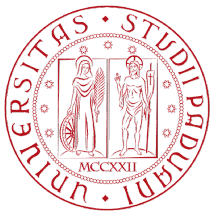


UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Dipartimento di Scienze Chimiche

Master Degree in Physics

Final Dissertation

Development of a machine learning potential for nucleotides in water

Thesis supervisor

Prof. Alberta Ferrarini

Thesis co-supervisor

Prof. Marialore Sulpizi

Candidate

Riccardo Martina

Academic Year 2019/2020

Abstract

Recent experimental research showed that nucleotides, under favorable conditions of temperature and concentration, can self-assemble into liquid crystals. The mechanism involves the stacking of nucleotides into columnar aggregates. It has been proposed that this ordered structure can favor the polymerization of long nucleotide chains, which is a fundamental step toward the so called "RNA world". In this thesis, starting from ab initio molecular dynamics simulations, at the density functional theory level, an all-atom potential for nucleotides in water, based on an implicit neural network representation, has been developed. Its stability and accuracy have been tested and its predictions on simple model systems have been compared with data generated both ab initio and using currently available empirical force field for nucleic acids.

Contents

Introduction

The origin of life [1] is one of the most challenging scientific questions of all times. The different time and length scales of the myriad of physico-chemical processes involved, ranging from the chemical reactions between small organic compounds, up to the supramolecular aggregation of complex biological structures, make it a formidable problem to tackle.

One of the open questions in the field regards the polymerization of long nucleic acid chains, like DNA and RNA strands, that seems to be very unfavorable in prebiotic conditions. Recently a series of experiments [2], [3], [4] have shown that ultrashort DNA oligomers and even free nucleosides can form ordered liquid crystal phases in solution. These ordered phases seem to be an interesting candidate as an environment for abiotic polymerization of long nucleotide chains [5] : a fundamental step toward the formation of the so called RNA world. The liquid crystal order could act as a guiding hand for the polymerization process, favoring the formation of long chains capable of biological activity.

The details of the structure of these aggregates at molecular level is difficult to obtain experimentally. Detailed insights can be obtained using molecular dynamics, a technique that consists in simulating the time evolution of a molecular system integrating classical equations of motion of the atomic nuclei. This requires the calculation of forces, that can be obtained from the electronic structure or by classical empiric potentials. The length and time scales involved in supramolecular aggregation place the present problem outside the domain of quantum mechanical calculations. Moreover, although accurate classical potentials have been developed for describing DNA and RNA polymers [6], they may not be transferable to single nucleosides in solution. In this thesis a first attempt to develop a model based on neural networks to calculate energies and forces for nucleosides in solution is presented. The model is trained on reference ab initio data and aims to bridge the gap between quantum accuracy and large-scale, long-time dynamics.

The neural network potential has been tested on three model systems: pure water, free dAMP in water and dAMP dimer. These systems have been studied also classically, using a state of the art force field, and ab initio, using density functional theory. The results obtained with these three methods are compared in order to assess the capability of the neural network potential of reproducing ab initio results.

The thesis is structured as follows.

- Chapter one is dedicated to a review of physical and chemical properties of nucleosides, in particular their capability of forming liquid crystal phases. The experimental results [2], [3], [4], are discussed.
- Chapter two presents the methods employed, namely: the general framework of molecular dynamics, classical force fields, density functional theory and neural network potentials.
- Chapter three describes the network training, along with the generation of the training set from ab initio calculations. The stability of the neural network model is discussed and its error is quantified. Relevant choices of parameters for the molecular dynamics simulations are also presented.
- Chapter four presents a comparison between molecular dynamics trajectories obtained ab initio, with classical force field, and with the neural network potential.

Chapter 1

Nucleoside phosphates: chemical properties and liquid crystal assemblies

1.1 Nucleoside phosphates: structure and nomenclature

Nucleoside phosphates [7] are molecules composed of a nucleobase linked to a sugar by the so called glycosidic bond, in turn linked to a phosphate group (that can be mono, di or triphosphate). The nucleobases that are present in biological systems are five: Guanine (G), Cytosine (C), Adenine (A), Timine (T) and Uracil (U). These are divided into two groups: pyridines (C, T, U), with one aromatic ring, and purynes (G, A), with two aromatic rings (figure 1.1). Nucleoside monophosphates are

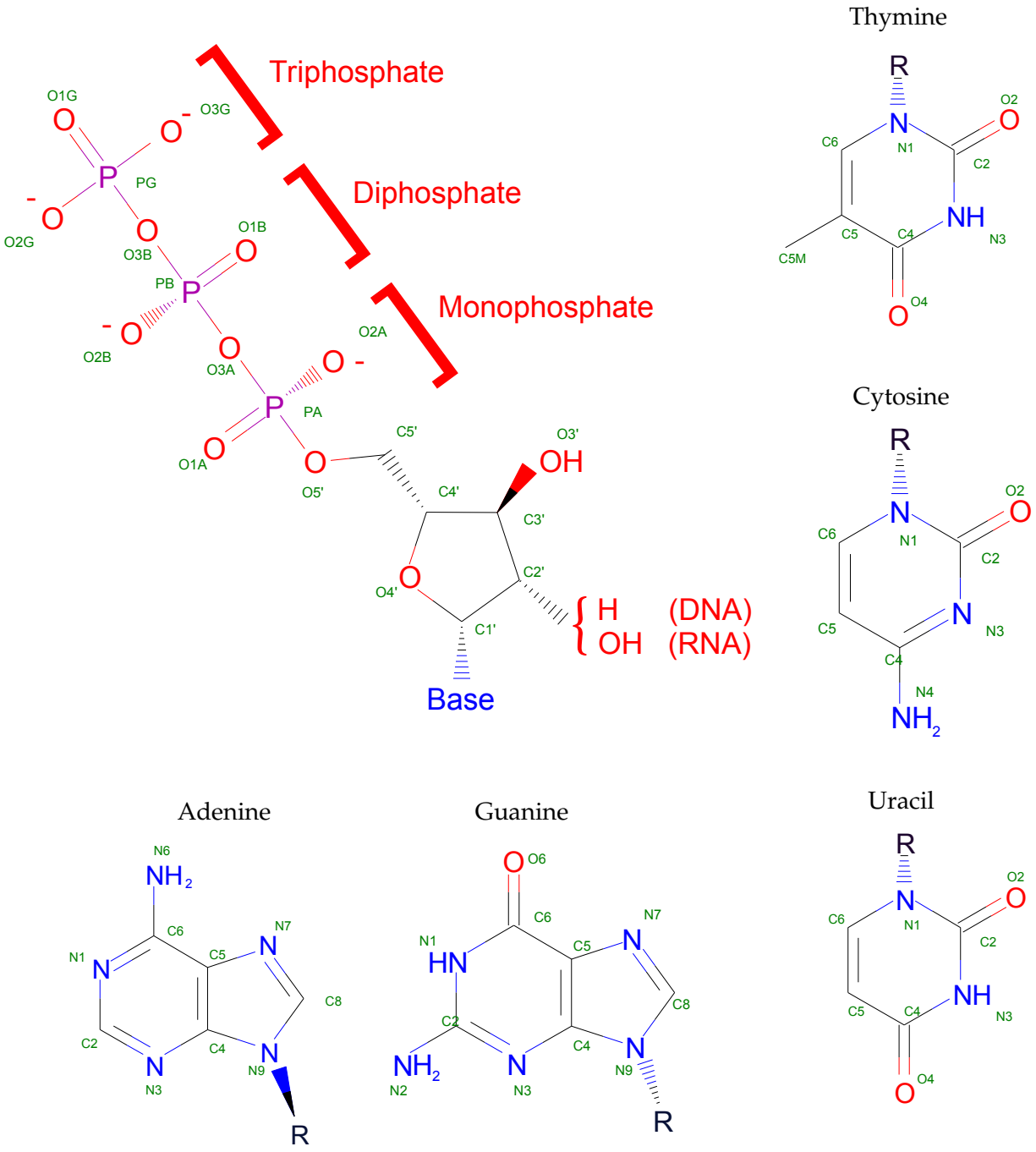


Figure 1.1: Structural formulas of nucleoside phosphates. Green labels are the naming conventions adopted in this work. Structural formulas and naming conventions are taken from the PDB database.

usually called nucleotides. Another important distinction comes from the sugar that is bound to the nucleobase: it can be ribose or deoxyribose. Nucleosides with ribose are called ribo-nucleosides, with deoxyribose deoxy-nucleosides. For sake of clarity the following abbreviations are used:

- NMPs, NDPs, NTPs (nucleoside mono, di and triphosphates) when speaking about different

nucleosides, in order to mark the difference; the word nucleotides will be used when only nucleotides are taken into accont;

- the abbreviations XMP, XDP, XTP, with X substituted by one of the five letters A, T, C, G, U, when speaking about a specific nucleoside phosphate;
- each abbreviation can be preceded by an *r* or a *d* to mark the difference between *r*ibonucleosides and *d*eoxyribonucleosides

For example, dADP denotes deoxy-adenosine-diphosphate, rNTPs the general class of ribo-nucleoside-triphosphates.

In the following sections some properties of nucleoside phosphates that are relevant in the context of this work will be pointed out.

1.2 Chemical and physical properties of nucleoside phosphates

Nucleoside phosphates display a rich variety of chemical properties and supramolecular interactions [7]. Nucleobases are capable of forming hydrogen bonds with complementary ones (A with T and U, G with C), a phenomenon called pairing; aromatic rings can pile together forming stacking complexes; the phosphate group is highly charged, it has a complex solvation shell and can bind cations; the molecule itself, being composed by three different parts (base, sugar and phosphate) has a certain internal flexibility, and the internal torsions are very important, for example, in determining the mechanical properties of the DNA double helix.

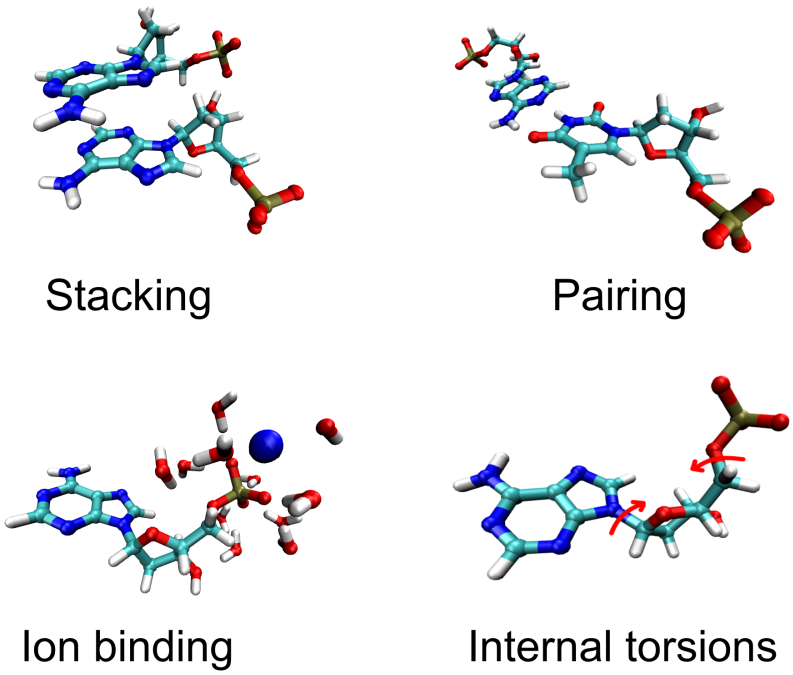


Figure 1.2: Most relevant interactions in nucleotides.

1.2.1 Phosphate group: acidity and phosphodiester bond

The phosphate acid dissociation constant is so low that it can be considered completely ionized at all pH values of interest [7]. The phosphorus atom in nucleotides is highly electrophilic and is prone to react with nucleophiles such as the hydroxyl group. The reaction between the phosphate group and the hydroxyl group of a second nucleotide produces the so called phosphodiester bond. Phosphodiester bonds can link many nucleotides together, forming a polymer (like the well known DNA

strand). In a nucleotide polymer the series of bonded phosphate groups and sugar rings is called backbone.

1.2.2 H bonding and W-C Pairing

The most known supramolecolar interaction between nucleobases is the hydrogen bonding. There are many hydrogen bonding configurations between nucleobases that are, in principle, possible [7]. These configurations are classified according to the edges that participate in H bonding. The nucleobases can form hydrogen bonds along three interaction edges that are called sugar, Hoogsten and Watson-Crick (W-C) edges, as shown in figure 1.3.

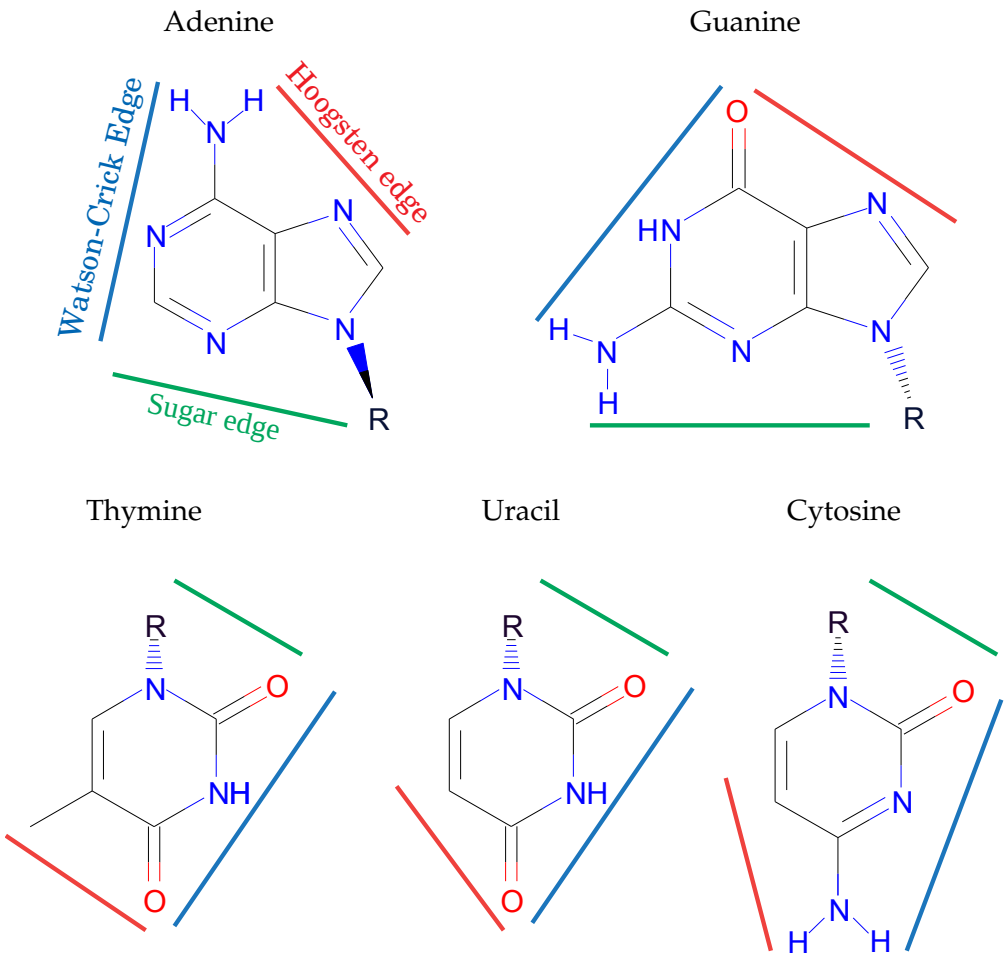


Figure 1.3: Hydrogen bond edges of the five nucleobases.

The most common configuration is the one between W-C edges: this is the configuration that is found in the double helix of the DNA. Nonetheless also the other interactions have their importance, as they were found to stabilize certain RNA conformations or more exotic DNA structures.

1.2.3 Aromatic stacking

Aromatic or π stacking are terms usually employed to name the interaction between aromatic rings. Stacking is the result of the interplay between three contributions: electrostatic interaction, short range exchange repulsion and dispersion interaction. The stacking interaction has long been known, but its strength and its contribution to the stability of DNA helices has not been understood until recently, mainly because the contribution of dispersion forces was not taken into account adequately [8]. The earliest model for the description of π stacking assumed it to be principally due to electrostatic interaction between the quadrupoles possessed by aromatic rings [9]. In the last years, as quantum mechanical calculations have become more capable of describing dispersion forces, their role in stacking interaction has been the subject of a large number of works. Many evidences suggest that this contribution can be even stronger than the quadrupole one [10].

1.2.4 H bond versus stacking

The relative importance of H bonding and aromatic stacking in the stabilization of the DNA double helix has been matter of debate for a long time. Both experimental and computational studies faced many difficulties in separating these energy contributions. In the last two decades, starting from the seminal work by Yakovchuk et. al. [11] the consensus has oriented towards considering the stacking interaction as the most important. Moreover some computational evidences suggest that the double helix is the preferred conformation of the DNA double strand because of the presence of aromatic stacking: without it the conformation would be ladder-like [12].

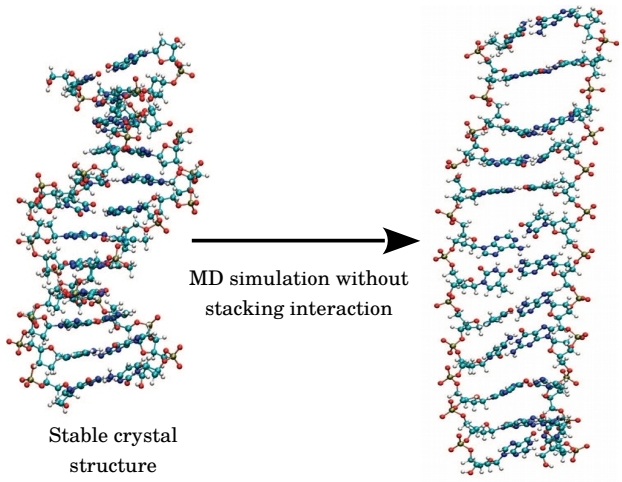


Figure 1.4: Change in DNA structure after artificial elimination of stacking interaction. (adapted from [12]).

1.2.5 Stacking aggregates in aqueous solution

It is known that free nucleoside phosphates in solution tend to form small stacking aggregates. Two seminal experimental works were carried out in the '60 using dNMPs [13] [14]
Stacking of dATP was subsequently observed and studied [15] [16]
In [17] values are proposed for the equilibrium constant of dimerization K_{dim} of dATP and dADP, in different conditions. At pH=6.9 the values proposed are:

$$\begin{array}{ll} \text{dATP} & K_{dim} = (8 \pm 6) \times 10^3 \\ \text{dADP} & K_{dim} = (5 \pm 3) \times 10^3 \end{array}$$

According to tese values even at low concentration nearly all the molecules are associated in stacking dimers: for example at an dATP concentration of 5×10^{-3} mol/L the 98.75% of the molecules is associated in stacking.

1.3 Prebiotic synthesis of nucleic acids

Nucleotides are the monomers of nucleic acids RNA and DNA. The ubiquity of RNA in many cellular function and the diversity of the roles that it fulfills (information carrying, protein synthesis, catalysis) led in the '60s to the so called "RNA world" hypothesis [18]: the idea is that RNA could have played a main role in the early stages of life origin, functioning both as genetic material and as a catalyst. The discovery in the '80s of RNA complexes with self-replication capacity has made this hypothesis the most widely accepted in the scientific community. Nonetheless many problems are still open: how did nucleotides assemble from their different molecular parts (ribose, phosphate, nucleobases) under the physical conditions of the earth billions of years ago? Then, after their synthesis, what mechanism made it possible the formation of the long chains that are capable of self-replication? This second question is particularly interesting from a soft-matter point of view. The smallest known rybozimes are around a hundred bases long [19]. Free polymerization reactions are known to give only small chains for entropic reasons: the resulting length distribution of the chains is a power law

with many short chains and few longer ones, with negligible probability of having chains with a biologically relevant length. So, what mechanism favored the formation of long RNA chains? In the last decades the research in the field has been focused on finding some autocatalytic process in which lengthening of the chains favors subsequent lengthening, in a positive feedback loop. A possible mechanism may be based on the capability of nucleosides to form ordered structures known as liquid crystals [5].

1.4 Liquid crystal phases

Liquid crystals (LCs) are intermediate phases between liquids and solids. They are fluids, in the sense that they have the mechanical properties of a fluid, but they possess long-range molecular order, like crystals. The basic building blocks of LCs are called mesogens: they can be simple molecules, complex supramolecular aggregate or even bigger objects, like the tobacco mosaic virus.

