Coupling of quantum chemistry models and high-performance algorithms for the global exploration of the energy landscape of atomic and molecular systems

Valentin Milia

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

Université de Toulouse

Doctorat de l'Université de Toulouse

préparé à Toulouse INP

Couplage de modèles de chimie quantique et d'algorithmes haute performance pour l'exploration globale du paysage énergétique de systèmes atomiques et moléculaires

Thèse présentée et soutenue, le 27 septembre 2024 par

Valentin MILIA

École doctorale
EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité
Informatique et Télécommunications

Unité de recherche
LAAS - Laboratoire d'Analyse et d'Architecture des Systèmes

Thèse dirigée par
Juan CORTES et Mathias RAPACIOLI

Composition du jury
M. Tony LELIÈVRE, Rapporteur, Ecole des Ponts ParisTech
M. Philippe CARBONNIERE, Rapporteur, Université de Pau et des Pays de l'Adour
Mme Corinne LACAZE-DUFAURE, Examinatrice, Université Toulouse III - Paul Sabatier
Mme Maud JUSOT, Examinatrice, Iktos
M. Juan CORTES, Directeur de thèse, CNRS Occitanie Ouest
M. Mathias RAPACIOLI, Co-directeur de thèse, CNRS Occitanie Ouest

Membres invités
Mme Nathalie Tarrat, CNRS Occitanie Ouest

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Institut National Polytechnique de Toulouse (INP Toulouse)*

Présentée et soutenue le *27/09/2024* **par :**

Valentin MILIA

Couplage de modèles de chimie quantique et d'algorithmes haute performance pour l'exploration globale du paysage énergétique de systèmes atomiques et moléculaires

---

JURY

Philippe Carbonniere	Professeur des universités	Rapporteur
Tony Lelièvre	Professeur des universités	Rapporteur
Juan Cortés	Directeur de Recherche	Directeur de Thèse
Mathias Rapacioli	Chargé de Recherche	Codirecteur de Thèse
Corinne Lacaze-Dufaure	Professeur des universités	Examinatrice
Maud Jusot	Cadre scientifique	Examinatrice
Nathalie Tarrat	Chargé de Recherche	Membre invité

École doctorale et spécialité :
MITT : Informatique et Télécommunications

Unité de Recherche :
Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS) et Laboratoire de Chimie et Physique Quantiques (LCPQ)

Directeur(s) de Thèse :
Juan CORTÉS et *Mathias RAPACIOLI*

Rapporteurs :
Philippe CARBONNIERE et *Tony LELIÉVRE*

Remerciements

Ce travail a été soutenu par le Centre National de la Recherche Scientifique (CNRS) dans le cadre de la subvention QUARTET 80|Prime. Ce travail a bénéficié de l'accès aux ressources HPC du centre de calcul CALMIP sous les allocations p19055 et p0059.

Contents

Acronyms

AG Algorithmes Génétiques.

ART Activation Relaxation Technique.

BBP Benzyl Butyl Phthalate.

BH Basin Hopping.

CG Coarse Grained.

CI-NEB Climbing Image Nudged Elastic Band.

CPR Conjugated Peak Refinement.

DBP Dibutyl Phthalate.

DEHP Di(2-ethylhexyl) phthalate.

DFT Density Functional Theory.

DFTB Density Functional Tight Binding.

DM Dynamique Moléculaire.

DTW Dynamic Time Warping.

EB Elastic Band.

EW-CI-NEB Energy Weighted Climbing Image Nudged Elastic Band.

FF Force Field.

GA Genetic Algorithms.

HCA Hierarchical Clustering Analysis.

HF Hartree-Fock.

HOMO Highest Occupied Molecular Orbital.

IGLOO Iterative Global Exploration and Local Optimization.

IR Infrared.

ISM Interstellar Medium.

LUMO Lowest Unoccupied Molecular Orbital.

MC Monte Carlo.

MD Molecular Dynamics.

MEP Minimum Energy Path.

MMC Metropolis Monte Carlo.

MoMA Molecular Motion Algorithms.

MSD Mean Square Deviation.

NEB Nudged Elastic Band.

OWD One Way Distance.

PAH Polycyclic Aromatic Hydrocarbon.

PCR Probabilistic Conformational Roadmap.

PES Potential Energy Surface.

PRM Probabilistic Roadmap.

PT Parallel Tempering.

RMSD Root Mean Square Deviation.

RRT Rapidly-exploring Random Tree.

RS Recuit Simulé.

SA Simulated Annealing.

SEP Surface d'Énergie Potentielle.

SMILES Simplified Molecular Input Line Entry Specification.

SSPD Symmetrized Segment-Path Distance.

TPS Transition Path Sampling.

T-RRT Transition-based Rapidly-exploring Random Tree.

US Umbrella Sampling.

VdW Van der Waals.

Chapter 1

Introduction

Contents

1.1 Introduction

A significant challenge in the field of atomic and molecular systems is to gain a deeper understanding of their fundamental properties. This challenge is particularly pronounced as the systems under study become increasingly complex and the necessity for efficient exploration of their energy landscape to predict their behavior in diverse physical and biological environments increases. This chapter will present the main concepts developed in this thesis in order to address the problem at hand and to contextualize the existing state-of-the-art methods. A theoretical framework will be developed to introduce the concept of electronic structures and define Schrödinger equations and the Born-Oppenheimer approximation. Subsequently, potential energy methods will be presented, with each method addressing different problems and varying mainly in terms of their accuracy and efficiency. Finally, global exploration methods will be introduced to illustrate the diversity of techniques available for investigating the potential energy surface (PES).

1.2 Atomistic/molecular modeling and electronic structure

Atoms are composed of a nucleus containing protons and neutrons, surrounded by electrons. The arrangement of electrons in the orbitals determines the chemical properties of the atom and its interactions with other atoms. The properties of atoms and molecules are governed by the laws of quantum mechanics, which describe the behavior of particles at the atomic and subatomic levels. The principles of quantum mechanics provide a framework for understanding the structure of atoms and molecules, the nature of chemical bonds, and the interactions between molecules.

Molecules are composed of two or more atoms held together by chemical bonds. The specific arrangement of the atoms within a molecule determine its shape and properties. Molecules exhibit a wide range of properties and behaviors, which depend on their composition, structure, and interactions. An understanding of the characteristics of molecules is essential for the prediction of their behavior and properties in various chemical processes.

1.2.1 Schrödinger equation

In 1926, Erwin Schrödinger, an Austrian physicist, introduced a wavefunction that describes how the quantum state of a physical system changes over time [139]. One of the pivotal achievements in the field of Quantum Chemistry is the formulation of the Schrödinger equation. This equation is one of the most important postulates of quantum mechanic and has played a crucial role in our understanding of the subatomic world.

The time-dependent Schrödinger equation is written as:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{R},\mathbf{r},t) = \hat{H}\Psi(\mathbf{R},\mathbf{r},t) \tag{1.1}$$

where $\hbar$ is the reduced Planck's constant, $\hbar = \frac{h}{2\pi}$, h is Planck's constant equal to $6.62607015.10^{-34}$ J.s., i is the imaginary unit, $\Psi(\mathbf{R},\mathbf{r},t)$ is the wavefunction of the system, which contains information about the position of the nuclei $\mathbf{R}$, the electrons $\mathbf{r}$ and time t, and $\hat{H}$ is the Hamiltonian operator, associated to the total energy of the system. For the particulary case of systems in a stationary state, meaning those where properties do not vary over time, temporal and spatial variables are separated. The state wave function (eigenfunction) is defined as:

$$\Psi(\mathbf{R},\mathbf{r},t) = e^{-i\frac{Et}{\hbar}}\psi(\mathbf{R},\mathbf{r}) \tag{1.2}$$

$\psi(\mathbf{R},\mathbf{r})$ represent the spatial contribution to the wavefunction and can be obtained by solving the time-independent Schrödinger equation. This equation is written as:

$$\hat{H}\psi(\mathbf{R},\mathbf{r}) = E\psi(\mathbf{R},\mathbf{r}) \tag{1.3}$$

The Hamiltonian operator plays a central role in the equation, dictating the dynamics

of the system by defining the energy landscape within which the system evolves.

In this form, the Hamiltonian operator applied to the wave function $\psi(\mathbf{r})$ equal to the energy E of the system multiplied by the wave function. The Hamiltonian operator is the sum of the kinetic operator and potential energy terms, and is defined as

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{en} + \hat{V}_{nn} \tag{1.4}$$

where e and n refer to the electronic and nuclear components, respectively, and the subscripts indicate the type of interaction (electron-electron, electron-nucleus, and nucleus-nucleus). The kinetic energy terms $\hat{T}_e$ and $\hat{T}_n$ are the operators associated to the kinetic energies for electrons and nuclei. The potential energy terms $\hat{V}_{ee}$ is the repulsive Coulomb interaction between electrons, $\hat{V}_{en}$ is the attractive nucleus-electron Coulomb interaction and $\hat{V}_{nn}$ the repulsive Coulomb interaction between nuclei. Defining a system of particles with N electrons and M nuclei, the five terms of the Hamiltonian (Eq. A.4) in atomic units ($\hbar = m_e = e = c = 1$) can be written as:

$$\hat{T}_e = -\frac{1}{2}\sum_{i=1}^{N} \nabla_i^2 \tag{1.5}$$

$$\hat{T}_n = -\frac{1}{2}\sum_{A=1}^{M} \frac{1}{M_A}\nabla_A^2 \tag{1.6}$$

$$\hat{V}_{en} = -\sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} \tag{1.7}$$

$$\hat{V}_{ee} = \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{1.8}$$

$$\hat{V}_{nn} = \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} \tag{1.9}$$

where r_{iA} is the distance between the i-th electron and the A-th nucleus, r_{ij} is the distance between the i-th and j-th electrons, R_{AB} is the distance between the A-th and B-th nuclei, Z_A is the atomic number of the A-th nucleus and M_A is the mass of the A-th nucleus. Finally the Laplacian operator ∇^2 is defined in cartesian coordinates as $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$.
In pratice, solving the Schrödinger equation analytically for systems with more than one electron is impossible. Therefore, numerical methods combined with approximations are used to solve this equation.

1.2.2 Born-Oppenheimer approximation

In 1927, Max Born and Robert Oppenheimer introduced the Born-Oppenheimer approximation, which simplifies the Schrödinger equation by treating the electronic and nuclear motions as independent variables [28]. The nuclear mass is much larger than

the electronic mass ($m_p \approx 1836 m_e$), making the electronic motion much faster than the nuclear motion, which leads to a separation of variables to describe electronic and nuclear motions. As a result, the electronic wavefunction can adjust instantaneously to changes in the nuclear positions, and electrons move in a potential field generated by the nuclei. The total wavefunction can be written as a product of electronic and nuclear wavefunctions:

$$\Psi(\mathbf{r}, \mathbf{R}) = \psi_n(\mathbf{R})\psi_e(\mathbf{r}; \mathbf{R}) \tag{1.10}$$

where $\psi_n(\mathbf{R})$ is the nuclear wavefunction, and $\psi_e(\mathbf{r}; \mathbf{R})$ is the electronic wavefunction that depends parametrically on the nuclear coordinates. The electronic problem can be solve independently from the nuclear motion. At a given position of the nuclei, the electronic problem can be obtain by solving a time-independent Schrödinger equation for the electrons only:

$$\hat{H}_e\psi_e(\mathbf{r}; \mathbf{R}) = E_e(\mathbf{R})\psi_e(\mathbf{r}; \mathbf{R}) \tag{1.11}$$

where E_e is the electronic energy and the electronic Hamiltonian operator is given as:

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{en} \tag{1.12}$$

Note that solving the equation A.11 leads to several solutions corresponding to different electronic states. In many cases and in particular in the following of this thesis, only the solution corresponding to the lowest energy eigenvalue is considered, also called the electronic ground state. The dynamics of the nuclei is governed by the potential energy obtained by adding the nuclear repulsion to the electronic energy:

$$E(\mathbf{R}) = E_e(\mathbf{R}) + \hat{V}_{nn} \tag{1.13}$$

According to the equation A.13, the potential energy of the system can be calculated for a given nuclear configuration. This approximation simplifies the Schrödinger equation by reducing the number of variables and allows the definition of a potential energy surface that describes the energy landscape of the chemical system. In addition to the former approximation, the nuclei are often considered as classical particles (punctual particles). They can therefore be treated from the Newton's law making use of the PES define in equation A.13.

1.3 Potential energy

To compute the potential energy of a system, several methods have been developed in the field of computational chemistry, varying in complexity and accuracy as presented in the figure A.1. The choice of the method size and computational cost are correlated, so mention only cost the level of accuracy required and the computational cost. The following section will present the principal methods utilized for the calculation of a system's potential energy, including those based on wave functions, as well as Density Functional Theory (DFT), Density Functional Tight Binding (DFTB), Force Field (FF), and Coarse-graining (CG).

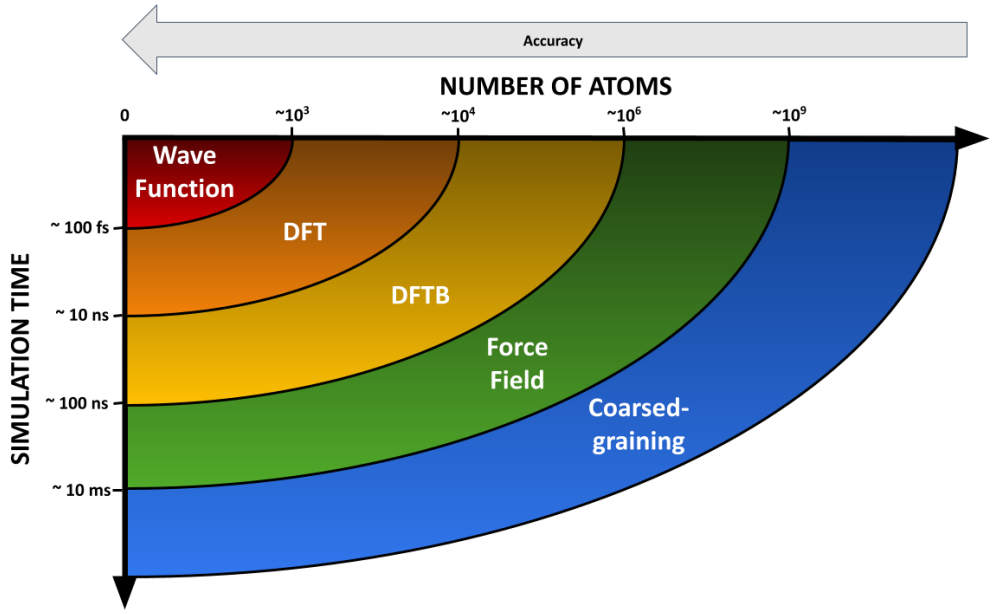


Figure 1.1: Potential energy scale.

1.3.1 Wave function based methods

The Hartree-Fock (HF) method [13, 81, 146, 61, 60] is a mean-field approach wherein electrons are presumed to evolve independently within an effective potential shaped by both nuclei and the ensemble of electrons. The method approximates the system's total multi-electron wave function as a product of single-electron wave functions. The computational scheme involves solving a set of N single-particle Schrödinger equations self-consistently.

Nevertheless, Fock identified a significant deficiency in Hartree's formulation—its noncompliance with the Pauli exclusion principle. This resulted in a wave function that was not antisymmetric with respect to particle exchange. To rectify this, Fock reformulated the wave function as a Slater determinant of single-electron functions, incorporating the fermionic characteristics of electrons and introducing the exchange energy term in the Hamiltonian. This resulted in the evolution of the original method into what is now widely recognized as the Hartree-Fock method.

Despite its significant contributions, the Hartree-Fock method does not account for correlation energy, the discrepancy between the exact quantum mechanical energy and the energy estimated by Hartree-Fock calculations.

1.3.2 Density Functional Theory (DFT)

Density Functional Theory (DFT) [144, 153] was introduced by Hohenberg and Kohn in 1964 and further developed by Kohn and Sham in 1965. DFT is a quantum mechanical theory used to investigate the electronic structure of many-body systems, primarily atoms, molecules, and condensed phases. Unlike methods that are based directly on the

wave function, DFT describes a system in terms of its electron density rather than its wave function.

The foundation of DFT is built upon two Hohenberg-Kohn (HK) theorems [86]:

1. The first HK theorem states that the ground-state properties of a many-electron system are uniquely determined by its electron density $\rho(\mathbf{r})$. This implies that all observable properties of the system are functionals of the electron density.

2. The second HK theorem provides a variational principle for the electron density. It states that the total energy functional $E[\rho]$ has its minimum value at the true ground-state electron density of the system.

Based on these theorems, Kohn and Sham developed a practical scheme known as the Kohn-Sham (KS) equations:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}), \tag{1.14}$$

where $\psi_i(\mathbf{r})$ are the Kohn-Sham orbitals, ε_i are their corresponding eigenvalues, and $V_{\text{eff}}(\mathbf{r})$ is the effective potential which includes the external potential, the Hartree potential, and the exchange-correlation potential. The effective potential is expressed as:

$$V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}[\rho(\mathbf{r})] + V_{\text{xc}}[\rho(\mathbf{r})]. \tag{1.15}$$

The exchange-correlation potential $V_{\text{xc}}[\rho(\mathbf{r})]$ is the most critical component in DFT calculations and incorporates all the many-body effects. Determining an accurate functional for V_{xc} is a major area of research within DFT.

DFT is considered more accurate and efficient than the Hartree-Fock (HF) method as it inherently includes electron correlation effects. It is widely used for calculating the electronic structure of molecules and predicting their properties.

1.3.3 Density Functional Tight Binding (DFTB)

Density Functional Tight Binding (DFTB) is a semi-empirical method that approximates the electronic structure of a system using a minimal basis set. The DFTB method was introduced by Elstner in 1998 [55]. It is based on the tight-binding approximation, which simplifies the electronic structure of a system. The DFTB method is a popular computational chemistry tool for calculating the electronic structure and predicting properties of molecules. It is particularly efficient for large systems and incorporates electron correlation effects. Chapter 2 will provide a more detailed presentation of DFTB.

1.3.4 Force Field (FF)

The Force Field (FF) method [105, 6] is a classical approach used in computational chemistry to estimate the potential energy of a system. This method, which is rooted

in classical mechanics, employs the classical equations of motion to describe how the positions and velocities of particles change over time. A FF method approximates the potential energy of a system based on the positions of its atoms. They are particularly effective for simulating large molecular systems due to their computational efficiency compared to quantum mechanical methods. This increase in efficiency is accompanied by a concomitant decrease in calculation accuracy. The potential energy U of a system in the force field method is typically expressed as a sum of contributions from bonded and non-bonded interactions:

$$U = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{non-bonded}}, \tag{1.16}$$

with each component defined as follows:

$$U_{\text{bond}} = \sum_{\text{bonds}} k_i^b (r - r_0)^2, \tag{1.17}$$

$$U_{\text{angle}} = \sum_{\text{angles}} k_i^\theta (\theta - \theta_0)^2, \tag{1.18}$$

$$U_{\text{dihedral}} = \sum_{\text{dihedrals}} k_i^\phi [1 + \cos(n\phi - \delta)], \tag{1.19}$$

$$U_{\text{non-bonded}} = \sum_{\text{non-bonded pairs}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right]. \tag{1.20}$$

Here, k_i^b, k_i^θ, and k_i^ϕ are force constants for bond lengths, bond angles, and dihedral angle respectively; r_0 and θ_0 are the equilibrium values for bond lengths and bond angles. ϕ is the dihedral angle, δ is the phase and n defines the number of minima or maxima between 0 and 2π. A_{ij} and B_{ij} are parameters for the Lennard-Jones potential describing van der Waals interactions, while q_i and q_j are the charges on atoms i and j, and r_{ij} is the distance between them. ε_0 is the vacuum permittivity.

The efficiency of FF methods enables the simulation of thousands to tens of thousands of atoms by simplifying the interactions between atoms through the use of springs. These methods are therefore indispensable for studies of large biomolecules such as proteins and nucleic acids. Furthermore, the incorporation of non-bonded interactions, such as van der Waals and electrostatic forces, enables a more precise representation of molecular dynamics and properties.

1.3.5 Coarsed-graining (CG)

Coarsed-graining (CG) is an approach used to approximate the potential energy of a system by reducing the number of degrees of freedom. The CG approach is based on the concept of effective interactions, which simplify the energy landscape of the system by grouping particles into coarse-grained beads. This method is widely used in computational chemistry to calculate the potential energy of molecules and predict their properties. CG is more efficient than atomistic methods in terms of the maximum size of system that can be simulated within a suitable timeframe for large systems. This

advantage is to put in perspective with the loss of precision.

The primary advantage of the coarse-graining method is its capacity to capture essential physical properties of a system while omitting fine details that do not significantly affect the overall behavior. By streamlining the computational model, CG methods can substantially speed up calculations, making it possible to simulate macroscopic phenomena and explore system behaviors on scales that are unattainable with conventional atomistic approaches.

Developing a coarse-grained model involves selecting the appropriate coarse-grained sites. Parameters are often derived from experimental data or high-level atomistic simulations and need to be adjusted to ensure that the CG model reproduces specific desired properties, such as phase behavior or diffusion coefficients. Once the model is developed, it is crucial to validate and refine it by comparing its predictions with experimental results or more detailed simulations, adjusting as necessary to enhance accuracy and reliability.

Coarse-graining is extensively applied in the study of biological macromolecules like proteins and nucleic acids [98, 137, 33], enabling researchers to investigate large-scale conformational changes and complex interactions over extended periods. It is also a critical tool in materials science, especially in the study of polymers and soft materials, where understanding the structure and dynamics at a large scale is vital.

Despite its numerous benefits, coarse-graining also poses certain challenges, primarily the loss of detailed atomic-level information which can be crucial for understanding specific properties such as reaction kinetics or detailed electronic attributes. Furthermore, the success of a coarse-grained model hinges on the careful balance between the details that are retained and those that are averaged out, necessitating a deep understanding of the system and the modeling techniques.

1.4 Global exploration methods

The exploration of the Potential Energy Surface is essential to access information about the most stable configurations, singular states or thermodynamics properties of a chemical system. The PES or energy hypersurface $E(\mathbf{R})$ is a representation of the potential energy of a system as a function of its geometry, which can be defined according to the atomic positions or internal coordinates. The PES, also called energy landscape, shows the stable configurations and the transition regions between different configurations. The PES can be used to determine the equilibrium geometry of a molecule, predict the properties of molecules, simulate chemical reactions and the activation energy required for chemical reactions to occur. To illustrate the concept, this surface can be represented graphically as a two-dimensional surface, where the energy is plotted as a function of the geometry of the chemical system (Fig. A.2). It should be noted that in reality, the PES is defined by the number of coordinates of the system under study.

Characteristic states can be defined on the PES such as the local minima. For such state, every first derivative of the energy with respect to geometric coordinates is equal to zero and every second derivative is positive. The global minimum is the most stable

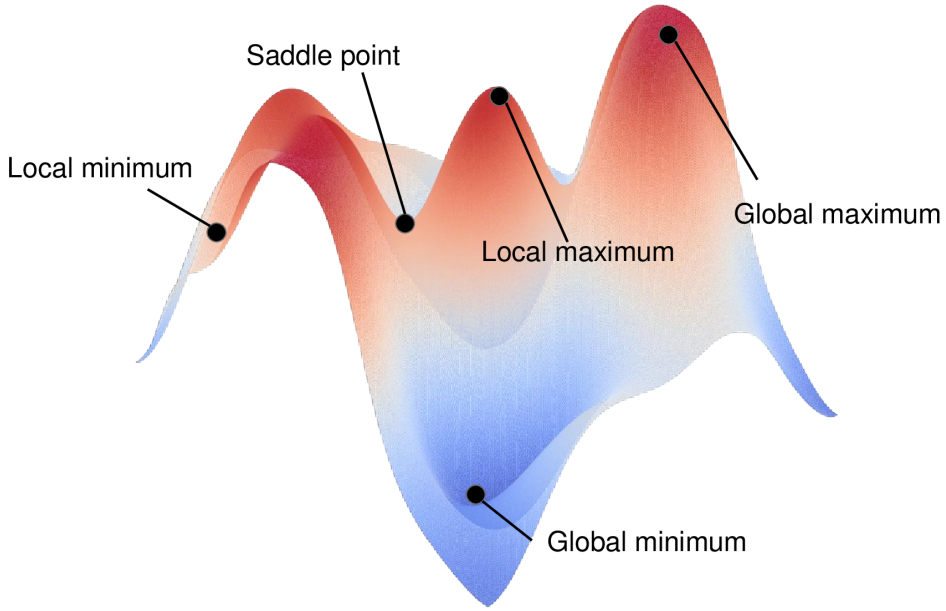


Figure 1.2: Potential Energy Surface of a molecule.

configuration of the system.

A saddle point on a PES is a critical point where the gradient (first derivative) of the energy with respect to all coordinates is zero, yet the Hessian matrix (second derivatives) exhibits a mixed signature and contains both positive and negative eigenvalues. The presence of at least one negative eigenvalue indicates a direction of instability, which distinguishes a saddle point from a local minimum.

Saddle points are classified by their *index*, which is the number of negative eigenvalues in the Hessian matrix at that point. This index determines the order of the saddle point:

- A *first-order* saddle point, often simply called a saddle point, has exactly one negative eigenvalue. This type of saddle point typically represents a transition state along a reaction pathway.
- Higher-order saddle points have more than one negative eigenvalue and represent more complex transition states that may involve simultaneous changes in multiple directions.

Understanding the order of saddle points is crucial for analyzing the pathway and mechanism of chemical reactions. For instance, identifying first-order saddle points is essential for locating transition states, which are pivotal for calculating activation energies and reaction rates.

Computing the potential energy along a single degree of freedom is relevant to understanding the properties of a molecule for a simple study of a H_2 system for example. However, for other systems, the PES becomes more complex and numerous degenerate basins may exist. In this context, exploring the PES becomes a challenging task as it involves searching for the most stable configurations of the molecule and the transition

states between different configurations. Several methods have been developed to explore the PES. These methods encompass a diverse array of techniques, including those based on Monte Carlo and Molecular Dynamics simulations, global optimization techniques, as well as robotic-inspired approaches derived from path planning. Algorithms focused on finding transition paths between energy basins, which are crucial for understanding the dynamics of chemical reactions and predicting reaction/transition rates, will be discussed in the Chapter 6.

1.4.1 Basic sampling methods

This section will introduce the most widely used sampling techniques employed to investigate the PES of molecules. Molecular Dynamic and Monte Carlo are effective for global optimization methods obtaining thermodynamic properties for both and kinetic properties for the Molecular Dynamics. Both of these methods employ the canonical ensemble as the sampling framework.

1.4.1.1 Monte Carlo (MC)

The Monte Carlo (MC) method was developed by Metropolis and Ulam in the 1940s with the objective of calculating multidimensional integrals [117]. The Monte Carlo (MC) method is a fundamental stochastic technique used to explore the potential energy surface of chemical systems by sampling configurations randomly. The most common Monte Carlo methods is the Metropolis Monte Carlo (MMC). MMC proposed by Metropolis et al. [118] is a widely used MC method that generates a sequence of configurations by accepting or rejecting proposed moves based on the Metropolis criterion. The Metropolis criterion is based on the Boltzmann distribution, which states that the probability of a system being in a particular state depends exponentially on its energy. The Metropolis criterion is given by:

$$P_{\text{accept}} = \min\left(1, \exp\left(-\frac{\Delta E}{k_B T}\right)\right) \tag{1.21}$$

where ΔE is the change in energy of the system, k_B is the Boltzmann constant equal to 1.380649 $10^{-23} m^2 kg s^{-2} K^{-1}$, and T is the temperature of the system. The Metropolis criterion ensures that the system moves towards lower energy states, corresponding to the most stable configurations of the molecule. The MMC method is widely used in computational chemistry to explore the PES, optimize molecular structures, and simulate chemical reactions. The power of this method lies in its simplicity and versatility, as it requires minimal assumptions about the system being studied. However, this method can be inefficient if the random sampling does not cover the significant regions of the PES effectively. This limitation is often mitigated by more sophisticated techniques in other variants of Monte Carlo methods. The efficiency of the Monte Carlo technique is highly dependent on the number of samples and the distribution from which these samples are drawn, making it crucial to ensure a wide and representative coverage of the state space to obtain accurate results.

1.4.1.2 Molecular Dynamics (MD)

Molecular Dynamics (MD) simulations are a powerful tool for studying the dynamic behavior of molecules by solving the classical equations of motion for the atoms in the molecule. MD simulations are based on Newton's laws of motion (Eq. A.22), which describe how the positions and velocities of particles change over time. The equations of motion are integrated numerically to simulate the motion of the atoms in the molecule. MD simulations can be used to explore the PES, optimize molecular structures, and simulate chemical reactions.

$$m_i \frac{d^2\mathbf{r}_i}{dt^2} = -\frac{\partial E}{\partial \mathbf{r}_i} = F_i \tag{1.22}$$

where m_i is the mass of the i-th atom, $\mathbf{r}_i$ is the position of the i-th atom, E is the energy of the system, and F_i is the force acting on the i-th atom. The force acting on the atoms is calculated from the gradient of the potential energy.
To model the system, the equation of motion needs to be integrated using various algorithms, most of which are based on Taylor series expansion. The Verlet algorithm is the most commonly used algorithm and is expressed as follows:

$$\begin{aligned} \mathbf{r}_i(t+\Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \mathbf{a}_i(t)\tfrac{\Delta t^2}{2} \\ \mathbf{v}_i(t+\Delta t) &= \mathbf{v}_i(t) + \tfrac{\Delta t}{2}\left(\mathbf{a}_i(t) + \mathbf{a}_i(t+\Delta t)\right) \end{aligned} \tag{1.23}$$

where $\mathbf{r}_i$ is the position of the i-th atom, $\mathbf{v}_i$ is the velocity of the i-th atom, $\mathbf{a}_i$ is the acceleration of the i-th atom, and Δt is the time step. The Verlet algorithm is a symplectic integrator that conserves energy and momentum, making it suitable for long-time simulations of molecular systems. One step Verlet algorithm is expressed as follow:

1. Initialization of step Δt, total simulation time $\mathcal{T}$
2. Initializing initial conditions : $t = 0$, $\mathbf{r}_i(0)$, $\mathbf{v}_i(0)$
3. Definition of function $\mathbf{a}_i$
4. While $t < \mathcal{T}$:
   (a) Calculation of $\mathbf{a}_i(t)$
   (b) Calculation of $\mathbf{r}_i(t + \Delta t)$
   (c) Calculation of $\mathbf{a}_i(t + \Delta t)$
   (d) Calculation of $\mathbf{v}_i(t + \Delta t)$
   (e) $t = t + \Delta t$

Each simulation takes place in statistical sets that define the thermodynamic quantities and their relationships. The most commonly used sets are the micro-canonical (NVE), canonical (NVT), and isothermal-isobaric (NPT) sets. In these sets, the parameters are kept constant according to the following nomenclature: E for energy, N for

number of atoms, P for pressure, T for temperature, and V for volume. For example, an NVT simulation will have a constant number of atoms, volume, and temperature. Appropriate thermostats and barostats, such as Nosé-Hoover and Andersen, are used to maintain pressure and temperature. However, this method can be inefficient for exploring the PES of complex systems with high energy barriers, as the system may become trapped in local minima. To overcome this limitation, enhanced sampling methods have been developed to accelerate the exploration of the PES and improve the accuracy of the results.

