,b}(\mathbf{x}_l,\boldsymbol{\theta}_j)\right)^2 + \lambda\sum_{j}^{N_{functions}}\widetilde{C_j}^2 \tag{3.18}$$

where w_b is the scaled weight of the b-th set of targets, of size M_b, calculated as

$$w_b = \frac{w'_b}{\sqrt{\frac{1}{M_b}\sum_l^{M_b}\left(y_{l,b} - \overline{y_b}\right)^2}} \tag{3.19}$$

In the equation 3.19, w'_b are the effective weights, subject to constraints $w'_b \in [0,1]$ and $\sum_b^{N_b} w'_b = 1$. The definition of their values is the topic of the next subsection.

The minimization of the weighted cost function with respect to the linear parameters is given by a normal equation that includes the weights:

$$\widetilde{\mathbf{C}} = \left(\mathbf{H}^T\mathbf{W}\mathbf{H} + 2M\lambda\mathbf{I}\right)^{-1}\left(\mathbf{H}^T\mathbf{W}\mathbf{y}\right) \tag{3.20}$$

being $\mathbf{W}$ the diagonal matrix containing the w_b values. In this case $M = \sum_b^{N_b} M_b$. The equations 3.9 and 3.10 must be modified as follows to take into account the weights

$$LOOCV_{error}(\lambda,\{\theta\}) = \sum_b^{N_b}\left(\frac{1}{M_b}\sum_l^{M_b}\left(\frac{y_l - \varphi_{est}(\mathbf{x}_l,\lambda,\{\theta\})}{1-h_l}\right)^2 w_b\right) \tag{3.21}$$

$$h_l = \frac{1}{M} + \frac{w_l(\boldsymbol{\varphi}_l - \bar{\boldsymbol{\varphi}})^2}{\sum_{l'}^{M} w_{l'}(\boldsymbol{\varphi}_{l'} - \bar{\boldsymbol{\varphi}})^2} \tag{3.22}$$

In order to be used as a reference data in the LRR-DE procedure, a quantity has to be expressed as a linear combination of the v functions or of the their derivatives. In the case of the forces on other atoms, this condition is satisfied by setting

$$y_{l,A,k} = F_{A,k}^{QM}(\mathbf{R}_l) - \sum_i^{N_{atoms}-2} f_{Ai,k}(\mathbf{R}_l) \tag{3.23}$$

and

$$H_{l,j,A,k} = -D_{Aj}\frac{\partial v_j(\{\boldsymbol{\theta}\},\mathbf{R}_l)}{\partial k} \tag{3.24}$$

where $F_{A,k}^{QM}$ is the *ab initio* k-component of the force on the atom A for the l-th configuration and $f_{Ai,k}$ is the k-component of the force on the atom A due to the i-th atom calculated with the force field kept constant in the fitting process.

The *ab initio* references for the contribution of the force field of the metal ion to the total energy of the system can be calculated as the difference

$$y_l = E_{tot}^{QM}(\mathbf{R}_l) - E_{env}^{QM}(\mathbf{R}_l) - E_M^{QM} \tag{3.25}$$

where $E_{env}^{QM}(\mathbf{R_l})$ is the energy of the $\mathbf{R_l}$ configuration without the metal ion and E_M^{QM} is the energy of the isolated metal ion. The elements of the $\mathbf{H}$ matrix in this case are

$$H_{j,l} = \sum_i^{N_{atoms}-1} D_{ij}v_j(\{\boldsymbol{\theta}\},\mathbf{R}_l) \tag{3.26}$$

The global $\mathbf{H}$ matrix, in the multi-objective fitting, is then the result of the concatenation of two or more matrices, each one relating to a specific quantity of reference.

Optimization of the weights in the multi-objective fitting

In multi-objective optimization, the *utopia point* [121], **OF°**, is defined as the vector of the single objective functions in which each component, $OF_b°$, corresponds to the global minimum in its relative space with respect to the variables to be optimized. In practice the utopia point is generally unattainable and two common approaches are adopted to address the problem: i) identify the set of Pareto solutions [122], leaving to the decision maker the choice on which to use ii) locate a compromise solution [123] minimizing the distance from the utopia point. Both alternatives involve some degree of arbitrariness. Here, as a criterion for obtaining the optimal weights, the second approach is adopted, using the Chebyshev metrics in a normalized space [124] as a method for calculating the distance from the utopia point:

$$OF_b^{norm}(\mathbf{C},\boldsymbol{\theta};\mathbf{w}) = \frac{OF_b(\mathbf{C},\boldsymbol{\theta};\mathbf{w}) - OF_b^{\circ}}{OF_b^{max} - OF_b^{\circ}} \tag{3.27}$$

The application of the Chebyshev distance involves the use of the *minimax criterion*:

$$\mathbf{w}_{opt} = \min\{\max\{\mathbf{OF}^{norm}(\mathbf{C},\boldsymbol{\theta};\mathbf{w})\}\} \tag{3.28}$$

It implies that the maximum component of the vector $\mathbf{OF}^{norm}(\mathbf{C},\boldsymbol{\theta};\mathbf{w})$ is minimized with respect to the weights. This choice aims at achieving the most balanced compromise solution. The result of the minimization depends on the choice of the OF_b^{max}, that corresponds to the worst acceptable value for the b-th objective function. The optimization of the equation 3.28 is performed using the simulated annealing algorithm [125]. The proposed variations for the weights are executed by applying an adaptive heuristics that reduces the number of the function evaluations and exploits the monotone relationship between w_b and OF_b^{norm}.

Chapter 4

Analysis of trajectories

The investigations reported in this PhD thesis have been performed by combining both classical MM techniques (mainly MD simulations) together with high, QM-level computations, giving rise to protocols which can be defined as "hybrid QM/MM" approaches. In this thesis, QM computations have been performed at the DFT level, using the B3LYP exchange-correlation functional [126] or the long-range corrected CAM-B3LYP [127]. The combination of the two level of theory has been anticipated in the previous section, where the functioning of the Joyce software and the LRR-DE procedure for the development of QM-based force field is illustrated: QM computations are used in fact to create all the input files (i.e., energy, gradients and Hessian matrix) which are used to built reliable parameters sets to be used in classical MD simulations. Using the developed FFs, MD simulations have been extensively carried out, to investigate the structural and dynamic properties of the solutes in the considered environments, by using analysis tools which are deepen in the following. The importance of a proper and effective force filed parameterization for the systems under investigation appears to be clear at this stage, since the computed properties (structural, dynamic and spectroscopic) are strictly affected by the force field parameters, which determine the traits of the simulated chemical entities.

4.1 Structural properties

Radial distribution functions (rdf or $g(r)$) describes how the density of a certain kind of particle (i.e., an atom type) varies from a reference one as a function of its distance. Rdf plots are very important, since they provide useful information about the atomic structure of a molecular systems. The rdf between generic atoms of type A and B has been computed as

$$g_{AB}(r) = \frac{\langle \rho_B(r) \rangle}{\langle \rho_B \rangle_{local}} = \frac{1}{\langle \rho_B \rangle_{local}} \frac{1}{N_A} \sum_{i \in A}^{N_A} \sum_{j \in B}^{N_B} \frac{\delta(r_{ij} - r)}{4\pi r^2} \tag{4.1}$$

where $\langle \rho_B(r) \rangle$ is the density of atom type B computed at distance r around A, and $\langle \rho_B \rangle_{local}$ is the density of B averaged over all spheres around A with radius equal to the half of the simulation box length. N_A and N_B are the number of A, B particles, and r_{ij} is the distance between particle i and j. The averaging is performed in time, i.e. over all the frames of the MD trajectory. In practice, the system is divided into spherical slices from r to $r + dr$, in order to obtain an histogram. In an homogeneous system the average number of atoms B in a spherical layer around A between r and r + dr is given by

$$dN_B = 4\pi r^2 \rho g_{AB}(r) dr \tag{4.2}$$

Integrating the right hand term one obtains the average coordination number of atoms B around A inside a sphere of radius r centered on A. The plot of the rdf profile gives important information on the location of first and second solvation shells, as well as on the intermolecular interactions that take place in molecular systems, thus to allow to use a cutoff value to separate different shells for subsequent analysis. Moreover, the Fourier transform of the rdf combination yields the total scattering functions that can be measured with a diffraction experiment (X-Ray or neutron).

The *radius of gyration* is well known to provide a rough measure of the compactness of a structure. It has been computed, using a standard GROMACS utility, as

$$R_g = \sqrt{\frac{\sum_i r_i^2 m_i}{\sum_i m_i}} \tag{4.3}$$

where m_i is the mass of atom i and r_i is the position of atom i with respect to the center of mass of the molecule.

4.2 Dynamical properties

Dynamical features are of great importance, since they allow to understand the influence of the embedding environment on the internal flexibility and global mobility of a solute in a solvent

Dynamic properties have been estimated mainly by using correlation functions. The *mean square displacement* (*MSD*) has been calculated as

$$MSD(t) = \lim_{t\to\infty} \langle [r(t_0+t) - r(t_0)]^2 \rangle \tag{4.4}$$

For molecules made up of more atoms, r refers to each atom i and the obtained MSD is averaged over these atoms. For large molecules, however, r can be simply related to the center of mass of the molecule. When a diffusive regime has been established by the solute, the diffusion coefficient D can be determined by using the Einstein relation [82]

$$D = \lim_{t\to\infty} \frac{1}{6t} MSD(t) \tag{4.5}$$

The diffusion coefficient D provides quantitative information, experimentally verifiable, on the motion of a solute in a solvent.

Giving three atoms i, j and k, the *rotational autocorrelation function* ($C_p(t)$) is calculated as the autocorrelation function of the vector p, defined as $p = ij$ x jk (i.e. the cross product of the two vectors ij and jk) according to

$$C_p(t) = \langle P_l(p(t0) \cdot p(t_0+t)) \rangle \tag{4.6}$$

where P_l is the first or the second Legendre polynomial read as

$$P_{1_{rot}} = cos(\phi(\tau)) \qquad ; \qquad P_{2_{rot}} = \frac{1}{2}(3cos^2(\phi(\tau)) - 1) \tag{4.7}$$

This measure describes the tumbling and spinning of a molecule due to the presence of the surrounding environment. When $C_p(t)$ is fitted to an exponential function (such as $C_p(t) = \exp(-t/\tau_p)$) it is

possible to estimate the correlation time τ_p, defined as the time needed to the investigated molecule to completely rotate around p.

The static dielectric constant ϵ of pyridine has been estimated in Chapter 6 on the basis of the fluctuations of the total dipole moment M of the simulation box, by computing the total dipole moment autocorrelation function with the following equation [128]

$$\epsilon = 1 + \frac{4\pi}{3\langle V\rangle k_b T}(\langle M^2\rangle \langle M\rangle^2) \tag{4.8}$$

and

$$M = \sum_i \mu_i \tag{4.9}$$

where μ_i is the molecular dipole moment of molecule i and V is the volume of the simulation box.

4.3 Hydrogen bond analysis

During a MD simulation, all the possible hydrogen bonds (HBonds) between donors (D) and acceptors (A) in a chemical systems are inspected. In particular, the atom at high electronegativity bound to the hydrogen atom has to be considered as D.

HBonds have been defined through the following standard geometric criterion

$$\begin{aligned} r &\leq r_{HB} \\ \alpha &\leq \alpha_{HB} \end{aligned} \tag{4.10}$$

where r is the distance between A and B, and α is the angle between atoms H, D and A at each frame; r_{HB} and α_{HB} are the corresponding reference values. By default, α_{HB} is set equal to 30°, while r_{HB} is 3.5 Å (which corresponds to the first minimum of the radial distribution function of SPC water [129].

In contrast with the geometric criterion employed in this works, it is worth noting that also continuous formulations of HBonds can be used. Continuous functions properly take into account the gradual decay of HBonding networks similar to the one of rdf profiles. One of these methods [130] substitutes the stepwise treshold of Eq. 4.10 with Gaussian functions. Each A-D interaction can be defined as a HBond assuming a score between 0 and 1, according to

$$f(r,\alpha) = g(;r_e;r_{hw}) \times g(\alpha;\alpha_e,\alpha_{hw}) \tag{4.11}$$

where r is the acceptor-hydrogen distance and α is still the $\widehat{\mathrm{HDA}}$ angle, and

$$g(x;x_e,x_{hw}) = \begin{cases} 1 & \text{if } x \leq x_e \\ \exp(-\frac{(\mathrm{x}-\mathrm{x_e})^2}{2\mathrm{x}_{\mathrm{hw}}^2} & \text{if } x > x_e \end{cases}$$

where x_e, x_{hw} are the maximum and half-width values of either the non-normalized radial or angular distribution function, obtained by a fit of the corresponding data. This function has been recently used for the description of formamide liquid [16] and nicotine molecule in water solution [131].

4.4 Absorption spectra

Absorption spectra have been computed on different and statistically uncorrelated configurations extracted from MD trajectories (see Chapters 7 and 8). Vertical transition energies are computed at the DFT level, producing single peak for each of the considered electronic transition, and convoluted with Gaussian functions in the energy domain using a properly chosen width at half maximum ($HWHM$) value, Δ_ν, according to [132]

$$\epsilon(\nu) \propto \sum_{i \in states} \frac{f_i}{\Delta_\nu} exp\left[\left(\frac{\nu - \nu_i^0}{\sigma_\nu}\right)^2\right] \tag{4.12}$$

where

$$\sigma_\nu = \left[2\sqrt{2ln(2)}\right]^{-1} \Delta_\nu \tag{4.13}$$

and f_i and ν_i^0 are the oscillator strength and the frequency of the *i-th* excitation, respectively. The distribution functions allow to gain the broadening usually observed in experimental spectra. The final spectrum is then converted to the wavelength domain ($\bar{\epsilon}(\lambda)$) for an easier comparison with experiments. Single signals computed on the MD snapshots, as extracted from the corresponding MD trajectories, are averaged according to:

$$\bar{\epsilon}(\lambda) = \sum_{c \in confs} \frac{\epsilon_c(\lambda)}{N_{confs}} \tag{4.14}$$

During the absorption computations, environment effects are considered in two different ways, using:

1. implicit solvation models (e.g., PCM). Here, solvent atomic coordinates are completely neglected during the vertical transition energies. The solute is placed within a cavity of proper size surrounded by a dielectric medium representing the environment. The molecule-shaped cavity is built by connecting spheres centered on the heavy atoms of the solute, by using the default set of atoms radii available in Gaussian software.

2. electrostatic embedding (EE). This is an explicit way of considering the enviroment, since atomic cordinates of the solvent are included during the spectroscopic computation [133, 134]. In particular, spheres centered at the solute center of mass are cut for each of the extracted snapshot: all the solvent molecules with at least one atom lying within the sphere radius from the solute are considered and replaced by the corresponding atomic charge in the calculation. The final system is then put into a PCM cavity, leading to a three-layers computation (DFT/point charges/PCM). Solute-solvents direct electrostatic interactions, which depend on atoms positions of the two conunterparts, are directly taken into account within this scheme.

4.5 Free energy calculations

Free energy difference estimation is a central task in chemistry. Many measurable properties (conformational equilibria, association constants, transition paths and so on) depend on the free energy of the system. Within the canonical ensemble, free energy is denoted as Helmoltz energy A; if the pressure rather than the volume is kept constant instead the same quantity is named as (G) Gibbs free energy. Experimentally, only the difference in free energy between two different states can be computed, so

that the quantities of interest in computational chemistry are actually ΔA and ΔG, which are numerically equivalent in condensed phase systems. In Chapters 5, 6 and 9 free energy differences have been computed by using two different techniques: the umbrella sampling [135] and the Bennet acceptance ratio (BAR) method [136].

In umbrella sampling, free energy difference are computed as a function of a reaction coordinate ξ, chosen by the user. A biased sampling is allowed by the application of a harmonic potential ω, to ensure a proper sampling along ξ. This operation is performed in different windows, the distribution of which overlap and each one characterized by a particular value of ξ. The resulting distribution is, however, non-Boltzmann. The corresponding Boltzmann averages can be extracted from the non-Boltzmann distribution, thus to obtain equlibrium quantities, according to

$$A_i(\xi) = -\frac{1}{\beta}\ln P_i^b(\xi) - \omega_i(\xi) + F_i \tag{4.15}$$

where β corresponds to $1/(k_bT)$ and $P_i^b(\xi)$ is the biased distribution of window i obtained from MD. The calculated $A_i(\xi)$ is also known as *potential of mean force*, (PMF). The value of ΔG is simply the difference between the highest and lowest values of the PMF curve. However, the constant F_i is undetermined. A common method to get F_i for extracting PMF is the Weighted Histogram Analysis Method (WHAM) [137], implemented in GROMACS, which determines the global, unbiased distribution P^u by minimizing the corresponding statistical error. Hence, the following formula is applied:

$$\exp(-\beta F_i) = \int P^u(\xi)\exp[-\beta w_i(\xi)]d\xi \tag{4.16}$$

The whole process is iterative and goes on until convergence.

BAR is a common method for computing free energy of hydration. The method estimates free energy differences relying on the output from simulations of different states, controlled by a parameter (λ) which weights all the force field contributions. From $\lambda = 1$ to $\lambda = 0$, interactions between the solute and the solvent is progressively turned off and the particles disappear from the medium. The free energy difference can be calculated directly if two states (say λ_i and λ_j) are close enough, by computing the Monte Carlo acceptance ratio of transition of i to j and viceversa. For a given configuration, free energy difference between state A and B is expressed as

$$\Delta A_{ij} = \frac{1}{\beta}\ln\frac{\langle\alpha\exp[-\beta U_i]\rangle_j}{\langle\alpha\exp[-\beta U_j]\rangle_i} \tag{4.17}$$

where α value is determined by minimizing the free energy difference between i and j. Then, solving numerically

$$\sum_{i=1}^{n_i}\frac{1}{1+\exp(\ln(\frac{n_i}{n_j})+\beta\Delta U_{ij}-\beta\Delta A)} - \sum_{j=1}^{n_j}\frac{1}{1+\exp(\ln(\frac{n_j}{n_i})+\beta\Delta U_{ji}-\beta\Delta A)} = 0 \tag{4.18}$$

the free energy ΔA is obtained. It is worth noticing that when the solute has nearly disappeared (i.e., λ value close to zero) the interaction energy is low, thus to allow particles to get close enough to collide each other. To avoid this issue, standard Coulomb and LJ potentials are replaced by the corresponding "soft-core" versions [138]. Soft core potentials available in GROMACS software (employed for the free

energy calculations performed in this work) are shifted versions of the regular potentials, so that the singularity in the potential (and the corresponding derivative) at zero distance is never reached.

4.6 Clustering analysis of structures

MD simulations produce a lot of information, such as atoms positions, forces and velocities. Although several properties are easy to compute directly from the MD trajectories, many others are difficult to extract and rationalize on the basis of the sampled molecular configurations [139]. *Cluster analysis* (or simply *clustering*) is defined as the process of partitioning a set of data objects into subsets [140]. The obtained subsets are called *clusters* and each cluster object is similar to the other objects of the same cluster, yet dissimilar to the objects belonging to other clusters. The clustering is based on proper metrics, which define the similarities and dissimilarities between the objects. Often, the used metric involves distances measure. Cluster analysis is defined also as a form of unsupervised learning, that is, the data which have to be clustered are not classified ("unlabeled" data), and as a consequence there is no univocal evaluation of the clustering results.

Four main clustering methods categories are usually defined. In *partitioning methods* (*k*-Means [141], k-Medoids [142]) a set of N objects are divided in K clusters ($N \leq K$), and the clusters are formed in order to optimize a partitioning criterion chosen by the user. The representative object of a cluster is referred as the *centroid* of that cluster, which is its center point. *Hierarchical methods* (such as BIRCH [143]) create a hierarchical decomposition of the set of data into clusters until a termination condition holds. They can be *agglomerative* or *divisive* , based on how the process is performed. In *density-based approaches* (DBSCAN [144], OPTICS [145]) the clusters sizes grow until a density treshold (i.e., the number of data objects in the "neighborhood" of a cluster) is above a treshold value. Unlike the other two, density-based approaches can lead to the formation of non-spherical clusters. The last category includes *grid-based methods* (as the STING method [146]), which convert the objects into points on a quantized grid. Such clustering procedures are cheap, since they do not depend on the number of the data points to be clustered, but only on the number of cells of the grid.

In Chapter 6, molecular conformations of pyridine extracted from MD trajectories have been exposed to a clustering procedure based on the *k*-Means algorithm, a centroid-based technique belonging to the partitioning methods. The procedure is composed by few step, which are repeated iteratively. K initial objects are randomly chosen as initial centroids of K clusters. Then, each object is assigned to the nearest cluster, i.e. each conformation is included within the cluster featured by its most similar centroid. The measure must be chosen by the user. The K centroids are recalculated as the mean of the membership cluster, and the data objects are assigned again to the closest centroids. Within each cluster, the objective function J to be minimized is the following

$$J = \sum_{j=1}^{K} \sum_{i=1}^{N} \| X_i^j - C_j \|^2 \tag{4.19}$$

where $\| X_i^j - C_j \|$ is the computed distance between the point X_i^j and the centroid C_j of the J-th cluster. The algorithm is therefore aimed at minimizing the intra-clusters variance. The whole algorithm is illustrated in Table 4.1.

1. *k*-Means algorithm

Input:

- k: the number of clusters,
- D: a set of data objects.

Output: a set of clusters.

Method:
1) chose randomly k objects from D as the initial cluster centroids;
2) **repeat**
3) (re)assign each object with the closest cluster
on the basis of the distance from the mean value of the objects in the cluster;
4) recalculate the mean value of the object of each cluster;
5) **until** no change;

TABLE 4.1: The k-Means algorithm for partitioning.

k-Means can be only applied if a mean of the data objects can be defined. Moreover, it is sensitive to data noise and outliers, since they could strictly affect the mean value.

One of the main limitation of the method is the necessity for the user to specify the number of clusters in advance. The performances are strictly dependent on the value of K, and in general there is no exact method for determining the correct number of clusters for a particular analysis. Common practice is to perform k-Means analysis multiple times, and comparing the quality of the clustering obtained from different values of K. A good estimate can be obtained by considering the Calinski-Harabasz (CH) index [147] in order to evaluate the optimal K, defined as

$$CH_K = \frac{[\mathbf{B}/K-1]}{[\mathbf{W}/n-K]} \tag{4.20}$$

where n is the total number of objects, $\mathbf{B}$ denotes the deviation between clusters for a K value, and $\mathbf{W}$ is the deviation of each object from the respective cluster centroid. The higher the CH_K value, the better is the solution.

Part II

Applications

This part presents the studies and corresponding results accomplished during this PhD. It is organized as it follows:

- Chapter 5 reports on an unbinding study of an intercalating drug (doxorubicin) from its biological counterpart (a DNA fragment). The computational investigation has been performed by using standard force field parameters taken from literature, with small changes. Related results has been shown during the presentation of the program Caffeine, developed in the SMART laboratory. Despite a fair agreement with previous theoretical works, the obtained results mainly serve the purpose of providing a realistic and reasonable data set for the software tools. Data visualization, a topic always intimately connected with chemistry and research in general, has to be considered as the surrounding scenario of this section.

- Chapter 6 shows how reliable bulk properties can be simulated only by refining the non-bonded part of a standard force field. Pyridine, a standard solvent routinely used in industry and in synthesis, has been considered as a case study. The transferability of the proposed model from the pure liquid to the aqueous solution is demonstrated, filling a gap in currently available force fields for pyridine, unable to describe both systems at the same level of accuracy.

- In Chapters 7 and 8 the building from scratch of the entire, QM-derived, intramolecular force fied of two chromophores allowed for effective MD simulations and spectroscopic studies. The interest on these molecular probes is related to the potential applications as photoactive species in luminescent solar concentrators (Chapter 7) and as microenvironment sensors (Chapter 8). In both cases, force field modeling has been made possible thanks to the Joyce software cited above. The former chapter reports also on the reliable modelling of environment effects during absorption spectra simulations of the investigated dyes within a polymeric matrix. The latter instead focused on the comparison between the dynamic and structural properties exhibited by the analyzed probe in different environments.

- In Chapter 9 the LRR-DE method (see Section 3.2) is validated through the parameterization of non-bonded models of metal ions in water solution. Beside the standard Lennard-Jones plus Coulomb functional, a more complicated and flexible form is parameterized to highlight the potentiality of the LRR-DE method in the optimization of models of general functional forms. The performances of the new models in reproducing thermodynamic and structural properties from MD simulations are of comparable or better quality respect to literature ones. It is noteworthy that this method can be easily extended in order to consider other environments different from water (e.g., protein catalytic sites).

Chapter 5

Dissociation of Doxorubicin from DNA Binding Site

The unbinding process of an anticancer agent, doxorubicin, from its biological counterpart (a DNA fragment) is investigated in this chapter. Standard Amber99sb force field [4] and GAFF [59] have been used, with small changes. This study has been performed in conjunction with the presentation of the software Caffeine [148], specifically tailored for molecular structures and data visualization with Virtual Reality (VR) systems such as VR theater and helmets. The doxorubicin-DNA complex has been considered as a well versed case study to demonstrate the benefits that can be obtained from immersive VR (IVR) along the several stages of a typical computational chemistry investigation.

5.1 Background

5.1.1 Visualization in chemistry

A detailed and yet compact representation of molecular structures, together with the inclusion of related properties in formulas and graphs, has always been at the heart of chemistry. Since the very beginning, chemists faced the growing amount of new, chemical entities by building their own terminology and ways of representation: it can been said in fact that chemistry was the first modern science able to create new objects to be studied and, as a consequence, needed a new language in addition to words and mathematical equations. The (only apparently) confused and complicated graphical conventions that nowadays worldwide chemists use to communicate each other are the consistent result of the process mentioned above. Afterwards, with the advent of computer machines, scientists of different areas have been able to take advantage of computer graphics techniques, which have allowed an extension of chemical language through the visual representation of molecular objects, ranging from small organic compounds to huge macromolecules of biological interest. In practice, two-dimensional schemes, drawn by means of only pencil and paper, have been juxtaposed by more sophisticated pictures and models, which reflect in many cases the chemical and physical features of the rendered molecular system. Such evolution has enabled the molecules to be perceived for what they really are, i.e. 3D objects with a well-established position in space of the relative atoms, thus being characterized in a more precise and effective manner. Molecular models such as ball-and-stick, licorice and vdW spheres are still used today despite their oldness and immediacy [149]. Even surfaces are a useful method of structure representation, mostly used when significant molecular properties such as shape, size and occupied volume are needed to be highlighted. Representation therefore plays

a key role in science and during the whole discovery process: it conveys information (such as the result of a QM computation) to human inspectors, possibly in an interactive way, by relying on human pattern recognition, and suggesting innovative points of investigation and new, previously unexplored scenarios [150]. More in general, scientific visualization has to reveal data, guiding researchers among knowledge acquisition: without molecular graphics and its drawing conventions many of numerical calculation would provide very little scientific insight by the sheer amount of numbers fed to the user [151].

Nowadays, the massive increase of computer graphics technologies for three dimensional IVR is allowing to achieve a further evolution in data representation and visualization [152]. In fact, it is now possible to create virtual 3D environments that extend users perception and increase users ability to quickly tackle with massive amounts of data coming from multiple and different sources: within such systems, users can directly interact with visualized data (by means of dedicated devices) in a more natural and friendly way compared to standard desktop systems with mouse and keyboard [153, 154]. VR tools include a vast array of devices at present, from cheap consumer grade ones to very costly specialized hardware. In the first category are included interactive sensors like the *Microsoft Kinect* [155] and the *Leap Motion* [156], current generation immersive helmets such as the *Oculus Rift* [157] and the *Vive* from HTC and Valve [158], or force-feedback devices like the Novint *Falcon 3D Touch controller* [159]. The second category instead accounts for virtual theaters such as the CAVE (Cave Automatic Virtual Environment) [160, 161], equipped with high-precision tracking sensors and driven by one or more powerful workstations. While CAVE-like systems represent some of the most advanced IVR system available today, they are (very) expensive fixed installations. For that reasons, they can be found only in few specialized research centers [162, 163].

FIGURE 5.1: (a) CAVE theater at Scuola Normale Superiore. (b) User wearing the Oculus Rift DK1 helmet.

5.1.2 The Caffeine molecular viewer: an overview

Caffeine is an innovative molecular viewer under continuous development at the SMART laboratory [164] at Scuola Normale Superiore in Pisa. It is aimed to enable the employment of different architectures, ranging from the standard computer desktop to more expensive IVR environments, within the field of high-accuracy molecular systems (and related data) representation. A particular field highly taken into account during the software's development is the accurate visualization of scientific numerical data while still keeping the system under investigation at its native atomic detail, under the user

control, allowing therefore for an intuitive and human friendly data fruition. Some peculiar features include the augmented reality-like visualization of 2D charted data and the interactive filtering of trajectories through the selection of chemical configurations of interest (called "key-frames"). The software can visualize both static and dynamic molecular structures (trajectories) read from PDB, XYZ and Gaussian Cube files. As many other molecular viewers, Caffeine supports the most diffused graphical representations methods, such as “all-atoms” visualization (balls-and-sticks, licorice, and vdW spheres) and ribbon diagrams for polypeptides and polynucleotides. In addition, volumetric datasets such as electron densities and molecular orbitals can be imported from Gaussian Cube files and visualized as isosurfaces.

5.2 The DOX/DNA system

5.2.1 Doxorubicin

Doxorubicin (DOX hereafter, structure in Figure 5.2a) is an anthracycline antibiotic, able to bind DNA by intercalation. The binding process is promoted by the peculiar structure of such chemical agent, which features fused hydrophobic ring systems that can insert between base pairs, creating favorable π-stacking with nucleobases and shelding from solvent molecules [165]. Apart from this planar, hydrophobic part, DOX exhibits also a hydrophilic, aminosugar moiety (daunosamine). The carbohydrate acts as an anchor, sitting in the DNA minor groove and interacting at the same time with nearby nucleobases. As a consequence of the intercalation event, Topoisomerase (either I or II) enzymes, which play key roles during DNA replication and trascription, are obstructed so to promote the disruption of DNA double strand and leading the cell to death [166].

Intercalating drugs are among the most employed drugs in anticancer therapy even today. However, despite their widespiread use, a detailed understanding of the formation and dissociation process of the DNA/drug complex remains elusive. From a computational chemistry point of view, MD simulations of DOX and related molecules [167, 168] have tried to elucidate the structural changes that take place in B-DNA upon intercalation. Enhanced sampling methods have been used to shed light on the intercalation pathway of daunomycin (closely related to DOX), suggesting a three-step process [169–171] which is compatible with experimental findings [172, 173].

5.2.2 Computational Details

Simulations were carried out with GROMACS 4.6.5 [98]. The PDB crystallographic structure 1D11 [174] containing one daunomycin molecule and a single palindromic DNA sequence of six nucleotides, was chosen as starting system. Daunomycin was modified adding a hydroxil group on the methyl-ketone side chain using GaussView [175] to obtain the DOX. Considering its palindromic sequence, the nucleic acid was duplicated, rotated of approximately 180° and moved with the aim of reproducing the HBonds pattern between complementary nucleobases. The so-built double-stranded system was protonated and immersed in a cubic box of roughly 16500 TIP3P [66] water molecules, adding Na^+ and Cl^- ions to achieve neutrality. The final system was composed by 50160 atoms put in a rectangular box of 493 nm^3. DNA was described using the Amber99sb force field. DOX intramolecular terms and vdW parameters were modeled according to the GAFF force field [59]. To parametrize a reliable charge set for the DOX

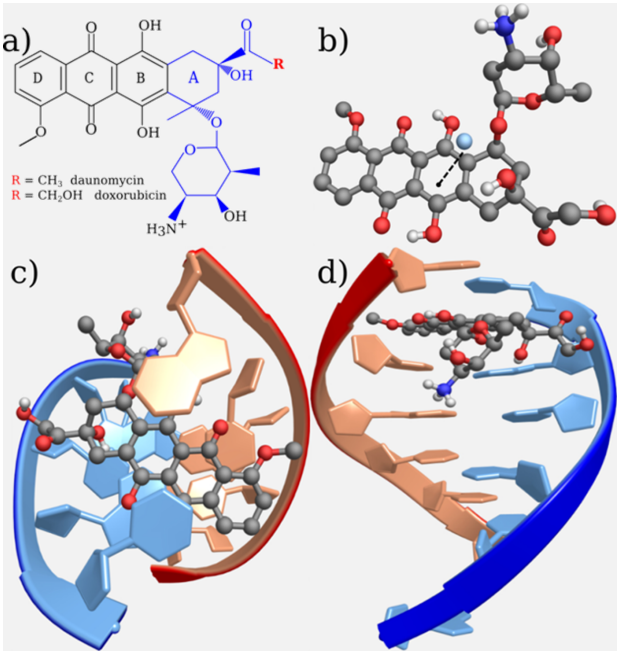


FIGURE 5.2: (a) Scheme of the daunomycin and doxorubixin compounds; the hydrophobic anthraquinone moiety is shown in black while the hydrophilic aminosugar part is in blue. (b) Top view doxorubicin drug, shown using a ball and stick representation with standard colours and omitting apolar hydrogen atoms. The position of the center of mass (COM), used in the umbrella sampling, is shown as a cyan atom relative to the center of ring B (see dashed line). (c-d) Top and side view of the starting structure (1D11) after the modeling of doxorubicin drug; the drug is intercalated between the two CG pairs. DNA backbone is shown either blue (5-3) or red (3-5), with the ending part of the ribbon thinner (C3) or thicker (C5); nucleotides are depicted using filled polygons omitting the detailed atomic structure.

molecule, the multi-conformation RESP (Multi-RESP [176]) procedure was applied. Such method takes into account more than one conformation of minimum energy via the equation

$$C(n) = \frac{\sum_i C_i(n) e^{\frac{-\Delta E_i}{kT}}}{\sum_i e^{\frac{-\Delta E_i}{kT}}} \tag{5.1}$$

where $C(n)$ is the Multi-RESP charge of the atom n; $C_i(n)$ is the RESP charge of atom n in the conformation i; E_i is the energy difference between the conformation i and the global minimum; k is the Boltzmann constant; T is the temperature (298.15 K). To ensure the stability of the terminal base pair in the intercalation site, additional weak harmonic bonds between the CG base pair above DOX were added; such bonds (each of 50 kJ mol^{-1}) were modeled in order to resemble the hydrogen bond coupling between the two nucleobases [177]. Long range electrostatic interactions were accounted by means of the PME method [89]. A cut-off of 14 Å was used for short range electrostatic Van der Waals interactions. LINCS [93] was used to constrain bond lengths and angles. The system was initially minimized by means of the steepest descent algorithm. Relaxation of solvent molecules and counter ions to 300 K was initially performed keeping solute atoms restrained to their initial positions with a force constant of 1000 kJ mol^{-1} nm^{-2}), for 5.0 ns in a NPT (using the Parrinello-Rahman barostat [96]) ensemble and using an integration time step of 2.0 fs. Then, the system was carried again to 0 K and progressively heated to 300 K in steps of 50 K. Starting from the last conformation of the equilibration step, DOX was pulled away from the intercalation site by the application of a harmonic potential of 125

kJ mol^{-1} nm^{-2} between the centers of mass (COMs) of DOX and of the four nucleobases delimiting the intercalation site. A pull rate of 10 nm ns^{-1} was used. The pulling process was allowed in each dimension. When a distance of 17.5 Å was reached, the DOX molecule was assumed to be completely separated from the DNA bundle. Along this path, system configurations were taken every 0.6 Å of COMs separation and used as starting point for the umbrella sampling, for a total of 25 simulation windows. Afterwards, a 200 ps equilibration run (at 300 K and 1 bar pressure) was performed for each selected window, according to the protocol reported by Lemkul et al [178]. In each window 10 ns of MD was performed, for a total of 250 ns. Instantaneous atomic forces along the umbrella sampling trajectories, together with the mutual displacements along the reaction coordinate, were stored every 2 ps. Results were analyzed using the WHAM method [137] to compute the PMF profile along the predefined reaction coordinate.

5.2.3 Results

The unbinding mechanism

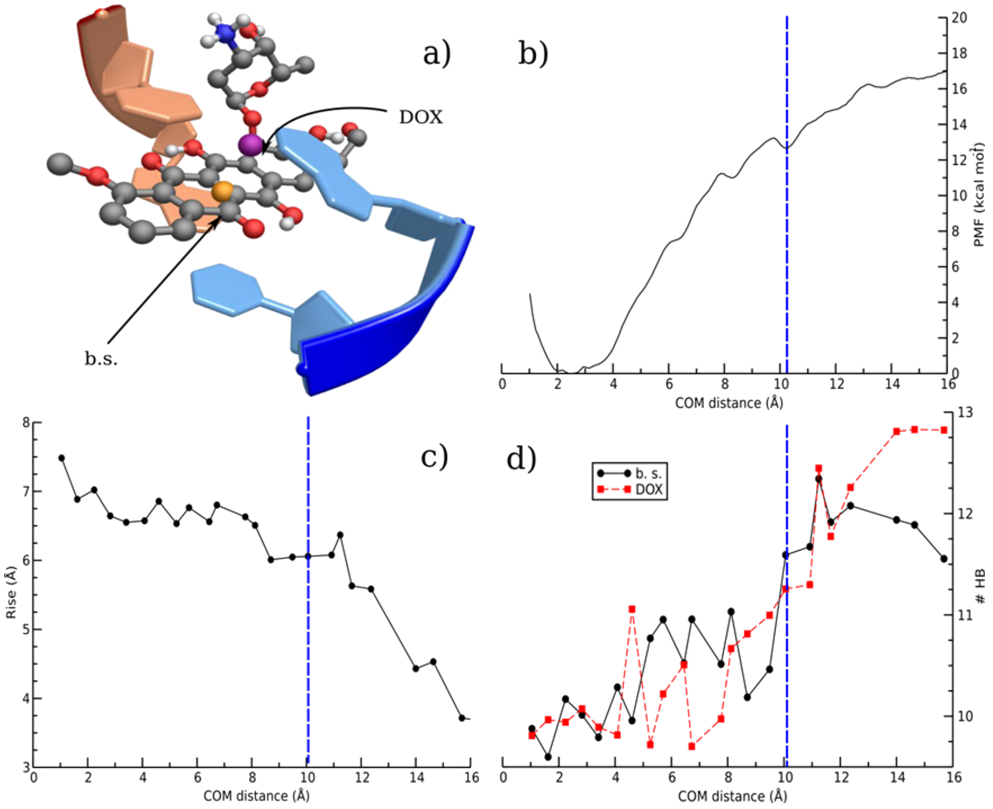


FIGURE 5.3: Binary complex dissociation process. Note that in panels (b-d) the position of the intermediate at 10 Å and final dissociated state along the sampling coordinate is highlighted with a blue dashed line. (a) Binding site of the DOX compound in the initial conformation. The position of the centers of mass of the binding site (b.s., indicated by the gold sphere) and of DOX (purple sphere) is shown. (b) Potential of mean force (PMF) curve associated to the distance between centers of mass of the binding base pairs and of the doxorubicin drug. (c) Calculated change in rise between DNA base pairs. (d) Number of hydrogen bonds between the binding site nucleobases (black) or DOX (red) with water molecules.

The unbinding process of DOX from the binary complex was investigated through the umbrella sampling method. In a first step, the drug was removed (pulling process) from its intercalation site applying a harmonic potential between the centers of mass (COMs) of DOX and of the binding site (see Figure 5.3a). For the latter, only the heavy atoms of the four nucleobases delimiting the intercalation site were considered. DOX was allowed to leave the intercalation site along each dimension, while the distance between the two considered COMs was chosen as the reaction coordinate. In Figure 5.3, panel b, the ΔG over the COMs distance is represented. The ΔG curve shows a large global minimum (assuming zero value for convenience) at about 2 Å and centered at approximately 2.5 Å. Interestingly, this situation is the same as in the crystallographic and in the fully equilibrated structure, where a distance between the two considered COMs of 2.69 Å is identified; the free energy minimum that emerges from the present investigation is thus confirmed by experimental data. The curve then increases rapidly as DOX is removed from the intercalation site. Two partially stable states are observed, one at about 8.5 Å and a second one at 10.2 Å. In the first, narrow well DOX seems to be still able to interact with DNA nucleobases through π-stacking interactions of its D aromatic ring, while the B-A rings are almost outside the intercalation site and thus accessible by the solvent. Roughly 3 kcal/mol are enough to overcome this state and enter in the next metastable state observed between 10 and 11 Å. This state could be associated to the intermediate one (IM) already found in previous studies [171] in the case of daunomycin: here, the rigid body of the DOX molecule lies on the plane defined by the two DNA backbones, while the intercalation site is still in an opened conformation (see Figure 5.4). Going forward on the chosen reaction coordinate, a rotation of DOX is observed at about 11 Å: this movement is sufficient to break the weak interactions that take place in the intermediate state, allowing the compound to definitively get rid of the macromolecular target. Finally, the energy profile establishes itself, and no more energy rises are detected as DOX approaches the bulk solvent. However, already at 12 Å it is very difficult to perform an appropriate sampling, because the ligand is completely dissociated from its biological target, and it is thus free to rotate and diffuse through the solvent. From Figure 5.3b it can be observed that 14 kcal/mol are necessary for DOX to reach the solvent. Such estimation is in good agreement with previous results of Lavery and co-workers on daunomycin [169, 171]. The measured difference should be reasonable, in view of the marginal differences between these two molecules.