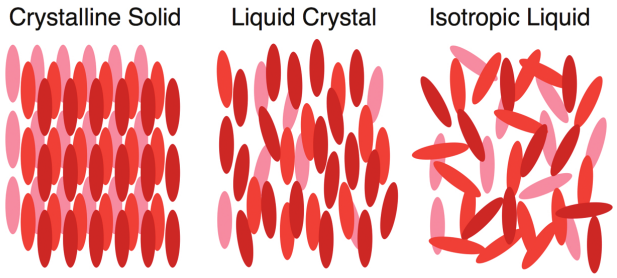


Figure 1.5: Liquid crystals are intermediate phases between isotropic liquids and crystalline solids.

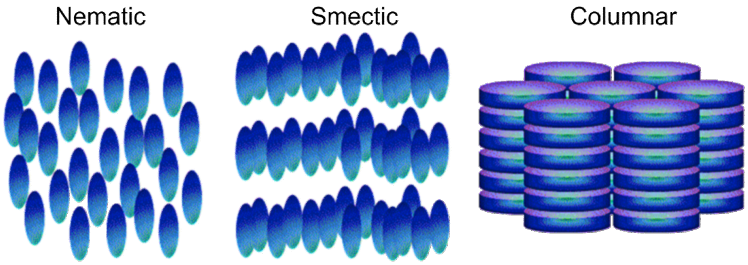


Figure 1.6: Classification of liquid crystals based on their long-range order.

From a structural point of view LCs can be classified by the type of ordering they possess. When mesogens are elongated molecules, such that a director vector $\mathbf{d}$ can be assigned to each one, their degree of alignment can be quantified by an order parameter. LCs that posses long range orientatonal order, i.e. with mesogens that are all meanly aligned along a common direction, are called nematic. LCs that posses orientatonal and two dimensional positional order, i.e. that form ordered layers, are called smectic.

As already said, mesogens can be single molecules or bigger aggregates. In columnar LCs molecules assemble into cylindrical structures that act as mesogens. Columnar phases can be nematic, with no other order than the mean orientation, or can can have a two dimensional lattice order in the plane perpendicular to the director.

1.4.1 Nucleoside phosphates LCs

Liquid crystal phases are frequently found in living systems. Cell membranes, for example, posses liquid crystal character and also DNA forms LC phases: the experiments that led to the discovery of the double helix structure in 1953 were made on DNA LCs. LC phases of DNA can be found also in vivo, for example in some bacteriophage, in bacteria and also in human sperms nuclei [20]. Recent experimental work has led to discover that also ultra-short DNA [2]and RNA [3] oligomers (4

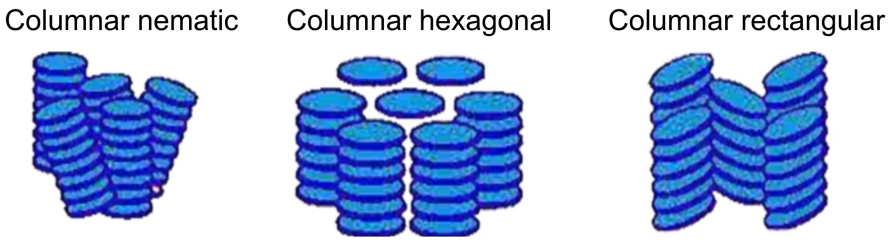


Figure 1.7: Different columnar liquid crystals.

nucleotites pairs) can form LC phases. Furthermore it has been shown [5] that such LC domains can act as templates in which the chemical ligation is enhanced by the physical supramolecular ordering. The authors of [5] propose a positive feedback coupling between self-assembly of LC domains and ligation, with the ligation stabilizing the LC domains and the LC domains favoring ligation: they named this mechanism "liquid crystal autocatalysis". This autocatalytic cycle favors the growth of DNA chains, up to biologically relevant lengths. In this hypothesis the LC acts as a guiding hand for the polymerization process, determined uniquely by the physical properties of nucleic acids themselves. Recently it has been shown that also single backbone-free nucleoside triphosphates [4] have

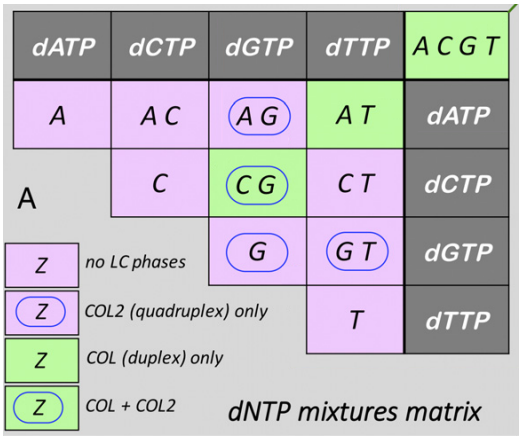


Figure 1.8: Table showing the NTPs combinations that produce different phases. Two LC phases can form: COL, which mesogens are of stacked base pairs, and COL2, composed by G-quadruplex, a supramolecular aggregate formed by guanine only. Reproduced from [4].

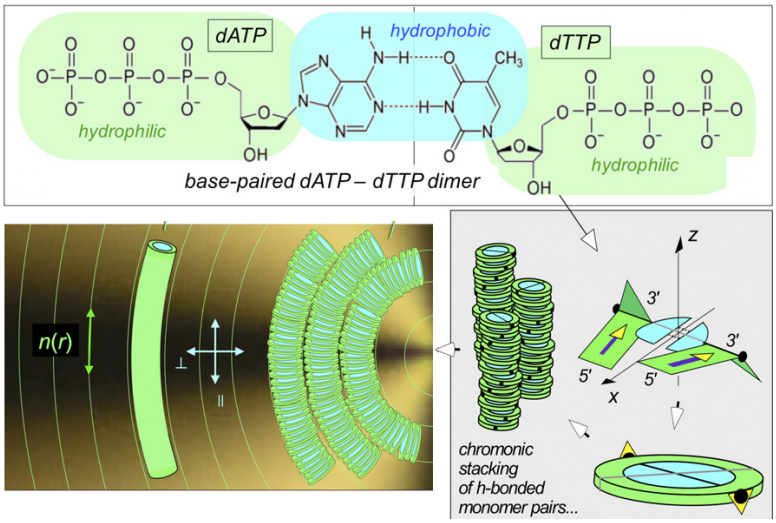


Figure 1.9: Hierarchy self-assembly mechanism proposed in [4]. (Image adapted from [4]).

the capability of forming LC domains. Furthermore WC selectivity seems to play an important role

in the formation of the LC phase: mixtures of non complementary nucleosides do not form LCs, as shown in figure 1.8. Based on this observation a simple hierarchical self-assembly mechanism has been proposed by the authors: the hydrogen bonding between base pairs reduces their solubility, promoting the chromonic stacking of bases, as shown in figure 1.9. Experiments on RNA and ribo-nucleosides have not been carried out at the moment, but it seems legitimate to expect that also ribo-nucleosides can display similar behaviors. Research is still ongoing.

1.5 Open problems

Many questions raised by the experimental research need to be addressed trough simulations. Aspects that remain to be unveiled are:

- the role of pairing and stacking and the mechanism of their mutual collaboration in forming the LC domains;
- the role of the phosphate group;
- the role of the ions in solution;
- whether helical structures already exist in LC phase.

This thesis is the first step toward a systematic study of LC aggregation of nucleosides by means of molecular dynamics, using recently developed machine learning techniques. The work is focused on the study of backbone-free nucleotides in water. Nucleotides have been chosen as model systems, instead of nucleoside triphosphates, in order to limit computational complexity, with the aim of extending the study to nucleoside di and triphosphates in the future.

Chapter 2

Computational methods

In this chapter an overview of the computational methods employed is provided. The general framework of Molecular Dynamics is introduced, along with the umbrella sampling technique; classical Force Fields are briefly described; DFT is revised with a focus on the implementation in the package CP2K; the machine learning methods employed for the training of the neural network potential are described.

2.1 Thermodynamic sampling

Consider a system of N atoms, described by the spatial coordinates $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2 \dots \mathbf{r}_N) \in \mathbb{R}^{dN}$, with associated momenta $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2 \dots \mathbf{p}_N) \in \mathbb{R}^{dN}$. At each point $(\mathbf{r}, \mathbf{p})$ in the $6N$ dimensional phase space a probability density $\rho(\mathbf{r}, \mathbf{p})$ can be associated, with the meaning:

$$\int_\Omega \rho(\mathbf{r}, \mathbf{p}) d^{3N}\mathbf{r} d^{3N}\mathbf{p} = \text{Probability of finding the system in a state within the region } \Omega \tag{2.1}$$

If ρ is constant in time then the system is said to be at equilibrium. A fundamental assumption of statistical mechanics is that, for an isolated system (where isolated means most of the times with constant energy, volume and number of particles) ρ is constant over all the accessible states (i.e. states for which $\mathcal{H}(\mathbf{r}, \mathbf{p}) = E$, where $\mathcal{H}$ is the system's Hamiltonian). A system with fixed energy, volume and number of particles is said to be in the microcanonical (NVE) ensemble.

From the assumption of equal probability in the microcanonical ensemble it can be derived that, for a system in diathermal contact with an ideal heat bath:

$$\rho(\mathbf{r}, \mathbf{p}) = \frac{1}{Z} e^{-\beta \mathcal{H}(\mathbf{r}, \mathbf{p})} \qquad\qquad Z = \int e^{-\beta \mathcal{H}(\mathbf{r}, \mathbf{p})} d^{3N}\mathbf{r} d^{3N}\mathbf{p} \tag{2.2}$$

This system is kept at constant volume, temperature and number of particles, and is said to be in the canonical (NVT) ensemble.

Since all the quantities of interest for the thesis do not depend on momenta, the following definition in configuration space will be adopted. Given an Hamiltonian that can be written as a sum of kinetic and potential terms $\mathcal{H}(\mathbf{r}, \mathbf{p}) = T(\mathbf{p}) + V(\mathbf{r})$, the canonical probability density in configuration space is expressed as:

$$\rho(\mathbf{r}) = \frac{1}{Z} e^{-\beta V(\mathbf{r})} \qquad\qquad Z = \int e^{-\beta V(\mathbf{r})} d^{3N}\mathbf{r} \tag{2.3}$$

Where $V(\mathbf{r})$ is the potential energy of the system, function of the N particle positions. The distribution $\rho(\mathbf{r})$ usually has this remarkable property: it is negligible in the vast majority of the configuration space except for small, circumscribed regions around the potential energy minima.

2.1.1 Metastable states, collective variables

It can happen that a system displays more potential energy minima separated by barriers, so high that its dynamics is dominated by long periods of time in which it remains in the close vicinity of one minimum, separated by fast transitions between them, during which the system crosses a low probability region. This is usually the case for molecular systems. When dealing with such multiple minima one might adopt a description based on collective variables (CV), namely, functions of the coordinates that take different values in the relevant metastable states. For example, for a chemical reaction in which a certain bond could break, an good candidate as CV would be the distance between the two atoms forming the bond.

Let the vector of *collective variables* be $\boldsymbol{\xi} = (\xi_1(\mathbf{r}), \xi_2(\mathbf{r}) \dots \xi_M(\mathbf{r})) \in \mathbb{R}^M$ with M<dN. The probability density in the $\boldsymbol{\xi}$ space is simply given by

$$\rho(\boldsymbol{\xi}) = \frac{1}{Z} \int \delta\big(\boldsymbol{\xi} - \boldsymbol{\xi}(\mathbf{r})\big) e^{-\beta V(\mathbf{r})} d^{3N}\mathbf{r} \tag{2.4}$$

This allows the definition of a free energy function associated to the collective variables:

$$U(\boldsymbol{\xi}) = -\frac{1}{\beta}\log\rho(\boldsymbol{\xi}) \tag{2.5}$$

The free energy $U(\boldsymbol{\xi})$ is completely defined by the choice of collective variables and the underlying potential energy. If the CVs are properly chesen, the free energy should have at least two dominating minima, corresponding to the metastable state that one wants to study.

2.1.2 Equilibrium averages

Given an observable $A(\mathbf{r}, \mathbf{p})$ its equilibrium average is defined as:

$$\langle A\rangle_\rho = \int \rho(\mathbf{r},\mathbf{p})A(\mathbf{r},\mathbf{p})d^{3N}\mathbf{r}d^{3N}\mathbf{p} \tag{2.6}$$

For a moment-independent observable it becomes:

$$\langle A\rangle_\rho = \int \rho(\mathbf{r})A(\mathbf{r})d^{3N}\mathbf{r} \tag{2.7}$$

When dealing with systems for which no analytical form of ρ is available the only possibility to compute averages is to explore the phase space numerically, with a sampling algorithm that ensures to visit each state with the correct probability, at least in the infinite time limit. Let $p(\mathbf{r}, \mathbf{p}, n)$ be the probability of the sampling algorithm to visit the state $(\mathbf{r}, \mathbf{p})$ at the timestep n. It is desirable that:

$$p(\mathbf{r},\mathbf{p},n) \xrightarrow[n\to\infty]{} \rho(\mathbf{r},\mathbf{p}) \tag{2.8}$$

If a suitable algorithm is available then a straightforward recipe to compute equilibrium averages is:

- Run the algorithm until a step n_{eq}, to let its probability p converge to the correct ρ.
- Run the algorithm until a step n_{end}, long enough to ensure a meaningful sampling of ρ.
- Compute: $\langle A\rangle = \dfrac{1}{n_{end}-n_{eq}}\displaystyle\sum_{i=n_{eq}}^{n_{end}} A(\mathbf{r}_i,\mathbf{p}_i)$

An algorithm suitable for this task, employed in the present work, is the molecular dynamics.

2.2 Molecular Dynamics

Molecular dynamics (MD) is a method of simulating molecular systems by solving Newton's equations of motion of the N atoms constituting the system:

$$m_i\frac{d^2\mathbf{r}}{dt^2} = \mathbf{f}_i(\mathbf{r}_1,\mathbf{r}_2\ldots\mathbf{r}_N) \tag{2.9}$$

where i is the particle index, m_i the mass of the $i-th$ particle, and f_i the force on it. Particles may be atoms or group of atoms in a coarse grained representation. The trajectory of the system is computed by numerical integration in discretized time domain: at each time step Δt the forces are computed and the the positions updated according to their values. Thus the main tasks to carry out during an MD simulation are force calculation and time integration.

2.2.1 Force calculation

MD is versatile technique that can be scaled to different sized systems, computing forces at different levels of physical accuracy:

- Ab Initio Molecular Dynamics (AIMD) is the most fundamental level: forces are computed solving the many body Schrödinger equation. This is the most reliable way to calculate energy and forces of a molecular system, but it is computationally demanding.
- Semiempirical methods are based on quantum mechanical formalism, but use many approximations and some parameters can be obtained from experimental data.
- Force Fields (FF) are sets of purely classical parameters, usually derived from ab initio reference calculations and experimental data, defining analytical interactions potentials between atoms.

AIMD and FF methods employed for this thesis are briefly reviewed in the next sections.

2.2.2 Time integration

MD is usually employed to extract statistical properties, like free energy and correlation functions; this means that the sampling of the correct thermodynamic ensemble is more important than the accurate calculation of the trajectories. Thus an integration technique suitable for MD should conserve the Hamiltonian: this ensures time reversibility of the algorithm and conservation of the probability in phase space. The family of methods that do this are called symplectic integrators [21].

All the simulations carried out for this thesis use the velocity-Verlet algorithm [22].

$$\begin{aligned} \mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\,\Delta t + \frac{1}{2}\,\mathbf{a}_i(t)\Delta t^2 \\ \mathbf{v}_i(t+\Delta t) &= \mathbf{v}_i(t) + \frac{\mathbf{a}_i(t)+\mathbf{a}_i(t+\Delta t)}{2}\Delta t \\ \mathbf{a}_i(t) &= \frac{1}{m_i}\mathbf{f}_i(\mathbf{r}_1(t), \mathbf{r}_2(t)\ldots\mathbf{r}_N(t)) \end{aligned} \tag{2.10}$$

Actually symplectic integrators do not conserve the real Hamiltonian of the system, but a slight perturbed version of it, so there is always a little error in the energy of a long MD simulation [21]. The error depends on the choice of the integration timestep: discrete timestepping leads to nonphysical resonance of degrees of freedom with typical frequencies close to the frequency of velocity updates. Thus the the maximum step that can be chosen is limited by the period of the fastest fundamental modes of the system under study.

Degree of freedom	Period [fs]
O-H, N-H stretch	9.8
C-H stretch	11.1
C=O stretch	19.6
H-O-H bend	20.8
C=C (aromatic) stretch	22.2

Table 2.1: Period of oscillations of some bonds that are present in nucleotides and water. Values from [21].

A usual rule of thumb is to take

$$\Delta t \simeq \frac{T_{fast}}{10}$$

where T_{fast} is the period of the fastest oscillation in the system. The fastest oscillation in organic molecules are usually due to bonds involving hydrogen atoms, as shown in table (2.1). The usual timestep employed in a MD simulation involving hydrogen atoms is $\Delta t = 0.5$ fs.

Sometimes constraint algorithms are employed to freeze some degrees of freedom, i.e. to keep bond lengths fixed at their equilibrium values, and larger timesteps (up to 2-5 fs) can be employed [21].

MD can reproduce system's dynamics faithfully and thus it provides a straightforward way to sample its probability density ρ. It has nonetheless a series of problems.

- Bulk systems are macroscopic: they contain a number particles of the order of the Avogadro number: simulating a whole bulk system is obviously impossible. The solution is to run the simulation under periodic boundary condition, in order to mimic an infinite system. When studying molecules in solution one has to choose a unit cell large enough to ensure that the "interesting" part of the system does not interact directly with its periodic image.
- The plain solution of the equations of motion gives an energy conserving trajectory, or, in a statistical mechanics language, a sampling in the microcanonical ensemble. To sample different ensembles, a coupling with a fictious external system (the reservoir) has to be introduced. All the simulations for this thesis were carried out in the canonical ensemble, so thermostat algorithms were employed.
- The computation of a MD trajectory can be too slow to accumulate a significant data in a real-world lapse of time, especially when dealing with metastable states. This problem can be circumvented using enhanced sampling techniques.

2.2.3 Thermostats

Thermostats are computational algorithms that couple MD calculation of the trajectories with an external heath bath. Denote with T_0 the target temperature and with $T(t)$ the instantaneous temperature of the system:

$$T(t) = \frac{1}{3Nk_B} \sum_i^N \frac{p_i^2}{m_i} \tag{2.11}$$

A thermostat that was used in the computations for this thesis is the Nosé-Hoover thermostat [23], in which the equation of motions are modified in this way:

$$m_i \frac{d^2\mathbf{r}}{dt^2} = \mathbf{f}_i(\mathbf{r}_1, \mathbf{r}_2 \dots \mathbf{r}_N) - \gamma_i(t)\mathbf{v}_i \qquad \frac{d\gamma(t)}{dt} = \frac{1}{Q}\big(3Nk_BT(t) - gT_0\big) \tag{2.12}$$

The rationale behind Nosé-Hoover scheme is to couple the degrees of freedom of the system to a fictious external degree of freedom. Nosé-Hoover equations of motions can be derived from the so called extended Hamiltonian, an Hamiltonian function containing the fictious degree of freedom. It can be then shown that Nosé-Hoover equations of motion give trajectories that are microcanonical for the extended system and canonical for the system.

Another thermostat that has been employed is the Langevin thermostat, which, in contrast to the deterministic Nosé-Hoover, includes a stochastic term. At each time step a stochastic and a friction term are added to the force:

$$\mathbf{f}_i(t) = \mathbf{f}_i^0(t) - \gamma\mathbf{v}_i(t) + \mathbf{\Phi_i} \tag{2.13}$$

where $\mathbf{f}_i^0(t)$ is the force due to the internal interactions and $\mathbf{\Phi_i}$ is a random vector sampled from a distribution that is gaussian for the modulus:

$$p\big(|\mathbf{\Phi_i}|\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{|\mathbf{\Phi_i}|^2}{2\sigma^2}\right\} \tag{2.14}$$

and uniform for the direction. The standard deviation and the friction coefficient are linked together by the equation $\sigma^2 = 2\gamma m_i k_B T$. With this choice the stochastic equations of motion give a correct canonical trajectory.

Nosé-Hoover and Langevin thermostat can also be used together (this "mixed" thermostat is implemented in the CP2K package used for ab initio calculations) [24].