1.4.2 Enhanced sampling methods

Enhanced sampling methods are techniques designed to improve the efficiency of exploring the PES by overcoming the limitations of standard sampling methods. These methods are particularly advantageous in systems with rugged energy landscapes, where the presence of high energy barriers can impede the convergence of simulations. The objective of enhanced sampling methods is to accelerate the exploration of the PES, enhance the sampling of rare events, and improve the accuracy of the results. These methods employ biasing potentials, reweighting schemes, or advanced algorithms to guide the simulation towards important regions of the energy landscape. The aforementioned approaches do not permit the direct acquisition of the system's thermodynamic properties; however, certain methodologies have been developed to identify these properties through the analysis and re-weighting of the simulation results.

1.4.2.1 Parallel Tempering (PT)

Parallel Tempering (PT), also known as *replica exchange method*, is an advanced technique designed to ameliorate ergodicity and convergence issues in MC and MD simulations. Initially introduced by Swendsen and Wang for MC simulations [152] and later adapted to MD simulations by Sugita and Okamoto [150], PT is widely utilized across various studies [50, 30, 151].

PT involves running multiple, simultaneous simulations at different temperatures. This method allows systems to exchange configurations at regular intervals, promoting the exploration of potential energy surfaces by enabling systems to overcome high energy barriers that would otherwise hinder simulation convergence. Such exchanges are governed by a carefully designed Metropolis-Hastings criterion, ensuring that the detailed balance is maintained and the thermodynamic equilibrium is not violated. The criterion for accepting a swap between replicas i and j with temperatures T_i and T_j is given by:

$$p = \min\left(1, \exp\left(-\Delta E\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\right)\right), \tag{1.24}$$

where $\Delta E = (E_j - E_i)$ and E_i and E_j are the energies of the replicas i and j, respectively.

PT proves a particular efficiency in systems with rugged energy landscapes, where numerous local minima are separated by high barriers. By allowing replicas at lower temperatures (and thus higher resolution and stability) to exchange information with

higher-temperature replicas, PT facilitates the crossing of energy barriers that would be insurmountable at lower temperatures alone. This mechanism significantly enhances the ability of the simulations to find the global minimum and accurately sample the PES.

1.4.2.2 Umbrella Sampling (US)

Umbrella Sampling (US) is a sophisticated computational technique developed to calculate the free energy profile along a specified reaction coordinate. Introduced by Torrie and Valleau [156], this method enhances the ability to explore PES effectively, particularly in regions that are typically difficult to sample due to high energy barriers or low probabilities of occurrence.

In the Umbrella Sampling approach, the potential energy of the system is deliberately biased along the reaction coordinate. This biasing is achieved through the introduction of an additional potential, known as the *umbrella potential*, which is designed to make less probable states more accessible. By modifying the landscape of the PES, US allows for more thorough sampling in regions of interest, such as transition states or intermediate states in a chemical reaction.

The process involves performing a series of simulations, each with a slightly different biasing potential applied to a particular segment of the reaction coordinate. The data collected from these simulations are then integrated using techniques such as the Weighted Histogram Analysis Method (WHAM) [101] to reconstruct the unbiased free energy profile.

This method is extensively utilized in computational chemistry to study free energy changes in chemical reactions, predict molecular properties, and understand complex biochemical pathways.

1.4.2.3 Metadynamics

Metadynamics is a powerful computational method designed to enhance the exploration of the PES and facilitate the calculation of free energy profiles. Introduced by Laio and Parrinello [102], Metadynamics employs a history-dependent bias potential to prevent the system from becoming trapped in local minima, a common challenge in molecular dynamics simulations.

The core mechanism of Metadynamics [16] involves the periodic addition of Gaussian potentials at the position of the system's current state in a selected reaction coordinate. This strategy discourages the system from revisiting previously sampled states by effectively creating a repulsive memory of these states. Each Gaussian potential is characterized by its width and height, which are critical for ensuring adequate exploration of the PES without sacrificing the resolution of important features.

As the simulation progresses, these Gaussian potentials accumulate, creating a bias that pushes the system to explore new regions. The system's tendency to revisit certain states decreases, allowing for a comprehensive exploration of the PES.

Metadynamics has become a widely used technique in computational chemistry for studying complex chemical reactions and predicting molecular properties. It is partic-

ularly valuable for mapping out free energy landscapes of molecular systems, exploring transition states, and understanding the energetics of biochemical pathways. The method's ability to provide deep insights into the thermodynamics of molecular interactions makes it an indispensable tool in the theoretical chemist's arsenal.

1.4.3 Global optimization methods

Global optimization methods are techniques employed to efficiently sample the PES and identify the most stable states of a chemical system. These methods are not designed to obtain the thermodynamic properties of the system; rather, they are employed to efficiently converge to the basins of the PES. These methods are not only applicable in chemistry for optimizing molecular structures but also across various fields like physics, economics, and operations research, where navigating complex functions to find optimal solutions is essential.

1.4.3.1 Basin Hopping (BH)

Basin Hopping (BH) proposed by Li and Scheraga [109] and Wales and Doye [162] is a MMC method that incorporates a local optimization stage. More precisely, BH (see Fig. A.3) generates a sequence of configurations by performing local optimization steps followed by random perturbations of the atomic positions. The local optimization step minimizes the energy of the system by adjusting the atomic positions to reach a local minimum on the PES. BH relies on a Metropolis criterion to accept or reject a configuration obtained after the local optimization step. The random perturbations introduce noise into the system, allowing it to escape from local minima and explore different regions of the PES. The BH method is efficient to explore the PES, identify the most stable configurations of the molecule.

1.4.3.2 Simulated Annealing (SA)

Simulated Annealing (SA) is a stochastic optimization technique inspired by the annealing process in metallurgy, where materials are heated and then gradually cooled to minimize their defects and increase ductility. This method, conceptualized by Kirkpatrick, Gelatt, and Vecchi [97], is designed to find the global minimum of a function over a large search space, making it ideal for complex optimization problems such as molecular structure optimization.

The SA algorithm [157] starts with a high initial temperature to allow for extensive exploration of the PES. This high-temperature phase helps the system to overcome and escape from local minima early in the optimization process. As the temperature decreases, the algorithm reduces the scale of exploration, fine-tuning the solution as it approaches lower energy states. The temperature reduction must be carefully controlled by a cooling schedule, which critically influences the balance between exploration and exploitation.

Key to the SA method is the acceptance of new states during the search process, which is governed by the Metropolis criterion (Eqn. A.21). This criterion allows the

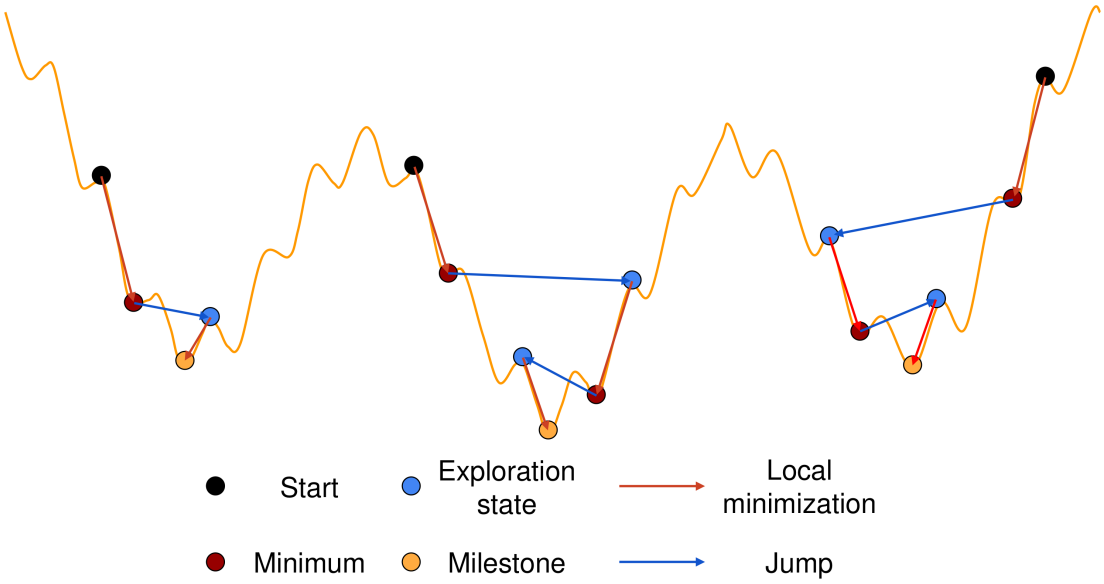


Figure 1.3: Basin Hopping method for exploring the PES.

algorithm to accept not only moves that lower the energy but also some that increase it, thus avoiding the trap of local minima in the early stages.

The process continues by cyclically modifying the system's configuration and gradually lowering the temperature until a minimum cooling temperature is reached or other stopping criteria are met.

1.4.3.3 Genetic Algorithms (GA)

Genetic Algorithms (GA) are a class of stochastic optimization methods that mimic the process of natural selection and evolution, as described by Darwin. This approach was formalized by Holland [87] and is particularly useful in computational chemistry for optimizing molecular structures. GAs operate by generating a diverse population of candidate solutions, each representing a possible configuration of the molecule under study.

The core of the GA method lies in its iterative process where the population evolves over multiple generations towards an optimal solution. This evolution is driven by genetic operators: selection, crossover, and mutation. Selection mimics natural survival pressures by preferring individuals with higher fitness levels, allowing them to pass their genes to the next generation. Crossover, or recombination, is a process where pairs of individuals exchange segments of their genetic material to produce new variants, combining beneficial traits from both parents. Mutation introduces random changes to individual genes, providing new genetic variations and helping the population avoid local minima by exploring new areas of the solution space.

Each iteration of the algorithm evaluates the fitness of all individuals in the population, typically measured by how well they solve the optimization problem or meet the

desired criteria. The genetic operators are then applied to create a new generation, ideally with higher average fitness than the previous one. Over successive generations, the population converges towards an optimal solution, mimicking the evolutionary process of adaptation.

The flexibility and effectiveness of genetic algorithms make them particularly suitable for problems where traditional optimization techniques struggle to perform well due to the complexity of the landscape involving numerous local optima.

1.4.4 Robotics-inspired methods

Motion planning is a fundamental problem in robotics that involves the identification of a collision-free path for a robot to move from an initial configuration to a goal configuration. This can also be applied to a robotic arm with a limited number of joints that is required to pick up an object, for example. A variety of algorithms have been developed to achieve these objectives. These algorithms have evolved beyond their original scope and have been employed in diverse fields, including industrial manufacturing, computer animation, and computational structural biology. For instance, they have been utilized in the context of protein folding and the optimization of molecular structures [26, 107, 124, 22]. In computational chemistry, these algorithms have been employed to efficiently explore the PES. Contrary to the global optimization methods, these algorithms are firstly designed to efficiently explore a high-dimensional space, but are not directly aimed at finding the global minimum. Nevertheless, the last method presented named Iterative Global Exploration and Local Optimization (IGLOO) [112] is a method that combines both motion planning algorithm and local optimization. Some of these algorithms are capable of identifying low energy states, as well as connecting them together in order to identify the transition paths between them. These particularities will be developed in greater detail in Chapter 6.

1.4.4.1 Probabilistic Roadmap (PRM)

The Probabilistic Roadmap (PRM) method, introduced by Kavraki et al. [96], is used to solve high-dimensional motion planning problems.
The PRM works by iteratively sampling a configuration of the configuration space. If the configuration is collision-free, it is added to the roadmap as a node. The new node is connected to the roadmap by finding its nearest neighbors. If the path between the new node and the nearest neighbors is collision-free, it is added to the roadmap as a straight line. These steps are iterated until a stopping criterion is reached. The roadmap then can be used to find a path between nodes using conventional graph search algorithms such as Dijkstra or A* [22]. Extensions of PRM involving energy calculations, presented next, have been proposed to explore the PES.

1.4.4.2 Probabilistic Conformational Roadmaps (PCR)

The Probabilistic Conformational Roadmaps (PCR) method proposed by Singh, Latombe, and Brutlag [145] is a PRM-based method that generates a roadmap by

accepting or rejecting new nodes using a probality function favoring low energy conformations. The probability function is evaluate as follow:

$$P(\text{accept}, q) = \begin{cases} 1 & \text{if } E_q < E_{\min} \\ \dfrac{E_{\max} - E_q}{E_{\max} - E_{\min}} & \text{if } E_{\min} \leq E_q \leq E_{\max} \\ 0 & \text{if } E_q > E_{\max} \end{cases} \tag{1.25}$$

where E_q is the energy of the conformation q, $E_{\min}$ and $E_{\max}$ are threshold values fixed for the system. For each edge e_{ij}, a weight is associated, representing the likelihood of the transition between the connected conformations. A series of intermediate conformations are generated along the path $\{q_i = c_0, c_1, .., c_n = q_j\}$ connecting the two nodes q_i and q_j (number of intermediate conformations is a parameter). The weight of the edge is computed as follow:

$$\begin{aligned} w(e_{ij}) &= -\sum_{i=0}^{n-1} \log(P_i) \\ P_i &= \frac{e^{-\frac{(E_{i+1}-E_i)}{KT}}}{e^{-\frac{(E_{i+1}-E_i)}{KT}} + e^{-\frac{(E_{i-1}-E_i)}{KT}}} \end{aligned} \tag{1.26}$$

where E_i is the energy of the conformation c_i, n the number of images, K is the Boltzmann constant and T is the temperature. PCR has been applied to find energetically favorable motions of bio-molecules [8].

1.4.4.3 Stochastic Roadmap Simulation (SRS)

The Stochastic Roadmap Simulation (SRS) method [10, 11, 35, 36, 9] is an improvement of PCR. The difference is the probability function, which is consistent with the Metropolis criterion [118]. The probability function is evaluated as follows:

$$P_{ij} = \begin{cases} \frac{1}{n_i} \exp(-\frac{\Delta E_{ij}}{KT}) & \text{if } \Delta E_{ij} > 0 \\ \frac{1}{n_i} & \text{otherwise} \end{cases} \tag{1.27}$$

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij} \tag{1.28}$$

where n_i is the number of neighbors of the node q_i, ΔE_{ij} is the energy difference between the nodes q_i and q_j, K is the Boltzmann constant and T is the temperature. SRS was used to predict ligand-protein interactions [8].

1.4.4.4 Rapidly-exploring Random Tree (RRT)

The Rapidly-exploring Random Tree (RRT) method, introduced by LaValle [103], is used to solve high-dimensional motion planning problems. The RRT works by iteratively sampling a configuration of the configuration space. If the configuration is collision-free, it is added to the tree as a node. The new node is connected to the tree by finding its nearest neighbor. The tree is expanded by adding new nodes in the direction of the random sample. The main difference with this method and the PRM is that the new

node is linked to the closest neighbor and not every neighbors where a path collision free exist. The RRT method will be discussed in more detail in the Chapter 2.

1.4.4.5 Transition-RRT (T-RRT)

The Transition-RRT (T-RRT) method proposed by Jaillet, Cortés, and Siméon [92, 93] is a RRT-based method that introduced a transition test to favor the exploration of low-energy regions of the PES. The transition test is based on the Metropolis criterion inspired from MC methods and is used to accept or reject new nodes based on the change in energy and the temperature of the system. Contrary to MC method, the temperature is auto-adaptative during the exploration to dynamically adjust the exploration of the PES. The algorithm keeps track of every rejection and acceptance of the transition test to adjust the temperature. The T-RRT method will be discussed in more detail in the Chapter 6.

1.4.4.6 Iterative Global Exploration and Local Optimization (IGLOO)

The Iterative Global Exploration and Local Optimization (IGLOO) method [112] combines the exploration of the PES with a local optimization. The IGLOO method is an iterative algorithm composed of three main steps: an exploration step, a local optimization step and a filtering step. The exploration step is performed using a RRT-based method to explore the PES. The local optimization step is performed using a local optimization method to minimize the potential energy of the molecules. The filtering step is used to remove redundant states and improve the efficiency of the exploration at the next iteration. IGLOO was successfully applied to predict the structure of disaccharide molecules on metal surfaces [1, 2]. IGLOO will be discussed in more detail in the Chapter 2.

1.5 Contribution summary

The thesis contains several contributions to the field of computational chemistry.

An overview of the different chapters is given here.

Chapter 2: This chapter presents the coupling of the IGLOO and DFTB methods for the exploration of the conformational space of molecules. IGLOO is inspired by robotics motion planning, while DFTB is an approximate quantum chemistry method. The implementation details entail interfacing software developed in our laboratories. The IGLOO method, implemented in the MoMA software suite, is coupled with the DFTB method, implemented in the deMonNano code. As a first study, the coupled approach was applied to the alanine dipeptide, a small peptide. The exploration identified the lowest energy conformations, thereby demonstrating the efficiency of the coupling in reducing computational costs while maintaining an accurate description of the chemical system.

Chapter 3: This chapter examines the PES of phthalate molecules using the IGLOO/DFTB coupling methodology introduced in the previous chapter. Phthalates are a family of compounds that are widely used in consumer products. It is important to understand their conformational behavior given the potential environmental and health impacts of these compounds. The chapter starts with an introduction to phthalates, emphasising their significance and the necessity for detailed energy landscape exploration. The methodology entails initializing IGLOO with a multitude of initial states to ensure comprehensive coverage and performing multiple independent runs to account for the stochastic nature of the method. This approach revealed a multitude of energy basins and facilitated the identification of stable conformations across a diverse range of phthalate molecules. Significant findings include the identification of various conformational minima, which were analyzed using both energetic and structural descriptors. The aforementioned descriptors facilitated an understanding of the interactions within phthalate molecules, including the effects of side-chain arrangements on molecular stability. Furthermore, the chapter compares DFTB and DFT calculations to validate the former's accuracy in representing phthalate energetics. The outcomes illustrate the efficiency of the IGLOO/DFTB approach in delineating the intricate potential energy landscapes of phthalates, furnishing valuable insights into their conformational dynamics and stability. The chapter proposes the further application of this methodology to other complex molecular systems, with the objective of generalizing the approach and refining molecular energy exploration techniques.

Chapter 4: This chapter presents an innovative algorithm for the generation of atomistic models of aromatic hydrocarbons on a large scale. The primary focus is the integration of molecular graph-based generation techniques with atom and fragment additions, with particular emphasis on the maintenance of predefined constraints on chemical structures. The introduction provides an overview of the significance of aromatic hydrocarbons in various scientific fields, including astrophysics and environmental science. It emphasizes the necessity for accurate models to simulate and understand their behavior in different environments. The methodology comprises two principal components: the SMILES Generator and the Structure Generator. The SMILES Generator algorithm is designed to produce a series of SMILES strings that adhere to specified constraints on the types and ratios of bonds and atoms. This is accomplished through a meticulous process that encompasses the selection of fragment types, the selection of atoms within the molecular graph, and the addition of fragments in order to incrementally construct the molecular structure. Subsequently, the Structure Generator algorithm assumes control to generate three-dimensional structures. This process involves the generation of initial, unoptimized structures from the SMILES outputs. These structures are then optimized through a series of steps aimed at minimizing self-collision and ensuring structural validity.

Chapter 5: This chapter examines the application of previously developed algorithms to the study of substructures of hydrogenated amorphous carbon polymers. The chapter

starts with an overview of the current understanding of substructures of hydrogenated amorphous carbon polymers in the interstellar medium (ISM), emphasising their detection through infrared absorption bands and their pivotal role in various physico-chemical processes in space. The methodology section outlines the production and analysis of hydrogenated amorphous carbon polymers. Subsequently, metrics for evaluating the generated structures are defined, focusing on both geometric and electronic descriptors. A Geometric descriptor as Hill-Wheeler parameters, which assess the shape deformation from a perfect sphere is defined, and electronic descriptors such as the HOMO-LUMO gap and London energy are computed, which provide insights into the electronic properties of the structures. The application of these descriptors has shown significant variations in the shapes and electronic properties of the substructures of hydrogenated amorphous carbon polymers structures.

Chapter 6: This chapter adresses the transition paths between low-energy conformations in molecular systems, presenting various computational techniques to map out these pathways. The chapter begins by discussing the theoretical foundation provided by Transition State Theory, emphasizing the importance of identifying the Minimum Energy Path (MEP) which represents the most favorable route for a reaction to proceed. A variety of computational methods are discussed to identify and analyze these transition paths. These include the dimer method for locating saddle points on the potential energy surface, and advanced methodologies such as the Nudged Elastic Band (NEB) method, which refines the path to minimize the energy along the reaction coordinate. A preliminary methodology to explore the diversity of transition paths between low-energy conformations is presented. The exploration of these paths is performed using the stochastic algorithm T-RRT, which generates numerous paths. A similarity measure is then applied to differentiate these paths, and a clustering method is then used to identify common patterns. Subsequently, a representative path of each cluster is selected for local optimization. The methodology is demonstrated on the alanine dipeptide molecule, and preliminary results are presented.

Chapter 2

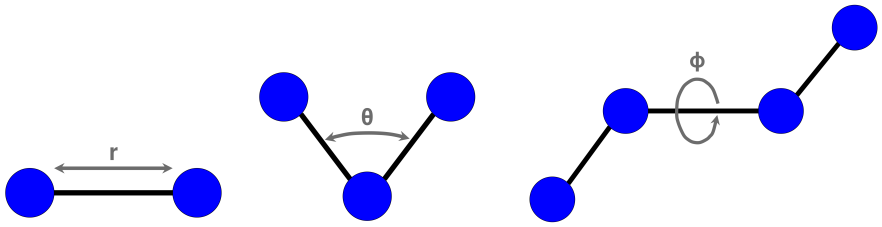
IGLOO/DFTB coupling

Contents

2.1 Introduction

2.1.1 Context: Conformational space exploration of molecules

Modeling is used to describe the process of representing a system in a way that is conducive to the study of its properties. The selection of an appropriate representation is dependent upon the size of the system or the level of detail required. An intuitive representation of a system is by its atomic coordinates, which are defined as the positions of the atoms in space. This representation is referred to as the all-atom representation. Although this representation is intuitive, it is not always the most efficient for exploring the conformational space of a molecule. An other way to represent a molecule is to use internal coordinates, which can be defined by a number of geometric parameters, including bond lengths, bond angles, and dihedral angles (Fig. 2.1). These parameters serve to determine the shape and structure of the molecule. The dihedral angle, also known as the torsion angle, is the angle between two planes, each defined by three atoms in the molecule. This angle is frequently employed in the investigation of biological molecules, such as proteins and nucleic acids, to track their conformational modifications. In some systems, the rigid geometry assumption can be employed to simplify the description of high-dimensional systems. This is achieved by reducing the



(a) Bond (b) Angle (c) Dihedral angle or Torsion

Figure 2.1: Degrees of freedom in a molecule. Blue circles represent atoms and black segments represent the bonds between them.

conformational space to a few degrees of freedom, such as dihedral angles, while other degrees are fixed at an equilibrium value. This assumption is applicable to systems where the impact of fluctuations in other degrees of freedom on the potential energy can be neglected.

Depending on the system and the type of exploration, dihedrals angles can be sufficient to represent the major modifications being prevalent on the total energy. While a representation based on dihedral angles is sufficient for the global exploration of the PES, a local optimization of the structure can be performed using a more detailed representation. Based on this assumption, we present the coupling of a robotics-inspired methods that explores the conformational space of molecules to find the lowest energy conformation with a quantum chemical method. Dihedral angles are employed to represent the conformational space of the molecule during the global exploration, while each low-energy conformation identified is then locally optimized using an all-atom representation. This strategy provides a good compromise between accuracy and computational cost. Both algorithms are presented, and the implementation details of the coupling are discussed.

2.1.2 State of the art: Exploration of the conformational space of molecules

The theoretical prediction of physico-chemical properties of molecules such as chemical reactivity, ionisation energies, spectroscopy often requires the knowledge of their low energy conformations. When the Born-Oppenheimer approximation can be applied, this involves the search for the most stable minima of the PES of the electronic ground state. In the case of flexible molecules (for instance biomolecules, polymers, chemicals, pollutants), the efficient exploration of their associated high-dimensional conformational space remains a challenging task, especially when a level of description of the potential close to an *ab initio* method is required. For this purpose, a cautious choice of the combination between an adequate exploration strategy (limiting the number of single point energy calculations to be performed, in particular over-sampling) and an appropriate level of description of the PES (compromise between the computational cost and requested accuracy) is mandatory.

A large majority of the PES exploration schemes [138, 27], that was presented on the chapter 1, rely on either Monte Carlo[78] (MC) methods or Molecular Dynamics simulations[79] (MD). MC is a stochastic approach often performed within the Metropolis algorithm where random displacements are accepted as a function of a temperature, while MD propagates the nuclei positions by solving Newton's equations of motion. These methods are robust in exploring a conformational landscape and can provide a thermodynamic interpretation. They can also be combined with periodic local minimizations to locate the bottom of the PES wells. Various strategies have been implemented to improve the efficiency of MC- and MD-based methods. This includes for instance simulated annealing [142], parallel tempering methods [152, 150, 151] or BH schemes [109, 162], used either in their standard [4, 38, 37] or improved versions [170, 171, 34]. A disadvantage of these schemes is that they do not keep track of the visited regions, which can lead to over-sampling of certain areas of the PES to the detriment of exploration of others. In the case of free energy reaction path calculations, methods keeping a knowledge of the visited space (e.g. umbrella sampling or metadynamics) have been developed to increase the exploration efficiency [168, 16]. These latter require however *a priori* knowledge of the reaction coordinates (collective variables), which prevents their use in a context of blind exploration of complex PES. In summary, there is still work to be done to develop efficient algorithms to discover potentially diverse energy basins, i.e. without prior knowledge of the system of interest and requiring little or no adaptation to a particular case study.

In recent years, methods inspired from robot motion planning algorithms have been proposed to efficiently explore the conformational space of molecular systems [22, 73, 143]. These methods construct data structures (trees or graphs) that encode the explored regions of the space, and avoid revisiting these regions. One of these algorithms is the Rapidly-exploring Random Trees (RRT) [104], which was subsequently extended to the exploration of energy landscapes aiming to find transition paths [93, 44, 56]. More recently, the RRT and Basin Hopping (BH) algorithms have been combined to find energy minima on a PES [135]. The strategy applied in this work, called Iterative Global exploration and LOcal Optimization (IGLOO) [112], iterates RRT-based exploration, local minimizations and filtering steps. The IGLOO algorithm will be detailed in the next sections.

Various levels of theory exist to compute the energy for the visited points of the PES. They range from high-level *ab initio* schemes with the wavefunction methods to lower-levels such as force field approaches [80]. In between, Density Functional Theory [125] (DFT) is the most common method used to study systems with tens to hundreds of atoms. Unfortunately, the computational cost of DFT, although much lower than that of wavefunction methods, remains a bottleneck in the framework of exhaustive PES explorations, when millions of single point energy calculations are intended to be done. An alternative method, namely the Density-Functional based Tight-Binding approach (DFTB)[128, 141, 55, 147], relying on several approximations of DFT, preserves the explicit quantum description of the electronic system while drastically reducing the computational cost thanks to the use of parameterized integrals and a minimal valence basis set. Its DFT ground usually makes it more transferable than force field models.

In this chapter, we report the coupling of a non-redundant conformational space exploration approach, namely the robotics-inspired IGLOO method, with the quantum chemical DFTB potential. This enable efficient discovery of diverse energy basins while preserving a highly-accurate level of description of chemical systems.

2.2 Exploration and energy calculation methods

In the following section each phase of the IGLOO algorithm is detailed and the coupling with the DFTB method is presented. In addition, several local optimizations were developed and are described in detail.

2.2.1 Rapidly exploring Random Tree (RRT) algorithm

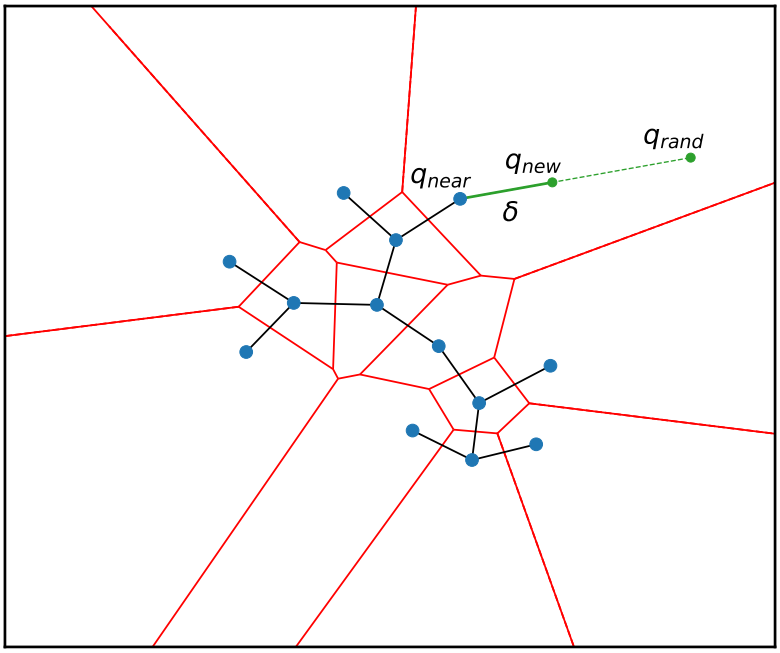


Figure 2.2: Illustration of the RRT algorithm. Blue point are nodes which are geometry of the system in the specific case of PES exploration. Edges that connect nodes between them are represented in black for the already explored tree and in green for the new explored nodes. Red segments are the frontier of the Voronoi cells.

The Rapidly-exploring Random Tree (RRT) [104] algorithm is a motion planning algorithm that builds a tree rooted at an initial state and grows it by adding new states in the direction of randomly sampled points in the configuration space. The tree is built by connecting the closest state to the sampled point, and the new state is added to the tree if it is reachable from the closest state. The algorithm is illustrated in Fig. 2.2 and algorithm 1. A node q_{rand} is sampled in the configuration space and the nearest node q_{near} in the tree is found. A new node q_{new} is then created by moving from q_{near} to q_{rand} with a step size of δ.

Algorithm 1: RRT-Exploration (Pseudo-code from [112])

input : Conformational space variables $\mathcal{C}$
Algorithm parameters and energy function P
Initial conformations for the exploration (roots) $\mathcal{Q}$
output: Set of trees containing conformations $\mathcal{T}$

$\mathcal{T} \leftarrow \mathcal{Q}$
while not `StoppingCriteria`$(P, \mathcal{T})$ **do**
  $\mathcal{T}_i \leftarrow$ `SelectTree`$(\mathcal{T})$
  $q_{\text{rand}} \leftarrow$ `SampleRandomConformation`$(\mathcal{C}, \mathcal{T}_i)$
  $q_{\text{near}} \leftarrow$ `GetNearestNeighbor`$(\mathcal{T}_i, q_{\text{rand}})$
  $q_{\text{new}} \leftarrow$ `ExtendTree`$(P, q_{\text{near}}, q_{\text{rand}})$
  if `ValidConformation`(P, q_{new}) **then**
    $\mathcal{T} \leftarrow$ `AddNode`$(\mathcal{T}_i, q_{\text{new}})$
return $\mathcal{T}$

For robot motion planning, a state is a configuration of the robot, and the tree is built in configuration space. In the context of molecular conformational space exploration, a state is a conformation of the molecule, and the tree is built in conformational space. In the following implementation, the conformational space is defined by the dihedral angles of the molecule.

