Facultade de Química

Universidade Vigo

Máster en Química Teórica y Modelización Computacional

Trabajo Fin de Máster - Curso 2020/21

Solute-solvent interactions in water: a QM/MM Energy Decomposition Analysis approach

Author:
Álvaro Pérez Barcia

Supervisors:
Marcos Mandado Alonso
Juan José Nogueira Pérez

July 2021

Contents

1 Introduction **1**

2 Methods **3**

2.1 Molecular Dynamics . . . 3

2.1.1 Basic Concepts in Molecular Dynamics . . . 3

2.1.2 Main Steps in a Molecular Dynamics Simulation . . . 5

2.2 Density Functional Theory . . . 7

2.2.1 Hohenberg and Kohn Theorems . . . 7

2.2.2 Kohn-Sham Equations . . . 8

2.2.3 Exchange-correlation Functional . . . 10

2.3 Energy Decomposition Analysis . . . 11

2.3.1 Rayleigh-Schrödinger Perturbation Theory . . . 11

2.3.2 Polarisation Theory . . . 14

2.3.3 Electron Density based EDA . . . 17

3 Computational Details **20**

4 Results and Discussion **23**

4.1 Convergence Study . . . 23

4.2 Nucleobases EDA . . . 27

5 Conclusions **32**

Bibliography **36**

Annex **37**

Chapter 1

Introduction

Intermolecular interactions are the driving force of many systems regardless of their small magnitude when compared with their intra-molecular counterparts. Their importance is crucial to protein binding events, solvation phenomena, catalysis applications or the design of new materials among many others. It is no wonder how these relatively small but still very relevant forces have shaped the chemical and physical reasoning since the early studies of van der Waals on the behaviour of real gases.

The search for a physically sound description of these interactions has favoured the application and development of radically different theoretical approaches to the problem, ranging from the force fields used in classical Molecular Mechanics (MM) to the most accurate approaches developed under the foundations of quantum mechanics (QM), where the contributions of Perturbation Theory (PT) have been particularly valuable to the field. Despite the seemingly opposed views on the matter from these theoretical perspectives, it is only through their conjunction by means of the hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) methods that larger, more realistic systems, can be described at reasonable accuracy.

Since their acceptance and consequent popularisation in the 1990s QM/MM methods have found extensive use in the simulation of biochemical systems[1], where the combination of quantum chemical accuracy with the extensive configurational sampling that MM is capable of has been key to their success. QM/MM techniques have thus enabled the extension of quantum chemical tools to the study of more complex systems, resulting in a qualitative leap forward in their application. In the framework of intermolecular interactions, the Energy Decomposition Analysis (EDA), an analytical tool partitioning the interaction energy into its electrostatic, polarisation and repulsive components, has seen application in the description of biochemical processes hand in hand with QM/MM methods[2].

The deep physical meaning that EDA gives to the interaction energy is very attractive as it can give further insight into poorly described mechanisms. Such is the case of the catalytic action of *myo-Inositol Oxygenase (MIOX)*, where *H. Hirao*[3] showed through an ONIOM(DFT/MM) EDA approach how the protein environment and dispersion forces influenced the O_2 binding to the active centre. Nonetheless, the application of the EDA scheme to bio-related systems can not only reveal its mechanistic insights, it has also the potential to contribute to rational drug design given the intimate relationship between interaction energy and drug functionalisation.

Regarding the environment description, the solvation process plays a crucial role in most chemical and biochemical systems. Given its importance, there is a long and rich trajectory in the computational treatment of solvation going from implicit models[4], where the solvent is not described explicitly but rather in a continuous manner usually by means of a distribution function, to explicit ones. EDA has a significant background when it comes to the description of solvation. Beginning in 1988, *Tomasi*

et al.[5] first attempted to apply the methodology to the condensed phase by introducing counterpoise corrections[6] to the description of dimeric interactions in solution. The use of implicit solvation models[4], particularly the Polarisable Continuum Model (PCM) model[7, 8], has been very popular when performing EDA. The work of *Gora et al.*[9] showed through an EDA-PCM approach how the electrostatic interaction is the main driver for the interaction energy in the water and hydrogen fluoride dimers. *Peifeng Su et al.*[10] extended the use of a new EDA-PCM scheme to the study of van der Waals, ionic, cation-π or metal-ligand binding through different test calculations with a variety of solvents. However, despite their success when describing solvation free energies and other thermodynamic properties [11], there are a number of inherent limitations to implicit solvation models. Most importantly, their inability to describe partially or completely inhomogeneous systems like ionic liquids or a solvated ligand in a partially exposed binding pocket[12].
Explicit solvation models treating each solvent molecule individually may overcome these limitations and, hence, they can be used in those situations that are challenging for their implicit counterparts. Ideally, the whole system would be treated fully *ab-initio* by means of Density Functional Theory (DFT), PT or other wavefunction methods. Nonetheless, a full QM treatment of the solute and the many solvent molecules surrounding it, as well as the different configurations required to appropriately sample configurational space, results in a prohibitive computational cost.
A natural alternative comes with the use of hybrid QM/MM methods [13, 14]. By treating the solute quantum-mechanically and the solvent by means of a MM force field, this technique comes as a compromise between computational efficiency and accuracy. An obvious complication that may arise from this approach is the treatment of the solute-solvent interaction, critical when an EDA is kept in mind. There are a number of components comprising these interactions, namely, permanent electrostatics, forward (MM $\rightarrow$ QM) and backward (QM $\rightarrow$ MM) polarisation, exchange-repulsion and dispersion. In this sense, several efforts have been made to develop new algorithms improving the description of permanent electrostatics and forward polarisation[15] as well as that of backward polarisation through the development of polarisable force fields based on fluctuating charges[16] and Drude oscillators [17] among others. Recent studies have been devoted to the development of accurate QM/MM partition schemes, where EDA can asses the performance of such partitions and, in last instance, contribute to MM force field parametrisation. The work of *Head-Gordon et al.*[12, 18] has been focused on this issue, where the evaluation of the AMOEBA[19] water force field and the QM/AMOEBA partition for solute-solvent interactions has been carried out with an EDA scheme.

In this framework, the work here developed is aimed at decomposing the solute-solvent interaction energy obtained with a QM/MM partition by means of the EDA. The novelty of the project is found in the EDA scheme used as it incorporates the external potential induced by the MM region, allowing its evaluation for each energy component. In this sense, a convergence study of each interaction energy component with increasing size of the QM region, which includes the solute and several solvent molecules, is performed to asses its impact on intermolecular interactions. Besides, the effect of the external MM potential on the convergence trend is also examined. The results derived from this study might be of interest for the set-up of a QM/MM EDA calculation as the magnitudes computed for the investigated test models reveal a clear dependency on both QM size and external MM potential.
The test systems consist of the five nucleobases; adenine, cytosine, guanine, thymine and uracil in water solution as well as two tetrahedric solutes, tetrafluoromethane (CF_4) and the ammonium cation (NH_4^+). The latter systems representing an insoluble and soluble molecule in liquid water, respectively, whilst the different functionalities of the nucleobases allow for a discussion of the impact that these structural differences have in the interaction energy, aside, of course, of their biological importance.

Chapter 2

Methods

The general procedure followed in this work could be summarised in a series of steps beginning by the configurational sampling of the studied systems, which is carried out by means of a classical Molecular Dynamics (MD) simulation, whose foundations can be found in Newtonian motion and statistical thermodynamics. In a second step, the QM/MM partition is performed in a variety of selected snapshots from the MD trajectory, where the QM region is computed with DFT. Finally, the EDA, which is strongly rooted in PT, is carried out on these geometries making use of the electron densities computed in the previous step. The following sections briefly introduce some of the theoretical bases that sustain the methods employed in this study.

2.1 Molecular Dynamics

In a MD simulation, the Newtonian equations of motion are numerically solved to generate a dynamic trajectory of a particular system, obtaining structural, dynamic and thermodynamic information. In classical MD, electrons are not considered explicitly, instead, only nuclear motion under an average potential is studied resulting in a drastic reduction of the computational cost in relation to that of QM methods, at the expense of accuracy. For this reason, MD is a suitable method for the study of large systems of up to millions of atoms and is used in this work to explore the different conformations that the solvated system adopts during a certain period of time.

For this reason, MD is an adequate tool to obtain the trajectory of the solvated systems, necessary for the later treatment of the different geometries with DFT and EDA.

2.1.1 Basic Concepts in Molecular Dynamics

For a system of N particles under a conservative potential, i.e., only dependent on the position of such particles, the trajectory is obtained by solving the differential equations embodied in Newton's second law

$$F_{x_i} = m_i \frac{d^2 x_i}{dt^2} = -\frac{dV_i}{dx_i} \tag{2.1.1}$$

where F is the force, m the particle's mass, x is the x-axis component of particle's position and V is the potential acting on particle i.

The interdependence of the nuclear motion makes impossible an analytical solution of eq. 2.1.1. Instead, an integrating algorithm using finite difference methods is used to numerically solve the equations. All of these algorithms assume that the positions and dynamic properties can be approximated as a Taylor series. In this context, the *Verlet algorithm*[20] is one of the most widely used method for integrating the equations of motion in MD. The Verlet algorithm uses positions and accelerations at time t to calculate the positions at a new times $t + \delta t$ and $t - \delta t$, $\boldsymbol{r}(t + \delta t)$ and $\boldsymbol{r}(t - \delta t)$. Expanding $\boldsymbol{r}(\boldsymbol{t})$ in a Taylor series up to second order:

$$\boldsymbol{r}(t + \delta) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\boldsymbol{a}(t)\delta t^2 \tag{2.1.2}$$

$$\boldsymbol{r}(t - \delta) = \boldsymbol{r}(t) - \boldsymbol{v}(t)\delta t + \frac{1}{2}\boldsymbol{a}(t)\delta t^2 \tag{2.1.3}$$

Summing both equations

$$\boldsymbol{r}(t + \delta) = 2\boldsymbol{r}(t) - \boldsymbol{r}(t - \delta t) + \boldsymbol{a}(t)\delta t^2 \tag{2.1.4}$$

Velocities are not present in the Verlet algorithm, they can be estimated as

$$\boldsymbol{v}(t) = \frac{\boldsymbol{r}(t + \delta t) - \boldsymbol{r}(t - \delta t)}{2\delta t} \tag{2.1.5}$$

This method, however, requires to know the position in a next step. The *Velocity Verlet algorithm*[20] calculates both the velocity and position at the same time step, not compromising precision:

$$\boldsymbol{r}(t + \delta t) = \boldsymbol{r}(t) + \boldsymbol{v}(t)\delta t + \frac{1}{2}\boldsymbol{a}(t)\delta t^2 \tag{2.1.6}$$

$$\boldsymbol{v}(t + \delta t) = \boldsymbol{v}(t) + \frac{1}{2}[\boldsymbol{a}(t) + \boldsymbol{a}(t + \delta t)]\delta t \tag{2.1.7}$$

The algorithm proceeds in three steps, the first one being the computation of the new positions according to eq. 2.1.6, then the accelerations at the new time $t + \delta t$ are determined from the forces computed at the new positions and, finally, the new velocities are obtained according to eq. 2.1.7.

Hence, a classical MD simulation requires the initial position of the particles and the initial velocities, the latter usually provided by a Maxwell-Boltzmann distribution. With respect to the initial positions, they can be obtained either from experiment or from a previous simulation, the latter is the case in this study.

The potential driving the behaviour of the modelled system is typically determined by a MM force field which, in the case of organic and biological systems, has usually the following general expression

$$\begin{aligned} V_{ij} = &\sum_{a=1}^{N_{bonds}} k_a(r - r_{eq})^2 + \sum_{b=1}^{N_{angles}} k_b(\theta - \theta_{eq})^2 + \sum_{c=1}^{N_{dihed.}} k_c(1 + \cos(nw - \gamma)) \\ &+ \sum_{i>j} \left[4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\epsilon\pi r_{ij}} \right] \end{aligned} \tag{2.1.8}$$

The first two terms on the right side of the eq. 2.1.8 describe bond stretching and bending as a harmonic potential with r_{eq} and θ_{eq} being the bond distance and angle at the equilibrium positions and k_a and k_b the force constants. Term three represents the dihedral torsion by means of a Fourier series where w is the dihedral angle, k_c the force constant and n is the number of minima with γ being its phase angle. The final term in eq. 2.1.8 describes non-bonding interactions by means of a Lennard-Jones potential and the Coulomb law.

From eq. 2.1.8 one can see how crucial it is to have a proper parametrization. For instance, the General Amber Force Field (GAFF)[21] bases its parametrization on more than 3000 MP2/6-31G* optimisations and 1260 MP4/6-311G(d,p) single point calculations in order to provide general applicability to the MD simulation of organic molecules.

Another important concept in MD simulations is the use of *Periodic Boundary Conditions* (PBC), which allow for a more accurate estimation of bulk properties from the simulation of finite systems. PBC alleviate many of the issues associated with finite size as the replication of the system in every direction implies that each particle interacts now with a periodic image of other particles in the system. To avoid artefacts, like the computation of a single interaction several times, a cut-off for non-bonded interactions is usually introduced, imposing a lower limit to the size of the simulation box.

2.1.2 Main Steps in a Molecular Dynamics Simulation

Despite the particular challenges that a specific MD simulation may present, there are at least three common steps to all of them[22].

Relaxation

The *Relaxation* step is the first of them all and it consists on the application of standard minimization algorithms like steepest descendent. The goal of this initial step is to avoid unreasonable displacements at the beginning of the simulation due to excessive forces at the initial configuration.

Equilibration

The *Equilibration* step follows. At this point, the simulation is carried out in the frame of a thermodynamic ensemble so that a significant amount of simulation time is devoted to reaching thermodynamic equilibrium. For instance, in this work the canonical (NVT) ensemble was used to thermalize the system prior to the actual MD simulation which was done in the isothermal-isobaric (NPT) ensemble. The use of these two ensembles is not casual as the first steps carried out in the NPT ensemble are set to equilibrate the density of the system after the NVT thermalisation step. To enforce the macroscopic conditions on our system it becomes necessary to use a thermostat and later a barostat. In the case of thermostat algorithms, they modify the Newtonian equations of motion acting on the instantaneous temperature which is computed as

$$K = \frac{1}{2}\sum_{i=1}^{N} m_i v_i^2 = \frac{3}{2} N k_b T \tag{2.1.9}$$

where K is the overall kinetic energy, N the total number of the particles in the system, k_b is the Boltzmann factor, v_i is the norm of the velocity vector of particle i and T the average temperature.
The *Langevin thermostat*[23] is used in this work and it supplements the microcanonical equations of motion by including Brownian dynamics, i.e., the effect of viscosity and random collision effects of an implicit solvent. It follows the general equation:

$$m a_i = F_i - \gamma_i p_i + f_i \tag{2.1.10}$$

where F_i is the force acting on particle i due to the interaction potential, p_i is the linear momentum, γ_i is the friction coefficient and f_i is a random force representing the damping between particles due to friction. These random numbers are chosen from a Gaussian distribution of variance

$$\sigma_i^2 = \frac{2m_i\gamma_i k_b T}{\delta t} \tag{2.1.11}$$

Thermodynamic properties are typically measured at constant pressure and temperature in the laboratory, thus, the coupling of a thermostat and a barostat is necessary to achieve the NPT ensemble that reproduces experimental conditions. The *Berendsen barostat*[24] is used in this work. Also known as the weak-coupling barostat it operates through volume rescaling by coupling the system to a weakly interacting pressure bath. This bath scales the volume periodically by a scaling factor, which produces more realistic fluctuations in the pressure as it slowly approaches the targeted value. According to statistical thermodynamics the pressure is computed as

$$P = \frac{2}{3V}\left(K + \frac{1}{2}\sum_{i>j}\boldsymbol{r}_{ij}\boldsymbol{F}_{ij}\right) \tag{2.1.12}$$

where K is the total kinetic energy and $\boldsymbol{F}_{ij}$ the force acting on particle i due to particle j. The Berendsen thermostat rescales the pressure by adding an extra term to the equations of motion

$$\left(\frac{dP}{dt}\right)_{bath} = \frac{P_0 - P}{\tau_P} \tag{2.1.13}$$

where P_0 is the reference pressure (the bath pressure) and τ_P is a time constant.
According to eq. 2.1.12 a change in pressure can be achieved by scaling interparticle distance (Virial term), i.e., by scaling the system's volume. For an isotropic system in a cubic box the box length and thus particle's coordinates are scaled according to

$$\mu = \left(1 - \frac{\delta t}{\tau_P}(P_0 - P)\right)^{\frac{1}{3}} \tag{2.1.14}$$

for a non-cubic cell, eq. 2.1.14 finds its tensor analogue.