2.2.4 Umbrella sampling

When a system has a sharp, dominating, energy minimum, the thermodynamic sampling is rather simple: a free MD simulation will correctly explore the relevant states around the minimum. If the system displays more metastable minima than a free MD sampling is not guaranteed to visit all the relevant regions of configuration space: it can happen that the mean escape time from a minimum is

larger than the timescale accessible to the MD simulation. In this case enhanced sampling methods are needed. These techniques try to make the system escape from the metastable states by adding a controlled bias potential. This bias potential has to be a function of a set of collective variables.

Suppose to run a MD simulation with a modified potential $\tilde{V}(\mathbf{r}) = V(\mathbf{r}) + B(\boldsymbol{\xi}(\mathbf{r}))$, where $B(\boldsymbol{\xi})$ is a term depending on CVs, called bias potential. The simulation will sample a biased probability density:

$$\tilde{\rho}(\boldsymbol{\xi}) = \frac{1}{Z}\int \delta\big(\boldsymbol{\xi} - \boldsymbol{\xi}(\mathbf{r})\big) e^{-\beta\left[V(\mathbf{r})+B(\boldsymbol{\xi}(\mathbf{r}))\right]} d^{3N}\mathbf{r} = e^{-\beta B(\boldsymbol{\xi})}\rho(\boldsymbol{\xi}) \tag{2.15}$$

The idea is to use the bias to let the MD explore states that are rarely accessed by plain MD and then reconstruct the real free energy function taking into account the effect of the bias.

Umbrella sampling [25] uses an harmonic potential as bias. Usually umbrella sampling is performed on one monodimensional collective variable at the time (distances between particles and centers of mass are common choices), and thus ξ is a scalar. The scheme consists in setting up a series of configurations of the system along the collective variables of interest and run a biased MD on each. Each run is usually called a window. Let the starting configuration be labeled with the index i and ξ_i the values of the collective variable at each starting configuration. The the $i-th$ window has a bias potential applied:

$$B_i(\xi) = \frac{k}{2}(\xi_i - \xi_i(t))^2 \tag{2.16}$$

The interesting collective variable path is then divided in bins. Let the bins be labelled with the index j, ranging from 1 to M, each centered at a value ξ_j of the collective variable, and denote with C_{ij} the total number of configurations from window i falling in the bin j. The real probability density at the bin j is:

$$\rho(\xi_j) = \omega_j \sum_i C_{ij} e^{-\beta B_i(\xi_j)} \tag{2.17}$$

ω_j is a weight factor that ensure normalization:

$$\sum_{j=1}^{M} \omega_j = 1 \tag{2.18}$$

The values of ω_j are computed with a procedure called Weighted Histogram Analysis (WHAM) [26], that constructs the set of weights in a self-consistent way. At the end the free energy profile along ξ is reconstructed from the normalized probability density simply by:

$$U(\boldsymbol{\xi}) = -\frac{1}{\beta}\log\tilde{\rho}(\boldsymbol{\xi}) \tag{2.19}$$

2.3 Classical Force Fields

A classical approach to force calculation uses a classical potential energy function that approximates the quantum ground-state potential energy. The classical description works well under the following conditions:

- the Born-Oppenheimer approximation is valid;
- the temperature is not too low;
- there is no bond breaking or forming;
- electrons are highly localized (metals and pi-bonded systems are delocalized).

Force fields [21] are sets of parameters defining a number of n-body interactions between atoms. Usually the interactions taken into account are divided into two categories: bonded, short-range interactions and non bonded, long-range interactions.

2.3.1 Bonded interactions

Bond interactions describe chemical bonding between atoms. They are of four types: pair stretch, angle bend, dihedral torsion and impoper dihedral interactions. Parameters defining the interactions are optimized to fit ab initio and/or experimental data.

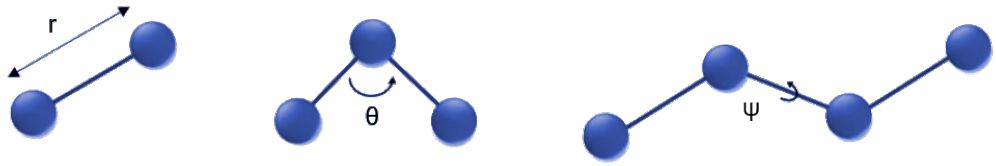


Figure 2.1: Definition of bond distance r, bond angle θ and torsion dihedral φ.

- Pair stretch interactions, between covalently bonded atoms:

$$V_{str}(r_{ij}) = \frac{k_{ij}}{2}(r_{ij} - r_{ij}^0)^2 \tag{2.20}$$

  where r_{ij} is the distance between bonded atoms and r_{ij}^0 the bond equilibrium distance.

- Angle bend interactions, between three consecutive covalently bonded atoms:

$$V_{bnd}(\theta_{ijk}) = \frac{k_{ijk}}{2}(\theta_{ijk} - \theta_{ijk}^0)^2 \tag{2.21}$$

  where θ_{ijk} is the angle formed by the three atoms and θ_{ijk}^0 the equilibrium angle.

- Dihedral interactions, between four consecutive covalently bonded atoms:

$$V_{dih}(\psi_{ijkl}) = \sum_{n=1}^{N}(1 + \cos(n\psi_{ijkl} - \phi_{ijkl})) \tag{2.22}$$

  where ψ_{ijkl} is the dihedral and ϕ_{ijkl} an offset parameter. N is usually equal to the number of minima the dihedral potential has: most of the times no more than two. Dihedral angle is equal to the angles between the planes containing the first three and the last three atoms. Dihedral potentials are generally weaker than the other: in some cases dihedrals can rotate through the full 360 degrees.

- Improper dihedral interactions, between four consecutive covalently bonded atoms:

$$V_{imp}(\varphi) = k(\psi_{ijkl} - \psi_{ijkl}^0)^2 \tag{2.23}$$

  Improper dihedrals are employed most of the times to enforce planar geometry, for example of aromatic rings.

2.3.2 Non bonded interactions

Nonbonded interactions are between atom pairs: they usually include the electrostatic interaction, the dispersion force and the short range repulsion.

- Electrostatic interaction is computed as the coulomb potential between point atomic charges:

$$V_{el}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon r_{ij}} \tag{2.24}$$

  the dielectric constant ϵ depends on the force field used. Most force fields use the vacuum permittivity ϵ_0. Atomic charges are usually extracted from ab initio gas phase calculations. Usually electrostatic is not calculated from the plain coulomb formula, which is computationally expensive, but instead using methods based on Ewald splitting between long-range and short-range terms [27]. Particle-Mesh Ewald (PME) [28] is the most efficient method for medium and large systems and it is commonly used.

- Dispersion and short range repulsion are modeled using the Lennard-Jones potential:

$$V_{LJ}(r_{ij}) = \varepsilon_{ij}\left[\left(\frac{r_{ij}^0}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij}^0}{r_{ij}}\right)^{6}\right] \tag{2.25}$$

  Usually not all the possible ε_{ij} and r_{ij}^0 are defined, but only their values for pairs of atoms of the same type. Interactions between different atom types are approximated using mixing rules.

2.3.3 Water models

Due to its ubiquity and its importance for molecular systems water is probably the compound that has been studied more in depth from a FF perspective. The set of FF parameters related to water molecules are so important that they have their own name: the water model. Many times the vast majority of the atoms in a simulation box are water oxygens and hydrogens, thus a good water model has to be not too computationally demanding. For this reason many water models are rigid, with angles and bond lengths fixed. If one is not interested in resonance spectra usually this can be a reasonable approximation. Some models include more than three charge centers, adding ghost charged particles. For the classical calculations here presented the TIP3P water model, a rigid three charges model, was used.

2.4 Ab initio methods

Although this thesis is mainly focused on machine learning, ab initio calculation are a crucial part of the whole work. Data for neural network training are generated ab initio: it is obvious that the accuracy of the final model will have as upper limit the accuracy of the reference ab initio calculations. From this the importance of having a solid and reliable ab initio method to generate training data. All the ab initio calculations performed for this work rely on Density functional theory (DFT). All the calculations were carried out with CP2K: a free, open source package for ab initio and classical molecular dynamics. A brief recall of the Kohn-Sham formulation of DFT is provided in the next subsection; a more in depth review of the implementation in the package CP2K is matter of the second one; while relevant choices of parameters are discussed in the third one.

2.4.1 Kohn-Sham Density Functional Theory

As usual in electronic structure calculations, the nuclei are treated as instantaneously fixed (Born-Oppenheimer approximation), generating a static external potential $V(\mathbf{r})$, in which the electrons are moving. The ground state of N interacting electrons is represented by a many-body wavefunction $\Psi_{gs}(\mathbf{r}_1, \mathbf{r}_2 \dots \mathbf{r}_N)$, a solution of the time independent Schrödinger equation:

$$\hat{H}\Psi = (\hat{T} + \hat{V} + \hat{U}) = \left[-\frac{\hbar}{2m}\nabla^2 + \sum_i V_i(\mathbf{r}_i) + \sum_{i<j} \frac{e^2}{|\mathbf{r}_i - \mathbf{r}_j|}\right]\Psi = E\Psi \tag{2.26}$$

Here the symbols $\hat{T}$, $\hat{V}$ and $\hat{U}$ represent the kinetic energy operator, the core-electron interaction and the electron-electron interaction.

A theorem due to Hohenberg and Kohn states that the energy of a quantum mechanical system is a unique functional of the electron density $n(\mathbf{r}) = \langle\Psi|\hat{n}(\mathbf{r})|\Psi\rangle$, where $\hat{n} = \sum_{i=1}^{N}\delta(\mathbf{r} - \mathbf{r}_i)$ is the density operator.

$$E[n] = T[n] + \int V(\mathbf{r})n(\mathbf{r})d\mathbf{r} + \frac{e^2}{2}\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + E_x[n] \tag{2.27}$$

Another fundamental result, the so called variatonal theorem, states that the expectation value of the hamiltonian on the ground state wavefunction Ψ_{gs} (i. e. the ground state energy) is smaller than the expectation value on every other possible wavefunction Ψ :

$$\langle\Psi_{gs}|\hat{H}|\Psi_{gs}\rangle < \langle\Psi|\hat{H}|\Psi\rangle \tag{2.28}$$

And so, by virtue of the Hohenberg and Kohn theorem:

$$E[n_{gs}] < E[n] \tag{2.29}$$

So one can solve (2.27) variationally to obtain the ground state electron density and the ground state energy, with the constraint on the number of particles:

$$\frac{\delta\Big(E[n] - \mu\big(\int n(\mathbf{r})d\mathbf{r} - N\big)\Big)}{\delta n} = 0 \tag{2.30}$$

The most known and commonly employed method to solve this variational problem is due to Kohn and Sham. For example, if the trial wavefunction is written as a Slater determinant $\Psi(\mathbf{r}_1, \dots \mathbf{r}_N) = \frac{1}{\sqrt{N}}\det[\phi_1\phi_2 \dots \phi_N]$, where ϕ_i are single particle wavefunctions (the so called orbitals); the kinetic energy of a non interacting system with this electron density is:

$$T_0[n] = -\frac{\hbar}{2m}\sum_{i=1}^{N}\int \phi^*(\mathbf{r})\nabla^2\phi(\mathbf{r})d\mathbf{r} \tag{2.31}$$

The idea is to write the functional (2.27) as:

$$E[n] = -\frac{\hbar}{2m}\sum_{i=1}^{N}\int \phi^*(\mathbf{r})\nabla^2\phi(\mathbf{r})d\mathbf{r} + \int V(\mathbf{r})n(\mathbf{r})d\mathbf{r} + \frac{e^2}{2}\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}' + E_x[n] + T[n] - T_0[n] \tag{2.32}$$

and to incorporate $T[n] - T_0[n]$ and the exchange energy $E_x[n]$ in a unique exchange-correlation (XC) functional $E_{xc}[n]$ which at this point contains all the terms of the functional that are not exactly known. The calculation of this term has to rely on approximations: usually the functional is written in this form:

$$E_{xc}[n] = \int \varepsilon_{xc}\big(n(\mathbf{r}), \nabla n(\mathbf{r}) \dots\big)d\mathbf{r} \tag{2.33}$$

Many formulations for ε_{xc} exists and the suitable one has to be chosen depending on the characteristics of the system under study. Exchange correlation functional are divided in a number of categories. The basic ones are:

- Local Density Approximation (LDA) functionals: ε_{xc} depends only on the local density $\varepsilon_{xc} = \varepsilon_{xc}\big(n\big)$
- Generalized Gradient Approximation (GGA) functionals: ε_{xc} depends on the density and its derivatives $\varepsilon_{xc} = \varepsilon_{xc}\big(n, |\nabla n|, \nabla^2 n\big)$

At this point it is possible to reformulate the variational problem (2.30) as a system of N variatonal problems in which the energy is minimized w.r.t. the complex conjugate of the orbitals ϕ_i subject to normalization constraint $\int |\phi_i|^2 d\mathbf{r} = 1$:

$$\frac{\delta\delta\Big(E[n] - \varepsilon_i\big(\int |\phi_i(\mathbf{r})|^2 d\mathbf{r} - 1\big)\Big)}{\delta\varphi_i^*} = \frac{\delta T_0}{\delta\varphi_i^*} + \left[V(\mathbf{r}) + \int \frac{e^2 n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \frac{\delta E_x[n]}{\delta n}\right]\frac{\delta n}{\delta\varphi_i^*} - \varepsilon_i\phi_i = 0 \tag{2.34}$$

Define $V_{xc}[n] = \frac{\delta E_x[n]}{\delta n}$; the Kohn-Sham equations follow:

$$\left[-\frac{\hbar}{2m}\nabla^2 + V(\mathbf{r}) + \int \frac{e^2 n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + V_{xc}[n]\right]\phi_i = \varepsilon_i\phi_i \tag{2.35}$$

These are Schrödinger equations of N non interacting electrons moving in an effective potential $V_{eff} = V(\mathbf{r}) + \int \frac{e^2 n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + V_{xc}$ and are much simpler than the original problem.It is possible to obtain a solution in an iterative way: guess an initial set of orbitals $\{\phi_i\}$, compute the effective potential, solve the Kohn-Sham equations getting new orbitals to compute a new density. The procedure is iteratively repeated until convergence.

2.4.2 Gaussian-Plane wave method, Quickstep code

The package CP2K implements DFT in a form called Gaussian and Plane wave (GPW) method [29]. The core of the implementation is a code called Quickstep [30]. The method relies on two representations of the electron density: one based on atom-centered gaussians-type orbitals (GTO) and another based on plane waves (PW) . Each representation has its own advantages: GTOs permits a faster evaluation on the kinetic and ionic core interaction matrix elements, while PWs are better suited for XC and electron-electron repulsion (Hartree) calculation. The two representations read:

$$\begin{array}{ll} \text{Gaussian representation} & \text{Plane wave representation} \\ \begin{cases} n(\mathbf{r}) = \sum P_{\mu\nu}\varphi_\mu(\mathbf{r})\varphi_\nu(\mathbf{r}) \\ \varphi_\mu(\mathbf{r}) = \sum_m C_{m\mu} g_m(\mathbf{r}) \end{cases} & n(\mathbf{r}) = \sum_{|\mathbf{G}|<G_c} n(\mathbf{G}) e^{-i\mathbf{G}\cdot\mathbf{r}} \end{array} \tag{2.36}$$

Where $n(\mathbf{G})$ are the Fourier coefficients of the electron density and $g(\mathbf{r})$ are primitive gaussian orbitals. PWs provide an efficient representation only if the density varies smoothly and slowly: representing rapid variation requires an high cutoff G_c, and building the PW representation would become a bottleneck in the calculations. For this reason this method is suited only for valence electrons: core electrons are to be frozen out by introducing an atomic pseudopotential V_{PP}. Atomic pseudopotentials are effective potential felt by valence electrons when the positive charge of the nuclei is screened by the core electrons. They are built following different possible procedures, in general fitting all electron calculations. The choice of the pseudopotential for the present work will be discussed later. The expressions for the energy functionals in GPW method are:

$$\begin{aligned} T[n] &= -\frac{1}{2}\sum_{\mu\nu}\langle\phi_\mu(\mathbf{r})|\nabla^2|\phi_\nu(\mathbf{r})\rangle \qquad E_I[n] = \sum_{\mu\nu}\langle\phi_\mu(\mathbf{r})|V_{PP}(\mathbf{r})|\phi_\nu(\mathbf{r})\rangle \\ E_C[n] &= \sum_{\mathbf{G}} \frac{n^*(\mathbf{G})n(\mathbf{G})}{\mathbf{G}^2} \qquad E_{xc}[n] = \int \varepsilon_{xc}\big(n(\mathbf{r}), \nabla n(\mathbf{r}) \dots\big) d\mathbf{r} \end{aligned} \tag{2.37}$$

The kinetic and pseudopotential matrix elements are computed analytically with recursive formulas discussed in [29], the Hartree terms are computed in reciprocal space in a Ewald-like manner and XC term is integrated numerically on a mesh.

2.5 Neural Network potentials

In the last years a new technique, based on Neural Networks (NN) for calculating energy and forces of molecular systems have emerged [31]. In this section an overview of NN architecture is given, and the Deep Potential method [32], that was used for this thesis is explained in detail.

A NN [33] consists in an ensemble of nodes, connected by links. Each node receives an input x and gives an output $y = f(x)$ according to an activation function f. Most of the times the activation function f is the same for every node. Each link is characterized by a *weight*: if a node is linked with n nodes each giving an output x_i and each link has weight w_i, the output of the node will be

$$y = f\left(\sum_{i=1}^{n} w_i x_i\right)$$

Sometimes so called *bias nodes* are be present. A bias node is a node that gives every time the same output. Then the output of a node is:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b_i\right)$$

Several NN architectures exist: the one that is mostly used, and that is used in the field of MD, is the feedforward, layered architecture [34]. The feedforward NNs are NNs that can be represented

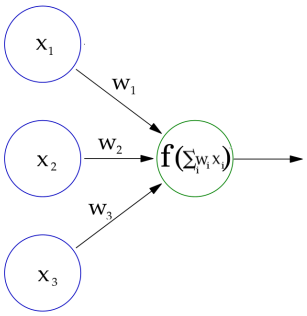


Figure 2.2: Depiction of the input-output scheme in a NN.

by a directed, acyclic graph. The main feature of feedforward NNs is that they do not contain closed loops: the output of a node cannot influence the node itself. For this reason the feedforward NNs are static, while their counterpart, the recurrent NNs, have a dynamic time-dependent behavior. Nodes in a feedforward neural network can be organized in layers: a node belonging to the layer i receives input only from the nodes belonging to the layer $i-1$ and gives his output only to the nodes belonging to the layer $i+1$. For historical reasons this architecture is called multilayer perceptron. From now on the term NN will be used implying that we are referring to a multilayer perceptron

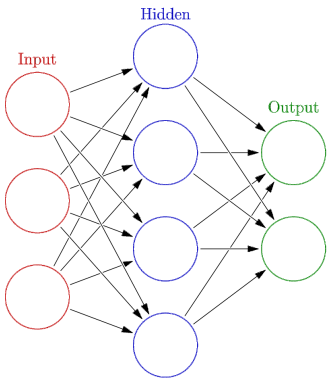


Figure 2.3: A multilayer perceptron with a single hidden layer.

architecture. At this point some nomenclature is nedeed. The input of the NN is, in general, a vector of real numbers $\mathbf{x} = (x_1,\ x_2 \ldots,\ x_n)$ and it is itself considered a layer, called input layer. Than there are the hidden layers, each of them receiving the input from the previous layer and firing its output to the subsequent one. The last layer is called output layer and its dimension gives the dimension of the output: a vector $\mathbf{y} = (y_1,\ y_2 \ldots,\ y_m)$. Call $\mathcal{N}$ the NN, which receives an input x giving an output $\mathcal{N}(\mathbf{x})$. The NN can be viewed as a composition of transformation laws, each one representing a single layer.

$$\mathcal{N}(\mathbf{x}) = \mathcal{L}^{(out)} \circ \mathcal{L}^{(N)} \circ \mathcal{L}^{(N-1)} \circ \ldots \mathcal{L}^{(1)}(\mathbf{x}) \qquad \begin{array}{rccc} \mathcal{L}^{(i)}: & \mathbb{R}^{N_{i-1}} & \to & \mathbb{R}^{N_i} \\ & \mathbf{x} & \to & f^{(i)}\big(W^{(i)}\mathbf{x}\big) \end{array} \tag{2.38}$$

where $W^{(i)} \in \mathbb{R}^{N_i \times N_{i-1}}$ and $f(.)$ represents the application of the activation function element-wise.