2.2.2 Iterative Global exploration and LOcal Optimization

The Iterative Global exploration and LOcal Optimization (IGLOO) algorithm performs a global exploration of the conformational landscape of molecules to find the lowest energy representatives. IGLOO relies on an exploration strategy originally proposed to solve motion planning problems in robotics. More precisely, it applies a variant of the RRT algorithm [104], adapted to the exploration of energy landscapes [40, 93]. Similarly to other techniques that perform global optimization by iterating local searches, such as the Basin-Hopping algorithm [109, 162], the RRT-based exploration is coupled with a local optimization technique with the aim of descending into the energy basins. The IGLOO algorithm interleaves global explorations and local minimization stages in an iterative manner, also including a filtering stage to reduce the number of states considered in subsequent iterations. A more detailed description of these stages will be provided below, together with the implementation details. More in-depth explanations about IGLOO can also be found in a recent work [112], which demonstrates the good performance of the algorithm compared to related methods for finding low-energy conformations of molecular systems. Note also that IGLOO was successfully applied to predict the structure of disaccharide molecules on metal surfaces [1, 2].

2.2.3 Density-Functional based Tight-Binding

The DFTB method is an approximated DFT scheme developed in the mid-90's [128, 141] following the pioneering work of Foulkes and Haydock[64]. It is derived from a Taylor expansion of the Kohn-Sham effective potential energy with respect to the electronic

density fluctuation aroung a reference density:

$$\rho = \rho_0 + \delta\rho \tag{2.1}$$

The molecular orbitals are expressed in a minimal valence basis set. At zeroth-order, the algebra is equivalent to standard non-consistent semi-empirical Tight-Binding [76] and the formal derivation allows for the tabulation of the Kohn-Sham and overlap matrices elements in the atomic basis as diatomic terms from reference DFT calculations. The energy consists in the usual "band structure" terms and a short-range repulsive term. The energy of the zeroth-order DFTB could be expressed as:

$$E_{\text{DFTB-0}[\rho_0]} = E_{band[\rho_0]} + E_{rep[\rho_0]} \tag{2.2}$$

The band energy term corresponds to the sum of the occupied molecular orbitals energies and is expressed as:

$$E_{band[\rho_0]} = \sum_{i}^{N_{\text{occ}}} \left\langle \psi_i \left| \hat{H}_0 \right| \psi_i \right\rangle \tag{2.3}$$

The ψ_i are the molecular orbitals and $\hat{H}_0$ is the Kohn Sham operator at the electronic reference density, expressed in the atomic basis set as:

$$\hat{H}_0 = -\frac{1}{2}\nabla^2 + v_{ext}(r) + \int \frac{\rho_0(r')}{|r - r'|} dr' + v_{xc}[\rho_0] \tag{2.4}$$

The repulsive term $E_{rep[\rho_0]}$ is expressed as:

$$E_{rep[\rho_0]} = -\frac{1}{2}\int\int \frac{\rho_0(r)\rho_0(r')}{|r - r'|} dr dr' + E_{\text{xc}}[\rho_0] - \int v_{xc}[\rho_0]\rho_0(r) dr + E_{NN} \tag{2.5}$$

where E_{NN} is the nuclear repulsion but in pratice, it is replaced by a pair potential repulsion contribution:

$$E_{rep[\rho_0]} = \frac{1}{2}\sum_{i \neq j} V_{rep}^{ij}(R_i - R_j) \tag{2.6}$$

where V_{rep}^{ij} is a repulsion potential for the pair of atoms i and j at a distance $R_i - R_j$. The values of each term are tabulated from DFT calculations on an isolated pair of atoms.

This approach was further extended by Elstner et al.[55] to include second order terms in the Taylor expansion. The second order DFTB energy is expressed as:

$$E_{\text{DFTB-2nd}[\rho_0]} = E_{\text{DFTB-0}[\rho_0]} + E_{\text{coul}[\rho_0]} \tag{2.7}$$

The correction term often called Coulomb term $E_{\text{coul}[\rho_0]}$ is expressed as:

$$E_{\text{coul}}[\rho_0] = \frac{1}{2}\int\int \left(\frac{1}{|r - r'|} + \frac{\delta^2 E_{\text{xc}}[\rho^0]}{\delta\rho(r)\delta\rho(r')} \right) \delta\rho(r)\delta\rho(r')\, dr dr' \tag{2.8}$$

At long distances, this correction accounts for the long-range electrostatic interactions between point charges and, at short distances, it also includes exchange-correlation contributions. In practice, it is expressed as a function of atomic charges:

$$E_{\text{coul}[\rho_0]} = \frac{1}{2}\sum_i \sum_j \Delta q_i \Delta q_j \gamma_{ij}(R_i - R_j) \tag{2.9}$$

where Δq_i is the charge deviation of the atom i from the neutral atom reference, γ_{ii} correspond to the Hubbard parameters of atoms U_A and interatomic terms γ_{ij} are computed using a "Coulomb shielded" expression.

Whereas the electronic problem can be solved with a single diagonalisation in the case of the zeroth-order DFTB, the introduction of second order contributions implies a self-consistent search, often referred to as Self-Consistent-Charge (SCC), for the electronic ground state density and energy. Indeed, the second order contribution induces charges dependence of the operator. The SCC extension allows DFTB to address problems for which the zeroth-order DFTB approach is not sufficient, in particular when atomic charges deviate from the neutral atoms reference and/or when the Coulomb interaction between atomic charges has a decisive role in the determination of structural or energetic properties. More recently, DFTB has also been improved by including third order terms in the Taylor expansion, which introduces a charge dependence of the chemical hardness[169, 71]. The third order DFTB energy is expressed as:

$$E_{\text{DFTB-3rd}[\rho_0]} = E_{\text{DFTB-2nd}[\rho_0]} + E_{\text{hard}[\rho_0]} \tag{2.10}$$

The hardness derivation term $E_{\text{hard}[\rho_0]}$ is expressed as:

$$E_{\text{hard}[\rho_0]} = \frac{1}{6}\int\int\int \frac{\delta^3 E_{\text{xc}}[\rho^0]}{\delta\rho(r)\delta\rho(r')\delta\rho(r'')}\delta\rho(r)\delta\rho(r')\delta\rho(r'')\, dr dr' dr'' \tag{2.11}$$

In practice, considering the DFTB hypothesis stating that integrals considering more than two centers are neglected, the hardness derivative term is expressed as:

$$E_{\text{hard}[\rho_0]} = \frac{1}{3}\sum_{ij} \Delta q_i^2 \Delta q_j \Gamma_{ij} \tag{2.12}$$

where Γij is the derivative of the γ function respecting the charge by introducing the Hubbard derivative parameter [53].

DFTB has been applied to compute various structural, energetic and thermodynamic properties as well as vibrational and electronic spectra, covering a wide range of systems like molecules, atomic or molecular clusters, extended materials or liquids [52, 54, 65, 53, 70, 100, 122, 123, 147, 132].

2.3 Implementation details

2.3.1 Coupling of the IGLOO and DFTB methods

Our implementation of the coupled IGLOO/DFTB method is based on the interfacing of softwares developed in our laboratories: (*i*) the IGLOO algorithm implemented in the Molecular Motion Algorithms (MoMA) software suite (https://moma.laas.fr/) and (*ii*) the DFTB energy calculation implemented in the deMonNano code (http://demon-nano.ups-tlse.fr).

A schematic view of the coupling between IGLOO (MoMA) and DFTB (deMon-Nano) is provided in Fig. 2.3 and the algorithm is detailed in algorithm 2. The master code MoMA which is written in C++ , initiates requests to and receives data from deMonNano which is written in Fortran. The two programs communicate through a wrapper, allowing intercompatibility of software based on different programming languages. The protocol used for the communication is an INET socket [167] (a programming interface that enables applications to send and receive data over a network, either locally or via the Internet). The server-client architecture is used, with MoMA acting as the server and deMonNano as the client. The interfaced software takes as input a set of parameters required to initialize the algorithms, as well as a chemical structure of a molecule. The main parameters of IGLOO are the number of (randomly sampled) initial states and the step size used for the first iteration of the algorithm, which aims to cover the conformational space roughly but globally. The step size is then self-adapted in subsequent iterations, shrinking as the exploration focuses in the low-energy basins. Other important parameters concern the stopping criteria. Several types of conditions are considered to determine the end of the iterative process performed by IGLOO. They are based on: (*i*) a maximum number of iterations; (*ii*) a limited computing time; (*iii*) estimated convergence, based on the evolution of the lowest energy value. These criteria are evaluated at the end of each iteration, and the first one to be satisfied stops the algorithm. In general, the parameters corresponding to the *i* and *ii* criteria are set to high values, so that the exploration iterates until estimated convergence is reached.

After the initialization, the IGLOO algorithm iterates three successive stages (see [112] for deeper explanations on the method):

1. Exploration: At each iteration, IGLOO explores the conformational space using a stochastic process starting from a set of states. For the first iteration, these states are randomly sampled using a strategy inspired from the Poisson disk sampling process to ensure good dispersion. A variant of the RRT algorithm is then applied to explore reachable regions of the conformational space by growing random trees rooted at the initial states. New states are added to the tree if they are below an energy threshold, which is determined automatically and decreases with each iteration of the algorithm. These “single point" energy calculations are made by deMonNano, with MoMA providing the coordinates of a given conformation.

2. Local minimization: The explored conformations are minimized locally. Energy minimization is performed by deMonNano using a conjugated gradient technique.

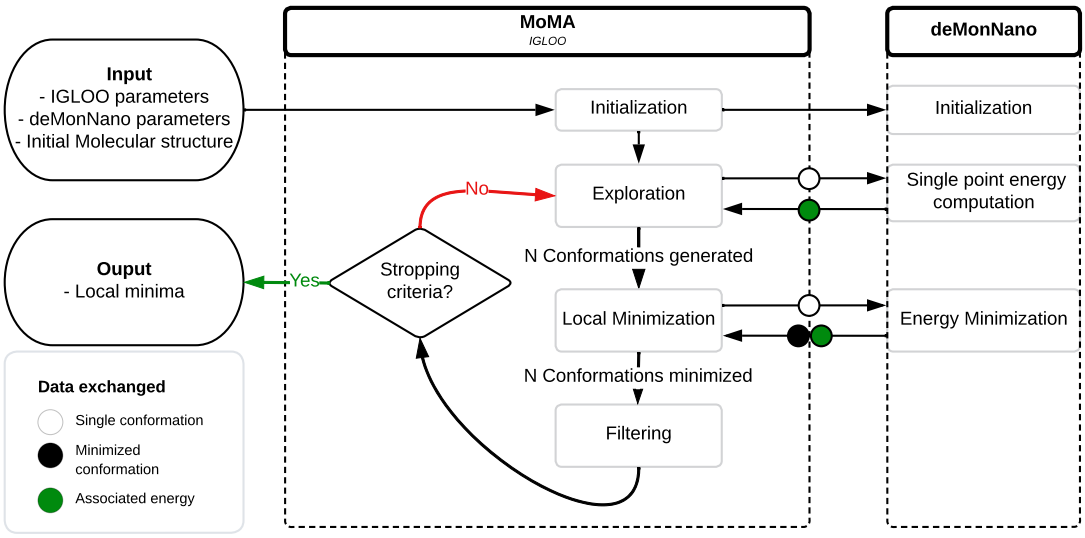


Figure 2.3: Schematic description of the IGLOO (MoMA)/DFTB (deMonNano) coupling.

In order to reduce computing time, several deMonNano executions can be performed in parallel, taking advantage of the independence of the calculations.

3. Filtering: This step enables dense areas to be cleaned up locally, with the aim to reduce the number of local minima from which the next iteration of the IGLOO algorithm is initialized.

In the present work, the degrees of freedom chosen for the IGLOO scheme are the molecular dihedral (torsion) angles, which are used both to perform geometric displacements and to characterize similarities between different conformations. The choice of dihedral angles is motivated by the fact that they are the most relevant degrees of freedom to describe the conformational space of a molecule. The dihedral angles are defined by the atoms forming the torsion. The torsion that is selected to represent the structure is that which represents the molecule's core. Then, the torsions are ordered in a list that is used to represent the conformation of the molecule. The list of torsions is used to compute the distance between two conformations, which is used to determine the next state in the tree during the exploration phase. The output of the IGLOO/DFTB method is a set of files containing the Cartesian coordinates of each local minimum found.

2.3.2 Local minimization schemes

A number of schemes were considered for the local minimization of the conformations generated by IGLOO. The details of each of these potential solutions are provided below.

Algorithm 2: IGLOO (Pseudo-code from [112])

input : Conformational space variables $\mathcal{C}$
Algorithm parameters and energy function P
output: Set of local minima $\mathcal{T}$

$\mathcal{T} \leftarrow \emptyset$; $n \leftarrow 0$
while not `MaxNumberRoots`(n, P) **do**
  $q \leftarrow$ `SampleRoot`$(\mathcal{C}, P, \mathcal{T})$
  $\mathcal{T} \leftarrow$ `AddMinimum`$(\mathcal{T}, q)$
  $n \leftarrow n + 1$
while not `StoppingCriteria`$(P, \mathcal{T})$ **do**
  $\mathcal{T}' \leftarrow$ `RRT-Exploration`$(\mathcal{C}, P, \mathcal{T})$
  foreach $q \in \mathcal{T}'$ **do**
    $q_{\text{new}} \leftarrow$ `LocalMinimization`$(\mathcal{C}, P, q)$
    $\mathcal{T} \leftarrow$ `AddMinimum`$(\mathcal{T}, q_{\text{new}})$
  $\mathcal{T} \leftarrow$ `FilterConformations`$(\mathcal{C}, P, \mathcal{T})$
  $P \leftarrow$ `UpdateParameters`$(\mathcal{T}, P)$
return $\mathcal{T}$

All-atom minimization using conjugated gradient : The first solution involved the use of an all-atom minimization approach, via the conjugate gradient method, which is a widely used iterative method for solving large systems of linear equations. This method has been implemented by Mathias Rapacioli [163] in the deMonNano code. The main idea of this method is to minimize the energy of a system by iteratively moving in a conjugated direction of the previous step. The algorithm is initialized with an initial configuration x_0 and the gradient of the energy g. Depending on the degree of freedom choosen, x_0 can be Cartesian atomic coordinates or dihedral angles values. In the *LineSearch* routine, the direction minimizing the energy is computed following a quadratic approximation as:

$$E = E(x) + \alpha \frac{\partial E}{\partial \alpha} + \frac{1}{2} \frac{\partial^2 E}{\partial \alpha^2} \tag{2.13}$$

where α is a parameter minimizing $E(x + \alpha \cdot direction)$ computed as $\alpha = -\frac{\partial E}{\partial \alpha} / \frac{\partial^2 E}{\partial \alpha^2}$. If the value of α exceeds a specified threshold, the threshold value is assigned to α while maintaining its sign, to prevent excessive displacement, which could block minimization. Then, the new configuration is computed as $x_{new} = x_{old} \pm \alpha \cdot direction$. The sign $\pm$ is determined by the configuration with the lowest energy. Note that after the first step, direction is equal to $-g_{+1} + \beta \cdot direction$.

The iterative process may be terminated if the maximum number of permitted steps has been reached. It should be noted that this type of convergence is not favored. The algorithm continues until convergence is achieved, which is determined by two parameters: the maximum gradient of the system and the difference between the previous and the current energy step. Both of these parameters must be below a specific threshold. The issue with this solution is that MoMA only considers dihedral angles as degrees of freedom, which means that the local minimization must be performed on a reduced

set of degrees of freedom. But this strategy is used for the local optimization of the structure obtained at the end of the IGLOO method.

Constrained local minimization using conjugated gradient: Therefore, the solution was to constrain all degrees of freedom except the dihedral angles. This solution was implemented during the thesis in the deMonNano code. To illustrate, consider an angle noted ABC, which represents a degree of freedom of the molecule. This angle could be subtracted from its contribution to the gradient.

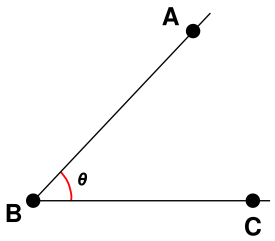


Figure 2.4: Illustration of the angle ABC.

Let us consider the angle ABC. The cosine of this angle is given by the following equation:

$$\cos(\theta) = \frac{\vec{BA} \cdot \vec{BC}}{||\vec{BA}|| \cdot ||\vec{BC}||} = f(\vec{R_A}, \vec{R_B}, \vec{R_C}) = constant \tag{2.14}$$

where $\vec{R_A}$, $\vec{R_B}$ and $\vec{R_C}$ are the positions of the atoms A, B and C. θ is the angle between the vectors $\vec{BA}$ and $\vec{BC}$ and has to be constant in order to constrain this degree of freedom. It should be noted that this method can be employed for other degrees of freedom. For instance, the cosine function can be replaced by a bond length or a cosine function on dihedral angles (this is implemented within the deMonNano code). For a molecule containing angles, the Langragian is given by the following equation:

$$\mathcal{L} = E - \lambda_1(f_1 - \tilde{f}_1) - \lambda_2(f_2 - \tilde{f}_2) - \cdots - \lambda_n(f_n - \tilde{f}_n) \tag{2.15}$$

where E is the total energy of the system, λ_i is the multiplier, $\tilde{f}_i$ is the targeted value of the energy of the system without modification on the angle i and f_i is the value of the energy value with a modification on the angle i in the current conformation. The Langragian of the system is reduced by the contribution of each deviation from a constrained angle. The gradient of the energy with respect to the angle ABC is then given by:

$$F = \frac{\partial \mathcal{L}}{\partial R} = F_R - \lambda_1 \frac{\partial f_1}{\partial R} - \lambda_2 \frac{\partial f_2}{\partial R} - \cdots - \lambda_n \frac{\partial f_n}{\partial R} \tag{2.16}$$

where F_R is the gradient of the energy at a minimization step and $\frac{\partial f}{\partial R}$ is the derivative of the energy respecting the coordinates. This derivative is computed using the method of finite differences.

For each angle, the deviation between the targeted value and the current value is computed and can be noted as:

$$\begin{aligned}
\Delta\cos(\theta_1) &= (f_1 - \tilde{f}_1) = \left|\frac{\partial f_1}{\partial R}\right| \left(-\lambda_1 \left|\frac{\partial f_1}{\partial R}\right| - \lambda_2 \left|\frac{\partial f_2}{\partial R}\right| - \cdots - \lambda_n \left|\frac{\partial f_n}{\partial R}\right|\right) \\
\Delta\cos(\theta_2) &= (f_2 - \tilde{f}_2) = \left|\frac{\partial f_2}{\partial R}\right| \left(-\lambda_1 \left|\frac{\partial f_1}{\partial R}\right| - \lambda_2 \left|\frac{\partial f_2}{\partial R}\right| - \cdots - \lambda_n \left|\frac{\partial f_n}{\partial R}\right|\right) \\
&\cdots \\
\Delta\cos(\theta_n) &= (f_n - \tilde{f}_n) = \left|\frac{\partial f_n}{\partial R}\right| \left(-\lambda_1 \left|\frac{\partial f_1}{\partial R}\right| - \lambda_2 \left|\frac{\partial f_2}{\partial R}\right| - \cdots - \lambda_n \left|\frac{\partial f_n}{\partial R}\right|\right)
\end{aligned} \tag{2.17}$$

Then, each λ_i can be obtained by solving the following equation system:

$$\begin{bmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & \ddots & \vdots \\ A_{n,1} & \dots & A_{n,n} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} \Delta\cos(\theta_1) \\ \vdots \\ \Delta\cos(\theta_n) \end{bmatrix} \tag{2.18}$$

where $A_{i,j} = \left|\frac{\partial f_i}{\partial R}\right| \left|\frac{\partial f_j}{\partial R}\right|$. Obtaining the values of the λ_i allows the minimization of the energy by substracting the contribution of the deviation from the constrained degrees of freedom using equation 2.15. Given the constraints of the system under consideration, which involve only dihedral angles, this approach is not well-suited to addressing the problem at hand. The test was carried out on a phthalate molecule, where an energy gradient lower than 10^{-3} Hartree was unattainable, due to the repercussions of the error affecting an excessive number of constrained degrees of freedom.

Dihedral angles minimization using conjugated gradient: To overcome this issue, a local minimization procedure was implemented that only considers dihedral angles. This method was made by computing the numerical gradient (Equation 2.19) of each flexible dihedral angles and then performing a local minimization using the conjugate gradient method.

$$\frac{\partial E}{\partial \theta_i} = \frac{E(\theta_i + \delta) - E(\theta_i - \delta)}{2\delta} \tag{2.19}$$

In this context, the value of the parameter δ is relatively small. This approach was implemented and tested during this thesis on phthalate structures. This method allows for the convergence to a local minimum. This method is interesting to explore the energy landascape of a constrained surface, and could be easly adapted to other degrees of freedom as bond length, angles, etc. However, due to the formalism of the numerical gradient, minimization is time-consuming, particularly in relation to the number of considered angles.

Dihedral angles minimization using Monte Carlo scheme: The solution presented in the thesis relies on a Monte Carlo scheme coupled to single-point energy computations using DFTB. This method relies on random displacement of the dihedral angles and then compute the energy of the new conformation. If the energy is lower

than the previous one, the new conformation is kept. Otherwise, the new conformation is kept with a probability of $e^{-\frac{\Delta E}{kT}}$ where ΔE is the difference of energy between the new and the previous conformation, k is the Boltzmann constant and T is a temperature (user defined parameter). The temperature is set to value permitting to explore low energy regions. This method offers a satisfactory balance between precision and efficiency. This method is employed in the local minimization step of the IGLOO algorithm. Furthermore, minimization steps are parallelized in the IGLOO code, which allows for a significant reduction of the computational time. The parallelization is performed using the OpenMP library, which is a shared-memory parallelization library. The library permits the splitting of minimization of each structure found at the exploration on many threads, the number of which is defined by the user.

2.4 Application to the alanine dipeptide

To illustrate the method, we applied the IGLOO/DFTB coupling to the alanine dipeptide molecule. The alanine dipeptide is a small peptide and is a well-known benchmark molecule for the exploration of the conformational space of peptides [93]. The alanine dipeptide has two main flexible dihedral angles, presented in the figure 2.5.

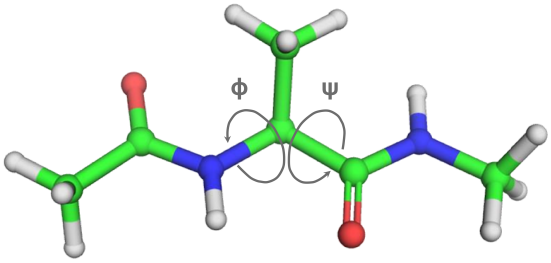


Figure 2.5: Alanine dipeptide molecule.

The PES of the alanine dipeptide is shown in Fig. 2.6 and was computed using a one-degree resolution. Each point was obtained using the zeroth-order DFTB formalism (matsci parameter from [114]). Given that the molecule has a low dimensional exploration space with only two degrees of freedom, basins could be found only by searching on the two dimentionals plot of the PES. However, in the case of a high-dimensional space, such methods will not be sufficient to identify the low-energy regions. Sampling was made using the IGLOO/DFTB method, and the lowest energy conformations were found. Only three exploration steps are shown in the Fig. 2.6. At each iteration, the exploration is conducted in the vicinity of energetic basins, thereby demonstrating the principal tenet of IGLOO, namely, that the method progressively approaches the low-energy regions. Three low-energy conformations were identified, corresponding to the three principal minima of the alanine dipeptide potential energy surface in the vacuum. The exploration of the alanine dipeptide potential energy surface exemplifies the efficiency of the IGLOO/DFTB coupling in exploring the conformational space of molecules

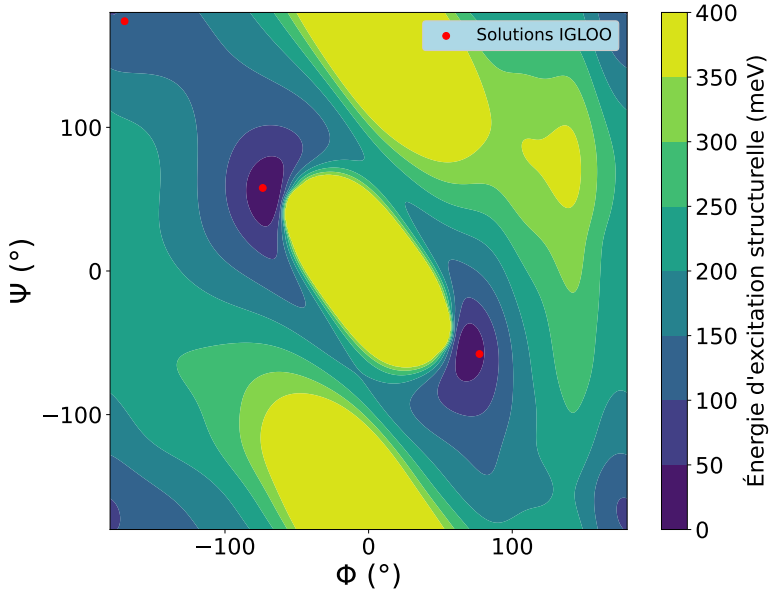


Figure 2.6: Alanine dipeptide PES exploration.

and identifying the lowest-energy conformations. The method is efficient and provides a satisfactory compromise between accuracy and computational cost.

2.5 Conclusion

In the present chapter, we have reported the coupling of a non-redundant conformational space exploration algorithm named IGLOO with the quantum chemical DFTB potential. Differents solutions were considered for the local minimization of the conformations generated by the exploration phase of IGLOO: an all-atom minimization approach, a constrained all-atom minimization approach and a local minimization procedure that only considers dihedral angles. These methods were not well suited for the problem at hand, but could be utilized for other application cases, for example for restrained a bond length of a molecule, or a minimization scheme restrained to a specific angle. Finally, a Monte Carlo minimization scheme on dihedral angles coupled to single-point energy computation using DFTB was retained for its efficiency and its balance between precision and efficiency. The method was applied to the alanine dipeptide molecule, and the results demonstrated the efficiency of the IGLOO/DFTB coupling in exploring the conformational space of molecules and identifying the lowest-energy conformations in a simple conformational space. The coupling reported will be applied to phthalate molecules to explore the conformational space of more complex molecules in the next chapter.

Chapter 3

Exploration of the potential energy surface of phthalate molecules

Contents

3.1 Introduction

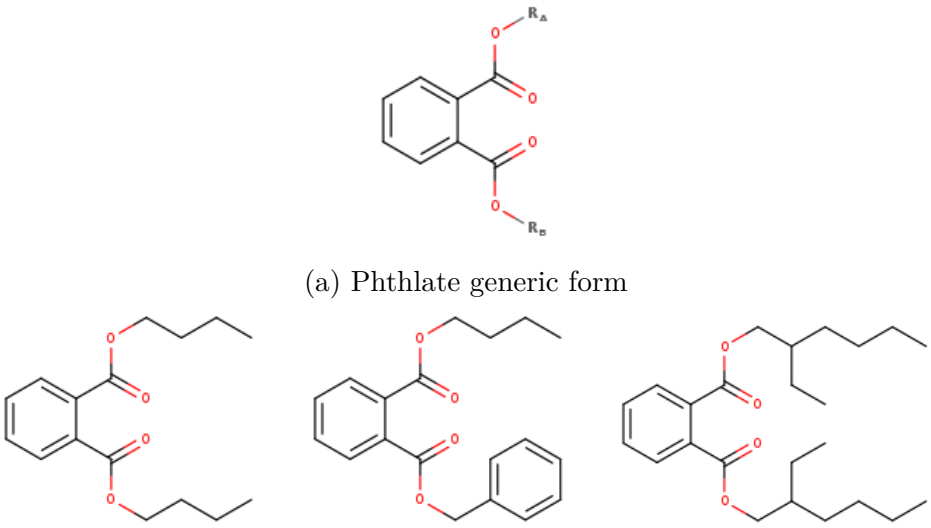
In this chapter, we consider three molecules of the phthalate family as an example of the application of the presented IGLOO/DFTB coupling. The energy landscapes of the selected structures show numerous degenerate basins where our exploration scheme could demonstrate its efficiency. The three investigated phthalate molecules are introduced, accompanied by a detailed presentation of the descriptors utilized in the study. Next, the structural excitation spectra of the molecules are studied, followed by an analysis of the conformational energy landscape, which shows a different behavior of the three molecules.

3.1.1 Phthalates family

Phthalates are commonly used in many consumer products such as PVC, coatings, adhesives, perfumes and cosmetics due to their plasticising properties [7, 149]. For example, phthalates prevent nail varnish from chipping, make perfumes last longer, make tool

handles stronger and more resistant, and increase the effectiveness of adhesives. Their global production is estimated at 3 million tonnes per year [108]. These molecules can be released into the environment during the production, use and disposal of products containing them, and these compounds can be found in water, air, soil and sediment. They have been associated with a variety of adverse effects on human health [83] depending on many factors, including dose, duration and route of exposure. In particular, they can act as endocrine disruptors by interfering with the natural hormones in the human body. Such effects were found in the premature development of breasts in Puerto Rican girls [39]. Some phthalates have been associated with reduced sperm quality in men, as well as birth defects in infants exposed in utero [95]. In this work, we focused on three representative phthalate molecules, namely the DEHP (di(2-ethylhexyl)) phthalate, which represents 50 % of the global phthalate production, the BBP (benzyl butyl) phthalate, and the DBP (dibutyl) phthalate which represent about one tenth of the DEHP production together. To our knowledge, no exhaustive exploration of their energy landscape has been reported in the literature.

Phthalates are esters of phthalic acid composed of an aromatic benzene ring with two ester groups on ortho position. They differ between each other by their terminal groups. The generic form of phthalates is shown in Fig. 3.1-(a) together with the three phthalate molecules explored in the present chapter, namely DBP (Fig. 3.1-(b)), BBP (Fig. 3.1-(c)) and DEHP (Fig. 3.1-(d)). DBP and DEHP have two identical terminal groups, while BBP has two different ones, a 4-carbon alkyl and a phenyl-terminated chain.



(a) Phthlate generic form

(b) DBP (c) BBP (d) DEHP

Figure 3.1: Generic form of phthalate (a) and the three molecules investigated in this work: dibutyl phthalate DBP (b), benzyl butyl phthalate BBP (c) and di(2-ethylhexyl) phthalate DEHP (d). Carbon skeleton in black and oxygen atoms in red. $R_{A,B}$ represent terminal groups of each side-chain.