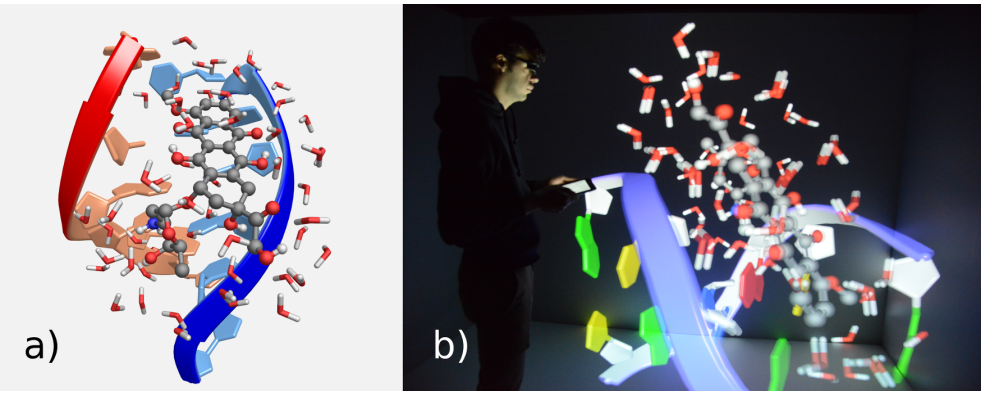


FIGURE 5.4: (a) Intermediate bound state of DOX with the nearest neighbor water molecules shown as licorice. (b) as as (a), within the CAVE.

Intercalation, and, subsequently, dissociation of anthracyclines from DNA have been demonstrated

to alter the overall topology of DNA [179]. A lot of parameters could be monitored to detect conformational changes along the time in which such processes take place. Here, the rise parameter, i.e. the distance along two base pairs along the DNA sequence, was considered as a measure of structural modification during the unbinding process. Figure 5.3c shows the average rise distance between the base pairs in the intercalation site as a function of the reaction coordinate in the simulated windows. High rise values (of approximately 7.5 Å) are detected for the intercalated state: then, as DOX approaches the bulk solvent, the distance between two consecutive bases decreases, reaching a final value of about 3.5 Å in the unbounded state, very close to the value of 3.4 Å featured by native B-DNA. It is also possible to observe that the intermediate state (at about 10 Å of COM distance) is still in an opened conformation, with a rise value close to 6 Å. Finally, the hydration of the binary system as a function of the reaction coordinate was taken into account (see Figure 5.3d) calculating the average number of hydrogen bonds between water molecules and either DOX or the four nucleobases of the binding site. Intercalation induces a decrease of the number of HBonds as DOX acts as an obstacle for water molecules, which cannot enter the binding site. At the same time, DOX is less hydrated as it approaches the DNA binding task: on average, two hydrogen bonds that take place in the DOX unbounded state lack in the intercalated state. It is interesting to note a peak of averaged number of hydrogen bonds for both the considered molecules in proximity of the intermediate state: in fact, at this point, DOX has already left the intercalation site, so to be considered solvent-exposed. In the meanwhile, the DNA intercalation site is still open and freely accessible by water molecules, that can hence interact with nucleobases. Then, after 10 Å of COM distance, the number of H-bonded water molecules slowly decreases. This is in agreement with the prior considerations: in fact, after this point, as it is proved in Figure 5.3c, the binding site reduces its size, because of the departure of DOX, thus decreasing water accessibility to the intercalation site nucleobases.

Studying the dissociation process in a IVR environment

The DOX/DNA investigation has been largely conducted by taking advantage of IVR technologies installed at Scuola Normale, and always using the Caffeine software. As regarding IVR tools currently available in the SMART laboratory, the CAVE [160, 161] and the Oculus Rift [157] are currently supported by Caffeine.

The CAVE is a cube-like room-sized IVR system, whose walls (from three to six) are projected with stereoscopic images. The position of the user within the CAVE is detected by a tracking system and used by the software to adjust the perspective of the displayed images, so to obtain a convincing and coherent stereoscopic visualization across the screens. When using the CAVE a new feedback, prioproception, is added to the perception of data. Proprioception is the capability to perceive and recognize the position of the own body in space, even without sight: the kinesthetic inputs from mechanoreceptors in muscles, tendons, and joints contribute to the human perception of limb position and limb movement in space. Within the CAVE 2D data charts are drawn in front of the user, immersed in the 3D scene and follows the movements of the user's head, in a way analogous to an augmented reality content (see Figure 5.5). The ability to "pinpoint" insights by using quantitative information becomes critical in a IVR environment where it is easier for the (inexperienced) user to be overwhelmed by various feedbacks. Moreover, the use of key-frames, by itself a useful feature, becomes critically important within a IVR environment since it allow the user to concentrate on important properties and provides

a guide during the visualization process to elaborate and save insights that otherwise would be lost upon leaving the IVR. Moreover, the user is able to interact with the system by means of a dedicated remote application for mobile devices. This application currently allows the user to rotate, translate and scale the molecular system, and to control the playback of frames.

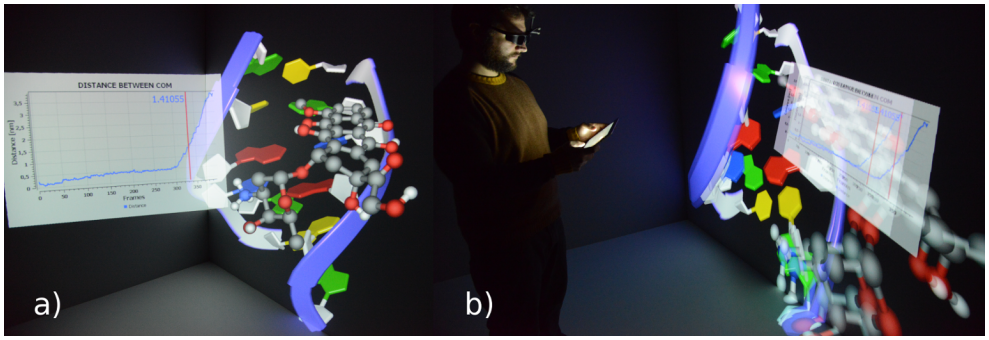


FIGURE 5.5: Dissociation of DOX from DNA bindingsite. (a) Simultaneous representation of charted data and molecules. The binary complex is on the right while a graph showing the distance between COMs is visible on the left with the red marker highlighting the current frame and distance value. For sake of clarity, the stereo mode of projectors was temporarily disabled to shoot this photo. (b) Selection of sensible structures as starting configurations in the umbrella sampling study performed within the CAVE.

In the analysis of DOX/DNA computations, the "CAVE" version of Caffeine has been used. In particular, the selection of the starting configurations for the umbrella windows was performed within such IVR environment: the ability to view, at the same time, a molecular conformation and the chart reporting the distance between COMS was exploited in order to effectively select sensible structures. Chosen structures have been marked as key-frames and used for the subsequent umbrella simulations. Then, appropriate conformations (such as the centroids of the single umbrella windows) have been used to reconstruct the whole unbinding process, from the intercalated to the completely unbound state, through the IM one, so as to build an artificial trajectory to be used in Caffeine to follow in the meanwhile both chemical structure evolution and related structural parameters. Finally, considering the PMF chart in Figure 5.3b, it is always possible to connect the current, visualized snapshot to its associated free energy value just switching from the COMs distance chart to the PMF one, thus increasing the user's understanding of the overall free energy study.

5.2.4 Conclusions

The dissociation of the binary complex DOX/DNA has been properly simulated by using extensive MD simulations through the application of the umbrella sampling method. Standard Amber99sb force field has been considered to describe DNA. Additional bonds were modeled in correspondence of the terminal base pairs, in order to prevent from disruption of the nucleic acid fragment. As stated at the beginning of this chapter, this system has been chosen as a proper platform for presenting Caffeine software and to show the potential of using IVR in computational medicinal chemistry. A simplified (to limit the computational cost), yet consistent, computational protocol has been developed to serve as a basis for illustrating the features of Caffeine to a wide audience. It is worth to observe anyway that,

qualitatively, the estimation of the free energy barrier (14 kJ/mol) is in good agreement with the previous results of Lavery and co-workers on daunomycin [169, 171]. Moreover, comparable profiles can be observed between the two systems as regarding structural parameters such as the rise value. The whole investigation envisions the idea of possible, productive and realistic employment of IVR technologies in computational chemistry, as well as the use of the Caffeine molecular viewer as a reliable front-end device in post-processing analysis.

Chapter 6

Fine-Tuning of Atomic Point Charges: the Case of Pyridine

A correct description of electrostatic contributions in force fields for classical simulations is mandatory for an accurate modeling of molecular interactions in pure liquids or solutions. Here, a new protocol for point charge fitting is proposed, aimed to take into the proper account different polarization effects due to the environment employing virtual sites and tuning the point charge within the polarizable continuum model framework. The protocol has been validated by means of MD simulations on pure pyridine liquid and on pyridine aqueous solution, reproducing a series of experimental observables and providing the information for their correct interpretation at atomic level.

6.1 Background

Despite its key role in describing the solvation ability of a substance, the static dielectric constant remains one of the most difficult bulk properties for classical simulations[180] and it is especially sensitive to the electrostatic part of the force field. Since the most widely used force fields still rely on partial atomic charges, several strategies have been employed to determine effective values for these quantities; one of the most adopted approaches relies in optimize their values to reproduce the QM derived electrostatic potential of a molecule[52]. The atomic charges depend strongly on the level of theory used in QM calculations and on the description of solvation effects. In this present study, the electrostatic parameters have been obtained through the Charge Model 5 (CM5) [54] taking into account the bulk solvent effects by means of Conductor-like Polarizable Continuum Model (C-PCM) [181]. This allows one to take into account the different polarization effect of various solvents (here pyridine or water) in tuning the effective atomic charges (as explained in the following). Furthermore, since the presence of hydrogen bonds (HB) has a remarkable effect on magnitude of the static dielectric constant (ϵ) [182], a thorough description of intermolecular interactions is necessary.

One of the problems in the HB description is its directionality. A possible solution is the use of off-site charges (so-called virtual sites or dummy atoms, VS hereafter) with a fixed position with respect to the generating atom that is meant to model the presence of lone pairs [16, 131]. In fact, an improved HB description of pyridine in aqueous solution has been obtained employing VS and adjusting the charges on the carbon atoms directly bound to nitrogen to preserve the molecular dipole moment [131]. However, to develop a model for pyridine able to describe the interactions both in aqueous solution and in pure liquid, the atomic charges on the whole molecule have been determined with a

fitting procedure using as reference parameters both the molecular dipole and quadrupole moments. Therefore, a method for the derivation of partial atomic charges and VS is proposed, which completely depends on properties determined at the QM level without any additional empirical parameters. It is demonstrated that the same strategy works for both pure (liquid) pyridine and its aqueous solution.

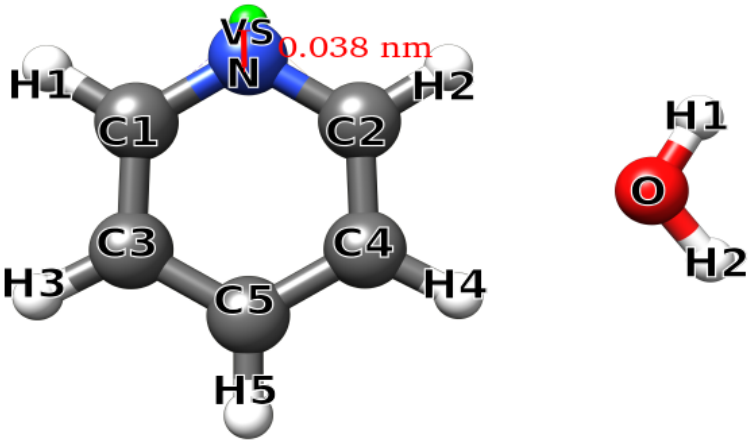
The pyridine aqueous solution, as well as the pure liquid, have been studied by means of classical simulations employing several models [131, 183, 184]. Although some of the tested force fields provide reasonable results, here the transferability of the proposed model may be highlighted: such model is able to simulate both pure pyridine and its aqueous solution, overcoming the limitations on previously reported force fields and delivering accurate structural and thermodynamic properties together with the static dielectric constant.

6.2 Methods

Classical MD simulations of both pure liquid and aqueous solution of pyridine were carried out using GROMACS 4.6.5 [98] and the OPLS/AA force field [183] to describe intramolecular and intermolecular potential with the exception of the pyridine atomic charges, which were estimated using the CM5 population analysis [54]. DFT calculations were performed at the B3LYP/6-31+G(d) level of theory using GAUSSIAN-09 [97] package and taking into account bulk solvent effects by means of the C-PCM [181] setting the reference solvent (pyridine or water) and imposing the value of the scaling factor for the sphere radius (α) to 1.05. A VS was located at the position of the centroid of the localized molecular orbital describing the sp2 lone-pair of the nitrogen atom using the Pipek-Mezey localization procedure [185] (the centroid distance from the nitrogen atom has been constrained during the simulation). The charge on the VS has been obtained adjusting the atomic charges of pyridine through a fitting procedure to reproduce the calculated dipole and quadrupole moments, as indicated in Table 6.1 (for atom labeling Table 6.1).

The simulation for the pyridine pure liquid was performed on a system containing 500 molecules, whereas the simulation of the aqueous solutions was performed on one pyridine and 512 TIP3P-FB [69] water molecules. In both cases, a cubic box with periodic boundary conditions was employed. After a steepest descent energy minimization the systems were heated up to 298.15 K for 200 ps (using the velocity-rescale thermostat[95] and τ = 0.1 ps) and then the time step and temperature coupling constant were increased to 2.0 fs and 0.2 ps respectively, and systems were let to converge to uniform density in a NPT ensemble (using the Parrinello-Rahman barostat [96] and τ = 0.1 ps).

Afterward production runs were run in the NVT ensemble, fixing the fastest degrees of freedom with LINCS algorithm (δt = 2.0 fs) [93]. The total sampling time was 50 ns for both the pure liquid and the aqueous solution. Electrostatic interactions were evaluated using the PME [90] method with a grid spacing of 1.2 Å and a spline interpolation of order 4. Harder et al. [3] have proposed an OPLS3 pyridine model, which also has an off-site charge on the nitrogen atom and reproduces well the hydration free energy (-4.3 kcal/mol). MD simulation using this force field was performed, in order to determine structural information on pure liquid and compare the obtained results with the proposed model. The main difference between OPLS3 and the presented model is the charge on nitrogen atom, which is not null in OPLS3 (+0.179 e) and the procedure employed to define the virtual site position [8]. An ab initio MD (AIMD) simulation of a pyridine aqueous solution at ambient temperature was



	q(e)					
	Solvent=Pyridine			Solvent=Water		
	CM5	Adjusted	Δ (CM5-Adj)	CM5	Adjusted	Δ (CM5-Adj)
C1-C2	0.035648	0.001047	0.034601	0.034855	-0.000150	0.035005
C3-C4	-0.090393	-0.113328	0.022935	-0.089909	-0.113072	0.023163
C5	-0.068839	-0.113375	0.044536	-0.068037	-0.113058	0.045021
N	-0.390529	0.000000	-0.390529	-0.393605	0.000000	-0.393605
VS	0.000000	-0.329297	0.329297	0.000000	-0.331658	0.331658
H1-H2	0.112816	0.151052	-0.038236	0.112965	0.151590	-0.038625
H3-H4	0.113452	0.095364	0.018088	0.114284	0.096017	0.018267
H5	0.116322	0.174402	-0.058080	0.117252	0.175946	-0.058694
μ (D)	-2.62941			-2.68354		
Q(XX) (D·Å)	-2.15265			-2.15585		
Q(YY) (D·Å)	7.75545			7.78872		
Q(ZZ) (D·Å)	-5.60280			-5.63287		

TABLE 6.1: The CM5 charges (e) for pyridine calculated at the B3LYP/6-31+G(d) level and used in the simulations (adjusted) taking into account the solvent (pyridine or water) effect by means of C-PCM. In the column Δ (CM5-Adj) is reported the difference between the calculated and used charges. The molecular dipole (μ) and quadrupole (Q) moments obtained from CM5 charges and employed during the fitting procedure in presence of VS are also reported. In the last row, the pyridine and water atoms labeling and the virtual site position are indicated.

carried out using the CP2K program [186] to further assess the reliability of this procedure to describe the hydrogen bond interactions and their directional character. A cubic box of size 11.77 Å was considered, containing 1 pyridine and 50 water molecules and subjected to periodic boundary conditions. 20 ps of simulation in the NVE ensemble was performed, using a time step of 0.1 fs. The electronic structure was calculated with DFT, utilizing the BLYP functional [187]. The TZV2P basis set was used in conjunction with the GTH pseudopotentials [188, 189]. A plane wave cutoff of 340 Ry was adopted for electron density. VdW interactions have been described by the method proposed by Grimme [190]. Dielectric constant and density have been evaluated using standard tools provided with GROMACS. To calculate the heat of vaporization, ΔH_{vap}, gas-phase simulations of 2 ns (δt = 0.2 fs) have been added for both systems. The ΔG_{hyd} were calculated using BAR method [136], performing the simulations with GROMACS. Structural analysis was performed with TRAVIS package [191].

The relative orientation of first neighbour pyridine molecules was also investigated by means of the *k*-Means clustering algorithm [141]. The optimal number of clusters was determined with the Calinski-Harabasz criterion.

6.3 Results

6.3.1 Aqueous solution

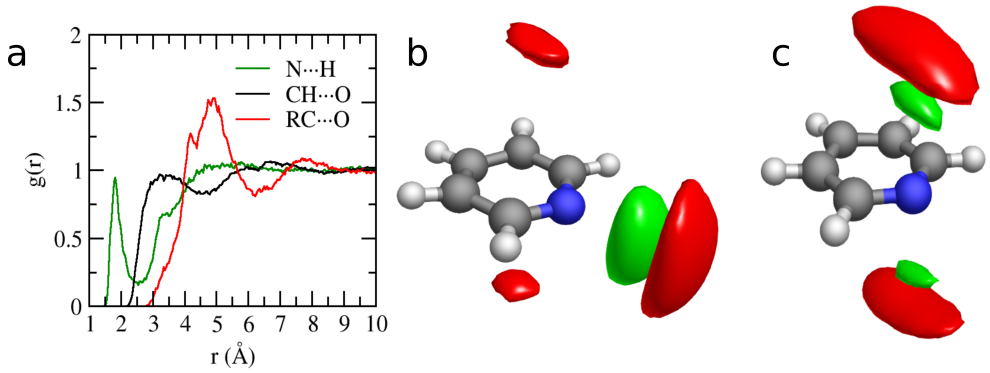


FIGURE 6.1: (a) Radial distribution functions between the pyridine nitrogen (green) and hydrogen (black) atoms and the water hydrogen and oxygen atoms respectively. In red the g(r) between the pyridine ring center (RC) and the oxygen atom of water. (b) Isosurface, obtained from MD simulation, of water oxygen (red), hydrogen atoms (green) around the pyridine molecule at an isovalue of 68 and 60 nm^{-3} respectively. c. Isosurface of water oxygen (red), hydrogen atoms (green) around the pyridine molecule at an isovalue of 68 and 60 nm^{-3} respectively. The results are obtained with the model deprived of VS.

The structure of pyridine in aqueous solution is characterized by the hydrogen bond interaction between the nitrogen atom of the heterocyclic ring and a hydrogen atom of the solvent molecules (Figure 6.1a). The oxygen atom also interacts with the hydrogen atoms of pyridine, leading to a weaker interaction, as confirmed by a longer interatomic distance (see Figure 6.1a). In fact, the first maximum in $g_{N\cdots H}(r)$ and $g_{CH\cdots O}(r)$ radial distribution functions (rdf) occurs around 1.8 Å and 3.2 Å respectively. The relative sdf of water hydrogen and oxygen atoms around the pyridine are reported in Figure 6.1b, corroborating the presence of a hydrogen bond between the nitrogen and the hydrogen atoms. The isovalue adopted to show the results does not lead to similar isosurfaces involving the pyridine hydrogen atoms, confirming the weaker character of the interaction with solvent. The oxygen atom density is localized above and below the aromatic ring suggesting the presence of HB $\cdots\pi$ interaction. In fact, the radial distribution function between the pyridine ring center (RC) and the oxygen atom (red line in Figure 6.1b confirms the presence of such an interaction. To demonstrate the usefulness of VS to mimic the lone pair of the sp2 nitrogen atom, in Figure 6.1c are also shown the same sdf obtained analyzing the simulations of the model when the VS is removed. In this case the simulation was performed using the CM5 charges (6.1) without the VS and the HB density distribution shows a lack in directionality, the sdf is not localized in the ring plane but above and below it. As described in Methods section, also an ab initio simulation, for purposes of comparison, was performed. Therefore, in Figure 6.2 the rdf between the pyridine nitrogen atom and the water oxygen atom obtained from three simulation is

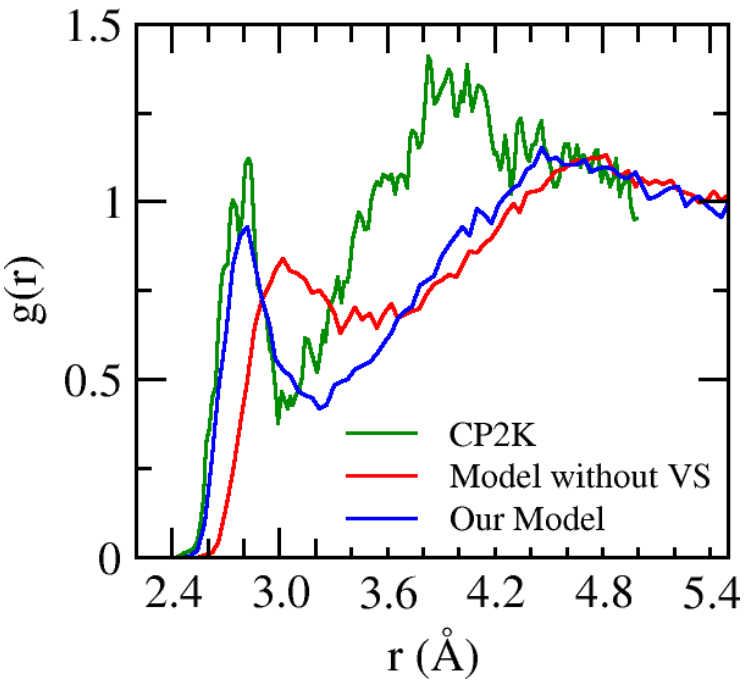


FIGURE 6.2: Radial distribution functions between the pyridine nitrogen and the water oxygen ($g_{N\cdots O}(r)$) atoms obtained from three simulations: this work (blue), CP2K (green) and the model without the virtual site (red).

reported: the here presented force field, CP2K and a simulation performed without VS. The comparison between CP2K and the present model (with VS) is remarkable and confirms the usefulness of VS to describe the hydrogen bond.

6.3.2 Pure pyridine

A first structural analysis of pure pyridine is reported in Figure 6.3a, where the rdf between nitrogen atoms is shown. The $g_{N\cdots N}(r)$ shows two distinct peaks, one around 4.9 Å and one around 5.9 Å. A similar rdf was obtained by Jorgensen and McDonald [39] and by Baker and Grant [192], who have previously studied liquid pyridine using the OPLS and OPLS-CS force fields respectively. Jorgensen and McDonald attributed the first peak to antiparallel contacts and the second to dimers that are offset stacked with parallel (head-to-tail) dipoles. According to Baker and Grant, the first peak results from a structure in which a pyridine molecule donates a hydrogen bond through the H1/H2 hydrogen atoms (for atom labeling see Table 6.1) and the second peak from one in which a pyridine molecule donates a hydrogen bond using the H3/H4 hydrogen atoms. Interpretation of the peaks in structural terms can be obtained by computing the combined distribution function (cdf) between the $g_{N\cdots N}(r)$ distance and the angle formed between the vectors connecting pyridine ring center and nitrogen atom. The cdf is shown in Figure 6.3b and permits to attribute the first peak to an antiparallel arrangement and the second to a parallel arrangement and to an interaction, which occurs between rings forming an angle up to 45°.

To better understand the relative orientation of pyridine molecules, *k*-Means[141] was applied. The first neighbour distances between the center of rings, the nitrogen and the C5 atoms was selected as clustering features; by inspection of Figure 6.3, an 8.0 Å cut-off for each feature was determined.

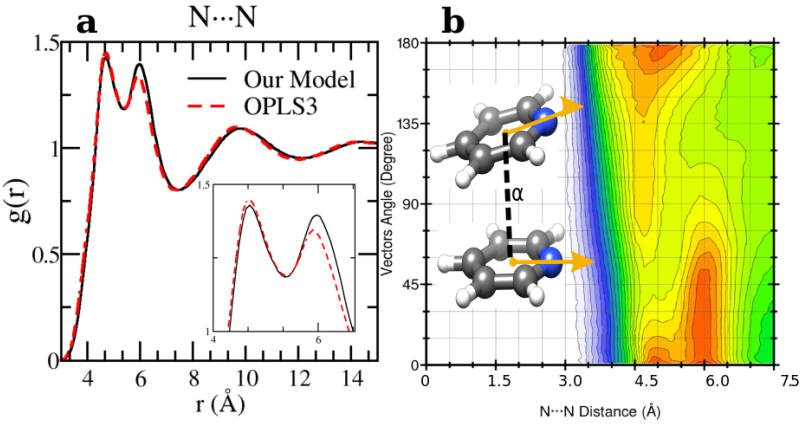


FIGURE 6.3: Radial distribution function between N···N in pure liquid (black: the proposed model; red: OPLS3). b. Combined distribution function between N···N distance and the angle a formed between the vectors connecting the center of the ring and the nitrogen atom (orange vectors in picture b). The results are obtained from the pure liquid pyridine simulation.

Application of the Calinski-Harabasz score [147] indicated the optimal number of clusters at NC = 2 containing 56% and 44% of data points, respectively. The first neighbour distances of the cluster centers show an approximately parallel orientation of molecules. The N···N and C5···C5 distances for the two cluster centers are of 5.78, 5.53 and 6.68, 6.87 Å respectively. This corresponds to the arrangements where the angle between two rings is 0° or around 45° (see insert in Figure 6.3b). In Figure 6.3a the $g_{N\cdots N}(r)$ g(r) calculated from the OPLS3 simulation is reported, which shows a slightly lower peak at around 5.9 Å suggesting that in the present case the parallel arrangement is more abundant. This observation will be useful to evaluate the differences between the calculated bulk properties.

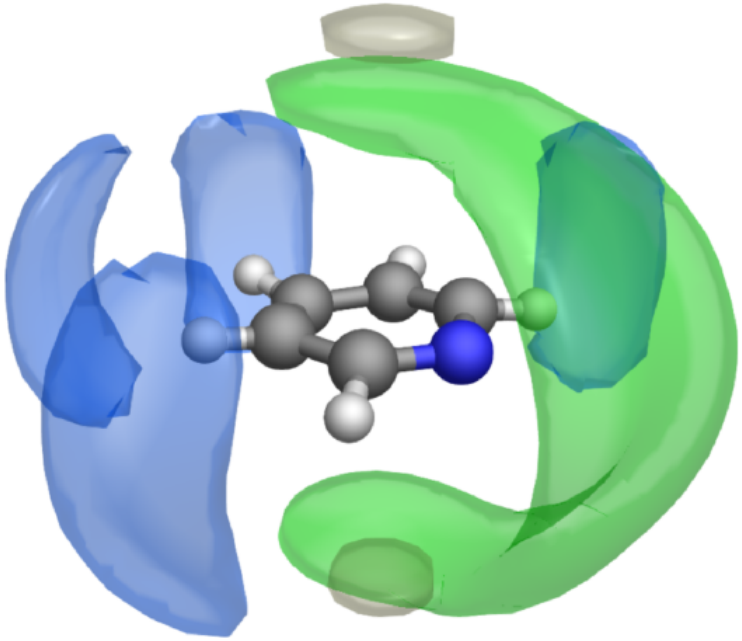


FIGURE 6.4: Isosurface of the pyridine nitrogen atom (blue), carbon atoms (gray) and hydrogen atoms (green) around the pyridine molecule at an isovalue of 15, 14 and 11 nm^{-3} respectively. The results are obtained from the pure liquid pyridine simulation.

The average densities of nitrogen, carbon and hydrogen atoms around the molecules are reported

in Figure 6.4. The hydrogen atoms density around nitrogen atom supports the hypothesis that the interactions in the pyridine liquid are dominates by N···HC hydrogen bonds. The same is true for the nitrogen atom density, localized near hydrogen atoms. This spatial distribution function deserves an additional consideration: the density does not point along the CH bond but is localized in the center of two CH bonds; this is due to the possibility of having simultaneous interactions between the nitrogen atom and two hydrogen atoms. Furthermore, the localization of hydrogen atoms (as well as the carbon) density above and below the aromatic ring, suggests that the arrangements are stabilized by the formation of CH···π interactions. It is also noteworthy that the N···HC interactions mainly involve, as expected, the H3, H4 and H5 hydrogens. Figure 6.5a sketches the cdf between two different N···H distances; this analysis confirms the simultaneous interaction in the pure liquid, as already shown by the sdf of Figure 6.4. Figure 6.5b shows also the relative cdf for the water solution between two N···H(water) distances; it is apparent that the interaction involves only one hydrogen atom and confirms the results of sdf analysis. Figure 6.5a-b reports also the single rdf, which points out a stronger interaction in the pyridine aqueous solution (shorter distance N···H involved) with respect to the pure liquid (higher maximum in the radial distribution function). In the following, results obtained for bulk properties (Table 6.2) are discussed. A correct reproduction of the static dielectric constant of the pure liquid is particularly important in view of the large use of pyridine as solvent for several processes of technological relevance [193]. Our pure liquid model, using CM5 charges (C-PCM) and a VS for sp2 nitrogen atom, allows one to obtain a dielectric constant value of 11.2 ± 0.2 in good agreement with the experimental value of 12.4 [193]. This result represents a non-negligible improvement with respect of the value of 6.7 ± 0.1 obtained using the standard OPLS force field [180]. Furthermore, the analysis of a MD simulation on pure pyridine performed in the present study employing the OPLS3 force field [3] leads to less accurate values of both static dielectric constant (9.1 ± 0.2) and density (1008.9 ± 0.2 kg/m3). Looking at Figure 6.3, the better value of dielectric constant could be due to a higher number of parallel pyridine or nearly parallel pyridine molecules issuing from the new force field. In Figure 6.6 the rdf between the nitrogen and the hydrogen atoms of pyridine is shown: the new model has a higher peak for the $g_{N\cdots H5}(r)$ which is in agreement with the parallel orientation of pyridine molecule.

Properties	This work	Exp.
ϵ	11.2 ± 0.2	12.4
ρ (kg/m^3)	995.3 ± 0.2	977.8
ΔH_{vap} (kcal/mol)	10.7 ± 0.2	9.61
ΔG_{hyd} (kcal/mol)	-4.23 ± 0.02	-4.7

TABLE 6.2: Computed values and experimental data for static dielectric constant, density (kg/m^3), heat of vaporization (kcal/mol) and free energy of hydration (kcal/mol) for pyridine molecule.

Other properties, often used to validate a force field, such as density and heats of vaporization (ΔH_{vap}), were also calculated and their computed values are in agreement with the experimental results. A density value of 995.3 ± 0.2 kg/m^3 was obtained (the experimental value is of 977.8 kg/m^3 [194]); instead, the calculated ΔH_{vap} value is equal to 10.7 ± 0.2 kcal/mol. Since the experimental value is 9.61 kcal/mol [194] and OPLS provides a value of 9.76 kcal/mol [15], in this case OPLS performs better than the new model.

Furthermore, the availability of the pyridine aqueous simulation, allows us to derive the free energy

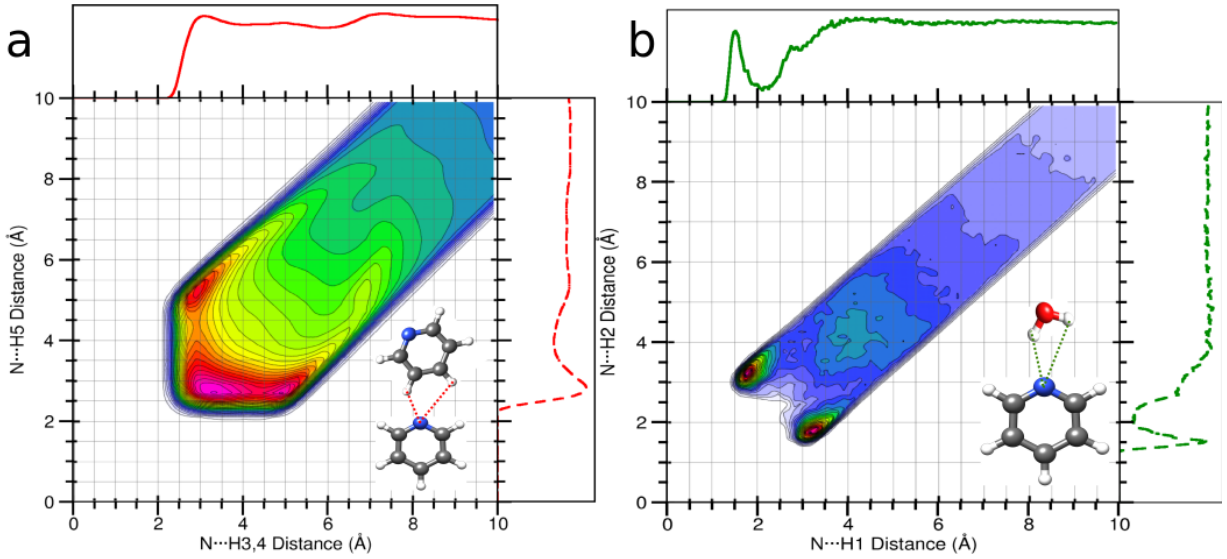


FIGURE 6.5: (a) Combined distribution function between N···H3,4 e N···H5, like indicated in insert, obtained from the pure pyridine liquid simulation. (b) Combined distribution function between N···H1 and N···H2 obtained from the water pyridine solution.

of hydration, ΔG_{hyd}. The result obtained is of -4.23 kcal/mol (Table 6.2) with an error around the 10% with respect to the experimental value of -4.7 kcal/mol [90].

6.3.3 Conclusion

In this study, an improved force field of pyridine, which allows one to describe both the pure liquid and the aqueous solution, is presented. The main feature of the new force field is the use of partial atomic charges that takes into account in an effective way the polarization effects of the environment, without adding any ad hoc correction term to the force field. By changing the dielectric constant value in the C-PCM protocol, it has been possible to describe through the same model the pyridine aqueous solution and the pure pyridine liquid. Furthermore, since the hydrogen bonds have been found to be important in describing the structure, a VS was introduced to mimic the lone pair of the sp^2 nitrogen atom. The VS allows one to describe the directional character of hydrogen bond interaction in agreement with reference AIMD simulation, without compromising the description of the pure liquid. In fact, the VS can describe both the interaction involving only a couple of atoms (pyridine-water) and a simultaneous interaction involving more than two atoms, like the interaction found in pure pyridine liquid, which may involve the nitrogen atom of one pyridine molecule and two hydrogen atoms of another.

Starting from these satisfactory structural results, also thermodynamic properties for the pure liquid was computed. The static dielectric constant, which is often poorly reproduced with standard force fields, has been properly reproduced. Although usually not included in force field validation, the dielectric constant represents a very important parameter governing the solvation capacity. At the same time, satisfactory results were obtained for density and vaporization enthalpy.

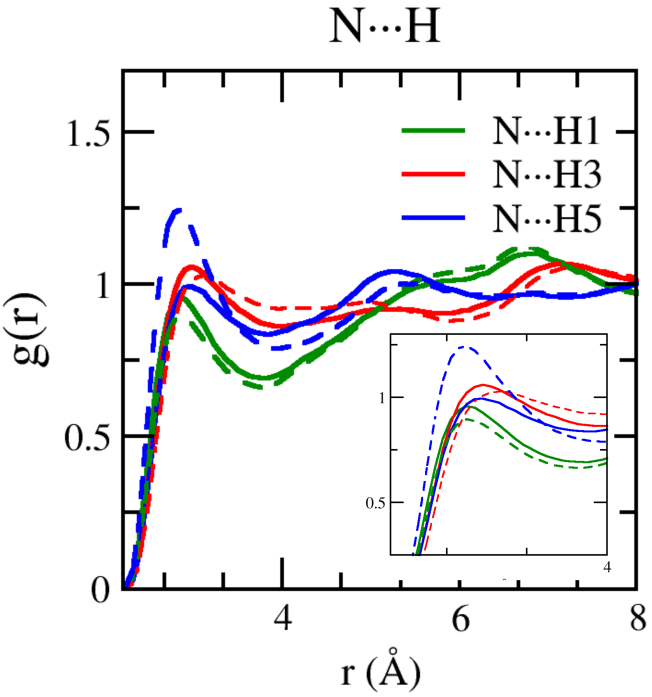


FIGURE 6.6: Radial distribution function between N and H1 (green), H3 (red), H5 (blue) hydrogen atoms. In continuous line OPLS3, in dashed line the model presented in this work.

Chapter 7

Modeling of Photoactive Dyes Within a Sunlight Harvesting Device

This chapter is aimed at a deeper investigation of two recently synthesized heteroaromatic fluorophores by means of a computational multilayer approach, integrating QM and MM. In particular, dispersion of the title dyes in a polymer matrix is studied in connection with potential applications as photoactive species in luminescent solar concentrators (LSCs). MD simulations, based on accurate QM-derived force fields, reveal increased stiffness of these organic dyes when going from $CHCl_3$ solution to the polymer matrix. QM/MM computations of UV spectra for snapshots extracted from MD simulations show that this different flexibility permits explaining the different spectral shapes obtained experimentally for the two different environments. Moreover, the general spectroscopic trends are reproduced well by static computations employing a polarizable continuum description of environmental effects.

7.1 Background

Commodity materials based on thermoplastic polymers derived from oil or renewable resources are everyday expanding their range of applications, thanks to their excellent thermo-mechanical properties, chemical stability, easy processing, sustainability and low cost [195]. A new class of materials is nowadays even more rapidly evolving due to the development of modern technologies that require combinations of properties. This great interest towards functional materials has driven a wide investigation on the design and development of new "intelligent" systems with unique mechanical and optical properties [196]. Several specific features can be finely tuned by an appropriate combination of a polymeric matrix with different molecular entities, by means of dispersion [197] or covalent functionalization [198, 199]. Such materials can be envisaged as stimuli responsive devices, which can find widespread applications in various fields [200, 201]. Many external factors, such as pressure, temperature, pH, viscosity or radiation, can in fact, alter the system's response [202]. Owing to these characteristics, optical responsive materials can be exploited in renewable energy technologies; over the last three decades increasing interest has been devoted to light harvesting devices. Recent findings on global warming [203] and nonrenewable energy resources have stimulated increasing interest towards the employment of alternative ways for energy. Sunlight stands indeed as an ideal asset that can be taken advantage of. Luminescent solar concentrators (LSCs) represent a way to decrease the cost of solar photovoltaics [204]. LSCs consist of a fluorescent dye dispersed in a thin slab of polymeric material. Upon solar irradiation, a fraction of the emitted light is collected at the edges of the device

where photovoltaic cells are located [205]. Organic fluorescent dyes bearing π-conjugated electron-donor and electron-acceptor moieties exhibit intra-molecular charge-transfer (ICT) [206] properties, and can therefore show the optical properties required by LSCs, such as high quantum yield and large Stokes shift [207].

Selection of the best molecular structure for such applications is not easily driven by a strategy based on experimental data alone, which can be very approximate. Therefore, computational modeling finds increasing use to carry out systematic studies aimed to gain a deeper understanding of the chemical–physical properties responsible for the experimental optical features. As a matter of fact, computational approaches nowadays allow simulation of spectral shapes at a very reasonable cost and with good results [208]. It is therefore not surprising that many and different computational methods, aimed at the reproduction of photophysical properties of organic probes dispersed in – or covalently bounded to – several environments, have been recently reported [209–212]: such methods, trying to reveal subtle phenomena that take place at the atomic level, hard to detect by means of experimental techniques alone, are becoming an invaluable tool to support experiments. However, the modeling of environmental effects, is often a crucial step, and it needs to be taken into account in order to correctly describe the photophysics of a target molecule: charge density can rearrange as a consequence of the electric field generated by the surrounding environment. Moreover, the conformational equilibrium of flexible molecules is often strongly tuned by solvent effects. Unfortunately, it is still arduous to treat the solvent using very accurate QM calculations, because of the computational cost of including a large reservoir of molecules constituting the environment. For these reasons, during the last few decades different approaches have been developed to deal with solvation, ranging from implicit methods, such as the PCM [99], to more demanding procedures, such as Monte Carlo (MC), and both classical and ab initio MD simulations, which explicitly feature solvent molecules in their spatial coordinates [213, 214].