Production

In a final step, data is finally collected for analysis, i.e., the computation of expectation values. This stage is reached once the properties of the system are converged to their averaged values, lacking any systematic dependence on the simulation time or the system's initial conditions. A criterion often used in biomolecular simulations, and employed in this work, to determine the starting point of the production phase is the analysis of the Root Mean Square Deviation (RMSD) of the simulated molecules as a function of time.

2.2 Density Functional Theory

Density Functional Theory (DFT) is one of the most popular methods in computational chemistry today as it provides a very good ratio between accuracy and computational cost, overcoming the poor treatment of electron correlation characteristic from the Hartree-Fock (HF) method without the high computational cost of most post-HF procedures. In this work, DFT is a central element, as it provides the electron densities that will be used to carry out the EDA.
DFT revolves around the electron density $\rho(\boldsymbol{r})$ as a knowledge of the density is all that is necessary for a complete determination of all molecular properties. The beginnings of DFT date back to 1965 with the formulation of the Hohenberg and Kohn theorems[25] which are stated in the following section.

2.2.1 Hohenberg and Kohn Theorems

Theorem 1. *The electron density $\rho(\boldsymbol{r})$ determines the external potential.*

Proof. [26] Reductio ab absurdum. Let there be two external potentials $v_1(\boldsymbol{r})$ and $v_2(\boldsymbol{r})$ arising from the same $\rho(\boldsymbol{r})$. Thus, there will be two Hamiltonians H_1 and H_2 with the same (ground state) density, but different wave functions ψ_1 and ψ_2. Applying the variational principle:

$$\begin{aligned} E_1^o < \langle\psi_2|H_1|\psi_2\rangle &= \langle\psi_2|H_2|\psi_2\rangle + \langle\psi_2|H_1 - H_2|\psi_2\rangle \\ &= E_2^o + \int \rho(\boldsymbol{r})[v_1(\boldsymbol{r}) - v_2(\boldsymbol{r})]d\boldsymbol{r} \end{aligned} \tag{2.2.1}$$

where E_1^o and E_2^o are the ground state energies of H_1 and H_2 respectively.
Analogously:

$$\begin{aligned} E_2^o < \langle\psi_1|H_2|\psi_1\rangle &= \langle\psi_1|H_1|\psi_1\rangle + \langle\psi_1|H_2 - H_1|\psi_1\rangle \\ &= E_1^o + \int \rho(\boldsymbol{r})[v_2(\boldsymbol{r}) - v_1(\boldsymbol{r})]d\boldsymbol{r} \end{aligned} \tag{2.2.2}$$

Summing the two inequalities we arrive at a contradiction:

$$E_1^o + E_2^o < E_2^o + E_1^o \tag{2.2.3}$$

Hence the external potential is uniquely determined by $\rho(\boldsymbol{r})$. □

We can express the energy as a functional of the density:

$$\begin{aligned} E[\rho] &= V_{ne}[\rho] + T[\rho] + V_{ee}[\rho] \\ &= \int \rho(\boldsymbol{r})v(\boldsymbol{r})d\boldsymbol{r} + T[\rho] + V_{ee}[\rho] \end{aligned} \tag{2.2.4}$$

where $T[\rho]$ is the kinetic energy, V_{ne} the nucleus-electron interaction energy and $V_{ee}[\rho]$ the electron-electron interaction energy.

Theorem 2. *The ground state density can be calculated, in principle exactly, using the variational method involving only the density.*

Proof. [26] Any approximate density $\tilde{\rho}$ determines its own wavefunction $\tilde{\psi}$. Applying the variational method to this wavefunction:

$$\langle\tilde{\psi}|H|\tilde{\psi}\rangle = \int \tilde{\rho}(\boldsymbol{r})v(\boldsymbol{r})d\boldsymbol{r} + T[\tilde{\rho}] + V_{ee}[\tilde{\rho}] = E[\tilde{\rho}] \geq E[\rho] \tag{2.2.5}$$

Applying the minimum condition to the energy constrained by the N representability of the density, i.e, $\int \rho(\boldsymbol{r})d\boldsymbol{r} = N$, where N is the number of electrons:

$$\delta E[\rho] - \mu\delta\left[\int \rho(\boldsymbol{r})d\boldsymbol{r} - N\right] = 0 \tag{2.2.6}$$

where μ is the Lagrange multiplier.

$$\int \frac{\delta E[\rho]}{\delta\rho(\boldsymbol{r})}\delta\rho(\boldsymbol{r})d\boldsymbol{r} - \mu\int \delta\rho(\boldsymbol{r})d\boldsymbol{r} = 0 \tag{2.2.7}$$

$$\mu = \frac{\delta E[\rho]}{\delta\rho(\boldsymbol{r})} = v(\boldsymbol{r}) + \frac{\delta T[\rho]}{\delta\rho(\boldsymbol{r})} + \frac{\delta V_{ee}[\rho]}{\delta\rho(\boldsymbol{r})} \tag{2.2.8}$$

Hence, the value of the Lagrange multiplier is defined at the minimum. □

From a practical point of view Theorem 2 states that having a good functional representation, we can then minimise it to get best electron density and structure.

2.2.2 Kohn-Sham Equations

The above description provides a method to minimise the energy by changing the corresponding density. Unfortunately the kinetic energy functional is unknown. For that reason, Kohn and Sham proposed the combination of the wavefunctions with the density approach[27]. They considered the single-determinantal wavefunction of a system of N non-interacting electrons constructed from one-electron functions, also known as orbitals, ϕ_i. This way the reference system of non-interacting electrons would have the same electron density as the real system, and thus the real energy could be derived from it. For such a system the single-determinantal kinetic energy and the electron density are exactly given by

$$T_s[\rho] = \sum_i^N \langle\phi_i| - \frac{1}{2}\nabla^2|\phi_i\rangle \tag{2.2.9}$$

$$\rho(\boldsymbol{r}) = \sum_i^N |\phi_i(\boldsymbol{r})|^2 \tag{2.2.10}$$

The orbitals obey an equation of the form

$$\left[-\frac{1}{2}\nabla^2 + v_s(\boldsymbol{r})\right]\phi_i = \epsilon_i\phi_i \tag{2.2.11}$$

and the energy of the system is given by

$$E[\rho] = T_s[\rho] + \int v_s(\boldsymbol{r})\rho(\boldsymbol{r})d\boldsymbol{r} \tag{2.2.12}$$

Returning to the problem with interacting electrons we can rewrite the energy by grouping all the unknown terms

$$\begin{aligned} E[\rho] &= \int v(\boldsymbol{r})\rho(\boldsymbol{r})d\boldsymbol{r} + T[\rho] + V_{ee}[\rho] \\ &= \int v(\boldsymbol{r})\rho(\boldsymbol{r})d\boldsymbol{r} + T_s[\rho] + J[\rho] + (T[\rho] - T_s[\rho]) + (V_{ee}[\rho] - J[\rho]) \\ &= \int v(\boldsymbol{r})\rho(\boldsymbol{r})d\boldsymbol{r} + T_s[\rho] + J[\rho] + E_{xc}[\rho] \end{aligned} \tag{2.2.13}$$

where $E_{xc}[\rho]$ is the exchange-correlation energy functional and its derivative, the exchange-correlation potential v_{xc}.

$$E_{xc} = (T[\rho] - T_s[\rho]) + (V_{ee}[\rho] - J[\rho]) \tag{2.2.14}$$

$$v_{xc}(\boldsymbol{r}) = \frac{\delta E_{xc}}{\delta \rho(\boldsymbol{r})} \tag{2.2.15}$$

and $J[\rho]$ is the Coulomb interaction term of the electron-electron interaction energy term ($V_{ee}[\rho]$):

$$J[\rho] = \frac{1}{2}\int\int \frac{1}{r_{12}}\rho(\boldsymbol{r_1})\rho(\boldsymbol{r_2})d\boldsymbol{r_1}d\boldsymbol{r_2} \tag{2.2.16}$$

Functional $E_{xc}[\rho]$ includes all the energy contributions which are not accounted by the previous terms, i.e.:

1. Electron exchange.
2. Electron correlation.
3. A portion of the kinetic energy needed to correct $T_s[\rho]$ to obtain the true kinetic energy $T[\rho]$.
4. Correction for self-interaction induced by the classical coulomb potential.

The above discussion allows to reformulate the problem in the following way

$$\left[-\frac{1}{2}\nabla^2 + v(\boldsymbol{r}) + \int \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|}d\boldsymbol{r'} + v_{xc}(\boldsymbol{r})\right]\phi_i = \epsilon_i\phi_i(\boldsymbol{r}) \tag{2.2.17}$$

These are the Kohn-Sham equations for the Kohn-Sham orbitals ϕ_i. They are solved in an iterative manner as the SCF equations, expanding the orbitals in terms of the basis set

$$\phi_i = \sum_\alpha c_{\alpha i}\eta_\alpha \tag{2.2.18}$$

$$\sum_\beta c_{\beta i}\left\langle \eta_\beta\right| - \frac{1}{2}\nabla^2 + v(\boldsymbol{r}) + \int \frac{\rho(\boldsymbol{r'})}{|\boldsymbol{r} - \boldsymbol{r'}|}d\boldsymbol{r'} + v_{xc}(\boldsymbol{r}) - \epsilon_i\left|\eta_\beta\right\rangle = 0 \tag{2.2.19}$$

Compared to the SCF procedure, the exchange term has now been replaced by the v_{xc} term. This approach solves the kinetic energy problem but the development of an exchange-correlation functional becomes necessary.

2.2.3 Exchange-correlation Functional

In this work the M062X exchange-correlation energy functional developed by *Prof. Donald Truhlar* at the Minnesota University [28] was used given its good performance when studying non-covalent interactions[29] as it proves its extensive testing in various databases.

The M06 series of functionals are classified as hybrid meta-generalised gradient approximations (hybrid meta-GGAs) as they include a fraction of exact HF exchange while the rest is derived from the meta-GGA approximation, where the second derivative of the electron density (Laplacian) is considered. In this case, this is done through a dependence on the kinetic energy density (τ). The M062X functional takes the form

$$E_{XC} = \frac{X}{100}E_X^{HF} + \left(1 - \frac{X}{100}\right) E_X^{DFT} + E_C^{DFT} \tag{2.2.20}$$

where E_X^{HF} is the non-local HF exact exchange, E_X^{DFT} is the local DFT exchange energy and E_C^{DFT} is the local DFT correlation energy. Parameter X is the percentage of exact HF exchange in the functional and it is optimised along with other parameters contained in the meta-GGA exchange and correlation functionals.

The local parts of the M06 functionals depend on three variables, namely the spin density (ρ_σ), reduced spin density gradient (x_σ) and spin kinetic energy density (τ_σ) where

$$x_\sigma = \frac{|\nabla \rho_\sigma|}{\rho_\sigma^{4/3}}, \quad \sigma = \alpha, \beta \tag{2.2.21}$$

$$\tau_\sigma = \frac{1}{2}\sum_i^{occ} |\nabla \psi_{i\sigma}|^2 \tag{2.2.22}$$

where $\psi_{i\sigma}$ are the spin orbitals.

A working variable (z_σ) and two working functionals $(\gamma(x_\sigma, z_\sigma))$, $(h(x_\sigma, z_\sigma))$ are also defined with (z_σ) dependent on both ρ_σ and τ_σ.

The M06 exchange functional is defined by

$$E_X^{M06} = \sum_\sigma \int \left[F_{X\sigma}^{PBE}(\rho_\sigma, \nabla\rho_\sigma) f(w_\sigma) + \varepsilon_{X\sigma}^{LSDA} h(x_\sigma, z_\sigma)\right] d\boldsymbol{r} \tag{2.2.23}$$

where $F_{X\sigma}^{PBE}$ is the exchange energy density according to the PBE exchange model and $\varepsilon_{X\sigma}^{LSDA}$ the local spin density approximation (LSDA) for exchange. $f(w_\sigma)$ is the spin kinetic-energy enhancement factor, a function of the spin kinetic energy density (τ_σ) and the spin density (ρ_σ). In the particular case of the M062X functional $h(x_\sigma, z_\sigma) = 0$.

The M062X correlation functional treats parallel and opposite spin correlation differently. The opposite spin-correlation is expressed as

$$E_C^{\alpha\beta} = \int e_{\alpha\beta}^{UEG}[g_{\alpha\beta}(x_\alpha, x_\beta) + h_{\alpha\beta}(x_{\alpha\beta}, z_{\alpha\beta})]d\boldsymbol{r} \tag{2.2.24}$$

whereas for parallel spin

$$E_C^{\sigma\sigma} = \int e_{\sigma\sigma}^{UEG}[g_{\sigma\sigma}(x_\sigma, x_\sigma) + h_{\sigma\sigma}(x_\sigma, z_\sigma)]D_\sigma d\boldsymbol{r} \tag{2.2.25}$$

$g_{\alpha\beta}(x_\alpha, x_\beta)$ is a parameterized function of the spin density gradient and $x_{\alpha\beta} \equiv x_\alpha^2 + x_\beta^2$ and $z_{\alpha\beta} \equiv z_\alpha + z_\beta$. D_σ is the self-interaction correction factor and $e_{\alpha\beta}^{UEG}$ and $e_{\sigma\sigma}^{UEG}$ are the uniform electron gas (UEG) correlation densities for both opposite and parallel spin cases. The total M062X correlation energy is given by

$$E_C^{M06} = E_C^{\alpha\beta} + E_C^{\alpha\alpha} + E_C^{\beta\beta} \tag{2.2.26}$$

2.3 Energy Decomposition Analysis

Intermolecular interactions play a crucial role in many chemical and biochemical processes despite its weakness when compared to those bonding molecules internally. The concept of intermolecular interaction appears already in the *Born-Oppenheimer* approximation as it is assumed that each interacting fragment presents the same geometry both when isolated and associated. The interaction energy between fragments A and B is defined as the difference between the energy of the dimer (E_{AB}) and the energies of the monomers

$$E_{int} = E_{AB} - (E_A + E_B) \tag{2.3.1}$$

Computing the interaction energy as suggested in eq. 2.3.1 is known as the *supermolecular approach* and given the incompleteness of the basis set, it is subject to the Basis Set Superposition Error (BSSE)[30]. Such error consists in an artificial improvement of the description of the fragments when they are forming the complex due to the basis set of its interacting partner. Therefore, the energy of fragment A will be lower in the complex AB than in its own basis. Despite being relatively small in absolute value, given the small magnitude of intermolecular interactions, the BSSE ends up often being larger than the interaction energy. The Counterpoise correction[6] is the standard procedure used to correct for the BSSE by computing all energies in the full basis set.

Moreover, the *supermolecular approach* does not provide any physical insight into the nature of the interatomical interactions. For a deeper understanding one must turn to perturbation or variational methods.

2.3.1 Rayleigh-Schrödinger Perturbation Theory

Rayleigh-Schrödinger perturbation theory (RSPT)[31] is an early formulation of time-independent perturbation theory proposed in the first third of the twentieth century. The general idea behind RSPT is that assuming the Hamiltonian of a system to be of the form

$$H = H^0 + \lambda H' \tag{2.3.2}$$

where λ is a small real number, the eigenstates of the Hamiltonial H should not be very different from those of H^0. Knowing the eigenstates of H^0 we can approximate those of H'.

There are three basic assumptions in RSPT (and an additional one in the case of non-degenerate RSPT):

1. All eigenstates and eigenenergies of H^0 are known.

$$H^0 \left|\psi_n^0\right\rangle = E_n^0 \left|\psi^0\right\rangle \tag{2.3.3}$$

2. Perturbation H' is known, meaning that H' can be written in terms of the complete basis of $|\psi_n^0\rangle$, i.e., the value of $\langle\psi_n^0|H'|\psi_m^0\rangle$ it is known for any m and n.
3. Only quantum states with discrete eigenenergies are considered.
4. *Non-degenerate perturbation theory:* a specific state $|\psi_n^0\rangle$ is considered. Here it is assumed that $|E_n^0 - E_m^0|$ is much larger than $\lambda H'$ for any other eigenstate $|\psi_m^0\rangle$.