3.1.2 Methodology

IGLOO was initialized with one hundred initial states for each molecule. This number was chosen to be large enough to ensure exhaustive exploration of conformational space.

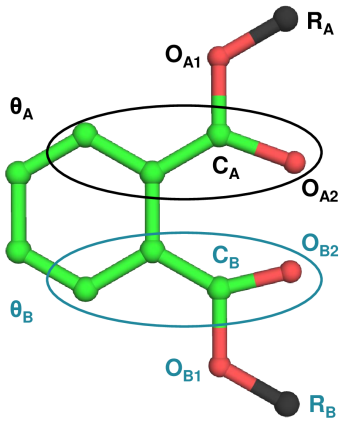
Ten independent runs of the IGLOO/DFTB coupling were performed for each molecule. Due to the stochastic nature of the exploration method, differences between runs are to be expected. However, the algorithm showed very good reproducibility in the case of BBP and DBP, finding nearly the same low-energy minima in each run. In the case of DEHP, due to the higher dimensionality and the multiplicity of possible intramolecular interactions, several new minima were generated with each additional run. Consequently, the algorithm was run ten more times (twenty in total). No new minima were discovered in the last few runs, which is reassuring in terms of exploration convergence. The data analyzed below are the concatenated results of all runs of the algorithm for each molecule.

Energy computations were carried out using the third-order DFTB formalism (3ob parameters)[71, 72], combined with an empirical dispersion correction [174]. A Fermi temperature of 100K was introduced to avoid convergence issues during the self-consistent scheme. Local minimizations were performed using an all-atom conjugated gradient technique, as detailed in the chapter 2. In order to validate the DFTB parameters and the dispersion correction, DFTB and DFT minima were compared on a structural and energetic basis (see section Structural and energy comparison between DFT and DFTB potentials). DFT calculations were performed using the Gaussian 16 set of programs[68]. The high-nonlocal and hybrid meta exchange-correlation M06–2X functional[173] was used together with a 6-311++G(d,p) basis set. This combination was chosen as it has been previously shown to describe phthalates energetics with satisfactory performance [129]. Structural parameters result from full geometry optimization in the gas phase, with no imposed constraint. Default SCF and geometry optimization criteria were used.

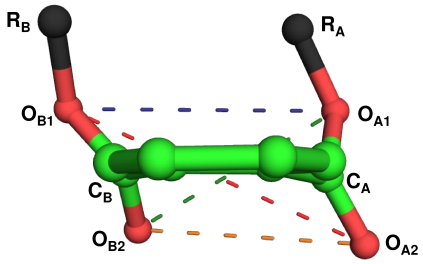
3.2 Geometric descriptors

The local minima obtained from the global exploration performed with the IGLOO/DFTB coupling were analysed on the basis of both energetic and structural descriptors. Descriptors characteristic of the relative orientation/organisation of the side-chains were defined from dihedral angles and interatomic distances (Fig. 3.2). Two dihedral angles were defined to describe the connection of the side-chains to the central aromatic ring (θ_{A} and θ_{B} in Fig. 3.2-(a)) as well as four oxygen-oxygen distances involving atoms belonging to two different chains: $d_{\mathrm{O_{A1}\text{-}O_{B1}}}$, $d_{\mathrm{O_{A2}\text{-}O_{B2}}}$, $d_{\mathrm{O_{A1}\text{-}O_{B2}}}$ and $d_{\mathrm{O_{A2}\text{-}O_{B1}}}$ (Fig. 3.2-(b)). The smallest value between these four distances defines the last descriptor, $dmin_{\mathrm{O\text{-}O}}$. Similarly, the $dmin_{\mathrm{C\text{-}O}}$ (Fig. 3.2-(c)) criterion was defined as the smallest distance between $d_{\mathrm{C_A\text{-}O_{B1}}}$, $d_{\mathrm{C_B\text{-}O_{A1}}}$, $d_{\mathrm{C_A\text{-}O_{B2}}}$ and $d_{\mathrm{C_B\text{-}O_{A2}}}$. In addition, we define (i) $d_{\mathrm{C\text{-}O_1}}$ as the smallest distance between $d_{\mathrm{C_A\text{-}O_{B1}}}$ and $d_{\mathrm{C_B\text{-}O_{A1}}}$ and (ii) $d_{\mathrm{C\text{-}O_2}}$ as the smallest distance between $d_{\mathrm{C_A\text{-}O_{B2}}}$ and $d_{\mathrm{C_B\text{-}O_{A2}}}$. Note that A and B are replaced by

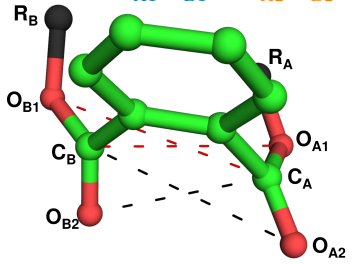
Bu (for butyl chain) and Be (for benzyl chain) for BBP because the chains are different and the atoms are not equivalent from one chain to the other.



(a) Dihedral descriptors: θ_{A} and θ_{B}



(b) O-O distance descriptors: $d_{\mathrm{O_{A1}\text{-}O_{B1}}}$, $d_{\mathrm{O_{A2}\text{-}O_{B2}}}$, $d_{\mathrm{O_{A1}\text{-}O_{B2}}}$ and $d_{\mathrm{O_{A2}\text{-}O_{B1}}}$

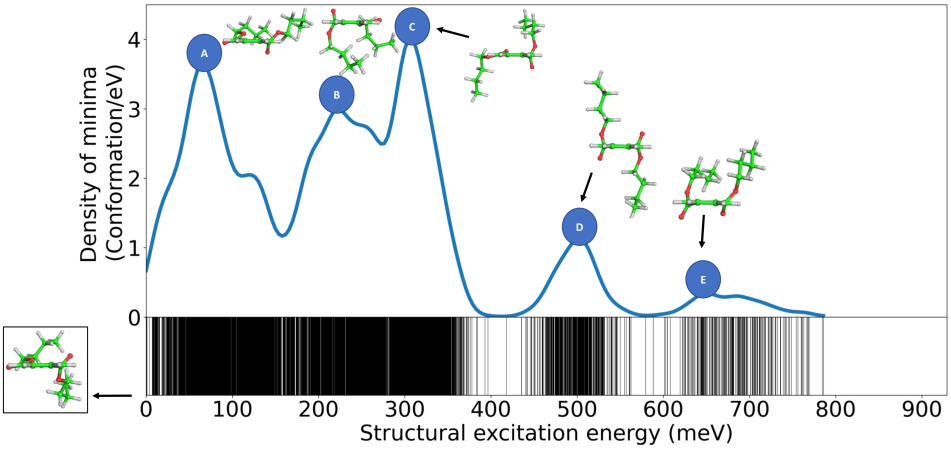


(c) C-O distance descriptors: $d_{\mathrm{C_A\text{-}O_{B1}}}$, $d_{\mathrm{C_B\text{-}O_{A1}}}$, $d_{\mathrm{C_A\text{-}O_{B2}}}$, $d_{\mathrm{C_B\text{-}O_{A2}}}$

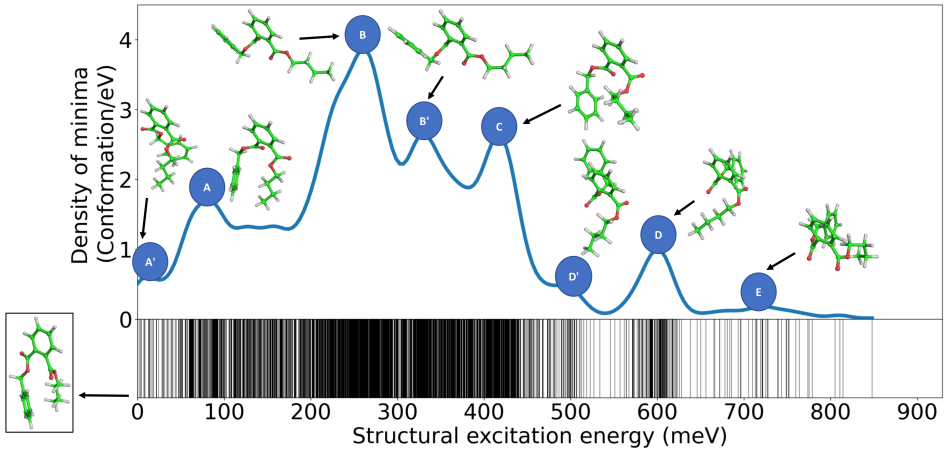
Figure 3.2: Structural descriptors: (a) dihedral angles and (b)-(c) interatomic distances. Carbon atoms are in green and oxygen atoms are in red. R balls in black represent the terminal group of each side chain.

3.3 Structural excitation spectra

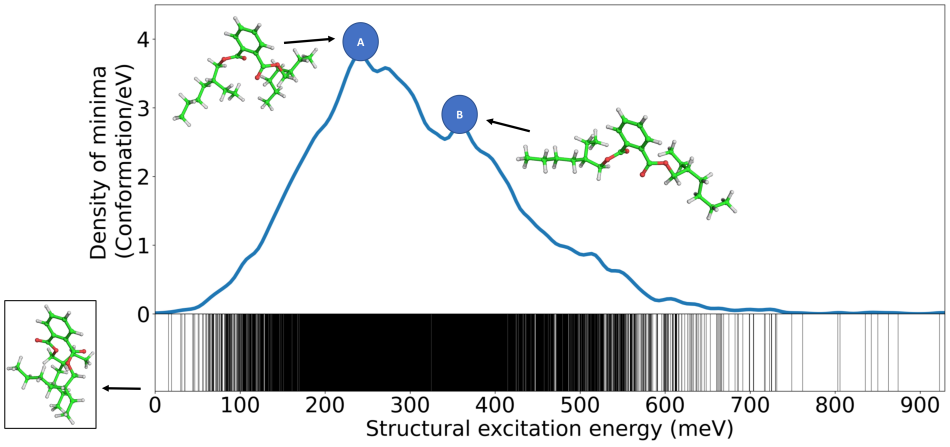
The structural excitation energy spectra obtained for the three investigated molecules are presented in Fig. 3.3, the zero energy reference being that of the global minimum. In the lower panels, each local minimum identified on the PES is represented by a bar. In the upper panels, these bar spectra are convoluted with a gaussian kernel in



(a) DBP



(b) BBP



(c) DEHP

Figure 3.3: Structural excitation energy spectra (bar and estimated density). For each molecule, main peaks are illustrated by their characteristic structures (multiple geometry could coexist in a peak). Global minima are depicted in the insets.

order to provide an estimation of the isomers density as a function of the energy. In the explorations, the highest energy isomers are located at 786 meV (DBP), 848 meV (BBP) and 929 meV (DEHP) above their respective global minimum. One should keep in mind that the present scheme aims at exploring the low energy regions of the PES, and the exploration of the highest energy regions is expected to be less exhaustive. As a result, the calculated isomers density at high energy is expected to be lower than the exact one. First, the general aspect of the isomer density distributions differ for the three molecules. These spectra can be characterized from the differences regarding their alternation of energy ranges exhibiting high/low isomer densities. For the DBP molecules (Fig. 3.3-(a)), the region of highest density extends from the global minimum energy to around 385 meV above it, with three peaks at 67 (peak **A**), 221 (peak **B**), and 308 meV (peak **C**). Another high-density peak is observed at 501 meV (peak **D**) surrounded by energy regions where almost no isomer was found. A final peak is identifiable at 642 meV (peak **E**), made up of few high-energy isomers. For the BBP molecule (Fig. 3.3-(b)), the isomer density is relatively low above the global minimum up to above 50 meV. In the 50-450 meV energy range in which the majority of isomers is found, three main high-density peaks are observed at around 81 meV, 260 meV, 418 meV (peaks **A**,**B**,**C**) and their variant (resulting from structures presenting small geometric variations with respect to those dominating the main peak) noted by **'** at 20 meV and 328 meV (peaks **A'** and **B'**). At higher energies, the density is low except around 601 meV and 727 meV (peaks **D** and **E**) and the variant at 504 meV (peak **D'**) where other peaks are observed. For the DEHP molecule, the general aspect of the isomer density distribution (Fig. 3.3-(c)) is drastically different from the two previous ones. A unique and extended high density region is observed between 50 and 550 meV above the global minimum energy, the largest isomer density being located around 240 meV (peak **A**). A minor shoulder is also observed at 373 meV (peak **B**).

Fig. 3.3 also shows the lowest-energy structure and isomers located at the maximum of the density peaks for each molecule, pointing out the conformational variability. Nevertheless, one should be aware that they may not be representative of all the isomers making up the corresponding peak, since various structural patterns can contribute to a given peak. Deriving general trends would require a deeper analysis of the interplay between characteristic structural patterns and energies that will be addressed in the following sections.

3.4 Structure-energy relationships

3.4.1 Distance based analysis

Although the three molecules are part of the same family, they exhibit different behaviours from an energetic point of view due to their terminal groups. Indeed, as the two alkyl chains are small in the DBP molecule, the interaction between the oxygen atoms is expected to drive the molecular energetics. The longer and ramified alkyl side-chains in DEHP would give rise to steric hindrance and to multiple possibilities for stabilisation through dispersive interactions. The interaction between the alkyl and aryl

chains in the BBP molecule can generate structures stabilized by Coulomb interaction between the negatively charged aromatic carbon atoms and the positively charged hydrogen atoms of the butyl chain. In order to investigate relationships between the isomers energies and their structures, we first focus on the distances between oxygen-oxygen and carbon-oxygen atoms belonging to different side-chains.

On the pie charts of Fig. 3.4, the isomers are classified in four families depending on which of the $d_{\mathrm{O_{A1}\text{-}O_{B1}}}$, $d_{\mathrm{O_{A2}\text{-}O_{B2}}}$, $d_{\mathrm{O_{A1}\text{-}O_{B2}}}$ or $d_{\mathrm{O_{A2}\text{-}O_{B1}}}$ distances is the smallest one ($dmin_{\mathrm{O\text{-}O}}$). It appears that the molecules with small side-chains (DBP and BBP) exhibit similar distributions. The $d_{\mathrm{O_{A2}\text{-}O_{B2}}}$ is rarely the smallest one, i.e. the smallest distance always involves at least one of the side-chains connected oxygens $\mathrm{O_{A1}}$ or $\mathrm{O_{B1}}$, and the isomer population is equally shared in three groups by the attribution of $dmin_{\mathrm{O\text{-}O}}$ to either $d_{\mathrm{O_{A1}\text{-}O_{B1}}}$, $d_{\mathrm{O_{A1}\text{-}O_{B2}}}$ or $d_{\mathrm{O_{A2}\text{-}O_{B1}}}$. The pie chart of the DEHP molecule is different, with four almost similar quartiles. In this case, the smallest distance is attributed to $d_{\mathrm{O_{A2}\text{-}O_{B2}}}$ in one case out of five, this increase of occurrence being probably due to steric hindrance that prevents the oxygen atoms $\mathrm{O_{A1}}$ and $\mathrm{O_{B1}}$ from approaching each other.

On the pie charts of Fig. 3.5, the isomers are classified in two families depending on which of the $d_{\mathrm{C\text{-}O_1}}$ or $d_{\mathrm{C\text{-}O_2}}$ distances is the smallest one ($dmin_{\mathrm{C\text{-}O}}$). In the case of BBP, $d_{\mathrm{C\text{-}O_1}}$ and $d_{\mathrm{C\text{-}O_2}}$ are both split in two subfamilies because $\mathrm{R_A}$ and $\mathrm{R_B}$ are different. For the three molecules, the pie charts are dominated by the family $dmin_{\mathrm{C\text{-}O}} = d_{\mathrm{C\text{-}O_1}}$. This can be interpreted from the fact that O_1 is less negatively charged than O_2, reducing coulomb repulsion with the slightly negatively charged COO function. DBP appears to have a higher contribution of $d_{\mathrm{C\text{-}O_1}}$ than the other two molecules.

The scatter plots in Figs. 3.4 and 3.5 allow to correlate oxygen-oxygen and carbon-oxygen distances with energy distributions. It appears that for DBP and BBP, several distinct high density regions can be identified regarding the energy correlation with either the values of $dmin_{\mathrm{O\text{-}O}}$ and $dmin_{\mathrm{C\text{-}O}}$ or the density of minima per oxygen-oxygen subgroup. In the case of DBP, the five different regions appear clearly, and they can be easily identified by clustering method, as can be done for instance with a k-means method (Fig. 3.8). These five regions correspond to different dominant O-O and C-O interactions. As expected due to the presence of identical side-chains, a good superposition of the $dmin_{\mathrm{O\text{-}O}}$ cross-interaction curves is observed (red and green curves Fig. 3.4), which reassures us about the quality of the exploration. We remind that the lowest energy region of the structural excitation spectrum (Figs. 3.3 and 3.6-(a)) corresponds to the peak **A** (from the most stable structure up to $\sim$ 150 meV). In these structures, the smallest distances between atoms from the COO functions always involve a positively charged carbon atom (with a charge of $\sim$ 0.66e) from one chain and a negatively charged oxygen atom $\mathrm{O_1}$ (with a charge of $\sim$ -0.35e) from the other chain (Fig. 3.6-(a)), consequently, $dmin_{\mathrm{C\text{-}O}}$ systematically involve an $\mathrm{O_1}$ atom ($d_{\mathrm{C-O_1}}$). This is consistent with the fact that the COO group is globally negatively charged (about -0.25 e) and that the charge carried by $\mathrm{O_1}$ atom is about -0.35 e whereas that of the $\mathrm{O_2}$ atom is about -0.56 e. For isomers belonging to peak **A**, the alkyl chains remains close even if they are from either side of the central phenyl group. Moreover, structures containing two $\mathrm{O_1}$ atoms pointing towards the carbon atom of the other chain COO group (highlighted by $dmin_{\mathrm{O\text{-}O}} = d_{\mathrm{O_{A1}-O_{B1}}}$) are particularly stable and gathered in the low energy region of

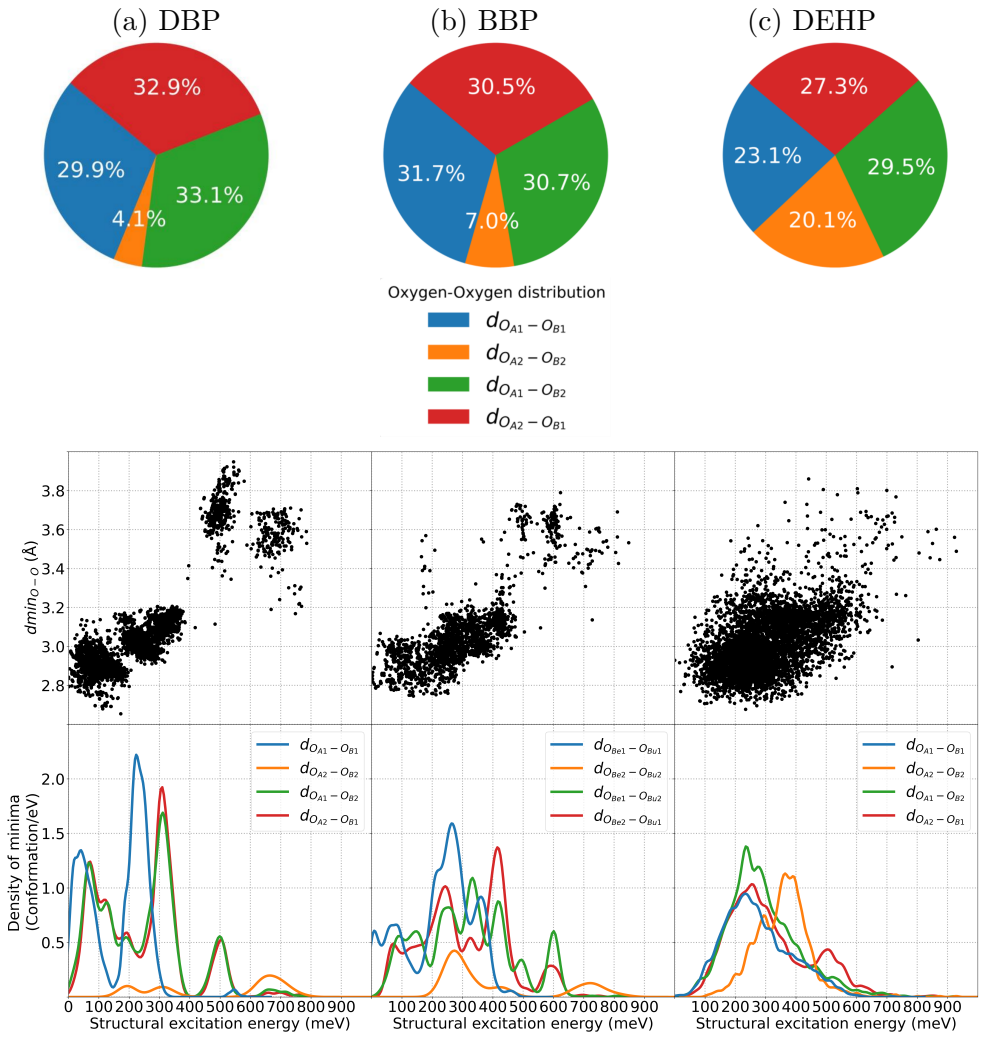


Figure 3.4: Global distribution of $dmin_{\text{O-O}}$ for all minima represented as a pie chart. The scatter plots report the $dmin_{\text{O-O}}$ values for each conformation and the curves represent the relative density of the conformations, both depicted as a function of the structural excitation energy. Isomers are classified according to the nature of $dmin_{\text{O-O}}$ (blue for $dmin_{\text{O-O}} = d_{\text{OA1-OB1}}$; orange for $dmin_{\text{O-O}} = d_{\text{OA2-OB2}}$; green for $dmin_{\text{O-O}} = d_{\text{OA1-OB2}}$ and red for $dmin_{\text{O-O}} = d_{\text{OA2-OB1}}$).

the peak **A** (i.e. below $\sim$ 50 meV, see Fig. 3.4-(a)). The second and third peaks (**B** and **C**) are mostly composed of structures where the two planes containing the COO groups are perpendicular to each others (Fig. 3.6-(b/c)). One COO group is parallel to the central phenyl and its O_1 (in peak **B**) or its O_2 (peak **C**) atom is involved in $dmin_{\text{C-O}}$. Again, the charge differences between O_1 and O_2 could explain the energy ranking between the two peaks. In addition, the induced relative orientations of the side-chains in peak **C** result in less favourable interactions between the two alkyl chains (Fig. 3.6-(c)). In the last two peaks (**D** and **E**), the two planes containing the COO groups are perpen-

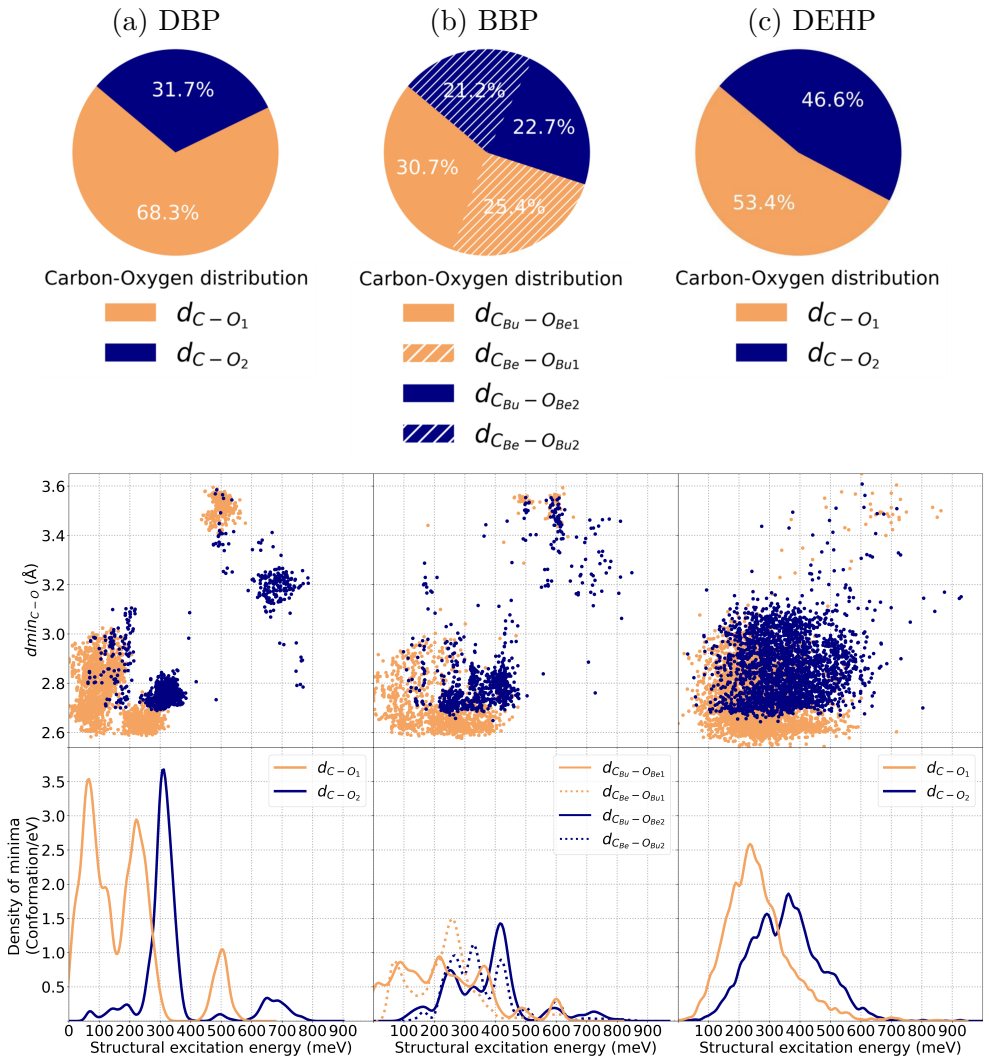
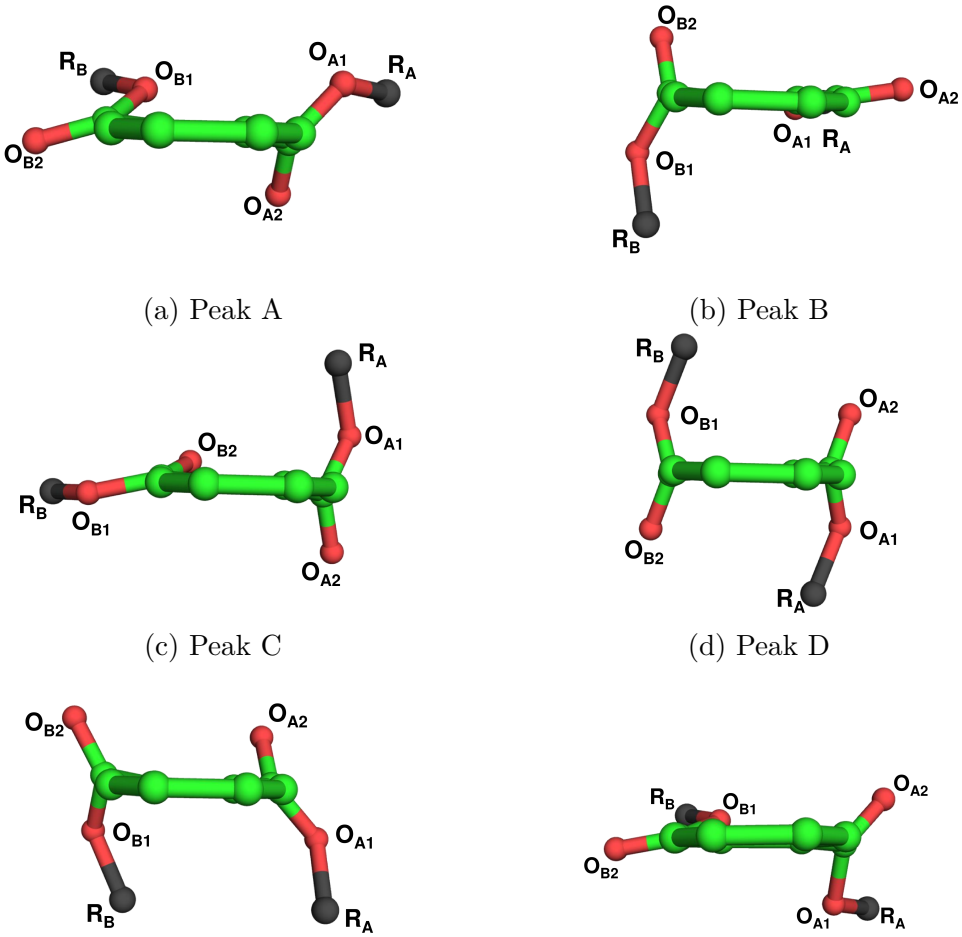


Figure 3.5: Global distribution of $dmin_{\text{C-O}}$ for all minima represented as a pie chart. The scatter plots report the $dmin_{\text{C-O}}$ values for each conformation and the curves represent the relative density of the conformations, both depicted as a function of the structural excitation energy. Isomers are classified according to the nature of $dmin_{\text{C-O}}$ (orange for $dmin_{\text{C-O}} = d_{\text{C-O}_1}$; blue for $dmin_{\text{C-O}} = d_{\text{C-O}_2}$).

dicular to the central phenyl, resulting in larger values for $dmin_{\text{C-O}}$ and $dmin_{\text{O-O}}$ than reported for the other peaks (Figs. 3.6-(d/e), 3.4-(a) and 3.5-(a)). These peaks differ by $dmin_{\text{O-O}} = d_{\text{O}_{A1}\text{-O}_{B2}}$ or $dmin_{\text{O-O}} = d_{\text{O}_{A2}\text{-O}_{B1}}$ in peak **D** or $dmin_{\text{O-O}} = d_{\text{O}_{A2}\text{-O}_{B2}}$ in peak **E**. This means that the side-chains are pointing in opposite (resp. similar) directions in peak **D** (resp. peak **E**). Although stabilizing interactions between the alkyl chains are almost absent in peak **D** structures whereas they are present in peak **E**, the steric hindrance between these chains results in shortening the distances between O_{A2} and O_{B2} (as can be seen from $dmin_{\text{O-O}} = d_{\text{O}_{A2}\text{-O}_{B2}}$) and therefore increasing the coulomb



(a) Peak A

(b) Peak B

(c) Peak C

(d) Peak D

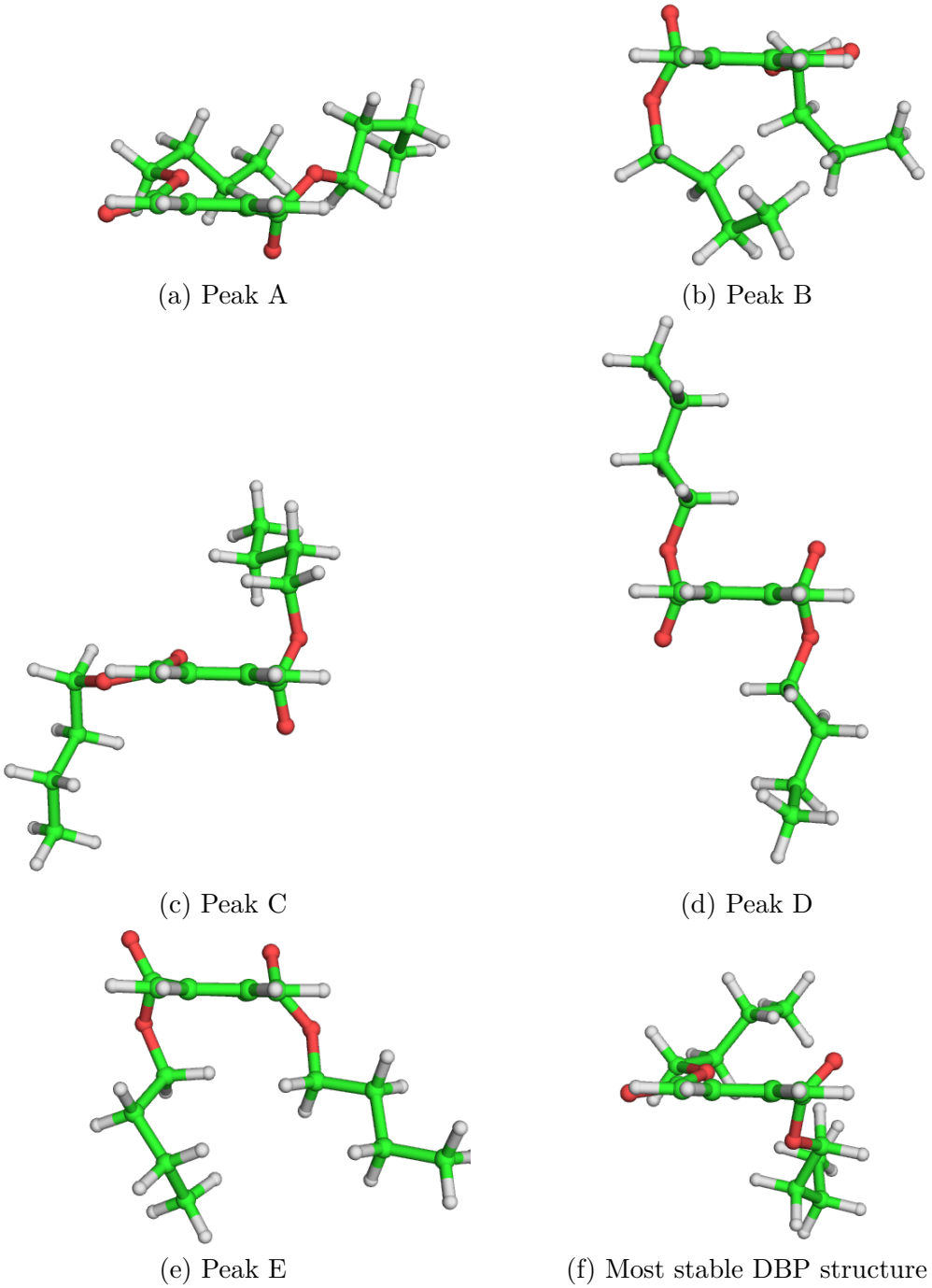
(e) Peak E

(f) Most stable DBP structure

Figure 3.6: Illustration of the ester groups relative orientation for characteristics structures of the DBP structural excitation energy spectrum main peaks (Fig. 3.3-(a)). The illustration depicts the relative orientation of the oxygen atoms. The DBP main peaks are also illustrated with their sidechains (Fig. 3.7).

repulsion between these oxygen atoms.