In a recent paper [215], a new class of alkynylimidazole-based fluorophores has been studied as promising luminescent dyes in LSCs. UV-Vis spectra of these molecules in tetrahydrofuran (THF) solution as well as in a poly(methyl methacrylate) (PMMA) thin film were experimentally determined, suggesting an enhanced rigidity of the investigated fluorophores in the polymer matrix. Absorption spectra were simulated in good agreement with experimental data, using the CAM-B3LYP long-range corrected hybrid density functional, in conjunction with the PCM to take bulk environmental effects into account. Moreover, a fast and reliable computational protocol was developed to simulate absorption and emission spectra of such luminescent species when dispersed in the rigid, polymeric environment: starting from the ground-state energy minimum, the main dihedral angle (i.e., the one between the phenyl and the imidazole moieties) was fixed at its ground-state value during the excitation (mimicking in such a way the caging effects of the hydrophobic bundle), while all other internal coordinates have been relaxed. As already stated, bulk electrostatic effects were taken into account by the PCM. As is well known, PCM offers the unquestionable advantage of providing an overall and reliable description of the surrounding environment, which is mainly characterized by its dielectric constant. However, the continuum method fails in the treatment of specific solute–solvent interactions, which depend on the spatial coordinates of the two interacting components [216]. For this reason, in the present

study, a recently developed protocol [217] was applied to address with improved accuracy the description of the polymeric bundle, and compared to the cheaper "static" approach described above in the reproduction of spectroscopic properties. The proposed approach involves computational methodologies ranging from MD simulations, based on accurate force fields, to full QM calculations, and explicitly simulating the surrounding solvent. As is well known, FF based simulations manage to mimic the dyes in a realistic environment, and follow the time-evolution of the whole system to its most likely configurations over time scales longer than 100 ns. However, since variations in the chromophore's structure significantly affect the computation of its spectroscopic properties, highly accurate and reliable FF parameters are required. For this reason the FF parameterization of the dye used in this work was performed using Joyce [9, 81], and all the intra-molecular parameters needed to describe the chromophore in its ground state were specifically derived from QM calculations by fitting optimized energies, gradients and Hessian matrices. Two different alkynylimidazole-based fluorophores bearing a different electron withdrawing group were investigated for their potential application in LSCs, namely the novel 5-((4-dicyanovinyl)ethynyl)-1-methyl-2-(4-nitrophenyl)-1H-imidazole, **a**, and the already reported25 5-((4-methoxy)ethynyl)-1-methyl-2-(4-nitrophenyl)-1H-imidazole, **b** (Figure 7.1). Alkynylimidazoles a and b were prepared according to a simple reaction sequence, starting from 1-methyl-1H-imidazole.

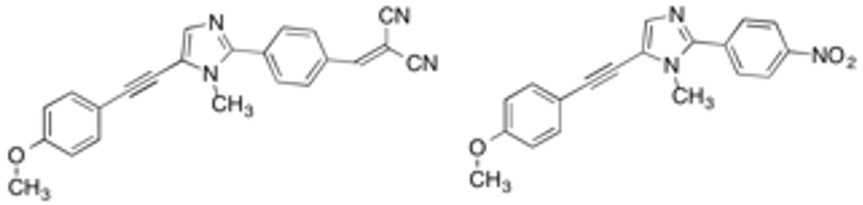


FIGURE 7.1: Structures of the investigated a (left) and b (right) alkynylimidazole fluorophores.

Evidence for the reduced flexibility of these organic dyes when dispersed in a polymeric bundle is gathered by comparison with chloroform solution, employing classical MD simulations. Then, attention is focused on reproducing absorption spectra in PMMA, and, in particular, on the comparison between different representations of the polymeric environment during such spectroscopic calculations.

7.2 Methods

7.2.1 General approach

The applied computational protocol (sketched in Figure 7.2) can be summarized as follows:

1. QM calculations are performed to sample the conformational space of the considered dyes to find their global minimum. Energies, first and second derivatives, are computed, together with CM5 [54] point charges. Next, partial geometry optimizations are performed, constraining specific internal coordinates at selected values, in order to describe soft degrees of freedom (i.e., flexible dihedrals). The obtained data are then used to optimize FF parameters.

2. Reliable models for both PMMA and $CHCl_3$ are built, by means of MM methods using FF parameters taken from the literature. MD simulation of the fluorophores in the two different considered

environments (i.e., in the real sized polymer matrix and in $CHCl_3$ solution) are then performed and analyzed.

3. Statistically uncorrelated snapshots are extracted from the MD trajectories, once the inspection of the internal motions of the dyes indicates that the accessible conformations have been sampled with sufficient accuracy.

4. Absorption spectra are computed (according to the protocol outlined in Section 4.4) and compared on one hand with experiments and on the other hand with the results obtained using a polarizable continuum representation of the environment (i.e., the polymeric matrix).

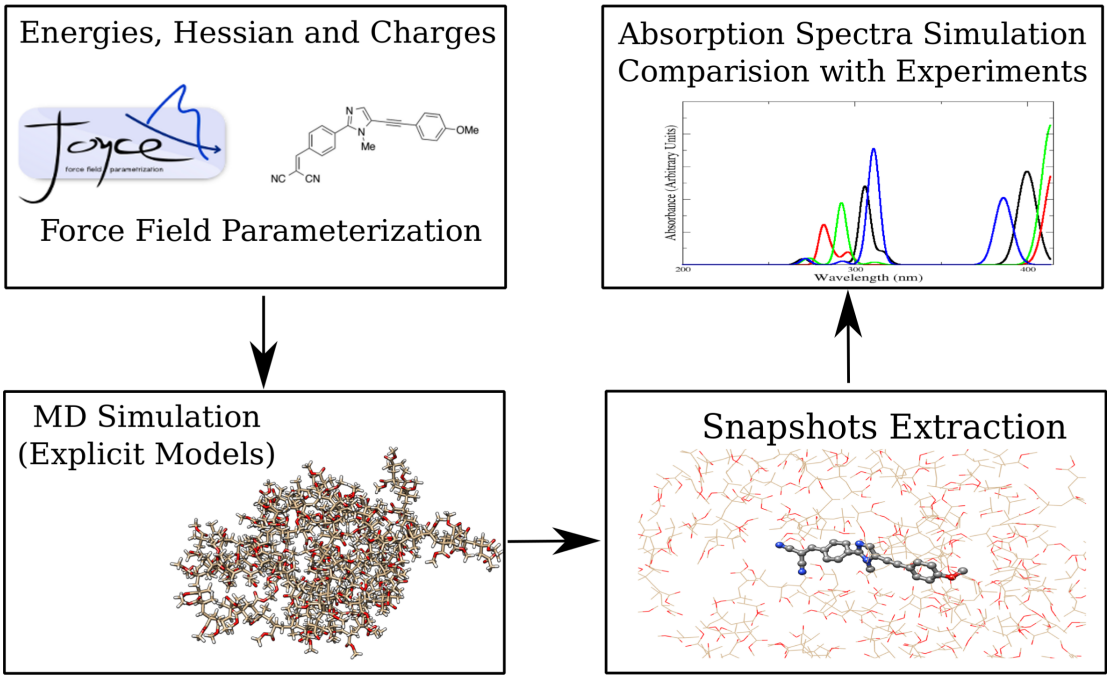


FIGURE 7.2: Sketch of the multilevel protocol employed in the present work.

7.2.2 QM calculations

QM calculations were performed to (i) obtain reference structures, for parameterizing classical FFs, and to (ii) compute absorption spectra.

For both compounds a and b, global energy minima were located at the DFT level, using the B3LYP exchange–correlation functional and the SNSD basis set [218, 219]. To reproduce experimental conditions, solvent effects were taken into account by means of the conductor-like variant of PCM [181]. In particular, butanoic acid was considered as the solvent, because of its dielectric constant (2.9931) similar to that of PMMA (3–3.3). Moreover, the Hessian matrix and harmonic vibrational frequencies were evaluated on the computed global minimum. In order to describe flexible terms in the fluorophores FFs and QM energy scans along flexible dihedrals were performed, all the other geometrical parameters being optimized at fixed values of the soft variable spaced by 30°. The structures obtained from

those relaxed scans were then used by Joyce software in order to fit dihedral FF parameters. Absorption spectra were computed for several (roughly, two-hundred for compound) statistically uncorrelated snapshots, extracted from MD simulations running in the polymeric environment. Environmental effects were taken into account by the so-called electrostatic embedding (EE) [133, 134] in which all the retained PMMA atoms are represented by their OPLS [183] point charges. At larger distances, bulk solvent effects were accounted for by means of the PCM. Electronic transitions were computed using the TD-DFT method, and considering the six lowest electronic states for each snapshot. The CAM-B3LYP functional and the SNSD basis set were used: this combination, in fact, has already been shown to be suitable for these kinds of systems [215]. Transition energies obtained for the MD snapshots were broadened by means of Gaussian functions, using a half width at half maximum (HWHM) of 0.25 eV. The final absorption spectrum is obtained by averaging the individual signals.

All the QM calculations were performed using the Gaussian 09 suite of programs [97].

7.2.3 Molecular modeling and MD simulations

A random sequence of atactic PMMA, in a completely linear conformation, was built by using an in-house Python script, thus obtaining a real-sized macromolecule of approximately 300 kDa (therefore, having 2920 monomers of methacrylic acid). Such conformation was then minimized using the conjugate gradient algorithm, until an energy threshold of 0.5 kJ/mol was reached. A replica of this minimized conformation was added and rotated, with the aim of increasing the complexity of the system. Preliminary MD runs were performed in vacuo, to equilibrate the polymer: the simulation time-step was initially set to 0.1 fs for the first 200 ps and then increased to 0.5 fs for the remaining 4.8 ns. A 14 Å cutoff was applied for both the electrostatic and the van der Waals (vdW) interactions. When a curled conformation was eventually reached by the polymer, periodic boundary conditions were applied in all the directions and the dye was manually introduced inside the polymer matrix and kept frozen at its equilibrium conformation. The system was coupled to a thermal bath at 300 K and to a pressure bath at 1 bar through weak coupling schemes [94], thus running in the NPT ensemble. Coupling constants were set to 0.1 ps and 0.5 ps, respectively. vdW forces were computed applying a cutoff distance of 13 Å whereas long-range electrostatic interactions were treated with the PME method. After 20 ns of MD simulation, the final density was found to be lower than 1.00 g/cm^{-3}, pretty far from the average experimental value (1.17–1.19 g cm^{-3}). To achieve a mass density closer to the experimental range, all the angles' force constant values were reduced by 50%. Moreover, the coupling constant to the barostat was reduced from 0.5 to 0.1 ps. Under these conditions, after 20 ns of simulation, a density value of 1.20 g cm^{-3} was obtained. Such collapsed structure was finally re-equilibrated by restoring the system under the initial conditions. After 5 ns, the density value was detected to be stable at 1.15 g m^{-3} (a reasonable value taking into account the complexity of the simulated system). The modeling process of the polimer matrix is summarized in Figure 7.3.

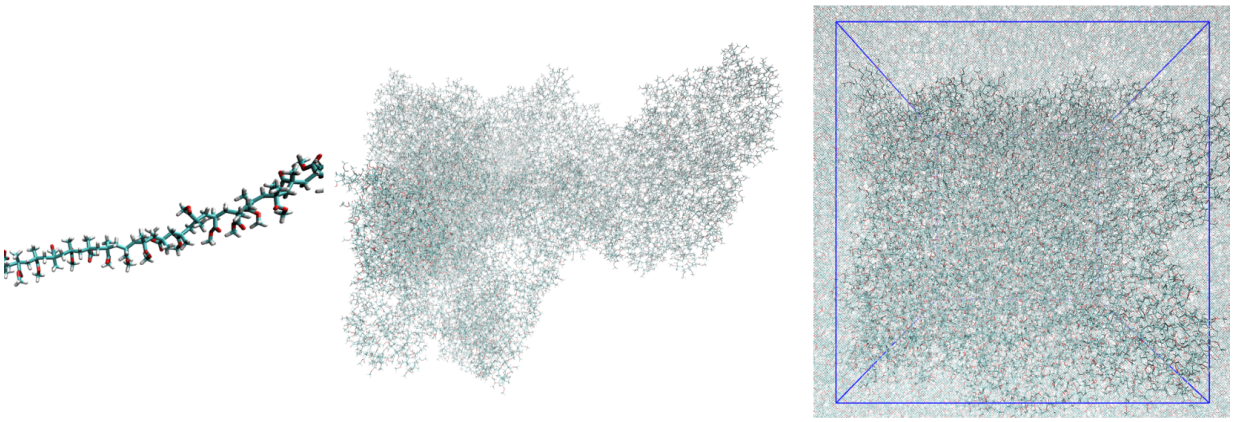


FIGURE 7.3: PMMA equilibration process. From left to right: a portion of the starting linear structure; tangled conformation obtained during the equilibration procedure; final equilibrated structure with PBC.

7.3 Results

7.3.1 Fluorophore force fields

Van der Waals (vdW) parameters were directly transferred from published force fields. For both **a** and **b**, CM5 charges were computed at geometries optimized in butanoic acid at the DFT/PCM level.

No evident differences were found on the two minima, as shown in Figure 7.4, where the **a** and **b** minima structures are superimposed. In particular, values of 28° and 30° for the dihedral angle between the imidazole and the pull-phenyl was found for **a** and **b**, respectively.

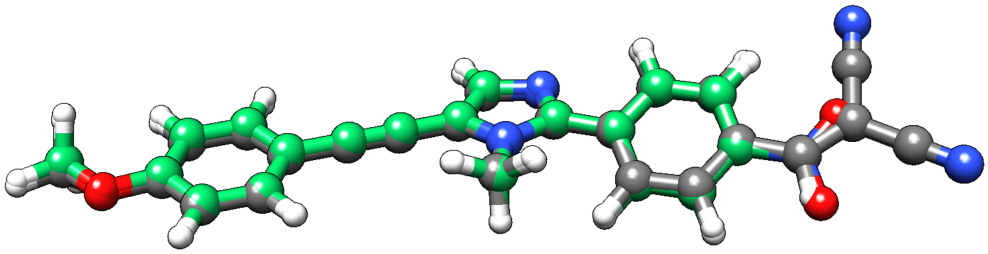


FIGURE 7.4: Superposition of **a** (carbons in grey) and **b** (carbons in green) computed minima.

Starting from **a**, and as already mentioned in Section 3.1, the parameterization was performed through the minimization of the Joyce merit function for each scanned internal coordinate, obtaining an overall standard deviation of 0.00624 kJ/mol. The same equilibrium geometrical parameters and force constants were used for equivalent internal coordinates only if contained within the same molecular block, e.g. the phenyl-pull or the phenyl-push moiety. The flexibility of the molecule under examination does not permit the use of only the absolute energy minimum to obtain a reliable FF. Therefore, a reference QM scan was performed for all the dihedral angles that could affect the overall conformation (and indicated in Figure 7.5 as $\delta 1$–5). The obtained energy profiles for the five dihedral angles are shown in Figure 7.5, where single point energies calculated at the QM level are compared to their FF counterparts.

The good agreement between both levels of theory (see Figure 7.5) points out that the Joyce FF performs a remarkable job in reproducing the structures and relative energies of different conformers.

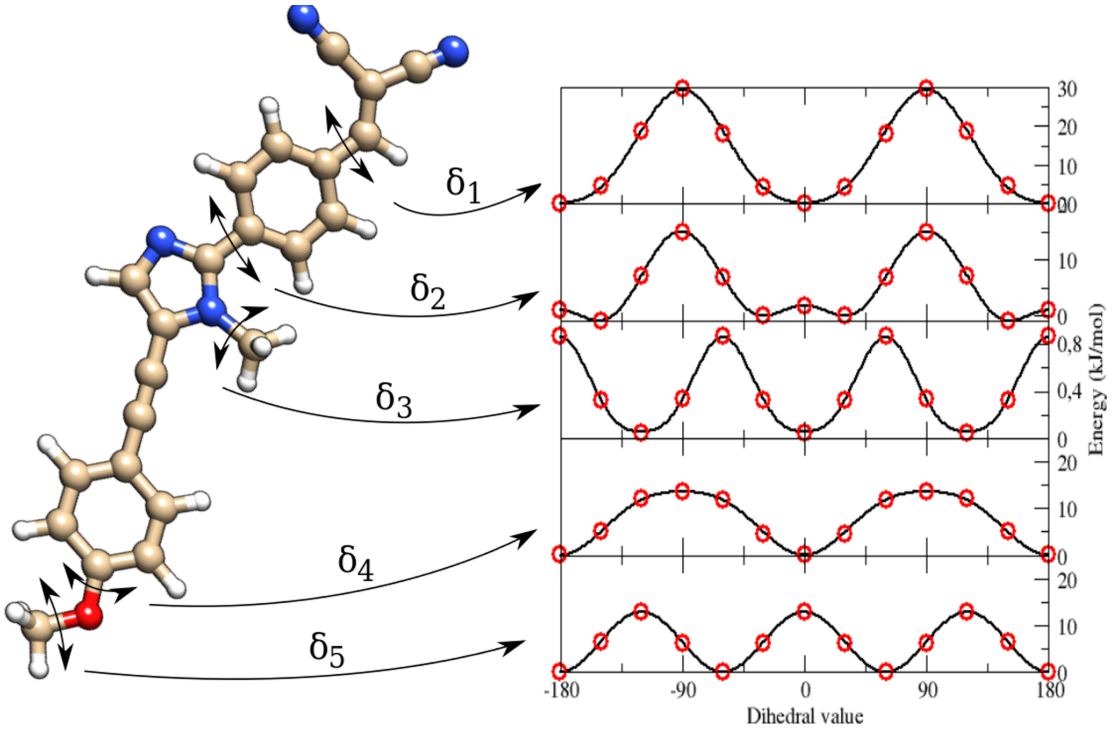


FIGURE 7.5: Left: Structure of the dye **a**. Right: Energy profiles along the five parameterized flexible dihedrals (from $\delta 1$, on top, to $\delta 5$, on bottom) at the QM (red circles) and the Joyce FF (continuous line) levels.

All the potential energy curves along the five torsional angles are symmetric about the origin. The $\delta 1$ and $\delta 4$ dihedral angles show energy minima at 0° and 180°, i.e. for planar arrangements of the whole molecule. The largest geometry variations are related to the $\delta 2$ dihedral angle, which shows four different energy minima at ±30° and ±150°. However, planar conformers (0° and 180°) are also quite stable due to the significant electron delocalization between the two connected rings. The capability of the FF of a to reproduce the vibrational behavior of the molecule was checked by comparing QM and FF harmonic frequencies. Figure 7.6 shows that the QM trend is well reproduced by the FF calculations, with a root-mean square deviation of 82 cm^{-1}, which can be considered satisfactory for the purposes of the present study.

With the aim of developing a FF also for the **b** dye, a somewhat different strategy was used: FF parameters for the common, planar structure were simply transferred from **a** to **b**. Then, reference geometrical parameters and bonding force constants relative to the nitro group were computed for the **b** molecule, freezing all the other parameters at the values optimized for a.

7.3.2 MD simulation analysis

The enhanced rigidity of the fluorophores upon dispersion in the polymer bundle was postulated to explain the reduced efficiency of quenching effects [215]. In fact, quantum yields determined in solution and PMMA have essentially the same value (≈0.1): while this parameter in solution is widely affected by large conformational changes that take place between the ground and the excited states, in the polymer matrix these structural conversions are not allowed. The goal of the MD simulations is therefore two-fold: on the one hand they permit the generation of statistically uncorrelated snapshots

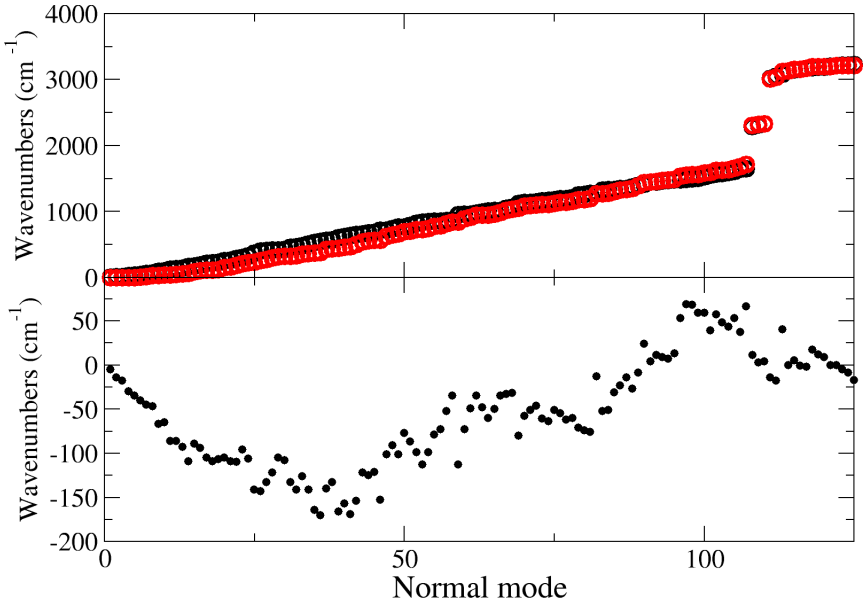


FIGURE 7.6: Top: Computed vibrational frequencies (both at the QM and FF levels) of fluorophore **a**. Bottom: Differences between the two descriptions.

on which spectroscopic properties can be computed and, on the other hand, they also permit demonstrating, at an atomistic level, the reduced flexibility of the dispersed molecules, due to polymer caging effects. In the following only the dye **a** is considered, since results related to dye **b** are almost identical to **a**. Interesting aspects are evidenced by inspection of Figure 7.7, where the population distribution of the $\delta 1$ and the $\delta 4$ flexible dihedral angles are plotted for the two investigated environments.

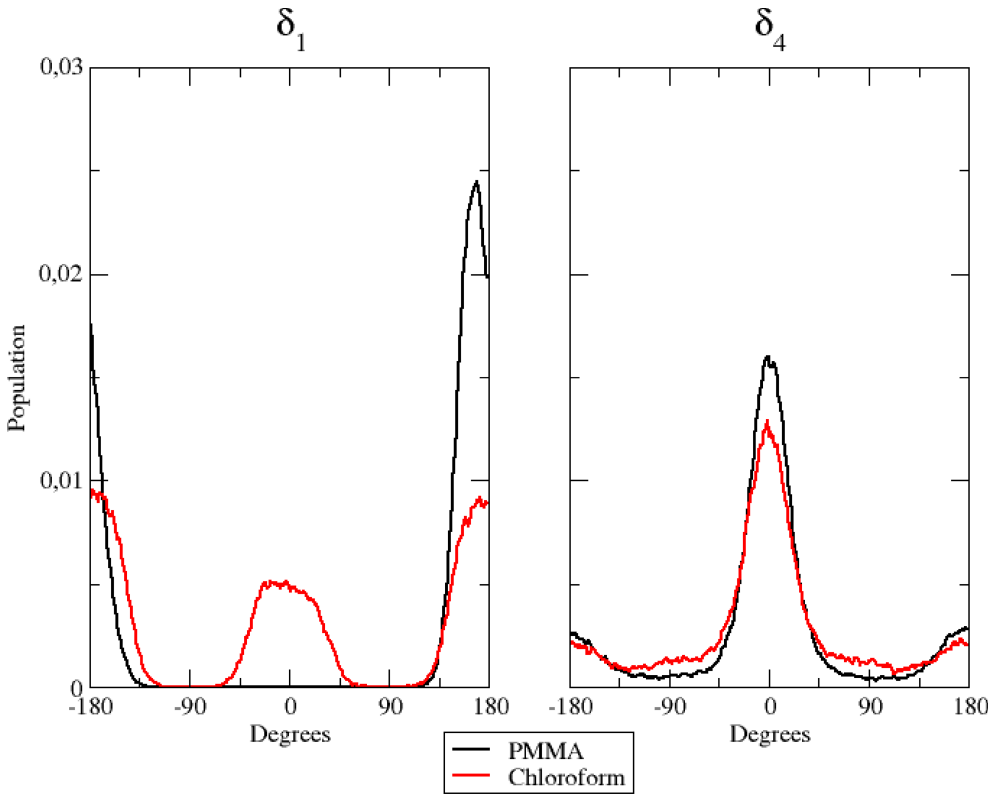


FIGURE 7.7: Population distribution of the $\delta 1$ (left panel) and the $\delta 4$ (right panel) flexible dihedrals in the two considered environments.

In the left panel, it is easy to note differences among the behavior of the dye when dispersed in the polymer matrix and in solution: in the former case, $\delta 1$ is frozen at approximately 180°, whereas both planar conformations (0° and 180°) are populated in the latter case. This fact could be due to the hindering effect exhibited by the polymer chains: their steric hindrance prevents the dye from any conformational rearrangement along the considered dihedral, since such torsion implies the movement of a large and bulky group (the dicyanovinyl moiety). Concerning next $\delta 4$ (right panel in Figure 7.7), no significant differences were found between the two different environments: nonetheless, in PMMA, the population is more peaked at 0° than in chloroform solution, highlighting again the constraining effect of the polymer matrix. This trend is confirmed by the behavior of $\delta 2$.

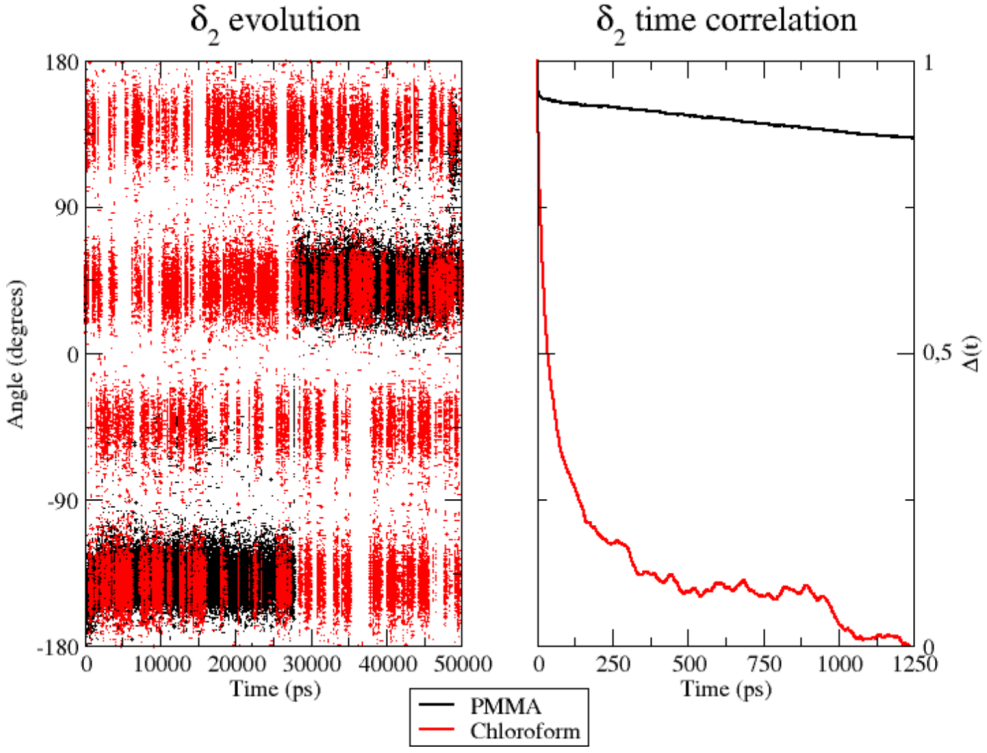


FIGURE 7.8: Left: Time evolution of the $\delta 2$ dihedral in the PMMA matrix (black) and chloroform solution (red). Right: Time correlation of the $\delta 2$ dihedral in the two environments.

As a matter of fact, the inter-conversion between the allowed conformations of $\delta 2$ is much faster in $CHCl_3$ than inside the polymer (Figure 7.8, left panel): indeed, during the considered 50 ns, only two well-defined transitions from one conformer to the other were observed, the first one occurring after more than 15 ns.

Furthermore, in $CHCl_3$, the four different conformers of the dye show comparable populations. In contrast, it is clear that 50 ns are not sufficient to obtain a fully equilibrated population of all the energy minima when the dye is within the PMMA thin film, because of the very slow structural conversions allowed by the polymeric entanglement. This observation was also confirmed by looking at the $\delta 2$ time autocorrelation function (ACF), shown in the right panel of Figure 7.8. This is apparent from the trend of the fluorophore to assume different conformations along $\delta 2$ in chloroform solution, with the $\Delta(t)$ value approaching zero after only 1 ns of simulation. Conversely, the same dihedral appears to be constant, remaining highly correlated for a much longer time, when the dye is within the polymer environment.

Further evidence of different flexibilities of the investigated dye among the PMMA chains and the chloroform solutions was provided by the analysis of other structural features, such as (i) the radius of gyration and (ii) the distance between the pull- and the push-phenyl rings, placed at the ends of the dye. Such parameters give an overall description of the internal mobility of the fluorophore in the medium under examination. The radius of gyration has been calculated according to Eq. 4.3, and its normalized trend is plotted in Figure 7.9, left panel.

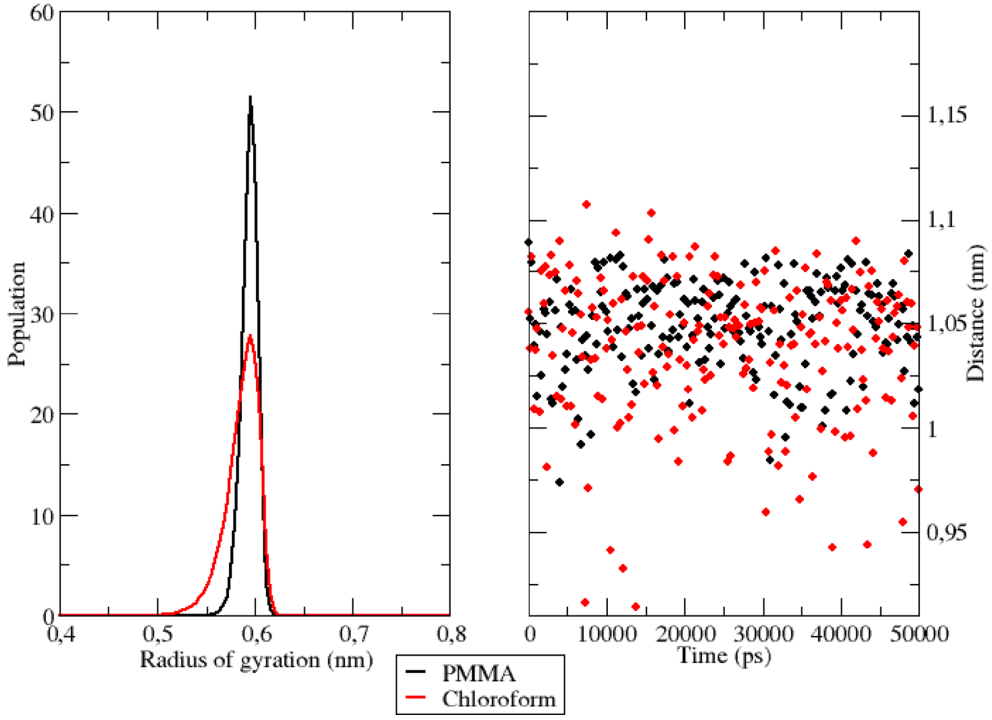


FIGURE 7.9: Left: Normalized gyration radius distribution of the investigated dye among the sampled time in the two different simulations. Right: Distance evolution between the two phenyl rings.

The hardness encountered by the dye is evident in the polymer simulation, where the radius of gyration is stable at 0.6 nm. Similar conclusions can be drawn by looking at the evolution of the distance between the push- and the pull-moieties during the simulation. As shown in Figure 7.9, right panel, the polymer does not allow the dispersed molecule to relax along its major axis, thus keeping the distance between the two considered groups around 1.05 nm for the whole time span. In contrast, in the chloroform solution, the distance between the two considered groups oscillates frequently, thanks to the higher flexibility of the surrounding environment.

The computation of dynamic properties in the two different media could permit gaining further insights also concerning the effects of the considered environments on the fluorophore's dynamics. The mean square displacement (MSD) of the center of mass of **a** was computed for dye **a** in the two surroundings.

Figure 7.10 shows the MSD variation during the first 50 ps. The behavior of the dye in chloroform is typical for a molecular solute in a liquid solvent, where a diffusive regime is quickly approached. In contrast, in the polymer matrix, the MSD value suddenly reaches a plateau: the solute is trapped, its motion being strongly hindered by the surrounding PMMA cage, which delays the establishment of a

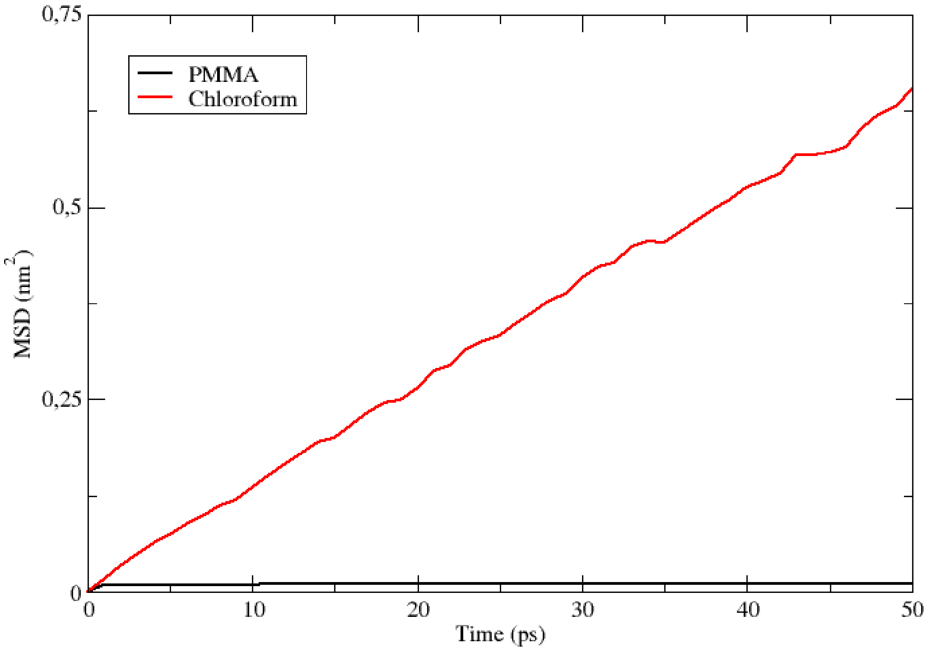


FIGURE 7.10: Comparison between the MSD(t) values computed in the PMMA matrix and chloroform solution.

diffusive regime. Moreover, the absence of a sub-diffusive behavior indicates that the polymer chains are also not diffusing, i.e. the polymer behaves as a plastic, thin film.

7.3.3 UV absorption spectra

All the results analyzed in the previous section point out a strong rigidity of the dyes induced by the caging effect of the polymeric matrix. It is, therefore, natural to assume that structural changes between the ground and excited electronic states are severely restricted in the polymer bundle. Keeping this in mind, it seems reasonable to simulate polymer hindering effects by the previously adopted static approach [215], where the $\delta 2$ dihedral angle is fixed to its ground-state value and bulk polymer effects are taken into account by the PCM. MD simulations on the contrary permit checking this simplified model by looking at the populations of the different conformers along the time.

Both the approaches mentioned above (the static and the dynamic one) were applied and compared with experiments. Inspection of Figure 7.11 shows that both the static and dynamic approaches overestimate the absorbance intensity at the maximum absorption wavelength. Nonetheless, the dynamic approach reproduces very well the first peak at approximately 300 nm, whose intensity is underestimated in the static model. This last approach shifts the absorption wavelength by about 11 nm, while the same experimental peak is underestimated by 7 nm using the dynamic approach. The same observations could be extended to system **b** (7.11 11, right panel). The overall shape of the experimental spectrum is again better reproduced by the dynamic simulation than by its static counterpart. The two approaches show, instead, comparable errors (of opposite sign) concerning the position of the peak maximum.

Computational and experimental findings are summarized in Table 7.1. Globally, inspection of the shape of the computed spectra evidences a good match with experiments, especially concerning the decay slope. Furthermore, the good agreement between the results obtained from atomistic and

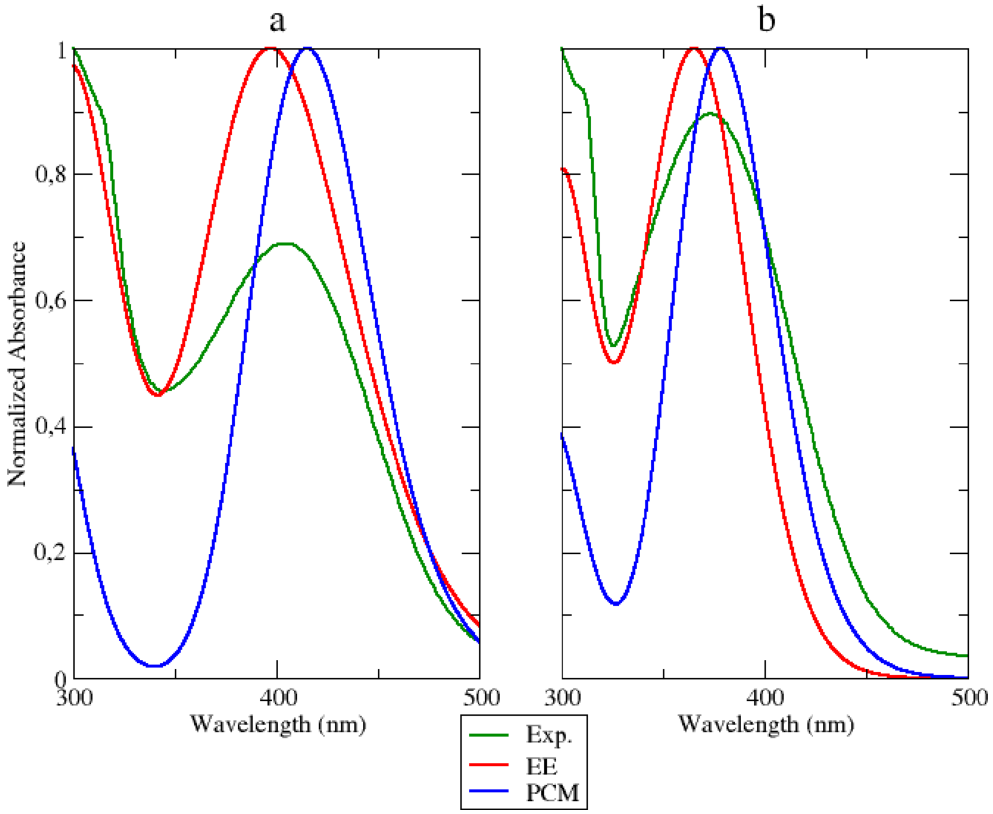


FIGURE 7.11: Absorption spectra of **a** (left) and **b** (right panel) computed using the dynamic approach (EE, continuous red line) and the static approach (PCM, blue line), compared to the experimental spectrum (Exp., continuous green line).

continuous descriptions of the environment suggests the absence of strong specific solute–solvent interactions (especially of electrostatic origin). Although the dynamic atomistic approach improves the reproduction of the spectral shape (especially in the area under 350 nm), it is worth noting that PCM performs a remarkable job in simulating the main features of the polymer embedding, permitting a preliminary semi-quantitative analysis of general trends by a fast approach.

System	Method	Absorption peak (nm)
	Experimental	404
a in PMMA	Static approach (PCM)	415
	Dynamic approach (EE)	397
	Experimental	373
b in PMMA	Static approach (PCM)	378
	Dynamic approach (EE)	365

TABLE 7.1: Wavelength values calculated at the lie peak for both **a** and **b** upon dispersion in the PMMA matrix.

7.3.4 Conclusion

Purposely tailored FFs have enabled us to follow the time evolution of two alkynylimidazole fluorophores when dispersed in a PMMA matrix intended to be part of a sunlight-harvesting device. Careful analysis of MD trajectories highlighted an overall rigidity of such hetero-aromatic molecular structures, due to the presence of the polymer chains, which hindered inter-conversions between different low-energy conformers. Analogous simulations in chloroform solution showed, instead, a significant flexibility of the dyes. These results confirm the hypothesis advanced recently to explain the different behavior of these dyes in solution and in the polymeric matrix.

Next, the attention was focused on the computation of absorption spectra of the same fluorophores within the polymer matrix. This was achieved by using two different methods, i.e. (i) a static approach, where spectroscopic computations were made only on the conformational minimum, modeling the environment using PCM, and (ii) a dynamic approach, where the same computations were made on several snapshots extracted from the MD trajectories, and explicitly representing the polymer through the so-called electrostatic embedding. Even if a more correct reproduction of absorption spectra was achieved by the application of the second, more demanding, protocol, it has to be noticed that a static approach also works pretty well. In particular, the static protocol is confirmed in this work to be a very useful tool in virtual screening campaigns, where the best structures for the above-mentioned applications have to be chosen by fast yet sufficiently accurate methods. It is, however, apparent that reliable optical features can be obtained using only force fields tailored for the correct reproduction of structural properties.

The proposed approaches could also be extended to the simulation of emission spectra: the validity of the CAM-B3LYP theoretical scheme in the computation of emission properties is currently under investigation, together with the parameterization of ad hoc FFs for the excited states of the two tested dyes, to be used during MD simulations in the dynamic approach. Future applications of the proposed procedures could also cover the inclusion of new, polar polymer matrices, as well as other condensed-phase environments, in order to further assess the validity of both implicit and explicit models.

Chapter 8

Computational Study of a Fluorescent Molecular Rotor in Various Environments

Fluorescent molecular rotors (FMRs) belong to an important class of environment sensitive dyes capable to effectively report on viscosity and polarity of their microenvironment. FMRs have found widespread applications in various research fields, ranging from analytical to biochemical sciences, for example in intracellular imaging studies or in volatile organic compound detection. Here, a computational investigation of a recently proposed FMR, namely the 4-(diphenylamino)phthalonitrile (DPAP), in various chemical environments is presented. A purposely developed force field is developed and then applied to simulate the rotor into a high- and low-polar solvent (i.e., acetonitrile and cyclohexane), a polymer matrix and a lipid membrane. Subtle effects of the molecular interactions with the embedding medium, the structural fluctuations of the rotor and its rotational dynamics are analyzed in some detail. Results correlate with available experimental data, thus supporting the reliability of the model, and provide further insights on the environment-specific properties of the dye. In particular, it is shown how molecular diffusion and rotational correlation times of the FMR are affected by the surrounding medium and how the molecular orientation of the dye becomes anisotropic once immersed in the lipid bilayer. Moreover, a qualitative correlation between the FMR rotational dynamics and the fluorescence lifetime is detected, a result in line with the observed viscosity dependence of its emission. Finally, optical absorption spectra are computed and successfully compared with their experimental counterparts.