Wavefunctions in RSPT

Since the eigenstates of H^0 form a complete basis, any quantum state can be written as a linear combination of $|\psi_m^0\rangle$.

$$|\psi_n\rangle = \sum_m a_m |\psi_m^0\rangle \tag{2.3.4}$$

The same is true for an eigenstate $|\tilde{\psi_n}\rangle$ of H

$$|\tilde{\psi_n}\rangle = \sum_m a_m |\psi_m^0\rangle \tag{2.3.5}$$

As λ is small, $|\tilde{\psi_n}\rangle$ is close to $|\psi_n^0\rangle$. Separating terms for $|\tilde{\psi_n}\rangle$

$$|\tilde{\psi_n}\rangle = a_n |\psi_n^0\rangle + \sum_{m\neq n} a_m |\psi_m^0\rangle \tag{2.3.6}$$

where $a_n \approx 1$ and $a_m \approx 0$ for $m \neq n$. Unnormalising eigenstates of H

$$|\psi_n\rangle = \frac{1}{a_n}|\tilde{\psi_n}\rangle = |\psi_n^0\rangle + \sum_{m\neq n}\frac{a_m}{a_n}|\psi_m^0\rangle = |\psi_n^0\rangle + \sum_{m\neq n} c_m |\psi_m^0\rangle \tag{2.3.7}$$

where it is known that $c_m \approx 0$ for a small λ. Since c_m is a function of λ ($c_m(\lambda)$) and λ is small, the Taylor series can be used

$$c_m = 0 + c_m^{(1)}\lambda + c_m^{(2)}\lambda^2 + c_m^{(3)}\lambda^3 + ... \tag{2.3.8}$$

Defining

$$|\psi_n^k\rangle = \sum_{m\neq n} c_m^{(k)} |\psi_m^0\rangle \tag{2.3.9}$$

we finally get

$$|\psi_n\rangle = |\psi_n^0\rangle + \lambda|\psi_n^1\rangle + \lambda^2|\psi_n^2\rangle + ... \tag{2.3.10}$$

Since eigenenergies are also functions of λ they can be expanded by a Taylor series, arriving at an equivalent expression

$$E_n = E_n^0 + \lambda E_n^1 + \lambda^2 E_n^2 + ... \tag{2.3.11}$$

The Schrödinger equation can be now re-written by introducing the expansions in eq.2.3.11 and 2.3.9

$$\begin{aligned}
&H|\psi_n\rangle = E|\psi_n\rangle \\
&(H^0+\lambda H')(|\psi_n^0\rangle + \lambda|\psi_n^1\rangle + \lambda^2|\psi_n^2\rangle + ...) = (E_n^0 + \lambda E_n^1 + \lambda^2 E_n^2 + ...)(|\psi_n^0\rangle + \lambda|\psi_n^1\rangle + \lambda^2|\psi_n^2\rangle + ...) \\
&H^0|\psi_n^0\rangle + \lambda(H^0|\psi_n^1\rangle + H'|\psi_n^0\rangle) + \lambda^2(H^0|\psi_n^2\rangle + H'|\psi_n^1\rangle) + ... = E_n^0|\psi_n^0\rangle + \lambda(E_n^0|\psi_n^1\rangle + E_n^1|\psi_n^0\rangle) + \\
&+ \lambda^2(E_n^0|\psi_n^2\rangle + E_n^1|\psi_n^1\rangle + E_n^2|\psi_n^0\rangle) + ...
\end{aligned} \tag{2.3.12}$$

Thus, in perturbation theory there are two sets of quantities to compute, the energy corrections at each order $E_n^0, E_n^1, E_n^2, ...$ and the wavefunction corrections at each order $\psi_n^0, \psi_n^1, \psi_n^2, ...$

First Order Corrections

$$H^0 |\psi_n^1\rangle + H' |\psi_n^0\rangle = E_n^0 |\psi_n^1\rangle + E_n^1 |\psi_n^0\rangle \tag{2.3.13}$$

The first order energy correction, E_n^1, is obtained by multiplying eq. 2.3.13 by $\langle\psi_n^0|$:

$$\begin{aligned}\langle\psi_n^0| H^0 |\psi_n^1\rangle + \langle\psi_n^0| H' |\psi_n^0\rangle &= \langle\psi_n^0| E_n^0 |\psi_n^1\rangle + \langle\psi_n^0| E_n^1 |\psi_n^0\rangle \\ \langle\psi_n^0| E_n^0 |\psi_n^1\rangle + \langle\psi_n^0| H' |\psi_n^0\rangle &= \langle\psi_n^0| E_n^0 |\psi_n^1\rangle + \langle\psi_n^0| E_n^1 |\psi_n^0\rangle\end{aligned} \tag{2.3.14}$$

Hence,

$$E_n^1 = \langle\psi_n^0| H' |\psi_n^0\rangle \tag{2.3.15}$$

Since ψ_n^0 is normalised, the so called *intermediate normalisation* has been enforced, the first order correction in energy is the expectation value of H'.

To compute the first order correction $|\psi_n^1\rangle$ to the wavefunction eq. 2.3.13 is multiplied by $\langle\psi_m^0|$ where $m \neq n$

$$\langle\psi_m^0| H^0 |\psi_n^1\rangle + \langle\psi_m^0| H' |\psi_n^0\rangle = \langle\psi_m^0| E_n^0 |\psi_n^1\rangle + \langle\psi_m^0| E_n^1 |\psi_n^0\rangle \tag{2.3.16}$$

Knowing that quantum states are orthogonal

$$E_m^0 \langle\psi_m^0|\psi_n^1\rangle + \langle\psi_m^0| H' |\psi_n^0\rangle = E_n^0 \langle\psi_m^0|\psi_n^1\rangle \tag{2.3.17}$$

So

$$\langle\psi_m^0|\psi_n^1\rangle = \frac{\langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} \tag{2.3.18}$$

where

$$|\psi_n^1\rangle = \sum_{m\neq n} c_m^{(1)} |\psi_m^0\rangle \implies c_m^{(1)} = \langle\psi_m^0|\psi_n^1\rangle = \frac{\langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} \tag{2.3.19}$$

Therefore

$$|\psi_n^1\rangle = \sum_{m\neq n} |\psi_m^0\rangle \frac{\langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} \tag{2.3.20}$$

Second Order Energy Correction

$$H^0 |\psi_n^2\rangle + H' |\psi_n^1\rangle = E_n^0 |\psi_n^2\rangle + E_n^1 |\psi_n^1\rangle + E_n^2 |\psi_n^0\rangle \tag{2.3.21}$$

Computing the energy correction, E_n^2 by multiplying eq. 2.3.21 by $\langle\psi_n^0|$

$$\begin{aligned}\langle\psi_n^0| H^0 |\psi_n^2\rangle + \langle\psi_n^0| H' |\psi_n^1\rangle &= \langle\psi_n^0| E_n^0 |\psi_n^2\rangle + \langle\psi_n^0| E_n^1 |\psi_n^1\rangle + \langle\psi_n^0| E_n^2 |\psi_n^0\rangle \\ \langle\psi_n^0| H^0 |\psi_n^2\rangle + \langle\psi_n^0| H' |\psi_n^1\rangle &= \langle\psi_n^0| E_n^0 |\psi_n^2\rangle + E_n^1 \langle\psi_n^0|\psi_n^1\rangle + E_n^2\end{aligned} \tag{2.3.22}$$

The second term in the right hand side of eq.2.3.22 equals to 0, since according to the definition of $|\psi_n^{(k)}\rangle$ given in eq.2.3.9, $|\psi_n^{(k)}\rangle$ is orthogonal to $|\psi_n^0\rangle$. Thus,

$$E_n^2 = \langle\psi_n^0| H' |\psi_n^1\rangle = \sum_{m\neq n} \langle\psi_n^0| H' |\psi_m^0\rangle \frac{\langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} = \sum_{m\neq n} \frac{\langle\psi_n^0| H' |\psi_m^0\rangle \langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} \tag{2.3.23}$$

To sum up, it has been assumed that for a Hamiltonian $H = H^0 + \lambda H'$ where all eigenstates $|\psi_n^0\rangle$ of H^0 are known as well as the expectation values $\langle\psi_{m1}^0|H'|\psi_{m2}^0\rangle$ for any eigenstate $|\psi_{m1}^0\rangle$ and $|\psi_{m2}^0\rangle$ of H^0 we can write both the eigenvalues and eigenstates of H as a power series expansion of λ

$$E_n = E_n^0 + \lambda E_n^1 + \lambda^2 E_n^2 + ... = E_n^0 + \lambda \langle\psi_n^0| H' |\psi_n^0\rangle + \lambda^2 \sum_{m\neq n} \frac{\langle\psi_n^0| H' |\psi_m^0\rangle \langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} + ... \quad (2.3.24)$$

$$|\psi_n\rangle = |\psi_n^0\rangle + \lambda |\psi_n^1\rangle + \lambda^2 |\psi_n^2\rangle + ... = |\psi_n^0\rangle + \lambda \sum_{m\neq n} |\psi_m^0\rangle \frac{\langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} + ... \quad (2.3.25)$$

It is worth noticing that the second order correction always results in a energy decrease with respect to the ground state energy when compared with the unperturbed term since

$$E_n^2 = \sum_{m\neq n} \frac{\langle\psi_n^0| H' |\psi_m^0\rangle \langle\psi_m^0| H' |\psi_n^0\rangle}{E_n^0 - E_m^0} = \sum_{m\neq n} \frac{\langle\psi_n^0| H' |\psi_m^0\rangle^2}{E_n^0 - E_m^0} \leq 0 \quad (2.3.26)$$

The numerator is non-negative as it is the product of $\langle\psi_n^0| H' |\psi_m^0\rangle$ by its complex conjugate and the denominator is strictly negative since n is the ground state of the unperturbed Hamiltonian, hence its energy E_n^0 is always lower than the eigenenergy of any other state.

2.3.2 Polarisation Theory

Polarisation theory[32] is the most conceptually simple perturbational approach to intermolecular interactions. It is a standard application of the Rayleigh-Schrödinger perturbation method to the eigenvalue problem for the electronic Hamiltonian H for the complex AB. It works under the assumption that at long distance the overlap between the monomer's wavefunctions can be ignored. As a result, a set of n_A electrons can be associated with fragment A and, consequently, a Hamiltonian H^A can be defined for fragment A in terms of its electrons, likewise for fragment B. The unperturbed Hamiltonian is thus defined as $H^0 = H^A + H^B$, whereas the perturbation will consist of the electrostatic interactions between the nuclei and electrons of A with those of B

$$V = \sum_{a\in A}\sum_{b\in B} \frac{e_a e_b}{4\pi\epsilon_0 r_{ab}} \quad (2.3.27)$$

here e_a is the charge of particle a, one of the particles of fragment A, and r_{ab} is the distance between it and particle b in molecule B.
Alternatively, V can be expressed in terms of the charge distribution. For fragment A it is defined as:

$$\hat{\varrho}_A = \sum_{a\in A} e_a \delta(\boldsymbol{r} - \boldsymbol{a}) \quad (2.3.28)$$

where vector $\boldsymbol{a}$ is the position vector of fragment A, $\delta(\boldsymbol{r} - \boldsymbol{a})$ is the Dirac delta function. The definition for monomer B is analogous. In these terms V can be reformulated as

$$V = \int\int \sum_{a\in A}\sum_{b\in B} \frac{e_a\delta(\boldsymbol{r}-\boldsymbol{a})e_b\delta(\boldsymbol{r'}-\boldsymbol{b})}{4\pi\epsilon_0|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r}d\boldsymbol{r'} = \int\int \frac{\hat{\varrho}_A(\boldsymbol{r})\hat{\varrho}_B(\boldsymbol{r'})}{4\pi\epsilon_0|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r}d\boldsymbol{r'} \quad (2.3.29)$$

As usual in perturbation theory both the wavefunction and the interaction energy are expressed as a power series in λ

$$\psi^{AB}(\lambda) = \psi_{pol}^{(0)} + \lambda\psi_{pol}^{(1)} + \lambda^2\psi_{pol}^{(2)} + ... \quad (2.3.30)$$

$$E_{int}(\lambda) = \lambda E^{(1)}_{pol} + \lambda^2 E^{(2)}_{pol} + ... \tag{2.3.31}$$

Regarding the unperturbed state $\psi^{(0)}_{pol}$, it is expressed as a simple product of eigenfunctions of H^0, $\psi^A_m\psi^B_n$. The different energy components of the interaction energy for a closed shell system are found using RSPT for the ground state of the system, labelled by $m = n = 0$, thus the unperturbed ground state would be $|\psi^{AB}_0\rangle = |\psi^A_0\psi^B_0\rangle$.

Electrostatic Interaction

It corresponds to the first order energy correction as this is the expectation value of the perturbation operator

$$E^{(1)}_{pol} = \langle\psi^{AB}_0|V|\psi^{AB}_0\rangle = \langle\psi^A_0\psi^B_0|V|\psi^A_0\psi^B_0\rangle \tag{2.3.32}$$

Expressing V in terms of eq. 2.3.29 we get

$$E^{(1)}_{pol} = \int\int \frac{\varrho_A(\boldsymbol{r})\varrho_B(\boldsymbol{r'})}{4\pi\epsilon_0|\boldsymbol{r}-\boldsymbol{r'}|} d\boldsymbol{r}d\boldsymbol{r'} \tag{2.3.33}$$

where $\varrho_A(\boldsymbol{r})$ is the expectation value of $\hat{\varrho}_A(\boldsymbol{r})$.

Induction and Dispersion Interactions

Induction and dispersion interactions correspond to the second-order energy correction. This term is separated into three terms considering excitations of each fragment at a time as well the simultaneous excitation of both fragments resulting in:

$$E^{(2)}_{pol} = E^A_{ind} + E^B_{ind} + E_{disp} \tag{2.3.34}$$

where

$$E^A_{ind} = \sum_{m\neq0} \frac{\langle\psi^{AB}_0|V|\psi^A_m\psi^B_0\rangle\,\langle\psi^A_m\psi^B_0|V|\psi^{AB}_0\rangle}{E^A_0 - E^A_m} \quad E^B_{ind} = \sum_{n\neq0} \frac{\langle\psi^{AB}_0|V|\psi^A_0\psi^B_n\rangle\,\langle\psi^A_0\psi^B_n|V|\psi^{AB}_0\rangle}{E^B_0 - E^B_n} \tag{2.3.35}$$

$$E_{disp} = \sum_{m\neq0}\sum_{m\neq0} \frac{\langle\psi^{AB}_0|V|\psi^A_m\psi^B_n\rangle\,\langle\psi^A_m\psi^B_n|V|\psi^{AB}_0\rangle}{(E^A_0 + E^B_0) - (E^A_m + E^B_n)} \tag{2.3.36}$$

Exchange-repulsion Energy

The *Polarisation approximation* described above works well at long range distances but fails at short range. The main reason for this failure is that the polarisation approximation misses the spin repulsion between fragments that occurs at short distances as a result of the wavefunction overlap. At this point exchange can no longer be ignored.

The need for antisymmetrisation complicates considerably the perturbational treatment of intermolecular interactions. By assigning electrons to each individual molecule in order to separate the Hamiltonian into an unperturbed part H^0 and a perturbation λV, $(H = H^0 + \lambda V)$, we obtain that H^0 and λV are not symmetric with respect to the permutation (P) of electrons between fragments while the total Hamiltonian is.

$$[P, H^0] \neq 0 \quad [P, \lambda V] \neq 0 \quad [P, (H^0 + \lambda V)] = 0 \implies [P, H^0] = -[P, \lambda V] \neq 0 \tag{2.3.37}$$

Hence, there is a non-zero first order quantity ($[P, \lambda V]$) equal to a zeroth-order quantity ($[P, H^0]$) which contradicts a basic assumption in RSPT as we can no longer collect terms in powers of λ.
Another problem arises when antisymmetrising the wavefunction to satisfy the Pauli principle:

$$|\psi_m^A \psi_n^B\rangle = \mathcal{A} |\psi_m^A\rangle |\psi_n^B\rangle \tag{2.3.38}$$

where $\mathcal{A}$ is the antisymmetriser operator acting on the Hartree product $|\psi_m^A\rangle |\psi_n^B\rangle$. A direct consequence derived form eq.2.3.38 is that the wavefunction $|\psi_m^A \psi_n^B\rangle$ is antisymmetric with respect to electron permutation between fragments, whereas $|\psi_m^A\rangle$ and $|\psi_n^B\rangle$ are only symmetric with respect to permutation of A and B electrons among themselves. In this situation, simple products are orthogonal whereas the antisymetrised products are not. Since Hermitian operators have orthogonal eigenstates there can not be a zeroth-order Hamiltonian with these antisymmetrised products as eigenstates. To overcome the problems derived from the inclusion of antisymmetrisation, a new version RSPT must be used. In this sense, *Symmetry Adapted Perturbation Theory* (SAPT)[33] has become a popular alternative.