The structural excitation spectrum of the BBP molecule is trickier to interpret. The structures between the global minimum (Fig. 3.9-(f)) and 100 meV give birth to two peaks **A** and **A'** (Fig. 3.9-(a/a')) for which the $dmin_{\text{C-O}}$ distance always involves an O_1 atom, which minimizes the Coulomb repulsion (note that the charges of the oxygen atoms are similar for the three molecules). The COO functions are parallel to each other for peak **A'** structures (with $dmin_{\text{C-O}}$ involving the C of the butyl chain noted C_{Bu} and O_1 of the benzyl chain noted O_{Be1}) whereas peak **A** structures contain a mix of parallel and perpendicular structures (with a mix of $dmin_{\text{C-O}} = d_{\text{C}_{\text{Bu}}\text{-O}_{\text{Be1}}}$ and $dmin_{\text{C-O}} = d_{\text{C}_{\text{Be}}\text{-O}_{\text{Bu1}}}$). Bringing the less negatively charged oxygen atoms O_1 closer together, as in peak **A'**, maximizes chain interactions and minimizes oxygen-oxygen



(a) Peak A

(b) Peak B

(c) Peak C

(d) Peak D

(e) Peak E

(f) Most stable DBP structure

Figure 3.7: Illustration of the side-chains relative orientation for characteristic structures of the main peaks of the DBP structural excitation energy spectrum.

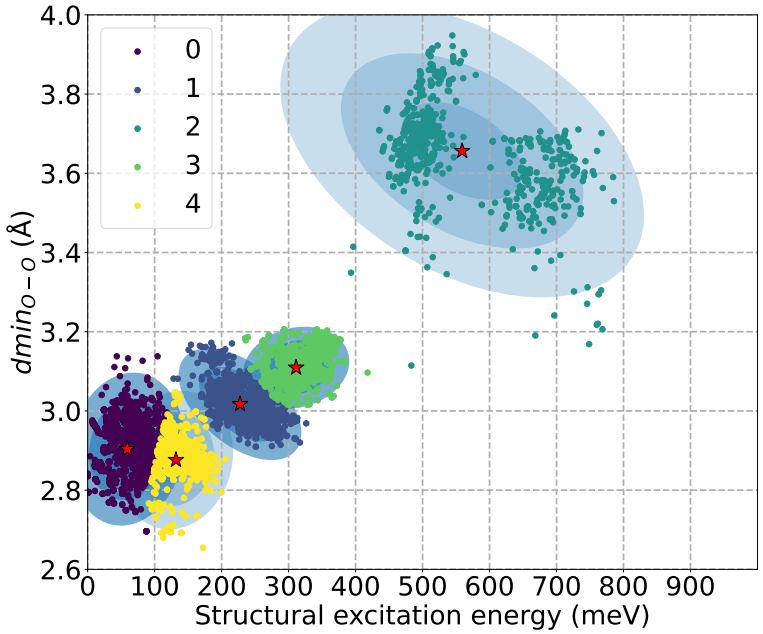
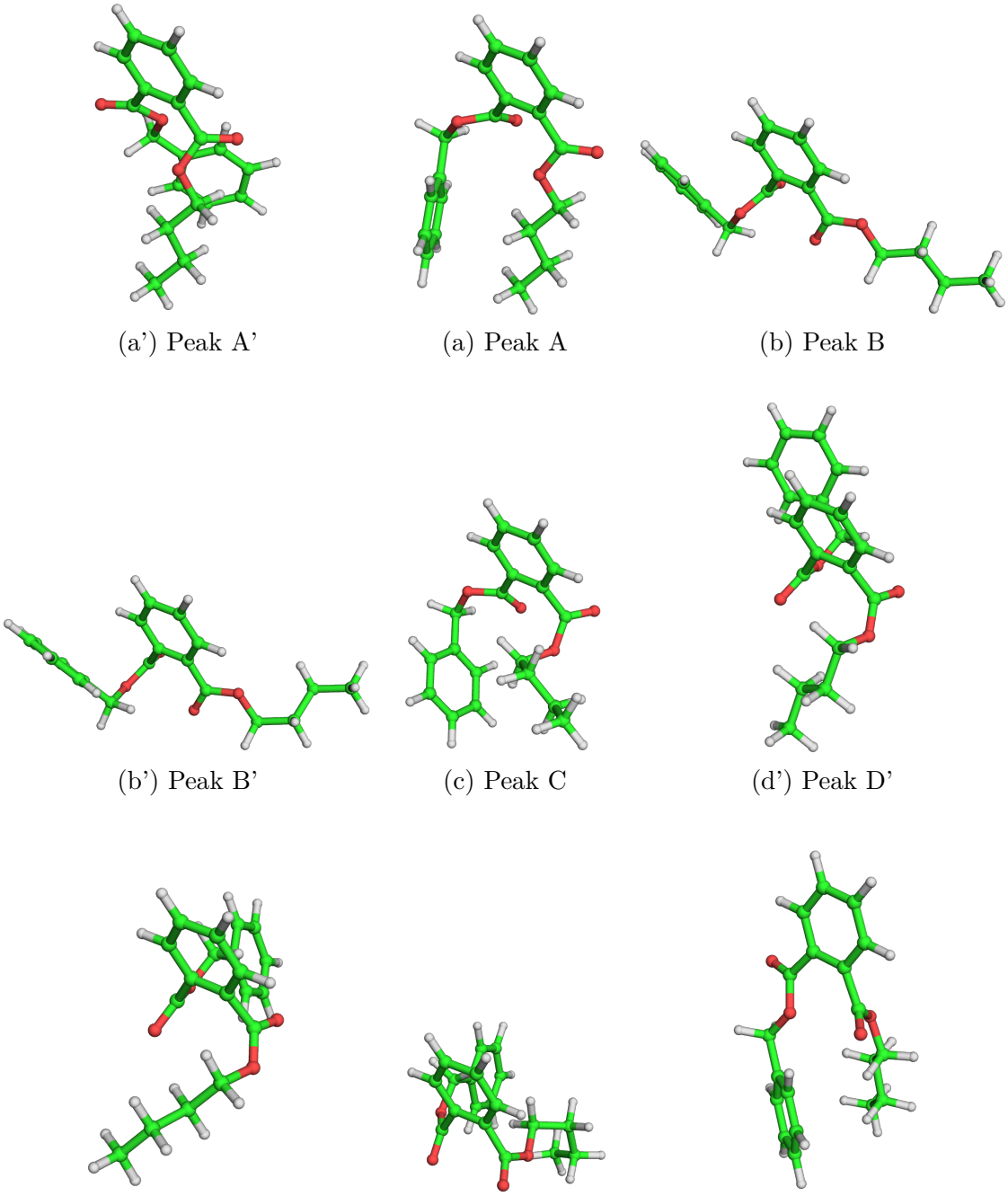


Figure 3.8: Clustering by k-means method of the point cloud corresponding to the plot of the $dmin_{\text{O-O}}$ distance of the DBP molecule as a function of its structural excitation energy. Red stars and blue ovals represent the center and the covariance of each cluster, respectively.

Coulomb repulsion. This explains the location of peak **A'** at lower energy than peak **A**. Note that, in addition to the peak **A'** features, the global minimum structure maximizes the interactions between the hydrogen atoms of its alkyl chain with the phenyl group of the other chain (Fig. 3.3-(b)). The structures present in peaks **B** and **B'** (Fig. 3.9-(b/b')) contain many various contributions, as can be seen from the $dmin_{\text{C-O}}$ and $dmin_{\text{O-O}}$ distribution plots (Fig. 3.4-(b) and Fig. 3.5-(b)) and it is therefore challenging to derive simple general trends. We note, however, that the smallest $dmin_{\text{C-O}}$ are obtained for $dmin_{\text{C-O}} = d_{\text{C-O}_1}$ and that $dmin_{\text{O-O}} = d_{\text{O}_{\text{A2}}\text{-O}_{\text{B2}}}$ is minority. Peak **C** (Fig. 3.9-(c)), on the other hand, is similar to peak **C** in DBP, with $dmin_{\text{C-O}}$ involving an O_2 atom and a perpendicular orientation between the two COO functions with one of the latter contained in the plane of the central phenyl. $dmin_{\text{O-O}}$ consists solely of cross interactions with a slight preference for $dmin_{\text{O-O}} = d_{\text{O}_{\text{Be2}}\text{-O}_{\text{Bu1}}}$. Similar patterns are observed for the structures belonging to peaks **D** and **D'** (Fig. 3.9-(d/d')) and those of the peak **D** of DBP, i.e. with two COO planes perpendicular to the central phenyl and side-chains pointing in opposite directions. Structures of the peak **D'** are slightly more stable thanks to the stacking of the two phenyl groups, which is associated to $dmin_{\text{O-O}} = d_{\text{O}_{\text{Be1}}\text{-O}_{\text{Bu2}}}$. Finally, representative structures of peak **E** of BBP (Fig. 3.9-(e)) are very similar to those of peak **E** of DBP. The COO groups planes are perpendicular to the central phenyl and parallel to each other, and the two carbonyl oxygen are on the same side. In this group, the contribution of $dmin_{\text{O-O}} = d_{\text{O}_{\text{A1}}\text{-O}_{\text{A2}}}$ and $dmin_{\text{O-O}} = d_{\text{O}_{\text{A2}}\text{-O}_{\text{A1}}}$ is almost absent in BBP although a minor contribution was observed in the DBP case.



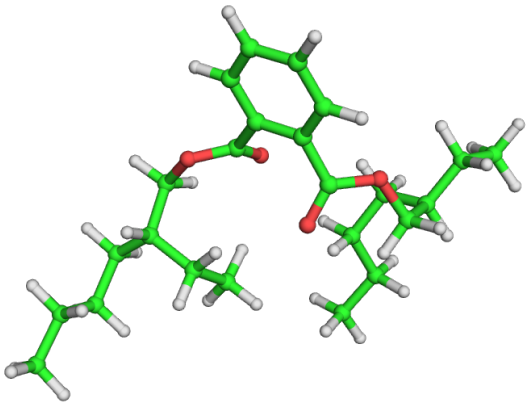
(a') Peak A' (a) Peak A (b) Peak B

(b') Peak B' (c) Peak C (d') Peak D'

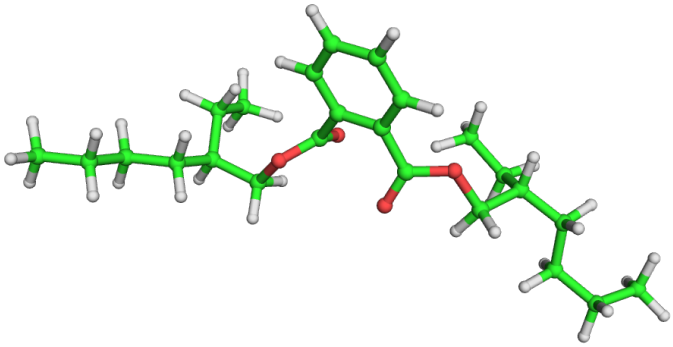
(d) Peak D (e) Peak E (f) Most stable BBP structure

Figure 3.9: Illustration of the side-chains relative orientation for characteristic structures of the main peaks of the BBP structural excitation energy spectrum.

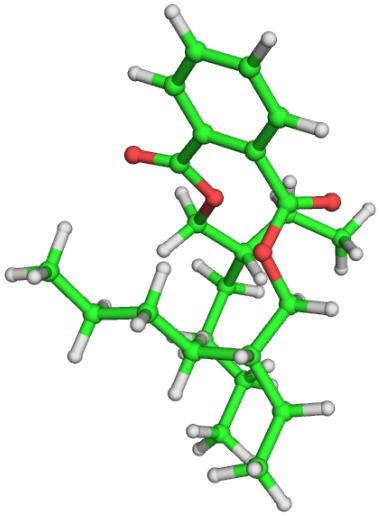
We remind, however, that the exploration could be incomplete at such high energies because the IGLOO scheme is mostly designed to explore the low energy regions of the PES.



(a) Peak A



(b) Peak B



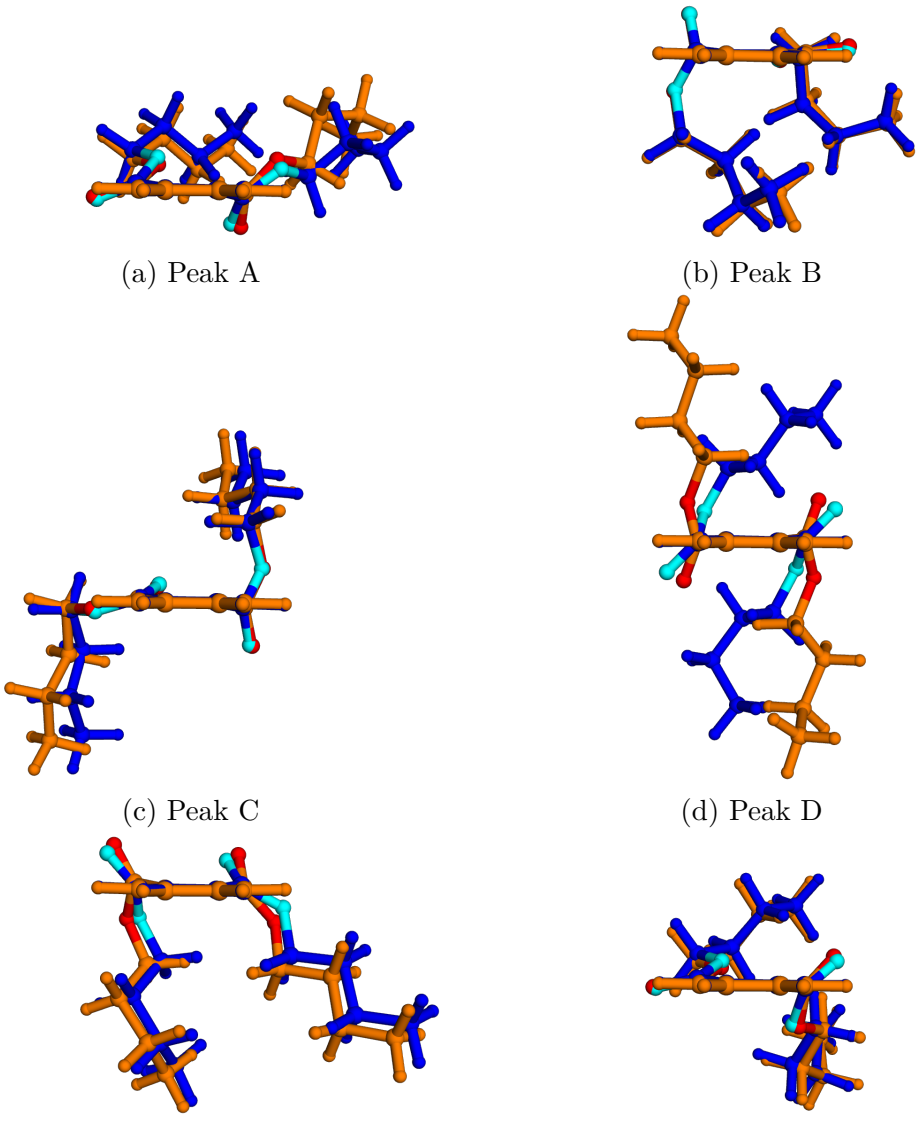
(c) Most stable DEHP structure

Figure 3.10: Illustration of the side-chains relative orientation for characteristic structures of the main peaks of the DEHP structural excitation energy spectrum.

The structural excitation energy curve for DEHP (Fig. 3.3 (c)) is less structured than the two previous ones. The $dmin_{\text{C-O}}$ of the most stable structure (Fig. 3.10-(c)) involves an O_1 atom, with the COO functions perpendicular to each other, and one of them being in the plane of the central ring. In addition, the side-chains are close to each other and on either side of the phenyl plane. From $\approx$ 75 meV, the chains can be placed on the same side of the phenyl plane. Another structure appears from $\approx$ 87 meV with $dmin_{\text{C-O}} = d_{\text{C-O}_2}$. Above this energy, the curve is a single broad distribution up to 600 meV above the global minimum. At least two substructures, noted **A** (Fig. 3.10-(a)) at $\approx$ 240 meV and **B** (Fig. 3.10-(b)) at $\approx$ 380 meV, can be identified. The $dmin_{\text{C-O}}$ and $dmin_{\text{O-O}}$ distributions of DEHP in the in Fig. 3.4-(c) and Fig. 3.5-(c) are large and overlapping. The large number of possible interactions between the long alkyl chains results in a continuum of isomers over the wide energy range for each previously discussed structural feature. While it was, to a large extend, easy to attribute a peak to a given structural characteristic for the two other molecules, the coexistence of structurally distinct isomers at a given energy hinders a detailed analysis of the DEHP structural excitation energy spectrum. Nevertheless, one can mention that peak **A** is dominated by structures exhibiting $dmin_{\text{C-O}} = d_{\text{C-O}_1}$ and an equal repartition between $dmin_{\text{O-O}} = d_{\text{O}_{\text{A1}}\text{-O}_{\text{B1}}}$, $dmin_{\text{O-O}} = d_{\text{O}_{\text{A1}}\text{-O}_{\text{B2}}}$ and $dmin_{\text{O-O}} = d_{\text{O}_{\text{A2}}\text{-O}_{\text{B1}}}$. Note that the slight difference between the $d_{\text{O}_{\text{A1}}\text{-O}_{\text{B2}}}$ and $d_{\text{O}_{\text{A2}}\text{-O}_{\text{B1}}}$ curves may be due to the difficulty of achieving complete exploration of high-dimensional conformational space. The presence of peak **B** is due to structures for which $dmin_{\text{C-O}} = d_{\text{C-O}_2}$ and $dmin_{\text{O-O}} = d_{\text{O}_{\text{A2}}\text{-O}_{\text{B2}}}$. This peak is higher in energy because it combines strong Coulomb repulsion (between the two most negatively charged O_2 atoms and between an O_2 atom and the COO group) and loss of dispersive stabilisation due to a large distance between the chains.

3.4.2 Structural and energy comparison between DFT and DFTB potentials

Additional DFT local minimizations were performed on the representative structures of the main peaks observed in Figure 3.3. Superimpositions of the corresponding DFTB and DFT structures are depicted in Figures 3.11, 3.12 and 3.13 for DBP, BBP and DEHP, respectively. Overall, the structural differences observed are minor except in the case of peak D of DBP, peak B of BBP and peak B of DEHP for which a slight modification of the orientation of the side-chains is observed. This is likely due to the existence of multiple of very close minima in these zones of the DFTB and DFT potential energy surfaces, as reflected by the continuum observed in the regions of these peaks in the DFTB structural excitation energy spectra (Figure 3.3). However, the structures concerned retain their main structural characteristics discussed above. A comparison of the DFTB and DFT energy ranking is given in Figures 3.14, 3.15 and 3.16 for DBP, BBP and DEHP, respectively. The hierarchy of minima is the same between the two methods, with only one inversion observed, between peaks B and C of the DBP. These results fully support our strategy of globally exploring the potential energy surfaces of



(a) Peak A

(b) Peak B

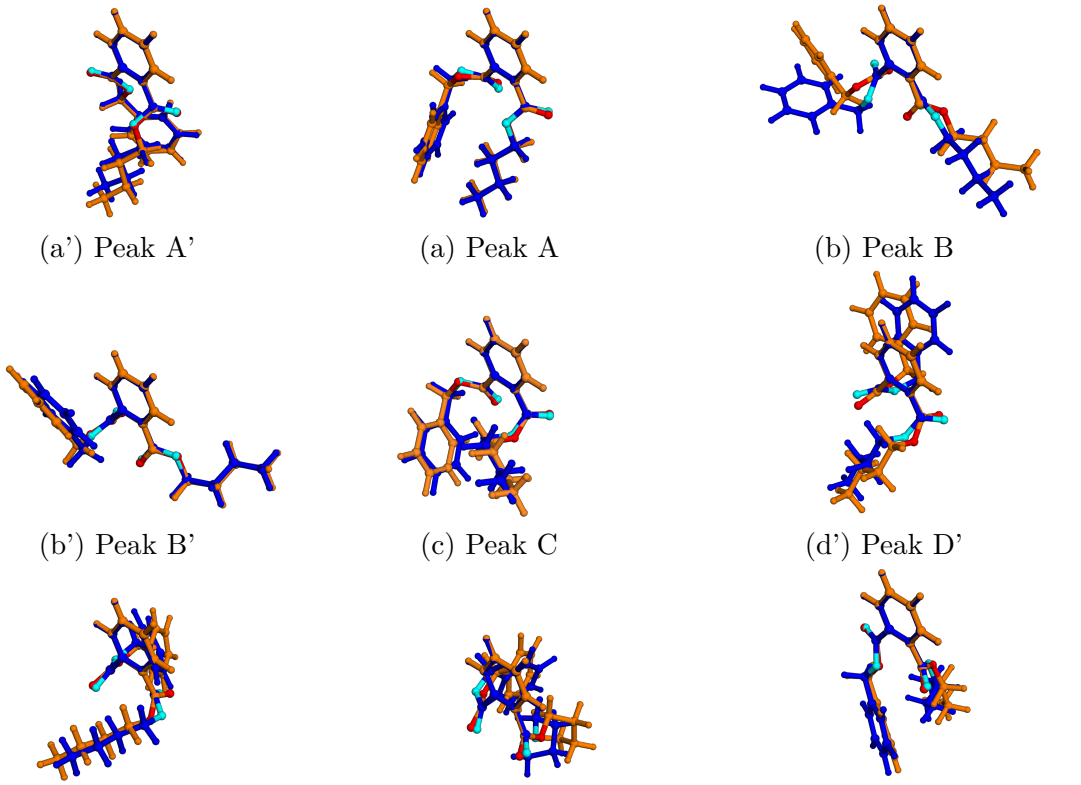
(c) Peak C

(d) Peak D

(e) Peak E

(f) Most stable DBP structure

Figure 3.11: Superposition of the representative structures of the main peaks observed in the structural excitation spectrum of DBP after local minimization at DFTB and DFT levels. DFTB: carbon and hydrogen in orange and oxygen in red. DFT: carbon and hydrogen in blue and oxygen in cyan.

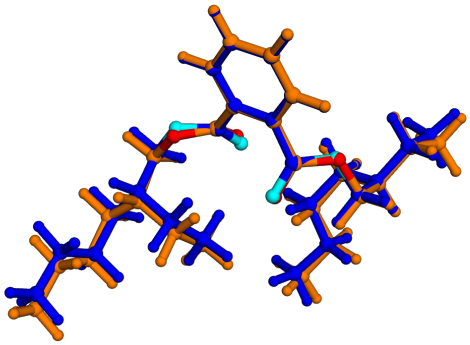


(a') Peak A' (a) Peak A (b) Peak B

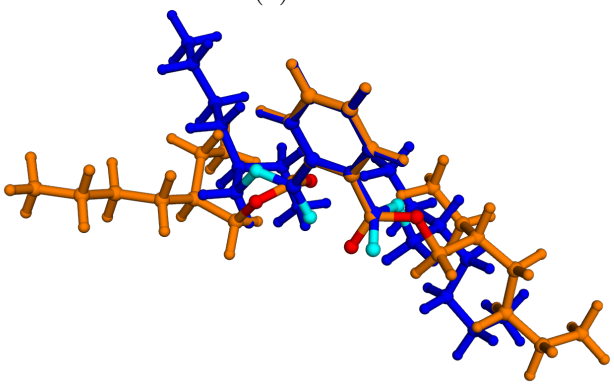
(b') Peak B' (c) Peak C (d') Peak D'

(d) Peak D (e) Peak E (f) Most stable BBP structure

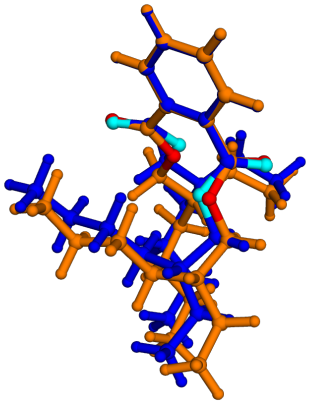
Figure 3.12: Superposition of the representative structures of the main peaks observed in the structural excitation spectrum of BBP after local minimization at DFTB and DFT levels. DFTB: carbon and hydrogen in orange and oxygen in red. DFT: carbon and hydrogen in blue and oxygen in cyan.



(a) Peak A



(b) Peak B



(c) Most stable DEHP structure

Figure 3.13: Superposition of the representative structures of the main peaks observed in the structural excitation spectrum of DEHP after local minimization at DFTB and DFT levels. DFTB: carbon and hydrogen in orange and oxygen in red. DFT: carbon and hydrogen in blue and oxygen in cyan.

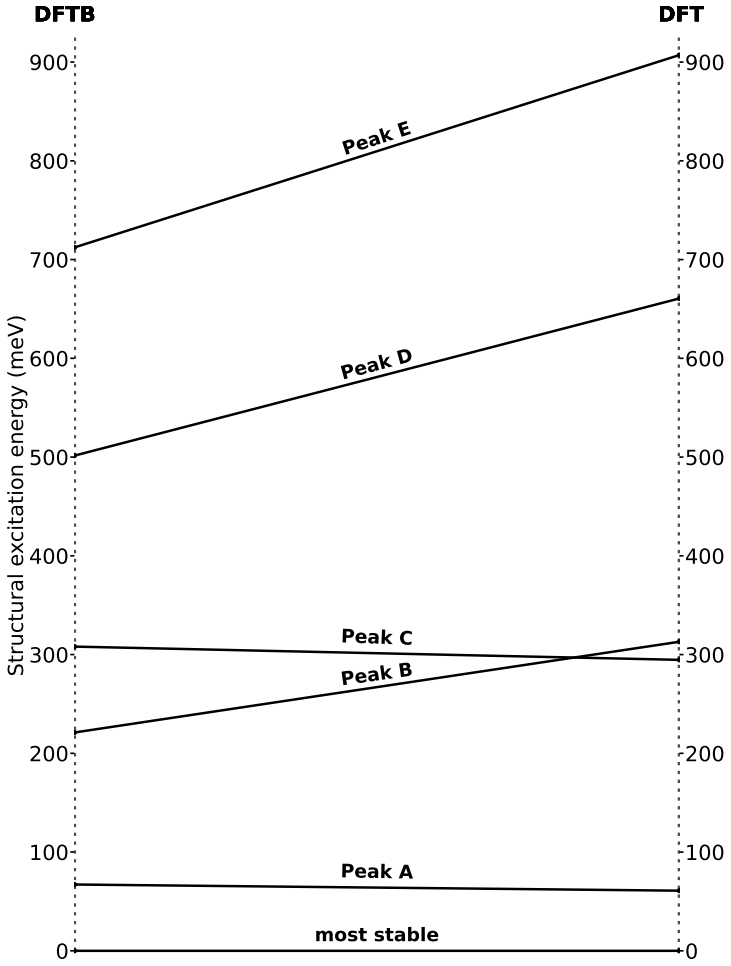


Figure 3.14: Comparison of DFT and DFTB energies of the characteristic DBP structures of the main peaks of the structural excitation energy spectra: lines connect the DFTB (left) and DFT (right) structural excitation energies (in meV) of the main peaks structures identified in figure 3.3. The DFTB(resp. DFT) structural excitation energy reference correspond to the lowest-energy structure computed at the DFTB(resp. DFT) level.