8.1 Background

During the last decades, the chemistry toolbox has been significantly expanded by the development of new organic fluorophores characterized by innovative structural and optical features. In particular, molecular dyes able to modulate their photophysical properties in response to different surrounding environments are nowadays employed in wide areas of chemical research, ranging from environmental to photovoltaics applications [220], and from biology to medicine [221–224]. In this context, fluorescent molecular rotors (FMR) have gained much attention owing to their simple synthesis and versatility [225–227]. Typically, FMRs are characterized by an electron acceptor moiety and an electron donor unit, which are connected by a flexible spacer with conjugated bonds. Such a chemical linker ensures, upon excitation, an electronic density shift from one unit to the other and confers to FMRs a peculiar

sensitivity towards local viscosity and polarity of the environment. In FMRs, emission is finely modulated by intramolecular structural changes occurring in the excited state, in addition to solvent dipolar relaxation. Moreover, in most FMRs the fluorescence signal stems from the competition between a locally (bright) excited state and a twisted intramolecular (either bright or dark) charge-transfer (TICT) state [225]. As a result, FMRs have been successfully employed as intracellular microviscosity detectors for $in\ vivo$ applications [228, 229]. This is relevant in view of the connection of plasma and cellular viscosity changes with biochemical processes and diseases [230].

Among the many FMRs reported to date, the recently synthetized 4-(diphenylamino) phthalonitrile (DPAP) turned out to be the prototype of a novel class of FMRs [231]. DPAP chemical structure presents a tertiary amine electron donor and two nitrile groups acting as electron acceptor moieties, embedded in a π-extended conjugated system (Figure 8.1). In contrast to most FMRs, which are characterized by a locally excited (LE)/TICT state mechanism, DPAP photophysical behavior is basically modulated by a free rotational motion of its phenyl rings [231]. DPAP spectroscopic response and its polarity and viscosity dependence have been already exploited in several applications [232–234]. Usually the $modus\ operandi$ of FMR photophysical mechanisms, including DPAP, are typically addressed by QM investigations of their electronic excited states and conformational changes. However their detailed structural and dynamical features in complex molecular environments and their specific interactions with the surroundings, which ultimately modulate the FMR spectroscopy, still remain largely elusive. Due to the several interplaying effects that directly and indirectly affect FMRs upon dissolution in condensed-phase systems, a thorough *in silico* investigation may help to properly identify the molecular determinants of the recorded experimental observables and possibly uncover the subtle relationship between molecular dynamics and spectroscopy [235, 236].

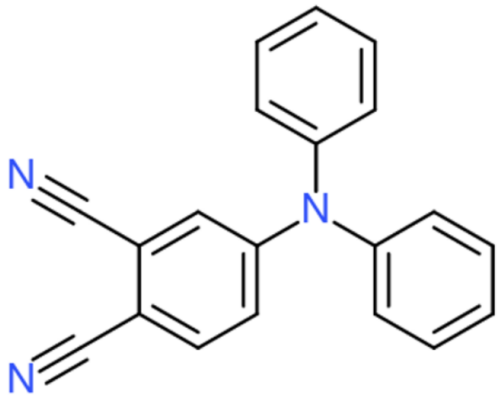


FIGURE 8.1: 4-(diphenylamino)phthalonitrile (DPAP) structure.

In this work, a MD study of DPAP in multiple environments has been carried out in order to describe the effect of the embedding medium on the structural and dynamical behaviour of the rotor. Acetonitrile (ACN), tetrahydrofuran (THF), *o*-xylene and cyclohexane were considered as solvents, to include a reasonable range of bulk properties (as static dielectric constant and viscosity). Additionally the study has been extended to include the atactic poly(methyl methacrylate) polymeric matrix (PMMA) and the hydrated 1,2-dioleoyl-*sn*-glycero-3-phosphocholine (DOPC) lipid bilayer, since applications of DPAP in both polymeric films [233] and in cell membrane environments [234] have been

reported in previous studies. Hence, DPAP has been chosen as an illustrative example for the computational treatment of a FMR in different target environments. A multistep computational protocol has been set out [235], which involves the development of a reliable ground-state molecular model of the rotor based on QM calculations, extensive atomistic simulations in all the environments and subsequent spectroscopic calculations of the optical absorption spectra. On the other hand, emission spectra, whose modeling would require the use of the excited-state force field, will be addressed in a future work, following the same computational procedure established in the present study. It is worth noting that a reliable force field is crucial in this context, owing to sensitivity of the dynamical and spectroscopic response to conformational changes of the FMR [231]. Analysis of the MD trajectories has allowed to shed some light on the structural and dynamic properties of the dye within the considered embeddings and concurrent local structural perturbations of the surroundings. Furthermore, the rotor mobility and rotational dynamics were scrutinized in view of the dye-environment specific interactions. Interestingly, in some of the environements a relation between molecular rotational dynamics and fluorescence lifetime has emerged from the present analysis. This result may have far-reaching implications for the possible exploitation of spectroscopic techniques to gather detailed molecular information in a large number of materials.

8.2 Methods

8.2.1 QM calculations and force field parameterization

All intramolecular terms were parameterized by using the Joyce software [9] by fitting energy gradients and Hessian matrix to corresponding QM data. In particular, bond, angle and stiff torsion terms were fitted to a QM Hessian matrix computed at DPAP optimized geometry, while torsional potentials of flexible dihedral angles were further refined through relaxed potential energy surface (PES) scan calculations. DPAP optimized geometries, energies and Hessian matrix were computed at the Density Functional Theory (DFT) level, according to the B3LYP exchange-correlation functional and the SNSD basis set [218, 219]. Bulk solvent effects were taken into account by means of the C-PCM [181]. Atomic partial charges were computed according to class IV CM5 charges [54] at the minimum energy configuration, whereas Lennard-Jones parameters were transferred from the OPLS/AA FF [2]. Acetonitrile and cyclohexane only were considered within the C-PCM to evaluate the influence of solvent polarity on DPAP structural and electronic properties, however FF parameters and atomic charges were finally based on acetonitrile owing to negligible differences in the obtained parameters (e.g., the largest atomic charge deviation was 2.98 x 10^{-2}e). Vertical transition energies were computed at the CAM-B3LYP/SNSD level of theory on selected configurations sampled during the MD simulations. All QM calculations were performed using the Gaussian 09 suite of programs [97].

8.2.2 MD simulations

Classical MD simulations of DPAP in different environments were performed using the GROMACS (ver. 4.6.5) software package [98]. The OPLS-AA FF [2] was used to describe intramolecular and intermolecular potential of tetrahydrofuran, *o*-xylene and acetonitrile with the exception of ACN atomic charges, which were estimated using the CM5 population analysis [54]. For cyclohexane, the general Amber

TABLE 8.1: Technical details of the performed Molecular Dynamics simulations.

Environment	N. of molecules	Box edge (nm)	Force Field
Acetonitrile	3053	6.62	OPLS-AA[2]
THF	962	5.24	OPLS-AA[2]
o-Xylene	556	4.82	OPLS-AA[2]
Cyclohexane	997	5.69	GAFF[59]
DOPC bilayer	200 DOPC 5791 H_2O	8.27 x 8.27 x 6.26	CHARMM[63]
PMMA	2 chains of 2920 monomers	9.45	OPLS-AA[2, 241]

FF (GAFF) for organic molecules [59] has been selected, since preliminary investigations showed that poor results were obtained in reproducing density with OPLS-AA. PMMA coordinates and topology were taken from the work of Chapter 7, using standard OPLS parameters. The hydrated (by means of TIP3P water molecules [237]) DOPC bilayer was modeled according to CHARMM FF [63], which is able to reproduce available experimental information on the structure and dynamics of phospholipid bilayers reasonably well.[238–240]

One DPAP molecule was solvated or embedded into a number of molecules representing the above mentioned environments enforcing periodic boundary conditions (see details in Table 8.1). In particular, in order to simulate DPAP in the hydrated DOPC bilayer, a rectangular box was chosen. After a steepest descent energy minimization, the systems were slowly heated up from an initial temperature of 150 K to 300 K for about 500 ps using the velocity-rescale thermostat [95] and a coupling constant (τ) equal to 0.1 ps. All systems, except DPAP in lipid bilayer, were equilibrated for 1 ns (with a timestep of 1.0 fs) in a NPT ensemble, using the Berendsen barostat [94], and the velocity-rescale thermostat with coupling constants of 1.0 ps and 0.1 ps respectively. In the case of lipid bilayer, the equilibration run lasted 6 ns according to a NPT ensemble, using the semi-isotropic pressure coupling. Afterwards, all production runs were carried out in the NVT ensemble, using the velocity-rescale thermostat (T=298.15 K and τ=0.1 ps) and increasing the integration timestep from 1.0 to 2.0 fs. Fastest degrees of freedom were constrained with the LINCS algorithm [93]. In the case of cyclohexane, only bonds with hydrogen atoms were kept rigid. The total sampling time was about 130 ns for all the systems. Electrostatic potential was described using the PME method [89], using a real-space cutoff of 1.4 nm and spline interpolation of order 4. VdW interactions were computed applying a cutoff of 1.4 nm. OPLS combination rules were used. System coordinates were stored every 500 steps (i.e., each picosecond). Trajectories analysis were performed with the TRAVIS package [191] and homemade scripts.

8.3 Results and Discussion

8.3.1 DPAP force field

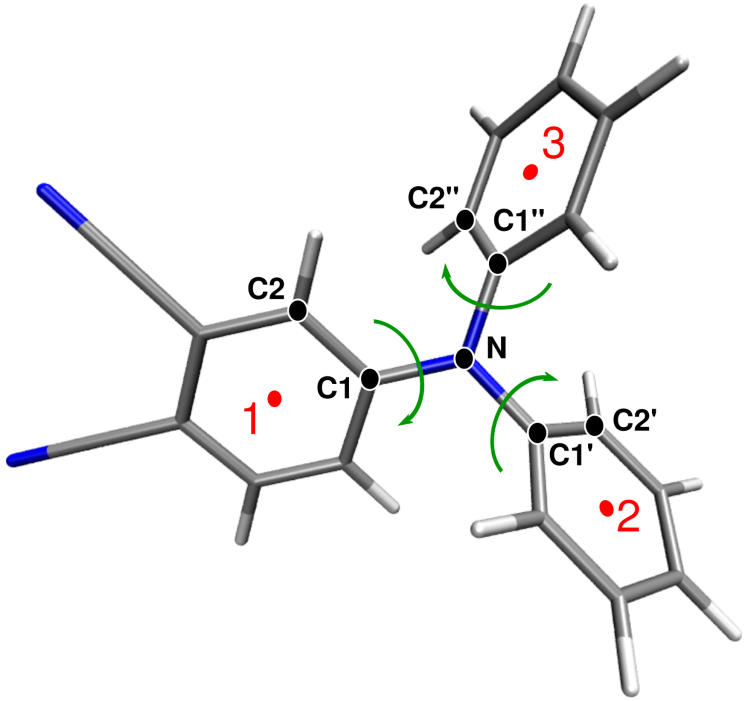


FIGURE 8.2: DPAP propeller-like conformation is indicated with green vectors. The rings centers (**1**, **2** and **3**), the *ipso* (C1, C1' and C1") and the *ortho* (C2, C2' and C2") carbon atoms are labeling in red and black respectively.

DPAP optimized structure adopts a propeller-like conformation in order to minimize steric hindrance among the three phenyl rings (see Figure 8.2). DPAP belongs to the C_1 point group since the presence of cyano substituents on one ring breaks the D_3 symmetry that otherwise characterizes the parent molecule (i.e. triphenylamine). The central NC1C1′C1″ moiety (for atom labeling see Figure 8.2) adopts a nearly planar geometry. From DFT calculations, the C1NC1′ and C1NC1″ angles are equal to about 121°, whereas C1′NC1″ is 117° in acetonitrile (i.e., the highest polar solvent) and in cyclohexane (lowest polar solvent), the difference between C1NC1′ (or C1NC1″) and C1′NC1″ being ascribed to the inductive and resonance effects of the two cyano groups on ring 1.[242] The three main degrees of freedom, characterizing DPAP conformational changes in solution, are the ring torsional angles with respect to the central amine group (i.e. NC1C1′C1″), hereafter referred to as dihedral **1** (C2C1NC1″), dihedral **2** (C1″NC1′C2′) and dihedral **3** (C2″C1″NC1′). Owing to the subtle interplay between structural conformation and photophysical properties, special attention was due to the parameterization of the torsional potentials along such dihedral angles, as described in the following.

Starting from the DFT optimized geometry of the dye, distinct relaxed PES scans along the dihedral angles were performed (Figure 8.3): each torsional angle was modified in multiple steps (at least 25), then the dye structure was relaxed keeping frozen the accounted dihedral angle to avoid spurious distorted conformations due to close interactions between the phenyl rings. Note that the potential energy curves (PEC) of dihedral angle **2** and **3** are equivalent, in this case. Solvent (i.e. acetonitrile) effects have been included implicitly in calculations. The obtained DFT PEC was used to refine torsional

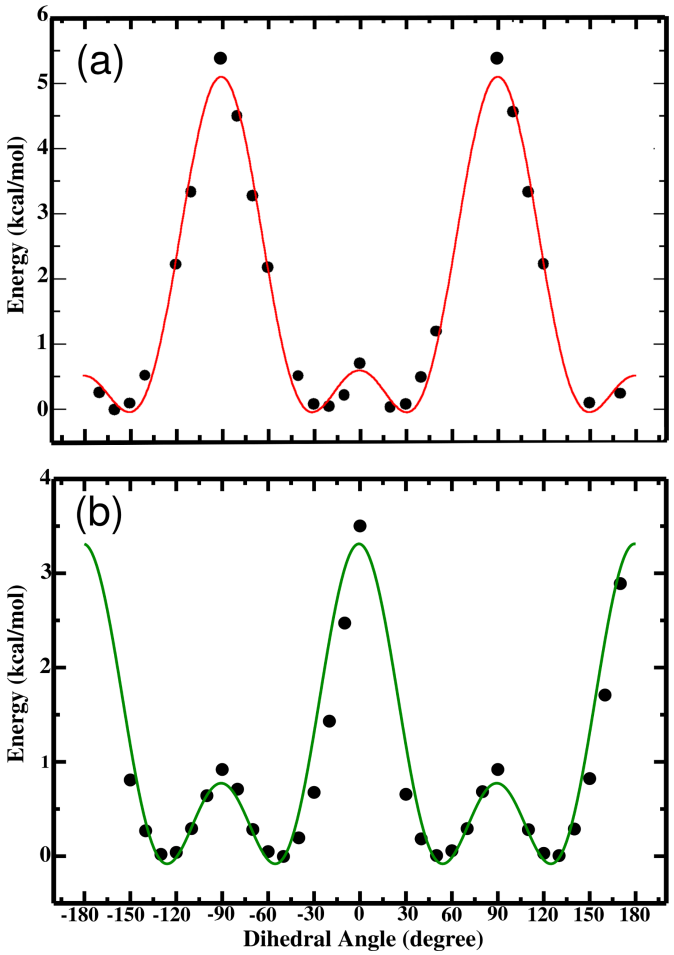


FIGURE 8.3: Torsional profile comparison between QM (black points) and Joyce (lines) for dihedral angles **1 (a)** (in red) and **2** and **3 (b)** (in green).

potential terms of the DPAP FF. Results are depicted in Figure 8.3, where DFT and MM PEC profiles show negligible differences. Both PEC are symmetric with respect to the planar geometry (at either 0° or 180°) and show four minima in correspondence to the propeller-like conformations.

Overall, the PEC profile is typical of aromatic amines: a similar profile, for example, has been obtained for the triphenylamine [243]. In particular, dihedral angle **2** and **3** show symmetry-related energy minima at ±130° and ± 50°, whereas dihedral **1** minima are located at ± 25° and ± 155°. These results are consistent with DPAP optimized geometry, in which **1**, **2** and **3** are, respectively, 22°, 51° and 125° in acetonitrile and 23°, 53° and 127° in cyclohexane. Two energy barriers characterize the interconversion among such energy minima, whose positions correspond to the planar and orthogonal geometry of the considered ring with respect to the central amine moiety: a small one of less than 1 kcal/mol and a larger one of about 3-5 kcal/mol. Interestingly, there is an apparent swap of such barriers in the two type of torsional angles: the highest energy barrier corresponds to the orthogonal configuration in case of dihedral **1** and to the planar configuration in case of dihedral **2** and **3**. As a consequence, the torsional potentials of the three dihedral angles cannot be treated equivalently. This peculiar observation is ultimately due to the resonance effect of the cyano groups on ring 1, which confer to an extra stabilization energy to the planar geometry. For the unsubstituted phenyl rings, the orthogonal conformation is energetically more favourable with respect to the planar one because the

steric hindrance is minimized. The situation is reversed in case of the di-substituted ring (i.e., ring 1).

8.3.2 DPAP in solutions

MD simulations of the dye in liquids, i.e. acetonitrile, tetrahydrofuran, *o*-xylene and cyclohexane, were carried out at normal conditions. DPAP intermolecular interactions and solvation have been analyzed in terms of the rdf profiles.

Figure 8.4 shows the rdf issuing from the center of mass (COM) of solvent molecules (acetonitrile, tetrahydrofuran, *o*-xylene or cylohexane) and the center of each of the three DPAP aromatic rings (RC). Noticeable differences are detected for ring 1 compared to ring 2/3. In the case of acetonitrile (Figure 8.4, blue line), the first main peak located at approximately 5 Å is clearly higher for the unsubstituted aromatic rings. Such peak takes a height of 1.25 and 0.9 for ring 2/3 and 1, respectively. A second less pronounced peak appears at 10 Å. A similar, but more structured, profile is found in cyclohexane (Figure 8.4, green line), where ring 1 is again less interacting with surrounding solvent molecules. The first peak is located at 6 Å of COM$\cdots$RC distance. Three more peaks are observed at 11, 15 and 20 Å. The four peaks are smooth and well resolved, thus indicating a well-defined solvent structure, as already pointed out by previous theoretical and experimental studies on cyclohexane liquid solution.[244, 245] The difference between ring 1 and 2/3 can be ascribed to the two cyano substituents, which turned away solvent molecules from ring 1 center. In case of THF and *o*-xylene solvents, the rdf profile is intermediate between the two more structurally different solvents, acetonitrile and cyclohexane, as can be admired in Figure 8.4. Ring 1 in particular, appears to be less solvated, compared to the cyclohexane case, with a peak height of 1.2 in both solvents. Looking at the other rings (2 and 3), the distribution is closer to the case of cyclohexane, though a lower interaction (of height 1.6 approximately) between ring center and THF COMs is detected.

To highlight specific interactions between the cyano groups and the solvent, the rdf between the cyano nitrogens and the solvent hydrogen atoms (i.e. the methyl hydrogen atoms of acetonitrile and the cyclohexane hydrogen atoms) are considered. The corresponding rdfs are depicted in Figure 8.5. The N$\cdots$H intermolecular interactions are well established in acetonitrile, with a well-defined peak (height of 1.3) located at approximately 2.6 Å, with an integral value computed at the end of the peak of 11. These results indicate that nitrogens are well disposed to interact with up to four acetonitrile molecules. A second peak is present at 4 Å. For cyclohexane, no specific interactions have been found which is consistent with its molecular symmetry. The same observation made in acetonitrile can be easily extended to *o*-xylene, since both corresponding structures present a CH_3 group.

The acetonitrile solution was also analyzed to observe the opposite interaction, i.e the one that involves DPAP aromatic ring H atoms and acetonitrile N atoms. In this case, acetonitrile acts as a hydrogen bonding acceptor. Figure 8.6 shows individual rdf for each H atom belonging to DPAP. Note that DPAP H atoms are considered equivalent under the assumption that the three rings may undergo free rotations. Steric obstruction and ring oscillations prevent solvent molecules from approaching H atoms in ortho position with respect to the tertiary amine nitrogen. This is the reason why the corresponding rdf are the lowest (see red, blue and green lines in Figure 8.6). The other hydrogen atoms instead are easily accessible and can interact with acetonitrile. In Figure 8.6 the hydrogen atom colored in black is the most disposed to interact with acetonitrile, being assisted by the nearby cyano groups.

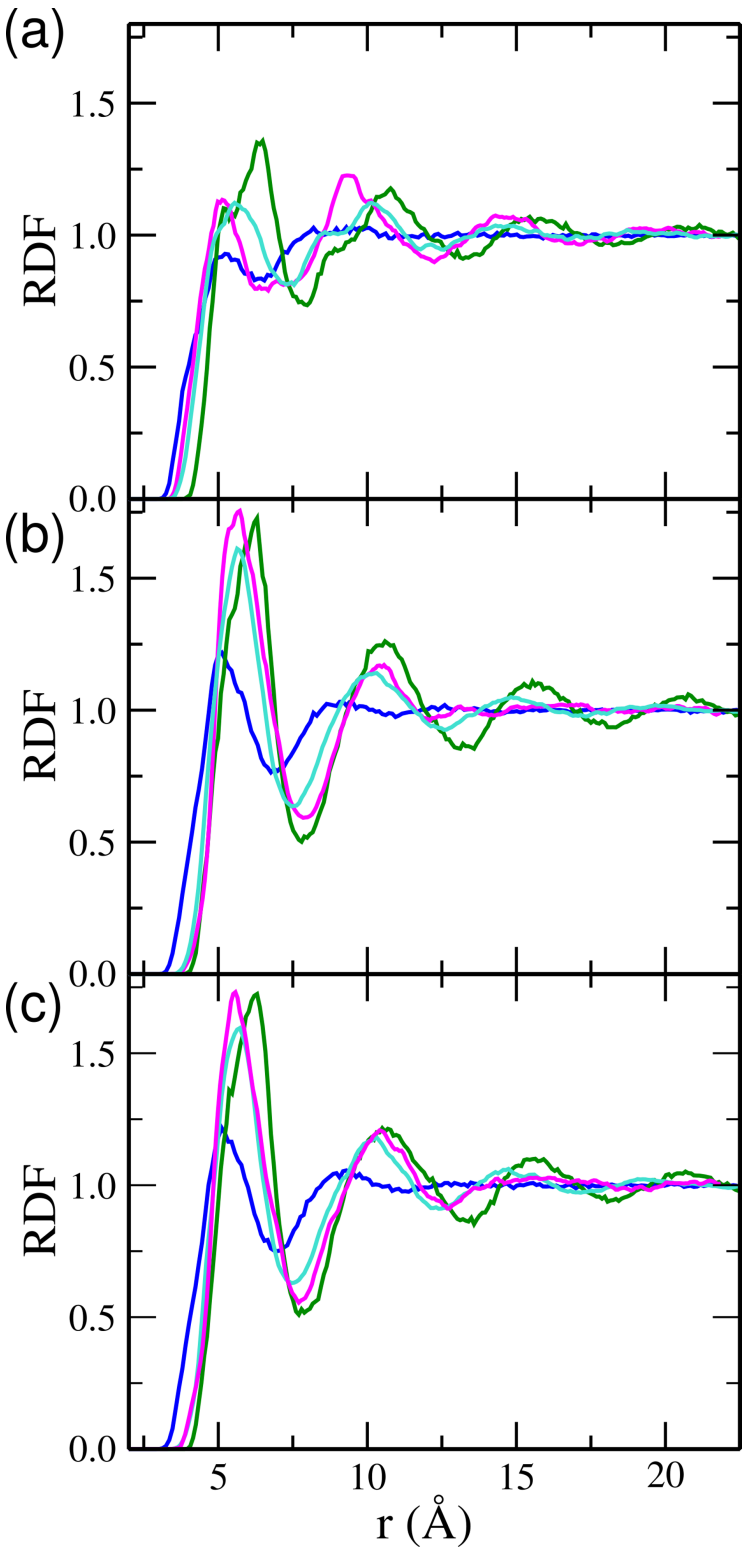


FIGURE 8.4: Radial distribution functions between DPAP ring centers (i.e., **1**, panel **(a)**; **2**, panel **(b)**; **3**, panel **(c)**) and acetonitrile (blue), *o*-xylene (magenta), tetrahydrofuran (cyan) or cyclohexane (green) center of mass.

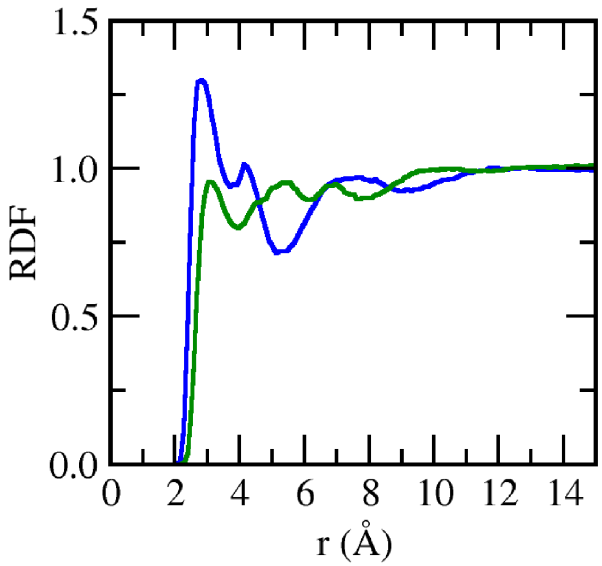


FIGURE 8.5: Radial distribution functions between DPAP (cyano) N and acetonitrile H atoms (blue) or cyclohexane H atoms (green).

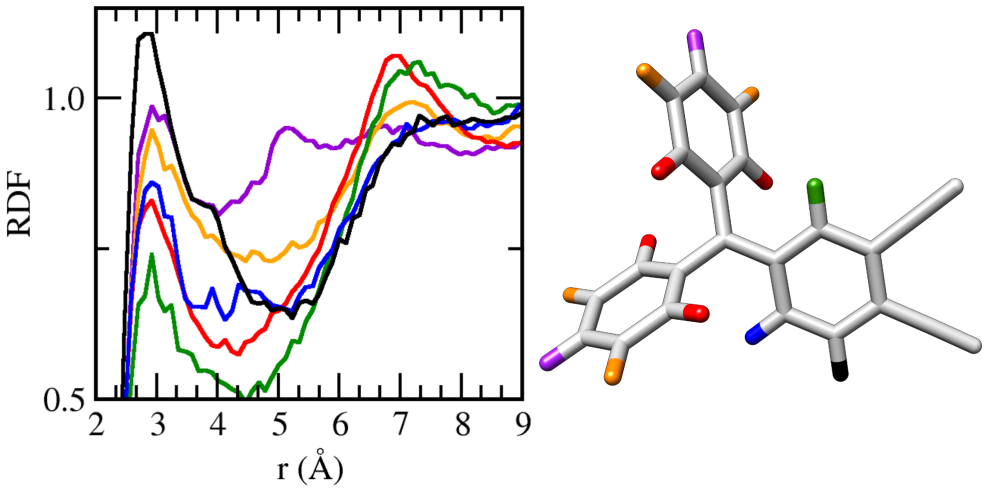


FIGURE 8.6: Radial distribution functions computed between different H atoms on DPAP and N atom on acetonitrile. Each H atom, and corresponding distribution, is highlighted following the color scheme depicted in the right panel.

Indeed, cyano substituents act as hydrogen bonding acceptors, thus facilitating the interaction with the ring H atom.

8.3.3 DPAP in polymeric matrix and lipid bilayer

DPAP was docked into the cavity of a pre-equilibrated atactic PMMA matrix following a similar protocol of a previous study [241]. The DPAP structure was embedded into the polymer matrix avoiding close contacts with the surrounding polymer chains and the system was minimized via the steepest descent algorithm until a energy threshold of 0.5 kJ/mol was reached. Within the simulated time interval, DPAP remained trapped into the PMMA cavity, displaying no translational motion and little reorientational freedom (*vide infra*).

The main structural features characterizing the rotor within the PMMA matrix were described by evaluating the rdf between DPAP cyano N atoms and PMMA methyl H atoms, and between DPAP H atoms and PMMA carbonyl O atoms. Inspection of Figure 8.7 reveals a noticeable structure arising from the interaction of the polymer hydrogen atoms and the nitrogen atoms of the cyano substituents, with an average distance evaluated from the last 2 ns of the MD simulation of 3.15 Å.

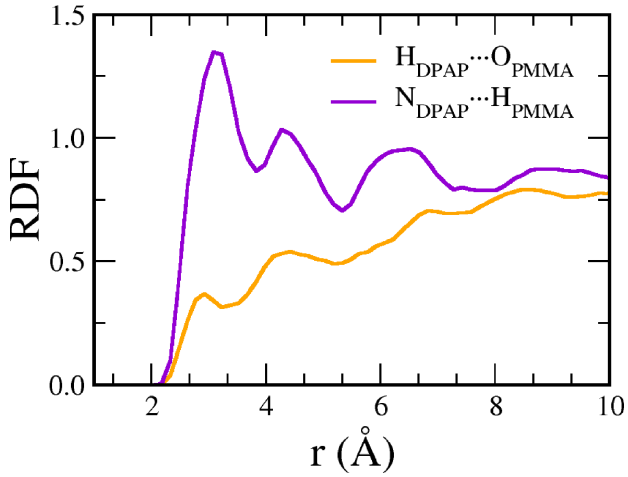


FIGURE 8.7: Radial distribution functions computed between DPAP H atoms and PMMA (carbonyl) O atoms (orange) and DPAP (cyano) N and PMMA (methyl) H atoms (violet).

Other distinguishable rdf peaks, due to the tangled structure of the polymer bundle, are located at 4.4, 6.4 and 9.2 Å. By comparison, a labile interaction takes place between the polymer carbonyl O and DPAP aromatic H atoms with a first low peak at about 3 Å.

On the other hand, in the study of the DPAP/DOPC membrane system, an initial configuration was obtained from a previous equilibrated membrane configuration containing one cholesterol unit, in which cholesterol was replaced by DPAP. In the starting configuration, DPAP was embedded within the lipid membrane at 2 Å depth from the hydrophilic interface. To characterize DPAP molecular dynamics in the hydrated DOPC bilayer, the lateral displacement and in-depth distance of the rotor from the lipid polar surface, as a function of time, were monitored (Figure 8.8). The lateral displacement (along the XY plane) shows a slow diffusive regime ≈15 Å in 100 ns of simulation. During the simulation, DPAP sequesters itself well within the dense membrane up to 13.43 Å from the lipid surface, which also accounts for its hindered rotations (see below). The average immersion distance was 7.05 Å

which is less than half of the average bilayer thickness (38 Å). This result is consistent with the experimental evidence issuing from a previous bioimaging study [246] in which DPAP was shown to localize preferentially within the cell membrane and into lipid vescicles according to its hydrophobic nature. A complete permeation was not observed within the present sampling, since it would require much longer timescales.

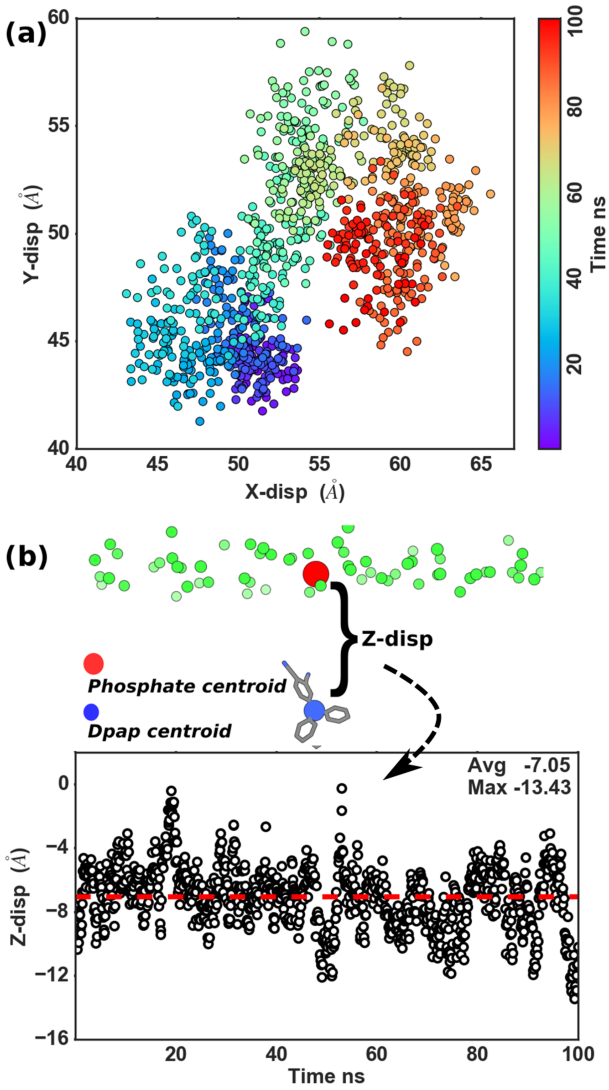


FIGURE 8.8: Displacement in XY plane **(a)** and Z dimension **(b)** of DPAP within the membrane. The DPAP centroid (blue circle in **b**) was taken as reference for the analysis. The DPAP permeates up to 13.43 Å from the lipid surface, defined considering the phosphate centroids (red circle in **b**).

To better analyze DPAP orientation within the lipid membrane during the MD simulation, the time evolution of the angle formed between the normal to the lipid bilayer and the normal to the DPAP NC1C1′C1″ moiety (see Figure 8.9) was evaluated.

As can be noted in Figure 8.9, most of the time DPAP molecular plane was oriented at an angle around 30° (or equivalently 150°) with respect to the lipid bilayer orthogonal axis, undergoing only two rotational transitions during the 100 ns time interval. The present analysis provided evidence of the anisotropic effect of the lipid environment, which, coupled to the intrinsic viscosity of the lipid alkyl chains, has severely hindered DPAP rotational dynamics. This result is not surprising since it appears

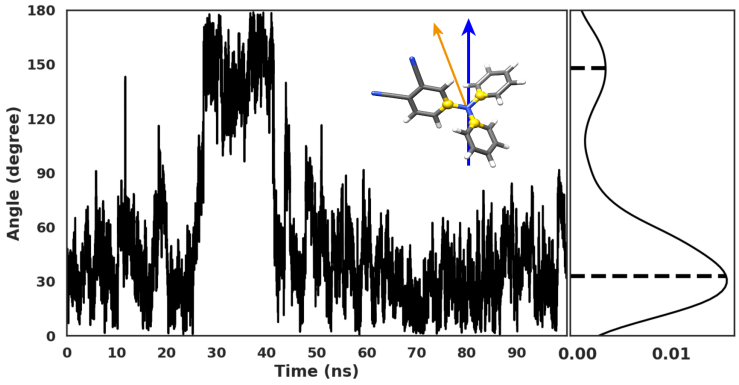


FIGURE 8.9: Evolution of the angle between the normal to the plane of lipid bilayer hydrophilic interface (vector in blue) and the axis perpendicular (vector in orange) to the plane defined by the three *ipso* carbon atoms (in yellow in insert) during 100 ns of simulation.

TABLE 8.2: Dynamical and spectroscopic properties of DPAP.

Environment	**D (10^{-5} cm^2 s^{-1})**	τ_{rot} **(ps)**	τ_{rot}^{dih} **(ps)**	μ **(mPa s)**	τ_{fl} **(ns)**[231]
Acetonitrile	3.68 ± 0.02	6.55±0.06	5.91±0.07	0.344[248]	2.61
THF	1.54 ±0.03	13.5±0.9	12.8±0.2	0.47[249]	12.9
o-Xylene	0.18 ± 0.02	84±7	80±5	0.81[250]	12.5
Cyclohexane	0.17 ± 0.03	62±4	62±4	0.887[251]	9.16
DOPC bilayer	0.024 ± 0.006	173±67	151±66	134-195[252]	14.3
PMMA	-	28600±8100	11000±7000	-[a]	12.3

[a] The PMMA viscosity has not been reported since the corresponding value strictly depends on the chain lengths.[253] In general, polymer viscosity is determined upon dissolution in a solvent.

consistent with what it is known for other organic compounds once embedded into lipid membranes, e.g. cholesterol [247].

8.3.4 Comparison of the structural and dynamic features of DPAP in multiple environments

The chemical environments considered in this work cover a broad spectrum of complex molecular embeddings, ranging from apolar/hydrophobic to polar/high-permittivity solvents and from low-density to highly viscous environments, including also a non-homogeneous and anisotropic system (i.e. the hydrated lipid membrane). Furthermore, the structural complexity and molecular weight of the embedding molecules increases considerably from acetonitrile to PMMA. In this section, the influence of the surrounding medium on the dynamic and structural properties of DPAP was scrutinized in some detail.

First, DPAP mobility was evaluated in terms of its self-diffusion constant in all embeddings. With

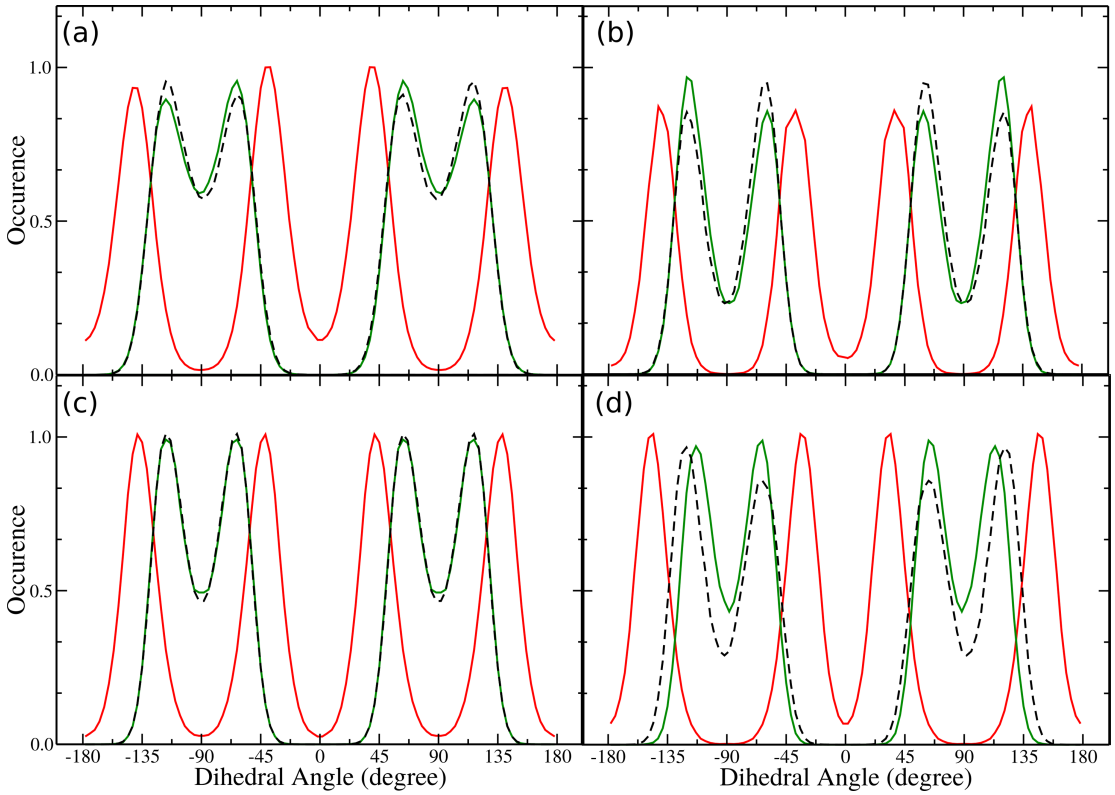


FIGURE 8.10: Dihedral distribution function of the three dihedral angles: in solid red line dihedral angle **1**, in solid green line dihedral angle **2** and in black dashed line dihedral angle **3** in **(a)** ACN, **(b)** Cyclohexane, **(c)** Hydrated 1,2-dioleoyl-*sn*- glycero-3-phosphocholine (DOPC) lipid bilayer, and **(d)** Poly(methyl methacrylate) polymeric matrix. (Note that the distributions have been symmetrized for the sake of comparison).

the exception of PMMA, noticeable translational motions were observed. The obtained diffusion constants, D, spanned several orders of magnitude (Table 8.2) as a result of medium viscosity and intermolecular interactions (as evidenced in the previous sections). In particular, a qualitative agreement with viscosity is apparent, as reported in Table 8.2. However, a quantitative relation, as predicted by the Stokes-Einstein equation [254] (which correlates the orientational diffusivity with viscosity) could not be obtained, likely due the formation of environment-specific interactions.

The distribution of the three flexible dihedral angles of DPAP (i.e. dihedral **1**, **2** and **3**, see description above), was evaluated as issuing from the MD simulations in all the considered environments. The angle distributions reflected the corresponding periodic torsional potentials based upon which the FF was derived, as shown in Figure 8.10. In all the MD simulations, DPAP selectively populated the three different torsional angles, with the highest occurrence falling within the minimum-energy configurations, while other geometries were progressively disfavoured according to the QM energy scan profile reported in Figure 8.3. In all cases except PMMA, the three aromatic rings were able to undergo a complete rotation, thus populating all minima predicted by the QM analysis. At this point, it is worth noting that DPAP rings may oscillate around their corresponding free energy minima but cannot rotate independently. Rather, structural transitions of the three dihedral angles may occur only in a concerted way. Such a coupled rotational motion is another sign of the steric hindrance among the aromatic rings.

In order to highlight the different intramolecular dynamics of DPAP in the selected environments, the time evolution and distribution of the dihedral angle **1** was evaluated and depicted along 1 ns in Figure 8.11.