SAPT avoids the non-orthogonality of the antisymmetrised product by taking as expansion function the simple product $|\psi_m^A\rangle |\psi_n^B\rangle$ without antisymmetrisation. $|\psi_m^A\rangle |\psi_n^B\rangle$ are eigenfunctions of H^0 in which some electrons are assigned to A and the rest to B. The antisymmetrisation effect is introduced separately. Denoting by ψ_k^{AB} a member of the set of unsymmetrised simple products, with $\psi_0^{AB} = |\psi_0^A\rangle |\psi_0^B\rangle$ we find that the first correction to the interaction energy corresponds to the exchange-repulsion energy

$$E_{er} = \langle \mathcal{A}\psi_0^{AB}| V |\mathcal{A}\psi_0^{AB}\rangle \tag{2.3.39}$$

An illustrative case is given by two hydrogen atoms in their ground (1s) state. If we consider the overlap of the charge density, the wavefunction of the system becomes

$$\psi = \sqrt{\frac{1}{2}} |a(1)b(2) - a(2)b(1)\rangle \tag{2.3.40}$$

where a is the normalised 1s wavefunction for atom A and equivalently for b. Defining the atomic Hamiltonians of A and B as $H^A(2)$ and $H^B(1)$, the interatomic interaction operator is given by

$$V = \frac{1}{r_{AB}} + \frac{1}{r_{12}} - \frac{1}{r_{A1}} - \frac{1}{r_{B2}} \tag{2.3.41}$$

where r_{AB} is the inter-nuclear distance, r_{12} the inter-electron distance, r_{A1} is the distance between nucleus A and electron 1 (assigned to B) and likewise for r_{B2}. It can be shown that the first order energy correction results in the electrostatic term provided by the polarisation approximation as well as the following extra terms:

$$\begin{aligned} E_{er} = &-\frac{S^2}{1-S^2}\left[\langle b| \frac{1}{r_A} |b\rangle + \langle a| \frac{1}{r_B} |a\rangle - \langle a(1)b(2)| \frac{1}{r_{12}} |a(1)b(2)\rangle\right] + \\ &+ \frac{1}{1-S^2}\left[S \langle a| \frac{1}{r_A} + \frac{1}{r_B} |b\rangle - K_{ab}\right] \end{aligned} \tag{2.3.42}$$

where $S = \langle a(1)|b(1)\rangle$ and K_{ab} is the exchange integral, i.e., $K_{ab} = \langle a(1)b(2)|1/r_{12}|a(2)b(1)\rangle$ which is a negative (attractive) contribution, while the term $\langle a|\,\frac{1}{r_A} + \frac{1}{r_B}\,|b\rangle$ represents the repulsion arising from the electron density overlap. The rest of the components can be thought of as corrections to the electrostatic energy. Overall E_{er} is a positive, thus repulsive contribution and for that reason it is known as the *exchange-repulsion* energy.

In short, the antisymmetrisation of the wavefunction modifies the electron density so that a repulsive force appears on the nuclei. In this example the effects of wavefunction antisymmetrisation have only been shown for the first energy correction but they also manifest in higher order terms, contributing to the induction and dispersion components of the interaction energy.

2.3.3 Electron Density based EDA

The energy decomposition scheme presented so far is based on Molecular Orbital theory and it undergoes serious difficulties with the construction of a supermolecular wavefunction from the fragment's wavefunctions. An interesting alternative is the use of a density based EDA, overcoming the difficulties derived from the complex treatment of the wavefunction[34]. In this framework, the EDA scheme performed in this work[29, 35] is carried out by means of the unperturbed monomer electron densities as well as the intermolecular deformation densities associated with each energy term.

This approach is possible since, recalling the principles of DFT and Pair-Density Functional Theory (PDFT), the energy of any molecular or atomic system can be explained in terms of the one and two-electron densities, the latter expressed as sum of the product of the one-electron densities and the exchange-correlation density. In the case of two interacting fragments both densities and the nuclear electrostatic potential and energy can be expressed in terms of their non-interacting values plus their changes along the interaction. This way we can think of the two electron deformation density, $\Delta\rho_{xc}(\boldsymbol{r})$, and the one-electron deformation density $\Delta\rho(\boldsymbol{r})$, which is indeed sum of two contributions. On one side, there is the effect of the Pauli exclusion principle, $\Delta\rho(\boldsymbol{r})_{Pau}$, which is the difference between the electron-density obtained from the antisymmetrised wavefunction and the electron-density obtained as a Hartree product. On the other hand, there is the electron polarisation due to the intermolecular interaction $\Delta\rho(\boldsymbol{r})_{Pol}$, given by the difference between the total and Pauli deformation densities.

Working with these deformation densities as well as the unperturbed electron densities from the monomers it is possible to separate the interaction energy in its different components.

Recalling the *supermolecular approach*, the energy of the AB complex can be expressed in terms of the one-electron and exchange-correlation densities, $\rho(\boldsymbol{r})$ and $\rho_{xc}(\boldsymbol{r}_1, \boldsymbol{r}_2)$ respectively, in the following way

$$\begin{aligned} E_{AB} = &-\frac{1}{2}\int \nabla^2\rho(\boldsymbol{r},\boldsymbol{r'})\Big|_{\boldsymbol{r'}=\boldsymbol{r}} d\boldsymbol{r} + \int \hat{v}_N\rho(\boldsymbol{r})d\boldsymbol{r} + \frac{1}{2}\int\int \frac{\rho(\boldsymbol{r}_1)\rho(\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1 d\boldsymbol{r}_2 + \\ &+\frac{1}{2}\int\int \frac{\rho_{xc}(\boldsymbol{r}_1,\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1 d\boldsymbol{r}_2 + \sum_{i=1}^{N-1}\sum_{j>i}^{N}\frac{Z_i Z_j}{|\boldsymbol{R}_i-\boldsymbol{R}_j|} \end{aligned} \tag{2.3.43}$$

where the one-electron density matrix $\rho(\boldsymbol{r},\boldsymbol{r'})\Big|_{\boldsymbol{r'}=\boldsymbol{r}}$ is used instead of the one-electron density for the calculation of the kinetic energy term and $\hat{v}_N$ represents the nuclei potential operator.

Operators and electron densities for the complex energy in eq.2.3.43 can be rewritten in terms of the non-interacting fragments. The last term in eq.2.3.43 can be rewritten as

$$\sum_{i=1}^{N_A-1}\sum_{j>i}^{N_A}\frac{Z_iZ_j}{|\boldsymbol{R}_i-\boldsymbol{R}_j|}+\sum_{k=N_A+1}^{N-1}\sum_{l>k}^{N}\frac{Z_kZ_l}{|\boldsymbol{R}_k-\boldsymbol{R}_l|}+\sum_{i=1}^{N_A}\sum_{k=N_A+1}^{N}\frac{Z_iZ_k}{|\boldsymbol{R}_i-\boldsymbol{R}_k|} \tag{2.3.44}$$

For the nuclei potential operator

$$\hat{v}_N=\hat{v}_{N_A}+\hat{v}_{N_B} \tag{2.3.45}$$

where N is the total number of nuclei in the complex and N_A the number of nuclei in fragment A, thus $N_B=N-N_A$. As for the electron densities

$$\rho(\boldsymbol{r})=\rho^A(\boldsymbol{r})+\rho^B(\boldsymbol{r})+\Delta\rho(\boldsymbol{r})_{Pol}+\Delta\rho(\boldsymbol{r})_{Pau} \tag{2.3.46}$$

$$\rho_{xc}(\boldsymbol{r}_1,\boldsymbol{r}_2)=\rho_{xc}^A(\boldsymbol{r}_1,\boldsymbol{r}_2)+\rho_{xc}^B(\boldsymbol{r}_1,\boldsymbol{r}_2)+\rho_x^{AB}(\boldsymbol{r}_1,\boldsymbol{r}_2)+\Delta\rho_{xc}(\boldsymbol{r}_1,\boldsymbol{r}_2) \tag{2.3.47}$$

where the last to terms in the right side of eq.2.3.46 correspond to the contributions from the Pauli repulsion and polarisation. Interfragment exchange and polarisation are represented by the last two terms in eq.2.3.47. Merging the nuclei and electron potentials associated with each non-interacting fragment, the unperturbed fragment potentials $\hat{v}_A$ and $\hat{v}_B$ can be defined

$$\hat{v}_A(\boldsymbol{r})=\hat{v}_{N_A}(\boldsymbol{r})+\int\frac{\rho_A(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{r'}|}d\boldsymbol{r};\quad \hat{v}_B(\boldsymbol{r})=\hat{v}_{N_B}(\boldsymbol{r})+\int\frac{\rho_A(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{r'}|}d\boldsymbol{r} \tag{2.3.48}$$

In terms of eq.2.3.43 to 2.3.48 the interaction energy of the system can be split in its electrostatic, exchange, repulsive and polarisation components ($E_{int}=E_{elec}+E_{exch}+E_{rep}+E_{pol}$). Each of them presented here below:

$$E_{elec}=\int\hat{v}_{N_A}\rho_B(\boldsymbol{r})d\boldsymbol{r}+\int\hat{v}_{N_B}\rho_A(\boldsymbol{r})d\boldsymbol{r}+\int\int\frac{\rho_A(\boldsymbol{r}_1)\rho_B(\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1d\boldsymbol{r}_2+\sum_{i=1}^{N_A}\sum_{k=N_A}^{N}\frac{Z_i^AZ_k^B}{|\boldsymbol{R}_i-\boldsymbol{R}_k|} \tag{2.3.49}$$

$$E_{exch}=\frac{1}{2}\int\int\frac{\rho_x^{AB}(\boldsymbol{r}_1,\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1d\boldsymbol{r}_2 \tag{2.3.50}$$

$$\begin{aligned}E_{rep}=&-\frac{1}{2}\int\nabla^2\Delta\rho_{Pau}(\boldsymbol{r},\boldsymbol{r'})\Big|_{\boldsymbol{r'}=\boldsymbol{r}}d\boldsymbol{r}+\int\hat{v}_A\Delta\rho_{Pau}(\boldsymbol{r})d\boldsymbol{r}+\\&+\int\hat{v}_B\Delta\rho_{Pau}(\boldsymbol{r})d\boldsymbol{r}+\frac{1}{2}\int\int\frac{\Delta\rho_{Pau}(\boldsymbol{r}_1)\Delta\rho_{Pau}(\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1d\boldsymbol{r}_2\end{aligned} \tag{2.3.51}$$

$$\begin{aligned}E_{pol}=&-\frac{1}{2}\int\nabla^2\Delta\rho_{Pol}(\boldsymbol{r},\boldsymbol{r'})\Big|_{\boldsymbol{r'}=\boldsymbol{r}}d\boldsymbol{r}+\int\hat{v}_A\Delta\rho_{Pol}(\boldsymbol{r})d\boldsymbol{r}+\int\hat{v}_B\Delta\rho_{Pol}(\boldsymbol{r})d\boldsymbol{r}\\&+\frac{1}{2}\int\int\frac{\Delta\rho_{Pol}(\boldsymbol{r}_1)\Delta\rho_{Pol}(\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1d\boldsymbol{r}_2+\int\int\frac{\Delta\rho_{Pau}(\boldsymbol{r}_1)\Delta\rho_{Pol}(\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1d\boldsymbol{r}_2+\\&+\frac{1}{2}\int\int\frac{\Delta\rho_{xc}(\boldsymbol{r}_1,\boldsymbol{r}_2)}{|\boldsymbol{r}_2-\boldsymbol{r}_1|}d\boldsymbol{r}_1d\boldsymbol{r}_2\end{aligned} \tag{2.3.52}$$

Since eq.2.3.50 and 2.3.51 both arise from the Pauli exclusion principle they are usually grouped under the label of *exchange-repulsion energy* or *Pauli energy*, E_{Pau}. Contrary, the polarisation term can be split in two contributions, namely induction and dispersion, by means of second order PT as discussed in the above section. To determine the expression of the induction term we start defining the *charge-induction* energy, E_{ch-ind}

$$E_{ch-ind} = \int \hat{v}_A \Delta\rho_{Pol}(\boldsymbol{r})d\boldsymbol{r} + \int \hat{v}_B \Delta\rho_{Pol}(\boldsymbol{r})d\boldsymbol{r} \tag{2.3.53}$$

which is indeed the sum of the second and third term on the right side of eq.2.3.52. From RSPT it is known that the second order energy correction is dependent on the first order correction to the wavefunction, the density in this case. Such expression is given by

$$\Delta\rho(\boldsymbol{r}) = \Delta\rho_A(\boldsymbol{r}) + \Delta\rho_B(\boldsymbol{r}) = 2\sum_{m\neq 0} \frac{\int \hat{v}_B \rho_A^{m0}(\boldsymbol{r})d\boldsymbol{r}}{E_A^0 - E_A^m}\rho_A^{m0} + 2\sum_{n\neq 0} \frac{\int \hat{v}_A \rho_B^{n0}(\boldsymbol{r})d\boldsymbol{r}}{E_B^0 - E_B^n}\rho_B^{n0} \tag{2.3.54}$$

which can be split into fragments A and B. Substituting eq.2.3.54 in eq.2.3.53 we find that the charge induction energy is given by

$$\begin{aligned} E_{ch-ind} =& 2\sum_{m\neq 0} \frac{\left(\int \hat{v}_B \rho_A^{m0}(\boldsymbol{r})d\boldsymbol{r}\right)^2}{E_A^0 - E_A^m} + 2\sum_{n\neq 0} \frac{\left(\int \hat{v}_A \rho_B^{n0}(\boldsymbol{r})d\boldsymbol{r}\right)^2}{E_B^0 - E_B^n} + \\ &+ 2\sum_{m\neq 0} \frac{\int \hat{v}_B \rho_A^{m0}(\boldsymbol{r})d\boldsymbol{r} \int \hat{v}_A \rho_A^{m0}(\boldsymbol{r})d\boldsymbol{r}}{E_A^0 - E_A^m} + 2\sum_{n\neq 0} \frac{\int \hat{v}_A \rho_B^{n0}(\boldsymbol{r})d\boldsymbol{r} \int \hat{v}_B \rho_B^{n0}(\boldsymbol{r})d\boldsymbol{r}}{E_B^0 - E_B^n} \end{aligned} \tag{2.3.55}$$

As derived from perturbation theory the first two terms in eq.2.3.55 correspond to the induction energy, hence rearranging terms

$$E_{ind} = \frac{1}{2}\left[E_{ch-ind} - 2\sum_{m\neq 0} \frac{\int \hat{v}_B \rho_A^{m0}(\boldsymbol{r})d\boldsymbol{r} \int \hat{v}_A \rho_A^{m0}(\boldsymbol{r})d\boldsymbol{r}}{E_A^0 - E_A^m} - 2\sum_{n\neq 0} \frac{\int \hat{v}_A \rho_B^{n0}(\boldsymbol{r})d\boldsymbol{r} \int \hat{v}_B \rho_B^{n0}(\boldsymbol{r})d\boldsymbol{r}}{E_B^0 - E_B^n}\right] \tag{2.3.56}$$

Finally replacing back the expression of the first order correction for the density (eq.2.3.54) we find the induction energy to be

$$E_{ind} = \frac{1}{2}\left[E_{ch-ind} - \int \hat{v}_A \Delta\rho_A(\boldsymbol{r})d\boldsymbol{r} - \int \hat{v}_B \Delta\rho_B(\boldsymbol{r})d\boldsymbol{r}\right] \tag{2.3.57}$$

which recalling the expression for E_{ch-ind} (eq.2.3.53) we can rewrite as

$$E_{ind} = \frac{1}{2}\left[\int \hat{v}_A \Delta\rho_B(\boldsymbol{r})d\boldsymbol{r} + \int \hat{v}_B \Delta\rho_A(\boldsymbol{r})d\boldsymbol{r}\right] \tag{2.3.58}$$

The discussion in this section has exposed how each of the interaction energy terms are computed, bare in mind that the dispersion term and further higher order corrections are determined as the difference between the polarisation and induction term, with the second order dispersion contributing the most to the term.