2.5.1 Activation functions

Typically activation functions are non-polynomial, non-even functions and are bounded at least inferiorly. Many activation functions have been discovered to have good performance. At the present state of the theoretical research there is only a little and qualitative understanding of why they perform well. A (non exhaustive) list of common activation function is:

- hyperbolic tangent: $f(x) = \tanh(x)$
- rectified linear unit (ReLU): f(x)= $\big\{0$ if $x < 0\ \ ;\ \ x$ if $x \leq 0\big\}$

- exponential linear unit (ELU): $\{a(e^x - 1) \text{ if } x < 0 \ ; \ x \text{ if } x \leq 0\}$
- gaussian error linear unit (GELU): $f(x) = x(\text{erf}(x) + 1)$

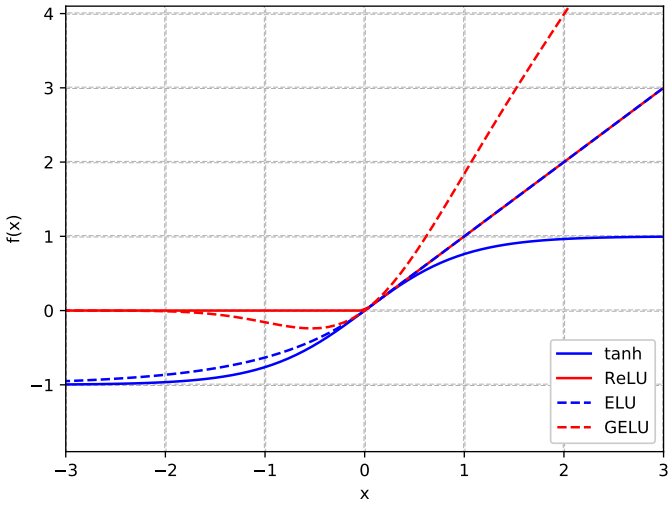


Figure 2.4: Plot near the origin of some commonly used activation functions.

When the activation function is non-polynomial, then a feedforward neural network with one hidden layer of arbitrary size can be proven to be a universal function approximator [35]. This is known as the Universal Approximation Theorem.

The most important feature of an activation function is its behavior in the vicinity the origin: usually the weights and the inputs are normalized in a way such that every node input stays around zero; when the inputs start to rise or low too much usually the model i

2.5.2 Training a NN

The NN are well suited for regression problems. A set of possible inputs is given along with their correct outputs, which are called labels. At each optimization step of the NN takes the training inputs, gives its output and the weights are updated depending on the difference between output and labels, following a certain learning algorithm; this process is repeated until the network reaches a convergence.

Call $\mathbf{x}_i^*$ the points in the training set, $\mathbf{y}_i^*$ the target labels and $\mathbf{y}_i = \mathcal{N}(\mathbf{x}_i^*)$ the final output of the network given a training input $\mathbf{x}_i$. A function $L(\mathbf{y}_i, \mathbf{y}_i^*)$, called loss function, is defined in order to represent the degree of mismatch between $\mathbf{y}_i$ and $\mathbf{y}_i^*$. A common loss function is the quadratic loss: $L(\mathbf{y}_i, \mathbf{y}_i^*) = |\mathbf{y}_i - \mathbf{y}_i^*|^2$

The weights $w_{ij}^{(n)}$ at learning step τ are updated according to:

$$w_{ij}^{(n)}(\tau + 1) = w_{ij}^{(n)}(\tau) - \Delta w_{ij}^{(n)}\big(L(\mathbf{y}_i(\tau), \mathbf{y}_i^*)\big) \tag{2.39}$$

The form of $\Delta w_{ij}^{(n)}$ is defined by a certain learning algorithm. A learning algorithm that is straightforward to implement for NNs is the gradient descent:

$$\Delta w_{ij}^{(n)} = -\eta \frac{\partial L(\mathbf{y}_i, \mathbf{y}_i^*)}{\partial w_{ij}^{(n)}} \tag{2.40}$$

where η is a weight called learning rate used to tune the length of algorithm's steps. The gradients are calculated with an scheme called backpropagation, obtained by systematic application of the chain rule [36].

In principle the learning can be conducted feeding the network with all the trainig data at each step,

calculating the mean loss, and updating the weights according to the learning rule. In practice this procedure makes the system fall in the first local minimum it finds, and, giving the high dimensionality of the parameter space, this local minimum has a negligible probability of been a "good" minimum. Usually neural networks are trained using stochastic learning algorithms. Stochasticity is good for learning, as it gives the algorithm the possibility to jump out from local minima and explore a bigger area of the parameter space, enhancing the probability of find good solutions. To achieve this the training set is divided in batches with B elements each; for each batch a mean loss is defined:

$$L_{batch}(\{\mathbf{y}_i\},\{\mathbf{y}_i^*\}) = \frac{1}{B}\sum_{i\in batch} L(\mathbf{y}_i,\mathbf{y}_i^*) \tag{2.41}$$

At each optimization step a batch is selected, L_{batch} is calculated and used to feed the learning algorithm. Each time all the batches have been used, the set is shuffled, new batches are defined and a new *training epoch* begins. This procedure is iterated until convergence. The batch size can affect the training efficiency: the smaller the batch size is, the bigger is the stochasticity in the training induced by the random choice of the elements in the batches. In the limit of the batch size equal to the training set size the training is completely deterministic, once the initial weights and biases are given.

In addition to stochasticity the more advanced optimization algorithms add to Δw memory terms that make the learning a non-markovian process. The package DeepMD, which was used for the present work, implements the so called Adam algorithm [37], a non-markovian stochastic gradient descent which has become a gold standard in the deep learning field.

When the model is trained it is a standard procedure to test it on data outside the training set. This data form the so called *test set*. In this way it is possible to control the phenomenon called *overfitting*: the tendency of the machine learning models to perform well on the data they were trained on and give worse performances on unseen data.

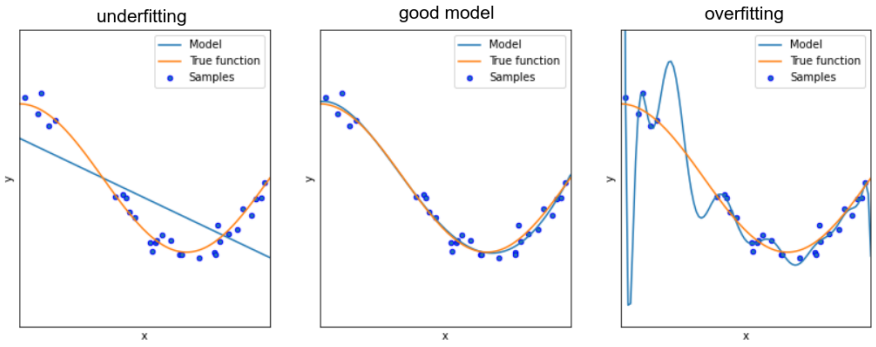


Figure 2.5: An illustration of overfitting.

2.5.3 NN for molecular dynamics

The first question to answer, walking into the field of NN applied to molecular dynamics is: why NNs? The age-old problem of MD is to replicate at the hemical accuracy level the interaction energies of a system, without relying on expensive quantum mechanics calculations. The idea is to fit a semi-empirical, completely classical, model on a (relatively) small set of data obtained from AIMD. The models that were mostly used up to now to tackle this problem pertain to the FF class that will be described in the last section of this chapter. The main problem with force-fields is their lack of transferability. Most of the times the force fields are developed in order to match certain structures (for example the known crystallographic structures of DNA) and are reliable only if the system is sufficiently close (at least locally) to them. If the system is far from these structures, an error is made, and the quantification of this error is mainly matter of intuition. The validity of a simulation with FF models can be inquired, many times, only qualitatively. The lack of transferability arises because the FFs are built up mostly with pair interactions (three and four-body interactions are present only for bonded atoms in form restraints on bond angles and dihedrals) while the real interactions have

complex multi-body nature. Some classes of more refined models, like embedded atom methods, bond order potentials, polarizable FFs and reactive FFs are based on the idea that the strength of an interaction may depend also on the local environment around an atom, but the choice of the functional form of this dependence is somehow arbitrary.

NN based methods are changing the state of the art. Trained on large sets of atomic configurations, and correspondent energies and forces, these methods can reproduce the original data accurately and give potential energy functions that are sufficiently transferable. Neural networks are naturally fit for regression problems, and the reconstruction of a potential energy surface from sparse samples calculated ab initio is indeed a regression problem. A great advantage w.r.t. classical force field is that NN models are agnostic, in the sense that no ad hoc choice on the analytical form of the potential energy has to be done. This can be a great advantage when dealing with very complex energy surfaces which functional form is difficult to guess.

2.5.4 Beheler-Parrinello architecture

Most of the NN models proposed for potential energy fitting follow the scheme developed by Beheler and Parrinello in [38]. Consider a system of N atoms and denote their positions as $\{\mathbf{r}_1,\ \mathbf{r}_2, \dots\ \mathbf{r}_N\}$, being each $\mathbf{r}_i \in \mathbb{R}^3$. The potential energy of the system is a function of $3N$ variables: $E = E(\mathbf{r}_1,\ \mathbf{r}_2, \dots\ \mathbf{r}_N)$. In the framework of Beheler-Parrinello architecture the potential energy is decomposed into a sum of atomic contributions:

$$E = \sum_i E_i(\{\mathbf{r}_{ij}\}_{i \neq j}) \tag{2.42}$$

where $\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i$ is the position of the atom j relative to the atom i. Here an approximation is done: only the atoms within a certain cutoff radius from the atom are considered in the computation of its contribution to the energy

$$E \simeq \sum_i E_i(\mathcal{R}_i) \tag{2.43}$$

where $\mathcal{R}_i = \{\mathbf{r}_{ij} \ \big|\ |\mathbf{r}_{ij}| \leq r_c\}_{i \neq j}$ the set of positions of atoms inside the cutoff radius, relative to atom i. This list describes the *local environment* of atom i. The Beheler-Parrinello architecture is an ensemble of N identical sub-architectures S_i, the i-th representing the function

$$E_i = E_i(\mathcal{R}_i)$$

This decomposition ensures the extensivity of the resulting energy function and make possible to apply a model trained on a small system to larger systems. Each sub-architecture is composed of a standard feedforward neural network and some sort of pre-processing unit. The naivest input for the NN would be the list $\mathcal{R}_i$, but this choice is not optimal from the computational point of view and it does not guarantee the translational, rotational, and permutational symmetry of the potential energy. Invariances can be of course learned by the NN itself, but this process requires a lot of redundant data (e.g. many identical but rotated-translated-permuted configurations) and the learning process can be slowed down by orders of magnitude. For this reason a preprocessing of the atomic positions is performed, and the positions are mapped into a quantity called local environment descriptor. In the last years a great effort has been put into the development of different local environment descriptors ans a large variety of possible choices is now available. [39] A brief overview of theory of local environment descriptors will be given in the next subsection. The model is trained in the classical way, defining a loss function and performing a stochastic gradient descent. If continuous and differentiable activation functions are used, the forces on an atom can be directly computed simply taking the derivative of the resulting energy w.r.t. the specific atomic positions and applying the chain rule to every NN layer. Therefore the model can be used to run a MD just as a classical FF, and the conservation of energy is ensured. Moreover, the simplicity of force calculation makes it feasible to learn simultaneously the energy and the forces, simply inserting a force deviation term in the loss function. The possibility of direct force learning improves the efficiency of the whole training: each ab initio snapshot can give much more information if forces are taken into account. Since in electronic structure calculations the forces are simply extracted by virtue of the Hallmann-Feynman theorem,

the simultaneous learning on forces and energies drastically reduces the required size of the training set, with negligible computational cost.
To sum up, the main advantages of the Beheler-Parrinello architecture are:

- The decomposition of the energy assures its extensivity and allows the generalization of the model at systems of different size.
- Forces can be directly computed by differentiation, an operation that is usually fast and simple to implement on feedforward NN models, giving an energy-conserving model.
- The model can be trained on forces and energies at the same time, preventing overfitting and requiring less AIMD snapshots in the learning set.

The main drawback is the difficulty in treating long-range interactions, given the explicit cutoff. This fact can limit the effectiveness of the Beheler-Parrinello networks in treating systems in which long-range correlations dominates the dynamics.

2.5.5 Local environment descriptors

Let $\mathcal{R} \in \mathbb{R}^{3N}$ be the local environment of a certain atom. The local environment descriptor is a map:

$$\begin{array}{rccc} \mathcal{D}: & \mathbb{R}^{3N} & \to & V \\ & \mathcal{R} & \to & \mathcal{D}(\mathcal{R}) \end{array} \tag{2.44}$$

where V is a generic real vector space. In practice the usual choices for V are $\mathbb{R}^M$ and $\mathbb{R}^{M_1 \times M_2}$. A good descriptor is:

- Invariant under translation, rotation, and permutation of indexes of atoms of the same kind.
- Surjective: different configurations must be mapped into different values of the descriptor. If the map $\mathcal{D}$ is also bijective the descriptor is said to be *complete*; it is said to be *overcomplete* if it is only surjective.
- Continuous and differentiable at least of class C^1, since its derivative is needed to calculate the forces and discontinuity in the forces are to be avoided.
- Computationally efficient.

Beheler and Parrinello originally proposed [38] to use a set of invariant functions (mainly combinations of gaussians and trigonometric functions) of the atomic coordinates as local descriptors. During the last decade many different descriptors has been developed [39].

2.5.6 The Deep Potential method

The NN model developed for this thesis is based on the package DeepMD [40]. This package implements the Deep Potential method [32], in particular its smooth variant [41]. The model pertains to the class of Beheler-Parrinello architectures. The potential energy is decomposed as in eq. (2.43). Thus, for each atom, a local environment is computed, then, for each $\mathbf{r}_{ij} = (x_{ij}, y_{ij}, z_{ij})$ in the environment the following quantity is calculated:

$$\tilde{\mathcal{R}}_{ij} = \left(\frac{s(r_{ij})}{r_{ij}}, \frac{x_{ij} s(r_{ij})}{r_{ij}}, \frac{y_{ij} s(r_{ij})}{r_{ij}}, \frac{z_{ij} s(r_{ij})}{r_{ij}} \right) \tag{2.45}$$

where $s(r_{ij})$ is a smooth cutoff function, that is needed in order to avoid discontinuities.

$$s(r_{ij}) = \begin{cases} 1 & \text{if} \quad r_{ij} \leq r_{cs} \\ \dfrac{1}{2 r_{ij}} \left[\cos\left(\pi \dfrac{r_{ij} - r_{cs}}{r_c - rcs} \right) + 1 \right] & \text{if} \quad r_{cs} < r_{ij} \leq r_c \\ 0 & \text{if } r_{ij} > r_c \end{cases} \tag{2.46}$$

The idea behind this choice is quite intuitive: most interactions decay at long distances as powers of $1/r$ ($1/r$ the coulomb interaction, $1/r^6$ the London force ecc...), thus a descriptor like that forces a physically meaningful convergence of the energies predicted by the NN at long distances. The presence of a smooth cutoff function serves to avoid discontinuities in the potential.

The vectors $\tilde{\mathcal{R}}_{ij}$ are organized in a matrix $\tilde{\mathcal{R}}_i$, each row of it being a vector $\tilde{\mathcal{R}}_{ij}$. The this matrix is processed to the final local environment descriptor $\mathcal{D}_i$. It is possible to demonstrate that the matrix $\tilde{\mathcal{R}}_i\tilde{\mathcal{R}}_i^T$ is invariant rotation and translation. Thus an invariant descriptor can be constructed as:

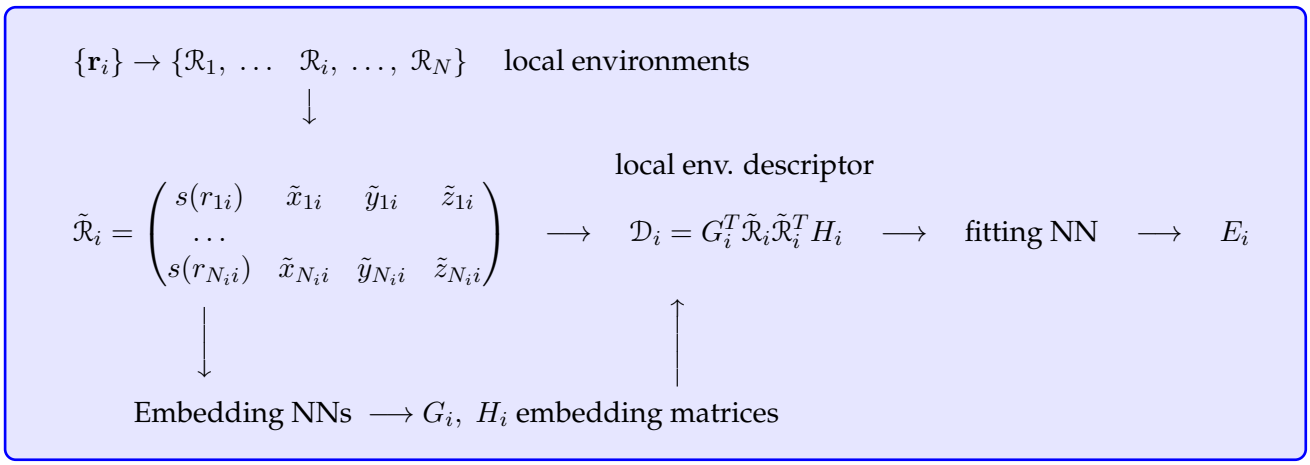
$$\mathcal{D}_i = G_i^T \tilde{\mathcal{R}}_i \tilde{\mathcal{R}}_i^T H_i \tag{2.47}$$

where $G_i \in \mathbb{R}^{N_i \times M_1}$ and $H_i \in \mathbb{R}^{N_i \times M_2}$ are themselves invariant. The local environment descriptor will be $\mathcal{D}_i \in \mathbb{R}^{M_1 \times M_2}$ and so, if $\mathcal{R}_i$ if $M_1 M_2 < 3N_i$, the information it carries will be compressed w.r.t the original loval environment $\mathcal{R}_i$: the matrices G_i and H_i can be viewed as information filters.

Those matrices have to be functions of variables that are invariants, an obvious choice is:

$$G_i = G_i\big(\{s(r_{ij})\}\big) \qquad H_i = H_i\big(\{s(r_{ij})\}\big) \tag{2.48}$$

In the Deep Potential architecture those functions are themselves trainable neural networks. These networks, called embeddings, are defined between pairs of different atom types. This means that the computational cost of training and running a network with k different atom types scales as $k(k+1)/2$. Defining more atom types for a single element, like in classical force field, could be beneficial for the model, since every embedding will learn less atomic environments, but can lead to much poorer performance in terms of computation time. The descriptor $\mathcal{D}_i$ is passed to a NN called fitting net, which is different for each atom type. So, a network with k atom types is composed of $k(k+1)/2$ embedding NN and k fitting NN. The whole Deep Potential algorithm is summarized in the following scheme.



The loss function of the model is defined as:

$$L = c_E \sigma_E^2 + c_f \sigma_f^2 \tag{2.49}$$

where σ_E^2 and σ_E^2 are the root mean squared errors on energy and forces.

$$\sigma_E^2 = \frac{1}{N}(E^{NN} - E)^2 \qquad \sigma_f^2 = \frac{1}{3N}|\mathbf{f}^{NN} - \mathbf{f}|^2 \tag{2.50}$$

where N the number of atoms in a given training sample.

Chapter 3

MD and network training

This chapter describes the construction, training and testing of the NN model. Premilinarly some details about the AIMD simulations used to train the network and on the classical MD simulations used to prepare equilibrated systems and to generate data for comparison are given.

3.1 Classical MD

For the simulations with classical FF the package GROMACS [42], version 2018.1, was used.

3.1.1 Force field

AMBER FF with OL15 refined parameters for nucleic acids has been employed [43]. This force field is considered, along with AMBER-bsc1, the current state of the art for simulation of the DNA double helix [6]. However this FF has been refined on double helical structures, so it may give imprecise results for free nucleotides. As already said, a classical FF specifically developed for nucleotides in solution does not exist. Moreover, since the AMBER-OL15 FF is developed for nucleotides bound by a backbone, an ad hoc adjustment of the phosphate charges was necessary. Unfortunately no charge parametrization for the monophasphate group exists in literature: the charges employed were optimized for the diphosphate head oxygens [44]. Classical simulations are not expected to give correct quantitative results, but they are only intended as a benchmark for the NN potential: a succesful NN potential should reproduce DFT results better than the classical FF.