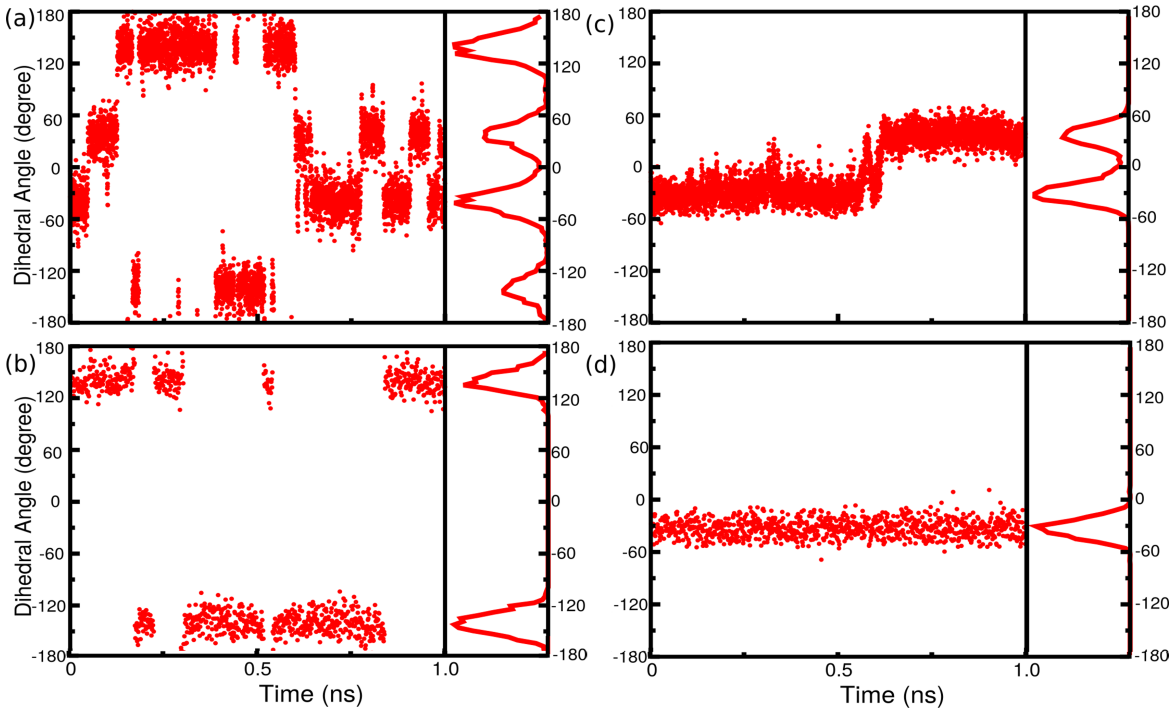


FIGURE 8.11: Time dependent dihedral distribution function for dihedral angle **1** for the first ns of simulation in ACN **(a)**, cyclohexane **(b)**, hydrated DOPC lipid bilayer **(c)** and PMMA polymeric matrix **(d)**.

In simple organic liquids (acetonitrile, cyclohexane, *o*-xylene and tetrahydrofuran) several complete rotations of the dicyano substituted ring were observed with ACN and cyclohexane yielding the fastest and lowest rotation rate, respectively. On the other hand, in the DOPC bilayer only a small amplitude oscillation (from -30° to +30°) was noticed and in the PMMA matrix no transitions were observed. A more quantitative analysis along the entire MD trajectories was carried out by evaluating the time autocorrelation function (ACF) of ring **1** torsional angle (Figure 8.12). Overall, the same trend discussed above was observed. The corresponding rotational correlation times ($\tau_{\mathrm{rot}}^{\mathrm{dih}}$) suggested the slowest dynamics to occur in the PMMA matrix (see results in Table 8.2). Note that the transition frequency of DPAP torsional angle is to be considered a direct consequence of the interaction with the environment, in addition to the intrinsic viscosity of the medium.

Furthermore, to characterize DPAP molecular rotations, the ACF of the axis perpendicular to the NC1C1′C1″ group (i.e., three *ipso* carbon atoms marked in yellow in inset of Figure 8.13b) was evaluated as a function of time. The ACF in acetonitrile decays more rapidly than in all other systems, showing a rotational correlation time (τ_{rot}) of about 6.55 ps (Table 8.2), followed by THF (τ_{rot} = 13.5 ps), cyclohexane (τ_{rot} = 62 ps) and *o*-xylene (τ_{rot} = 84 ps). This result appears to be consistent with the interactions between investigated solvent molecules and DPAP rings highlighted in Figure 8.4. The rotational motions appeared highly retarded in the more viscous systems as the membrane or the polymeric embedding. Here, it is worth noting that DPAP maintained a higher degree of rotational

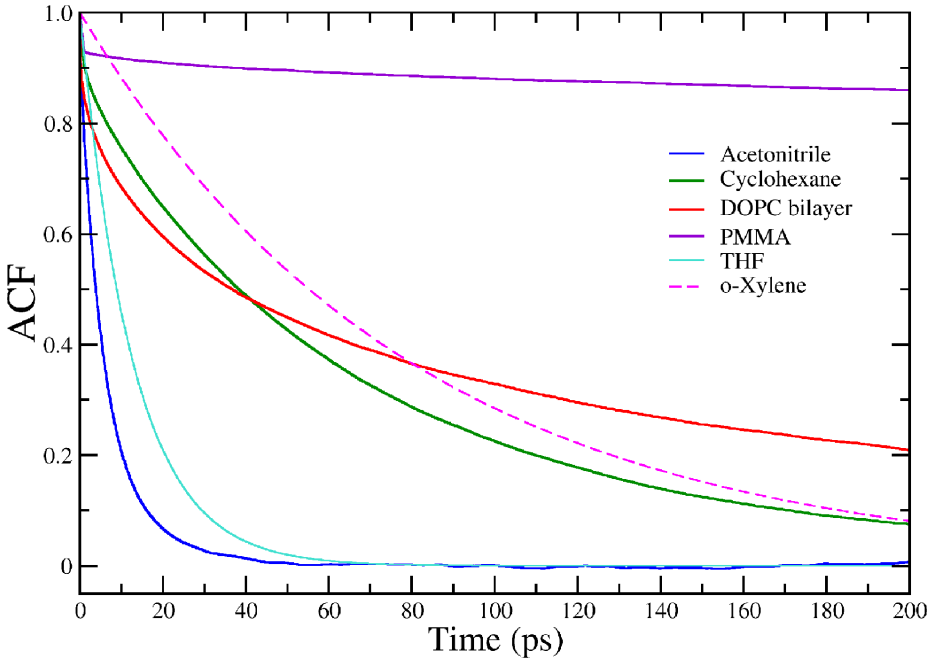


FIGURE 8.12: Autocorrelation function of the vector perpendicular to the ring 1 (Figure 8.2). Data related to the first 100 ps of simulation are reported.

freedom into the lipid bilayer than into the polymeric matrix (both intramolecular ring rotations and whole-molecule rotations, see Table 8.2). However, rotational (and translational) motions of DPAP in PMMA could not be sampled satisfactorily within the simulated time interval, owing to the high inertial mass of the polymeric matrix. Therefore, the estimated correlation times (i.e., $\tau_{\text{rot}}^{\text{dih}}$ and τ_{rot}) have to be considered qualitatively more than quantitatively.

Previously, DPAP fluorescence quantum yield was observed to follow approximately a Förster-Hoffmann relation [255] when plotted against viscosity in a set of low-dielectric and increasingly viscous solvents [231]. Here, excluding the DPAP/PMMA and DPAP/THF systems, a good degree of correlation between DPAP rotational dynamics (τ_{rot}) and the observed fluorescence lifetime (τ_{fl}) was noted in the same environments under investigation, as reported in Table 8.2 and depicted in Figure 8.14. This finding is consistent with the view that non-radiative processes are disfavored in more viscous and less interacting embeddings, thus enhancing the half-life time of the corresponding excited states. As a matter of fact, entrapping fluorescent dyes into nanoparticles, such as silica-based particles, is a fruitful strategy exploited in imaging applications to achieve extended emission lifetimes. In this study, the role of the environment in modulating DPAP fluorescence signal emerged as connected with the capability to hinder or enhance DPAP rotational dynamics.

8.3.5 Optical absorption spectra of DPAP

Theoretical absorption spectra of DPAP in all environments were evaluated by carrying out spectroscopic calculations, at the CAM-B3LYP/SNSD level of theory, on 200 molecular configurations extracted from MD trajectories. Spectra were generated from the convolution of vertical excitation energy calculations on the first few excited states using an empirical half-width-half-maximum (HWHM) parameter to better match experiments. The four liquids (for which experimental absorption were

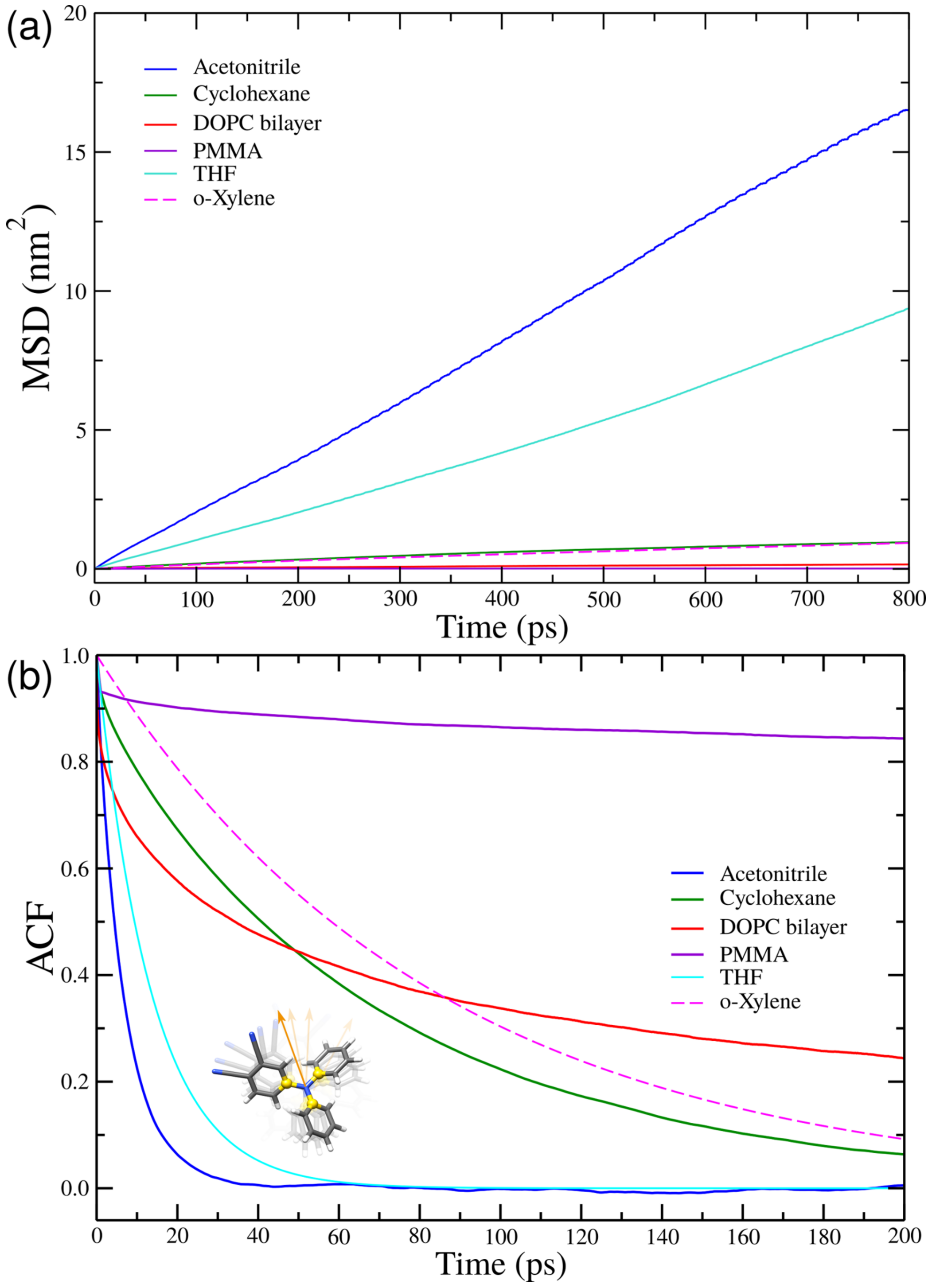


FIGURE 8.13: **(a)** Mean square displacement (nm^2) of DPAP in acetonitrile (blue line), tetrahydrofuran (cyan line), *o*-xylene (magenta dashed line), cyclohexane (green line), DOPC bilayer (red line) and PMMA (violet line). Data related to the first 800 ps of simulation are reported. **(b)** Autocorrelation function of the vector perpendicular to the plane defined by the three *ipso* carbon atoms (in yellow), shown in the insert. Data related to the first 200 ps of simulation are reported.

available) were modeled with the PCM, with a HWHM of 0.2 or 0.1 eV. The simulated absorption spectra are depicted in Figure 8.15 and compared to the experimental counterparts taken from Ref. [231] (spectra are normalized for comparison). The main, broad peak, located at 329 nm and 327 nm in acetonitrile and cyclohexane, respectively, corresponds essentially to the $S_1 \leftarrow S_0$ and $S_2 \leftarrow S_0$ transitions. The same transition accounts for the main peaks at 327 nm and 324 nm in the case of *o*-xylene and THF. In all solvents, the main band is well reproduced by the theoretical calculations: in acetonitrile, the theoretical spectrum is slightly redshifted by about 8 nm, while in cyclohexane deviation is only

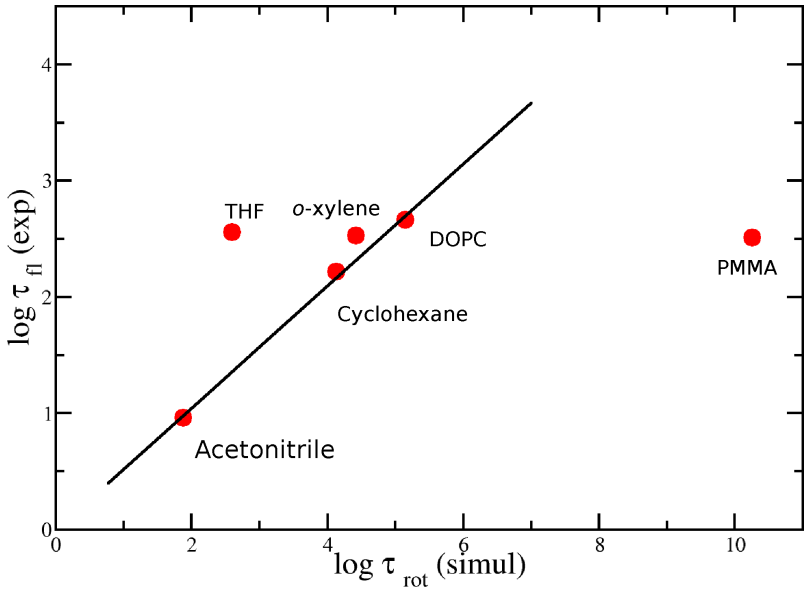


FIGURE 8.14: Correlation between DPAP rotational correlation time (ps) and fluorescence lifetimes (ns) in the six considered environments.

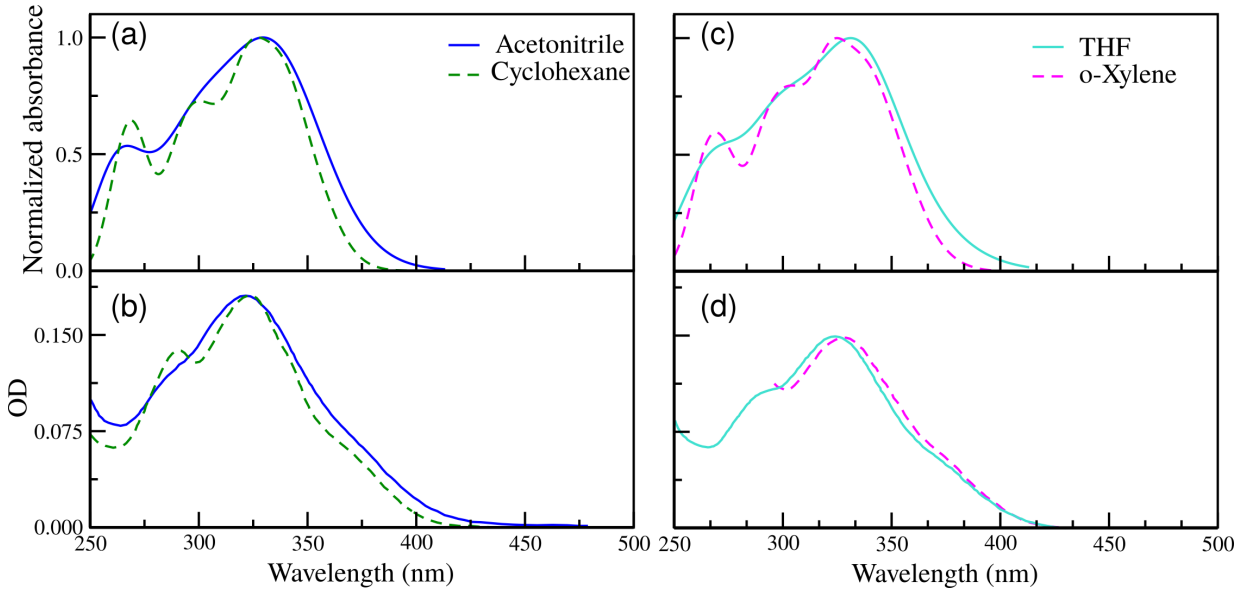


FIGURE 8.15: Comparison between theoretical (a, c) and experimental (b, d) absorption spectra of DPAP in acetonitrile (continuous blue line), cyclohexane (dashed green line) *o*-xylene (dashed magenta line) and tetrahydrofuran (continuous cyan line).

3 nm. Similar deviations of 2 nm and 7 nm are observed for the other two low-dielectric solvents. A second peak, which has been assigned to the $S_3 \leftarrow S_0$ transition, appears at smaller wavelengths. The peak maximum is well defined in cyclohexane (292 nm) and *o*-xylene, and fairly reproduced by present computations (299 nm for cyclohexane; 298 nm for *o*-xylene). The same transition is not appreciable in acetonitrile as well as in THF, and the corresponding peak is merged in the first wide band. Finally, in the four solvents another optical band at shorter wavelengths resulted from calculations did not match an experimental counterpart in the considered region of the electromagnetic spectrum. Results on the maximum absorption peaks are summarized in Table 8.3. Similitudes between experimental spectra are well preserved in the theoretical ones, confirming the DPAP solvent-independent features of this

kind of spectroscopy within the accounted wavelength region.[256]
To verify if further improvements could be achieved in modeling absorptions, the electrostatic embedding [133, 134] was considered to describe solvent molecules during vertical energies computations. Within such paradigm, solvent residues with at least one atom within 20 Å from DPAP have been replaced by the respective atomic charge (according to the corresponding force field) during the QM computation. The EE wavelengths at maximum absorption for acetonitrile and cyclohexane were located at 341 and 327 nm, respectively. This result is in agreement with the PCM case for cyclohexane. Also for acetonitrile, anyway, the peak is in line with the previous investigation, since it is within the corresponding statistical error reported in Table 8.3. Despite the absorption spectra of DPAP recorded in various solvents were found not to change substantially, in contrast to the emission ones, their theoretical reproduction was not expected to be trivial, since excitation energies are quite sensitive to the dihedral angle **1** both in polar and apolar solvents.[231] Hence, it was largely important to base spectroscopic calculations on reliable molecular structures which also provide a representative sampling of the FMR configurational space.

Environment	**Absorption peak (nm)**	
	Theory	Experiment[231]
Acetonitrile	329±20	321
THF	331±18	324
o-Xylene	325±17	327
Cyclohexane	327±15	324
DPOC	323±12	-
PMMA	321±13	-

TABLE 8.3: Maximum absorption peak wavelength (nm)

Absorption spectra were also computed in the PMMA polymeric matrix and DOPC membrane (HWHM value of 0.2 eV), even if no experimental counterparts were available in this case (Figure 8.16). These two environments are characterized by low dielectric permittivity: PMMA is about 2.8-3 and a similar value can be predicted for the membrane, since DPAP is embedded in the lipophilic layer throughout the MD simulation (Figure 8.8). In both cases, the dielectric medium was simulated by adopting the butanoic acid (ϵ= 2.9931) within the PCM formalism. Note that PCM was already shown to well reproduce the electrostatic effects of PMMA environment in a previous work [241]. Maximum absorption wavelength was found at about 321 nm for both systems, in line with the calculations for the solvents reported above. The other two peaks take place at about 298 and 268 nm and appear as

shoulders of the main first absorption band. However, in the membrane case, the decay at lower wavelength values is faster, and the $S_3 \leftarrow S_0$ transition peak is better defined. It is worth noting that such subtle differences are only due to the DPAP configurations extracted from MD simulations, since the description of the environment (i.e. the dielectric continuum) is the same.

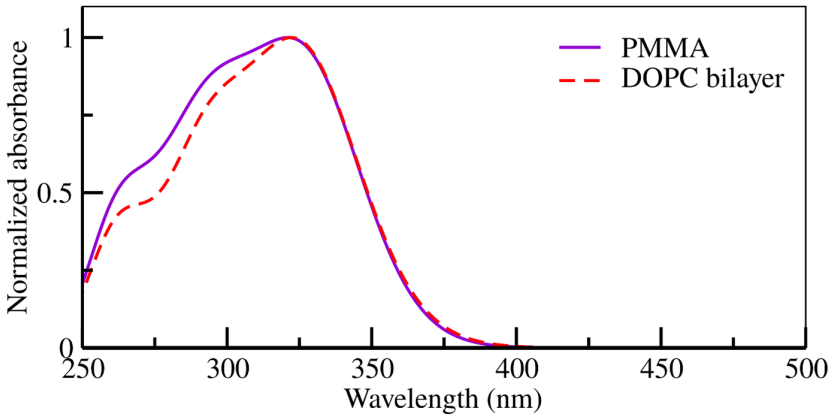


FIGURE 8.16: Absorption spectra of DPAP in membrane (red) and in PMMA (violet line) using PCM as environment model.

8.4 Conclusions

A molecular model of DPAP, a recently proposed FMR fruitfully employed in various imaging and detection applications, has been presented and investigated through extensive MD simulations in different environments. In each case, DPAP has shown peculiar structural and dynamical features, as well as specific interactions with the environment. In the lipid membrane, DPAP has displayed an anisotropic molecular orientation and a jump-diffusion rotational dynamics. This seems consistent with the observed alignment of cholesterol and other organic compounds, once embedded into lipid bilayers. The reliability of the present DPAP model was further tested by simulating the absorption optical spectra including the effect of the embedding, obtaining an overall good agreement with available spectroscopic data.

The subtle effects issuing from medium viscosity and environment-specific interactions on the dynamic properties of the rotor have been especially highlighted and discussed in view of DPAP molecular mobility. Self-diffusion constant and rotational correlation time of the present FMR are, indeed, strongly modulated by the environment to the extent that they may vary by several orders of magnitude. This is of particular interest since the rotational relaxation of fluorescent dyes can be related to various optical properties, such as fluorescence lifetime, emission intensity, fluorescence depolarization, etc., as well as to properties of the micro-environment, such as the viscosity. Here, a simple quantitative relation between the viscosity of the embedding medium and the FMR dynamics could not be obtained. This result comes as no surprise since the existence of specific dye-environment interactions, the non-continuous rotational dynamics and the shape of the rotor are all factors contributing to appreciable deviations from the ideal Stokes-Einstein-Debye model, as already noted in previous studies (see, e.g., Ref. [257]).

Nevertheless, the present study has unraveled a possible correlation between DPAP rotational correlation time and its fluorescence lifetime in all considered environments (except PMMA and THF): the more retarded the rotational relaxation, the longer the emission lifetime. Intriguingly, this finding may provide a molecular insight on the effective control exerted by the FMR rotational dynamics towards the competition between radiative and non-radiative decay processes, which ultimately modulates the dye fluorescence signal. Accordingly, specific intermolecular interactions more suited to interfere with molecular rotations seem to be the key factor modulating the emission response of the present FMR. While this point would necessitate further investigation and validation, for example by modeling the dye excited state and emission dynamics, if confirmed it could be used to gather detailed molecular dynamics information on the dye, as well as on its interactions with the environment, through standard spectroscopic techniques.

Chapter 9

Validation of the LRR-DE procedure

LRR-DE is s novel statistical procedure which has been developed in order to optimize the parameters of non-bonded force fields of metal ions in soft matter. The criterion for the optimization is the minimization of the deviations from *ab initio* forces and energies calculated for model systems. The method exploits the combination of the linear ridge regression and the cross-validation techniques with the differential evolution algorithm. Wide freedom in the choice of the functional form of the force fields is allowed since both linear and non-linear parameters can be optimized. The method has been described in Section 3.2. In this chapter, the methodology has been validated using the force field parameterization of five metal ions (Zn^{2+}, Ni^{2+}, Mg^{2+}, Ca^{2+}, and Na^{+}) in water. To this end, LRR-DE has been combined with a novel sampling procedure aimed at maximize the dissimilarity of the instances included in the training set and the coverage of the conformational space of the investigated molecular system.

9.1 Background

Metal ions are omnipresent in proteins, where they serve fundamental functional roles, including structural and catalytic functions [258]. In the complex scenario of force field development, the classical modeling of metal ions is still maybe regarded as a stand-alone issue. Three different approaches are commonly used for the description of metal ionic species within MD simulations, which are summarized in the following.

1. The *non-bonded model* [259] treats the metal ion as a simple sphere, characterized by its atomic charge and the vdW parameters. The interactions with surrounding ligands is described by means of classical Coulomb and LJ potentials.

2. The *bonded model* [260, 261] models the metal ion as covalently bound to its coordinating atoms, thus to explicitly considers chemical bonds, angles and diheadrals with the first coordination sphere atoms.

3. The *dummy-atoms model* [262] connects additional sites with the central ion, at the specific geometry to be attained. Dummy atoms act as virtual sites, ad they mimic somehow valence electrons. No real bonding with the coordination sphere is considered, and, as in the non-bonded model case, the interactions with other atoms is modeled thorugh Coulomb plus LJ potentials.

Among them, the non-bonded model is the most commonly used, because of its easy implementation in MM software; moreover, it allows in principle to describe changes of ligands at the metal coordination center. Ions parameters for non-bonded models in biomolecular FFs have been historically developed in water solution, as done by Stote and Karplus [263] and, more recently, Jensen and Jorgensen [264]. Subsequently, the optimized parameters are transferred within metalloprotein catalitic sites [265]. Major challenges are related to the proper treatment of non-negligible QM effects, which are hard to include within classical descriptions [266]. In this context, the limits of the simple electrostatic plus Lennard-Jones (LJ) model emerge, and a transition to more flexible, multi-parameters potential (e.g., by means of polarization) becomes necessary [262, 267, 268]. Therefore, the availability of techniques capable to optimize FFs of any functional form can be crucial.

9.1.1 Current status of parameterization procedures of non-bonded metal ions force fields

The generation of metal ions FFs has been extensively discussed by Li and Merz in a recent review [269]. Methods for parameterizing non-bonded FFs of metal ions are based primarily on the reproduction of experimental thermodynamic and structural quantities. In the pioneering work of Aqvist [270], the parameterization of 12-6 LJ potentials of a set of ions was performed using the hydration free energies as a reference and the FEP method [271] to calculate the MM estimates. Babu and Lim [272] used the same method exploiting the relative HFEs with respect to the Cd^{2+} value to generate the FFs of 24 divalent metal ions. Joung and Cheatham [273] parametrized 12-6 Lennard-Jones potentials of monovalent ions employing as reference HFE, crystal lattice energies and crystal lattice constants. Li, Merz and co-workers [259, 274] developed the parameters of over fifty metal ions reproducing HFE, ion-oxygen distances (IOD) and coordination numbers (CN). In general, the methods that employ experimental references suffer of two difficulties: i) the availability of data is usually limited to a reduced number of solvents, sometimes only to water. ii) the exploration of the parameter space, usually performed through a grid search, requires a MD or Monte Carlo simulation for each trial solution, making the process inefficient and applicable only to simple functional forms. Both problems can be solved using QM data as target values in the fitting. However, only a very small number of methods based on QM references has been developed. The more significant ones are the works of Floris et al. [275] and Wu at al. [268]. The method proposed by Floris et al. optimizes the ion-water potential reproducing *ab initio* energies calculated for $[M(H_2O)_n]^{q+}$, where the number of the explicit water molecules (n) is one or two, and the rest of the solvent is described by the PCM. Therefore, the performances of the method are dependent on the quality of the solvent description. Moreover, the application of PCM precludes the possibility to parametrize the FFs in heterogeneous environments. These limitations have been overcome in the recent application of the force-matching method by Wu et al. to parametrize the short-long effective functions (SLEF) model in protein environment. In the Wu et al. methodology a squared deviations cost function defined with respect to a sample of QM/MM references is minimized using a local optimizer. The procedure here presented maintains the desirable properties of the Wu et al. approach and introduces further advances in order to generate transferable non-bonded pairwise force fields to model metal ions interactions in metalloproteins. In fact, the multi-objective optimization allows a tight control on the performances of the model. The application of a regularized cost function and the tuning of the hyperparameters through the leave-one-out cross validation protect

from overfitting. The combination of algebraic and metaheuristic optimization ensures the efficient detection of the global minimum of the cost function in the parameter space.

9.2 GRASP sampling

In order to build the training set for the fitting, a set of representative configurations of the environment of the metal must be selected. The sampling must be performed carefully to obtain a general and balanced model maintaining the size of the training set such as the computational cost of the technique is affordable.

Assuming to have a criterion for deciding if a set of configurations is better than another, the selection of the best training set would be a NP-hard problem of combinatorial optimization. Therefore, the sampling issue can be separated in three distinct problems: i) generate the candidate configurations to be included in the training set ii) propose a model to determine the fitness of a training set iii) identify an approximated procedure to solve the combinatorial problem of maximization of the fitness. In particular, the application of the greedy randomized adaptive search procedure [276] (GRASP) is exploited for the third step. For this reason, the whole procedure is called GRASP sampling, which funcitoning is sketched in Figure 9.1.

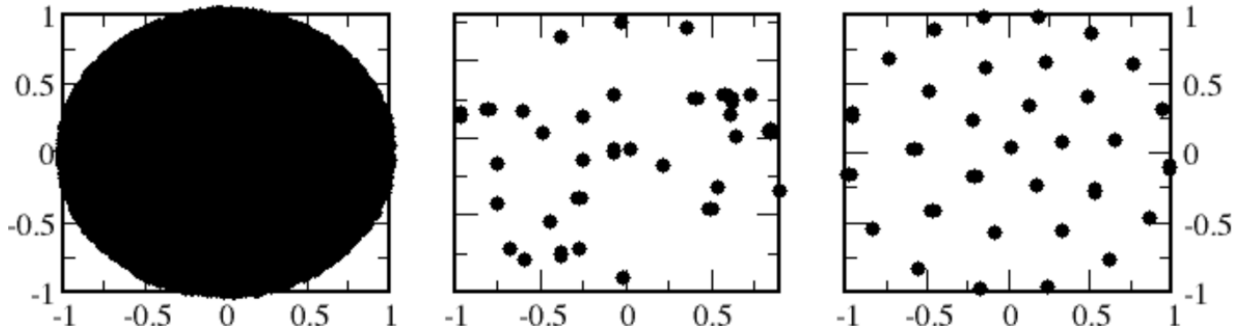


FIGURE 9.1: An illsutrative example of the use of GRASP to maximize the dissimilarity of the instances of a chosen dataset. The GRASP selection (right panel) from a pool (left) is compared to a random selection (central panel).

9.2.1 Generation of the candidate configurations

In this work, the generations of the candidate configurations has been performed with the parallel tempering [277–279] technique using a pre-existing FF. It is suitable for this purpose because it explores a large portion of the free-energy landscape of a molecular system. The configurations are drawn for each replica at regular intervals of extension comparable to the time scale of the significant events of the system considered and proceeds up to obtain a sufficiently large pool of candidates (tens of thousands). Alternative approaches aimed to the generation of the candidate configurations can be considered, such as the extraction from a MD trajectory or a metadynamics sampling, as well as from pre-existent databases. Moreover, procedures which iteratively improve the FF through subsequent sampling and fitting steps can be used in order to correctly reproduce the physics of the investigated system, as proposed by previous works [13, 120].

9.2.2 The metal-centric dissimilarity score

Since the aim of the present work is the optimization of the FF of a specific atom, the evaluation of the fitness of a possible training set should be focused on the environment of that atom. More specifically, the configurations included in the training set should maximize the representativeness of the situations in proximity of the metal ion. In order to achieve this goal, a dissimilarity score of a set of configurations focused on the neighborhood of the metal is proposed.

As descriptor of the l-th configuration, the vector $\boldsymbol{d}_l$ is used, whose elements are the Euclidean distances between the metal and all other atoms. Each i-th component of the vector $\boldsymbol{d}_l$ is transformed applying a Gaussian kernel as follows

$$k_{li} = \exp\left[-\frac{d_{li}^2}{2\sigma^2}\right] \tag{9.1}$$

where the parameter σ is a measure of the distance from the metal which identifies the most significant region to sample.

The Euclidean distance between the l-th and j-th configurations in the k-space is

$$\delta_{lj} = \|\boldsymbol{k}_l - \boldsymbol{k}_j\| \tag{9.2}$$

δ_{lj} is invariant with respect to the translations or rotations of the system, moreover it satisfies the coincidence axiom ($\delta_{lj} = 0$ if and only if $l \equiv j$) and the symmetry condition ($\delta_{lj} = \delta_{jl}$). As consequence of the transformation of the equation 9.1, $k_{li} - k_{ji}$ is amplified with respect to $d_{li} - d_{ji}$ where the first derivative of the Gaussian function is larger. This fact occurs in correspondence of $d = \sigma$. For $d \to 0$ and $d \gg \sigma$, conversely the differences in the k-space vanish, therefore in those zones the information is compressed.

The metal-centric dissimilarity score of the set $\{\boldsymbol{k}\}$, constituted by N_{TS} configurations selected among N_{PT} candidates, is defined as the mean value of the N_{loc} distances from the N_{loc} nearest configurations weighed for an exponential factor, $2^{N_{loc}-j}$:

$$DS(\{\boldsymbol{k}\}, N_{TS}, N_{PT}) = \frac{\frac{1}{N_{TS}}\sum_l^{N_{TS}} \frac{1}{N_{loc}}\sum_j^{N_{loc}} 2^{N_{loc}-j}\delta_{lj}}{\sum_j^{N_{loc}} 2^{N_{loc}-j}} \tag{9.3}$$

In this formula, the j-th configuration is the j-th nearest one to the l-th configuration. The exponential weight has two roles: i) it assigns more importance to the nearest configuration in the score ii) it makes the score near independent from the N_{loc} value. The dissimilarity score is related to the inverse of the local density of the points in the k-space. If the configurations are selected in order to maximize the score, for a given value of N_{TS}, a stratified sampling of the environment of the metal is obtained. That is, the distributions of the distances between the metal ion and the atoms included in the spherical shell centered in $d = \sigma$ are flatter than the distributions generated by the parallel tempering simulations. The maximization of the coverage offered by a stratified sampling increases the probability to perform the fitting in interpolation regime instead of extrapolation regime.

The dimensionality reduction and permutational symmetry

The number of the atoms of the system of interest is generally in the order of thousands, nevertheless the dimensionality of the $\boldsymbol{k}$ vector can be reduced without loss of information because only few components have a value different from zero. This measure assures that the calculation of the dissimilarity score is affordable in the combinatorial optimization step.

To assure the permutational invariance of the dissimilarity score, a further arrangement of the $\boldsymbol{k}$ vector is necessary, namely all the equivalent atoms are placed in ordered positions with respect to the distance from the metal ion. In this context, two atoms of the same element are considered permutationally invariant if they can exchange their positions through a move compatible with the dynamics of the system.

9.2.3 The combinatorial optimization of the training set

The maximization of the dissimilarity score (equation 9.3) with respect to the candidate configurations is performed exploiting an adapted form of the greedy randomized adaptive search procedure [276] (GRASP). GRASP is a combinatorial optimization method consisting in two main phases repeated iteratively: construction and local search. In the first one a greedy randomized adaptive strategy is employed to build a feasible solution that is refined by a subsequent local search. The operation is repeated saving the best solution found.

The construction phase starts selecting randomly one configuration from the candidate set. In order to extract the second configuration, the Euclidean distances from the first one are calculated for all the remaining candidates, and a configuration is selected randomly from the subset of the instances further than the 99-th percentile. In analogous way the following N_{loc} - 2 configurations are chosen. From this point on the construction phase proceeds iteratively by cumulative addition of a new element that maximizes the dissimilarity score. The new element is selected from a list of ordered candidates, $\{s_i\}$, composed evaluating the dissimilarity score of the set $\{\{S_{i-1}\}+s_i\}$, where $\{S_{i-1}\}$ is the partial solution composed by $i-1$ elements. The selection of the subsequent element is performed according to an exponential probability distribution that attributes the maximum value to the first element of the ordered list.

In the local search phase, the solution is modified one element at a time and the trial solution is accepted if the dissimilarity score increases. This operation is performed using three different strategies: random substitution, proposal of new elements sorted by distance from the centroids of the partial solution and local refinement.

9.3 Computational details

Classical simulations were performed under periodic boundary conditions, using GROMACS 4.6.5 [280].

The systems were composed by one metal ion, surrounded by 2178 water molecules, leading to a cubic box of size 40 Å. The rigid TIP3P model [281] has been used to describe the water molecules. Fastest degrees of freedom were constrained with the LINCS algorithm [93]. In the sampling step, the metal ions have been modeled using parameters developed by Åqvist [270] (Mg^{2+}, Na^{+}), Merz

[282] (Zn^{2+}) and Li [259] (Ni^{2+}, Ca^{2+}). Each system was minimized using the steepest descent algorithm implemented in GROMACS using a convergence threshold on the root-mean-square forces of 1 $kJ{\cdot}mol^{-1}{\cdot}cm^{-1}$. Systems were slowly heated from 0 to 298 K in NVT ensemble for 1 ns, and then equilibrated in NPT conditions for 1 ns to reach uniform density. The final structure for each system was considered as the starting point for the parallel tempering simulations. A number of 25 replicas was employed, covering a temperature range of 100 K, from 298 to 398 K. Temperature distribution of single replicas in the chosen range was established so as to attempt an exchange rate of 0.25 [283]. Each replica was equilibrated in the NPT ensemble for 500 ps, using the stochastic velocity rescale algorithm [95]. The time step was set equal to 2 fs. Production runs were conducted in the NVT ensemble for 5 ns, for a total simulation time of (5ns × 25 replicas) 125 ns. Electrostatic interactions were described through the PME method, whereas van der Waals interactions were considered applying a cutoff of 10 Å. The FFs generated with the LRR-DE procedure were tested by initially equilibrating the systems in the NPT ensemble for 500 ps. After that, production run time was set to 5 ns in the NVT ensemble. Cutoff values of 19 Å for both LJ and real-space PME were used. Non-standard models were tested using tabulated potentials.

Radial distribution functions were computed using standard tools available in GROMACS software on the last 1.5 ns of simulation. Free energy of hydration values were computed using the Bennett acceptance ratio (BAR) [284] implemented in GROMACS. A number of 21 windows was used. Corresponding λ values were chosen ranging from 0 to 1, in steps of 0.05. Each window was run for 500 ps. The first 100 ps was considered to equilibrate the systems, and therefore not included in the final HFE computation. All the QM calculations were performed using the Guassian 09 package [97] at the B3LYP/cc-pVDZ level. Singlet spin-state has been considered for all the ions except Ni^{2+}, for which the triplet has been found to be more stable than the singlet spin state in the QM model systems. All the parameters optimized with the LRR-DE are provided in Appendix A.

9.4 Validation

The algorithm has been validated by applying it to the parameterization of the force fields of five metal ions in water: Zn^{2+}, Ni^{2+}, Mg^{2+}, Ca^{2+}, and Na^{+}. The TIP3P model [281] has been used to describe the water molecules. The whole adopted protocol is summarized in Figure 9.2.

Particular attention has been addressed in the case of the zinc ion, which has been used as reference for the calibration of the method. QM/MM calculations on large spherical clusters and pure QM calculations on clusters of lower size have been initially considered as references. Although the first type of calculation reproduces more closely the actual situation in which common FFs are used, it involves two disadvantages: i) a bias is introduced by the MM part of the calculation, and ii) the number of atoms involved is very large, increasing the computational cost of the fitting. Therefore, in this work, pure QM calculations on small clusters have been chosen as references, verifying that the size of the model systems was sufficiently large through a systematic study with variable number of water molecules (see next section). As level of theory, the B3LYP functional in combination with the cc-pVDZ basis set has been selected.