Chapter 3

Computational Details

The studied systems consist in a series of solute molecules; the five nucleobases, the ammonium cation and tetrafluoromehtane, all of them located at the centre of a truncated-octahedron simulation box fitting a sphere of radius 20 $\mathring{A}$ surrounded by a series of water molecules ranging from 2566 to 3051 depending on the simulated system. Water molecules were added with the TIP3P[36] solvation model, where each molecule is represented as a rigid monomer and the dimerisation energy between monomers m and n, E_{mn} is computed by means of a Coulomb and Lennard-Jones potential (between oxygens). Each monomer is parametrised as shown in Table 3.1 here below

Table 3.1: TIP3P water parameters. $r(OH)$ is the hydrogen-oxygen distance in $\mathring{A}$, $\angle HOH$ is the corresponding angle in degrees, A is the Lennard-Jones ”size of the particle parameter” in $kcal\mathring{A}^{12}/mol$, C is the Lennard-Jones ”size of the particle parameter” in $kcal\mathring{A}^{6}/mol$ and q is the atomic charge

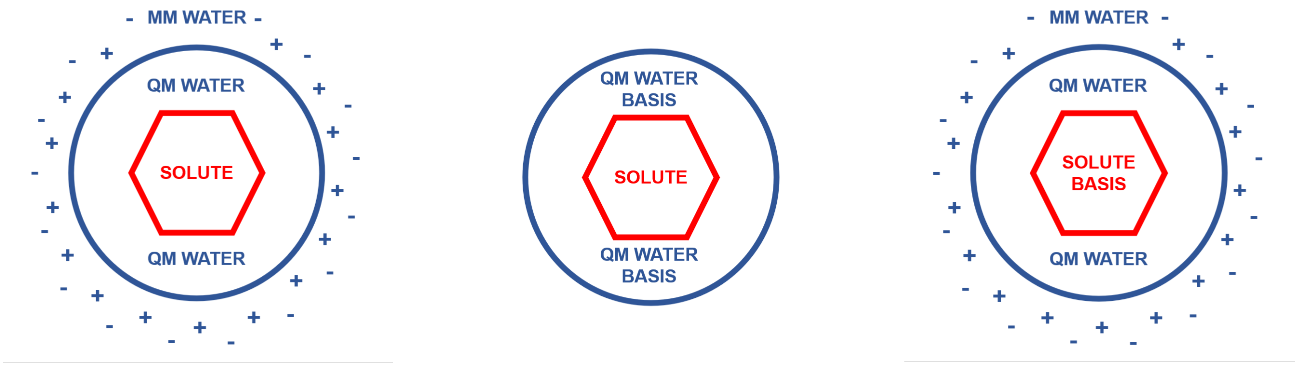
TIP3P water parameters

	r(OH)	$\angle HOH$	A x 10^{-3}	C	q(O)	q(H)
TIP3P	0.9572	104.52	582.0	595.0	-0.834	0.417

The geometries of the solutes and their charges were obtained from a previous geometry optimisation carried out with the Gaussian09 package [37] at the M062X/cc-pVDZ level. An MD simulation of 100ns was carried out for each simulated system in the NPT ensemble at a constant pressure and temperature of 1bar and 300K. The Berendsen barostat and the Langevin thermostat were used for this purpose. The Lennard-Jones potential cut-off was set to 12 $\mathring{A}$ and electrostatic interactions were computed with the particle-mesh Ewald method[38]. A preceding 0.5ns thermalisation step was carried out in the NVT ensemble to bring the system to the desired temperature of 300K, itself preceded by a short relaxation step of 2000 simulation steps. Periodic boundary conditions where imposed in every MD simulation. All MD simulations where carried out with AMBER18[39], where the GAFF[21] was employed for the simulation of the solutes.

Single point QM and QM/MM calculations were performed with Gaussian16[40] at the M062X/cc-pVDZ level on selected geometries of the previously obtained MD trajectory to determine the electron densities that will be later required to perform the EDA. The AMBER program Cpptraj[41] was employed for the selection of the geometries along the trajectories.

For the nucleobases, 100 different configurations where picked from the last 50ns of the MD simulation, each of them equally spaced by a 0.5ns simulation time interval. An in-house python script was then used to define the QM and MM regions where, out of the total number of water molecules, a sphere



(a) Complex **(b)** Fragment 1 **(c)** Fragment 2

Figure 3.1: Schematic representation of the fragments required for the EDA.

containing the solute at its centre and a total of 100 from the closer water molecules were included in the QM region, while the remaining solvent was represented as point charges in the MM region.
As for CF_4 and NH_4^+ they where only employed as test systems for the convergence study of the different energy components with QM size and, thus, it was deemed unnecessary to carry out a configurational sampling for these systems along the MD trajectory. Therefore, instead of a 100 molecular snapshots, only one was selected from the last 50ns of the MD trajectory. An increasingly larger QM region was defined on these systems ranging from 10 to 240 QM water molecules for the convergence study.
All single point calculations were performed defining three different fragments for each geometry of the MD trajectory: one for the whole system including the point charges of the MM region (complex), another containing the water molecules of the QM region and the basis functions of the solute as well as the point charges (fragment 2) and, lastly, the solute with the basis functions of the QM water molecules (fragment 1). The inclusion of the basis sets was done with ghost atoms, in order to account for the BSSE. An schematic representation of each considered fragment can be seen in Figure3.1.

Once the electron densities from the fragments defined above were computed, the interaction energy decomposition was performed with a locally developed Fortran90 code implementing the EDA scheme presented in Section 2.3.3. Such code computes the deformation densities required for the EDA and it extends its general treatment of the electron densities to the particular case of a QM/MM calculation. To do so, the external potential generated by the MM region, $\hat{V}_{ff}$, is added to the nuclear potential of one of the fragments, according to Figure 3.1 to fragment 2, resulting in

$$\hat{v}_{N_{F2}}^{QM/MM}(\boldsymbol{r}) = \hat{v}_{N_{F2}}(\boldsymbol{r}) + \hat{V}_{ff} \tag{3.0.1}$$

$\hat{V}_{ff}$ is also included for the calculation of the internuclear potential of the electrostatic energy (eq. 2.3.49).

For the convergence study of the different interaction energy components with increasing QM size, in addition to the scheme considered above, an alternative one where the MM charges were not considered was also implemented. This way, a comparison between QM and QM/MM convergence was obtained. To better asses the convergence of the data, the *Cauchy* criterion for sequence convergence was employed. The definition is provided here

Definition. *A sequence (a_n) of real numbers is a Cauchy sequence if for each $\epsilon > 0$ of $\mathbb{R}$, there is a natural number n_0 such that*

$$|a_p - a_q| < \epsilon$$

for any p,q $\geq n_0$. Any sequence of real numbers is convergent if and only if it is a Cauchy sequence.

Through this criterion a minimum threshold for data convergence can be defined and thus the convergence of the interaction energies can be discussed with this threshold and not uniquely based on the graphical representation of the data.

Chapter 4

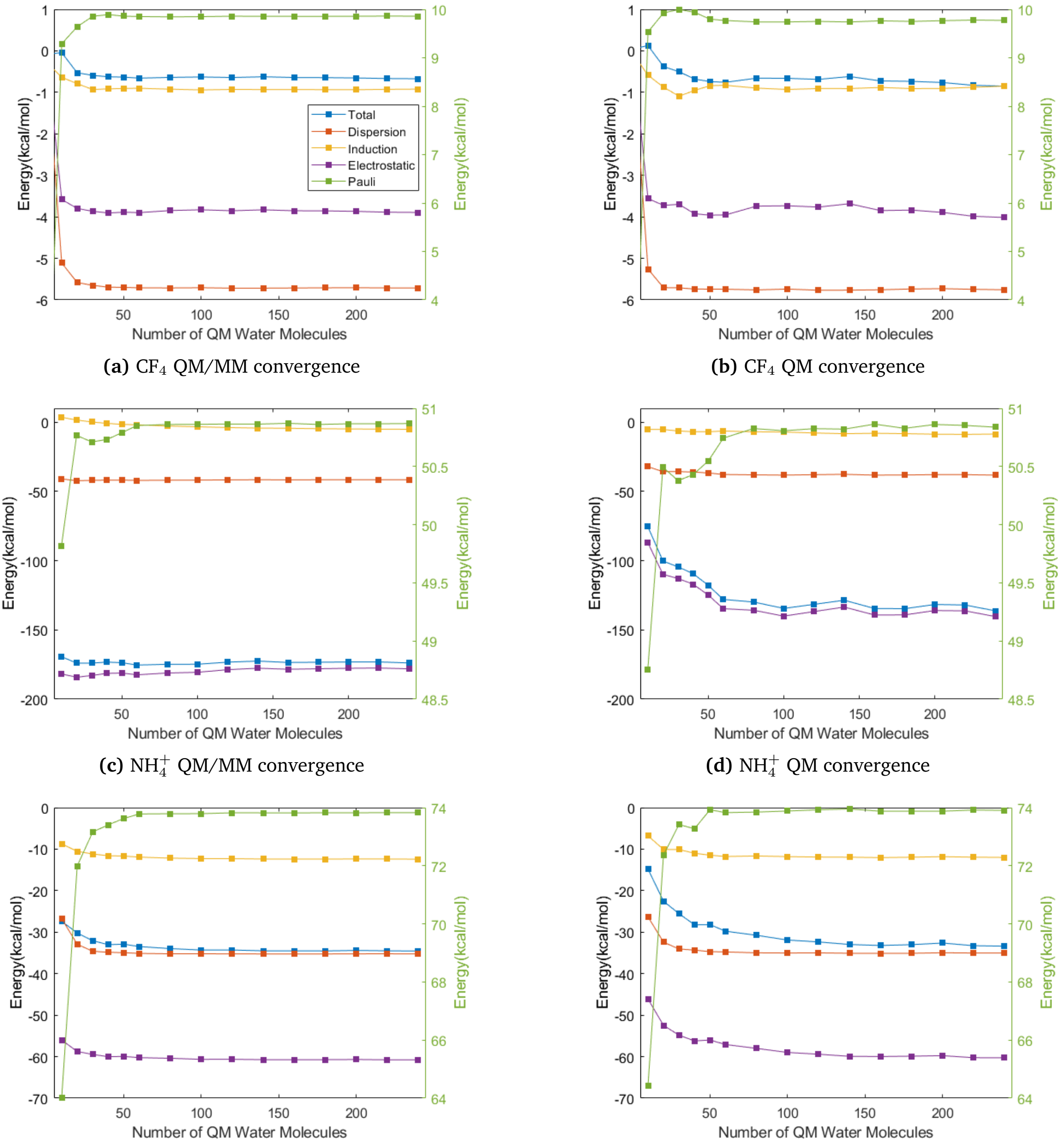
Results and Discussion

4.1 Convergence Study

As stated in the above section, we start by studying an adequate QM/MM partition for our system. For this purpose, a convergence study of the different components of the interaction energy is carried out. The aim of this study is twofold: on one hand, we need to establish an adequate size for the QM region so that we can compute the different components of the interaction energy for the different snapshots of the MD trajectory at their converged value. On the other hand, we want to evaluate the impact of the MM charges on the behaviour of the interaction energy. For this reason the convergence study compares the trend followed by each energy component within the QM/MM approach with the one described by a pure QM formulation.

Given the long range character of some interaction energy components, most importantly the electrostatic component, the converge study of the nucleobases with QM region size with was complemented with additional systems of varying polarity. In particular, CF_4 and NH_4^+ were considered as they represent two extreme cases of a hydrophobic and hydrophilic system in a water solution. Moreover, their highly symmetric (tetrahedral) character prevents trend deviations caused by strong interaction of water molecules with a particular region of the solute, as it may be the case for the nucleobases where certain functional groups favour strong local interactions with the solvent. Both tetrahedral molecules will undergo the same treatment as the nucleobases, beginning with a 100ns classical MD simulation and followed by the random selection of a single trajectory snapshot where a progressively larger QM solvation sphere is defined, ranging from 10 to 240 water molecules. The results of these simulations are shown in Figure 4.1 below.

By looking at Figure 4.1, we can see how the hydrophobic system CF_4 shows a better overall convergence than its hydrophilic counterpart NH_4^+. Comparing Figure 4.1a and 4.1c we can see that the electrostatic component of the interaction energy is the main driver for the increase in the total interaction energy of NH_4^+ with respect to that of CF_4, in accordance with the ionic character of the former. The fact that the electrostatic component decays with r^{-n}; $n = 1, 2, 3$ depending on the nature of the interaction, *i.e.*, is a long range interaction, explains the overall worsen of the convergence trend, requiring more explicit water molecules. The Pauli repulsion term shows a significant increase in the hydrophilic system going from a value of 9.86 kcal/mol for CF_4 to 50.87 kcal/mol for NH_4^+ when the 240 water solvation sphere is considered. Given the short range nature of the repulsive interaction, its increase is synonymous with a shorter distance between solute and solvent.



(a) CF_4 QM/MM convergence

(b) CF_4 QM convergence

(c) NH_4^+ QM/MM convergence

(d) NH_4^+ QM convergence

(e) Adenine QM/MM convergence

(f) Adenine QM convergence

Figure 4.1: Convergence of the interaction energy components in kcal/mol with increasing number of QM water molecules surrounding the CF_4 (a and b), NH_4^+ (c and d) and adenine (e and f) solutes. Figures a, c and e include convergence with point charges surrounding the QM region while b, d and f do not. Total interaction energy in blue, dispersion in red, induction in yellow, electrostatic in purple and Pauli in green. The Pauli repulsion term is represented in the scale shown in the right vertical axis.

As for the induction term, it shows a positive value for the first few points in the case of the ammonium cation in the presence of MM point charges as it can be seen from Figure 4.1c. These values are clearly nonphysical since, according to perturbation theory, the second order energy correction from which the induction component originates must always be negative. Besides, induction is an attractive force representing the distortion of a molecule's charge distribution due to the electric field generated by its neighbours. Taking a look at Figure 4.1d one finds that the induction component is no longer positive for any QM solvation sphere, for this reason it is reasonable to think that the lack of polarisability of the TIP3P force field may be responsible for the unexpected induction values in Figure 4.1c. Furthermore, since the ammonium cation is charged and smaller in size than CF_4 it has a larger polarisation effect on its environment, meaning that the lack of polarisability of TIP3P penalises the induction term to a greater extent in the case of NH_4^+.
Regarding the values of the induction and dispersion terms, it is worth noting that both molecules are tetrahedric and thus they have no net dipole moment, which explains the small value of the induction term in both cases. As for the dispersion component, it is the main contribution to the attractive interaction energy in the case of CF_4. Nonetheless, its value is notably larger for the ammonium cation, the reason behind being the smaller interparticle distance as the solute-solvent interaction is greater for the cationic system.

Comparing now the energy components of each system obtained under the presence of MM point charges and those obtained only within the QM region, i.e., comparing Figures 4.1a, 4.1c and 4.1e with Figures 4.1b,4.1d and 4.1f, respectively, we can observe how the convergence trend is not as clear in the absence of point charges. To aid the discussion, Tables 4.1 and 4.2 present the energy value for the largest solvation sphere of 240 water molecules, which approaches a radius of 12 $\mathring{A}$, as well as the determined threshold for each component for a QM region containing 100 water molecules. To determine the convergence threshold, the *Cauchy* criterion for sequence convergence is used, its definition can be found in Section 3. According to it, for a given natural number n_0 we can determine the value of a real number $\delta > 0$ satisfying the condition $\delta < \epsilon$ such that our sequence is a *Cauchy* sequence. To do so, a particular n_0 is selected, in our case a_{n_0} would be the energy term corresponding to a QM sphere of a particular size, e.g., the QM sphere of 100 water molecules. To determine the threshold (δ), the set of all possible differences for any $p, q \geq n_0$ are computed and the maximum of such set is selected as threshold, thus satisfying the condition $|a_p - a_q| \leq \delta < \epsilon$ for all $p, q \geq n_0$. In the previous example, elements a_p and a_q would be the energy terms corresponding to the QM spheres of 100, 120, 140 water molecules and so on. As one could expect from the trends described in Figure 4.1, the greater the value of n_0, the smaller the threshold δ.

Table 4.1: Energy convergence thresholds and energy values for the 240 water QM sphere (W240) both in *kcal/mol* for all interaction energy components of the CF_4 system. Data is presented for the system including MM point charges (QM/MM) and lacking them (QM).

	QM/MM		QM	
Energy Term	Threshold (δ)	Energy (W240)	Threshold (δ)	Energy(W240)
Total	0.0480	-0.6730	0.2310	-0.8500
Electrostatic	0.0640	-3.8940	0.3310	-4.0150
Induction	0.0190	-0.9230	0.0810	-0.8500
Dispersion	0.0100	-5.7140	0.0380	-5.7570
Repulsion	0.0120	9.8580	0.0390	9.7710

Table 4.2: Energy convergence thresholds and energy values for the 240 water QM sphere (W240) both in *kcal/mol* for all interaction energy components of the NH_4^+ system. Data is presented for the system including MM point charges (QM/MM) and lacking them (QM).

	QM/MM		QM	
Energy Term	Threshold (δ)	Energy (W240)	Threshold (δ)	Energy(W240)
Total	2.2900	-173.8310	7.6550	-136.2180
Electrostatic	3.0910	-178.0840	6.8470	-140.3360
Induction	1.8200	-5.1630	1.7010	-8.5880
Dispersion	0.3110	-41.5940	0.5910	-38.2430
Repulsion	0.0100	50.8690	0.0570	50.8380

Data in Table 4.1 shows a clear trend in convergence related to each interaction energy component, the long range components having a worse convergence than the short range Pauli repulsion, which shows the clearest trend along with the dispersion component for CF_4. The lack of MM charges polarising the electron density of the QM region worsens significantly the data convergence for the same size of the QM water sphere. Once again, the long range interactions are the most affected. As a consequence, including the polarisation effects of the MM point charges results in a better description of the system, implying that a smaller QM region can be used to achieve a good degree of data convergence, decreasing computational cost in the most demanding step of the simulation, which is indeed the computation of the electron density of the fragments.
The same is true for the ammonium cation as shown in Table 4.2. However, unlike CF_4, we do not find a good agreement between the energy components obtained with and without MM point charges. The reason for this deviation is found in the required neutrality of the *Ewald summation*[42] used in the MD simulation. The *Ewald summation* computes electrostatic interactions when PBC are used. It speeds up the summation of the coulombic interactions and ensures the absolute convergence of the series.
These required neutrality implies that a counterion, in this case Cl^-, must be included in the system during the MD simulation. As a result, the geometries used in the convergence study include this ion, which is treated as a point charge in the MM region when the QM/MM partition is carried out. This way, when performing the EDA without the point charges the counterion is eliminated resulting in the energy differences that can be seen in Table 4.2.