3.1.2 MD parameters

The Nosé-Hoover thermostat has been employed. The electrostatic interactions were treated using PME algorithm. The bond involving hydrogen were constrained using LINCS algorithm [45] and the integration timestep was set to 2 fs. All the simulations were carried out in periodic boundary conditions.

3.2 AIMD

For the DFT simulations the package CP2K [46], version 6.1, was used.

3.2.1 DFT parameters

The BLYP exchange correlation functional has been used. BLYP is a GGA functional that combines Becke's exchange functional [47] with the correlation functional developed by Lee, Yang and Parr [48]. This functional is known to reproduce well the properties of liquid water: in particular it gives a good description of the hydrogen bond [49]. The GTH pseudopotential [50] has been employed, along with the DZVP-MOLOPT-GTH basis set [51].

Gaussian orbitals in CP2K are represented on multiple grids, the sharpest on finer grids, the smoother on sparser ones. The thickness of the grids can be set choosing the reciprocal space cutoff of the finer one. Normally a cutoff around 280 Ry (CP2K's preset value) is considered adequate for having energy and forces correctly converged and has been employed for pure water [52]. However this cutoff seems too low for the systems under study: the Na cations tend to remain frozen in their initial positions during the MD run and there is a constant energy drift. The cutoff has consequently been increased to 600 Ry, a value at which no ion freezing is observed and no energy drift displays. This is in line with other values employed in literature for simulations involving cations [53].

3.2.2 Dispersion correction

As pointed out in the first chapter the dispersion forces play an important role in the interaction between nucleotides. There are two main ways to treat these forces in the framework of DFT. One is to incorporate them completely ab initio in the exchange-correlation functional, the other is to add an explicit attraction term between atoms, parametrized to fit ab initio calculations. Since the inclusion of dispersion ab initio is computationally too expensive for our purposes the second way was chosen. In particular the D3 correction parametrized by Grimme et al. [54] has been employed.

D3 consist in a classical correction that depend on interatomic distances. It incorporates the $1/r^6$ and $1/r^8$ dependencies that are found when dispersion interaction is treated by means of perturbation theory, with terms fitted on reference, fully ab initio, data. This method has become widely popular in the last years and has proven to be reliable in the simulation of nucleobase stacking [55]. As suggested in [55] the variant of this method with the damping proposed by Becke and Johnson, with a cutoff range of 8 Å was adopted.

3.2.3 MD parameters

The mixed Langevin Nosè-Hoover thermostat has been used. The integration timestep has been set to 0.5 fs. All the simulations were carried out in periodic boundary conditions.

3.2.4 Generation of the training set

The training data for the NN were taken from DFT MD trajectories with different starting configurations of nucleotides solvated in water. The reference from which the starting configurations were built is the crystallographic structure of the Dikerson dodecamer [56], available from the Protein Data Bank [57]. To have the starting configuration the interesting part has been isolated, the residual phosphodiesteric bon cut and the missing phosphate oxygen added. The structure were then relaxed in gas phase and solvated using GROMACS before starting the DFT MD run. Sodium ions were added in order to balance the phosphate charges. The starting configuration that were used are:

- Free nucleotides: dAMP and dTMP in cubic boxes of 15^3 Å.
- Stacking dimers: dAMP/dAMP, dTMP/dTMP and dAMP/dTMP in cubic boxes of 22^3 Å.
- Pairing dimer: dAMP-dTMP in a $20 \times 30 \times 20$ Å^3 box.
- Pairing-stacking quadruplet: dAMP/dTMP - dTMP/dAMP in a $20 \times 30 \times 20$ Å^3 box.

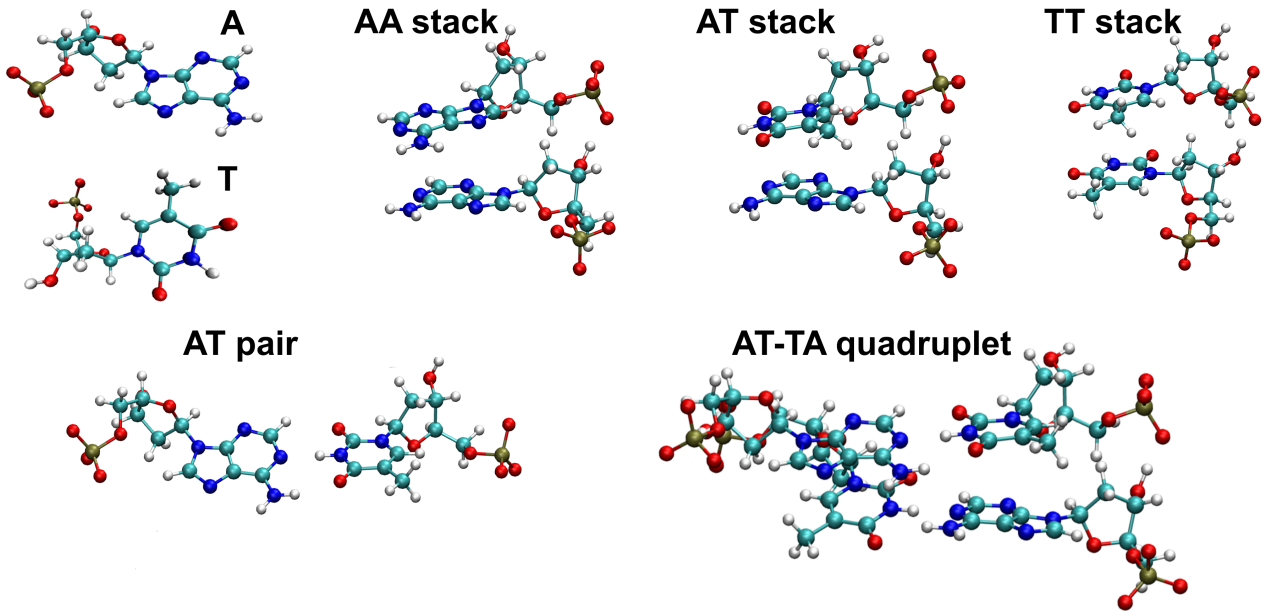


Figure 3.1: Starting structures for dataset generation.

For every starting configuration two NVT trajectories have been computed: one at 300 K and one at 1000 K. The rationale behind this choice is to have both a thick sampling of the regions around the equilibrium, but also an out of equilibrium sampling of configurations that are difficult to explore at ambient temperature. A ~ 6 ps trajectory of a system containing only water, generated at 1000 K, has been added to that data to have a better sampling of solvent interactions. Table 3.1 reports in detail the features of each training subset.

system	n. frames	time [ps]	box [Å^3]
A	16370	~ 8.2	15^3
T	12180	~ 6.4	15^3
A/A	30560	~ 15.3	22^3
A/T	18900	~ 9.5	22^3
T/T	21350	~ 10.7	22^3
A-T	8670	~ 4.3	$20 \times 30 \times 20$
QUAD	30300	~ 15.2	$20 \times 30 \times 20$
WATER	12744	~ 6.4	15.6404^3
Total	151074	~ 75.5	

Table 3.1: Size of each system in training set.

3.3 NN potential training

For the neural network training the package DeepMD [40], version 1.1.4 has been used. All the simulations with NN potential were performed with LAMMPS [58].

3.3.1 Network testing

The network hyperparameters have been chosen after a series of numerical experiments on training performance and model stability. First of all the network has to give a stable, converging learning. Typical root mean square errors (RMSE) on the test set of a NN potential trained for bulk systems are $\sigma_E \sim 10^{-3}$ eV and $\sigma_f \sim 10^{-1}$ eV/Å [59]. RMSEs are here defined as

$$RMSE_E = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{N_i}\sqrt{(E_i^{NN} - E_i)^2} \qquad RMSE_f = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{3N_i}\sqrt{|\mathbf{f}_i^{NN} - \mathbf{f}_i|^2} \tag{3.1}$$

where M is the size of the test batch and N_i the number of atoms in each tested frame.

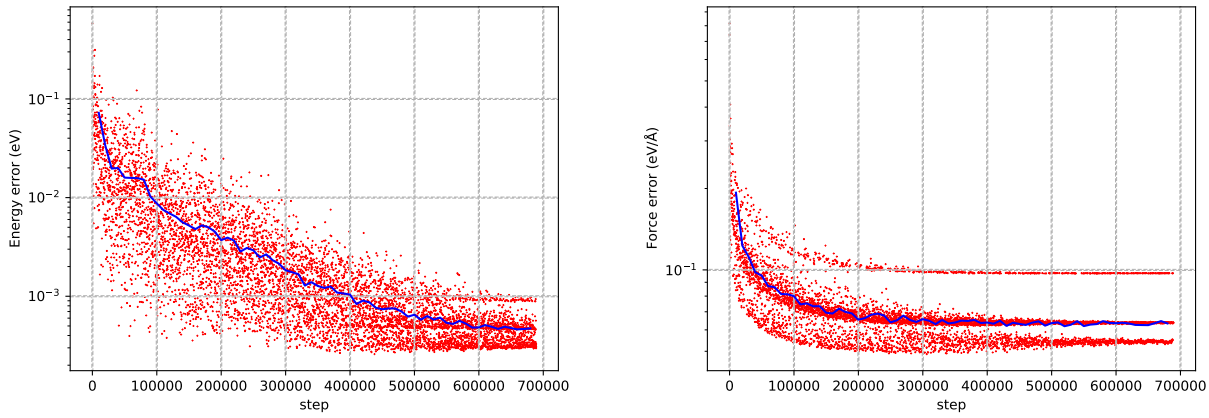


Figure 3.2: Typical learning curves for a correctly converging training. The full lines are moving averages on 100 steps windows-

Figure 3.2 shows typical learning curves for energy and forces. The splitting of the points on different lines is due to the fact that DeepMD conducts the test, at each step, on a single system and the model performs differently on different systems. The energy contribution associated to the complex bond geometry of a nucleotide is obviously much more difficult to describe than the simple water molecule: in figure 3.2, for example, the net was trained on three systems: A, A/A and WATER. A contains 101 water molecules for one nucleotide, A/A contains 320 water molecules: 160 for each nucleobase. Thus the error on the nucleotide contribution to the energy is more "diluted" by the lower error on water. To be sure that this splitting is effectively due to this effect and not some other unknown

factor one can simply check that the error on different systems scales with the inverse of the number of water molecules per nucleotide. In this example the final error on the force on the system A is ~ 0.1 and the one on A/A is ~ 0.06; their ratio is ~ 0.6. The inverse ratio between the number of water molecules per nucleotide in the two systems is $100/160 = 6.38$ so the "dilution of error" appears to explain well this difference.

3.3.2 Stability of the model

After a correctly converged training, the stability of the trajectories that the models generates has to be checked. In this case the test set has little significance in unveiling overfitting: the test set is taken from the same trajectories as the training one, so it is possible to spot overfitting only in regions very close to the training, something that is very unlikely to happen. Moreover, as discussed in [60] networks that are trained both on a function and its derivatives do not produce the classical oscillatory overfitting, while they are more prone to point-like failures, in narrow zones of the input space. Those failures are generated from groups of neurons whose responses cancel at all training points [60]. This neurons tend to develop spurious energy minima that can be dramatically wrong outside the region covered by the training set. This failures are much more difficult to identify with a simple test set than classical overfitting.

Given that, the only way to effectively test a model is directly through a MD run. The simplest test one can think of is a stability test: run a MD and see if something blatantly unphysical happens. The typical instability that has been observed is the breaking and formation of bonds, or an abrupt modifications of their geometry. In this case the network predicts spurious energy minima for unphysical conformations. Since energies involved in covalent bonds are quite high (50-100 kcal/mol) this failure can be detected simply looking at the time evolution of the potential energy of the system. See for example figure 3.3.

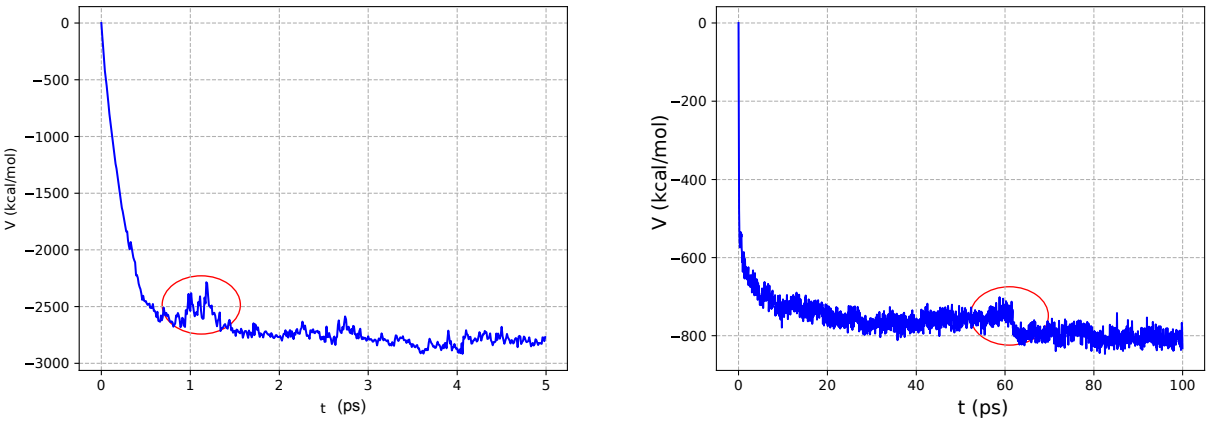


Figure 3.3: How failures typically affect the time evolution of the potential energy: on the right a whole rearrangement of the geometry of a dAMP molecule, on the right the breaking of an O-H covalent bond of a water molecule. Energy is shifted giving a zero value to the starting configuration.

Stability seems to be strongly affected by the choice of the activation function. In general unbounded activation functions give networks that have good performance on training and test set, but are very prone to dramatic point-like failures. Unbounded activation funtion that have been tested are GELU and ELU. Bounded activation functions like the hyperbolic tangent give more stable models, even if they seem to learn slower than unbounded ones. This intuitively makes sense: nodes with unbounded activation function can give arbitrary high outputs and thus develop deep (sometimes hundreds of kcal/mol) spourios minima (see for example figure 3.4). The use of unbounded activation functions was then discarded. Then the effect of the network size was studied.

One of the hardest things to achieve was to obtain potential capable of modeling nucleobases in water as well as pure water. Many models that had a good stability on systems with solvated nucleotides gave bad performances on pure water: the phenomenon that was observed most of the times is the

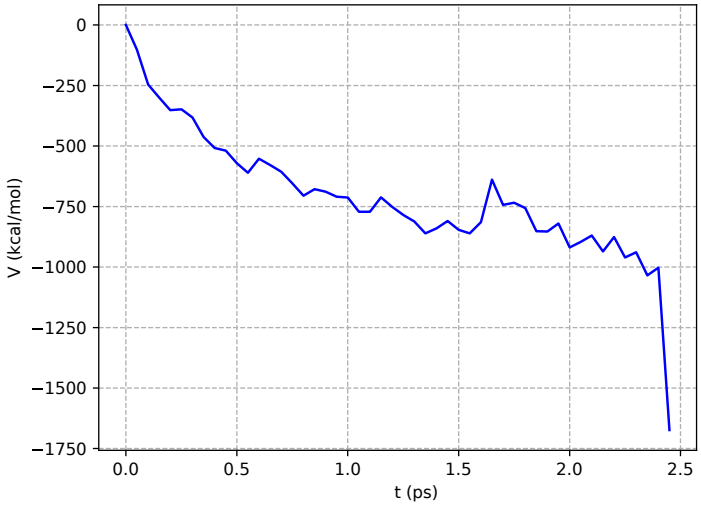


Figure 3.4: Typical failures of networks with unbounded activation functions: the system falls abruptly in a spurious potential energy well.

formation of vacuum bubbles in the simulation box. Interestingly these bubbles form without abrupt changes in the potential energy. Figure 3.5 shows the time evolution of the potential energy of a

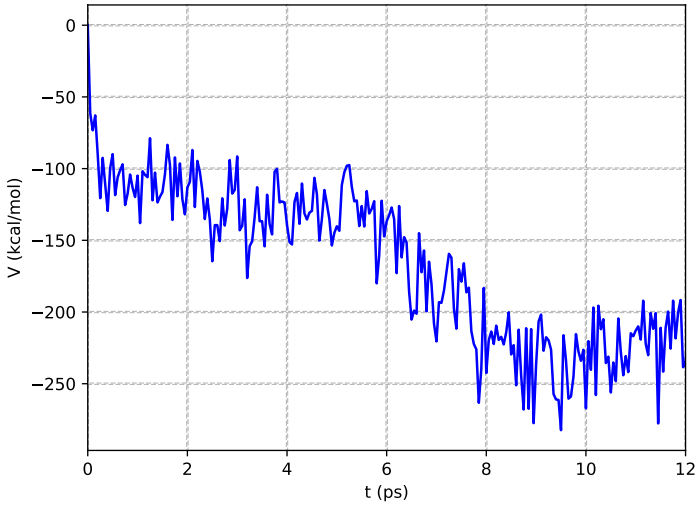


Figure 3.5: Potential energy change during the formation of a vacuum bubble in water.

system with 128 water molecules: for the first 6 ps the model seems to reproduce the properties of water: an inspection of this first part of the trajectory reveals that the O-O and O-H radial distribution functions are qualitatively correct. Than, at 6 ps, a bubble nucleates, and enlarges for 4 ps. The model appears to have a systematic tendency to form surfaces, giving them a negative energy, and the energy gain in creating a surface is high enough to compress the rest of the water.

This is quite surprising, since NN potentials have shown to be capable of reproducing properties of pure water with high accuracy [40], [61], [62]. The only difference in this case is that the training is carried out simultaneously on pure water and water solvating molecules. An intuitive explanation to this failure could be that, since nucleotides exclude a certain volume to water molecules, some neurons are wrongly learning that "empty is better" from the systems containing nucleotides.
The occurrence of this type of failure appears to be linked with the network size. The developers of DeepMD suggest the use of networks that are rather big (three layers with 100-200 neurons per layer for the fitting network and embedding matrices with $M_2 \sim 100$) [40] and their tests on water and metals give good results with networks of this size. In the tests carried out for this thesis such big

networks perform very poorly, and the bubble phenomenon occurs systematically. Smaller networks are less prone to this type of failure.

Moreover, with big networks, it appears that after a certain amount of steps (usually after 500000 steps) the prosecution of the training starts to worsen the model. This can be due to the rich variety of complex atomic environment of the systems under study: group of neurons become highly specialized in modeling certain atomic environment, giving bad performances on others; the larger the network, the more nodes can overspecialize in different environments.

After all it seems that models with $M_1 \sim 10 - 20$ and fitting networks with two layers and $20 - 40$ neurons per layer give better results in term of stability. In literature there is no clear consensus about the right size of the networks for MD applications: many works used small networks with few tenths of nodes [63] [64] [65] [66] [59]. In particular a recent study of water permation in boron nitride [59] used a network with two fitting layers and 20 nodes per layer. This is the system found in the literature that comes closest to the ones studied in this thesis: it certainly lacks of the richness of atomic environments of the present ones, but it contains four different atomic elements (the present ones have six) and water interacting with another compound. Thus the choice of a small network seems to be rather founded.

The architecture that was chosen is the following.

- Fitting NN: 2 layers with 40 neurons each.
- Embedding NN: 2 layers with 10 and 20 neurons.
- DImension of the embedding matrices: $M_1 = 20$, $M_2 = 4$
- Atom types: carbon (C), nitrogen (N), phosphorus (P), nucleotide oxygen (On), nucleotide hydrogen (Hn), water oxygen (On), water hydrogen (Hn), sodium (Na); 8 atom types in total.

Table 3.2 reports the test RMS errors of the model on each system in the dataset.

system	$RMSE_E$ [eV]	$RMSE_f$ [eV/Å]
A	8.1×10^{-3}	9.3×10^{-2}
T	8.4×10^{-3}	9.8×10^{-2}
A/A	6.5×10^{-3}	7.3×10^{-2}
A/T	6.1×10^{-3}	7.8×10^{-2}
T/T	6.3×10^{-3}	47.5×10^{-2}
A-T	5.7×10^{-3}	6.5×10^{-2}
QUAD	7.2×10^{-3}	8.1×10^{-2}
WATER	2.3×10^{-3}	5.2×10^{-2}
Mean	6.3×10^{-3}	7.6×10^{-2}

Table 3.2: Test error on the dataset's systems.