A large basin of candidate configurations has been generated using parallel tempering and a set of 160 elements has been extracted through the GRASP sampling procedure. The appropriate size of the

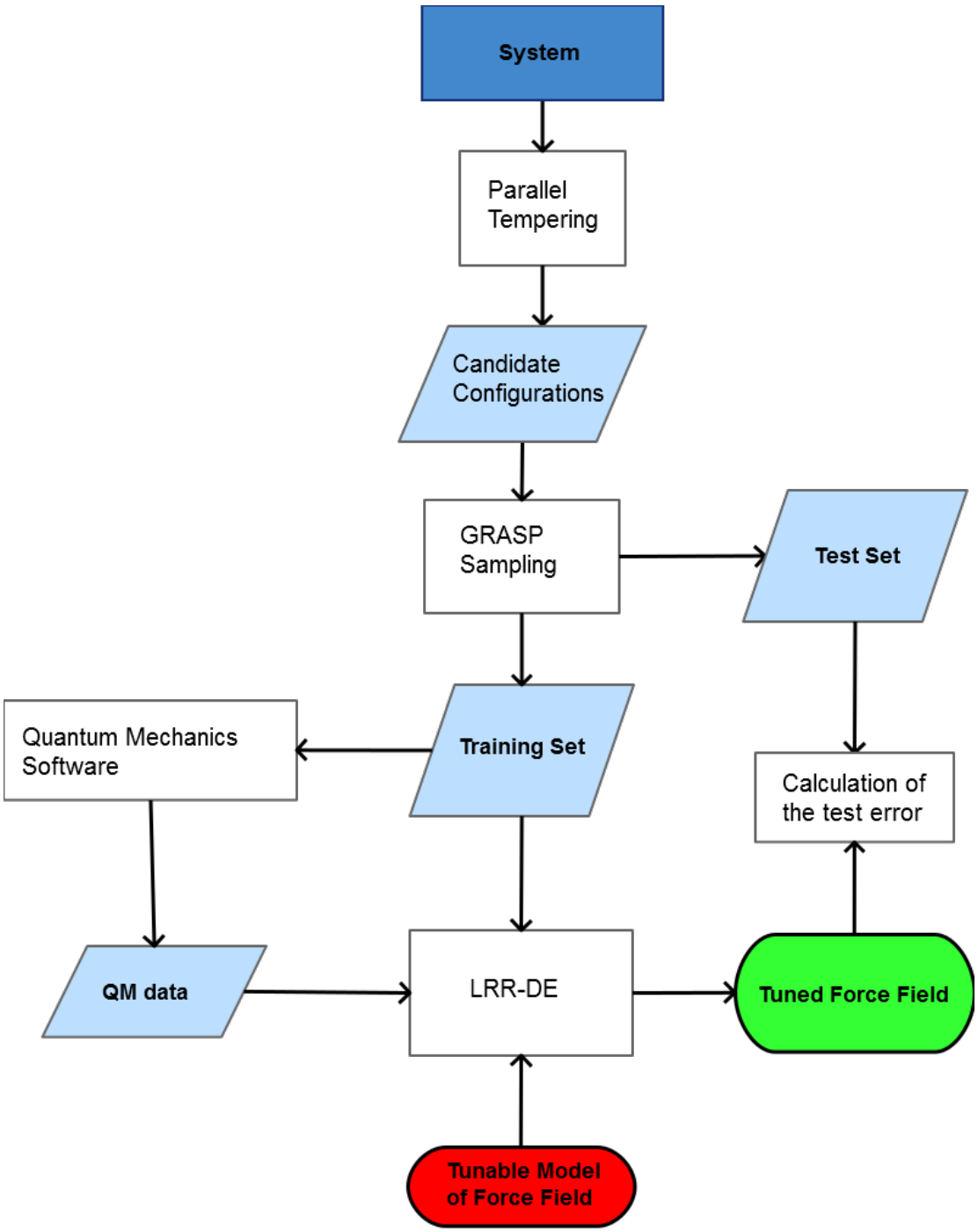


FIGURE 9.2: High level flowchart of the proposed algorithm.

training set (N_{train}) has been identified by performing a statistical convergence test, applying the LRR-DE method to the fitting of the forces and the energies for the cluster $[Zn(H_2O)_{128}]^{2+}$ case. Starting from a training set of 8 elements and incrementing the size progressively, the three linear parameters of the 12-6-1 FF have been optimized with respect to the QM references. For each size of the training set, 256 independent training sets have been generated selecting randomly N_{train} configurations without repetition, from the total 160 available and using the remaining 160-N_{train} configurations as test set. The averages of the resulting mean squared errors are shown in Figure 9.3. The graphs confirm that the LOOCV error (red lines) constitutes a better estimate of the test error (green lines) with respect to the training set errors (blue line). In fact, the LOOCV errors and test errors converge to the same value when the size of the training set is greater than 60. The values of the test errors decrease rapidly when the training set size is small, and converge to a constant value when N_{train} is greater than 100 instances. Therefore, in all the following fittings training sets of 120 elements have been employed, so as to provide sufficient generality to the obtained models.

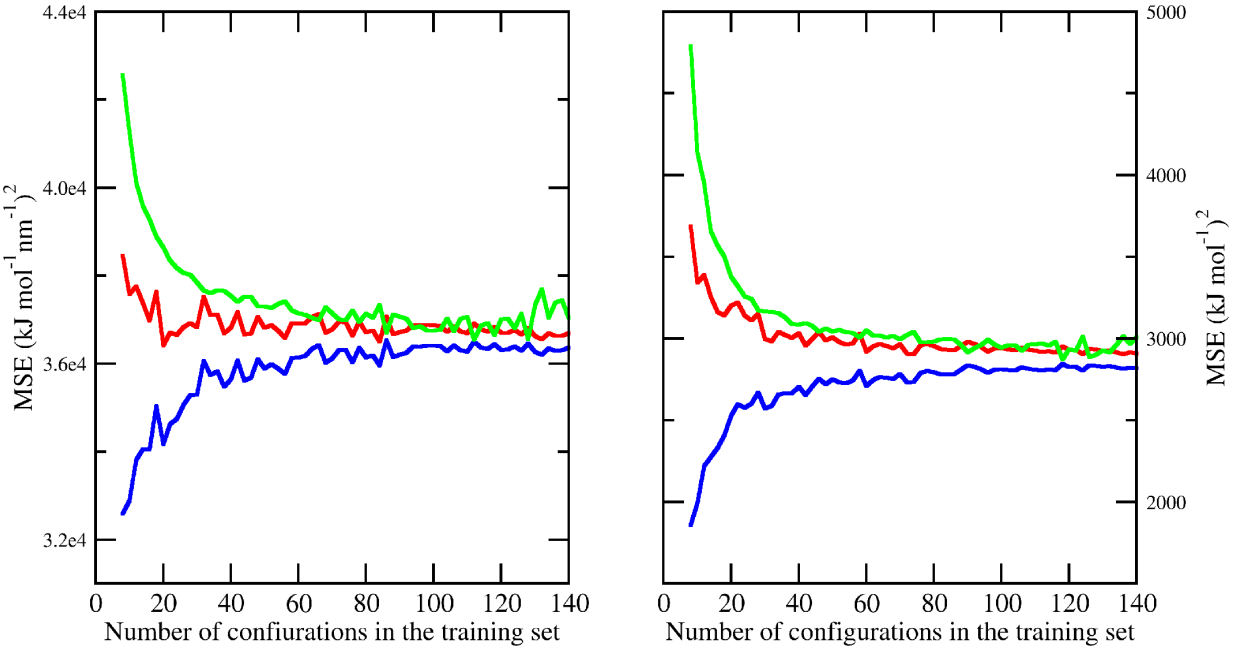


FIGURE 9.3: Mean of 256 tests of the MSE for the training set (blue line), leave-one-out cross validation (red line), and test set (green line), increasing progressively the size of the training set. The model 12-6-1 has been used to perform the fitting. In each test the elements of the training set are selected randomly from the 160 configurations and the remaining are used as test set.

9.4.1 Systematic comparative study of binary potentials

In order to calibrate the methodology, the optimization of the parameters of twelve binary pairwise models (see Table 9.2) has been performed using as reference systems the $[Zn(H_2O)_n]^{2+}$ clusters with n equal to 6, 16, 32, 64, and 128, water molecules. The models consist of a repulsive term, activated only for the zinc-oxygen interaction, and the Coulomb potential. The clusters are built extracting the n closest water molecules to the zinc ion for each of the 160 sampled configurations (see Figure 9.4). The results have been compared to AMBER99 [270], Li *et al.* [259] (Li, hereafter) and Hartree-Fock (HF) estimates. In standard conditions of temperature and pressure, the coordination number of the

zinc ion in bulk water is six [285] and the mean number of molecules included in the first and second spheres of coordination is about 30 [286]. Therefore, the smallest cluster considered corresponds to the extraction of the first sphere of coordination, the $[Zn(H_2O)_{16}]^{2+}$ cluster is representative of the first shell of coordination and part of the second one, and the larger clusters include all the molecules of the first two coordination spheres and beyond. For the largest clusters, $[Zn(H_2O)_{128}]^{2+}$, the average distance of the furthest oxygen is 9.6 Å. The parameterization of the force fields has been executed in single-objective mode with the forces on the zinc ion as output references for each cluster, in two-objective mode, contemplating simultaneously forces on the zinc ion and energies of the same cluster, and in four-objective mode considering all the possible couplings of forces and energies for two types of clusters. In all cases, the resulting force fields have been tested on the QM forces on zinc ion, the energies (y_l in equation 3.26) and the forces on the nearest oxygen and hydrogen atoms for all the clusters, so as to evaluate their capacity in predicting quantities unused in the fitting. As a significant case, the 12s-1 FF data are shown in Appendix B (Tables B.1, B.2, B.3, B.4) and analyzed below.

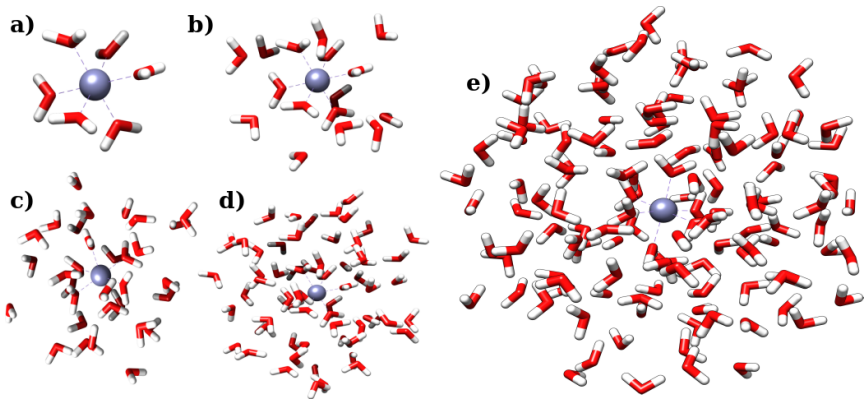


FIGURE 9.4: Representative structures of the extracted clusters containing 6 (a), 16 (b), 32 (c), 64 (d) and 128 (e) water molecules.

Table B.1 reports the MAEs obtained training the 12b-1 force field with the single-objective fitting. The LRR-DE procedure allows to obtain the optimal reproduction of the fitted quantities for the considered model, achieving errors about four times lower than the standard force fields AMBER99 and Li. This can be better appreciated by observing Figure 9.5, where the comparison between the QM forces and those predicted by the model for a test set of 40 instances not included in the training set is shown. As consequence of the Newton's third law, also the errors of the forces on the oxygen atoms are drastically reduced with respect to the AMBER estimates. In addition, from the data it emerges that the forces in clusters of different size than the one used in the training can be reproduced with good accuracy. On the other hand, the force fields trained on the forces produce high errors in the prediction of the energies, indicating that these models are not sufficiently general. In order to overcome this drawback, the transition to a multi-objective fitting is necessary. Table B.2 reports the results for the two-objective fittings, considering simultaneously the forces and the energies for a given cluster. The inclusion of the energies in the output references allows to obtain a remarkable reduction of the MAEs for this quantity at the price of a moderate yet acceptable increase in error on the forces. Even more general force fields can be generated if the fitting is performed on data of clusters of two different sizes (Tables B.3 and B.4). In fact, the MAEs resulting from the four-objective fitting are considerably lower than the errors produced by AMBER99 and Li for both the forces and the energies. From the tables

Test set	14-1	12-1	10-1	8-1	6-1	14b-1	12b-1	10b-1	8b-1	6b-1	Exp-1	Exp2-1
E($6H_2O$)	62.93	47.92	31.47	46.79	150.95	49.85	51.59	53.88	57.08	61.65	38.96	34.45
E($16H_2O$)	48.96	47.97	49.55	61.90	105.31	62.61	63.37	64.39	65.81	67.84	57.77	55.63
E($32H_2O$)	47.78	46.17	45.04	46.24	57.78	46.00	46.23	46.56	47.03	47.80	44.67	44.12
E($64H_2O$)	50.95	49.08	46.69	44.29	43.61	43.77	43.80	43.85	43.94	44.12	43.61	43.59
E($128H_2O$)	50.35	48.36	46.45	45.72	53.72	45.28	45.43	45.63	45.95	46.51	44.52	44.26
F(Zn, $6H_2O$)	327.64	274.35	223.32	185.69	198.66	177.43	178.21	179.42	181.62	186.26	174.29	174.07
F(Zn, $16H_2O$)	327.38	282.93	241.33	208.37	212.90	203.20	203.36	203.76	204.68	206.98	203.66	204.60
F(Zn, $32H_2O$)	321.05	274.73	231.99	200.83	206.63	195.50	195.77	196.27	197.44	199.95	196.09	197.21
F(Zn, $64H_2O$)	322.22	275.33	231.20	198.14	202.86	193.49	193.69	194.08	194.84	197.11	194.02	195.35
F(Zn, $128H_2O$)	321.93	274.64	230.37	198.69	201.52	193.50	193.63	193.96	194.89	197.42	194.27	195.60
F(6O, $6H_2O$)	1326.13	1231.41	1117.21	984.41	876.50	938.52	942.65	948.91	959.33	979.15	920.07	916.11
F(6O, $16H_2O$)	1404.42	1309.95	1196.19	1064.08	956.60	1018.41	1022.52	1028.76	1039.13	1058.86	1000.02	996.08
F(6O, $32H_2O$)	1429.27	1334.73	1220.76	1088.18	980.36	1042.32	1046.45	1052.71	1063.12	1082.94	1023.88	1019.93
F(6O, $64H_2O$)	1422.20	1327.51	1213.41	1080.74	972.86	1034.88	1039.01	1045.27	1055.68	1075.50	1016.42	1012.47
F(6O, $128H_2O$)	1420.76	1326.11	1211.99	1079.30	971.52	1033.51	1037.63	1043.89	1054.28	1074.06	1015.08	1011.13
F(12H, $6H_2O$)	351.11	362.23	381.43	419.41	512.29	421.22	422.85	425.12	428.55	434.59	411.19	406.56
F(12H, $16H_2O$)	398.68	409.88	429.15	466.91	558.59	468.70	470.31	472.56	475.96	481.95	458.77	454.18
F(12H, $32H_2O$)	403.08	414.87	435.00	474.14	568.32	475.99	477.65	479.96	483.47	489.63	465.74	461.00
F(12H, $64H_2O$)	402.56	414.09	433.86	472.39	565.41	474.23	475.87	478.17	481.63	487.71	464.08	459.40
F(12H, $128H_2O$)	402.28	413.73	433.40	471.74	564.75	473.57	475.21	477.50	480.97	487.05	463.45	458.80

TABLE 9.1: Mean absolute errors (kJ/mol for the energies, E, and kJ/(mol nm) for the forces, F) in the prediction of the quantities indicated in the first column, for the model indicated in the first row trained using the four-objective fitting 4-o($32H_2O$/$128H_2O$).

Model	Analytical expression	Number of non-linear parameters
14-1	$\frac{C_1^O}{d^{14}} + \frac{C_2}{d}$	0
12-1	$\frac{C_1^O}{d^{12}} + \frac{C_2}{d}$	0
10-1	$\frac{C_1^O}{d^{10}} + \frac{C_2}{d}$	0
8-1	$\frac{C_1^O}{d^{8}} + \frac{C_2}{d}$	0
6-1	$\frac{C_1^O}{d^{6}} + \frac{C_2}{d}$	0
14b-1	$\frac{C_1^O}{(d+\theta_O)^{14}} + \frac{C_2}{d}$	1
12b-1	$\frac{C_1^O}{(d+\theta_O)^{12}} + \frac{C_2}{d}$	1
10b-1	$\frac{C_1^O}{(d+\theta_O)^{10}} + \frac{C_2}{d}$	1
8b-1	$\frac{C_1^O}{(d+\theta_O)^{8}} + \frac{C_2}{d}$	1
6b-1	$\frac{C_1^O}{(d+\theta_O)^{6}} + \frac{C_2}{d}$	1
Exp-1	$C_1^O \exp[-\theta_O d] + \frac{C_2}{d}$	1
Exp2-1	$C_1^O \exp[-\theta_{O,1} d + \theta_{O,2} d^2] + \frac{C_2}{d}$	2

TABLE 9.2: Analytical expressions of the two-terms models tested in the systematic comparative study.

it can be noticed that the deviations of the HF forces from the B3LYP references are lower than those provided by model 12b-1, which however reproduces the energies more accurately. Therefore, as general recipe for the production of the force fields tested in the MD applications four-objective fittings have been exploited, employing as system references the clusters $[M(H_2O)_{32}]^{n+}$ and $[M(H_2O)_{128}]^{n+}$ (4-o($32H_2O/128H_2O$) hereafter). The multi-objective fittings produce larger errors in the prediction of the forces on oxygen atoms with respect to the single-objective ones. However, they remain considerably lower if compared with AMBER. For this reason the forces on the coordinating atoms have not been included as output references in the systematic study.

Table 9.1 shows the comparison of the performances of the twelve potentials considered in the systematic study for the 4-o($32H_2O/128H_2O$) fitting. Except for the case 6-1, that produces larger errors, all the potentials provide comparable results in the reproduction of the energies. Conversely, the errors for the forces are more dependent on the repulsive part of the potential employed. More specifically, the use of a repulsive term dependent on a non-linear parameter guarantees better performances. Notable is the modest result of the 12-1 model, that exploits the repulsive term of the Lennard-Jones potential, only the r^{-14} term produces a worse agreement with the QM forces.

The proposal of a novel functional form for the force fields of the metal ions goes beyond the scope of this work, however, the systematic comparative study provides the following useful indications in this regard: i) the optimal charge to reproduce the forces on the metal ion is lower than the formal charge (Table B.1) ii) in the single-objective fitting, the optimal charge for the smallest cluster is lower with respect to the clusters of larger size (Table B.1) iii) the introduction of the energies in the references has the effect to increase the value of the optimal charge over the formal charge (Tables B.2, B.3, B.4) the use of a repulsive term including a non-linear parameter is necessary to achieve good performances in the reproduction of the forces (Table 9.1). Tables B.1, B.2, B.3, and B.4 are all related to the data of the model 12b-1, however, the behavior described in the points i), ii) and iii) is common to all the tested potentials. From a physical point of view, these results can be justified by the effects of the charge

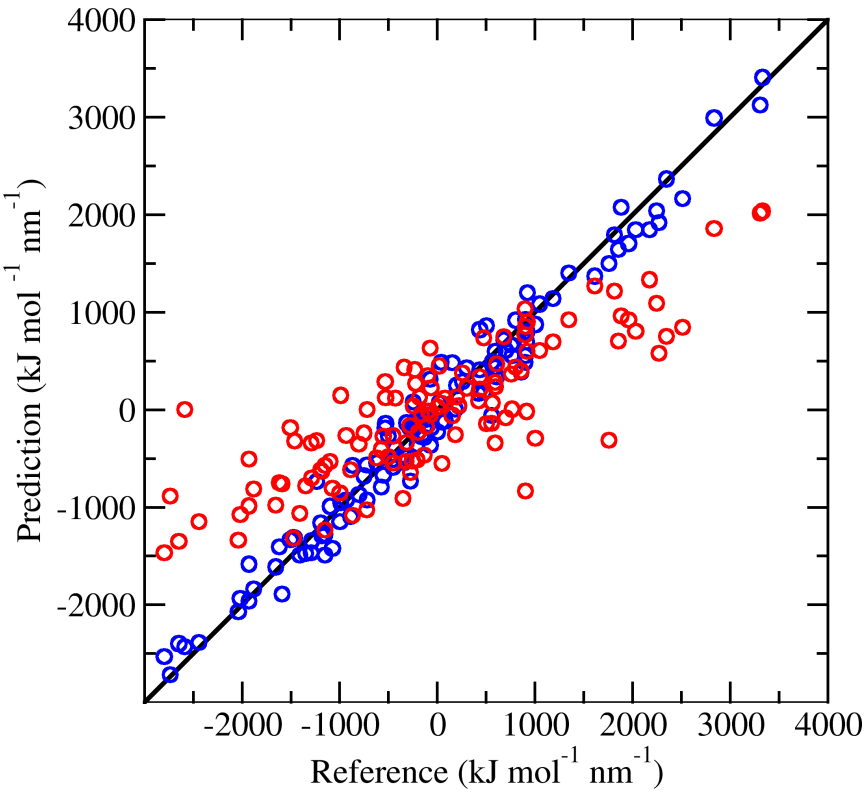


FIGURE 9.5: Graphical comparison of the prediction of 12b-1 model (blue points) and Li (red points) predictions of the forces on the zinc ion in the $[Zn(H_2O)_{128}]^{2+}$ cluster with respect to the B3LYP/cc-pVDZ reference for a test set of 40 configurations.

transfer from the ion to the coordinating water molecules and of the non-point-like structure of the ion. Both effects are expected to vanish at large distances, where a Coulomb potential generated by the formal point charge describes adequately the ion interactions. Therefore, a generic three-terms force field for metal ions that implements the indications emerged from the systematic comparative study has the form:

$$V_{tot} = V_{rep}(\theta) + \frac{q_F}{r} + \frac{f(r)_{dump} g(r)_{flex}}{r} \tag{9.4}$$

where $V_{rep}(\theta)$ is the repulsive part of the potential, q_F is the formal charge of the ion, $f(r)_{dump}$ is a dumping function which goes to zero beyond the first coordination shell of the ion, and $g(r)_{flex}$ is a function that provides the necessary flexibility to meet simultaneously the points i), ii) and iii). An explicit form for the potential of the equation 9.4 is

$$V_{tot} = \frac{C_1}{(r-\theta_1)^{12}} + \frac{q_F}{r} + C_2 \frac{e^{-\theta_2 r}(1 - e^{-\theta_3 (r-\theta_4)^2})}{r} \tag{9.5}$$

The third term of this FF, here labeled as 12b-1_F-E1Gr, recalls the physical meaning pursued by the model proposed by Wu *et al.* [268] to describe zinc charge interactions in metalloenzymes catalytic sites. In this context, the employment of such functional form has only illustrative purposes to highlight the potentiality of the LRR-DE method in the optimization of models of general functional forms. Therefore, the 12b-1_F-E1Gr FF has been tested and compared to the 12-6-1 (Lennard-Jones

combined with Coulomb potential, optimizing the charge value) and 12-6-1_F (Lennard-Jones combined with Coulomb potential, with the charge of the ion set equal to the formal charge) FFs only in terms of structural properties.

9.4.2 Parameters Optimization and Molecular Dynamics simulations

The 12-6-1, 12-6-1_F, and 12b-1-1EGr force fields have been trained using the 4-o($32H_2O$/ $128H_2O$) fitting for the cases of the Zn^{2+} ion in water, and tested in MD simulations. Properties derivable from MD have been also compared with Babu and Lim parameters [272], as well as the already cited Li parameters: performances of these two sets of parameters have been re-evaluated in this paper by applying the same computational protocol reported in the corresponding original work. Uniform values have been assigned to the weights of the objective functions, and they have been optimized according to the procedure explained in subsection 3.2.3 only if unsatisfactory errors for a QM reference have been observed. Table 9.3 reports the mean absolute errors of the three-terms models with respect to QM references of the zinc-water clusters. The larger flexibility of the 12b-1_F-E1Gr force field allows to reduce the errors, in particular in the reproduction of the forces. The accuracy of the 12-6-1 forces is significantly higher than the 12-6-1_F ones. This behavior is a consequence of the fact that the LRR-DE optimization provides a negative parameters for the r^{-6} term when the charge is a free parameter. Thus, the r^{-6} term contributes in describing the repulsion and as a consequence the QM forces are reproduced with the 6-1 quality (Table 9.1). Both 12-6-1_F and 12-6-1 FFs overcome the performance provided by Li. The same trend is observed in the results of the MD simulations. In fact, the radial distribution function (g(r)) between zinc ion and water oxygen obtained with the 12-6-1_F model presents a slightly better agreement with the EXAFS data [287] than the Li prediction (see Figure 9.6). A more consistent improvement is achieved with the 12-6-1 and 12b-1_F-E1Gr potentials, as can be appreciated in Figure 9.6. Broadly speaking, the ion-oxygen distance is well reproduced by the employed models, and their performances are better than polarizable models [288, 289], which predict lower values when compared to the experimental ones. As the flexibility of the FF is improved, the height of the peak diminishes while its width increases, thus becoming comparable with results provided by *ab initio* investigations [286, 289], as well as with the experimental EXAFS data. Moreover, it is worth noticing that the g(r) peak obtained with the 12-6-1 model (where only three parameters - i.e., Lennard-Jones parameters and the electrostatic charge - are optimized) is in line with the one predicted by even more flexible models, such as the one proposed by Chillemi *et al* [287], where 9 parameters are optimized.

As regarding the hydration free energy (HFE), the prediction provided by the 12-6-1 model is considerably more accurate than the estimates of 12-6-1_F and Li. In particular, the deviation between the experimental reference and the 12-6-1 estimate is lower than 5 kJ/mol, while for 12-6-1_F and Li is larger than 100 kJ/mol. Taking into account the parameters generated by Babu for the zinc ion, the LRR-DE 12-6-1_F model offers a slightly worse reproduction (of 0.01 Å) of the IOD value. However, an opposite behavior is observed in the reproduction of HFE, where the performance of Babu FF is poor, with a deviation from the experimental value of roughly 175 kJ/mol.

The whole protocol already used for zinc ion has been extended to the optimization of the FF models for Ni^{2+}, Mg^{2+}, Ca^{2+}, and Na^{+} ions in bulk water. In the cases of Ca^{2+} and Na^{+} forces on coordinating oxygens have been included in the training, thus performing a six-objective fitting. Such measure

Test set	12-6-1$_F$	12-6-1	12b-1$_F$-E1Gr	12b-1	AMBER99 [270]	Li [259]	HF
E(6H_2O)	49.12	77.48	70.75	51.59	140.97	116.71	153.47
E(16H_2O)	49.36	74.34	59.45	63.37	86.14	68.36	146.89
E(32H_2O)	47.73	49.84	41.48	46.23	93.35	74.38	136.58
E(64H_2O)	51.42	44.35	43.33	43.80	86.11	69.96	129.74
E(128H_2O)	49.38	47.92	43.81	45.43	95.75	77.10	113.38
F(Zn, 6H_2O)	280.69	194.72	141.48	178.21	748.83	627.34	65.13
F(Zn, 16H_2O)	287.11	210.88	184.37	203.36	724.32	604.13	71.97
F(Zn, 32H_2O)	279.35	204.67	180.99	195.77	724.55	603.82	74.96
F(Zn, 64H_2O)	280.30	201.49	180.38	193.69	726.41	605.57	73.59
F(Zn, 128H_2O)	279.54	201.66	179.97	193.63	724.29	604.22	73.81
F(6O, 6H_2O)	1242.78	1003.25	611.03	942.68	1766.17	1650.26	767.28
F(6O, 16H_2O)	1321.28	1082.85	692.46	1022.52	1842.67	1727.30	808.86
F(6O, 32H_2O)	1346.07	1107.02	714.92	1046.45	1867.99	1752.40	807.00
F(6O, 64H_2O)	1338.86	1099.57	707.54	1039.01	1861.30	1745.70	808.01
F(6O, 128H_2O)	1337.46	1098.12	706.39	1037.63	1859.89	1744.27	808.02
F(12H, 6H_2O)	356.05	451.17	356.05	422.85	356.05	356.05	527.17
F(12H, 16H_2O)	403.66	498.40	403.66	470.31	403.66	403.66	617.26
F(12H, 32H_2O)	408.32	506.54	408.32	477.65	408.32	408.32	630.00
F(12H, 64H_2O)	407.67	504.33	407.67	475.87	407.67	407.67	629.47
F(12H, 128H_2O)	407.36	503.70	407.36	475.20	407.36	407.36	629.30

TABLE 9.3: Mean absolute errors (kJ/mol for the energies and kJ/(mol nm) for the forces) in the prediction of the quantities indicated in the first column, for the model indicated in the first row.

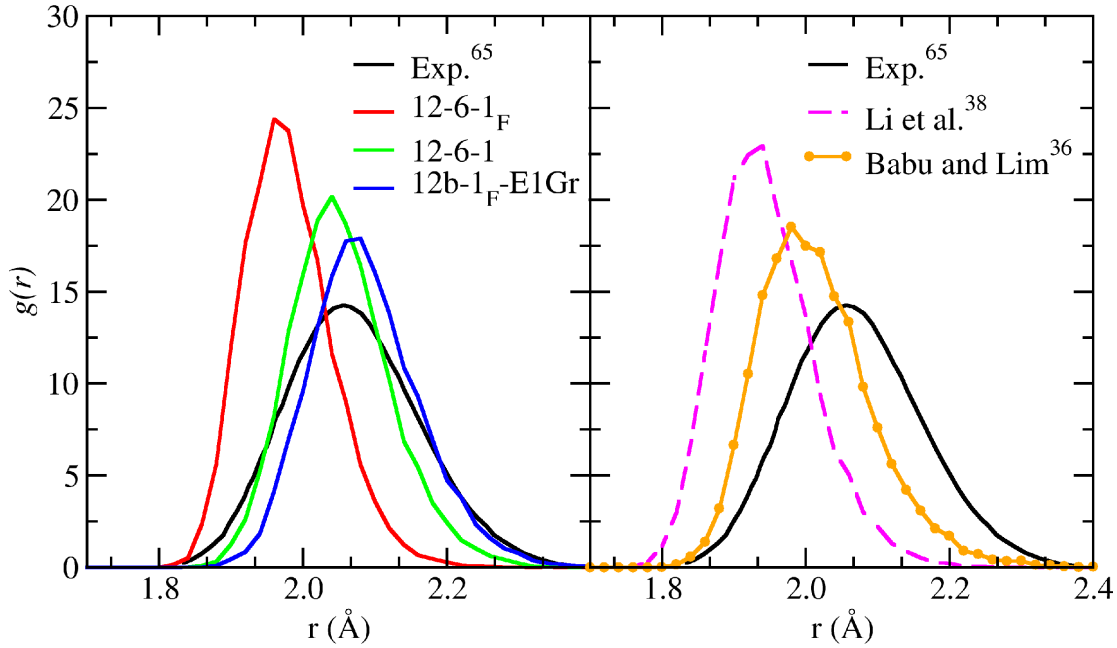


FIGURE 9.6: Radial distribution functions between zinc ion and water oxygens, using the 12-6-1$_F$, 12-6-1, 12b-1$_F$-E1Gr (left panel) and Li[259] and Babu[272] models (right). In both panels, the comparison with the experimental profile[287] is provided.

turned out to be necessary, since errors on these quantities were larger than expected. Table 9.4 collects the estimates of the position of the first peak of the radial distribution function, the HFE, and the coordination number of the ion for the five considered systems.

The optimized 12-6-1$_F$ on nickel divalent cation overcomes the performances of Li and Babu in terms of HFE estimation. Compared to the 12-6-1$_F$ force field, Babu offers a better agreement with the experimental IOD value (of 0.01 Å). As regarding the 12-6-1 model, the experimental HFE is exceeded by 102.72 kJ/mol; however the prediction of the ion-oxygen distance (IOD) is improved of 0.9 Å with respect to the Li [267] data. The experimental IOD is reproduced even better by 12b-1$_F$-E1Gr model, and the experimental ion-water oxygen radial distribution function [287] is reproduced with quite good accuracy (see Figure 9.7).

For the magnesium ion, the peak of the g(r) provided by the MD simulation with the 12-6-1$_F$ model has a deviation larger than 0.02 Å from the experimental data with respect to the Li prediction. The g(r) is correctly reproduced by Babu parameters. On the other hand, 12-6-1$_F$ gives a reduction of the error of about 35 kJ/mol and 108 kJ/mol in the HFE estimate with respect to the Li and Babu force fields, respectively. As for the zinc case, also for the magnesium ion, the 12-6-1 model produces a large improvement in the HFE prediction. Among all the considered FFs, the 12b-1$_F$-E1Gr model provides the best agreement with the experimental data for the g(r) peak position. The 12-6-1 and 12b-1$_F$-E1Gr FFs give estimates in good agreement with the state-of-art pairwise potentials [290] and polarizable models [289]. Moreover, a satisfactory comparability with AIMD and QM/MM simulations, which predict ion-oxygen distance values between 2.08 and 2.13 Å[289], is observed.

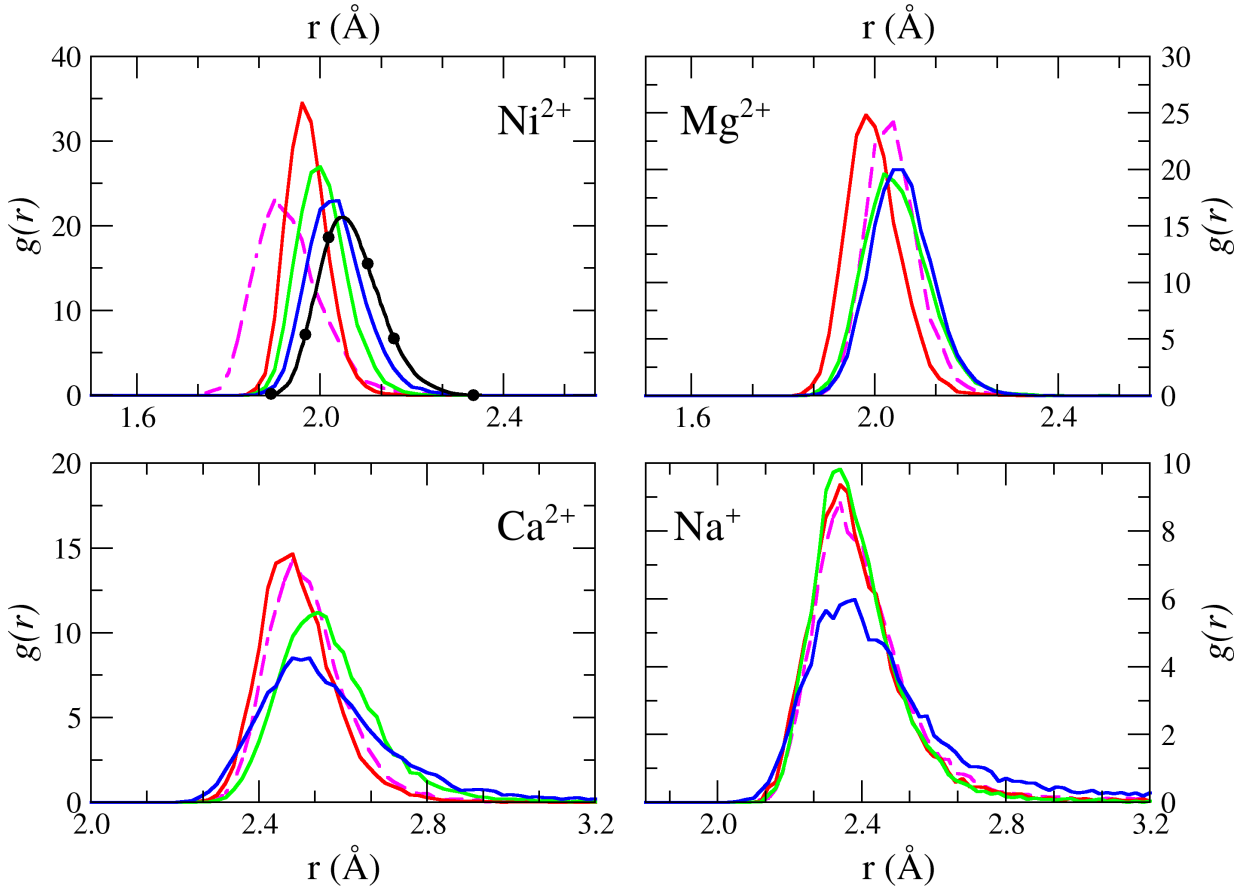


FIGURE 9.7: Radial distribution functions between water oxygens and Ni^{2+}, Mg^{2+}, Ca^{2+} and Na^{+} ions. The 12-6-1$_F$ (red line), 12-6-1 (green line) and 12b-1$_F$-E1Gr (blue line) models considered in this work are compared to Li (dashed magenta line) and experimental (black line) estimates.

The 12-6-1$_F$ model for the calcium ion trained with the LRR-DE procedure offers better performances than Li and Babu in the prediction of the peak position as well as in the HFE estimation. Again, the 12-6-1 force field produces a further increase in accuracy for the HFE, at the expense of a slightly worse result in the g(r) peak position. Also the performance of the 12b-1$_F$-E1Gr model is less satisfactory than in the previous cases. However, such measures are in line with those coming from a recent AIMD simulation which provides a metal-ion distance in the first coordination shell of 2.51±0.07 [291].

In the case of the sodium ion, LRR-DE performances are compared to the ones offered by Joung and Cheatham parameters [273]. The LRR-DE models give an excellent agreement with the experimental data in the prediction of the peak position of the radial distribution function. QM/MM simulations conducted on the hydrated ion present a certain variability in the computed metal-oxygen distance, ranging from 2.33 to 2.42 [292–294]. The HFE values calculated with the 12-6-1 and 12-6-1$_F$ models produce a decrease of accuracy of 15 kJ/mol with respect to the Cheatham estimate. However, it is worth noting that sodium Lennard-Jones parameters were specifically optimized by Joung and Cheatham in order to reproduce the HFE value [273].

All the computed g(r) profiles for Ni^{2+}, Mg^{2+}, Ca^{2+} and Na^{+} are shown in Figure 9.7. The second hydration shell (i.e., the second peak in the IOD profile) is highlighted in Figure 9.8. No notable difference can be appreciated in the peak position for the FFs tested. However, on average, the height of the peak predicted by Babu appears to be lower. In all cases the coordination numbers of the ions are consistent with those experimentally observed. Properties computed with Babu and Li parameters coincide with previous investigations [259].

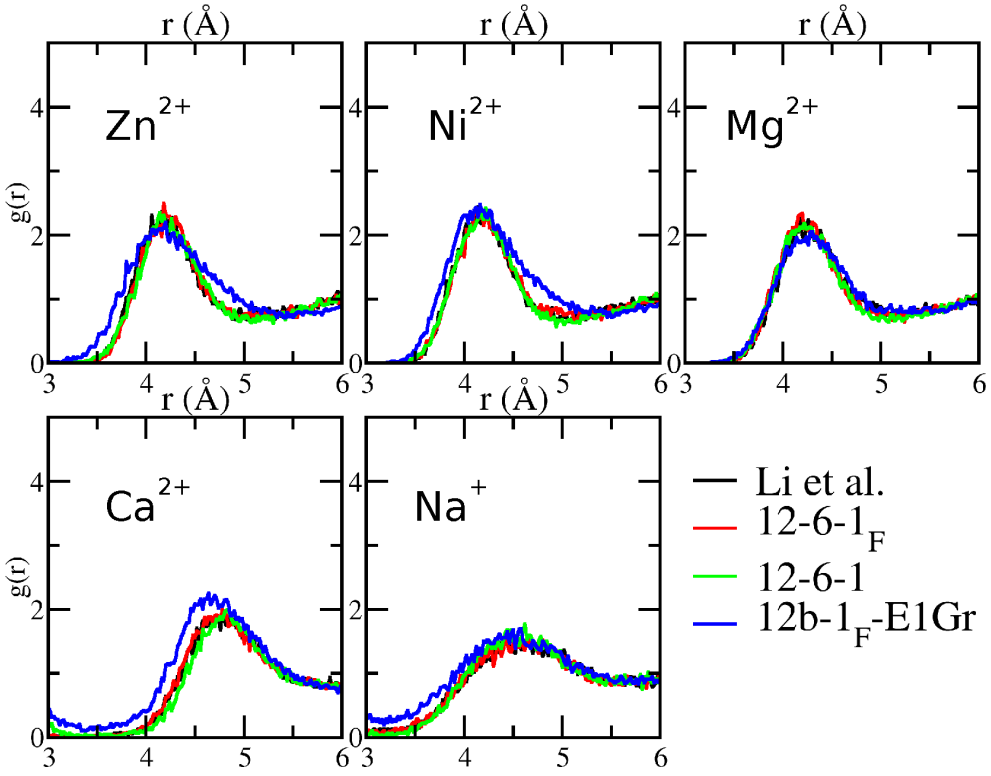


FIGURE 9.8: Second peak from radial distribution functions for the metal ion-water oxygens interaction in Molecular Dynamics simulations using the Li (black line, Zn 2+ , Ni 2+ , Mg 2+ , Ca 2+), Babu and Lim (orange, Zn 2+ , Ni 2+ , Mg 2+ , Ca 2+), Joung and Cheatham (magenta, Na +), 12-6-1 F (red), 12-6-1 (green) and 12b-1 F -E1Gr (blue) models.

All the experimental quantities explored in this investigation are reproduced with good accuracy by all the tested LRR-DE optimized models even if not directly considered during the parameter fitting procedure. Therefore, the discussed optimization method has been proved to be general, since the considered structural and thermodynamic properties are reproduced with comparable accuracy with respect to standard FFs, directly optimized in order to reproduce such quantities. Since the presented method is designed to refine only the FF of the metal ion, the optimized parameters are specific for the description within the surrounding considered environment, and the general performances are affected by the implicit approximations included in this model. The obtained FFs have not been tested in environments not considered in the training and the quality of the performances are not guaranteed in such cases. Transferable FFs can be generated using this methodology, if an appropriate sampling which considers interactions with a wide range of atom types is executed.