Concerning the choice of an adequate size for the EDA of the nucleobases, the same analysis discussed above is performed for the adenine molecule. To simplify the analysis, the convergence threshold for the energy components is studied only for the total interaction energy as it is clear from the previously studied cases that the global term reflects the converge trend described by each individual component. By taking several points along the trajectory described by the interaction energy we can asses the data convergence up to such point. Table 4.3 summarises the convergence trend by showing the normalised convergence thresholds with respect to the total interaction energy for a QM region of 240 water molecules. This way, the convergence thresholds can be compared with the limit value at a QM sphere of 240 water molecules.

Table 4.3: Normalised convergence threshold for the total interaction energy with increasing size of the QM region for the adenine molecule.

	W10	W20	W30	W50	W100	W200
Threshold	0.2071	0.1261	0.0723	0.0478	0.0072	0.0041

We can see a rapid increase in data convergence when going from 10 to 50 water molecules in the QM region, from there on there is a significant stagnation in the convergence rate. For this reason, 100 water molecules is deemed a good approximation to the converged value of the interaction energy components, as it represents a good compromise between accuracy and computational cost.

Outlining the ideas here exposed, we have seen that the inclusion of MM point charges polarising the electron density of a system improves the convergence of the interaction energy components, reducing the requirement for QM solvent molecules and thus saving computational resources. It is also clear that a significant number of QM water molecules are necessary to appropriately compute the interaction energies, as much as 100 water molecules were deemed necessary to converge the long range energy terms to an order of 10^{-3} when compared to the limit value of the interaction energy at 240 QM waters for the adenine molecule.

It is also worth noticing that some significant improvements in the description of the system could be made by including a polarisable forcefield for the solvent. Besides, quantum mechanical treatment of the counterions would also be desirable given their strong interaction with the solute.

4.2 Nucleobases EDA

As suggested above, the EDA is performed for each of the nucleobases surrounded by a QM region containing 100 water molecules and the remaining MM point charges. To asses the conformational freedom of the system, the EDA is carried out on a sample of a hundred MD trajectory snapshots whose statistical treatment will be performed under the assumption of normality. The energies obtained for the case of the adenine molecule are presented as a histogram in Figure 4.2. Equivalent representations can be found in the Annex section for the remaining nucleobases.

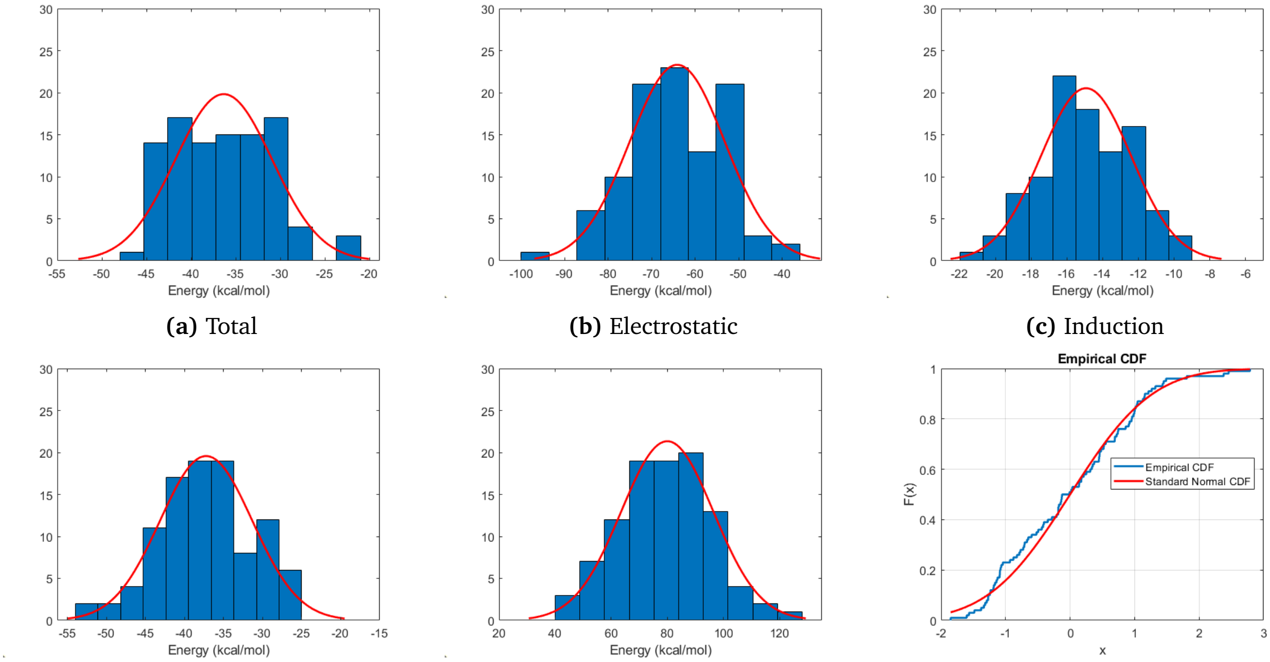
To prove that the assumption of normality is reasonable, which is purely based on the relatively large size of the sample, the *Kolmogorov–Smirnov*[43, 44] (KS) normality test is performed on each data sample. This statistic is a null hypothesis non-parametric test that is used to compare if the Empirical Cumulative Distribution Function (ECDF) for a given set of n samples follows that of a hypothetical (continuous) CDF, in this case, the one of a normal distribution function. The test statistic is defined as

$$D = \underset{1\leq i\leq n}{sup} |(\hat{F}_n(x_i) - F_0(x_i))|$$

where $\hat{F}_n(x)$ is the ECDF and $F_0(x)$ is the hypothetical CDF. Thus the test accepts the null hypothesis if the statistic is lesser or equal than a given tabulated value dependent on the hypothetical distribution at a given significance level. The comparison between the ECDF for the total interaction energy and the normal CDF can be seen in Figure 4.2f. For all the studied systems the KS test validated the null hypothesis for the total interaction energy component, confirming the normality of the sampled data at 5% significance. Since the normality assumption has been proven correct, the population statistics can be expressed within a certain confidence interval as shown in Table 4.4.

As we can see from Table 4.4 the sample mean shows how the repulsive Pauli term is the main contribution to the total interaction energy at 79.95 kcal/mol, while the electrostatic term is the largest within the attractive components at -64.14 kcal/mol, which can be related to adenine's ability to form hydrogen bonds in a polar solvents like water.

Regarding the sample's Standard Deviation (STD), we can see that is proportional to the sample mean. If we calculate the Relative Standard Deviation (RSTD) we see that all components fall under the 16% to 21% range, in particular, RSTDs of 15.9%, 16.9%, 17.1% and 20.6% are found for the dispersion,



(a) Total **(b)** Electrostatic **(c)** Induction

(d) Dispersion **(e)** Pauli **(f)** KS Test

Figure 4.2: Histogram for each interaction energy component with energy in *kcal/mol* in the horizontal axis and relative data frequency in the vertical axis with a fitted normal distribution function (red) for the adenine molecule.(e) ECDF and normal CDF used in the KS test.

Table 4.4: Sample mean ($\bar{x}$) and standard deviation (s) in *kcal/mol*. Population mean (μ) and standard deviation (σ) also in *kcal/mol* for a confidence interval of 95% for each interaction energy component for the adenine molecule. The percentage contribution of each energy component to the attractive energy according to the sample mean is shown in the left column.

		$\bar{x}$	μ	s	σ
Electrostatic	55.1%	-64.1421	[-66.3149, -61.9692]	10.9508	[9.61484, 12.7212]
Induction	12.9%	-14.9324	[-15.4337, -14.4312]	2.52619	[2.21802, 2.93462]
Dispersion	32.0%	-37.2431	[-38.4158, -36.0705]	5.90984	[5.18888, 6.86532]
Pauli		79.9528	[76.6893, 83.2162]	16.4470	[14.4406, 19.1061]
Total		-36.3354	[-37.4133, -35.2575]	5.43246	[4.76973, 6.31075]

induction, electrostatic and Pauli components respectively. These high RSTDs point out that even for a relatively simple system like the one studied here, the configurational sampling provided, in this case, by a classical MD trajectory, proves to be relevant when using quantum mechanical tools like the EDA. The highest RSTD is that of the repulsive Pauli component which is a short-range term and as such it is the most sensitive to small configurational variations when particles are closely interacting.

For comparison among the different nucleobases, the distribution functions for each energy component as well as the total interaction energies are presented in Figure 4.4. The same pattern concerning the arrangement of the energy terms is observed in all systems, where the Pauli repulsion and the electrostatic component contribute the most to the interaction energy and the STDs are proportional to the mean of the distribution. Also, Table 4.6 aids the discussion by displaying the sample means for

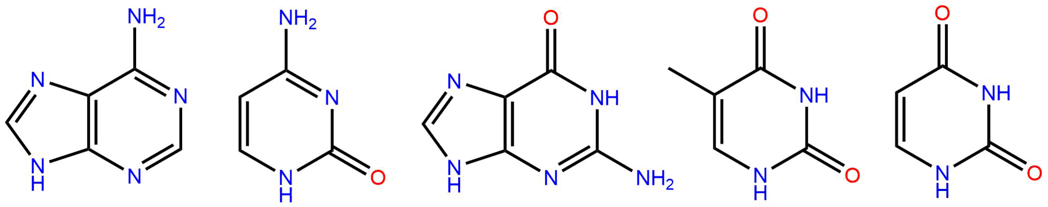
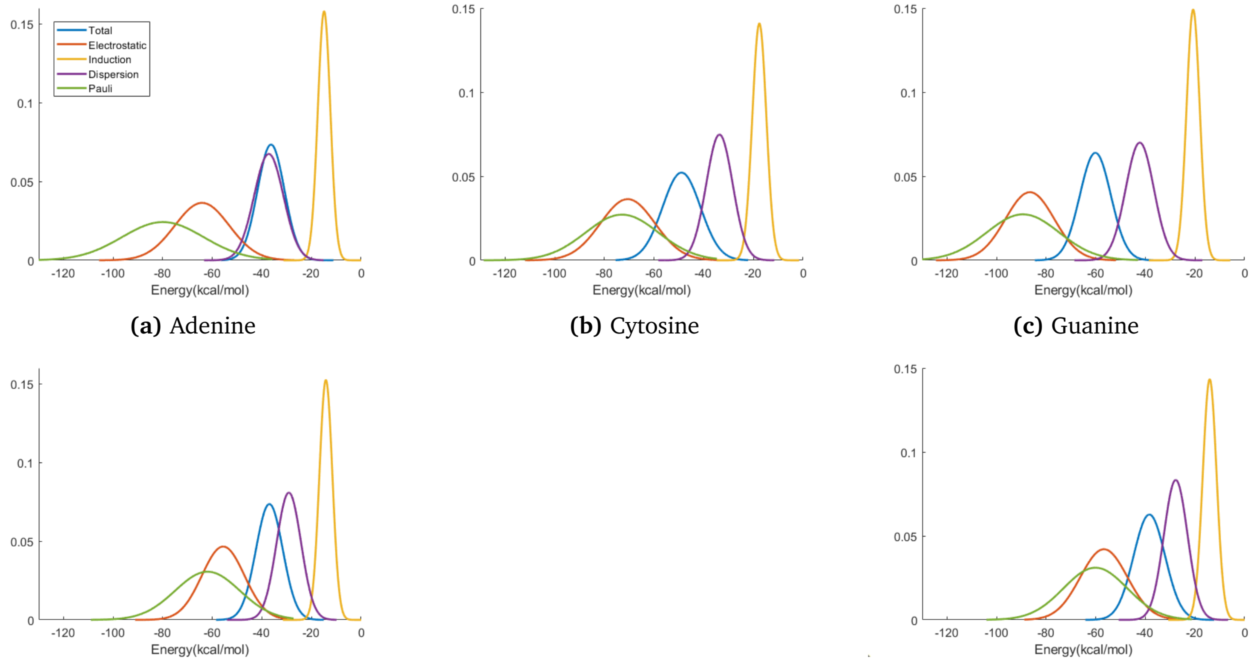


Figure 4.3: Schematic representation of the five nucleobases. From left to right: adenine, cytosine, guanine, thymine and uracil.

100 MD snapshots for all nucleobases. A more complete analysis of the remaining nucleobases in the fashion of Table 4.4 can be seen as an Annex in Table 5.1.



(a) Adenine

(b) Cytosine

(c) Guanine

(d) Thymine

(e) Uracil

Figure 4.4: Normal distributions for the sampled components of the interaction energy of each system. The distributions are normalised with the x axis representing the energy in kcal/mol. The Pauli energy is represented with opposite sign for comparison with the remaining energy components.

When it comes to the total interaction energy we find that the five nucleobases follow the trend: guanine > cytosine > uracil > adenine ∼ thymine. This relative order allows for a limited interpretation given the structural similarities between the purine and pyrimidine bases, a scheme showing the structure of the five nucleobases is shown in Figure 4.3. Taking a closer look at the different energy components, we can see the effect that this slightly different structural features induce in the interaction energy. When it comes to the electrostatic energy component guanine and cytosine exhibit the higher values of -86.59 kcal/mol and -70.63 kcal/mol, respectively, with adenine closely behind at -64.14 kcal/mol. The fact that both guanine and cytosine have a carbonyl and primary amine group allows the formation of a significant number of hydrogen bonds both of donor and acceptor nature, possibly explaining the high value of the electrostatic term in both molecules.

A remarkable observation is the the high value for the repulsive Pauli interaction energy in adenine, responsible for the relatively small interaction energy of this molecule in water despite the relatively high attractive contributions.

Dispersion follows an interesting trend as the purine basis (adenine and guanine) show the higher values followed by cytosine, thymine and uracil. This order generally agrees with the trend in electron delocalisation for the nucleobases in vacuum [45, 46]. However, the difference of 5.03 kcal/mol favouring the dispersion interaction of guanine instead of adenine is in disagreement with the trend described in the bibliography. The effect of solvation on the electron delocalization of the solute might be the cause for this deviation.
To further investigate the relation between the dispersion and electron delocalisation, the Delocalisation Indices (DIs)[47], a quantum chemical tool based on the n-electron density function with great sensitivity for the description of electron delocalization patterns, are computed for each sample. Table 4.5 compares the 6-centre (Δ_6) and 5-centre (Δ_5) DIs in vacuum and in water solution for each of the rings comprising the five nucleobases.

Table 4.5: 6-centre (Δ_6) and 5-centre (Δ_5) delocalization indices in $a.u.$ computed for each of the rings comprising the five nucleobases; adenine, cytosine, guanine, thymine and uracil. The values of Δ_6 and Δ_5 are shown for a single geometry in vacuum (the energy minimum), whereas the average for a sample of 100 geometries is shown for the water solution.

	Adenine	Cytosine	Guanine	Thymine	Uracil
			Vacuum		
Δ_6	0.6468	0.2898	0.1949	0.0908	0.1010
Δ_5	0.6615	-	0.7867	-	-
			Water Solution		
Δ_6	0.5709	0.3684	0.1967	0.1518	0.1758
Δ_5	0.2854	-	0.7991	-	-

From Table 4.5 we can see how solvation affects differently the electron delocalisation in the purine and pyrimidine bases. A general increase can be observed for the latter, whilst a remarkable decrease happens to the adenine molecule. As for guanine, no significant variation occurs. For the bicyclic structures the overall delocalisation is the summation of each indivual ring contribution, yielding a value of 0.9868 for guanine and 0.8563 for adenine. These results are now in perfect agreement with the trend described by the dispersion energy component of the interaction energy as it is clear from Table 4.6, as the ordering of the nucleobases according to the DIs now matches that of the dispersion energy.

Table 4.6: Sample mean ($\bar{x}$) in *kcal/mol* for each interaction energy component for the adenine, cytosine, guanine, thymine and uracil molecules.