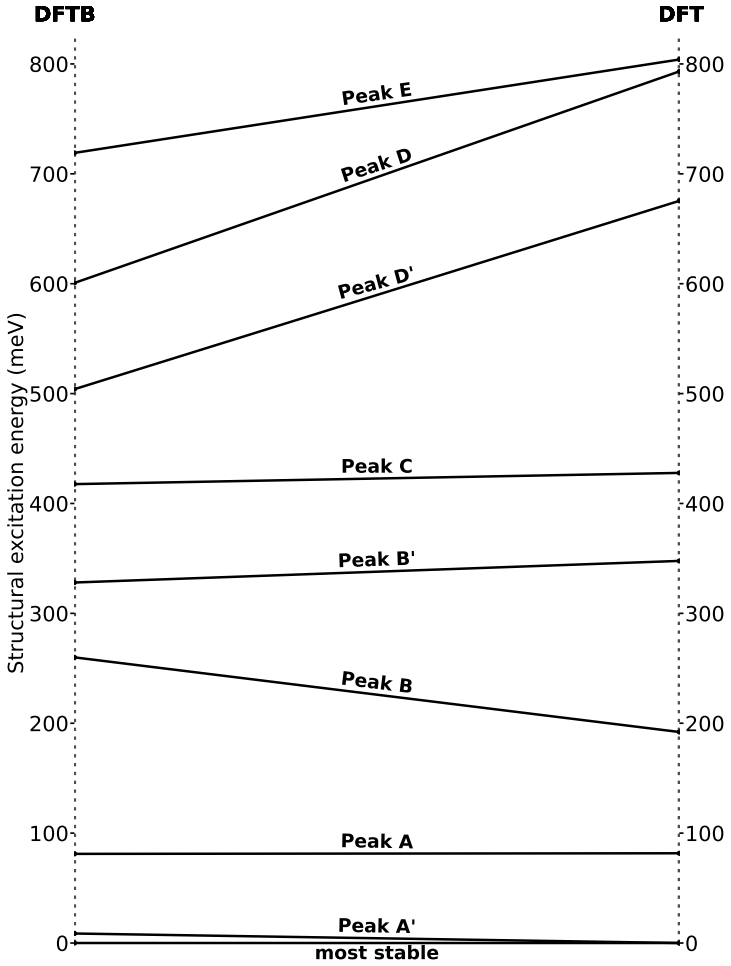


Figure 3.15: Comparison of DFT and DFTB energies of the characteristic BBP structures of the main peaks of the structural excitation energy spectra: lines connect the DFTB (left) and DFT (right) structural excitation energies (in meV) of the main peaks structures identified in figure 3.3. The DFTB(resp. DFT) structural excitation energy reference correspond to the lowest-energy structure computed at the DFTB(resp. DFT) level.

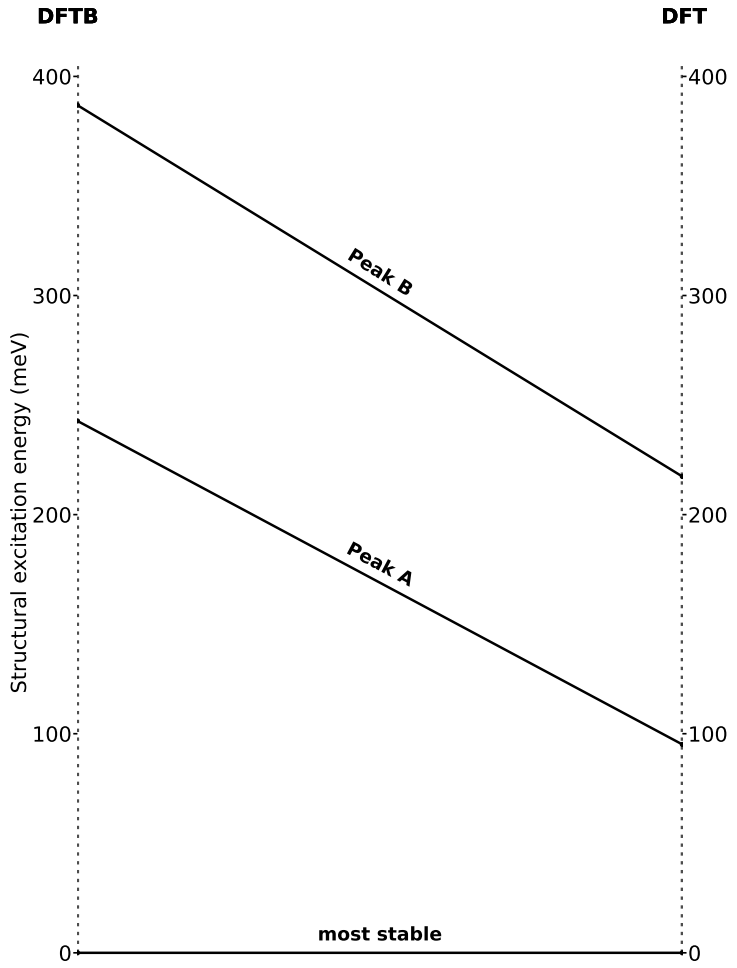
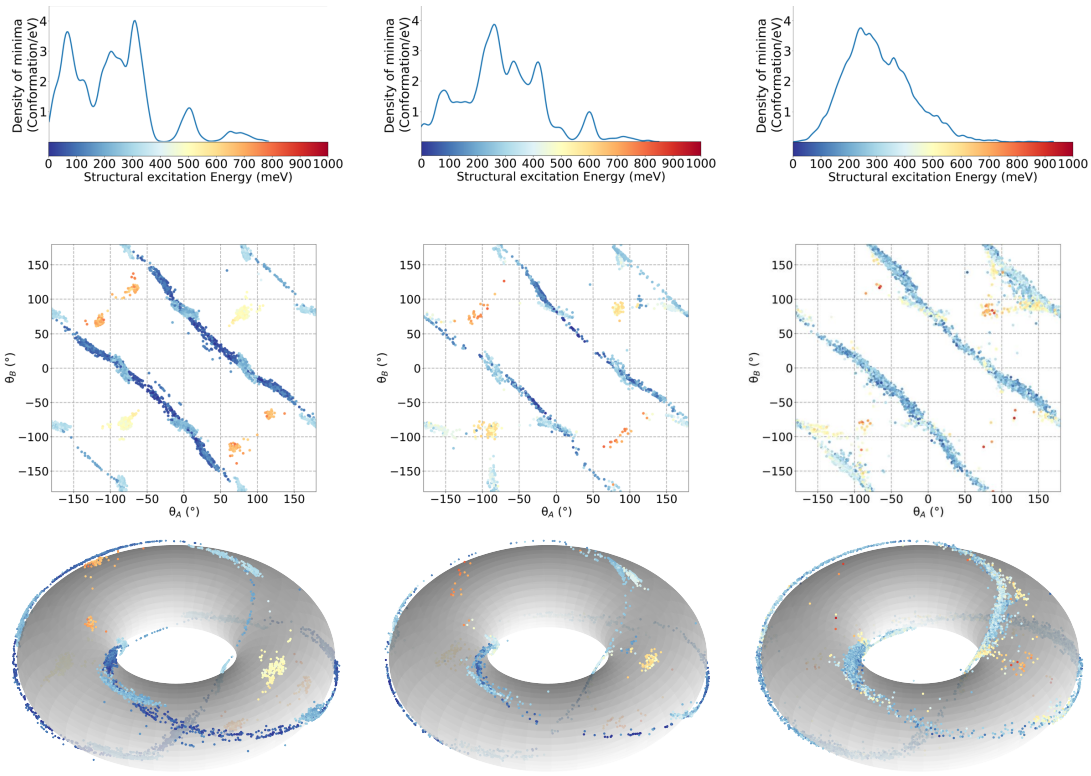


Figure 3.16: Comparison of DFT and DFTB energies of the characteristic DEHP structures of the main peaks of the structural excitation energy spectra: lines connect the DFTB (left) and DFT (right) structural excitation energies (in meV) of the main peaks structures identified in figure 3.3. The DFTB(resp. DFT) structural excitation energy reference correspond to the lowest-energy structure computed at the DFTB(resp. DFT) level.

phthalate molecules at the DFTB level.

3.4.3 Dihedral angle based analysis

The distribution of the conformations resulting from the IGLOO/DFTB exploration can also be visualized on a two-dimensional (2D) projection with respect to the dihedral angles θ_A and θ_B (see Fig. 3.2 for their definition). The projections for the three molecules are presented in Fig. 3.17, where each conformation corresponds to a point colored as a function of its structural excitation energy. For clarity, the spectra on Fig. 3.3 are presented again on top of the color-bars in Fig. 3.17. The figure shows two types of projections for these 2D angular distributions. The first one, in the center of the figure, is a classical representation on a Euclidean plane. In the second one, at the bottom of the figure, the conformations are projected on the surface of a two-dimension torus. This type of representation is less usual but better suited to the visualization of angular values due to their periodicity.



(a) DBP (b) BBP (c) DEHP

Figure 3.17: Distribution of the conformations resulting from the IGLOO/DFTB exploration. Top: Structural excitation spectra (Fig. 3.3) with color-bar. Middle: Two-dimensional (2D) projection with respect to dihedral angles θ_A and θ_B. Bottom: Projection on the surface of a two-dimension torus. In these plots, each conformation corresponds to a point colored as a function of its structural excitation energy (upper panel color-bar).

An initial structural analysis of the molecules reveals symmetries in conformational space that should be found in the dihedral angle analysis. Changing the signs of both θ_A and θ_B is equivalent to performing a symmetric projection of the atom coordinates with respect to a plane passing through the benzene ring. As this would lead to the same isomer, the 2D Euclidean projections in the central row of Fig. 3.17 should be symmetric with respect to the $y = x$ axis (ascending diagonal). In addition, when the two terminal groups are identical, exchanging the values of θ_A and θ_B also leads to the same structure and, as a consequence, symmetry with respect to the $y = -x$ axis (downward diagonal) should also appear. The difference between the two side-chains of BBP induces a loss of this second type of symmetry, particularly visible between the lower left and upper right bands in the corresponding 2D Euclidean plot. All the previously mentioned expected symmetries are recovered in Figure 3.17, which is reassuring regarding the quality of the global exploration. The figure shows similar angular distribution for the three molecules. All the low-energy conformations (colored in blue) are grouped within parallel bands in the Euclidean projection, or rings on the surface of the torus. Note that each ring is divided into two bands (one long and one short) on the Euclidean plane because periodicity is not taken into account. Note also that some isomers of higher energies (colored in yellow, orange and red) are located between these bands/rings. The energetic grouping of DBP and BBP isomers shown in Fig. 3.3, can also be observed through well defined colored regions on the 2D projections of Fig. 3.17-(a-b). In the plot corresponding to DEHP (Fig. 3.17-(c)), energy basins are less clearly identifiable, reflecting, as previously mentioned, the complex competition between several weak stabilizing interactions.

The 2D projections clearly show that θ_A and θ_B are strongly correlated. Considering only the points in the blue bands/rings and using linear regression, we obtained $\theta_B + \theta_A = c$ with $|c|$ in the [85-92°] range and correlation coefficients larger than 0.97, the two bands/rings differing by the sign of c. This correlation between the two dihedral angles is illustrated in the animation provided at https://zenodo.org/records/12646922. The reason for this correlation is probably due to the fact that low-energy conformations tend to maximize the oxygen-oxygen distances between side-chains and, therefore, when one chain rotates, the other does so accordingly. In our previous analysis of the structural excitation spectrum, the high densities were mostly interpreted in the light of $dmin_{O-O}$ and $dmin_{C-O}$ values, the later being strongly linked to the θ_A and θ_B values.

Finally, we can imagine transitions between conformations projected onto the two low-energy bands/rings, passing through the high-energy yellow-red regions. However, finding these transitions would require a variant of the IGLOO algorithm, focused on sampling transition paths rather than low-energy basins. This work will be presented in the chapter 6.

3.5 Conclusion

The methodology presented in Chapter 2 has been applied to the exploration of the conformational potential energy surface of three molecules representative of the phthalate family: butyl benzyl phthalate (BBP), dibutyl phthalate (DBP) and di-(2-ethylhexyl)

phthalate (DEHP). This choice was motivated by their high impact on human health. The results show that BBP, DBP and DEHP, despite belonging to the same family and being close in size, present different conformational landscape properties. The general aspect of the structural excitation energy spectra shows different isomer density distributions for the three molecules. The DBP spectrum has well defined peaks while the DEHP one exhibits a continuum of close-energy states. The BBP spectrum is at the crossroads between these two previous behaviours. These differences have been rationalized making use of descriptors based on distances and dihedral angles.

DBP lower-energy structures are mostly governed by oxygen-oxygen coulomb interactions. In the case of BBP, original structures, where the positively charged hydrogen atoms of the butyl side-chain point toward the negatively charged aromatic carbon atoms, allow to maximize coulomb interactions stabilization. Finally, DEHP long and ramified side-chains induce steric hindrance and dispersive interactions, these latter being at the origin of competitions between plenty of isomers. These interactions drive the geometric properties of the investigated phthalate molecules leading either to peaks (DBP and BBP) or to a broad feature (DEHP) in characteristic O-O and C-O distances distribution plots and to a strong correlation between the two dihedral angles describing the side-chains orientation for the three molecules.

One should note that the phthalate molecules have been studied here in the gas phase and that further research could provide a protocol for finding conformations that could exist under more realistic conditions. The effects of the environment could be incorporated through QM-MM explicit [172] or implicit [90] solvent scheme. The IGLOO/DFTB coupling implemented in this work, allowing the identification of low energy minima of a molecule with no a priori knowledge of its potential energy surface, could be extended in the future to the blind search of the minimum energy path between selected structures.

3.6 Data and Software Availability

Data presented in this chapter have been deposited on ZENODO:
https://zenodo.org/records/10040725.
As mentioned in the “Implementation details" section, the combined IGLOO/DFTB approach was implemented on the basis of the Molecular Motion Algorithms (MoMA) software suite (https://moma.laas.fr/) and the deMonNano code (http://demon-nano.ups-tlse.fr). Software binaries and user guidelines are available at:
https://gitlab.laas.fr/moma/binaries/igloo-dftb-coupling.

Chapter 4

Large-scale generation of atomistic models of aromatic hydrocarbons

Contents

4.1 Introduction

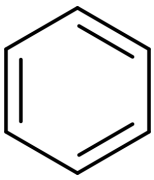
Molecules are categorized into various families based on their structural features, bonding patterns, and chemical properties. This classification helps in understanding their behavior, reactivity, and applications in different fields of science and technology.

One primary distinction in molecular families is between inorganic and organic molecules. Inorganic molecules are primarily composed of elements other than carbon, such as metals, and often participate in ionic bonding. Conversely, organic molecules are characterized by the presence of carbon atoms, linked predominantly by covalent bonds, forming a vast array of structures from simple hydrocarbons to complex biomolecules.

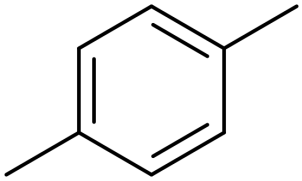
Among organic molecules, hydrocarbons are the simplest and most fundamental class, consisting solely of carbon and hydrogen. Hydrocarbons are further divided into two major categories based on the types of bonds between the carbon atoms: saturated and unsaturated hydrocarbons. Saturated hydrocarbons, or alkanes, have single bonds and are symbolized by the general formula C_nH_{2n+2}. Unsaturated hydrocarbons include alkenes and alkynes, which contain double and triple bonds, respectively.
Transitioning to a more specialized family of molecules, aromatic hydrocarbons, also known as arenes, constitute a crucial class of hydrocarbons distinguished by the presence of one or more aromatic rings. These rings are planar, cyclic structures with delocalized π-electrons that adhere to Huckel's rule, which states that aromatic compounds must have a specific number of π-electrons (4n + 2, where n is a non-negative integer) in a closed loop of continuously overlapping p-orbitals. Aromatic hydrocarbons can be classified into:

- **Monocyclic Aromatic Hydrocarbons** (Fig. 4.1): These contain a single aromatic ring, such as benzene, toluene (methylbenzene), and xylene (dimethylbenzene).

- **Polycyclic Aromatic Hydrocarbons (PAHs)** (Fig. 4.2): These consist of multiple fused aromatic rings, such as naphthalene (two fused benzene rings), anthracene (three fused benzene rings), and phenanthrene (three fused benzene rings).



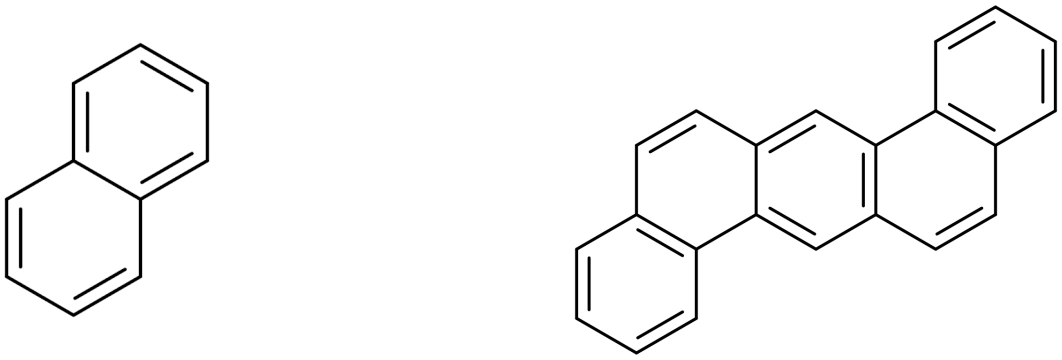
(a) Benzene



(b) Xylene (dimethylbenzene)

Figure 4.1: Illustration of Monocyclic Aromatic Hydrocarbons

This chapter presents a novel algorithm for the large-scale generation of atomistic models of aromatic hydrocarbons. The algorithm leverages the capabilities of molecular graphs combined with a three-dimensional structure generator to obtain diverse structures, with a different arrangement molecular structures. The primary contribution of this work is the development of a two-part structure generation algorithm. The algorithm is based on the generation of molecular graphs and the addition of atoms and fragments to the graph. The development of this algorithm enables the creation of extensive molecular databases, which are essential for applications in environmental science, astrophysics, and materials science.



(a) Naphthalene

(b) Dibenz[a,h]antracene

Figure 4.2: Illustration of Polycyclic Aromatic Hydrocarbons

4.1.1 Context: Aromatic hydrocarbons

Polycyclic aromatic hydrocarbons (PAHs) (discussed above) are a class of organic compounds composed primarily of carbon and hydrogen atoms. The carbon atoms form the skeleton of the rings, while hydrogen atoms are attached to the periphery to complete the molecular structure. PAHs can be thought of as fragments of graphene, with a honeycomb carbon structure bordered by hydrogen atoms. Among PAHs, naphthalene ($C_{10}H_8$), with its two aromatic rings, is considered to be the simplest PAH, while benzene (C_6H_6), although aromatic and the elemental building block of PAHs, is not considered as a PAH.

PAHs are broadly divided into two main groups: peri-condensed and cata-condensed PAHs. Peri-condensed PAHs have a compact structure in which most carbon atoms are part of two or three different rings, forming nearly circular molecules such as coronene ($C_{24}H_{12}$), circumcoronene ($C_{54}H_{18}$), or circumcircumcoronene ($C_{96}H_{24}$). In contrast, cata-condensed PAHs have an open structure, with carbon atoms belonging to a maximum of two rings, resulting in linear or branched chain configurations (Fig. 4.2).

On Earth, PAHs occur naturally in petroleum and coal, resulting from the chemical transformation of natural "product" molecules. They are also formed during the combustion [67] of carbonaceous fuels such as wood, coal, diesel, fat, tobacco, and even during the cooking of food. As a result, PAHs are found in vehicle exhaust, tobacco smoke, and charred food. Notably, certain PAHs have been identified as carcinogenic and mutagenic [75, 62], highlighting the importance of their study in the context of public health and air pollution. Additionnaly, PAHs can be found in complex structures such as soot. For example, PAHs have been identified in soot from domestic coal-burning stoves [166].

Soot, often referred to as black carbon, is a complex mixture of tiny particles composed primarily of carbon. Soot is a major component of air pollution and can be found in both urban and rural environments from a variety of sources, including vehicles, industrial processes, and residential heating [25].

Soot formation is a process that involves the pyrolysis and incomplete combustion of

hydrocarbon fuels. During this process, a series of complex chemical reactions occur that result in the nucleation of small particles, followed by their growth and agglomeration into larger aggregates. These particles consist of a core of elemental carbon with various organic compounds [23].

Soot particles pose significant environmental and health risks. Environmentally, they contribute to air pollution and climate change. Soot particles can absorb solar radiation, affecting the Earth's radiation balance and contributing to global warming. They also play a role in the formation of acid rain and can lead to the deterioration of materials and surfaces[130]. From a health perspective, soot particles are of concern because of their ability to penetrate deep into the lungs and bloodstream, leading to respiratory and cardiovascular disease. Long-term exposure to soot particles has been linked to increased rates of asthma, bronchitis, heart attacks, and even premature death [127].

Experimental analysis such as Infrared spectroscopy can give an information about the distribution of the functional groups inside a soot sample for example. To identify specifically the type of structures presents inside a particle, a database demonstrating a variety of aromatic hydrocarons is required. This database can be used to study the properties of these molecules, such as their electronic and geometric descriptors, and to compare them with experimental data. To create such databases, algorithms are developed with various strategies to generate a large number of structures, following constraints on chemical elements, functional groups, etc.

4.1.2 State of the art: structure generation algorithms

A structure generation algorithm is a method for generating a set of structures that satisfy a set of constraints. Depending on the specifications, such algorithms can generate structures from different families of molecules. Algorithms developed by Wahab et al. for the Compas project (illustrated in Fig. 4.3) are able to generate polycyclic aromatic systems [160], cata-condensed hetero-polycyclic aromatic systems [115], and peri-condensed polybenzenoid hydrocarbons [159]. These structure generations algorithms based on the computation of random molecular graphs use CaGe [29], a chemical graph generation software. This software generates all possible unoptimized molecules according to size and chemical elements constraints. Structures are then locally optimized using xTB [15]. A Density Functional Theory (DFT) optimization is also performed to ensure the quality of the structures using ORCA[121, 120]. At each step, many filters are applied, such as removing structures where bonds are created during xTB minimization, or structures with a linear stretch longer than six rings (hexacene for example) because they have non-negligible open-shell character [155, 18, 99] in the ground state, and such molecules are relatively unstable. The generation of random structures following constraints is a difficult task and the quality of the structures obtained is not always satisfactory, since the algorithm of the Compas project has to refine the structures obtained in order to converge to a database of structures that could exist.

The innovative algorithms introduced by Leguy et al. in their EvoMol framework provide a novel methodology for the unbiased generation of molecular structures. These algorithms employ an evolutionary strategy for molecular generation, enabling the explo-

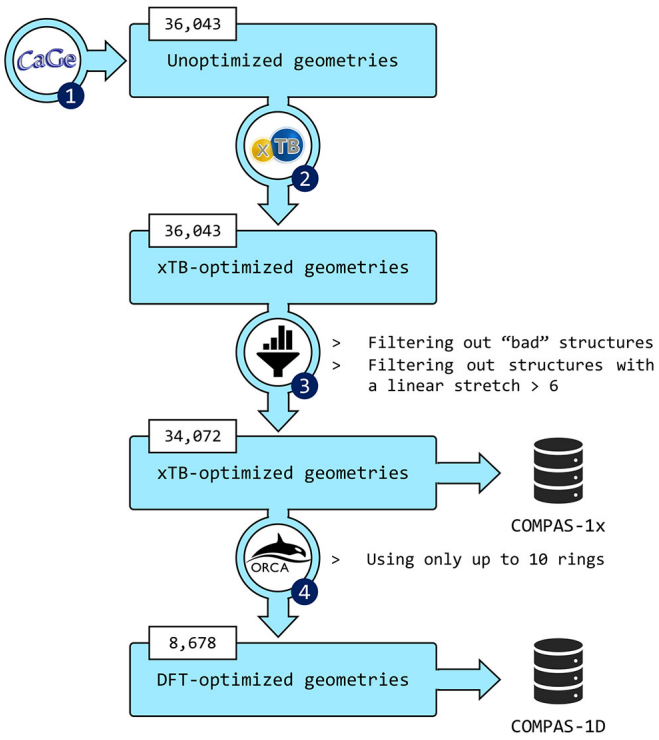


Figure 4.3: Illustration of the Flowchart of the data-generation process in the Compas project from [160].

ration of both known and uncharted chemical spaces without the reliance on pre-existing datasets. The fundamental principle underlying the EvoMol method is the implementation of a set of seven generic (depicted in Fig. 4.4) chemically meaningful mutations, which are applied at the atomic level to construct molecular graphs in a sequential manner [106]. EvoMol's generation process begins with simple molecules, often starting

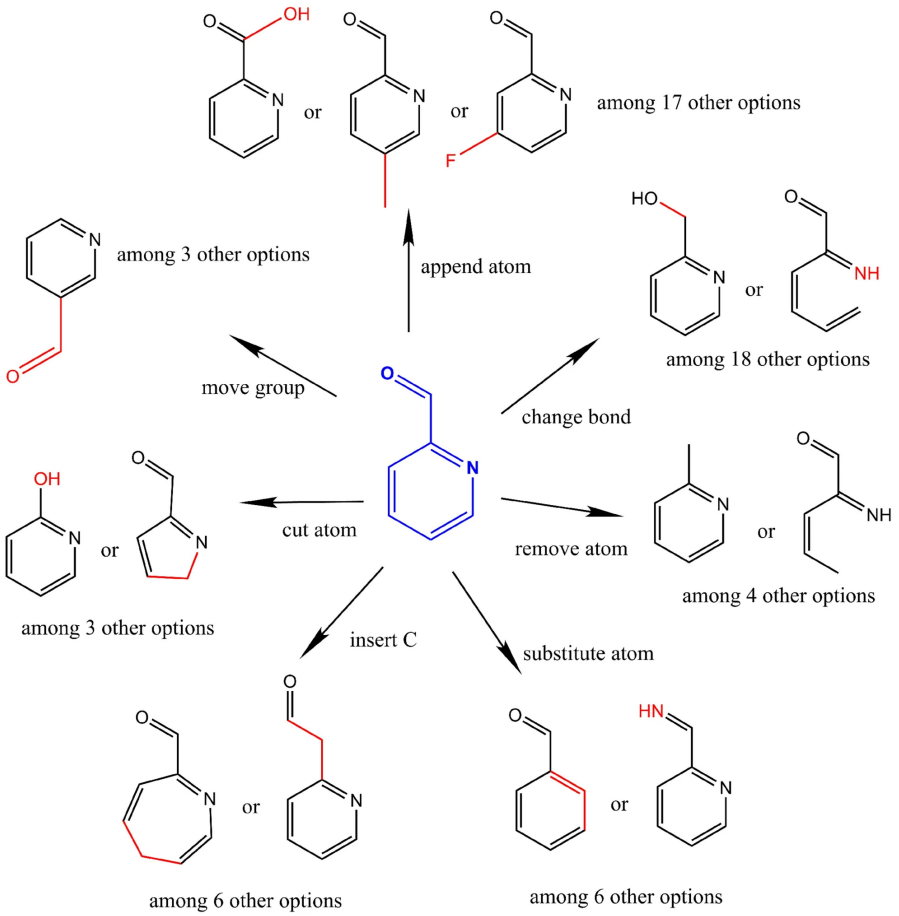


Figure 4.4: Illustration of the Flowchart of the data-generation process in the EvoMol project from [106].

from methane, and employs mutations to expand the molecular structure in a controlled manner. The generated molecular structures are guaranteed to be chemically valid at each step due to the rigorous application of constraints that prevent the formation of chemically impossible configurations. This is facilitated by the use of RDKit, a software toolkit that aids in the manipulation and validation of chemical structures. RDKit enables EvoMol to maintain a consistent representation of molecules as molecular graphs, thus facilitating the subsequent stages of optimization and property evaluation. Subsequently, EvoMol performs optimization based on specific target properties, such as HOMO and LUMO energies. These properties are of great importance for applications in organic electronic materials, where the role of molecular orbitals is significant. The optimization process is conducted using DFT calculations in order to ensure that the

generated molecules exhibit the desired properties.
The algorithms may be employed to generate structures based on experimental observations or to be browsed by experimentalists for the purpose of comparing properties derived from their research.

4.1.3 A new strategy for Structure Generation

The primary objective is to randomly generate structures devoid of any a priori knowledge and to investigate the impact of the structure on electronic and geometric descriptors. Moreover, depending on the study, it may be advantageous to impose constraints on certain properties. The presented algorithm will consider constraints on the number of carbons in the structure, the ratio of aromatic C=C bonds to C-X bonds, the ratio of olefinic CH_2 to aliphatic CH_3, and the ratio of aliphatic CH_2 to aliphatic CH_3. These constraints were choosen to generate structures according to the article from Dartois et al. [41]. This article will be detailed in the application of the algorithm presented in Chapter 5.

As with the other methods discussed, the new algorithm developed (depicted in Fig. 4.5) in this chapter is based on molecular graphs. The generation process is based on the addition of atoms and fragments to the molecular graph. The algorithm is capable of generating structures that adhere to predefined constraints on the ratios of various chemical bonds. Three-dimensional structures are generated by a process developed in the Structure Generator section.

From an initial set of constraints, defined as the number of carbons by structures and ratios on the functional groups (defining the type of bond added to the graph), the developed algorithm is divided into two parts:

- The first part combine a SMILES Generator and a SMILES Filtering processes, which generates a set of Simplified Molecular Input Line Entry System (SMILES), respecting a set of constraints given by ratios and a total number of carbon atoms. SMILES is a notation system for describing the structure of chemical species using short ASCII strings (see Fig. 4.5 top right). SMILES strings can uniquely represent molecular structures and are a compact and convenient way to encode molecular information. A SMILES string typically contains atoms, represented by their chemical symbols, and bonds between them, represented by various symbols such as '-', '=', and '#' for single, double, and triple bonds, respectively. Number can be used to indicate bonds between atoms that do not follow a linear or simple sequence. These numbers are essentially "labels" that help close cycles or rings in the molecular structure. For example, the SMILES string for coronene ($C_{24}H_{12}$) is 'C1=CC2=C3C4=C1C=CC5=C4C6=C(C=C5)C=CC7=C6C3=C(C=C2)C=C7', for ethanol (C_2H_6O) the SMILES string is 'CCO'. SMILES are filtered based on criteria detailed bellow and are then used to generate a 3D structure.

- The second part combine a Structure Generator and a Structure Filtering processes, which generates a set of 3D structures from the set of SMILES codes

Figure 4.5: Illustration of the Flowchart of the data-generation process. A box with the structure generator and filtering is generated for each SMILES generated at the previous step. These box will generated the number of 3D structures requested

obtained from the first part. From a SMILES, a number N_S of structures is generated. For N_S>1, an acceptation test is performed based on distance between the structures. Only structures above a distance threshold defined by the user are kept.

The following sections provide a detailed account of each component of the methodology.

4.2 SMILES Generator

4.2.1 Overview

The SMILES generator is an algorithm designed to generate a set of molecular graph. The algorithm, which is detailed in Fig. 4.6, generates a set of SMILES strings, respecting the number of SMILES requested (N_S), and a set of constraints on the ratios of various chemical bonds:

$$ratio_A = \frac{N_{aromaticC=Cbonds}}{N_{C-Xbonds}} \quad ratio_B = \frac{N_{olefinicCH2}}{N_{aliphaticCH3}} \quad ratio_C = \frac{N_{aliphaticCH2}}{N_{aliphaticCH3}} \tag{4.1}$$

Each ratio is defined to generate structures that adhere to the infrared observations presented in the article [41] and will be more explained in the next chapter.