If a model has proven to be stable, than a more refined test is performed, training multiple models with different weight initialization. One model serves to evolve the dynamics of a reference trajectory, while the others are used only to compute energies and forces and check how much their predictions deviate from the reference. A good indicator to look at, as suggested by the developers [40] is the maximum force deviation (MFD). If m+1 different models are used for testing then MFD for a given trajectory frame is defined as:

$$MFD = \max_j \sum_{k=1}^{m} \frac{1}{m} |\mathbf{f}_j^{(0)} - \mathbf{f}_j^{(k)}| \tag{3.2}$$

Where the index j runs over the atoms in the simulation, $|\mathbf{f}_j^{(0)}$ is the force on the atom j predicted by

the model that evolves the dynamics and $|\mathbf{f}_j^{(k)}$ are the forces predicted by the other m testing models. This indicator is useful because it highlights large deviations on single atoms.

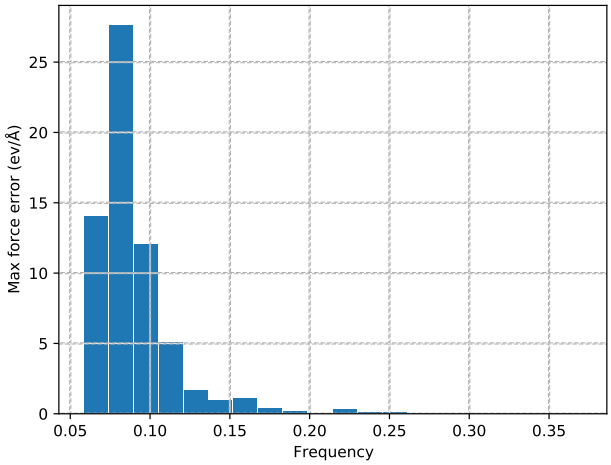


Figure 3.6: Potential energy change during the formation of a vacuum bubble in water.

Figure 3.6 reports the histogram of the MFD for a simulation performed on the system AA. The mean value is $\langle MFD \rangle = 0.1$ eV/Å. This value can be taken as a reasonable estimate of the error made in the force calculation.

3.3.3 MD parameters

For MD simulations with NN potential the Nosè-Hoover thermostat implemented in LAMMPS has been used. The integration timestep has been set to 0.5 fs. All the simulations were carried out in periodic boundary conditions.

Chapter 4

Analysis of MD trajectories

In this chapter the results obained from MD simulations with the NN potential are discussed and compared with the ones given by the OL15 FF and DFT. The model systems that have been studied are pure water and a single dAMP molecule in water. The analysis is focused primarly on the structural properties, characterized using average bond lenghts, dihedral angles and distribution functions. All the radial distribution functions were calculated using the software VMD [67]; the distributions involving angular components and the correlation functions using TRAVIS [68] [69]; the order parameters using the ORDER python package.

4.1 Water

The first model system is a cubic box of 128 water molecules of side 15.6404 Å. The simulations were carried out in the NVT ensemble. DFT water trajectory has been gently provided by the prof. Sulpizi's group. The trajectory is 40 ps plus long and the system was equilibrated for 10 ps before production. The NN and classical MD simulations were carried out from the same starting configuration of the DFT one and both trajectories are 100 ps long, after 10 ps of equilibration.

The simplest way to characterize the local structure of liquid water is through radial distribution functions. The following notation is adopted: let the atoms be labeled with indexes $i \in \{1, 2 \dots N\}$ and the sets of atoms of the same type be denoted with capital letters $A,\ B \dots$. Atomic types can be simply the atom elements or more specific categories, like the oxygens of water molecules, or the aromatic carbon atoms.
The radial distribution function (rdf) between two atomic types is defined as:

$$g_{AB}(r) = \frac{V}{N_A N_B} \left\langle \sum_{i \in A} \sum_{j \in B} \delta\big(r - |\mathbf{r}_i - \mathbf{r}_j|\big) \right\rangle \tag{4.1}$$

Water is characterized using oxygen-oxygen and oxygen-hydrogen rdfs.

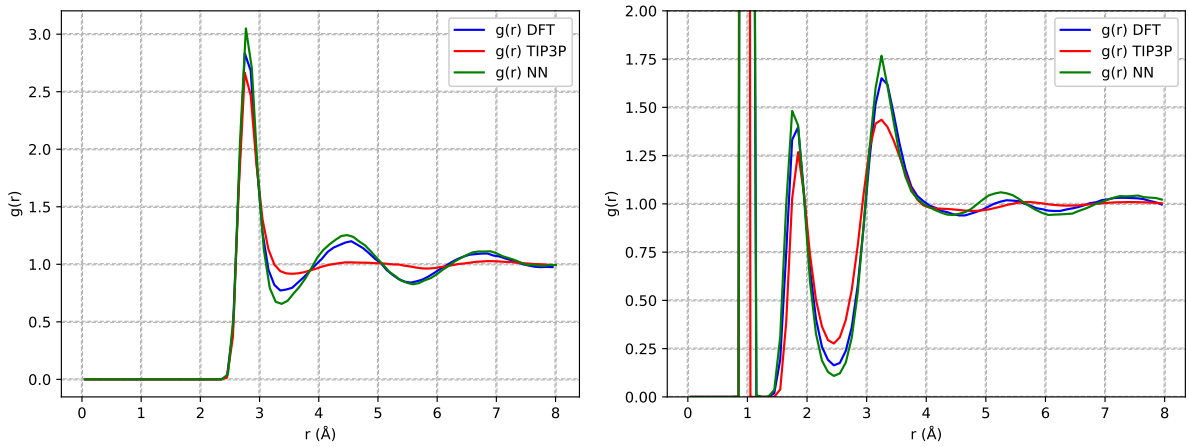


Figure 4.1: Water O-O (left) and O-H (right) radial distribution functions.
O-O first peak integrals: $I_1^{DFT} = 4.7 \quad I_1^{OL15} = 4.7 \quad I_1^{NN} = 4.6$
O-H second peak integrals (hydrogen bond peak): $I_1^{DFT} = 2.0 \quad I_1^{OL15} = 1.9 \quad I_1^{NN} = 2.0$

As shown in figure 4.1, the MD simulations performed with the NN potential reproduces quite well the rdfs obtained ab initio. However the precision of this result is lower than other published ones, obtained with similar methods, for example in [40]. This is probably due to the difficulty in learning simultaneously water-water interactions and water-solute interactions. Another descriptor that gives important information is the angular distribution function (adf). The angular distribution function between two vectors is defined as:

$$g_{\mathbf{r}_1,\mathbf{r}_2}(\theta) = \frac{1}{\sin(\theta)} \left\langle \sum_{i \in A} \sum_{j \in B} \delta\big(\theta - \theta(\mathbf{r}_1, \mathbf{r}_2)\big) \right\rangle \tag{4.2}$$

In the case of water two adfs are particularly important. The first one is the adf between the vector going from an hydrogen to the oxygen of the same molecule and the vector going from the same hydrogen to an oxygen of another water molecule; it will be called, for simplicity, OHO adf. The second one is between the vectors going from an hydrogen to the oxygen of the same molecule and the vector going from the oxygen of another molecule to an hydrogen of the same molecule; it will be called OHOH adf. These vectors are depicted in figure 4.2.

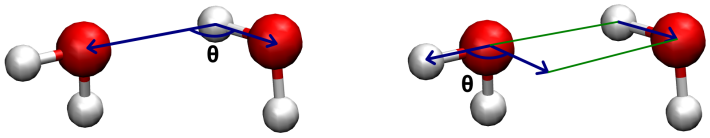


Figure 4.2: Depiction of the vectors defining OHO (left) and OHOH (right) adfs.

The rationale behind these definitions is to obtain information about the directionality of the hydrogen bond. Usually the hydrogen bond distances are reproduced quite well by FFs, as can be seen in

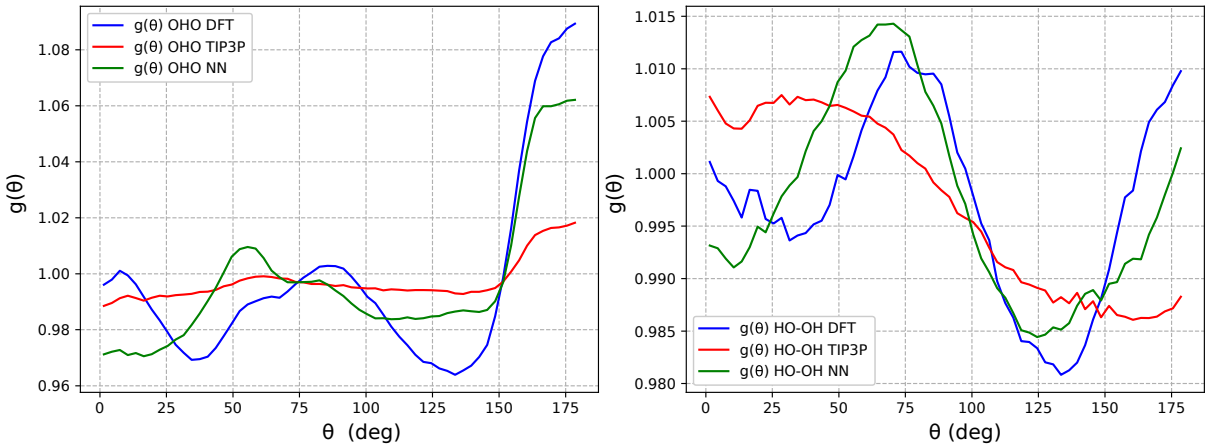


Figure 4.3: OHO (left) and OHOH (right) adfs.

figure 4.1, but they fail in the description of its directionality (figure 4.3); hydrogen bonds in FFs are the result of the interplay between coulomb and Lennard-Jones interactions, and this two potentials are completely centrosymmetric: they cannot reproduce the directionality of a real hydrogen bond, which orientation is determined by the oxygen lone pairs. The NN appears to do better, capturing at least qualitatively the preferred mutual orientations of water molecules.

The last descriptor that has been computed is the tetrahedral parameter q, as defined in [70]. This parameter is calculated locally for every oxygen atom (here labeled as i) as:

$$q_i = 1 - \frac{3}{8} \sum_{ji=1}^{3} \sum_{k=j+1}^{4} \left(\cos \psi_{ijk} + \frac{1}{3} \right)^2 \tag{4.3}$$

where ψ_{ijk} is the angle formed by the three oxygens and the sum runs over the 4 nearest neighbors of the atom i. If q_i is equal to one than the $i-th$ oxygen is the center of a perfect tetrahedron with its 4 nearest neighbors at the corners. Its average $\langle q \rangle$ is an order parameter of the water-tetrahedral ice transition [71]. Figure 4.4 shows the probability distribution of the q parameter obtained using the three methods. Even in this case the predictions of the NN potential are in line with DFT data. In general it seems that the model is capable of describing the local structure of water at a satisfactory level of precision.

Another interesting aspect is the dynamics of the hydrogen bond formation and breaking: this as-

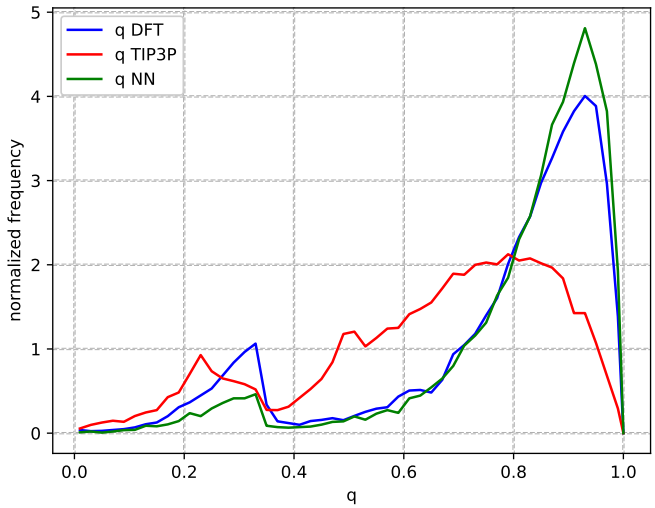


Figure 4.4: Distribution of the tetrahedral parameter q.

pect can be studied observing the changes in mutual orientation of water molecules in time. Reorientation of water molecules is connected with the oscillation and rearrangements of the hydrogen bonds [72]. Water reorientation has two components: a fast oscillatory one, called libration, with typical timescales of ~ 0.1 ps, due to thermal motion of the acceptor hydrogen atom around its equilibrium bonded position; and a slow one, due to hydrogen bond breaking and slow rotations of the whole H-bonded water dimer, with a typical timescale of 10 ps. The simplest way to quantify of

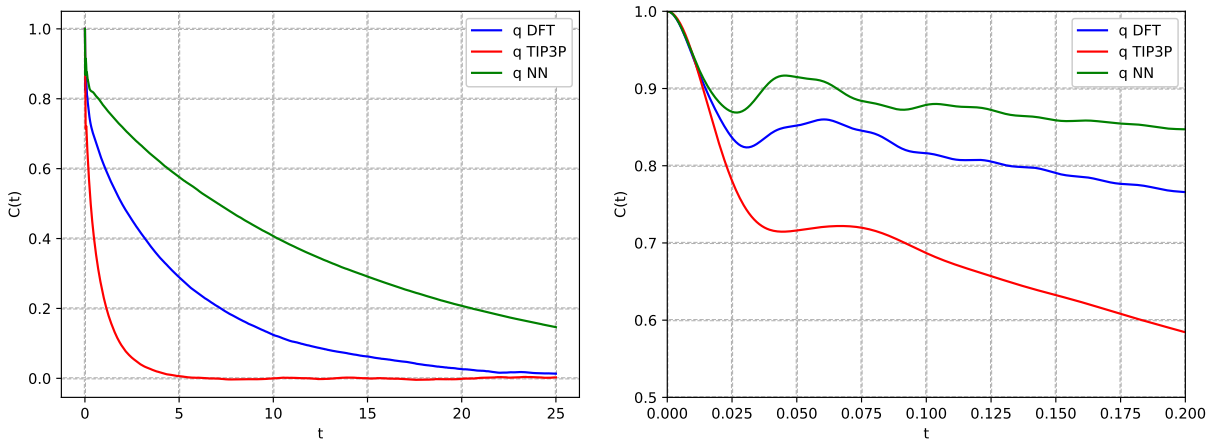


Figure 4.5: Orientational autocorrelation function $C_2(t)$ (left), a focus of the same graph on the sub-picosecond region (right).

reorientation dynamics is to compute a time autocorrelation function of the molecular orientation. Usually the time correlation function is defined in terms of the second Legendre polynomial of the dot product between a vector defining water orientation at time t_0 and at time $t_0 + t$ [72].

$$C_2(t) = \langle P_2\big(\mathbf{u}(t_0) \cdot \mathbf{u}(t_0 + t)\big)\rangle \tag{4.4}$$

Here $\mathbf{u}$ is as an adimensional orientation vector. Figure 4.5 shows the results obtained with the three different methods. The NN potential gives a good description of the libration component, but seems to fail in describing correctly the long time decorrelation. The decay time of the hydrogen bond is usually quantified as the integral of the correlation function: $\tau = \int_0^\infty C_2(t)dt$ [72]. The results in the

present case are:

$$\begin{aligned} \tau_{DFT} &= 4.21 \text{ ps} \\ \tau_{TIP3P} &= 0.69 \text{ ps} \\ \tau_{NN} &= 10.83 \text{ ps} \end{aligned}$$

The hydrogen bonds predicted by the NN potential "live longer" than the ones predicted ab initio. Classical force field, on the other hand, produces a very fast decorrelation compared to DFT data.

4.2 Free dAMP in water: molecular conformations

Free dAMP in water is a rather complex system to describe: the NN has to learn in a consistent way many different interactions at the same time. Reproducing the intramolecular interactions is by its own a challenging task, and the interaction with the surrounding water and ions is even more complex. In this section the question on how good the NN potential can reproduce the internal conformations of dAMP in water is addressed. The system that has been studied consists in dAMP molecule solvated with 101 water molecules and 3 sodium ions, in a cubic box of side 15 Å, starting from the same initial configuration a DFT, a NN, and a classical FF trajectory were generated. All the trajectories were equilibrated for 10 ps. The DFT trajectory is 20 ps long, while the other two are 100 ps.

4.2.1 Bond lengths

The simplest parameter to look at is the mean bond length. In figure 4.6 the mean bond lengths predicted by DFT are plotted against the ones predicted by NN potential.

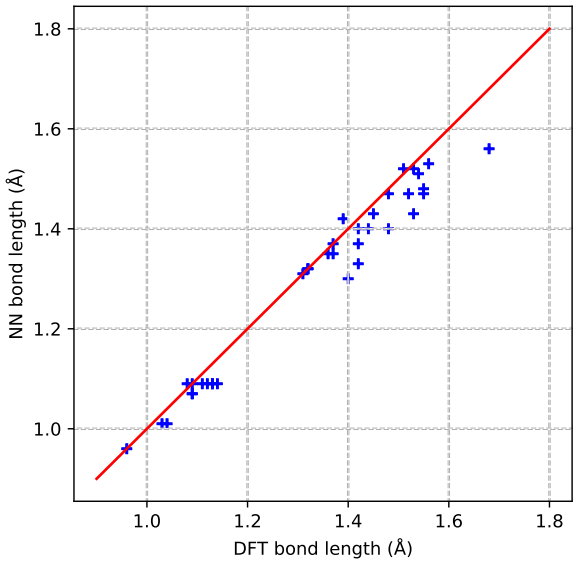


Figure 4.6: Correlation of bond lengths predicted by DFT and by NN.

The RMS error on bond lengths is $\sigma_{bond} = 0.047$ Å. It can be noticed that there is a tendency to slightly underestimate bond lengths. It is worth to point out that the NN predicts some bonds to have exactly the same lengths: a closer look reveals that this bonds are all between identical atom pairs. For example the seven C-H bonds of the ribose are all predicted to be 1.09 Å long: the network tends to describe all hydrogen atoms bonded to a ribose carbon in the same way.

4.2.2 Nucleobase dihedrals

Other important parameters are the nucleobase dihedral angles. Their equilibrium values and probability distribution say how much the nucleobase remains flat during a simulation.

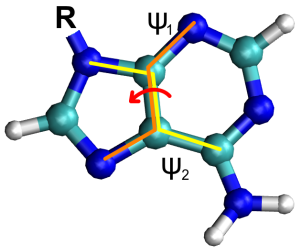


Figure 4.7: Definition of ψ_1 and ψ_2 dihedral angles.