	Adenine	Cytosine	Guanine	Thymine	Uracil
Electrostatic	-64.1421	-70.6258	-86.5936	-55.6232	-56.7488
Induction	-14.9324	-17.612	-20.7593	-14.2076	-14.0889
Dispersion	-37.2431	-33.6711	-42.2701	-29.1435	-27.7829
Pauli	79.9528	72.9155	89.3883	61.9451	60.1718
Total	-36.3354	-48.9534	-60.186	-36.999	-38.4177

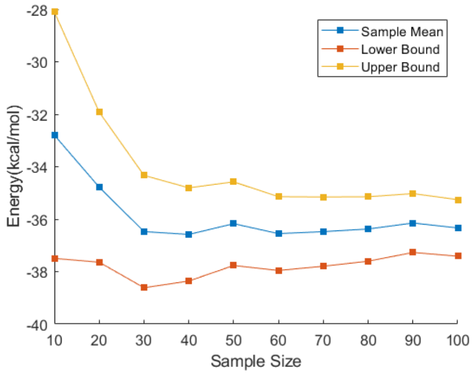


Figure 4.5: Sample mean as well as upper and lower bounds to the population mean at a 95% confidence interval for the total interaction energy (kcal/mol) with increasing sample size for the adenine molecule.

A final question may arise concerning the number of MD snapshots that are necessary to achieve a sufficiently precise description of the interaction energies for our population. So far, we have performed the data analysis for a sample of 100 MD snapshots, from which we can infer the population parameters μ and σ with reasonable accuracy. Nonetheless, and taking into account the high computational cost of performing the EDA for each snapshot, it would be interesting to estimate the minimal sample from which we can infer relevant population statistics without sacrificing accuracy. For this purpose, the same statistical procedure described previously is now performed for samples of increasing size from 10 to 100 molecular snapshots in increments of 10 elements for each sample. Thus, the study of the convergence of the sample and population statistical parameters, with a focus on the sample mean, for the total interaction energy is assessed. The results are shown in Figure 4.5 and a summary is also given in Table 4.7 below for the adenine molecule. Table 5.2 in the Annex section shows equivalent data for the remaining nucleobases.

The theoretical background for this approach can be found in the *Central Limit Theorem*, which states that if $X_1, X_2, ..., X_n$ are n independent random variables, equally distributed, drawn from a population with overall mean μ and finite variance σ^2, then the limiting form of

$$Z_n = \sqrt{n}\frac{(\bar{X} - \mu)}{\sigma}$$

is a standard normal distribution, where $\bar{X}$ is the sample mean. As a consequence, when $n \to \infty$ then $\bar{X} \to \mu$. This is indeed what is shown both in Table 4.7 and Figure 4.5. Not only does the sample mean approach a certain value, the length of the interval in which the population mean can be found for a 95% confidence interval is progressively reduced with increasing n. If we were to choose a smaller sample than our initial treatment of 100 MD snapshots, then a sample with 30 to 50 geometries would be an excellent compromise between accuracy and computational cost. As we can see from Table 4.7, for this range of samples we achieve a stable value for the $\bar{x}$ and s while the interval for the population parameters is kept relatively narrow at a 95% confidence interval.

Table 4.7: Sample mean ($\bar{x}$) and standard deviation (s) in *kcal/mol*. Population mean (μ) and standard deviation (σ) also in *kcal/mol* for a confidence interval of 95% for each interaction energy component for the adenine molecule with increasing sample size.

Sample size	$\bar{x}$	μ	s	σ
10	-32.7979	[-37.4959, -28.0999]	6.56738	[4.51727, 11.9895]
20	-34.7799	[-37.6422, -31.9176]	6.11582	[4.65102, 8.93259]
30	-36.4739	[-38.6155, -34.3324]	5.73518	[4.56753, 7.70989]
40	-36.5828	[-38.3591, -34.8065]	5.55407	[4.54968, 7.13163]
50	-36.1735	[-37.7669, -34.5801]	5.60672	[4.68348, 6.98672]
75	-36.2706	[-37.5413, -34.9999]	5.52309	[4.75880, 6.58214]
100	-36.3354	[-37.4133, -35.2575]	5.43246	[4.76973, 6.31075]

Chapter 5

Conclusions

In the above pages, the decomposition of the interaction energy components in the framework of QM/MM has been discussed for a series of solutes in a water solution. Through this approach, the EDA is extended from the study of simple dimeric systems to larger, more complex, ones where its physical insight can ease the description of otherwise inaccessible phenomena. The adaptation of the EDA from a purely QM formulation to a MM environment proves to be an effective approach for the description of intermolecular interactions in large systems.

The convergence study performed on the test systems CF_4 and NH_4^+ reveals how the data convergence is not only dependent on the short or long character of the interaction but also on the presence of the point charges embedded in the MM region. As much as 100 QM water molecules are deemed necessary for a sufficiently accurate description of the interaction energies for a solute in water solution.
Following the converge study, the EDA is performed for a sample of 100 MD trajectory snapshots, whose statistical treatment reveals the importance of the configurational sampling when computing interaction energies as a significant STD is found for the obtained data. The discussion of the nucleobases behaviour in water is aided by the computation of the DIs, this quantum-mechanical tool allows to target the relation between the electron delocalisation of the nucleobases in a solvated environment and the dispersion component of the interaction energy.

A general conclusion that could be extracted from this work is the that the use of the EDA benefits from a QM/MM approach, as the polarisation exerted by the MM environment leads to a better description of the interaction energy, speeding up its convergence with QM region size and thus reducing the cost derived from its computation. Nonetheless, the MM force field here employed is not polarisable, which leads to unreasonable trends for the ammonium cation in water solution. A significant improvement could be achieved with the use of a polarisable force field and a quantum-mechanical treatment of the counter ions.
The good performance of the EDA for the nucleobases in water solution proves the validity of this method and opens the door for application in complex systems, particularly those of biological interest as the QM/MM-EDA scheme here exposed has the potential to overcome the limitations that some of the most widely used solvation continuum models still suffer today.

Bibliography

[1] H. M. Senn and W. Thiel. QM/MM Methods for Biomolecular Systems. *Angewandte Chemie International Edition*, 48(7):1198 – 1229, 2009. 1

[2] M. J. S. Phipps, T. Fox, C. S. Tautermannb and C. K. Skylaris. Energy decomposition analysis approaches and their evaluation on prototypical protein–drug interaction patterns. *Chemical Society Reviews*, 98(10):3177–3211, 2015. 1

[3] H. Hirao. The Effects of Protein Environment and Dispersion on the Formation of Ferric-Superoxide Species in myo-Inositol Oxygenase (MIOX): A Combined ONIOM(DFT:MM) and Energy Decomposition Analysis. *The Journal of Physical Chemistry*, 115(38):11278–11285, 2011. 1

[4] J. Tomasi, B. Mennucci and R. Cammi. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews*, 105(8):2999–3093, 2005. 1, 2

[5] R. Cammi, F. J. Olivares del Valle and J. Tomasi. Decomposition of the interaction energy with counterpoise corrections to the basis set superposition error for dimers in solution. Method and application to the hydrogen fluoride dimer. *Chemical Physics*, 112(1):63–74, 1988. 2

[6] S. F. Boys and F. Berbardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics*, 19(4):553–566, 1970. 2, 11

[7] S. Miertus, E. Scrocco and J. Tomasi. Electrostatic interaction of a solute with a continuum. A direct utilizaion of AB initio molecular potentials for the prevision of solvent effects. *Chemical Physics*, 55(1):117–129, 1981. 2

[8] S. Miertus and J .Tomasi. Approximate evaluations of the electrostatic free energy and internal energy changes in solution processes. *Chemical Physics*, 65(2):234–245, 1982. 2

[9] R. W. Gora, W. Bartkowiak, S. Roszak and J. Leszczynski. Intermolecular interactions in solution: Elucidating the influence of the solvent. *Journal of Chemical Physics*, 120(6):2802–2813, 2004. 2

[10] R. Cammi, F. J. Olivares del Valle and J. Tomasi. Free energy decomposition analysis of bonding and nonbonding interactions in solution. *The Journal of Chemical Physics*, 137:034111, 2012. 2

[11] C. J. Cramer and D. G. Truhlar. Free energy decomposition analysis of bonding and nonbonding interactions in solution. *Accounts of Chemichal Research*, 41(6):760–768, 2008. 2

[12] Y. Mao, Y. Shao, J. Dziedzic, C. K. Skylaris, T. Head-Gordon and M. Head-Gordon. Performance of the AMOEBA Water Model in the Vicinity of QM Solutes: A Diagnosis Using Energy Decomposition Analysis. *Journal of Chemical Theory and Computation*, 13(5):19631979, 2017. 2

[13] J. Gao. Hybrid Quantum and Molecular Mechanical Simulations: An Alternative Avenue to Solvent Effects in Organic Chemistry. *Accounts of Chemichal Research*, 29(6):298–305, 1996. 2

[14] E. Shaw, C. J. Woods and A.J. Mulholland. Compatibility of Quantum Chemical Methods and Empirical (MM) Water Models in Quantum Mechanics/ Molecular Mechanics Liquid Water Simulations. *The Journal of Physical Chemistry Letters*, 1(1):219–223, 2010. 2

[15] M. S. Gordon, M. A. Freitag and P. Bandyopadhyay. The Effective Fragment Potential Method: A QM-Based MM Approach to Modeling Environmental Effects in Chemistry. *The Journal of Physical Chemistry A*, 105(2):219–223, 2001. 2

[16] F. Lipparini, C. Cappelli and V. Barone. Linear Response Theory and Electronic Transition Energies for a Fully Polarizable QM/Classical Hamiltonian. *Journal of Chemical Theory and Computation*, 8(11):4153–4165, 2012. 2

[17] E. Boulanger and W. Thiel. Toward QM/MM Simulation of Enzymatic Reactions with the Drude Oscillator Polarizable Force Field. *Journal of Chemical Theory and Computation*, 10(4): 1795–1809, 2014. 2

[18] Y. Mao, O. Demerdash, M. Head-Gordon and T. Head-Gordon. Assessing Ion–Water Interactions in the AMOEBA Force Field Using Energy Decomposition Analysis of Electronic Structure Calculations. *Journal of Chemical Theory and Computation*, 12(11):5422–5437, 2016. 2

[19] J. W. Ponder, C. Wu, P. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, R. A. DiStasio Jr., M. Head-Gordon, G. N. I. Clark, M. E. Johnson and T. Head-Gordon. Current Status of the AMOEBA Polarizable Force Field. *The Journal of Physical Chemistry B*, 114(8):2549–2564, 2010. 2

[20] L. Verlet. Computer ”Experiments” on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, 1967. 4

[21] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case. Development and Testing of a General Amber Force Field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004. 5, 20

[22] E. Braun, J. Gilmer, H. B. Mayes, D. L. Mobley, J. I. Monroe, S. Prasad and D. M. Zuckerman. *Best Practices for Foundations in Molecular Simulations [Article V1.0] v1.0*, 2018. 5

[23] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, first edition, 1991. 5

[24] H. J. Berendsen, J. V. Postma, W. F. van Gunsteren, A. DiNola and J. Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics.*, 81(8):3684–3690, 1984. 6

[25] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864, 1964. 7

[26] P. O. Widmark and B. O. Roos. *European Summerschool in Quantum Chemistry. Book II.* Lund University, fourth edition, 2005. 7, 8

[27] K. I. Ramachandran, G. Deepa and K. Namboori. *Computational Chemistry and Computational Modelling. Principles and Applications.* Springer, first edition, 2005. 8

[28] Y. Zhao and D. G. Truhlar. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theoretical Chemistry Accounts*, 120(120):215–241, 2008. 10

[29] N. Ramos-Berdullas, I. Pérez-Juste, C. Van Alsenoy and M. Mandado. Theoretical study of the adsorption of aromatic units on carbon allotropes including explicit (empirical) DFT dispersion corrections and implicitly dispersion-corrected functionals: the pyridine case. *Physical Chemistry Chemical Physics*, 17(1):575–587, 2015. 10, 17

[30] S. F. Boys and F. Bernardi. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Molecular Physics*, 19(4):553–566, 1970. 11

[31] E. Schrödinger. Quantisierung als Eigenwertproblem (Quantization as an eigenvalue problem). *Annalen der Physik*, 80(13):437–490, 1926. 11

[32] A. J. Stone. *The Theory of Intermolecular Forces*. Oxford University Press, first edition, 1996. 14

[33] B. Jeziorski, R. Moszynski and K. Szalewicz. Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes. *Chemical Reviews.*, 94(7):1887–1930, 1994. 16

[34] Q. Wu, P. W. Ayers and Y. Zhang. Density-based energy decomposition analysis for intermolecular interactions with variationally determined intermediate state energies. *The Journal of Chemical Physics*, 131(16):164112, 2009. 17

[35] M. Mandado and J. M. Hermida-Ramón. Electron Density Based Partitioning Scheme of Interaction Energies. *Journal of Chemical Theory and Computation*, 7(3):633–641, 2011. 17

[36] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2): 926–935, 1983. 20

[37] N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao and D. J. Fox. *Gaussian09*. Gaussian, Inc., 2016. 20

[38] T. Darden, D. York and L. Pedersen. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, 98(12):10089–10092, 1993. 20

[39] A. Kovalenko, T. Kurtzman, T. S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D. J. Mermelstein, K. M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D. R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C. L. Simmerling, J. Smith, R. Salomon-Ferrer, J. Swails, R. C. Walker, J. Wang, H. Wei, R. M. Wolf, X. Wu, L. Xiao, D. M. York D. A. Case, I. Y. Ben-Shalom, S. R. Brozell, D. S. Cerutti, T. E. Cheatham III, V. W. D. Cruzeiro, T. A. Darden, R. E. Duke, D. Ghoreishi, M. K. Gilson, H. Gohlke, A. W. Goetz, D. Greene, R. Harris, N. Homeyer, Y. Huang, S. Izadi, and P. A. Kollman. *AMBER 2018*. University of California, San Francisco, 2018. 20

[40] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian16 Revision C.01, 2016. Gaussian Inc. Wallingford CT. 20

[41] D. R. Roe and T. E. Cheatham. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, 9(7):3084–3095, 2013. 20

[42] P. P. Ewald. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Chemical Reviews*, 369(3):253–287, 1921. 26

[43] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933. 27

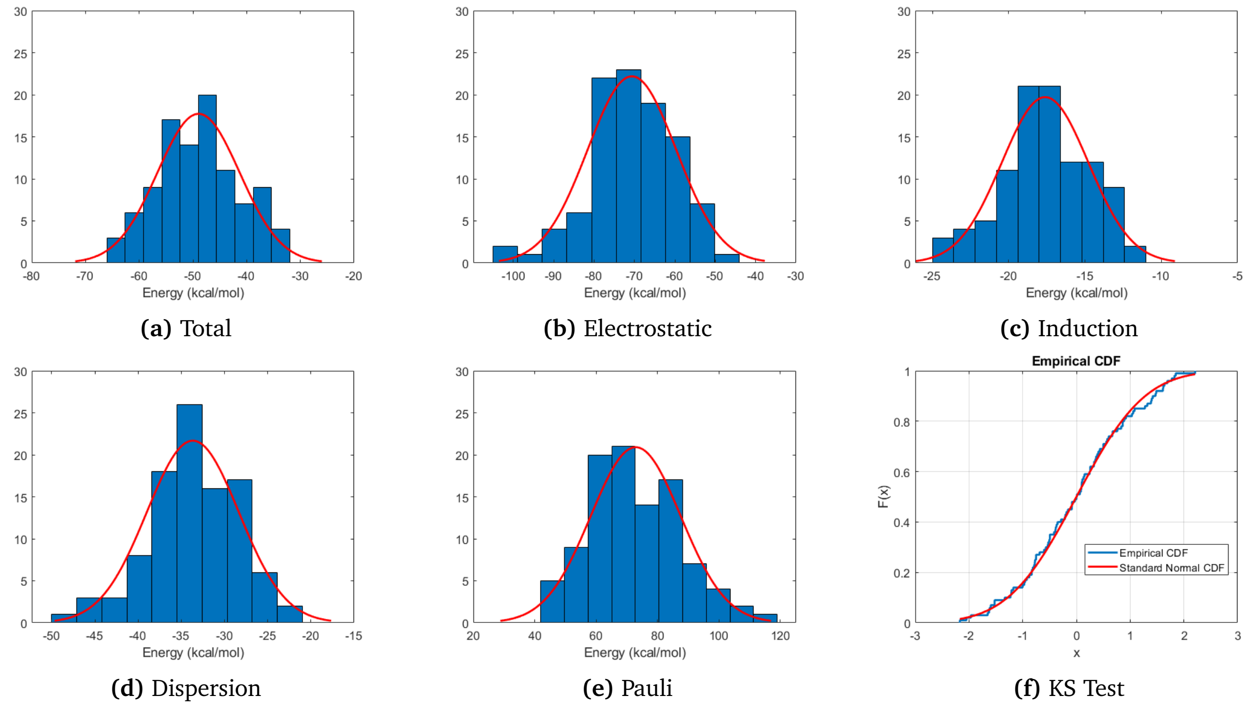
[44] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19(2):279–281, 1948. 27