In the algorithm, each structure is defined as a RDKit *mol* object, which is a graph representing the connectivity of the molecule with the constituting atoms and bonds that connect them. For a generation, three main phases are performed until convergence on the total number of carbon in the graph (N_C), as follows: (i) Selection of the fragment type, (ii) selection of the atom between each available carbon in the graph, and (iii) addition of the fragment to the graph.

4.2.2 Fragment type selection

The fragment type selection phase is the process of choosing the type of fragment to be added to the graph. Three types of fragments may be incorporated into the graph: a single-bonded carbon, a double-bonded carbon, or an aromatic fragment. The selection of the fragment type is contingent upon the constraints defined by the ratios $ratio_A$, $ratio_B$, and $ratio_C$. In order to determine which fragment should be appended to the graph, a hierarchy is created between the constraints (ratios of 4.1) that must be satisfied. A ratio is considered as satisfied if it is upper than the targeted ratio.
The hierarchy is randomly generated (one possible choice is $ratio_C > ratio_A > ratio_B$) and the first constraint that is not satisfied is used to determine the type of fragment to be added. In the event that all constraints are satisfied, the default option is to add a single bonded carbon atom. For example, if the order is $ratio_B > ratio_C > ratio_A$, the first ratio to be satisfied is $ratio_B$, which represents the ratio between olefinic CH_2 and aliphatic CH_3 carbons. Subsequently, the next step is to determine to how the graph will be modified according to the fragment selected.

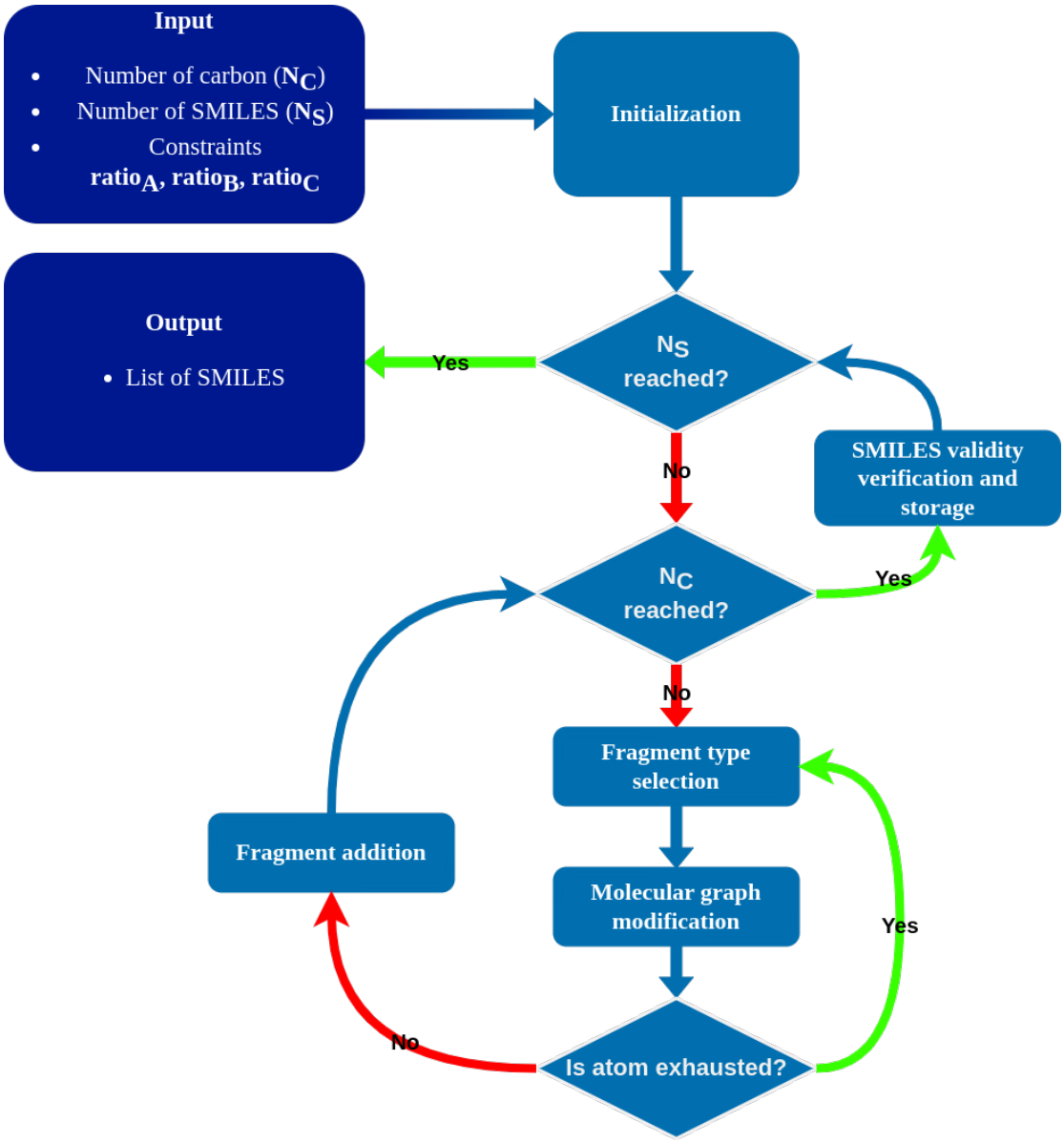


Figure 4.6: Illustration of the Flowchart of the SMILES Generator.

4.2.3 Molecular graph modification

The molecular graph modification phase may be executed in two distinct manners: a modification of a bond within the graph (for instance, a single bond becoming a double bond) or a selection of a carbon atom within the graph and the type of fragment to be incorporated. The selection will depend on which ratio must be satisfied in order to converge to the desired chemical composition. It should be noted that for modifications that change a bond type in the graph (from single to double bond for example), no atom is selected and a new fragment type selection is made with recalculated ratio. Note that only carbon atom bonds are discussed below, and hydrogens bonds are implicit. Furthermore, the following discussion will focus on the strategy used to increase each ratio, given that they are correlated and, thus, increasing one will result in a reduction of the others. The different options could be defined as:

- **Ratio$_\mathbf{A}$** : In order to increase and reach the desired ratio, a random carbon atom is selected within the graph. Following this, an aromatic fragment will be added at the subsequent phase. In the event that the graph is empty, no atom will be selected and an aromatic fragment will be added.

- **Ratio$_\mathbf{B}$** : In the event that the graph is empty, the simplest option is to add an ethene. Then, to increase this ratio in order to reach the desired value, several possibilities exists. Two possible option are: (i) increase the number of olefinic CH_2 carbons or (ii) decrease the number of aliphatic CH_3 carbons. For the first option (i), a carbon atom, not in an aromatic ring is searched and a carbon atom is double bonded to this one at the subsequent phase. For the second option (ii), if an ethyl exist, the single bond between the carbon atoms of this group is converted from a single bond to a double bond. Note that if (i) and (ii) are possible, a random selection is made between them.

- **Ratio$_\mathbf{C}$** : As discussed for the previous ratio, two options are possible to increase the ratio: (i) increase the number of aliphatic CH_2 carbons or (ii) decrease the number of aliphatic CH_3 carbons. For the first option (i), if a terminal double bond between carbon atoms is found, the bond is converted to a single bond. For the second option (ii), if a methyl group is found, a carbon is added to the graph and single bonded to this group at the subsequent phase. As for the previous ratio, if (i) and (ii) are possible, a random selection is made between them. If the graph is empty, an ethane will be added to the graph.

4.2.4 Fragment addition

If an atom was selected at the molecular graph modification step (i.e. bonds were not modified at the previous step), with the aim to be the linking atom from the graph to the fragment, a valence check is conducted on this atom. For instance, if a single bond will be added, the selected atom must be available to form a single bond. A classical valence rule is employed for the addition of a single or double bonded carbon atom or an aromatic fragments. A specific rule is incorporated if a double bond is

introduced between a new carbon atom and the selected atom. A carbon atom is considered to be exhausted if it already has a double bond and another double bond is attempting to be added. This condition is created to prevent the formation of a series of double bonds, which is not pertinent to the subject matter to be discussed in the subsequent chapter. In both cases of single bond or double bond addition, the underlying principle is straightforward. A new atom is introduced into the graph, and the subsequent bond is established between the new atom and the selected atom. Two distinct scenarios exist with regard to the addition of aromatic fragments. If the selected atom is aromatic, the aromatic fragment will be a closure benzene ring, as defined in Fig. 4.7. It is assumed that the aromaticity of the resulting structure will be preserved. If the addition is not possible, the graph is preserve and a new round of fragment type selection is made. Alternatively, the added fragment is selected from the set of possible fragments (see Fig. 4.8): benzene ring, naphthalene, anthracene, phenanthrene, pyrene, benzo[c]phenanthrene, chrysene, or tetracene. A single bond is then created between the selected atom and the aforementioned fragment.

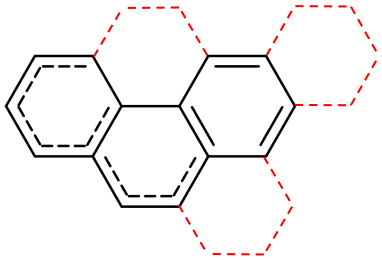
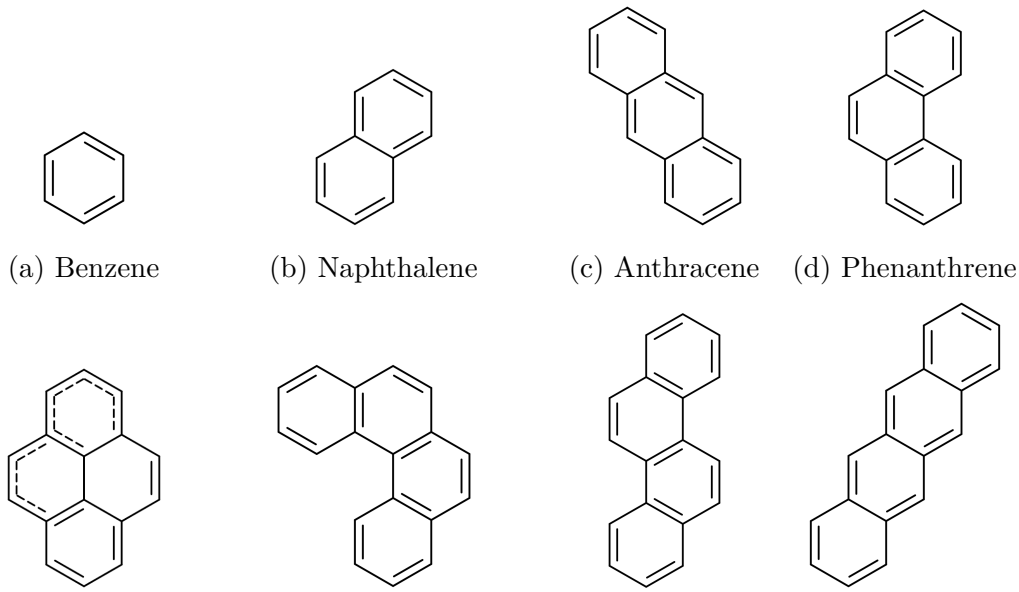


Figure 4.7: Addition of an aromatic fragment to the molecular graph for an aromatic carbon. The red part represents the ring added to the graph.

For the case depicted in Figure 4.7, a problem occur with the RDKit library. To bond an aromatic fragment illustrated in red dotted lines, two atoms in the graph are required. The bonding process starts with one atom of the aromatic fragment connected to the graph. Each atoms are identified by their index, which is updated when a element is added to the graph, resulting in a wrong connection with the second atom of the aromatic fragment and the graph. To avoid this problem, each atom is now identified by its map number, which is a list of unique identifiers for each atom in the graph and is not automatically updated when new atoms are added. The map number of new atoms is then created after the end of the bonding process.
The algorithm iterates until the number of carbon atoms in the graph equals the targeted number of carbons (N_C) and the number of SMILES generated is equal to the number of SMILES targeted (N_S). Note that there is a protection in the algorithm if for more than hundred iteration in a row, it is impossible to add atoms on the graph. In this case the algorithm will reset the graph and start again. This procedure is used to avoid an infinite loop in the algorithm. To improve the performance of the algorithm, different instances of the code are run in parallel (embarrassingly parallel) using multi-threading. Each thread is independent, with different ratios. Generated SMILES are stored and used in the next part of the algorithm to generate 3D structures.



(a) Benzene (b) Naphthalene (c) Anthracene (d) Phenanthrene

(a) Pyrene (b) Benzo[c]phenanthrene (c) Chrysene (d) Tetracene

Figure 4.8: Addition of an aromatic fragment to the molecular graph to a non aromatic carbon

4.3 Structure Generator

4.3.1 Overview

For each SMILES, a set of three-dimensional structures is generated. A preliminary, unoptimized structure (in PDB format) is generated from the SMILES using RDKit, which serves as input (see Fig. 4.9). The Structure Generator (SG) is an iterative algorithm that generates a set of structures, respecting the requested number of structures (N_{Struct}), the targeted distance between structures, and the number of fails allowed (N_{fails}). At each iteration of the SG algorithm, three principal steps are carried out: (i) structure sampling, (ii) geometric relaxation, and (iii) structure validation and energy minimization. The algorithm iterates until the number of structures generated is equal to the number of structures requested, or stops because it is impossible to generate a structure based on the constraint given for this SMILES. The generated structures are then stored in a database for further analysis.

4.3.2 Structure sampling

In order to perform the structure sampling phase, an object is constructed, containing the connectivity information of the molecule from the initial PDB file and the list of the dihedral angles of the molecule. For a torsion around a bond (a dihedral angle), several combinations of atoms can be defined. Only one torsion is defined around a bond, the remaining atoms are defined inside rigids (see Figure 4.10).

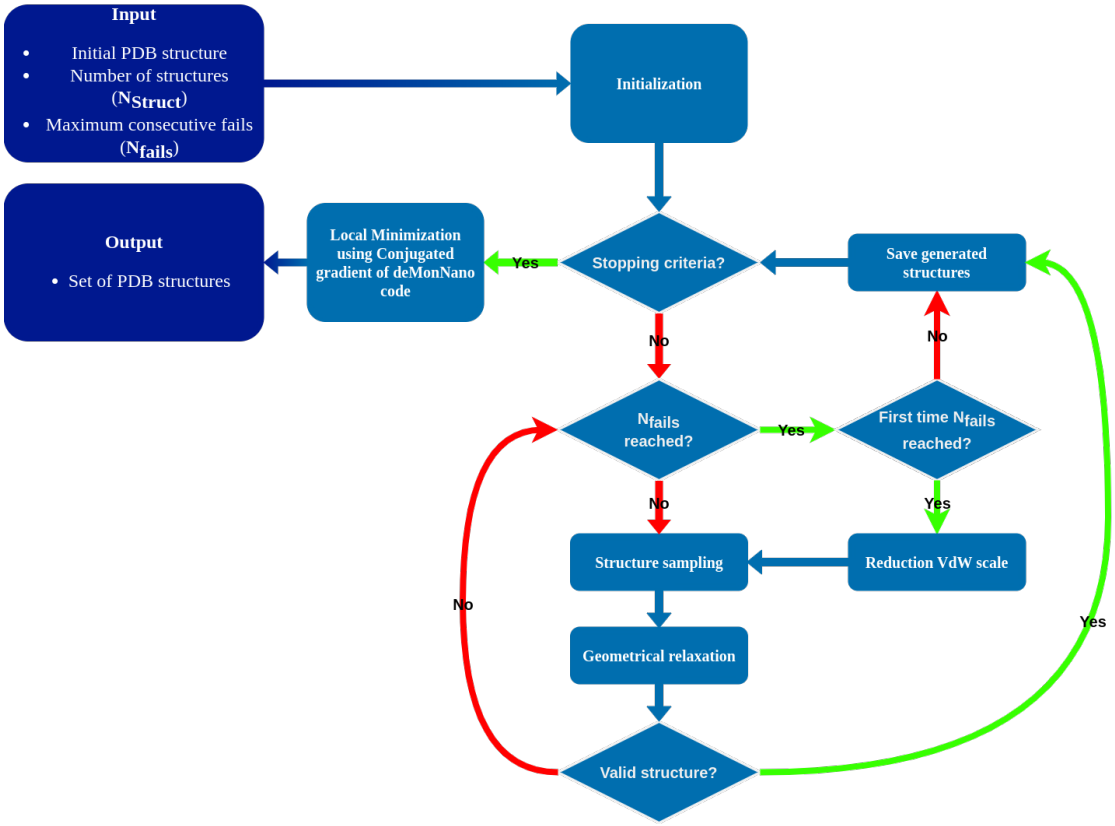


Figure 4.9: Illustration of the Flowchart of the Structure Generator.

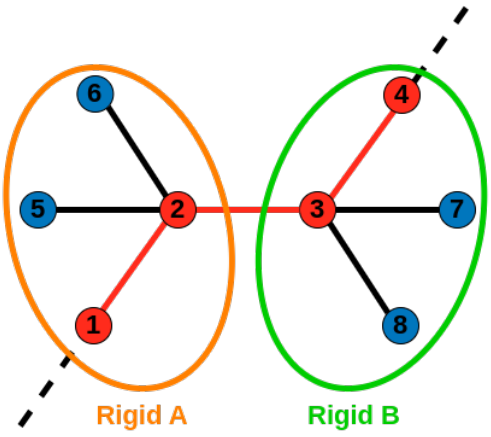


Figure 4.10: Illustration of the rigid definition in a molecule. Black dotted lines represent bonds to the rest of the molecule. Red lines represent the bonds between atoms composing the dihedral angle, including the atoms 1,2,3 and 4. Rigid A is composed of atoms 1,2,5 and 6 while Rigid B is composed of atoms 3,4,7 and 8. When a rotation is performed on the dihedral angle, a move is performed on the entire rigid connected to this dihedral angle, which can be rigid A or B. The other part of the molecule is also rigid and follow the movement.

When a rotation is performed on a dihedral angle, collisions are activated to the rigids directly connected to atoms of this dihedral angle (i.e. a collision check could be performed only on atom for which the collision is activated). Collisions are defined atom by atom, for which a Van der Waals (VdW) sphere is associated. A user parameter named VdW scale, is a ratio used on the VdW sphere to define a collision sphere around each atom. If there is an intersection between collision spheres of two atoms, the latter are considered to be in collision. This scale is lower during the sampling phase than during the minimization phase. This is done to reduce the time required to sample a structure, which is a demanding step, especting the relaxation step to remove constraints of the molecule.
The sampling scheme follow an iterative process, by assigning random values between $-\pi$ and π to the dihedral angle of the structure, starting with the dihedral angle that is the closest to the center of mass (determined on the unoptimized structure given by RDKit) and extending to the extremity of the structure. Upon the allocation of a value to a dihedral angle, the associated rigid collisions are activated, while the remaining rigids for which their dihedral angles have not yet received a value are deactivated. At each step, if a collision is detected, the value of the dihedral angle in question is modified, and the collision is tested again. If the number of failures for this dihedral angle exceeds a specified threshold (defined by the user), the dihedral angle allocation value phase is restarted from the previous dihedral angles. Collisions are deactivated for every atoms that are in the associated rigid.
If multiple fails happen at the same dihedral angle, the algorithm could restart from the previous n-1 dihedral angle (where n is the position of the actual dihedral angle in the dihedral list) to the n-i st dihedral angle (where i is the number of fails occurring for this dihedral angle). For example, if this is the second time the algorithm has failed to assign a value to a dihedral angle due to collisions, it will start the assignment process two dihedral angles back. The algorithm will iterate until all dihedral angles have a value and no collision are detected under a certain threshold. A security is added to prevent a state where the algorithm is stuck in an infinite loop at the same dihedral angle, by reseting the entire structure. Geometric relaxation is then performed on the structure.

4.3.3 Geometric relaxation

After the structure sampling phase, a collision test is conducted on the generated structure. A collision may be identified due to the fact that the VdW scale is greater at this stage than at the previous one (discussed in the previous section). The algorithm will identify which atoms are colliding and attempt to eliminate the collision through a minor perturbation to the dihedral angles associated with these rigids (i.e., the closest dihedral angles to these atoms). Should a perturbation result in an increase in the number of collisions, the structure will revert to its previous state. Furthermore, the subsequent perturbation will be greater than the previous one until a conformation is identified that exhibits a lower number of collisions than the best one obtained during iteration. This scheme continues until a threshold of $\frac{\pi}{2}$ perturbation (empirically determined) is

reached. Above this threshold, a low convergence rate was observed in the direction of a less collision-prone state. Upon the identification of a less collision-prone structure, the perturbation value is reset to its initial value, and the algorithm continues. The algorithm iterates until no collisions are found or the number of iterations reaches a specified threshold (user defined parameter). Subsequently, the structure is returned and a test is conducted to ascertain whether it has no collision and is at a certain distance from the existing structures.

4.3.4 Structure validation and energy minimization

To ensure that structures are sufficiently different from each other, the mean square deviation (MSD) between the generated structure and the existing structures is calculated and defined by the equation 4.2.

$$MSD_{ij} = \frac{1}{N_A} \sum_{k=1}^{N_A} (\theta_k^i - \theta_k^j)^2 \tag{4.2}$$

where MSD_{ij} is the distance between structure i and structure j, N_A is the number of dihedral angles in the structure, θ_k^i is the k-th dihedral angle value in structure i, and θ_k^j is the k-th dihedral angle value in structure j. Note that the distance between dihedral angles is between $-\pi$ and π. For example, the distance between $-\pi$ and π is zero. Subsequently, the algorithm compares the MSD between the generated structure and the existing structures. If the MSD is above a specified threshold defined by the user, the structure is accepted. If the structure does not meet the requisite criteria, it is rejected and N_{fails} is incremented. In the event that the number of fails exceeds the maximum number of fails, the algorithm will reduce the VdW scale of the minimizer, only once. This is done because for the same database of SMILES, a variety of structures exists, some more branched than others. The algorithm tries to generate structure for the whole database at once, without any user intervention. It should be noted that the VdW scale is set higher initially to optimize as much as possible the distance between atoms and reduce the computational time required for the local minimization performed with deMonNano (discussed below). This change will decrease the sampling step speed, but increase both geometric relaxation and deMonNano local minimization.
The reduction of the VdW scale parameter is a challenging step, especially for ramified structures. It is a compromise between allowing for the incorporation of structures that are not overly self-colliding (i.e. no bond created during a local minimization using the conjugated gradient from deMonNano for example) and ensuring that the structures are not rejected outright (i.e. structures that could be afterward relaxed using a geometric minimizer). In the event that the number of failures exceeds the required number of failures and the algorithm has already reduced the VdW scale, the algorithm will stop and return the set of structures that have been generated thus far (if no structure was generated, the initial structure will be returned). From the returned structures, a local energy minimization using the conjugated gradient technique of deMonNano code [131] is performed (discussed in the Chapter 2). This method is employed to relax every degrees of freedom of each structures. Finally, the set of generated structures are returned.

4.4 Conclusion

In this chapter, a novel algorithm for the large-scale generation of atomistic models of aromatic hydrocarbons was developed. This innovative approach provides a robust tool for generating diverse molecular structures without prior bias, while adhering to specific chemical constraints. The integration of the SMILES formalism introduces a new level of flexibility and efficiency. The dynamic creation and straightforward encoding of molecular structures enabled by this feature facilitate the widespread sharing and reproduction of data across different platforms, which is precious for collaborative research endeavors. Moreover, the Structure Generator enables the generation of three-dimensional structures from SMILES-generated structures. The algorithm is designed to ensure the structural integrity of the generated models, while also providing a mechanism for optimizing the geometry of the structures to relax atomics collisions. The applications of these generated models extend beyond mere theoretical interest. In the fields of environmental science and astrophysics, for instance, such databases could be employed to identify structures from experimental analysis using infrared techniques for example. Such information is of great importance to deciple chemical processes occurring on Earth and in cosmic environments. Notwithstanding, the chapter also acknowledges ongoing challenges, particularly in terms of computational efficiency and algorithmic complexity. Future work could be conducted on the addition of new fragments possibly including other types of atoms, the management of radical structures that are removed from the algorithm, and the addition of new constraints to the algorithm. Next chapter will focus on the application of the algorithm to the generation of a database of aromatic hydrocarbons.

Chapter 5

Application to Hydrogenated amorphous carbon polymer

Contents

5.1 Introduction

In the vast expanses of galaxies, the spaces between stars, collectively known as the interstellar medium (ISM), are far from empty. Comprising approximately 10% of a galaxy's stellar mass [74], the ISM is a vibrant mix of gases, complex molecules, and dust. Of particular interest is the dust component, which, despite constituting only about 1% of the gas mass, plays a crucial role in the ISM by influencing a myriad of physicochemical processes. The ability of dust particles to absorb and scatter stellar radiation, re-emitting it at longer wavelengths, positions them as central actors in the galactic theater. Two principal types of refractory interstellar dust are recognized: silicates and carbonaceous dust.

In astrophysics, complex structures may be discovered and studied through observations made by telescopes at different wavelengths, from the ultraviolet to the radio. For example, polycyclic aromatic hydrocarbons (PAH) are important molecules in astrophysics, as they are possible starting materials for abiotic syntheses of materials required by the earliest forms of life [134]. In order to identify molecules from observations, researchers require access to a large database containing a wide range of structures with

their key descriptors. For example, studying the infrared (IR) ISM spectrum can be compared to a database of IR spectra to identify each contribution in that spectrum. To obtain such database, a first step is to generate a large number of structures, respecting constraints based on previous IR analysis. As detailed in the Chapter 4, an algorithm was developed to generate a database of structures, respecting constraints based on the analysis from Dartois et al. [41].

For this purpose, a database has been generated and is intended to serve as a reference for future studies necessitating a collection of Hydrogenated amorphous carbon polymer (a-C:H) substructures unit with specific ratios of functional groups. a-C:H, which are defined as PAH precursor [41], are the best candidate to explain the Diffuse ISM absorption observed in our galaxy and other galaxies [41]. Diffuse ISM refers to a less dense region of the ISM. This database is then analyzed based on geometric and electronic descriptors to provide insights into the properties of a-C:H substructures. The results are compared to the properties of PAHs and other carbonaceous structures to identify the similarities and differences between these materials.

5.1.1 Aromatic hydrocarbons in astrophysics

Aromatic hydrocarbons are a class of molecules that play a crucial role in astrophysics. Each of the following structures were defined in the Chapter 4.

Polycyclic aromatic hydrocarbons (PAHs)
The importance of PAHs ranges from astrophysics to environmental science. In astrophysics, PAHs are considered as important constituents of the ISM [154], participating in space chemistry and influencing the formation of other molecules. Their study, in particular through vibrational and emission spectra, helps to understand the chemical composition and processes in the universe [154, 133, 59].

PAHs contribute significantly to the heating of the ISM through the photoelectric effect, where ultraviolet photons eject electrons from PAH molecules, leading to the heating of gas [158]. This process is crucial for maintaining the thermal balance in various astrophysical environments, influencing star formation and the lifecycle of cosmic dust [88]. Furthermore, PAHs are hypothesized to be responsible for the unidentified infrared (UIR) bands observed in many astrophysical objects, including HII regions, planetary nebulae, and the diffuse ISM [5].

Soots
In astrophysics, soot-like particles, often referred to as cosmic dust, play a crucial role in various cosmic phenomena. These particles, primarily composed of carbonaceous materials, are formed in the outflows of carbon-rich stars through nucleation and growth processes similar to those observed in terrestrial soot formation [45]. The study of cosmic dust provides insights into the lifecycle of stars and the evolution of galaxies.

Cosmic soot particles contribute to the ISM by acting as catalysts for the formation of molecular hydrogen (H_2), the most abundant molecule in the universe. These particles provide surfaces where hydrogen atoms can combine to form H_2, a process that is

essential for star formation [161]. Additionally, cosmic dust grains absorb and re-emit stellar radiation, affecting the thermal balance and chemistry of the ISM [48].

Hydrogenated amorphous carbon (a-C:H)

Hydrogenated amorphous carbon represents a significant component of the diffuse ISM in our Milky Way and other galaxies. Composed of intertwined carbon and hydrogen atoms, a-C:H grains are primarily detected through their distinctive infrared absorption bands, which arise from the vibrations of C-H bonds. The spectral signatures of a-C:H grains vary across different sightlines in the galaxy, prompting intriguing questions about their distribution, properties, and the underlying mechanisms driving their spatial and evolutionary variations. Given the difficulties in directly accessing interstellar dust, a significant portion of our knowledge is derived from observations conducted remotely via telescopes and from laboratory simulations that replicate ISM conditions. These studies are further enhanced by analyses of extraterrestrial materials that are more readily available on Earth, such as meteorites and interplanetary dust particles collected from the upper atmosphere. The combination of these approaches provides a comprehensive understanding of the complex processes occurring within the ISM.

The presence of a-C:H grains in the diffuse interstellar medium but their absence in dense interstellar clouds suggests that these particles undergo significant transformations depending on their environment. This discrepancy represents a key focus of this research, as it may provide insight into the lifecycle of interstellar dust and its impact on the ISM. Moreover, the optical emissions from a-C:H grains following UV or visible light absorption, known as photoluminescence, demand a detailed study to ascertain their contribution to the broader spectrum of interstellar emissions. Laboratory characterizations and astrophysical relevance of this material have revealed that a-C:H plays a crucial role in the lifecycle of dust, influencing phenomena such as hydrogen formation, extended red emission, and acting as a precursor to polycyclic aromatic hydrocarbons (PAHs) [77, 69, 140, 46, 21, 94].

Recent observations have revealed that Galactic diffuse interstellar dust features are prevalent in external galaxies [91, 148]. These ISM dust features constitute a significant fraction of the matter (at least 5 to 30% of the carbon cosmic abundance [47, 136]) that is essential for the evolution of dust and solid-phase chemistry in the ISM. The ultraviolet photoproduction of an interstellar dust analog, which Dartois et al. refer to as a-C:H, has been analyzed in detail to understand its potential role in the 2175 Å extinction bump [21].

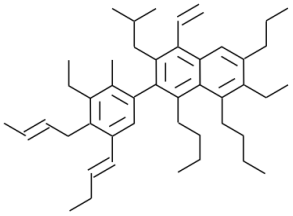
Martín-Doménech, Dartois, and Caro investigated the diffusion of photo-produced hydrogen (H_2) in a-C:H as a function of temperature, demonstrating that the desorption and diffusion of hydrogen in these amorphous carbons are temperature-dependent. The researchers illuminated key mechanisms for the photolysis of ISM dust [116].