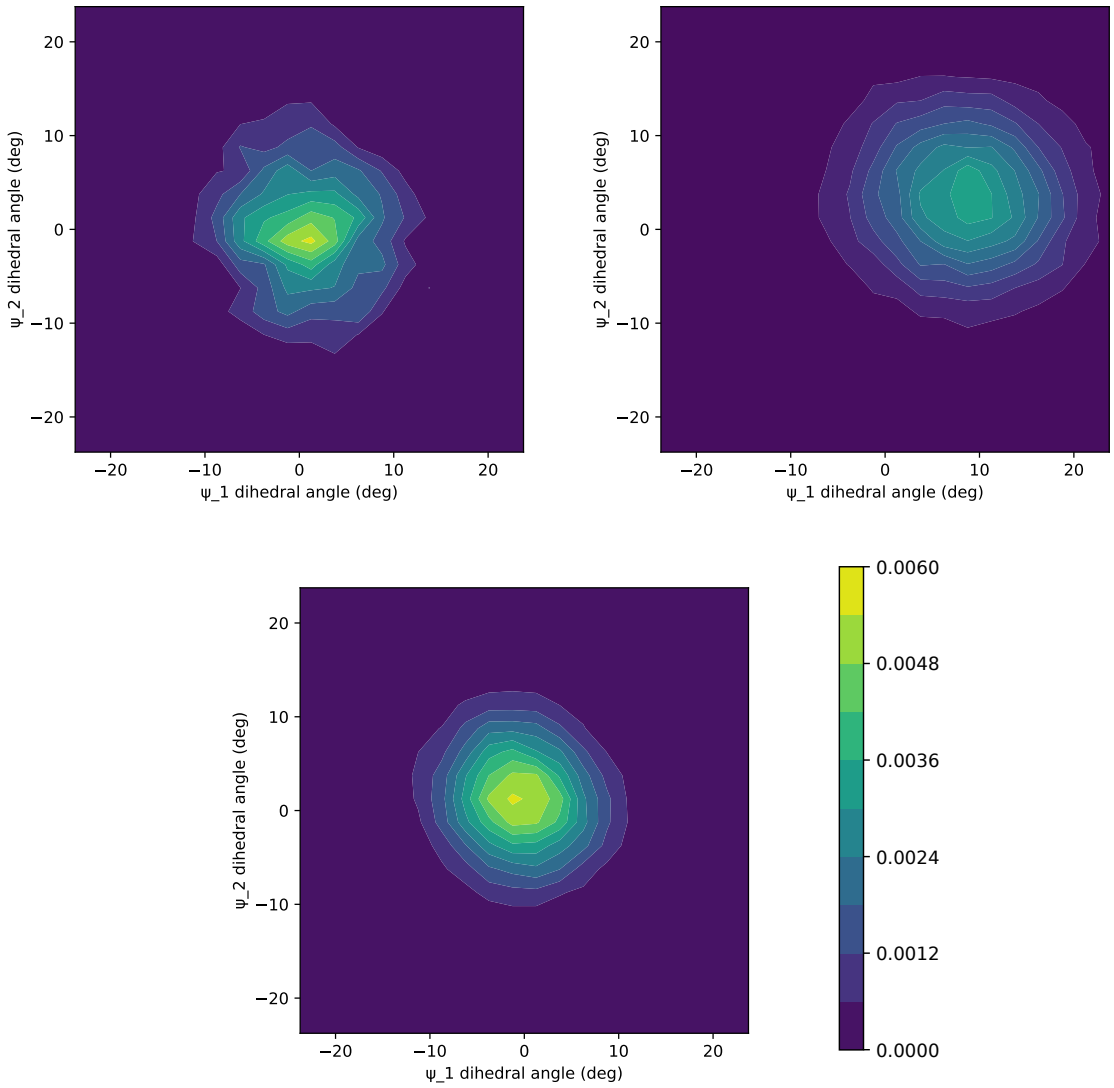


Figure 4.8: Iso level plot of the probability distribution of the two dihedral angles ψ_1 and ψ_2 computed from DFT MD (up left), NN MD (up right) and FF MD (down).

In figure 4.7 the two dihedral angles that were chosen for this analysis are defined. Figure 4.8 shows the joint probability distribution of the two dihedrals calculated from a DFT, NN and FF simulations. Figure 4.9 reports instead the separate distributions of the two angles calculated with the three methods. The values have been shifted in order to have a zero angle when all the three vectors defining the dihedrals lie on the same plane. The NN predicts a shift of the minimum from the correct value, i.e. the nucleobase is not flat at the NN energy minimum, but there is a spurious torsion of its bonds. The OL15 FF gives correct nucleobase torsions.

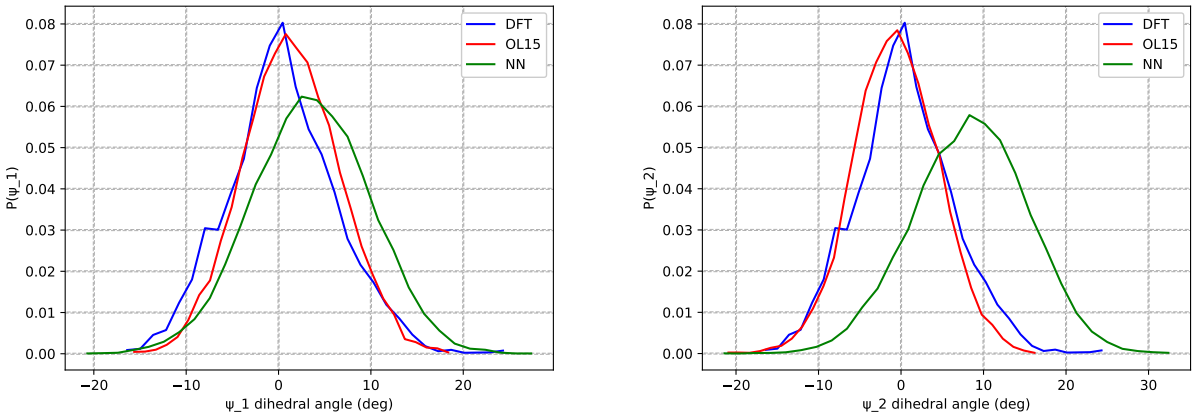


Figure 4.9: Comparison of the ψ_1 (left) and ψ_2 (right) distributions.

4.2.3 Base-ribose and ribose-phosphate torsional angles

Two other important angles are the ribose-phosphate and the nucleobase-phosphate dihedrals. Their value defines the mutual orientation of the three parts of the nucleotide.

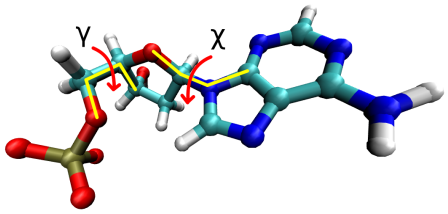


Figure 4.10: Definition of χ and γ dihedral angles.

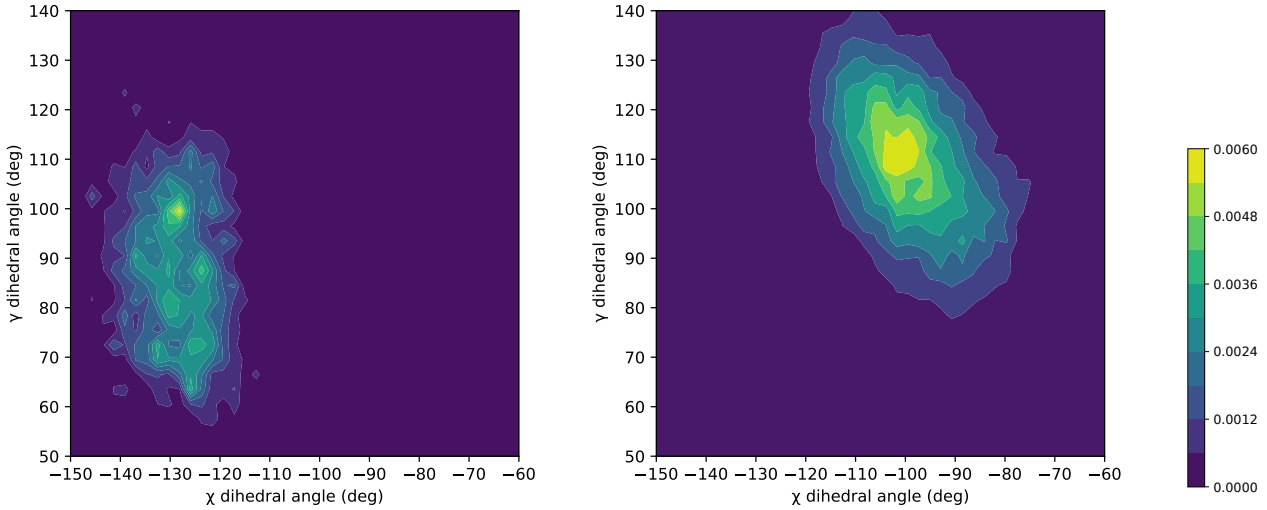


Figure 4.11: Iso level plot of the joint $\chi - \gamma$ distribution, computed from DFT MD (left) and NN MD (right).

Figure 4.10 depicts the definition of the two dihedral angles. In figures 4.11 and 4.12 the distribution of the dihedrals from NN MD is compared with the distribution from DFT MD. Mean values of the dihedrals from DFT simulations are: $\langle\chi\rangle = -116°$ $\langle\gamma\rangle = 58°$, while the NN predicts $\langle\chi\rangle = -97°$ $\langle\gamma\rangle = 103°$. The NN fails in describing both dihedrals, in particular γ is shifted by $45°$.

These dihedrals are not correctly described by the classical force field either: since OL15 parameters are tuned on helical structures of DNA, χ and γ are strongly restrained in order to reproduce

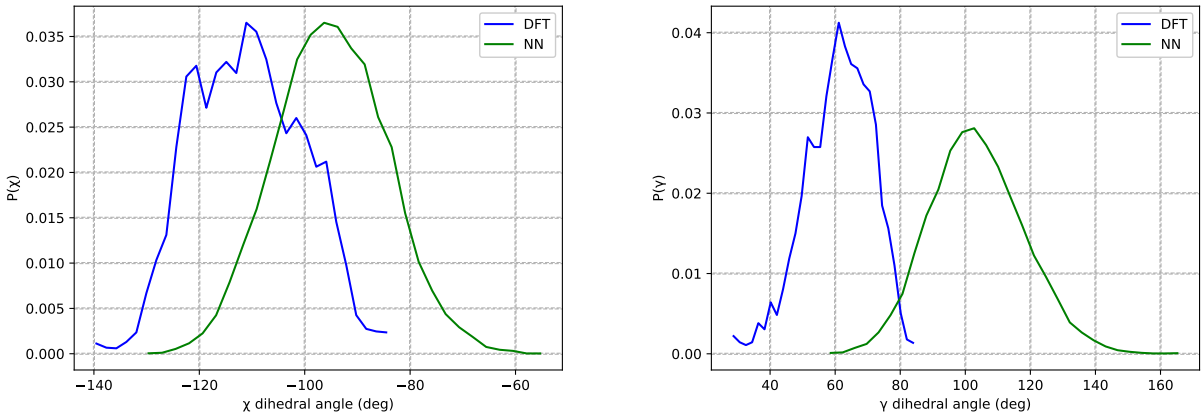


Figure 4.12: Comparison of the χ (left) and γ (right) distributions obtained ab initio and with NN potential.

this geometries. This can be seen in figure 4.13, that shows the joint probability distribution of the two dihedrals calculated from a 100 ns trajectory evolved with OL15 force field.The four different peaks correspond to different conformation that are found in biological polynucleotides [73] are not compatible with backbone free dAMP in solution.

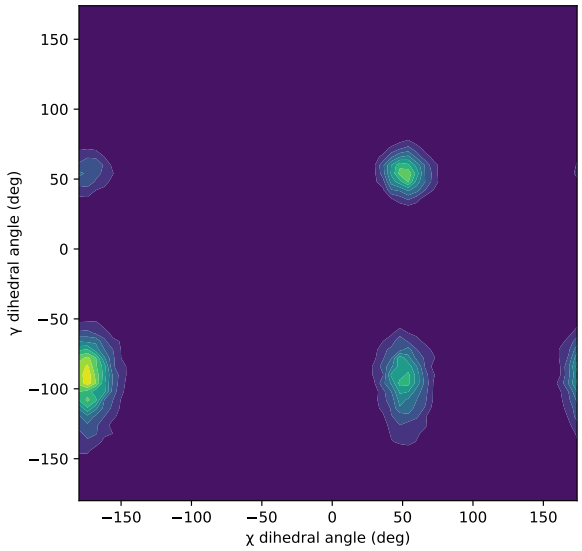


Figure 4.13: Iso level plot of the $\chi - \gamma$ distribution during a 100 ns classical simulation.

4.2.4 Phosphate group solvation shell

Water interacts with dAMP mainly forming hydrogen bonds: the phosphate group is a strong hydrogen bond acceptor, and some nitrogens of the nucleobase are acceptors too. We begin by studying the interaction between solvent and phosphate group. Figure 4.14 shows the rdf between water hydrogens and phosphate oxygens. The NN appears to reproduce quite well the solvation of the phosphate, giving the right value for the first peak integral, as well as matching the structure of the outer solvation shells. The OL15 FF, in contrast, tends to overestimate the strength of water-phosphate H-bonds, ginving a higher value for the first peak integral.

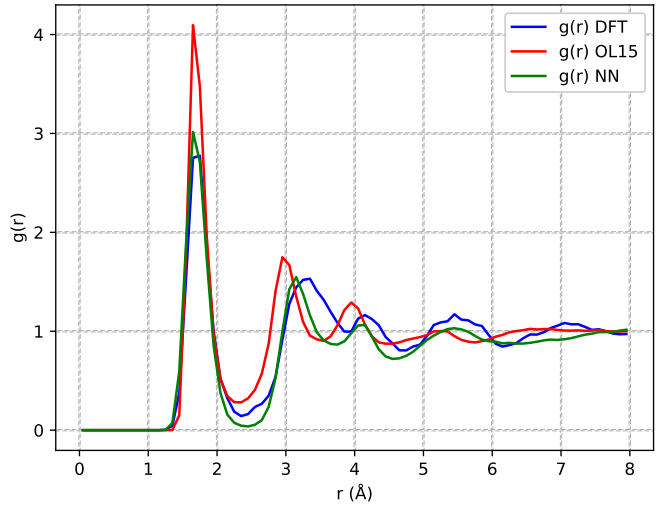


Figure 4.14: Op-Hw radial distribution function.
First peak integrals: $I_1^{DFT} = 2.7 \quad I_1^{OL15} = 3.2 \quad I_1^{NN} = 2.6$.

4.2.5 Interaction with ions

Another interesting feature of the phosphate group is its interaction with sodium ions. The phosphate group is charged, it can directly bind ions, or it can interact with them through the water molecules in the solvation shell.

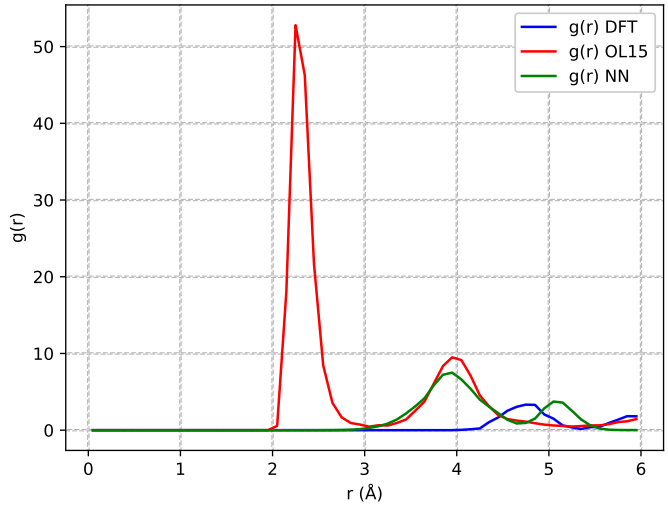


Figure 4.15: Op-Na radial distribution function.

Figure 4.15 shows the rdf between sodium ions and phosphate oxygens. The FF appears to predict an incorrect binding of the ions. The first sharp peak in the rdf predicted by the FF is at 2.4 Å, a distance that is very close to the Van der Waals radius of sodium, which is 2.3 Å. Thus the sodium can be considered completely bound to the phosphate. This does not happens in the MD simulations with DFT and NN potential: the peak at 3.8 Å that is seen in the NN rdf indicates that the sodium ion interacts with the phosphate group through the solvation shell. A closer look at the trajectory reveals that both DFT and NN predicts a configuration in which the ion is bound to the N7 nitrogen of the nucleobase and to the phosphate through a water molecule, forming a bridge-like structure, as shown in figure 4.16.

This structure correspond to the peak at 4.8 Å of NN rdf and at 5.2 Å of the DFT rdf in figure 4.15.

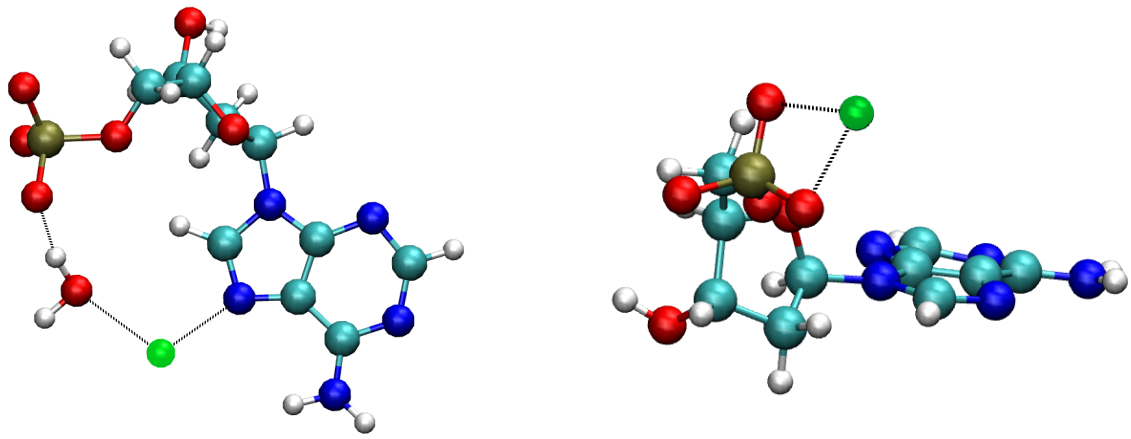


Figure 4.16: Different binding configurations predicted by DFT and NN (left) and by FF (right).

This structure remains stable for the whole DFT MD, while in the NN simulation the sodium moves after some times and remains bound only to the phosphate, giving the 3.8 Å peak in the rdf.

4.2.6 Adenine solvation shell

The solvation shell of the adenine has been studied looking at water hydrogen - acceptor nitrogen, and water oxygen - aminic hydrogen radial distribution functions.

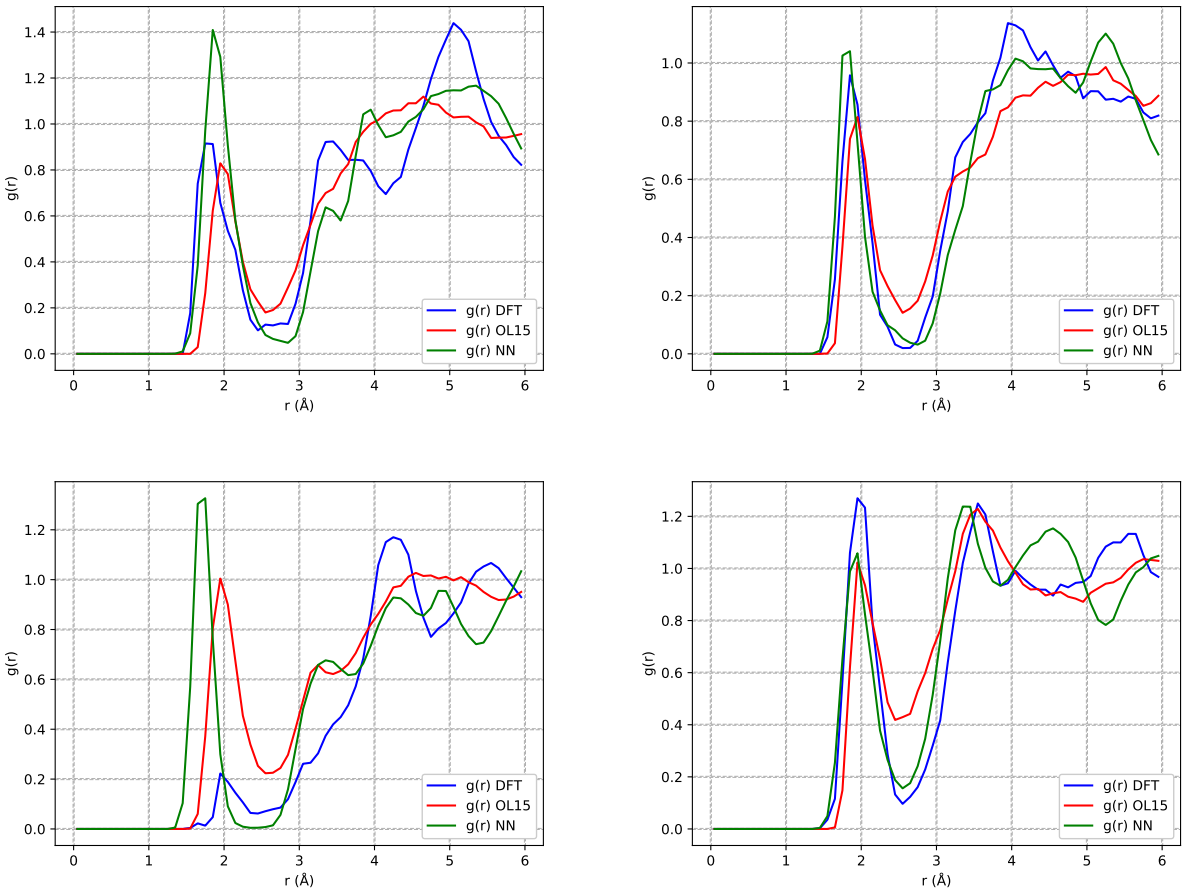


Figure 4.17: N1-Hw (up left), N3-Hw (up right) N7-Hw (down left) and H6-Ow (down right) radial distribution functions.