[45] P. Cysewski. An ab initio study on nucleic acid bases aromaticities. *Journal of Molecular Structure*, 714(1):29–34, 2005. 30

[46] O. Huertas, J. Poater, M. Fuentes-Cabrera, M. Orozco, M. Sola and F. J. Luque. Local Aromaticity in Natural Nucleobases and Their Size-Expanded Benzo-Fused Derivatives. *Journal of Physical Chemistry A*, 110(44):12249–12258, 2006. 30

[47] M. Mandado, M. J. González-Moa and R. A. Mosquera. QTAIM n-center delocalization indices as descriptors of aromaticity in mono and poly heterocycles. *Journal of Computational Chemistry*, 28 (1):127–136, 2007. 30

Annex



(a) Total

(b) Electrostatic

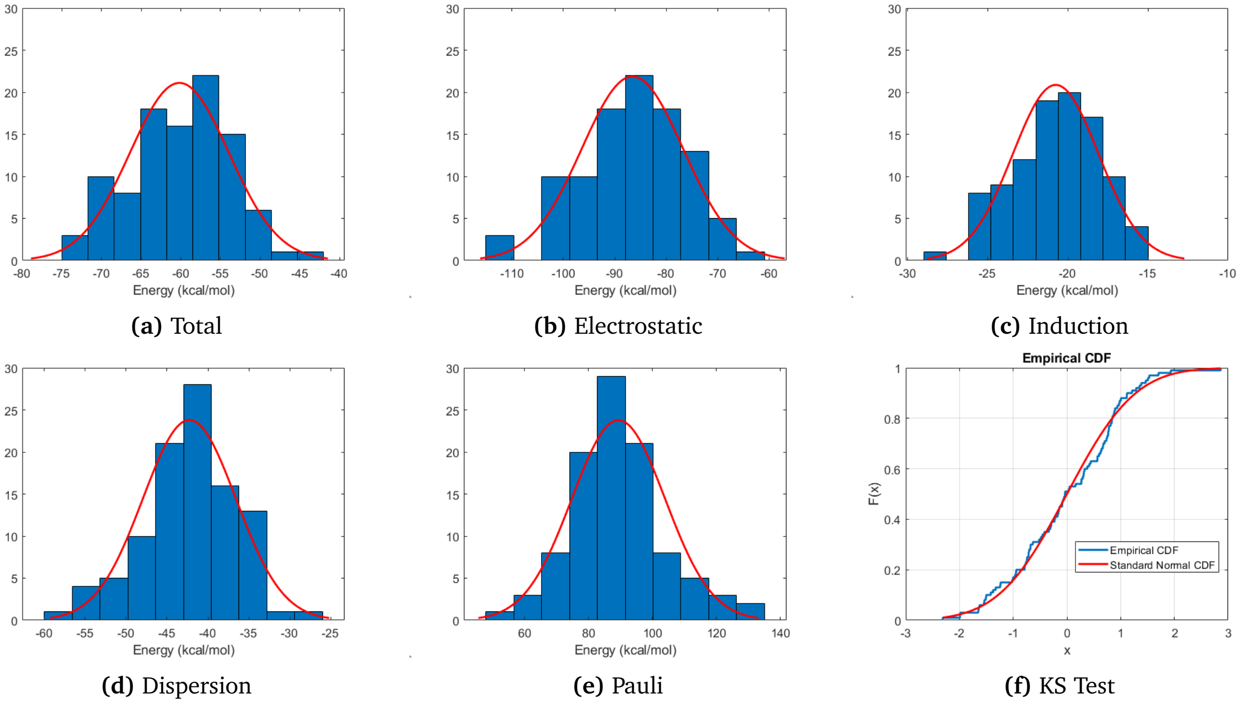
(c) Induction

(d) Dispersion

(e) Pauli

(f) KS Test

Figure 5.1: Histogram for each interaction energy component with energy in *kcal/mol* in the horizontal axis and relative data frequency in the vertical axis with a fitted normal distribution function (red) for the cytosine molecule.(e) ECDF and normal CDF used in the KS test.



(a) Total

(b) Electrostatic

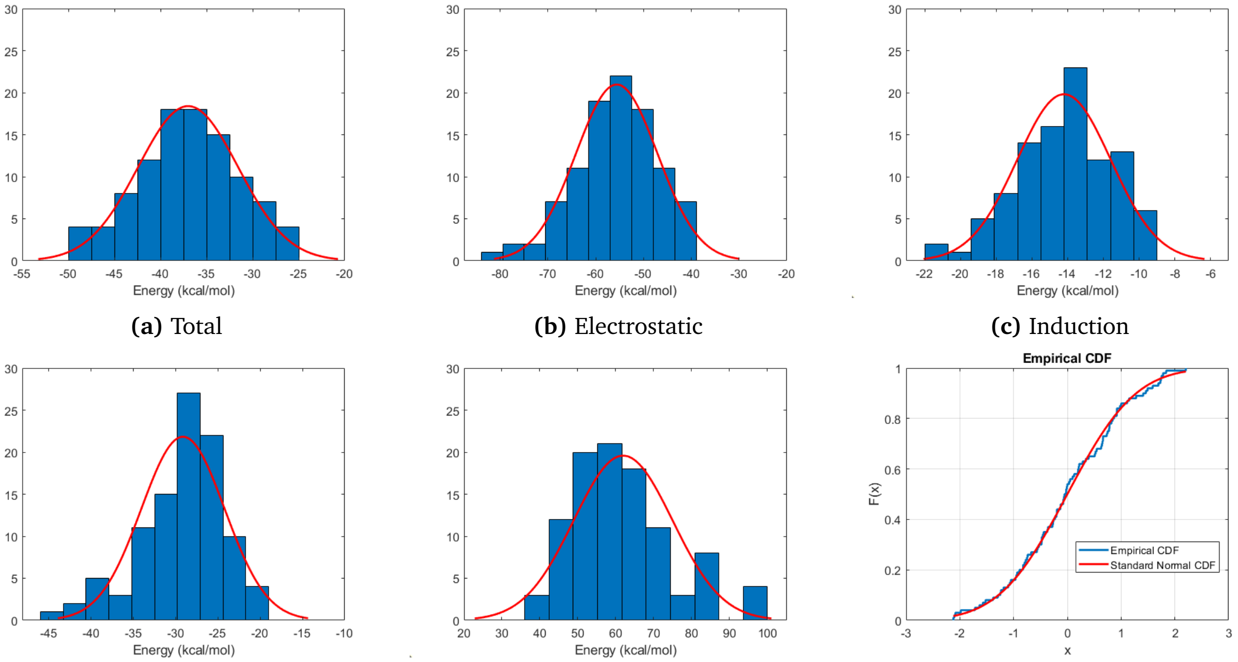
(c) Induction

(d) Dispersion

(e) Pauli

(f) KS Test

Figure 5.2: Histogram for each interaction energy component with energy in *kcal/mol* in the horizontal axis and relative data frequency in the vertical axis with a fitted normal distribution function (red) for the guanine molecule.(e) ECDF and normal CDF used in the KS test.



(a) Total

(b) Electrostatic

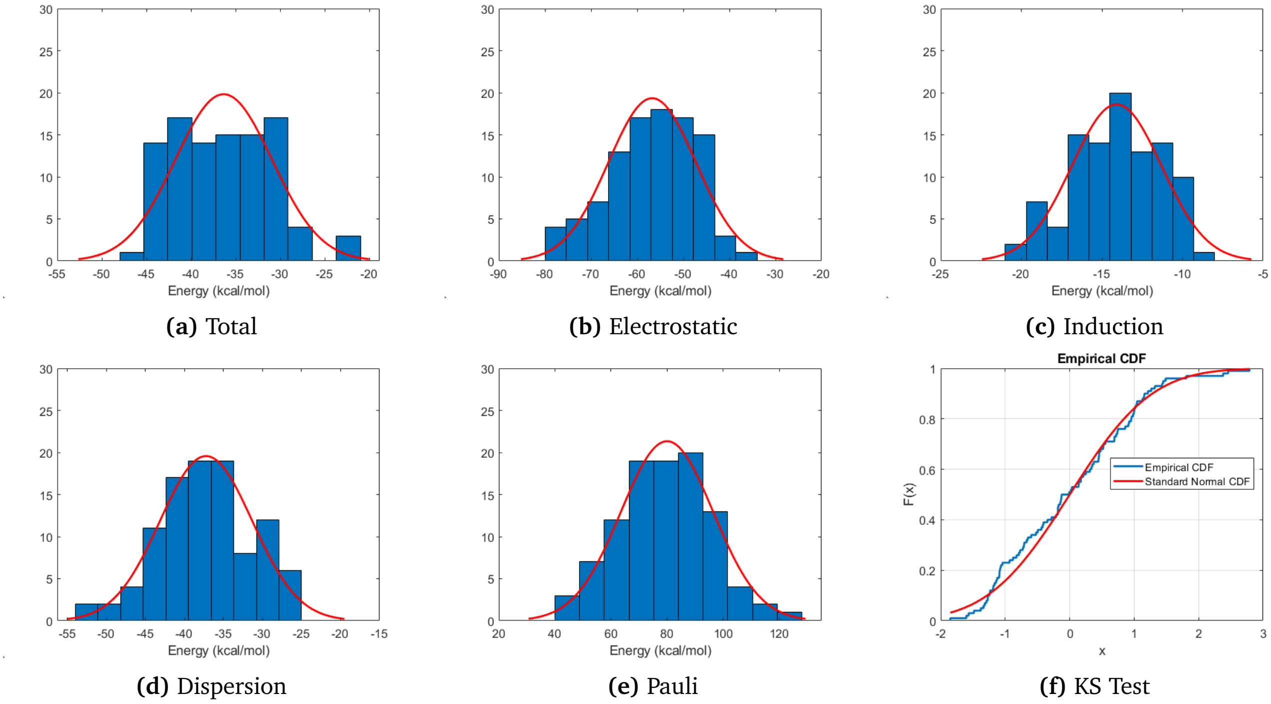
(c) Induction

(d) Dispersion

(e) Pauli

(f) KS Test

Figure 5.3: Histogram for each interaction energy component with energy in *kcal/mol* in the horizontal axis and relative data frequency in the vertical axis with a fitted normal distribution function (red) for the thymine molecule.(e) ECDF and normal CDF used in the KS test.



(a) Total

(b) Electrostatic

(c) Induction

(d) Dispersion

(e) Pauli

(f) KS Test

Figure 5.4: Histogram for each interaction energy component with energy in *kcal/mol* in the horizontal axis and relative data frequency in the vertical axis with a fitted normal distribution function (red) for the uracil molecule.(e) ECDF and normal CDF used in the KS test.

Table 5.1: Sample mean ($\bar{x}$) and standard deviation (s) in *kcal/mol*. Population mean (μ) and standard deviation (σ) also in *kcal/mol* for a confidence interval of 95% for each interaction energy component for the cytosine, guanine, thymine and uracil molecules. The percentage contribution of each energy component to the attractive energy according to the sample mean is shown in the left column.

		Cytosine			
		$\bar{x}$	μ	s	σ
Electrostatic	57.0%	-70.6258	[-72.8019, -68.4497]	10.9669	[9.62902, 12.7400]
Induction	14.4%	-17.612	[-18.1741, -17.0499]	2.83294	[2.48734, 3.29096]
Dispersion	28.6%	-33.6711	[-34.7296, -32.6125]	5.33500	[4.68416, 6.19753]
Pauli		72.9155	[70.0002, 75.8307]	14.6922	[12.8999, 17.0676]
Total		-48.9534	[-50.4712, -47.4357]	7.64915	[6.71601, 8.88583]
		Guanine			
		$\bar{x}$	μ	s	σ
Electrostatic	57.9%	-86.5936	[-88.5481, -84.6392]	9.84992	[8.64830, 11.4424]
Induction	13.9%	-20.7593	[-21.2904, -20.2283]	2.67612	[2.34965, 3.10878]
Dispersion	28.2%	-42.2701	[-43.4010, -41.1393]	5.69912	[5.00387, 6.62053]
Pauli		89.3883	[86.4919, 92.2846]	14.5968	[12.8160, 16.9567]
Total		-60.186	[-61.4235, -58.9485]	6.23687	[5.47601, 7.24521]
		Thymine			
		$\bar{x}$	μ	s	σ
Electrostatic	56.2%	-55.6232	[-57.3219, -53.9244]	8.56119	[7.51678, 9.94532]
Induction	14.4%	-14.2076	[-14.7271, -13.6881]	2.61825	[2.29884, 3.04156]
Dispersion	29.4%	-29.1435	[-30.1223, -28.1646]	4.93307	[4.33127, 5.73063]
Pauli		61.9451	[59.3586, 64.5317]	13.0357	[11.4454, 15.1432]
Total		-36.999	[-38.0745, -35.9236]	5.41991	[4.75872, 6.29618]
		Uracil			
		$\bar{x}$	μ	s	σ
Electrostatic	57.5%	-56.7488	[-58.6304, -54.8673]	9.48249	[8.32569, 11.0156]
Induction	14.3%	-14.0889	[-14.6415, -13.5362]	2.78517	[2.44540, 3.23547]
Dispersion	28.2%	-27.7829	[-28.7327, -26.8332]	4.78668	[4.20273, 5.56056]
Pauli		60.1718	[57.6229, 62.7206]	12.8456	[11.2786, 14.9225]
Total		-38.4177	[-39.6793, -37.1561]	6.35821	[5.58255, 7.38618]

Table 5.2: Sample mean ($\bar{x}$) and standard deviation (s) in *kcal/mol*. Population mean (μ) and standard deviation (σ) also in *kcal/mol* for a confidence interval of 95% for each interaction energy component for the cytosine, guanine, thymine and uracil molecules with increasing sample size.

Cytosine

Sample size	$\bar{x}$	μ	s	σ
10	-48.3456	[-52.2315, -44.4597]	5.43216	[3.73643, 9.91700]
20	-48.7224	[-51.6069, -45.8379]	6.16330	[4.68713, 9.00194]
30	-48.8443	[-51.5230, -46.1655]	7.17388	[5.71333, 9.64396]
40	-48.6705	[-50.9805, -46.3605]	7.22285	[5.91668, 9.27440]
50	-48.9593	[-50.9358, -46.9829]	6.95460	[5.80942, 8.66636]
75	-48.3283	[-49.9830, -46.6735]	7.19208	[6.19684, 8.57117]
100	-48.9534	[-50.4712, -47.4357]	7.64915	[6.71601, 8.88583]

Guanine

Sample size	$\bar{x}$	μ	s	σ
10	-60.3503	[-64.2996, -56.4010]	5.52080	[3.79740, 10.0788]
20	-60.9629	[-63.5573, -58.3685]	5.54346	[4.21575, 8.09662]
30	-61.0160	[-62.8485, -59.1836]	4.90742	[3.90830, 6.59712]
40	-60.7887	[-62.5737, -59.0037]	5.58129	[4.57197, 7.16657]
50	-60.8726	[-62.4784, -59.2668]	5.65025	[4.71984, 7.04096]
75	-60.6715	[-62.1328, -59.2102]	6.35118	[5.47230, 7.56902]
100	-60.1860	[-61.4235, -58.9485]	6.23687	[5.47601, 7.24521]

Thymine

Sample size	$\bar{x}$	μ	s	σ
10	-36.3844	[-41.7678, -31.0010]	7.52550	[5.17631, 13.7386]
20	-37.3375	[-40.3397, -34.3352]	6.41487	[4.87845, 9.36937]
30	-36.9739	[-39.1399, -34.8079]	5.80053	[4.61958, 7.79774]
40	-36.6432	[-38.4877, -34.7986]	5.76764	[4.72463, 7.40586]
50	-36.3869	[-37.9834, -34.7904]	5.61758	[4.69256, 7.00025]
75	-36.5225	[-37.8019, -35.2431]	5.56059	[4.79111, 6.62683]
100	-36.9990	[-38.0745, -35.9236]	5.41991	[4.75872, 6.29618]

Uracil

Sample size	$\bar{x}$	μ	s	σ
10	-43.0368	[-48.5325, -37.5411]	7.68249	[5.28429, 14.0252]
20	-39.9470	[-43.7630, -36.1311]	8.15359	[6.20073, 11.9089]
30	-39.5185	[-42.3319, -36.7051]	7.53434	[6.00040, 10.1285]
40	-39.2756	[-41.5166, -37.0346]	7.00724	[5.74006, 8.99755]
50	-39.3768	[-41.2498, -37.5038]	6.59053	[5.50529, 8.21268]
75	-38.5395	[-39.9910, -37.0880]	6.30859	[5.43561, 7.51827]
100	-38.4177	[-39.6793, -37.1561]	6.35821	[5.58255, 7.38618]