		IOD	MAE(IOD)	HFE	MAE(HFE)	CN
Zn^{2+}	Li *et al.*	1.93	0.16	-1849.33	105.85	6
	Babu and Lim	1.98	0.11	-1779.53	175.65	6
	12-6-1_F	1.97	0.12	-1801.39	153.79	6
	12-6-1	2.04	0.05	-1960.62	5.44	6
	12b-1_F-E1Gr	2.07	0.02			6
	Exp.	2.09 ± 0.006		-1955.18		6
Ni^{2+}	Li *et al.*	1.92	0.14	-1874.01	105.86	6
	Babu and Lim	1.98	0.08	-1800.74	179.13	6
	12-6-1_F	1.97	0.09	-2022.78	42.91	6
	12-6-1	2.01	0.05	-2082.59	102.72	6
	12b-1_F-E1Gr	2.03	0.03			6
	Exp.	2.06 ± 0.01		-1979.87		6
Mg^{2+}	Li *et al.*	2.03	0.06	-1724.23	105.85	6
	Babu and Lim	2.08	0.01	-1651.10	178.98	6
	12-6-1_F	2.01	0.08	-1759.79	70.29	6
	12-6-1	2.03	0.06	-1871.92	41.84	6
	12b-1_F-E1Gr	2.06	0.03			6
	Exp.	2.09 ± 0.004		-1830.08		6
Ca^{2+}	Li *et al.*	2.49	0.03	-1399.97	105.01	8
	Babu and Lim	2.61	0.15	-1328.19	166.79	8.3
	12-6-1_F	2.46	0.00	-1431.76	73.22	8
	12-6-1	2.53	0.07	-1551.43	46.45	8
	12b-1_F-E1Gr	2.50	0.04			8
	Exp.	2.46		-1504.98		8
Na^{+}	Joung and Cheatham	2.34	0.01	-374.89	10.05	5.87
	12-6-1_F	2.34	0.01	-389.53	24.69	5.91
	12-6-1	2.34	0.01	-389.95	25.11	5.95
	12b-1_F-E1Gr	2.35	0.00			5.72
	Exp.	2.35± 0.06		-364.84		5-6

TABLE 9.4: Simulated IOD peak (Å), free energy of hydration (HFE, kJ/mol) and coordination number (CN) values using the developed parameters for the considered force fields. Results are compared with the ones of Li *et al.*, Babu and Lim, Joung and Cheatham, and experimental data. The mean absolute errors are with respect to the experimental references.

9.5 Conclusion

Force Field	MAE(IOD peak)	MAE(HFE)
Li *et al.*	0.10	105.64
Babu and Lim	0.09	177.64
12-6-1_F	0.07	85.05
12-6-1	0.06	49.11
12b-1_F-E1Gr	0.03	

TABLE 9.5: Mean absolute deviations from the experimental references for the position of the first IOD peak (MAE(IOD peak)) and the hydration free energy (MAE(HFE)) for the considered divalent ions. MAE(IOD peak) is expressed in Å and MAE(HFE) in kJ/mol.

A novel statistical procedure (called LRR-DE) has been developed to optimize the parameters of a model so as to reproduce the general behavior of a system, given a representative data set. The fundamental feature of the method is the combination of the linear ridge regression and cross-validation techniques with the metaheuristic algorithm differential evolution. This machinery allows to optimize both linear and non-linear parameters of a model of generic functional form. The application of the regularization and the cross-validation avoids the problem of overfitting, if the training set is chosen properly. This aim is achieved applying the GRASP sampling, a combinatorial technique capable to maximize the dissimilarity of the elements of the data set. A methodology based on LRR-DE has been derived to parametrize the non-bonded force fields of metal ions, using *ab initio* quantities as references. From the calibration phase, performed on the case of the zinc ion in water, a general protocol for the fitting has been identified. This involves the use of both the forces and the energies computed on clusters of different sizes as references. The application of the multi-objective optimization is optionally activated and further reference data can be considered if unsatisfactory errors for a certain type of data have been obtained. The validation of the methodology has been performed exploiting the cases of five ions in water, for which several quantitative results of comparison, both experimental and computational, are available. The performances of the force fields trained with the LRR-DE have proved to be of comparable or better quality with respect to standard FFs, as the summary Table 9.5 attests. The possibility of the LRR-DE procedure to use as reference QM forces and energies of different systems simultaneously offers great margins of applicability to the method. In particular, the method is suitable for the optimization of FF of metal ions in heterogeneous environment, such as in the case of protein cofactors, for which experimental thermodynamic data are usually unavailable. The procedure can be applied to generate transferable FFs, if an appropriate sampling which considers interactions with a wide range of atom types is executed. Moreover, the capacity of the method to tune generic models makes it the ideal tool for optimizing FF with more sophisticated functional forms than those commonly used in MD programs.

Conclusions and perspectives

This thesis has been focused on the accurate classical modeling of chemical systems through the development of reliable force fields for MD simulations. Such task has been accomplished by the employment of several computational tools, ranging from state-of-art algorithms for atomic charges fitting to statistical learning techniques for the refininment of LJ parameters. The application in MD simulations of the proposed models allowed for the theoretical investigation of a wide range of chemical systems (biomolecules, flexible dyes and rigid organic molecules) thus to shed light on the related subtle phenomena that take place at the atomic level and correlate them with experimental observables.

The accuracy offered by literature models has been evaluated at first, and it has been proved to be adequate for the description of structural and termodynamic phenomena of biomolecular systems (Chapter 5). This purpose has been addressed by setting up a computational protocol for the simulation of the dissociation process of a ligand-receptor complex, using immersive technologies for a deeper understanding of the investigated event.

Then, it has been demonstrated that current limits of literature force fields, as the one related to trasferability from one environment to another one, can be overcomed, thus allowing to reproduce a series of chemical properties with sufficient accuracy and at reasonable cost by employing classical simulations in multiple solvents. A protocol based on first-principle computations has lead to the optimization of the description of intermolecular interactions by properly taking into account environment effects. The procedure, applied and validated on pyridine (Chapter 6), can be easily extended to other organic solvents of industrial interest (aniline, DMSO, and so on) in order to fill current gaps in available force fields of common use and improve the description of several molecules for which accurate models are still missing.

The application of computational tools for the optimization of specifically tailored ground-state force fields for flexible and large molecules has allowed for the simulation of absorption spectra and other structural and dynamical features with fair agreement with experimental data. Future works will be devoted to the generation of excited-states force fields, using the same protocol already applied for the alkynylimidazole dyes and DPAP in Chapters 7 and 8: such models can be easily employed for the analysis of excited-state dynamics as well as for the computation of the fluorescence, since the importance of emission properties in these kind of molecular probes.

At final, a novel procedure for the development of non-bonded models of metal ions in water solution has lead to the obtaining of new parameters with comparable or even better quality respect to the literature ones (Chapter 9). The method will be applied in the near future to more complex and heterogeneous systems, such as metallic catalytic sites and organometallic systems. Moreover, the fitting procedure can be extended to optimize the intramolecular part of a force field.

In general, the satisfactory comparison with available experimental data proves the quality of the

proposed approaches, as well as helps in the understanding of the complex mechanisms that govern the investigated phenomena. The protocols developed in this work can be further extended to the in silico investigation and prediction of the technological properties of other molecular species of industrial interest, thus accelerating their set-up and fabrication processes. On the other hand, the presented computational strategies and models may be integrated with other theoretical tools within virtual screening campaigns, aimed at a more effective search of novel chemical entities with the desired physico-chemical traits.

Part III

Appendices

Appendix A

Force field parameters

A.1 AP dyes

This section of Appendix A reports structures, atom indices and Joyce force field parameters of the dyes studied in Chapter 7. LJ parameters are transferred from OPLS-AA. Charges have been computed using CM5 at the B3LYP/SNSD level of theory. Non-bonded interactions of the 1-4 type have been excluded.

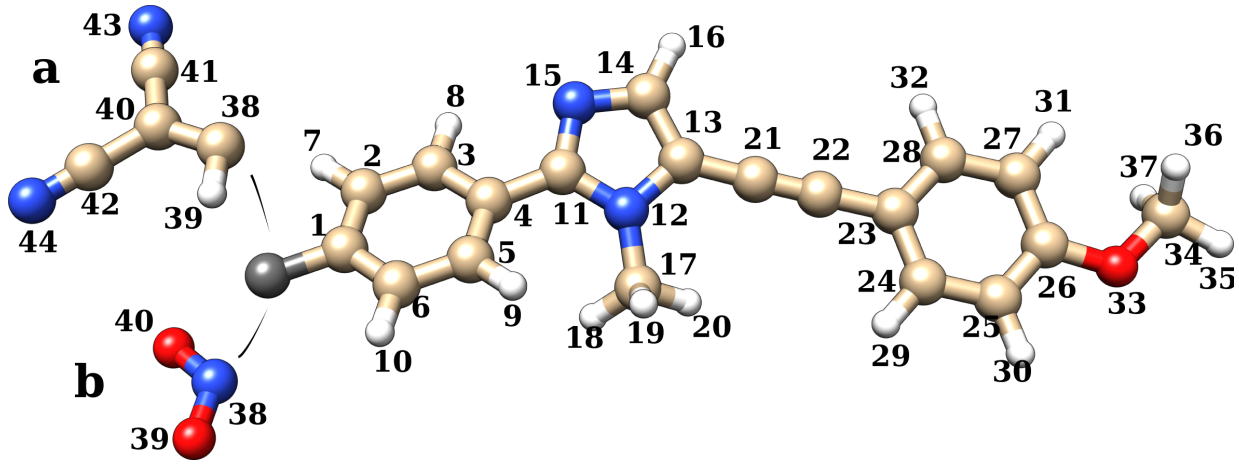


FIGURE A.1: Main scaffold, structures and atom numbers used in the topology files of the two investigated dyes.

Atom		Charge	Lennard-Jones		Atom		Charge	Lennard-Jones	
#	Type	q(e)	ϵ	σ	#	Type	q(e)	ϵi (kJ/mol)	σ (nm)
1	CA	-0.014692	0.355	0.29288	26	CA	0.097115	0.355	0.29288
2	CA	-0.077903	0.355	0.29288	27	CA	-0.115905	0.355	0.29288
3	CA	-0.080194	0.355	0.29288	28	CA	-0.085843	0.355	0.29288
4	CA	0.016234	0.355	0.29288	29	HA	0.111585	0.242	0.12552
5	CA	-0.092009	0.355	0.29288	30	HA	0.114977	0.242	0.12552
6	CA	-0.074794	0.355	0.29288	31	HA	0.112094	0.242	0.12552
7	HA	0.111753	0.242	0.12552	32	HA	0.112284	0.242	0.12552
8	HA	0.115161	0.242	0.12552	33	OS	-0.231589	0.029	0.58576
9	HA	0.114037	0.242	0.12552	34	CT	-0.12639	0.355	0.27614
10	HA	0.117279	0.242	0.12552	35	HV	0.117978	0.25	0.12552
11	CA	0.236292	0.355	0.29288	36	HV	0.106994	0.25	0.12552
12	N	-0.278203	0.325	0.71128	37	HV	0.106948	0.25	0.12552
13	CA	0.101919	0.355	0.29288					

Atom		Charge	Lennard-Jones		Atom		Charge	Lennard-Jones	
#	Type	q(e)	ϵ	σ	#	Type	q(e)	ϵi (kJ/mol)	σ (nm)
14	CA	0.02153	0.355	0.29288			**a**		
15	N	-0.408453	0.325	0.71128	38	CL	-0.022113	0.355	0.317984
16	HA	0.126081	0.242	0.12552	39	HA	0.13151	0.242	0.12552
17	CM	-0.105589	0.355	0.76144	40	CD	0.019259	0.355	0.317984
18	HC	0.121021	0.25	0.12552	41	CN	0.211037	0.33	0.276144
19	HC	0.120628	0.25	0.12552	42	CN	0.201518	0.33	0.276144
20	HC	0.11887	0.25	0.12552	43	N	-0.356313	0.32	0.71128
21	CK	-0.065676	0.33	0.87864	44	N	-0.369598	0.32	0.71128
22	CK	-0.044361	0.33	0.87864			**b**		
23	CA	-0.019544	0.355	0.29288	38	NO	0.062924	0.325	0.50208
24	CA	-0.087302	0.355	0.29288	39	ON	-0.188437	0.296	0.71128
25	CA	-0.107599	0.355	0.29288	40	ON	-0.186432	0.296	0.71128

Bond							
i	j	b_{ij}^{eq} (nm)	k_{ij}^{s} (kJ/mol)	i	j	b_{ij}^{eq} (nm)	k_{ij}^{s} (kJ/mol)
1	2	0.1414	315699.207	25	26	0.1404	347563.214
2	3	0.1383	355365.18	26	27	0.1401	347563.214
3	4	0.1412	322199.123	23	28	0.1405	333025.144
4	5	0.1407	322199.123	27	28	0.1393	348716.415
1	6	0.1412	315699.207	24	29	0.1086	336132.25
5	6	0.1386	355365.18	25	30	0.1086	336132.25
2	7	0.1083	338818.342	27	31	0.1083	336132.25
3	8	0.1084	338818.342	28	32	0.1086	336132.25
5	9	0.1083	338818.342	26	33	0.136	332011.511
6	10	0.1087	338818.342	33	34	0.1427	264301.282
4	11	0.146	320098.726	34	35	0.109	313438.804
11	12	0.1376	310883.584	34	36	0.1096	313438.804
12	13	0.1389	309257.087	34	37	0.1096	313438.804
13	14	0.1391	369872.098				
11	15	0.1337	370395.568			**a**	
14	15	0.1354	332448.31	1	38	0.1443	331364.784
14	16	0.1081	346296.187	38	39	0.1088	330151.178
12	17	0.1461	246654.559	38	40	0.137	324566.43
17	18	0.109	318908.58	40	41	0.1428	327488.968
17	19	0.1091	318908.58	40	42	0.1428	327488.968
17	20	0.1094	318908.58	41	43	0.1162	1119155.778
13	21	0.1405	406572.33	42	44	0.1162	1119155.778
21	22	0.1218	872113.995			**b**	
22	23	0.1421	408943.544	1	38	0.1464	154137.162
23	24	0.1412	333025.144	38	39	0.1231	470490.51
24	25	0.1385	348716.415	38	40	0.1231	470490.51

Angle

i	j	k	θ^{eq}_{ijk}	k^{θ}_{ijk}	i	j	k	θ^{eq}_{ijk}	k^{θ}_{ijk}
1	2	3	120.73	253.9221	22	23	24	120.97	173.8313
2	1	6	117.63	193.1521	22	23	28	120.87	173.8313
1	2	7	120.67	317.778	23	24	25	120.85	157.5618
2	1	43	125.51	335.0622	24	23	28	118.16	126.5979
2	3	4	121.43	245.4664	23	24	29	119.46	313.9231
3	2	7	118.6	331.085	24	25	26	120.35	222.8489
2	3	8	119.92	331.085	25	24	29	119.69	333.5413
3	4	5	118.05	173.1986	24	25	30	120.86	333.5413
4	3	8	118.65	334.2055	25	26	27	119.59	211.48
3	4	11	118.12	352.7555	26	25	30	118.79	306.2763
4	5	6	120.53	245.4664	25	26	33	115.89	309.6693
4	5	9	120.81	334.2055	26	27	28	119.75	222.8489
5	4	11	123.8	352.7555	26	27	31	121.13	306.2763
1	6	5	121.62	253.9221	27	26	33	124.52	309.6693
1	6	10	119.19	317.778	23	28	27	121.3	157.5618
6	1	43	116.85	335.0622	23	28	32	119.38	313.9231
6	5	9	118.63	331.085	28	27	31	119.13	333.5413
5	6	10	119.19	331.085	27	28	32	119.32	333.5413
4	11	12	126.64	281.2499	26	33	34	118.63	92.5806
4	11	15	122.48	236.5187	33	34	35	105.78	508.21
11	12	13	106.84	233.2102	33	34	36	111.15	508.21
12	11	15	110.88	263.8464	33	34	37	111.14	508.21
11	12	17	129.13	223.0908	35	34	36	109.5 3	21.1039
12	13	14	105.33	181.1476	35	34	37	109.5 3	21.1039
13	12	17	123.78	190.9795	36	34	37	109.68	321.1039
12	13	21	123.8	228.1906					
13	14	15	110.55	208.6338				**a**	
13	14	16	127.02	236.8634	1	38	39	114.32	489.3336
14	13	21	130.87	229.9816	1	38	40	131.6	107.8855
11	15	14	106.41	210.1101	39	38	40	114.08	248.6584
15	14	16	122.43	340.9489	38	40	41	119.06	167.314
12	17	18	109.83	477.4827	38	40	42	125.76	167.314
12	17	19	108.37	477.4827	41	40	42	115.18	206.1414
12	17	20	111.24	477.4827	40	41	43	178.44	684.4887
18	17	19	108.99	328.7599	40	42	44	179.95	684.4887
18	17	20	109.66	328.7599				**b**	
19	17	20	108.7	328.7599	1	38	39	118.2	443.4879
13	21	22	178.62	141.9975	1	38	40	118.2	443.4879
21	22	23	179.78	57.2501	39	38	40	123.65	1110.8395

Rigid torsion

i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}	i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}
1	2	3	4	0	41.625	17	12	13	21	0	147.627
2	3	4	5	0	112.976	21	13	14	16	0	91.483
3	4	5	6	0	112.976	23	24	25	26	0	100.257

Rigid torsion

i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}	i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}
4	5	6	1	0	41.625	24	25	26	27	0	81.101
2	1	6	5	0	102.567	25	26	27	28	0	81.101
6	1	2	3	0	102.567	26	27	28	23	0	100.257
7	2	3	8	0	45.925	24	23	28	27	0	92.593
8	3	4	11	0	124.124	28	23	24	25	0	92.593
11	4	5	9	0	124.124	22	23	24	29	0	113.375
9	5	6	10	0	45.925	29	24	25	30	0	41.63
11	12	13	14	0	98.884	30	25	26	33	0	118.527
12	13	14	15	0	189.666	33	26	27	31	0	118.527
13	14	15	11	0	289.976	31	27	28	32	0	41.63
12	11	15	14	0	269.018	22	23	28	32	0	113.375
15	11	12	13	0	160.815				**a**		
4	11	12	17	0	8.72	39	38	40	42	0	137.51

Flexible torsion

i	j	k	l	γ^{δ}_{ijkl}	k^{δ}_{ijkl}	n^{δ}	i	j	k	l	γ^{δ}_{ijkl}	k^{δ}_{ijkl}	n^{δ}
5	4	11	12	0.0	0.365	1	27	26	33	34	180.0	6.799	2
				0.0	-7.023	2					180.0	1.016	4
				0.0	0.195	3	26	33	34	35	180.0	2.088	3
				0.0	3.293	4					**a**		
				0.0	-0.126	5	2	1	38	39	0.0	-7.292	2
				0.0	0.203	6					0.0	1.096	4
11	12	17	18	0.0	-0.405	3					**b**		
				0.0	0.067	6	2	1	38	39	0.0	-6.413	2
										0.0	0.652	4	

A.2 DPAP

This section reports structures, atom indices and Joyce force field parameters of DPAP moelcular rotor (Chapter 8). LJ parameters are transferred from OPLS-AA. Charges have been computed using CM5 at the B3LYP/SNSD level of theory. Non-bonded interactions of the 1-4 type have been excluded.

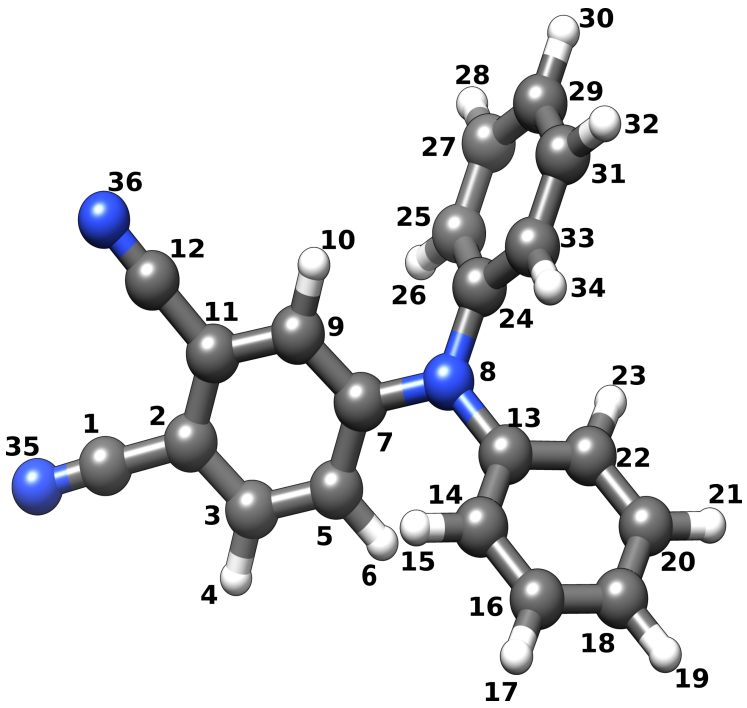


FIGURE A.2: Main scaffold, structures and atom numbers used in the topology files of DPAP molecule.

Atom		Charge	Lennard-Jones		Atom		Charge	Lennard-Jones	
#	Type	q(e)	ϵ	σ	#	Type	q(e)	ϵi (kJ/mol)	σ (nm)
1	CN	0.194079	0.365	0.62760	19	HA	0.109806	0.242	0.12552
2	CB	0.001146	0.355	0.29288	20	CA	-0.096189	0.355	0.29288
3	CA	-0.063950	0.355	0.29288	21	HA	0.111658	0.242	0.12552
4	HA	0.129524	0.242	0.12552	22	CA	-0.094065	0.355	0.29288
5	CA	-0.082391	0.355	0.29288	23	HA	0.114456	0.242	0.12552
6	HA	0.120872	0.242	0.12552	24	C	0.099678	0.355	0.29288
7	C	0.137370	0.355	0.29288	25	CA	-0.093872	0.355	0.29288
8	N	-0.295163	0.330	0.71128	26	HA	0.114647	0.242	0.12552
9	CA	-0.074139	0.355	0.29288	27	CA	-0.095871	0.355	0.29288
10	HA	0.123468	0.242	0.12552	28	HA	0.111831	0.242	0.12552
11	CB	0.022105	0.355	0.29288	29	CA	-0.100695	0.355	0.29288
12	CN	0.207369	0.365	0.62760	30	HA	0.110007	0.242	0.12552
13	C	0.099989	0.355	0.29288	31	CA	-0.095871	0.355	0.29288
14	CA	-0.094065	0.355	0.29288	32	HA	0.111831	0.242	0.12552
15	HA	0.114456	0.242	0.12552	33	CA	-0.093872	0.355	0.29288
16	CA	-0.096189	0.355	0.29288	34	HA	0.114647	0.242	0.12552
17	HA	0.111658	0.242	0.12552	35	NN	-0.400313	0.320	0.71128
18	CA	-0.101147	0.355	0.29288	36	NN	-0.382804	0.320	0.71128

Bond							
i	j	b_{ij}^{eq} (nm)	k_{ij}^{s} (kJ/mol)	i	j	b_{ij}^{eq} (nm)	k_{ij}^{s} (kJ/mol)
1	2	0.1424	325609.235	18	20	0.1396	327121.563
2	3	0.1403	317335.500	20	21	0.1086	334783.052
3	4	0.1085	334783.052	13	22	0.1399	281475.110
3	5	0.1384	327121.563	20	22	0.1395	327121.563
5	6	0.1083	334783.052	22	23	0.1086	334783.052
5	7	0.1413	281475.110	8	24	0.1434	265578.413
7	8	0.1388	265578.413	24	25	0.1400	281475.110
7	9	0.1411	281475.110	25	26	0.1086	334783.052
9	10	0.1083	334783.052	25	27	0.1394	327121.563
2	11	0.1415	230073.749	27	28	0.1086	334783.052
9	11	0.1391	317335.500	27	29	0.1397	327121.563
11	12	0.1433	325609.235	29	30	0.1086	334783.052
8	13	0.1434	265578.413	29	31	0.1397	327121.563
13	14	0.1400	281475.110	31	32	0.1086	334783.052
14	15	0.1086	334783.052	24	33	0.1399	281475.110
14	16	0.1394	327121.563	31	33	0.1395	327121.563
16	17	0.1086	334783.052	33	34	0.1086	334783.052
16	18	0.1397	327121.563	1	35	0.1164	1119383.487
18	19	0.1086	334783.052	12	36	0.1161	1119383.487

Angle									
i	j	k	θ_{ijk}^{eq}	k_{ijk}^{θ}	i	j	k	θ_{ijk}^{eq}	k_{ijk}^{θ}
1	2	3	120.53	826.0873	14	16	18	120.33	603.0479
1	2	11	121.41	392.6911	17	16	18	120.17	323.8368
2	1	35	179.60	242.1940	16	18	19	120.18	323.8368
2	3	4	119.28	357.5271	16	18	20	119.62	603.0479
2	3	5	121.17	306.8322	19	18	20	120.20	323.8368
3	2	11	118.07	490.1432	18	20	21	120.15	323.8368
4	3	5	119.55	323.8368	18	20	22	120.33	603.0479
3	5	6	119.08	323.8368	21	20	22	119.51	323.8368
3	5	7	121.15	716.7162	13	22	20	119.97	716.7162
6	5	7	119.76	286.0881	13	22	23	119.65	286.0881
5	7	8	121.35	608.1875	20	22	23	120.38	323.8368
5	7	9	117.91	425.4345	8	24	25	120.36	608.1875
8	7	9	120.75	608.1875	8	24	33	119.84	608.1875
7	8	13	121.13	338.8713	24	25	26	119.72	286.0881
7	8	24	121.20	338.8713	24	25	27	119.96	716.7162
7	9	10	120.15	286.0881	25	24	33	119.79	425.4345
7	9	11	120.82	478.0010	26	25	27	120.32	323.8368
10	9	11	119.03	357.5271	25	27	28	119.50	323.8368
2	11	9	120.88	490.1432	25	27	29	120.32	603.0479
2	11	12	120.29	392.6911	28	27	29	120.17	323.8368
9	11	12	118.82	826.0873	27	29	30	120.18	323.8368
11	12	36	179.93	242.1940	27	29	31	119.63	603.0479
8	13	14	120.38	608.1875	30	29	31	120.19	323.8368

Angle

i	j	k	θ^{eq}_{ijk}	k^{θ}_{ijk}	i	j	k	θ^{eq}_{ijk}	k^{θ}_{ijk}
8	13	22	119.83	608.1875	29	31	32	120.15	323.8368
13	8	24	117.67	338.8713	29	31	33	120.33	603.0479
13	14	15	119.72	286.0881	32	31	33	119.51	323.8368
13	14	16	119.97	716.7162	24	33	31	119.96	716.7162
14	13	22	119.78	425.4345	24	33	34	119.65	286.0881
15	14	16	120.31	323.8368	31	33	34	120.39	323.8368
14	16	17	119.50	323.8368					

Rigid torsion

i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}	i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}
11	2	3	4	-179.3	79.017	15	14	16	18	-179.0	87.040
11	2	3	5	-0.1	88.959	14	16	18	20	-0.2	73.958
35	1	2	3	1.8	0.001	17	16	18	19	0.4	30.534
1	2	11	12	0.7	105.609	16	18	20	21	-179.9	87.040
2	3	5	6	-179.3	32.981	19	18	20	22	179.8	87.040
3	2	11	9	0.0	97.751	18	20	22	13	0.5	48.999
4	3	5	7	179.2	61.357	21	20	22	23	-0.1	30.534
3	5	7	9	0.1	35.342	33	24	25	26	179.0	67.394
6	5	7	9	179.5	67.394	8	24	33	34	-1.2	31.209
5	7	9	11	-0.2	60.572	24	25	27	28	179.9	61.357
8	7	9	10	-0.8	31.209	8	24	25	27	-179.3	179.055
7	9	11	2	0.2	22.464	25	24	33	31	0.1	35.342
10	9	11	2	-179.2	79.017	26	25	27	29	-178.9	87.040
9	11	12	36	13.8	0.001	25	27	29	30	179.6	87.040
8	13	14	16	-179.3	179.055	28	27	29	31	-179.4	87.040
8	13	22	20	178.6	179.055	27	29	31	33	-0.4	73.958
13	14	16	17	179.9	61.357	30	29	31	32	0.3	30.534
22	13	14	16	-0.7	35.342	29	31	33	24	0.5	48.999
14	13	22	23	-179.9	67.394	32	31	33	34	-0.1	30.534

Improper dihedrals

i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}	i	j	k	l	ϕ^{eq}_{ijkl}	k^{ϕ}_{ijkl}
2	1	3	11	-0.0	52.740	20	16	19	18	-0.1	245.364
5	2	4	3	0.5	275.811	22	18	21	20	0.3	245.364
7	3	6	5	0.4	268.761	23	13	20	22	-0.0	345.666
9	5	8	7	-0.1	541.118	24	8	25	33	0.8	541.118
8	7	13	24	0.0	28.016	27	24	26	25	-0.2	253.992
11	7	10	9	-0.4	210.894	29	25	28	27	-0.5	245.364
12	2	9	11	-0.4	22.646	31	27	30	29	-0.1	245.364
13	8	14	22	0.8	541.118	33	29	32	31	0.3	245.364
16	13	15	14	-0.1	253.992	34	24	31	33	-0.0	345.666
18	14	17	16	-0.5	245.364						

Flexible torsion													
i	j	k	l	γ^{δ}_{ijkl}	k^{δ}_{ijkl}	n^{δ}	i	j	k	l	γ^{δ}_{ijkl}	k^{δ}_{ijkl}	n^{δ}
5	7	8	13	0.00	-4.632	2	7	8	24	33	0.00	3.642	2
				0.00	0.324	3					0.00	-0.015	3
				0.00	2.880	4					0.00	2.707	4
9	7	8	24	0.00	-4.632	2	7	8	13	22	0.00	3.642	2
				0.00	0.324	3					0.00	-0.015	3
				0.00	2.880	4					0.00	2.707	4
13	8	24	25	0.00	3.642	2	24	8	13	14	0.00	3.642	2
				0.00	-0.015	3					0.00	-0.015	3
				0.00	2.707	4					0.00	2.707	4

A.3 Metal ions

Here, the force field parameters optimized for the metal ions Zn^{2+}, Ni^{2+}, Mg^{2+}, Na^{2+} and Ca^{2+} are reported. When atomic charge is optimized, LJ parameters are written as A_{ij}^{12} and B_{ij}^{6} instead of standard σ and ϵ, (as they compared in Eq. 1.9). Since the optimization has been conducted in water environment, the reported parameters are intended to work in water solution. Moreover, since water model has not vdW, the LJ interactions involve only metal ions and water oxygens.

	σ (nm)	ϵ (kJ/mol)
Zn^{2+}	0.3516905	0.000204
Ni^{2+}	0.1145527	744.7884
Mg^{2+}	0.1939073	0.734308
Na^{2+}	0.2150201	1.686356
Ca^{2+}	0.2577062	2.029290

TABLE A.12: Metal ions LRR-DE optimized parameters for the 12-6-1_F model.

	A^{12} (kJ mol^{-1} nm^{12})	B^{6} (kJ mol^{-1} nm^{6})	q (e)
Zn^{2+}	3.405e-08	-0.004349	2.300773
Ni^{2+}	1.232e-07	-0.000057	2.213392
Mg^{2+}	3.344e-08	-0.005197	2.321638
Na^{2+}	3.689e-07	0.000620	1.021000
Ca^{2+}	6.562e-07	-0.008399	2.271904

TABLE A.13: Metal ions LRR-DE optimized parameters for the 12-6-1 model.

	C1 (kJ mol^{-1} nm^{12})	C2 (e)	θ_1 (nm)	θ_2 (nm^{-1})	θ_3 (nm^{-2})	θ_4 (nm)
Zn^{2+}	2.912e-05	696.482913	-0.099025	15.281931	0.270012	0.126974
Ni^{2+}	4.1724190	4.9229e-07	-0.021800	10.000000	7.000000	0.100000
Mg^{2+}	-22.109731	-91.616616	-0.795994	20.000000	10.00000	0.270071
Na^{2+}	5.782e-06	2.4124543	-0.057371	8.4124410	6.089116	0.182687
Ca^{2+}	1.085e-04	11.3382033	-0.101856	10.1049636	8.342623	0.218861

TABLE A.14: Metal ions LRR-DE optimized parameters for the 12b-1-E1Gr model.

Appendix B

Systematic study of zinc potentials

This section reports the computed mean absolute errors (kJ mol^{-1} for the energies, E, and kJ mol^{-1} nm^{-1} for the forces, F) in the prediction of the quantities indicated in the first column. The training set data is indicated in the first row. Reported data are relative to the optimized 12b-1 model. The values of the errors for the same data set used in the training process are marked in bold. F(6O, $n$$H_2O$) are the forces on the six oxygen atoms closest to the zinc ion for the cluster with n water molecules. F(12H, $n$$H_2O$) are the forces on the twelve hydrogen atoms closest to the zinc ion for the cluster with n water molecules.

	Training set						
Test set	F(Zn, $6H_2O$)	F(Zn, $16H_2O$)	F(Zn, $32H_2O$)	F(Zn, $64H_2O$)	F(Zn, $128H_2O$)	AMBER99	Li [259]	HF
E($6H_2O$)	761.97	530.02	551.62	559.56	576.14	140.97	116.71	153.47
E($16H_2O$)	949.67	650.46	691.86	703.52	726.39	86.14	68.36	146.89
E($32H_2O$)	998.71	669.10	719.59	732.93	758.65	93.35	74.38	136.58
E($64H_2O$)	1060.05	702.01	761.09	776.03	804.42	86.11	69.96	129.74
E($128H_2O$)	1083.49	706.46	771.29	787.29	817.46	95.75	77.10	113.38
F(Zn, $6H_2O$)	**92.95**	117.31	113.81	111.44	111.08	748.83	627.34	65.13
F(Zn, $16H_2O$)	191.20	**175.86**	177.88	178.36	179.11	724.32	604.13	71.97
F(Zn, $32H_2O$)	168.91	160.55	**159.70**	159.41	159.70	724.55	603.82	74.96
F(Zn, $64H_2O$)	167.94	162.12	161.04	**160.58**	160.79	726.41	605.57	73.59
F(Zn, $128H_2O$)	165.53	160.95	159.46	158.92	**158.96**	724.29	604.22	73.81
F(6O, $6H_2O$)	312.62	284.89	269.23	261.93	252.14	1766.17	1650.26	767.28
F(6O, $16H_2O$)	302.87	357.11	338.21	329.26	316.50	1842.67	1727.30	808.86
F(6O, $32H_2O$)	286.98	370.51	349.89	340.15	325.87	1867.99	1752.40	807.00
F(6O, $64H_2O$)	285.65	363.77	343.25	333.63	319.68	1861.30	1745.70	808.01
F(6O, $128H_2O$)	285.94	362.59	342.19	332.58	318.42	1859.89	1744.27	808.02
F(12H, $6H_2O$)	259.02	301.66	285.63	282.95	278.63	356.05	356.05	527.17
F(12H, $16H_2O$)	298.62	347.19	329.60	326.57	321.70	403.66	403.66	617.26
F(12H, $32H_2O$)	293.34	348.17	329.14	325.80	320.35	408.32	408.32	630.00
F(12H, $64H_2O$)	296.21	349.11	330.48	327.21	321.91	407.67	407.67	629.47
F(12H, $128H_2O$)	296.36	349.04	330.52	327.29	322.08	407.36	407.36	629.30
Optimized charge	1.529	1.787	1.709	1.694	1.670			

TABLE B.1

TABLE B.2

	Training set							
Test set	2-o($6H_2O$)	2-o($16H_2O$)	2-o($32H_2O$)	2-o($64H_2O$)	2-o($128H_2O$)	AMBER99	Li [259]	HF
E($6H_2O$)	**23.39**	32.42	55.39	72.37	96.84	140.97	116.71	153.47
E($16H_2O$)	49.05	**40.46**	53.40	67.63	92.18	86.14	68.36	146.89
E($32H_2O$)	88.38	55.27	**41.09**	45.45	60.94	93.35	74.38	136.58
E($64H_2O$)	120.15	76.97	44.57	**42.50**	50.37	86.11	69.96	129.74
E($128H_2O$)	159.29	108.18	61.61	49.05	**43.38**	95.75	77.10	113.38
F(Zn, $6H_2O$)	**191.19**	185.30	179.20	176.33	173.07	748.83	627.34	65.13
F(Zn, $16H_2O$)	211.04	**206.28**	202.59	200.78	198.25	724.32	604.13	71.97
F(Zn, $32H_2O$)	209.46	201.74	**196.89**	195.42	192.96	724.55	603.82	74.96
F(Zn, $64H_2O$)	205.33	198.76	194.40	**192.81**	190.19	726.41	605.57	73.59
F(Zn, $128H_2O$)	203.45	198.02	193.99	192.08	**189.89**	724.29	604.22	73.81
F(6O, $6H_2O$)	980.81	963.04	921.68	891.74	855.18	1766.17	1650.26	767.28
F(6O, $16H_2O$)	1060.42	1042.81	1001.63	971.78	935.36	1842.67	1727.30	808.86
F(6O, $32H_2O$)	1084.55	1066.84	1025.49	995.56	959.06	1867.99	1752.40	807.00
F(6O, $64H_2O$)	1077.10	1059.38	1018.04	988.11	951.59	1861.30	1745.70	808.01
F(6O, $128H_2O$)	1075.69	1058.00	1016.69	986.76	950.18	1859.89	1744.27	808.02
F(12H, $6H_2O$)	480.17	462.79	447.00	443.84	439.29	356.05	356.05	527.17
F(12H, $16H_2O$)	527.06	509.90	494.26	491.13	486.62	403.66	403.66	617.26
F(12H, $32H_2O$)	535.97	518.37	502.29	499.07	494.43	408.32	408.32	630.00
F(12H, $64H_2O$)	533.35	515.96	500.16	497.00	492.44	407.67	407.67	629.47
F(12H, $128H_2O$)	532.67	515.32	499.51	496.35	491.78	407.36	407.36	629.30
Optimized charge	2.385	2.335	2.288	2.279	2.266			

	Training set							
Test set	4-o($6H_2O/16H_2O$)	4-o($6H_2O/32H_2O$)	4-o($6H_2O/64H_2O$)	4-o($6H_2O/128H_2O$)	4-o($16H_2O/32H_2O$)	AMBER99	Li [259]	HF
E($6H_2O$)	**25.61**	**28.53**	**29.24**	**30.89**	35.95	140.97	116.71	153.47
E($16H_2O$)	**41.35**	44.18	46.61	52.51	**44.60**	86.14	68.36	146.89
E($32H_2O$)	58.82	**42.31**	41.40	43.98	**42.66**	93.35	74.38	136.58
E($64H_2O$)	78.20	47.29	**44.14**	44.39	50.92	86.11	69.96	129.74
E($128H_2O$)	107.01	61.87	53.79	**45.98**	71.39	95.75	77.10	113.38
F(Zn, $6H_2O$)	**185.57**	**186.28**	**188.44**	**192.61**	181.84	748.83	627.34	65.13
F(Zn, $16H_2O$)	**208.47**	210.93	212.95	216.36	**205.24**	724.32	604.13	71.97
F(Zn, $32H_2O$)	203.07	**203.76**	205.45	208.34	**198.47**	724.55	603.82	74.96
F(Zn, $64H_2O$)	199.92	200.09	**202.17**	205.84	196.22	726.41	605.57	73.59
F(Zn, $128H_2O$)	199.22	200.70	202.97	**206.32**	196.13	724.29	604.22	73.81
F(6O, $6H_2O$)	993.07	1006.81	1012.74	1023.06	967.19	1766.17	1650.26	767.28
F(6O, $16H_2O$)	1072.67	1086.38	1092.28	1102.56	1046.95	1842.67	1727.30	808.86
F(6O, $32H_2O$)	1096.83	1110.57	1116.51	1126.82	1070.99	1867.99	1752.40	807.00
F(6O, $64H_2O$)	1089.38	1103.12	1109.05	1119.37	1063.54	1861.30	1745.70	808.01
F(6O, $128H_2O$)	1087.95	1101.68	1107.61	1117.93	1062.14	1859.89	1744.27	808.02
F(12H, $6H_2O$)	450.33	420.78	412.58	400.61	439.72	356.05	356.05	527.17
F(12H, $16H_2O$)	497.57	468.27	460.14	448.28	487.05	403.66	403.66	617.26
F(12H, $32H_2O$)	505.69	475.54	467.16	454.89	494.87	408.32	408.32	630.00
F(12H, $64H_2O$)	503.49	473.78	465.48	453.39	492.87	407.67	407.67	629.47
F(12H, $128H_2O$)	502.86	473.12	464.85	452.82	492.22	407.36	407.36	629.30
Optimized charge	2.298	2.210	2.185	2.148	2.267			

TABLE B.3

	Training set						
Test set	4-o($16H_2O/64H_2O$)	4-o($16H_2O/128H_2O$)	4-o($32H_2O/64H_2O$)	4-o($32H_2O/128H_2O$)	4-o($64H_2O/128H_2O$)	AMBER99	Li [259]	HF
E($6H_2O$)	33.15	32.69	56.33	51.59	75.72	140.97	116.71	153.47
E($16H_2O$)	**46.97**	**51.79**	58.16	63.37	75.96	86.14	68.36	146.89
E($32H_2O$)	41.29	42.82	**42.31**	**46.23**	51.19	93.35	74.38	136.58
E($64H_2O$)	**44.95**	43.36	**42.59**	43.80	**45.09**	86.11	69.96	129.74
E($128H_2O$)	57.99	**47.83**	53.13	**45.43**	**44.09**	95.75	77.10	113.38
F(Zn, $6H_2O$)	183.09	187.52	177.81	178.21	174.69	748.83	627.34	65.13
F(Zn, $16H_2O$)	**207.22**	**210.86**	202.25	203.36	200.09	724.32	604.13	71.97
F(Zn, $32H_2O$)	199.97	202.71	**195.81**	**195.77**	193.75	724.55	603.82	74.96
F(Zn, $64H_2O$)	**197.27**	200.33	**193.65**	193.69	**191.55**	726.41	605.57	73.59
F(Zn, $128H_2O$)	197.39	**200.87**	193.18	**193.63**	**191.03**	724.29	604.22	73.81
F(6O, $6H_2O$)	987.41	1007.83	922.47	942.65	890.63	1766.17	1650.26	767.28
F(6O, $16H_2O$)	1067.08	1087.40	1002.41	1022.52	970.67	1842.67	1727.30	808.86
F(6O, $32H_2O$)	1091.20	1111.59	1026.27	1046.45	994.44	1867.99	1752.40	807.00
F(6O, $64H_2O$)	1083.75	1104.14	1018.82	1039.01	986.99	1861.30	1745.70	808.01
F(6O, $128H_2O$)	1082.31	1102.69	1017.48	1037.63	985.64	1859.89	1744.27	808.02
F(12H, $6H_2O$)	424.58	409.41	438.89	422.85	435.14	356.05	356.05	527.17
F(12H, $16H_2O$)	472.03	457.00	486.22	470.31	482.50	403.66	403.66	617.26
F(12H, $32H_2O$)	479.41	463.91	494.02	477.65	490.20	408.32	408.32	630.00
F(12H, $64H_2O$)	477.62	462.27	492.03	475.87	488.27	407.67	407.67	629.47
F(12H, $128H_2O$)	476.95	461.66	491.38	475.21	487.61	407.36	407.36	629.30
Optimized charge	2.221	2.175	2.264	2.216	2.253			

TABLE B.4

Bibliography

[1] A. R. Leach. *Molecular modelling: principles and applications*. Pearson Education, 2001.