In figure 4.18 the nomenclature of these atoms, introduced in section 1.1, is recalled. In figure 4.17 the three rdfs centered on the acceptor nitrogens, plus the rdf centered on the two equivalent aminic

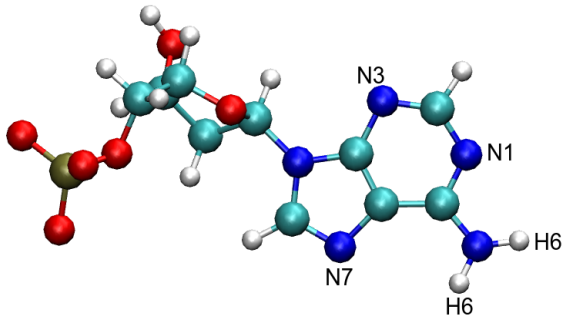


Figure 4.18: Nomenclature of donor nitrogens and aminic hydrogens.

hydrogens are plotted. The NN reproduces well the rdfs obtained from DFT MD, except for the first peak in the N7 rdf: this is due to the stability of the binding configuration previously described, that lasts for the whole DFT trajectory, preventing the water from forming H-bonds with the N7 nitrogen. The DFT trajectory is probably too short to observe the unbinding of the sodium ion and the formation of H-bonds on N7, but this is probably what would happen during a longer trajectory. This is probably the source of the discrepancy in the third plot of figure 4.17.

4.3 dAMP-dAMP stacking dimer

At the moment the NN potential is not capable of reproducing correctly the stacking interaction. During a DFT MD carried out for 20 ps the stacking dimer remains stable, while in MD runs with the NN potential the stacking dimer dissociates after 10 ps. To have a better, although approximate, idea of the strength of the stacking interaction, a free energy calculation by umbrella sampling has been performed with classical OL15 force field. The collective variable used is simply the distance between the centers of mass of the nucleobases.

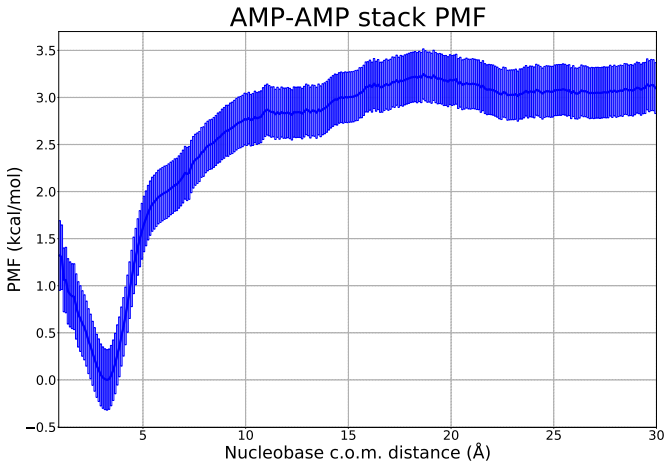


Figure 4.19: dAMP/dAMP stack PMF.

Figure 4.19 shows the PMF profile along the center of mass distance. The estimated free energy of dimerization is $\Delta G_{dim}^{dAMP} = -3.2 \pm 0.4$ kcal/mol. This value can be compared with the dimerization constant for ADP that was calculated from experimental data in [17] (see also section 1.2.5). The value proposed for the dimerization constant is $K_{dim}^{dADP} = (5 \pm 3) \times 10^3$, at which corresponds a free energy of dimerization $\Delta G_{dim}^{dADP} = -5.4 \pm 0.2$ kcal/mol. The value calculated for dAMP dimers is not completely reliable, being computed using the classical FF, but it is of the same order of magnitude of an experimental value for a very similar molecule, so it can be taken at least as a rough estimation of the stacking interaction strength.

Conclusions and future perspectives

The use of machine learning approaches to bridge the gap between AIMD accuracy and long timescales is currently an active, rapidly developing, area of research.

Liquid crystal aggregation of nucleotides in solution is a phenomenon that involves many different interactions and effects at different scales. The description of such molecules, their mutual interactions and their solvation can correctly be achieved by quantum mechanical calculations, but the time-scales that are accessible to these methods is not long enough to study complex supramolecular aggregates like liquid crystals. Using a neural network it is possible, in principle, to fit ab initio reference data and to obtain a potential energy function that is accurate and fast to evaluate. Using Deep Potential, a neural network architecture that was previously used only on water, solids and gas-phase molecules, the possibility of simulating nucleotides in water has been explored. The model has been trained on a set of ab initio trajectories of nucleotides in water; it was capable of fitting the training set with a mean accuracy on 10^{-3} eV on energies and 0.1 eV/Å on the forces. At the current state the model is capable of carrying out a stable molecular dynamics simulation for a time of the order of hundreds of picoseconds.

The model has been tested on three model systems: pure water, a solvated dAMP molecule and a solvated dAMP stacking dimer. The properties of the solvent are correctly reproduced, including the hydrogen bond formation and its directionality. The correlation time of water orientation is overestimated: this indicates the presence of spurious energy barriers that make the hydrogen bond kinetics slower. The geometry of the dAMP molecule is predicted correctly, and the solvation properties of the nucleobase and phosphate group are well reproduced. Learning correctly dihedral angles seems to be one of the main difficulties: in particular the model gives incorrect predictions for the ribose-phosphate torsion and the nucleobase is not predicted to be perfectly planar. The stacking interaction between different dAMP molecules is not reproduced at the moment: the stacking dimer is stable during a 20 ps ab initio trajectory, while it dissociates after 10 ps of a trajectory evolved with the NN potential.

Future developments could consist in the augmentation of the training set, using iterative procedures as described in [40], and the construction of models with more atom types associated to different hybridization states of the same elements. Another approach could be to combine a classical force field with a neural network potential, like in [74]. The idea in this case is to promote the force field to a DFT accuracy adding a neural network term to it. This approach could solve some stability problems like bond breaking in long simulations. NN appears to perform well in reproducing intermolecular interactions like the hydrogen bond, but has difficulties in learning complex covalent geometries. An approach that combines NN and FF together would possibly take benefit of the strengths of the two approaches: stability and simplicity of FF binding geometries and good description of many body intermolecular interactions of NN potentials.

Bibliography

[1] Norio Kitadai and Shigenori Maruyama. Origins of building blocks of life: A review. *Geoscience Frontiers*, 9(4):1117 – 1153, 2018.

[2] Tommaso P. Fraccia, Gregory P. Smith, Lucas Bethge, Giuliano Zanchetta, Giovanni Nava, Sven Klussmann, Noel A. Clark, and Tommaso Bellini. Liquid crystal ordering and isotropic gelation in solutions of four-base-long DNA oligomers. *ACS Nano*, 10(9):8508–8516, 2016.

[3] Giuliano Zanchetta, Tommaso Bellini, Michi Nakata, and Noel A. Clark. Physical polymerization and liquid crystallization of RNA oligomers. *Journal of the American Chemical Society*, 130(39):12864–12865, 2008.

[4] Gregory P. Smith, Tommaso P. Fraccia, Marco Todisco, Giuliano Zanchetta, Chenhui Zhu, Emily Hayden, Tommaso Bellini, and Noel A. Clark. Backbone-free duplex-stacked monomer nucleic acids exhibiting watson–crick selectivity. *Proceedings of the National Academy of Sciences*, 115(33):E7658–E7664, 2018.

[5] Marco Todisco, Tommaso P. Fraccia, Greg P. Smith, Andrea Corno, Lucas Bethge, Sven Klussmann, Elvezia M. Paraboschi, Rosanna Asselta, Diego Colombo, Giuliano Zanchetta, Noel A. Clark, and Tommaso Bellini. Nonenzymatic polymerization into long linear RNA templated by liquid crystal self-assembly. *ACS Nano*, 12(10):9750–9762, 2018.

[6] Rodrigo Galindo-Murillo, James C. Robertson, Marie Zgarbová, Jiří Šponer, Michal Otyepka, Petr Jurečka, and Thomas E. Cheatham. Assessing the current state of AMBER force field modifications for DNA. *Journal of Chemical Theory and Computation*, 12(8):4114–4127, 2016.

[7] Stephen Neidle. 2 - the building-blocks of DNA and RNA. In Stephen Neidle, editor, *Principles of Nucleic Acid Structure*, pages 20 – 37. Academic Press, New York, 2008.

[8] Chelsea R. Martinez and Brent L. Iverson. Rethinking the term "pi-stacking". *Chem. Sci.*, 3(7):2191–2201, 2012.

[9] Christopher A. Hunter and Jeremy K. M. Sanders. The nature of π-π. interactions. *Journal of the American Chemical Society*, 112(14):5525–5534, 1990.

[10] Stefan Grimme. Do special noncovalent π–π stacking interactions really exist? *Angewandte Chemie International Edition*, 47(18):3430–3434, 2008.

[11] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–574, 01 2006.

[12] Jiří Černý, Martin Kabeláč, and Pavel Hobza. Double-helical $\rightarrow$ ladder structural transition in the B-DNA is induced by a loss of dispersion energy. *Journal of the American Chemical Society*, 130(47):16055–16059, 2008.

[13] Paul O. P. Ts'o, Ingelore S. Melvin, and Alfred C. Olson. Interaction and association of bases and nucleosides in aqueous solutions. *Journal of the American Chemical Society*, 85(9):1289–1296, 1963.

[14] Arthur D. Broom, Martin P. Schweizer, and Paul O. P. Ts'o. Studies of the association of purine nucleosides by vapor pressure osmometry and by proton magnetic resonance. *Journal of the American Chemical Society*, 89(14):3612–3622, 1967.

[15] Will E. Ferguson, Charles M. Smith, E.T. Adams, and Grant H. Barlow. The temperature-dependent self-association of adenosine 5'triphosphate in 0.154 M NaCl. *Biophysical Chemistry*, 1(5):325 – 337, 1974.

[16] James L. Weaver and Robert W. Williams. Raman spectroscopic measurement of base stacking in solutions of adenosine, AMP, ATP, and oligoadenylates. *Biochemistry*, 27(25):8899–8903, 1988.

[17] Fernando Peral and Ernesto Gallego. The self-organization of adenosine 5'-triphosphate and adenosine 5'-diphosphate in aqueous solution as determined from ultraviolet hypochromic effects. *Biophysical Chemistry*, 85(1):79 – 92, 2000.

[18] Michael P. Robertson and Gerald F. Joyce. The origins of the RNA world. *Cold Spring Harbor Perspectives in Biology*, 2010.

[19] Elizabeth A. Doherty and Jennifer A. Doudna. Ribozyme structures and mechanisms. *Annual Review of Biochemistry*, 69(1):597–615, 2000.

[20] Nathalie Sartori Blanc, Alfred Senn, Amélie Leforestier, Françoise Livolant, and Jacques Dubochet. DNA in human and stallion spermatozoa forms local hexagonal packing with twist and many defects. *Journal of Structural Biology*, 134(1):76 – 81, 2001.

[21] Tamar Schlick. *Molecular Modeling and Simulation: An Interdisciplinary Guide*. Springer-Verlag, Berlin, Heidelberg, 2002.

[22] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76(1):637–649, jan 1982.

[23] D. J. Evans and B. L. Holian. The nose–hoover thermostat. *The Journal of Chemical Physics*, 83(8):4069–4074, 1985.

[24] Andrew Jones and Ben Leimkuhler. Adaptive stochastic methods for sampling driven molecular systems. *The Journal of Chemical Physics*, 135(8):084125, 2011.

[25] Johannes Kästner. Umbrella sampling. *WIREs Computational Molecular Science*, 1(6):932–942, 2011.

[26] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[27] Paul Peter Ewald. Die berechnung optischer und elektrostatischer gitterpotentiale. *Annalen der Physik*, 369(3):253–287, 1921.

[28] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh ewald: An nlog(n) method for ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993.

[29] Jürg Hutter Gerald Lippert and Michele Parrinello. A hybrid gaussian and plane wave density functional scheme. *Molecular Physics*, 92(3):477–488, 1997.

[30] Joost VandeVondele, Matthias Krack, Fawzi Mohamed, Michele Parrinello, Thomas Chassaing, and Jürg Hutter. Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Computer Physics Communications*, 167(2):103 – 128, 2005.

[31] Michael Gastegger and Philipp Marquetand. Molecular dynamics with neural-network potentials. 12 2018.

[32] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.*, 120:143001, Apr 2018.

[33] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan 2015.

[34] Murat Sazli. A brief review of feed-forward neural networks. *Communications, Faculty Of Science, University of Ankara*, 50:11–17, 01 2006.

[35] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251 – 257, 1991.

[36] Stuart Dreyfus. The numerical solution of variational problems. *Journal of Mathematical Analysis and Applications*, 5(1):30 – 45, 1962.

[37] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[38] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007.

[39] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013.

[40] Han Wang, Linfeng Zhang, Jiequn Han, and Weinan E. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*, 228:178 – 184, 2018.

[41] Linfeng Zhang, Jiequn Han, Han Wang, Wissam Saidi, Roberto Car, and Weinan E. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4436–4446. Curran Associates, Inc., 2018.

[42] D. van der Spoel H. J. C. Berendsen and R. van Drunen. Gromacs - a message-passing parallel molecular-dynamics implementation. *Computer Physics Communications*, 91(1-3):43–56, 1995.

[43] Jiří Šponer Michal Otyepka Thomas E. Cheatham Rodrigo Galindo-Murillo Petr Jurečka Marie, Zgarbová. Refinement of the sugar–phosphate backbone torsion beta for AMBER force fields improves the description of Z- and B-DNA. *Journal of Chemical Theory and Computation*, 11(12):5723–5736, 2015.

[44] Kristin L. Meagher, Luke T. Redman, and Heather A. Carlson. Development of polyphosphate parameters for use with the AMBER force field. *Journal of Computational Chemistry*, 24(9):1016–1025, 2003.

[45] Herman J. C. Berendsen Johannes G. E. M. Fraaije Berk Hess, Henk Bekker. LINCS: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[46] Thomas D. Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V. Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z. Khaliullin, Ole Schütt, Florian Schiffmann, Dorothea Golze, Jan Wilhelm, Sergey Chulkov, Mohammad Hossein Bani-Hashemian, Valéry Weber, Urban Borštnik, Mathieu Taillefumier, Alice Shoshana Jakobovits, Alfio Lazzaro, Hans Pabst, Tiziano Müller, Robert Schade, Manuel Guidon, Samuel Andermatt, Nico Holmberg, Gregory K. Schenter, Anna Hehn, Augustin Bussy, Fabian Belleflamme, Gloria Tabacchi, Andreas Glöß, Michael Lass, Iain Bethune, Christopher J. Mundy, Christian Plessl, Matt Watkins, Joost VandeVondele, Matthias Krack, and Jürg Hutter. Cp2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics*, 152(19):194103, 2020.

[47] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100, Sep 1988.

[48] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B*, 37:785–789, Jan 1988.

[49] Dario Alfè Michael J. Gillan and Angelos Michaelides. Perspective: How good is DFT for water? *The Journal of Chemical Physics*, 144(13):130901, 2016.

[50] S. Goedecker, M. Teter, and J. Hutter. Separable dual-space gaussian pseudopotentials. *Phys. Rev. B*, 54:1703–1710, Jul 1996.

[51] Joost VandeVondele and Jürg Hutter. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *The Journal of Chemical Physics*, 127(11):114105, 2007.

[52] Matthew J. McGrath J. Ilja Siepmann Joost VandeVondele Michiel Sprik Jürg Hutter Bin Chen Michael L. Klein Fawzi Mohamed Matthias Krack Michele Parrinello I-Feng W. Kuo, Christopher J. Mundy. Liquid water from first principles: Investigation of different sampling approaches. *The Journal of Physical Chemistry B*, 108(34):12990–12998, 2004.

[53] Olivia Lynes, Jonathan Austin, and Andy Kerridge. Ab initio molecular dynamics studies of hydroxide coordination of alkaline earth metals and uranyl. *Phys. Chem. Chem. Phys.*, 21:13809–13820, 2019.

[54] Stefan Grimme, Jens Antony, Stephan Ehrlich, and Helge Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction DFT-D for the 94 elements H-Pu. *The Journal of Chemical Physics*, 132(15):154104, 2010.

[55] Kruse Holger, Pavel Banáš, and Jiří Šponer. Investigations of stacked DNA base-pair steps: Highly accurate stacking interaction energies, energy decomposition, and many-body stacking effects. *Journal of Chemical Theory and Computation*, 15(1):95–115, 2019.

[56] AV Fratini, ML Kopka, HR Drew, and RE Dickerson. Reversible bending and helix geometry in a B-DNA dodecamer: CGCGAATTCGCG. *The Journal of biological chemistry*, 257(24):14686—14707, December 1982.

[57] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[58] Steve Plimpton. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics*, 117(1):1 – 19, 1995.

[59] Hossein Ghorbanfekr, Jörg Behler, and François M. Peeters. Insights into water permeation through hbn nanocapillaries by ab initio machine learning molecular dynamics simulations. *The Journal of Physical Chemistry Letters*, 11(17):7363–7370, 2020.

[60] L. Raff S. T. S. Bukkapatnam A. Pukrittayakamee, M. Hagan and R. Komanduri. Practical training framework for fitting a function and its derivatives. *IEEE Transactions on Neural Networks*, 22(6):936–947, 2011.

[61] Hsin-Yu Ko, Linfeng Zhang, Biswajit Santra, Han Wang, Weinan E, Robert A. DiStasio Jr, and Roberto Car. Isotope effects in liquid water via deep potential molecular dynamics. *Molecular Physics*, 117(22):3269–3281, 2019.

[62] Grace M. Sommers, Marcos F. Calegari Andrade, Linfeng Zhang, Han Wang, and Roberto Car. Raman spectrum and polarizability of liquid water from deep neural networks. *Phys. Chem. Chem. Phys.*, 22:10592–10602, 2020.

[63] Jörg Behler. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angewandte Chemie International Edition*, 56(42):12828–12840, 2017.

[64] Kyoungmin Min and Eunseog Cho. Neural network interatomic potential for predicting the formation of planar defect in nanocrystal. *The Journal of Physical Chemistry C*, 124(17):9424–9433, 2020.

[65] Tobias Morawietz, Andreas Singraber, Christoph Dellago, and Jörg Behler. How van der waals interactions determine the unique properties of water. *Proceedings of the National Academy of Sciences*, 113(30):8368–8373, 2016.

[66] Bingqing Cheng, Edgar A. Engel, Jörg Behler, Christoph Dellago, and Michele Ceriotti. Ab initio thermodynamics of liquid and solid water. *Proceedings of the National Academy of Sciences*, 116(4):1110–1115, 2019.

[67] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD – visual molecular dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[68] Martin Brehm and Barbara Kirchner. TRAVIS - a free analyzer and visualizer for monte carlo and molecular dynamics trajectories. *Journal of Chemical Information and Modeling*, 51(8):2007–2023, 2011.

[69] M. Brehm, M. Thomas, S. Gehrke, and B. Kirchner. TRAVIS - a free analyzer for trajectories from molecular simulation. *The Journal of Chemical Physics*, 152(16):164105, 2020.

[70] JR Errington and PG Debenedetti. Relationship between structural order and the anomalies of liquid water. *Nature*, 409(6818):318—321, January 2001.

[71] Elise Duboué-Dijon and Damien Laage. Characterization of the local structure in liquid water by various order parameters. *The Journal of Physical Chemistry B*, 119(26):8406–8418, 2015.

[72] Damien Laage, Guillaume Stirnemann, Fabio Sterpone, Rossend Rey, and James T. Hynes. Reorientation and allied dynamics in water and aqueous solutions. *Annual Review of Physical Chemistry*, 62(1):395–416, 2011.

[73] Miroslav Krepl, Marie Zgarbová, Petr Stadlbauer, Michal Otyepka, Pavel Banáš, Jaroslav Koča, Thomas E. Cheatham, Petr Jurečka, and Jiří Šponer. Reference simulations of noncanonical nucleic acids with different variants of the AMBER force field: Quadruplex DNA, quadruplex RNA, and Z-DNA. *Journal of Chemical Theory and Computation*, 8(7):2506–2520, 2012.

[74] Huziel E. Sauceda, Michael Gastegger, Stefan Chmiela, Klaus-Robert Müller, and Alexandre Tkatchenko. Molecular force fields with gradient-domain machine learning (gdml): Comparison and synergies with classical force fields. *The Journal of Chemical Physics*, 153(12):124109, 2020.