[2] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[3] E. Harder, W. Damm, J. Maple, C. Wu, M. Reboul, J. Y. Xiang, L. Wang, D. Lupyan, M. K. Dahlgren, J. L. Knight, et al. *J. Chem. Theory Comput.*, 12(1):281–296, 2015.

[4] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. *Proteins: Struct., Funct., Bioinf.*, 65(3):712–725, 2006.

[5] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. *Proteins: Struct., Funct., Bioinf.*, 78(8):1950–1958, 2010.

[6] N. Schmid, A. P. Eichenberger, A. Choutko, S. Riniker, M. Winger, A. E. Mark, and W. F. van Gunsteren. *Eur. Biophys. J.*, 40(7):843–856, July 2011.

[7] H. Heinz, T. J. Lin, R. Kishore Mishra, and F. S. Emami. *Langmuir*, 29(6):1754–1765, 2013.

[8] R. T. Cygan, J. J. Liang, and A. G. Kalinichev. *J. Phys. Chem. B*, 108(4):1255–1266, 2004.

[9] V. Barone, I. Cacelli, N. De Mitri, D. Licari, S. Monti, and G. Prampolini. *Phys. Chem. Chem. Phys.*, 15(11):3736–3751, 2013.

[10] S. Grimme. *J. Chem. Theory Comput.*, 10(10):4497–4514, 2014.

[11] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J. C. Gumbart. *J. Comput. Chem.*, 34(32):2757–2770, December 2013.

[12] R. M. Betz and R. C. Walker. *J. Comput. Chem.*, 36(2):79–87, January 2015.

[13] L. P. Wang, J. Chen, and T. Van Voorhis. *J. Chem. Theory Comput.*, 9(1):452–460, January 2013.

[14] D. Cole, J. Z. Vilseck, J. Tirado-Rives, M. C. Payne, and W. L Jorgensen. *J. Chem. Theory Comput.*, 12(5):2312–2323, 2016.

[15] L. S. Dodda, J. Z. Vilseck, K. J. Cutrona, and W. L. Jorgensen. *J. Chem. Theory Comput.*, 11(9):4273–4282, 2015.

[16] M. Macchiagodena, G. Mancini, M. Pagliai, and V. Barone. *Phys. Chem. Chem. Phys.*, 18(36):25342–25354, 2016.

[17] P. M. Morse. *Phys. Rev.*, 34(1):57, 1929.

[18] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. States, S. Swaminathan, and M. Karplus. *J. Comput. Chem.*, 4(2):187–217, 1983.

[19] F. London. *Trans. Farad. Soc.*, 33:8b–26, 1937.

[20] John Edward Jones. In *Proc. Roy. Soc. (London)*, volume 106, pages 463–477. The Royal Society, 1924.

[21] Richard A Buckingham. In *Proc. Roy. Soc. (London)*, volume 168, pages 264–283. The Royal Society, 1938.

[22] H. A. Lorentz. *Annalen der physik*, 248(1):127–136, 1881.

[23] D. Berthelot. *Compt. Rendus*, 126:1703–1706, 1898.

[24] B. E. F. Fender and G. D. Halsey Jr. *J. Chem. Phys.*, 36(7):1881–1888, 1962.

[25] M. Waldman and A. T Hagler. *J. Comput. Chem.*, 14(9):1077–1084, 1993.

[26] A. Nikitin, Y. Milchevskiy, and A. Lyubartsev. *J. Phys. Chem. B*, 119(46):14563–14573, 2015.

[27] S. A. Lebedeff. *J. Chem. Phys.*, 40(9):2716–2721, 1964.

[28] C. E. Cloete, T. W. Smuts, and K. De Clerk. *J. Chromatogr. A*, 120(1):29–34, 1976.

[29] B. N. Srivastava and I. B. Srivastava. *J. Chem. Phys.*, 36(10):2616–2620, 1962.

[30] S. Lifson and M. Levitt. *Computers & Chemistry*, 3(2-4):49–50, 1979.

[31] A. Bondi. *J. Phys. Chem.*, 68(3):441–451, 1964.

[32] G. Gafner, F. H. Herbstein, and C. M. Lee. *Acta Cryst.*, 28(5):422–426, 1972.

[33] R. Chauvin. *J. Phys. Chem.*, 96(23):9194–9197, 1992.

[34] L. Pauling. *The nature of the chemical bond and the structure of molecules and crystals: an introduction to modern structural chemistry*, volume 18. Cornell university press, 1960.

[35] M. Oobatake and T. Ooi. *Progress of theoretical physics*, 48(6):2132–2143, 1972.

[36] C. D. Berweger, W. F. van Gunsteren, and F. Müller-Plathe. *Chem. Phys. Lett.*, 232(5-6):429–436, 1995.

[37] T. van Westen, T. H. Vlugt, and J. Gross. *J. Phys. Chem. B*, 115(24):7872–7880, 2011.

[38] X. Daura, A. E. Mark, and W. F Van Gunsteren. *J. Comput. Chem.*, 19(5):535–547, 1998.

[39] W. L. Jorgensen and N. A. McDonald. *J. Mol. Struct. (THEOCHEM)*, 424(1-2):145–155, 1998.

[40] F. J. Salas, G. A. Mendez-Maldonado, E. Núñez-Rojas, G. E. Aguilar-Pineda, H. Domínguez, and J. Alejandre. *J. Chem. Theory Comput.*, 11(2):683–693, 2015.

[41] G. Liang, P. C. Fox, and P. Bowen. *J. Comput. Chem.*, 17(8):940–953, 1996.

[42] D. Yin and A. D. MacKerell. *J. Comput. Chem.*, 19(3):334–348, 1998.

[43] S. D. Fried, L. P. Wang, S. G. Boxer, P. Ren, and V. S. Pande. *J. Phys. Chem. B*, 117(50):16236–16248, 2013.

[44] J. Gasteiger and M. Marsili. *Tetrahedron*, 36(22):3219–3228, 1980.

[45] A. K. Rappe and W. A. Goddard III. *J. Phys. Chem.*, 95(8):3358–3363, 1991.

[46] A. K. Rappé, C. Casewit, K. S. Colwell, W. A. Goddard III, and W. M. Skiff. *J. Am. Chem. Soc.*, 114(25):10024–10035, 1992.

[47] S. Geidl, T. Bouchal, T. Raček, R. S. Vařeková, V. Hejret, A. Křenek, R. Abagyan, and aroslav Koča. *J. Cheminform.*, 7(1):59, 2015.

[48] R. S. Mulliken. *J. Chem. Phys.*, 23(10):1833–1840, 1955.

[49] F. L. Hirshfeld. *Theor. Chem. Acc.*, 44(2):129–138, 1977.

[50] P. Löwdin. *J. Chem. Phys.*, 18(3):365–375, 1950.

[51] U. C. Singh and P. A. Kollman. *J. Comput. Chem.*, 5(2):129–145, 1984.

[52] C. I. Bayly, P. Cieplak, W. Cornell, and P. A. Kollman. *J. Phys. Chem.*, 97(40):10269–10280, 1993.

[53] J. W. Storer, D. J. Giesen, C. J. Cramer, and D. G Truhlar. *J. Comput. Aided Mol. Des.*, 9(1):87–110, 1995.

[54] A. V. Marenich, S. V. Jerome, C. J. Cramer, and D. G. Truhlar. *J. Chem. Theory Comput.*, 8(2):527–541, 2012.

[55] H. Sun. *J. Phys. Chem. B*, 102(38):7338–7364, 1998.

[56] S. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. *J. Am. Chem. Soc.*, 106(3):765–784, 1984.

[57] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. *J. Am. Chem. Soc.*, 117(19):5179–5197, 1995.

[58] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. *J. Chem. Theory Comput.*, 11(8):3696–3713, August 2015.

[59] J. Wang, R. M Wolf, J. W. Caldwell, P. A Kollman, and D. A. Case. *J. Comput. Chem.*, 25(9):1157–1174, 2004.

[60] W. L. Jorgensen, J. D. Madura, and C. J. Swenson. *J. Am. Chem. Soc.*, 106(22):6638–6646, 1984.

[61] L. Monticelli and D. P. Tieleman. *Biomolecular simulations: Methods and protocols*, pages 197–213, 2013.

[62] S. E. Feller and A. D. MacKerell. *J. Phys. Chem. B*, 104(31):7510–7515, 2000.

[63] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, . Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. D. Mackerell. *J. Comput. Chem.*, 31(4):671–690, 2010.

[64] J. Hermans, H. J. C. Berendsen, W. F. Van Gunsteren, and J. P. M. Postma. *Biopolymers*, 23(8):1513–1518, 1984.

[65] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren. *J. Comput. Chem.*, 25(13):1656–1676, 2004.

[66] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. . *Chem. Phys.*, 79(2):926–935, 1983.

[67] H. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. In *Intermolecular forces*, pages 331–342. Springer, 1981.

[68] K. Toukan and A. Rahman. *Phys. Rev. B*, 31(5):2643, 1985.

[69] L. P. Wang, T. J. Martinez, and V. S Pande. *J. Phys. Chem. letters*, 5(11):1885–1891, 2014.

[70] A. Pedone, G. Malavasi, M. C. Menziani, A. N. Cormack, and U. Segre. *J. Phys. Chem. B*, 110(24):11780–11795, 2006.

[71] A. C. T. Van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard. *J. Phys. Chem. A*, 105(41):9396–9409, 2001.

[72] K. D. Nielson, A. C. T. van Duin, J. Oxgaard, W. Q. Deng, and W. A. Goddard. *J. Phys. Chem. A*, 109(3):493–499, 2005.

[73] O. Rahaman, A. C. T. Van Duin, V. S. Bryantsev, J. E. Mueller, S. D. Solares, W. A. Goddard III, and D. J. Doren. *J. Phys. Chem. A*, 114(10):3556–3568, 2010.

[74] J. L. Banks, G. A. Kaminski, R. Zhou, D. T. Mainz, B. J. Berne, and R. A. Friesner. *J. Chem. Phys.*, 110(2):741–754, 1999.

[75] P. Ren and J. W. Ponder. *J. Comput. Chem.*, 23(16):1497–1506, 2002.

[76] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr, et al. *J. Phys. Chem. B*, 114(8):2549–2564, 2010.

[77] P. Drude. *Annalen der Physik*, 306(3):566–613, 1900.

[78] M. Jana and A. D. MacKerell. *J. Phys. Chem. B*, 119(25):7846–7859, 2015.

[79] F. Ercolessi and J. B. Adams. *EPL (Europhys. Lett.)*, 26(8):583, 1994.

[80] P. Maurer, A. Laio, H. W. Hugosson, M. C. Colombo, and U. Rothlisberger. *J. Chem. Theory Comput.*, 3(2):628–639, 2007.

[81] I. Cacelli and G. Prampolini. *J. Chem. Theory Comput.*, 3(5):1803–1817, 2007.

[82] M. P. Allen and D. Tildesley. *Computer simulation of liquids*. Oxford university press, 1989.

[83] L. Verlet. *Phys. Rev.*, 159(1):98, 1967.

[84] R. W. Hockney, S. P. Goel, and W. Eastwood. *J. Comput. Phys.*, 14(2):148–158, 1974.

[85] K. W. Kratky. *J. Comput. Phys.*, 37(2):205–217, 1980.

[86] D. Beglov and B. Roux. *J. Chem. Phys.*, 100(12):9050–9063, 1994.

[87] Y. Li, G. Krilov, and B. Berne. *J. Phys. Chem. B*, 109(1):463–470, 2005.

[88] G. Mancini, G. Brancato, B. Chandramouli, and V. Barone. *Chem. Phys. Lett.*, 625:186–192, 2015.

[89] T. Darden, L. Perera, L. Li, and L. Pedersen. *Structure*, 7(3):R55 – R60, 1999.

[90] T. Darden, D. York, and L. Pedersen. *J. Chem. Phys.*, 98(12):10089–10092, 1993.

[91] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. *J. Comput. Phys.*, 23(3):327–341, 1977.

[92] S. Miyamoto and P. A. Kollman. *J. Comput. Chem.*, 13(8):952–962, 1992.

[93] B. Hess, H. Bekker, H. C. Berendsen, and J. G. E. M. Fraaije. *J. Comput. Chem.*, 18(12):1463–1472, 1997.

[94] H. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. Di Nola, and J. R. Haak. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[95] G. Bussi, D. Donadio, and M. Parrinello. *J. Chem. Phys.*, 126(1):014101, 2007.

[96] M. Parrinello and A. Rahman. *J. Appl. Phys.*, 52(12):7182–7190, 1981.

[97] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, B. G. Janesko, F. Lipparini, G. Zheng, J. L. Sonnenberg, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J.A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, P. V. Parandekar, N. J. Mayhall, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, , and D. J. Fox. Gaussian 09 Development Version and Revision i.04p. Gaussian Inc. Wallingford CT 2010.

[98] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl. *Bioinformatics*, 29(7):845–854, 2013.

[99] J. Tomasi, B. Mennucci, and R. Cammi. *Chem. Rev.*, 105(8):2999–3094, 2005.

[100] S. K. Burger, M. Lacasse, T. Verstraelen, J. Drewry, P. Gunning, and P. W. Ayers. *J. Chem. Theory Comput.*, 8(2):554–562, 2012.

[101] J. M. Seminario. *Int. J. Quantum Chem.*, 60(7):1271–1277, 1996.

[102] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, and K. R. Müller. *J. Chem. Theory Comput.*, 9(8):3404–3419, 2013.

[103] J. Behler. *Phys. Chem. Chem. Phys.*, 13(40):17930–17955, 2011.

[104] V. Botu, R. Batra, J. Chapman, and R. Ramprasad. *J. Phys. Chem. C*, 121(1):511–522, 2016.

[105] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J. A. K. Suykens. *LS-SVMlab Toolbox User's Guide: version 1.7*. Katholieke Universiteit Leuven, 2010.

[106] R. Storn and K. Price. *J. Glob. Optim.*, 11(4):341–359, 1997.

[107] S. Das and P. N. Suganthan. *IEEE Trans. Evolut. Comput.*, 15(1):4–31, 2011.

[108] A. N. Tikhonov and V. Y. Arsenin. *Solutions of ill-posed problems*. Winston, 1977.

[109] A. N. Tikhonov. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

[110] A. E. Hoerl. *Chem. Eng. Prog.*, 58(3):54–59, 1962.

[111] A. E. Hoerl and R. W. Kennard. *Technometrics*, 12(1):55–67, 1970.

[112] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[113] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[114] J. Vesterstrom and R. Thomsen. In *Evolutionary Computation, 2004. CEC2004. Congress on*, volume 2, pages 1980–1987. IEEE, 2004.

[115] R. K. Ursem and P. Vadstrup. In *Evolutionary Computation, 2003. CEC'03. The 2003 Congress on*, volume 2, pages 790–796. IEEE, 2003.

[116] J. Zhang, Q. Niu, K. Li, and G. W. Irwin. *IFAC Proceedings Volumes*, 44(1):14717–14722, 2011.

[117] N. Di Pasquale, S. Davie, and P. L. A. Popelier. *J. Chem. Theory Comput.*, 12(4):1499–1513, 2016.

[118] O. T. Unke, M. Devereux, and M. Meuwly. *J. Chem. Phys.*, 147(16):161712, 2017.

[119] S. Izvekov, M. Parrinello, C. Burnham, and G. A. Voth. *J. Chem. Phys.*, 120(23):10896–10913, 2004.

[120] O. Akin-Ojo, Y. Song, and F. Wang. *J. Chem. Phys.*, 129(6):064108, 2008.

[121] R. T. Marler and J. S. Arora. *Struct. Multidiscip. O.*, 26(6):369–395, 2004.

[122] E. Zitzler and L. Thiele. *IEEE Trans. Evolut. Comput.*, 3(4):257–271, 1999.

[123] P. L. Yu and G Leitmann. *J. Optim. Theory. Appl.*, 13(3):362–378, 1974.

[124] S. S. Rao and T. I. Freiheit. *ASME . Mech. Des*, 113(3):286–291, 1991.

[125] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. *Science*, 220(4598):671–680, 1983.

[126] Becke A. D. *J. Chem. Phys.*, 98(7):5648–5652, 1993.

[127] Yanai T., Tew D. P., and Handy N. C. *Chem. Phys. Lett.*, 393(1):51 – 57, 2004.

[128] T. M. Nymand and P. Linse. *J. Chem. Phys.*, 112(14):6386–6395, 2000.

[129] D. C. Rapaport. *Molecular Physics*, 50(5):1151–1162, 1983.

[130] M. Pagliai, G. Cardini, R. Righini, and V. Schettino. *J. Chem. Phys.*, 119(13):6655–6662, 2003.

[131] M. Pagliai, G. Mancini, I. Carnimeo, N. De Mitri, and V. Barone. *J. Comput. Chem.*, 38(6):319–335, 2017.

[132] X. H. Zhang, L. Y. Wang, G. H. Zhai, Z. Y. Wen, and Z. X. Zhang. *J. Mol. Struct. (THEOCHEM)*, 906(1):50–55, 2009.

[133] G. G. Hall and C. M. Smith. *Int. J. Quantum Chem.*, 25(5):881–890, 1984.

[134] C. M. Smith and G. G. Hall. *Theor. Chim. Acta*, 69(1):63–69, 1986.

[135] G. M. Torrie and J. P. Valleau. *J. Comput. Phys.*, 23(2):187–199, 1977.

[136] C. H. Bennett. *J. Comput. Phys.*, 22(2):245–268, 1976.

[137] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[138] T. C. Beutler, A. E. Mark, R. C. van Schaik, P. R. Gerber, and W. F. van Gunsteren. *Chem. Phys. Lett.*, 222(6):529–539, 1994.

[139] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham. *J. Chem. Theory Comput.*, 3(6):2312–2334, 2007.

[140] J. Han, J. Pei, and M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[141] J. MacQueen et al. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[142] L. Kaufman and P. Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987.

[143] T. Zhang, R. Ramakrishnan, and M. Livny. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM, 1996.

[144] M. Ester, H. P. Kriegel, J. Sander, X. Xu, et al. In *Kdd*, volume 96, pages 226–231, 1996.

[145] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander. In *ACM Sigmod record*, volume 28, pages 49–60. ACM, 1999.

[146] W. Wang, J. Yang, R. Muntz, et al. In *VLDB*, volume 97, pages 186–195, 1997.

[147] T. Caliński and J. Harabasz. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[148] A. Salvadori, A. Brogni, G. Mancini, and V. Barone. In *Augmented and Virtual Reality: First International Conference, AVR 2014, Lecce, Italy, September 17-20, 2014, Revised Selected Papers*, pages 333–350, cham, 2014. Springer International Publishing.

[149] J. Brickmann, E. T. Exner, M. Keil, and J. R. Marhöfer. *Molecular Modeling Annual*, 6(2):328–340, 2000.

[150] M. Valle. *Int. J. Quantum Chem.*, 113(17):2040–2052, 2013.

[151] N. Luehr, A. G. B. Jin, and T. J. Martínez. *J. Chem. Theory Comput.*, 11(10):4536–4544, 2015. PMID: 26574246.

[152] J. D. Hirst, D. R. Glowacki, and M. Baaden. *Faraday Disc.*, 169:9–22, 2014.

[153] J. E. Stone, A. Kohlmeyer, K. L. Vandivort, and K. Schulten. In G. Bebis, R. Boyle, B. Parvin, D. Koracin, R. Chung, R. Hammound, M. Hussain, T. Kar-Han, R. Crawfis, D. Thalmann, D. Kao, and L. Avila, editors, *Advances in Visual Computing: 6th International Symposium, ISVC 2010, Las Vegas, NV, USA, November 29 – December 1, 2010, Proceedings, Part II*, pages 382–393, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[154] R. Sharma, M. Zeller, V. I. Pavlovic, T. S. Huang, Z. Lo, S. Chu, Y. Zhao, J. C. Phillips, and K. Schulten. *IEEECGA*, 20:29–37, 2000.

[155] Microsoft. `http://www.xbox.com/en-US/xbox-one/accessories/kinect-for-xbox-one`.

[156] Leap Motion Inc. `https://www.leapmotion.com`.

[157] Oculus VR LLC. `https://www.oculus.com/en-us/rift/`.

[158] HTC and Valve. `http://www.htcvive.com`.

[159] Novint Technologies Inc. `http://www.novint.com/index.php/novintfalcon`.

[160] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart. *Commun. ACM*, 35(6):64–72, June 1992.

[161] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '93, pages 135–142, New York, NY, USA, 1993. ACM.

[162] `http://www.keele.ac.uk/pharmacy/digital/kave/`.

[163] `http://www.iumsc.indiana.edu/graphics/XMView/`.

[164] Scuola Normale Superiore. `http://smart.sns.it`.

[165] W. A. Denny and B. C. Baguley. *Curr. Top. Med. Chem.*, 3(3):339–353, 2003.

[166] K. M. Tewey, T. C. Rowe, L. Yang, B. D. Halligan, and L. F. Liu. *Science*, 226(4673):466–468, 1984.

[167] M. Trieb, C. Rauch, F. R. Wibowo, B. Wellenzohn, and K. R Liedl. *Nucleic Acids Res.*, 32(15):4696–4703, 2004.

[168] H. Lei, X. Wang, and C. Wu. *J. Mol. Graph. Model.*, 38:279–289, 2012.

[169] A. Mukherjee, R. Lavery, B. Bagchi, and J. T. Hynes. *J. Am. Chem. Soc.*, 130(30):9747–9755, 2008.

[170] R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, and C. Laughton. *Nucleic Acids Res.*, 38(1):299–313, 2009.

[171] M. Wilhelm, A. Mukherjee, B. Bouvier, K. Zakrzewska, J. T. Hynes, and R. Lavery. *J. Am. Chem. Soc.*, 134(20):8588–8596, 2012. PMID: 22548344.

[172] J. B. Chaires, N. Dattagupta, and D. M. Crothers. *Biochemistry*, 21:3927–3932, 1982.

[173] J. B. Chaires, S. Satyanarayana, D. Suh, I. Fokt, T. Przewloka, and W. Priebe. *Biochemistry*, 35(7):2047–2053, 1996.

[174] A. H. J. Wang, G. Ughetto, G. J. Quigley, and A. Rich. *Biochemistry*, 26(4):1152–1163, 1987. PMID: 3567161.

[175] Gaussian. `http://www.gaussian.com/g_prod/gv5.htm`.

[176] P. Cieplak, W. D. Cornell, C. Bayly, and P. A. Kollman. *J. Comput. Chem.*, 16(11):1357–1377, 1995.

[177] G. Mancini, I. D'Annessa, A. Coletta, G. Chillemi, Y. Pommier, M. Cushman, and A. Desideri. *PLoS ONE*, 7(12):1–10, 12 2012.

[178] J. A. Lemkul and D. R. Bevan. *J. Phys. Chem. B*, 114(4):1652–1660, 2010. PMID: 20055378.

[179] F. Yang, S. S. Teves, C. J. Kemp, and S. Henikoff. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1845(1):84 – 89, 2014.

[180] C. Caleman, P. van Maaren, M. Hong, J. S. Hub, L. T. Costa, and D. van der Spoel. *J. Chem. Theory Comput.*, 8(1):61–74, 2011.

[181] V. Barone and M. Cossi. *J. Phys. Chem. A*, 102(11):1995–2001, 1998.

[182] S. Goldman and C. Joslin. *J. Phys. Chem.*, 97(47):12349–12355, 1993.

[183] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives. *J. Am. Chem. Soc*, 118(45):11225–11236, 1996.

[184] I. Cacelli, A. Cimoli, P. R. Livotto, and G. Prampolini. *J. Comput. Chem.*, 33(10):1055–1067, 2012.

[185] J. Pipek and P. G. Mezey. *J. Chem. Phys.*, 90(9):4916–4926, 1989.

[186] J. Hutter, M. Iannuzzi, F. Schiffmann, and J. VandeVondele. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 4(1):15–25, 2014.

[187] A. D Becke. *Phys. Rev. A*, 38(6):3098, 1988.

[188] S. Goedecker, M. Teter, and J. Hutter. *Phys. Rev. B*, 54(3):1703, 1996.

[189] C. Hartwigsen, S. Gœdecker, and J. Hutter. *Phys. Rev. B*, 58(7):3641, 1998.

[190] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. *J. Chem. Phys.*, 132(15):154104, 2010.

[191] M. Brehm and B. Kirchner. *J. Chem. Info. Model.*, 51(8):2007–2023, 2011. PMID: 21761915.

[192] C. M Baker and G. H. Grant. *J. Chem. Theory Comput.*, 3(2):530–548, 2007.

[193] E. F. V. Scriven and R. Murugan. *Kirk-Othmer Encyclopedia of Chemical Technology*, 2005.

[194] Y. Marcus. *The properties of solvents*, volume 16. Wiley Chichester, 1998.

[195] J. Harmsen and J. B. Powell. *Sustainable development in the process industries: cases and impact*. ohn Wiley & Sons, 2011.

[196] F. Ciardelli, G. Ruggeri, and A. Pucci. *Chem. Soc. Rev.*, 42(3):857–870, 2013.

[197] S. K. Yesodha, C. K. S. Pillai, and N. Tsutsumi. *Progress in Polymer Science*, 29(1):45–74, 2004.

[198] G. S. Kumar and D. C. Neckers. *Chem. Rev.*, 89(8):1915–1925, 1989.

[199] A. Natansohn and P. Rochon. *Chem. Rev.*, 102(11):4139–4176, 2002.

[200] S. L. R. Barker, D. Ross, M. Tarlov, M. Gaitan, and L. E. Locascio. *Analytical chemistry*, 72(24):5925–5929, 2000.

[201] K. Komori, H. Matsui, and T. Tatsuma. *Bioelectrochemistry*, 65(2):129–134, 2005.

[202] A. Pucci, R. Bizzarri, and G. Ruggeri. *Soft Matter*, 7(8):3689–3700, 2011.

[203] J. Houghton. *Reports on Progress in Physics*, 68(6):1343, 2005.

[204] W. G. H. M. van Sark. *Renewable Energy*, 49:207–210, 2013.

[205] M. G. Debije and P. P. C. Verbunt. *Adv. Ener. Mater.*, 2(1):12–35, 2012.

[206] F. Bureš. *RSC Advances*, 4(102):58826–58851, 2014.

[207] B. C. Rowan, L. R. Wilson, and B. S. Richards. *IEEE ournal of selected topics in quantum electronics*, 14(5):1312–1322, 2008.

[208] J. Bloino. *J. Phys. Chem. A*, 119(21):5269–5287, 2015.

[209] S. Fantacci, F. De Angelis, A. Sgamellotti, A. Marrone, and N. Re. *J. Am. Chem. Soc.*, 127(41):14144–14145, 2005.

[210] V. Barone, J. Bloino, S. Monti, A. Pedone, and G. Prampolini. *Phys. Chem. Chem. Phys.*, 13(6):2160–2166, 2011.

[211] G. Prampolini, F. Bellina, M. Biczysko, C. Cappelli, L. Carta, M. Lessi, A. Pucci, G. Ruggeri, and V. Barone. *Chem.-Eur. J.*, 19(6):1996–2004, 2013.

[212] N. A. Murugan, R. Apostolov, Z. Rinkevicius, J. Kongsted, E. Lindahl, and H. JÅgren. *J. Am. Chem. Soc.*, 135(36):13590–13597, 2013.

[213] M. Sulpizi, P. Carloni, J. Hutter, and U. Rothlisberger. *Phys. Chem. Chem. Phys.*, 5(21):4798–4805, 2003.

[214] M. Pu and T. Privalov. *Chem.-Eur. J.*, 21(49):17708–17720, 2015.

[215] V. Barone, F. Bellina, M. Biczysko, J. Bloino, T. Fornaro, C. Latouche, M. Lessi, G. Marianetti, P. Minei, A. Panattoni, et al. *Phys. Chem. Chem. Phys.*, 17(40):26710–26723, 2015.

[216] O. B. Malcıoglu, A. Calzolari, R. Gebauer, D. Varsano, and S. Baroni. *J. Am. Chem. Soc.*, 133(39):15425–15433, 2011.

[217] N. De Mitri, G. Prampolini, S. Monti, and V. Barone. *Phys. Chem. Chem. Phys.*, 16(31):16573–16587, 2014.

[218] V. Barone, P. Cimino, and E. Stendardo. *J. Chem. Theory Comput.*, 4(5):751–764, 2008.

[219] V. Barone and P. Cimino. *J. Chem. Theory Comput.*, 5(1):192–199, 2008.

[220] A. Hagfeldt, G. Boschloo, L. Sun, L. Kloo, and H. Pettersson. *Chem. Rev.*, 110(11):6595–6663, 2010. PMID: 20831177.

[221] M. A. Haidekker, A. G. Tsai, T. Brady, H. Y. Stevens, J. A. Frangos, E. Theodorakis, and M. S. Intaglietta. *Am. J. Physiol. Heart Circ. Physiol.*, 282(5):H1609–H1614, 2002.

[222] G. Signore, R. Nifosi, L. Albertazzi, B. Storti, and R. Bizzarri. *J. Am. Chem. Soc.*, 132(4):1276–1288, 2010. PMID: 20050646.

[223] G. Brancato, G. Signore, P. Neyroz, D. Polli, G. Cerullo, G. Abbandonato, L. Nucara, V. Barone, F. Beltram, and R. Bizzarri. *J. Phys. Chem. B*, 119(20):6144–6154, 2015. PMID: 25902266.

[224] M. Gonçalves and Sameiro T. *Chem. Rev.*, 109(1):190–212, 2009.

[225] M. A. Haidekker and E. A. Theodorakis. *J. Biol. Eng.*, 4(1):11, 2010.

[226] C. E. Kung and J. K. Reed. *Biochemistry*, 28(16):6678–6686, 1989. PMID: 2790023.

[227] M. A. H. Alamiry, E. Bahaidarah, A. Harriman, T. Bura, and R. Ziessel. *RSC Adv.*, 2:9851–9859, 2012.

[228] R. W. Sinkeldam, N. J. Greco, and Y. Tor. *Chem. Rev.*, 110(5):2579–2619, 2010. PMID: 20205430.

[229] M. K. Kuimova. *Phys. Chem. Chem. Phys.*, 14:12671–12686, 2012.

[230] 208.

[231] M. Koenig, G. Bottari, G. Brancato, V. Barone, D. M. Guldi, and T. Torres. *Chem. Sci.*, 4:2502–2511, 2013.

[232] M. Koenig, T. Torres, V. Barone, G. Brancato, D. M. Guldi, and G. Bottari. *Chemical Communications*, 50(85):12955–12958, September 2014. 00000.

[233] P. Minei, M. Koenig, A. Battisti, M. Ahmad, V. Barone, T. Torres, D. M. Guldi, G. Brancato, G. Bottari, and A. Pucci. *J. Mater. Chem. C*, 2:9224–9232, 2014.

[234] M. Koenig, B. Storti, R. Bizzarri, D. M. Guldi, G. Brancato, and G. Bottari. *J. Mater. Chem. C*, 4(14):3018–3027, March 2016. 00000.

[235] M. Pavone, G. Brancato, G. Morelli, and V. Barone. *ChemPhysChem*, 7(1):148–156, 2006.

[236] V. Barone, M. Biczysko, and G. Brancato. In ohn R. Sabin and Erkki Brändas, editor, *Advances in Quantum Chemistry*, volume Volume 59, pages 17–57. Academic Press, 2010.

[237] M. W. Mahoney and W. L. Jorgensen. *J. Chem. Phys.*, 112(20):8910–8922, 2000.

[238] P. Tieleman, I. Vattulainen, E. Tajkhorshid, J. W. Essex, M. B. Ulmschneider, K. Schulten, J. L. Robertson, B. Roux, and S. Khalid. *Molecular Simulations and Biomembranes*. RSC Biomolecular Sciences. The Royal Society of Chemistry, 2010.

[239] D. T. Warshaviak, M. J. Muellner, and M. Chachisvilis. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1808(10):2608 – 2617, 2011.

[240] S. Jo, J. B. Lim, J. B. Klauda, and W. Im. *Biophysical ournal*, 97(1):50 – 58, 2009.

[241] G. Del Frate, F. Bellina, G. Mancini, G. Marianetti, P. Minei, A. Pucci, and V. Barone. *Phys. Chem. Chem. Phys.*, 18:9724–9733, 2016.

[242] A. Domenicano and P. Murray-Rust. *Tetrahedron Letters*, 20(24):2283 – 2286, 1979.

[243] I. Reva, L. Lapinski, N. Chattopadhyay, and R. Fausto. *Phys. Chem. Chem. Phys.*, 5:3844–3850, 2003.

[244] M. Falkowska, D. T. Bowron, H. G. Manyar, C. Hardacre, and T. G. A. Youngs. *ChemPhysChem*, 17(13):2043–2055, 2016.

[245] A. V. Anikeenko, A. V. Kim, and N. N. Medvedev. *J. Struct. Chem.*, 51(6):1090–1096, Dec 2010.

[246] R. Bizzarri, personal communications.

[247] A. Kessel, N. Ben-Tal, and S. May. *Biophys. J.*, 81(2):643 – 658, 2001.

[248] G. Moumouzias, D. K. Panopoulos, and G. Ritzoulis. *J. Chem. Eng. Data*, 36(1):20–23, 1991.

[249] D. S. Gill, J. Singh, R. Ludwig, and M. D. Zeidler. *J. Chem. Soc. Faraday Trans.*, 89(21):3955–3958, 1993.

[250] J. G. Speight et al. *Lange's handbook of chemistry*, volume 1. McGraw-Hill New York, 2005.

[251] C. Wohlfarth. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[252] Y. Wu, M. Stefl, A. Olzynska, M. Hof, G. Yahioglu, P. Yip, D. R. Casey, O. Ces, J. Humpolickova, and M. K. Kuimova. *Phys. Chem. Chem. Phys.*, 15:14986–14993, 2013.

[253] R. Kumar and S. S. Sekhon. *Ionics*, 14(6):509, 2008.

[254] P. J. W. Debye. *Polar molecules*. Chemical Catalog Company, Incorporated, 1929.

[255] T. Förster and G. Hoffmann. *Zeitschrift fr Physikalische Chemie*, 75:63 – 76, 1971.

[256] M. A. Haidekker, T. P. Brady, D. Lichlyter, and E. A. Theodorakis. *Bioorganic Chemistry*, 33(6):415 – 425, 2005.

[257] S. A. Rice and G. A. Kenney-Wallace. *Chem. Phys.*, 47(2):161–170, April 1980. 00073.

[258] A. C Rosenzweig. *Chem. Biol.*, 9(6):673–677, 2002.

[259] P. Li, B. P. Roberts, D. K. Chakravorty, and K. M. Merz. *J. Chem. Theory Comput.*, 9(6):2733–2748, 2013.

[260] M. B. Peters, Y. Yang, B. Wang, L. Fusti-Molnar, M. N. Weaver, and K. M. Merz. *J. Chem. Theory Comput.*, 6(9):2935–2947, 2010.

[261] F. Rose, M. Hodak, and J. Bernholc. *Sci. Rep.*, 1:11, 2011.

[262] F. Duarte, P. Bauer, A. Barrozo, B. A. Amrein, M. Purg, J. Åqvist, and S. C. L. Kamerlin. *J. Phys. Chem. B*, 118(16):4351–4362, 2014.

[263] R. H. Stote and M. Karplus. *Proteins: Struct., Funct., Bioinf.*, 23(1):12–31, 1995.

[264] K. P. Jensen and W. L. Jorgensen. *J. Chem. Theory Comput.*, 2(6):1499–1509, November 2006.

[265] T. O. Wambo, L. Y. Chen, S. F. McHardy, and A. T. Tsin. *Biophys. Chem.*, 214:54–60, 2016.

[266] R. Deeth, A. Anastasi, C. Diedrich, and K. Randell. *Coord. Chem. Rev.*, 253(5):795–816, 2009.

[267] P. Li and K. M. Merz. *J. Chem. Theory Comput.*, 10(1):289–297, 2013.

[268] R. Wu, Z. Lu, Z. Cao, and Y. Zhang. *J. Chem. Theory Comput.*, 7(2):433–443, 2010.

[269] P. Li and K. M. Merz. *Chem. Rev.*, 117(3):1564–1686, 2017.

[270] J. Aqvist. *J. Phys. Chem.*, 94(21):8021–8024, 1990.

[271] R. W. Zwanzig. *J. Chem. Phys.*, 22(8):1420–1426, 1954.

[272] C. S. Babu and C. Lim. *J. Phys. Chem. A*, 110(2):691–699, 2006.

[273] I. S. Joung and T. E. Cheatham III. *J. Phys. Chem. B*, 112(30):9020–9041, 2008.

[274] P. Li, L. F. Song, and K. M. Merz. *J. Phys. Chem. B*, 119(3):883–895, 2014.

[275] F. Floris, M. Persico, A. Tani, and J. Tomasi. *Chem. Phys. Lett.*, 199(6):518–524, 1992.

[276] T. A. Feo and M. G. C. Resende. *J. Glob. Optim.*, 6(2):109–133, 1995.

[277] R. H. Swendsen and J. S. Wang. *Phys. Rev. Lett.*, 57(21):2607–2609, 1986.

[278] U. H. E. Hansmann. *Chem. Phys. Lett.*, 281(1):140–150, 1997.

[279] Y. Sugita and Y. Okamoto. *Chem. Phys. Lett.*, 314(1):141–151, 1999.

[280] B. Hess, C. Kutzner, D. Van Der Spoel, and E. Lindahl. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[281] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. *J. Chem. Phys.*, 79(2):926–935, 1983.

[282] S. C. Hoops, K. W. Anderson, and K. M. Merz. *J. Am. Chem. Soc.*, 113(22):8262–8270, 1991.

[283] A. Patriksson and D. van der Spoel. *Phys. Chem. Chem. Phys.*, 10(15):2073–2077, 2008.

[284] C. H. Bennett. . *Comput. Phys.*, 22(2):245 – 268, 1976.

[285] Y. Marcus. *J. Chem. Soc., Faraday Trans.*, 87(18):2995–2999, 1991.

[286] E. Cauët, S. Bogatko, J. H. Weare, J. L. Fulton, G. K. Schenter, and E. J. Bylaska. *J. Chem. Phys.*, 132(19):194502, 2010.

[287] G. Chillemi, P. D'Angelo, N. V. Pavel, N. Sanna, and V. Barone. *J. Am. Chem. Soc.*, 124(9):1968–1976, 2002.

[288] J. C. Wu, J. P. Piquemal, R. Chaudret, P. Reinhardt, and P. Ren. *J. Chem. Theory Comput.*, 6(7):2059–2070, 2010.

[289] S. Riahi, B. Roux, and C. N. Rowley. *Can. J. Chem.*, 91(7):552–558, 2013.

[290] M. T. Panteva, G. M. Giambaşu, and D. M. York. *J. Comput. Chem.*, 36(13):970–982, 2015.

[291] S. Bogatko, E. Cauët, E. Bylaska, G. Schenter, J. Fulton, and J. Weare. *Chem.-Eur. J.*, 19(9):3047–3060, 2013.

[292] S. S. Azam, Z. Ul-Haq, and M. Q. Fatmi. *J. Mol. Liq.*, 153(2):95–100, 2010.

[293] A. Bankura, V. Carnevale, and M. L. Klein. *J. Chem. Phys.*, 138(1):014501, 2013.

[294] A. Bankura, V. Carnevale, and M. L. Klein. *Mol. Phys.*, 112(9-10):1448–1456, 2014.