Structures with similar IR spectra to the diffuse ISM are shown in [41]. A comparision of these structures with the diffuse ISM (DISM) spectrum has been presented in this article by Dartois et al. The authors have calculated the ratio of functional groups present in the DISM structures. Using IR spectra, the authors define the

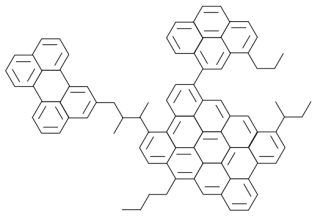
ratio of aromatic C=C bonds to C-X bonds ($ratio_A$), the ratio of olefinic CH2 to aliphatic CH3 ($ratio_B$), and the ratio of aliphatic CH2 to aliphatic CH3 ($ratio_C$) of the observed structures. From this analysis, the authors propose a model of the substructure unit of a-C:H particles. The generation of a-C:H substructures represents a key objective that motivates the development of the algorithm presented in the Chapter 4. This algorithm can create different structures, having different IR spectrum. It is noteworthy that the observed spectra from space are composed of a variety of molecules and conformations. Obtaining a variety of structures is an interesting opportunity to analyze the impact of structural differences on the IR spectrum.

5.1.2 Methodology

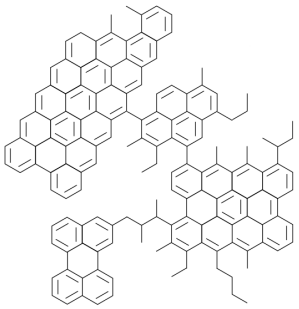
In their study, Dartois et al. (2005) produced a-C:H through the photolysis of various organic molecule precursors at low temperatures. The substructures unit of a-C:H are believed to be representative of a constituent from the DISM. From this film with IR analysis, the olefinic and aliphatic ratios were estimated. Results show that the a-C:H substructures exhibited a ratio of olefinic CH_2 to aliphatic CH_3 between 0.05 and 0.1. The aliphatic CH_2 to aliphatic CH_3 ratio was evaluated between 1.80 and 2.20. The aromatic contribution was evaluated and the results suggested that the ratio of aromatic C=C in the a-C:H network is between 0.05 and 0.2.



(a) Species A



(b) Species B

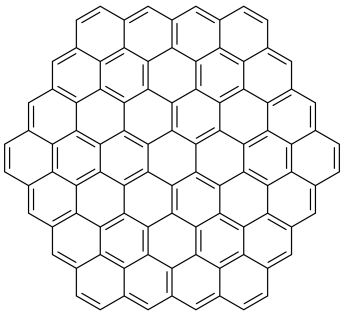


(c) Species C

Figure 5.1: Structures taken from [41]. The design of species A was carried out on ratios of olefinic, aliphatic, and aromatic compounds, as presented in [41]. The design of species B and C was guided by the generally accepted structure of carbonaceous interstellar dust, as outlined in Pendleton and Allamandola [126].

To facilitate comparison of the a-C:H spectra to the DISM spectra, three species (presented in Figure 5.1) were selected that largely adhere to the detailed ratios. The species A (Fig. 5.1a) was designed to respect the ratios, while the B and C (Fig. 5.1b and 5.1c) were selected from the generally accepted structure of carbonaceous interstellar dust, as outlined in Pendleton and Allamandola (2002).

The generation of random structures that respect this ratio represents a significant challenge. The algorithm developed in the previous chapter was employed here to gen-



(a) Circumcircumcoronene ($C_{96}H_{24}$)

(b) Linear structure ($C_{96}H_{194}$)

Figure 5.2: Illustration of the limit cases.

erate a-C:H substructures that adhere to the ratios of olefinic, aliphatic, and aromatic compounds. The database was generated with the same number of carbon atoms as the Circumcircumcoronene ($C_{96}H_{24}$, Fig. 5.2a). The number of carbon atoms was selected to ensure that the constraints were respected, with a relatively high number of atoms being necessary for this purpose. In addition, a set of linear structure ($C_{96}H_{194}$, Fig. 5.2b), again with 96 carbons atoms, was generated for comparision. The two aforementioned structures served as limiting cases for the purpose of comparing the generated structures. The circumcircumcoronene is characterized by a complete aromatic character, whereas the linear structure is that of an alkane.

In the following section, the database is firstly analyzed to determine the distribution of the structures based on the ratios of olefinic, aliphatic, and aromatic compounds. Geometric and electronic descriptors are then computed to analyze the shapes and the electronic properties of the structures. The results are compared to the properties of circumcircumcoronene and linear structures to identify the similarities and differences between these materials.

5.2 Database analysis

The study was conducted in accordance with the stipulated ratios and total number of carbon atoms. This results in the following constraints:

- The total number of carbon atoms is fixed to 96 (same as in the circumcircumcoronene PAH, see Fig. 5.2a)
- $ratio_A$: $N_{aromaticC=Cbonds}/N_{aromaticC-Xbonds}$ between 0.05 and 0.2
- $ratio_B$: $N_{olefinicCH2}/N_{aliphaticCH3}$ between 0.05 and 0.1
- $ratio_C$: $N_{aliphaticCH2}/N_{aliphaticCH3}$ between 1.80 and 2.20

Generated vs theoretical distribution

Prior to generating the structure of the a-C:H molecules, a numerical analysis was conducted to ascertain the distribution of the structures following the different ratios. The results are presented in Fig. 5.3. To generate this figure, all possible combinations with 96 carbon atoms were created and only atom-based ratios were calculated. Calculating the $ratio_A$ is more challenging because it is based on the number of aromatic bonds divided by the total number of bonds in the structure. This necessitates a complete knowledge of the connectivity of the structure. Consequently, the $ratio_A$ is not considered in this plot. This distribution could be contrasted with the distribution of the database (Fig. 5.4).

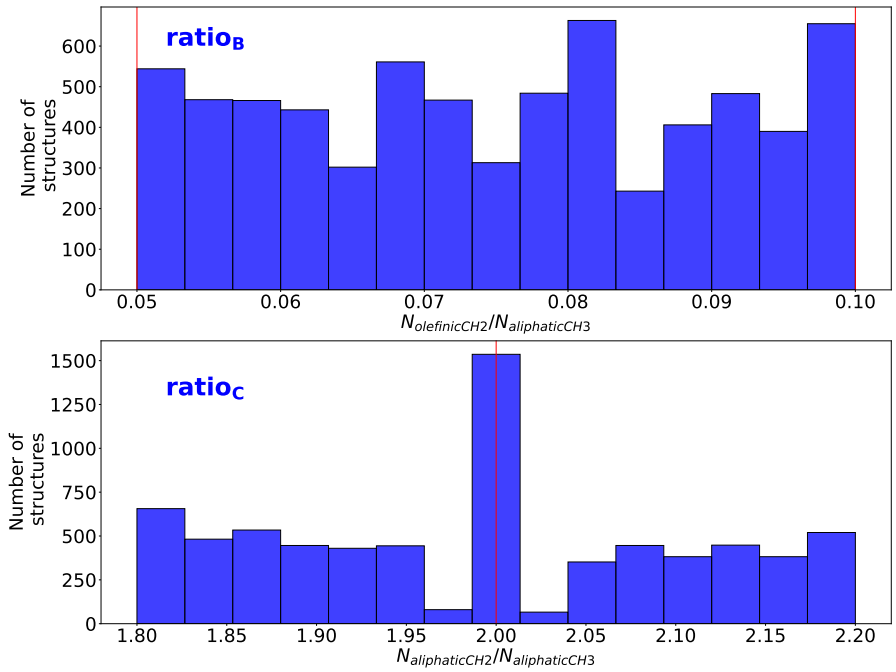


Figure 5.3: Theoretical distribution of ratios B and C for the generated structures. The red lines represent the constraints given in the article by Dartois et al. [41].

As can be observed, the distribution between the two figures exhibit some differences in area containing few structures in Fig. 5.3. For example, no structure was found in the database (Fig. 5.4) with a $ratio_C = 1.97$. This is due to the fact that the ratio A is considered in the second figure, and the addition of this constraint leads to an impossibility to generate structures that have a $ratio_B = 0.065$ for example. Moreover, the distribution of the structures is not uniform, meaning that for a bar in one ratio, the distribution of the structures in the other ratios could be different. This is illustrated in Fig. 5.5a, for the $ratio_A$ between 0.05 and 0.06, distributions of $ratio_B$ and $ratio_C$ are different from the Fig. 5.4. The same comment can be made for the $ratio_A$ between 0.06 and 0.07 in the Fig. 5.5b.

It should be noted that each figures presented at this step represent the distribution of the generated SMILES, which is not the distribution of the 3D structures generated at

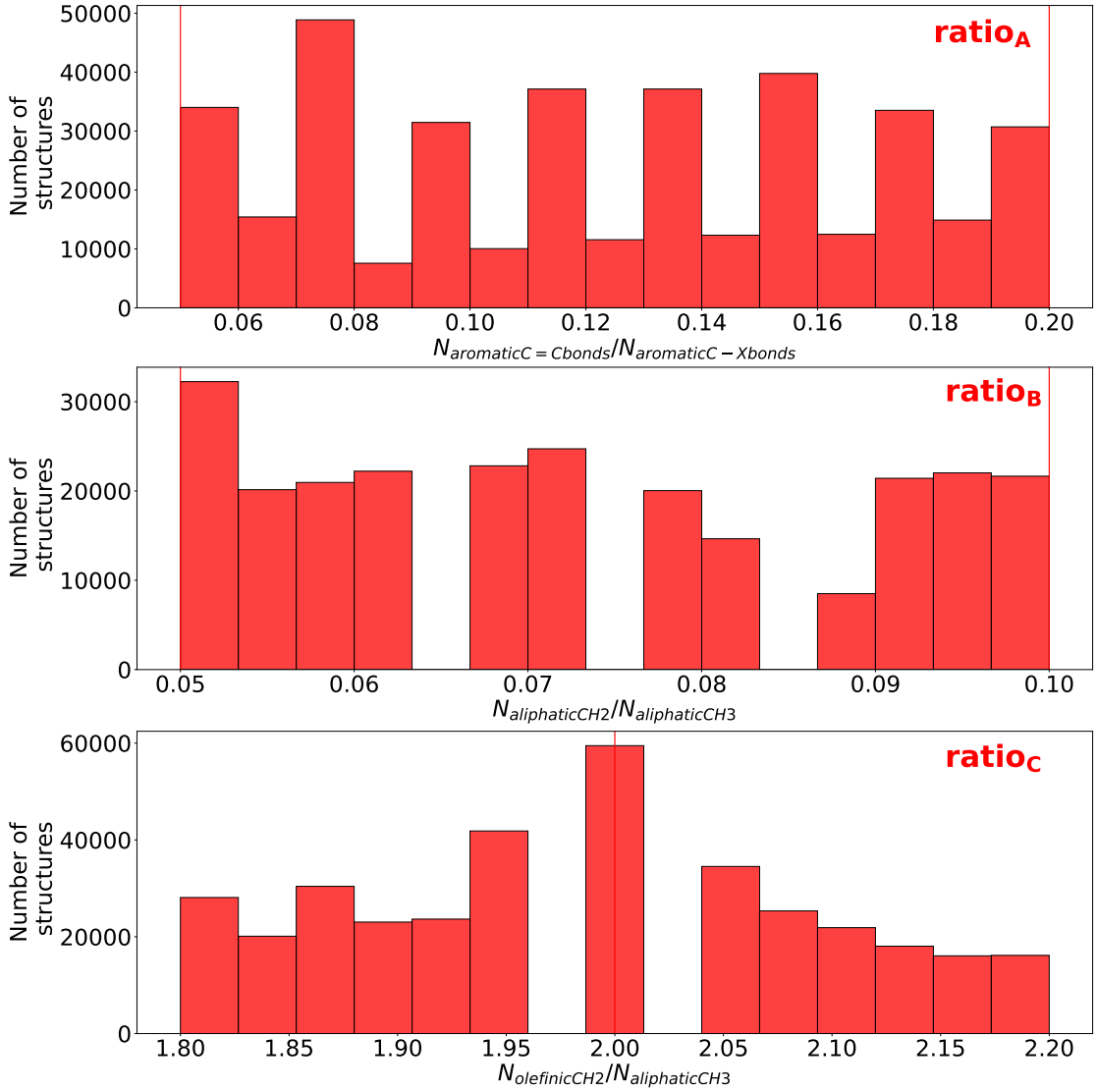
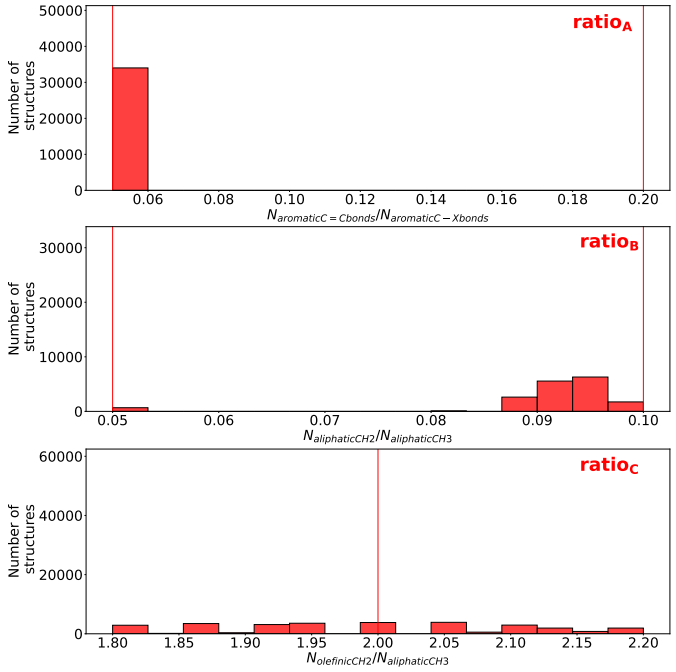
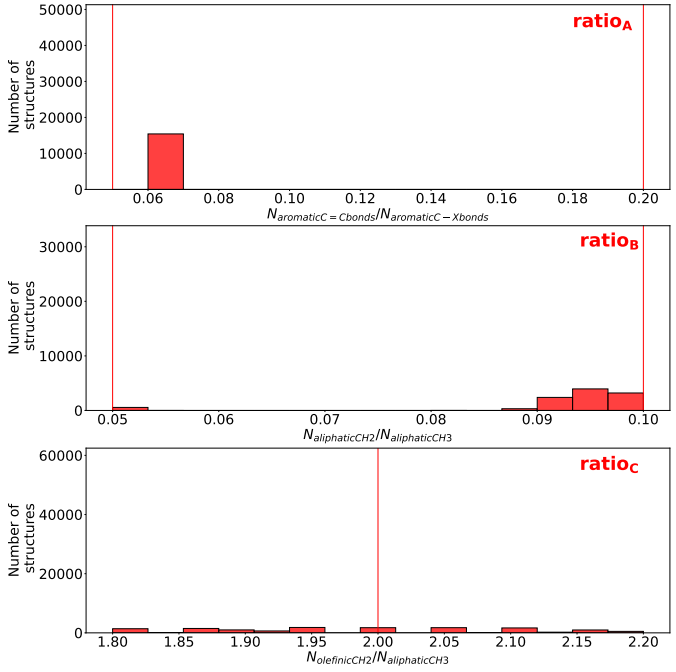


Figure 5.4: Distribution of the ratios for the generated structures. The red lines represent the constraints given in the article by Dartois et al. [41]. Each bar represents every structure available at a given interval ratio without considering the other ratios. For example, for the bar of the $ratio_A$ between 0.05 and 0.06, $ratio_B$ could assume values between 0.05 and 0.10.



(a) Structures at a $ratio_A$ between 0.05 and 0.06, while $ratio_B$ and $ratio_C$ are not fixed.



(b) Structures at a $ratio_A$ between 0.06 and 0.07, while $ratio_B$ and $ratio_C$ are not fixed.

Figure 5.5: In this figure, the value of $ratio_A$ is fixed, while the values of $ratio_B$ and $ratio_C$ are not. This illustrates that for a given ratio, the distribution of structures is not uniform across the other ratios.

the next step.

Filtering and selection strategy
Structures containing radical fragments have been removed from the database because they are not relevant for the study developed in this chapter. In addition, this type of structure requires a specific electronic analysis to determine its multiplicity in order to obtain its energetic descriptors and to be able to compare it with other structures in the database. Removing this type of structure leads back to removing structures with an odd number of atoms, since the number of carbon atoms is fixed at 96.

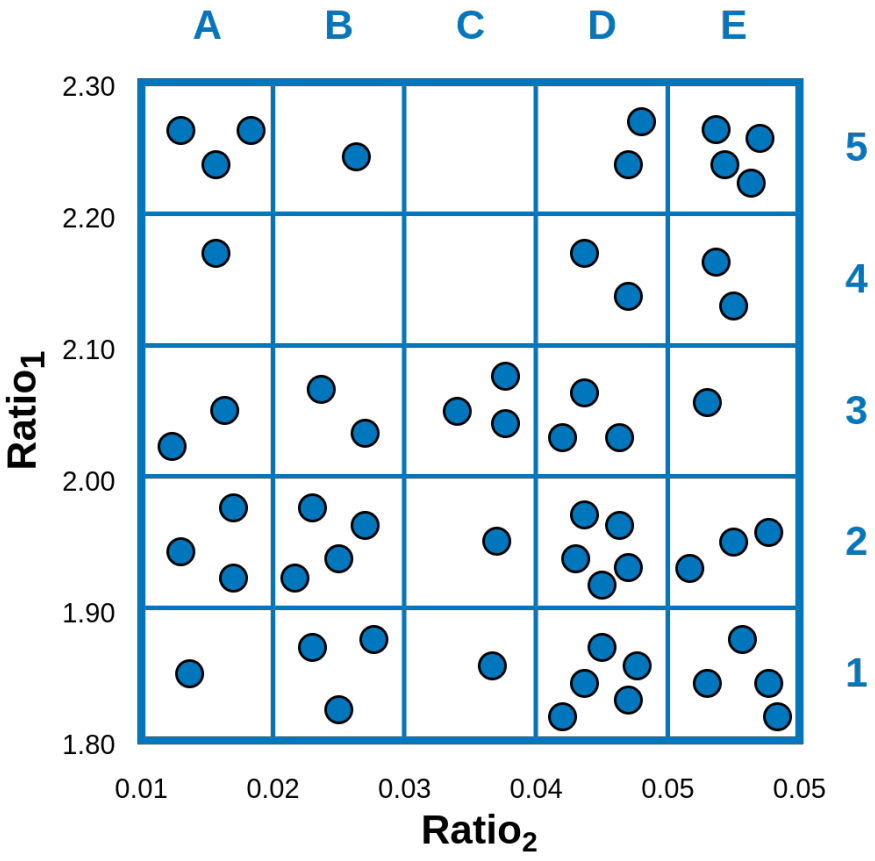


Figure 5.6: Database distribution for structures only based on two ratios. Each point represent a structure, which is contained in a square (i.e. interval for the two ratios). For example, in this database, 3 structures are available in A5, for a $ratio_1$ between 2.20 and 2.30 and a $ratio_2$ between 0.01 and 0.02 while no structure was generated in the range of the square C5.

Subsequently, the structures to be generated are selected according to the procedure outlined in Fig. 5.6. As illustrated in this figure, a database comprising only two ratios is depicted. It is noteworthy that the identical strategy is employed for the aforementioned database. In this figure, each conformation, defined by its ratio, is

represented by a point in a square. Each square is defined by the interval of the two ratios. The strategy employed for the selection of structures is an iterative process, whereby ratios are examined. The initial step is to search for a structure within A1. A single structure is identified and transferred to the generation stage. This structure couldn't be selected during subsequent iterations of the selection stage. This process is repeated for each square, unless no structure is available in that square. It is important to note that if no structure is found in a given interval, that interval is excluded from further selection. Then, the selected SMILES are used in the structure generator to obtained the 3D structures.

5.3 Descriptors definition

The generated structures obtained are analyzed based on both geometric and electronic descriptors. Each of them was chosen because it represents a specific property of the generated database. The definitions of these properties can be found in the following subsections.

5.3.1 Geometric and structural descriptors

A pertinent geometric descriptor for characterizing the database are the Hill-Wheeler parameters [31]. These parameters define the asphericity, i.e. the deformation of the structure from a perfect sphere. In order to discuss the asphericity, it is necessary to define the center of mass of the molecule. This is calculated as follows:

$$R_{com} = \frac{\sum_{k=1}^{N} m_k r_k}{\sum_{k=1}^{N} m_k} \tag{5.1}$$

where m_k is the mass of the atom k. This center of mass is used to recenter the molecule. Then the inertia tensor is defined as:

$$I = \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix} \tag{5.2}$$

Each component are generally defined as:

$$I_{ij} = \sum_{k=1}^{N} m_k (\|r_k\|^2 \delta_{ij} - x_{i,k} x_{j,k}) \tag{5.3}$$

where $r_k = (x_{1,k}, x_{2,k}, x_{3,k})$ is the position of atom k with mass m_k. The indices i, j take on values of 1, 2, or 3 for the Cartesian coordinates $x_{1,k}, x_{2,k}, x_{3,k}$, respectively. The Kronecker delta symbol is defined as:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{5.4}$$

Using Equations 5.2, 5.3 and 5.4, diagonal elements of the inertia matrix is given by:

$$\begin{array}{l} I_{11} = \sum_{k=1}^{N} m_k(x_{2,k}^2 + x_{3,k}^2) \\ I_{22} = \sum_{k=1}^{N} m_k(x_{1,k}^2 + x_{3,k}^2) \\ I_{33} = \sum_{k=1}^{N} m_k(x_{1,k}^2 + x_{2,k}^2) \end{array} \tag{5.5}$$

and off-diagonal elements are defined as:

$$\begin{array}{l} I_{12} = I_{21} = -\sum_{k=1}^{N} m_k x_{1,k} x_{2,k} \\ I_{13} = I_{31} = -\sum_{k=1}^{N} m_k x_{1,k} x_{3,k} \\ I_{23} = I_{32} = -\sum_{k=1}^{N} m_k x_{2,k} x_{3,k} \end{array} \tag{5.6}$$

Principal moments of inertia are defined as the eigenvalues of the tensor I, which are ordered as $I_1 \geq I_2 \geq I_3 \geq 0$. These are computed by the diagonalization of the tensor I. The Hill-Wheeler parameters are then defined as:

$$I_p = \frac{2}{3} r_c^2 \left[1 + \beta \sin\left(\gamma + \frac{(4k-3)\pi}{6} \right) \right], \quad p = 1, 2, 3 \tag{5.7}$$

where r_c is the root mean square radius defined by:

$$r_c = \left(\frac{1}{N} \sum_{k=1}^{N} r_k^2 \right)^{1/2} \tag{5.8}$$

The shape parameter β is defined in the range $[0, 1]$, and it is used to measure the oblateness of the cluster. Meanwhile, the parameter γ is defined in the range $[0, \pi/3]$, and is used to measure the cluster triaxiality. A perfect sphere is defined by the parameters $\beta = 0$ and whatever value for the γ parameter. An axially symmetric prolate ellipsoid is defined by the parameters $0 \leq \beta \leq 1$ and $\gamma = 0$, while an axially symmetric oblate ellipsoid is defined by the parameters $0 \leq \beta \leq 1$ and $\gamma = \pi/3$. The Hill-Wheeler parameters are illustrated in Figure 5.7.

These parameters are computed as follows:

$$\begin{array}{l} \beta = \frac{I_1 - I_2}{\sqrt{3}\sin(tga) I_{mean}} \\ \gamma = \frac{180}{\pi} tga \end{array} \tag{5.9}$$

In practice, when the principal momenta of inertia are known, $tga = \arctan\left(\sqrt{3} \frac{I_2 - I_1}{2I_3 - I_2 - I_1} \right)$ and $I_{mean} = \frac{I_1 + I_2 + I_3}{3}$. The determination of these two parameters enables the analysis of the variety of shapes exhibited by the generated structures.

Other straightforward descriptors are defined to analyze the database, such as the number of aromatic units by structure, which defines the number of aromatic cycles in a structure. For instance, the coronene ($C_{24}H_{12}$) molecule has seven aromatic units, while the molecule in Fig. 5.8 has six aromatic units. Moreover, the number of islands is defined as the number of disconnected aromatic fragments that are not connected by

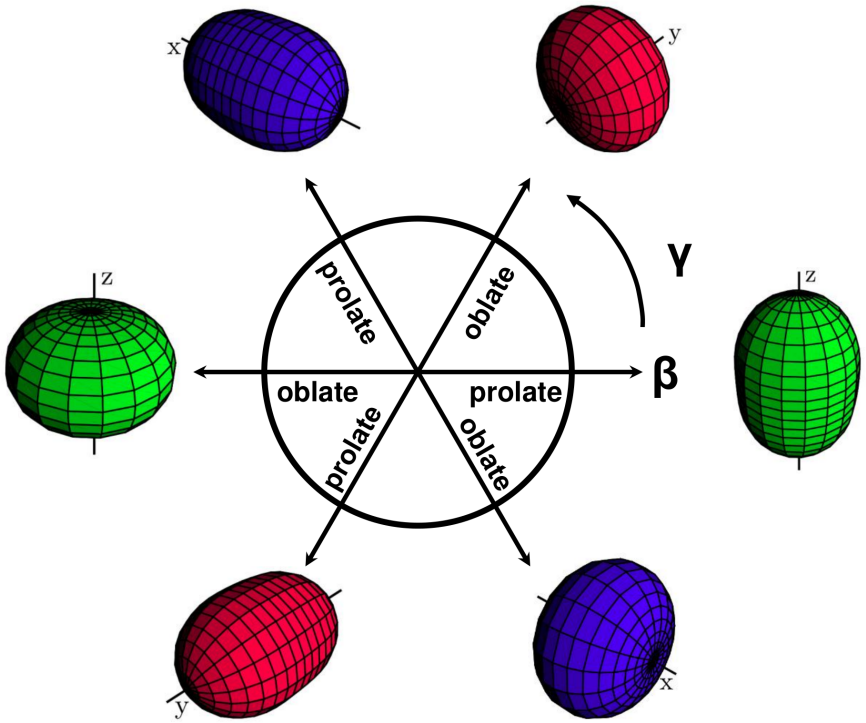


Figure 5.7: Hill-Wheeler parameters, β and γ are the asphericity parameters, respectively. Adapted from Fortunato[63]. Note that $\gamma = n\frac{\pi}{3}$ with $n = 0, 1, 2, 3, 4, 5$

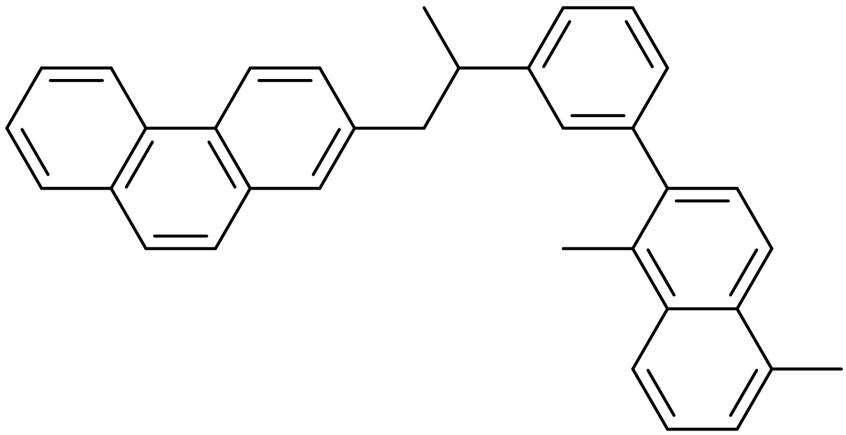


Figure 5.8: Illustration of $C_{35}H_{30}$.

an aromatic bond together. For example, the coronene has only one island, while the $C_{35}H_{30}$ molecule represented in Fig. 5.8 has three islands. Then, the maximum size of the islands for a molecule is defined as the number of aromatic units in the largest island. For instance, the maximum size of the islands for coronene is seven, while that of the molecule in Fig. 5.8 is three. The last simple descriptor used in the following anaylis is the number of atoms for each structures. The number of carbon atoms being fixed at 96, this descriptor characterizes the number of hydrogen atoms in the structure.

5.3.2 Electronic descriptors

In addition to geometric descriptors, electronic descriptors are employed to characterize the generated structures. A number of descriptors were calculated, including the band energy, the repulsive energy, the electronic energy, the Fermy energy level, and the (HOMO)-(HOMO-1) gap. However, these descriptors did not exhibit the requisite behavior to effectively characterize the elements within the database.
The first discussed descriptor is the gap HOMO-LUMO, which is the difference between the energy of the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). This descriptor is available on the deMonNano code and is illustrated in Fig. 5.9. In addition, the London energy is a dispersion energy, named after the physicist Fritz London who first described it in the 1930s [51]. It is a type of van der Waals force that is induced between atoms or between non polar molecules. This force arises due to the fluctuations in electron distribution within atoms and molecules, leading to temporary dipole moments.

Subsequently, the ionization energy is employed, serving as a metric for the energy required to remove an electron from an atom or molecule, thereby forming a positively charged ion (cation). This is a pivotal electronic descriptor that furnishes data regarding the stability and reactivity of a system. The ionization energy can be calculated as the discrepancy in energy between the cationic and the neutral system. Finally, electronic

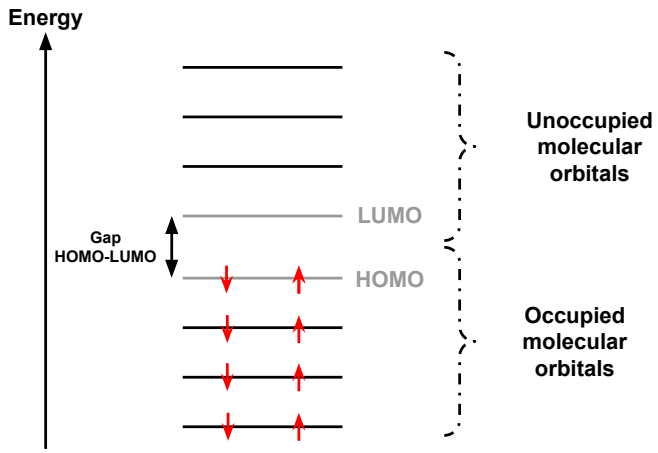


Figure 5.9: Illustration of the gap between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO)

affinity represents the energy change when an electron is added to a neutral atom or molecule, resulting in the formation of a negatively charged ion (anion). It is a measure of the tendency of an atom or molecule to accept an electron. The electronic affinity can be calculated as the difference in energy between the neutral and the anionic system. The computation of both ionization energy and electronic affinity is based on the geometry of the neutral form, with the addition or removal of an electron.

5.4 Results

From the presented descriptors, an analysis is made comparing both structural and electronic descriptors. The results are presented in the following sections. For each figure there are extreme cases which are the circumcircumcorone in orange and the linear structures in green which contain the same number of carbon atoms as the circumcircumcoronene.

5.4.1 Geometry-based analysis

The Hill-Wheeler parameters β and γ are employed to analyze the generated structures. The results are presented in Fig. 5.10. The histograms on the right and top of the plot show the distribution of β and γ, respectively. The circumcircumcoronene is oblate, a result that is anticipated given the planar nature of the structure. Most of the linear structures are prolate with some exceptions. As displayed, structures in the database exhibit a wide range of shapes. Most of the generated structures are rather spherical, as indicated by the small values of β. Nevertheless, some structures exhibit prolate characteristics, as evidenced by the elevated values of β, indicating a degree of asphericity.

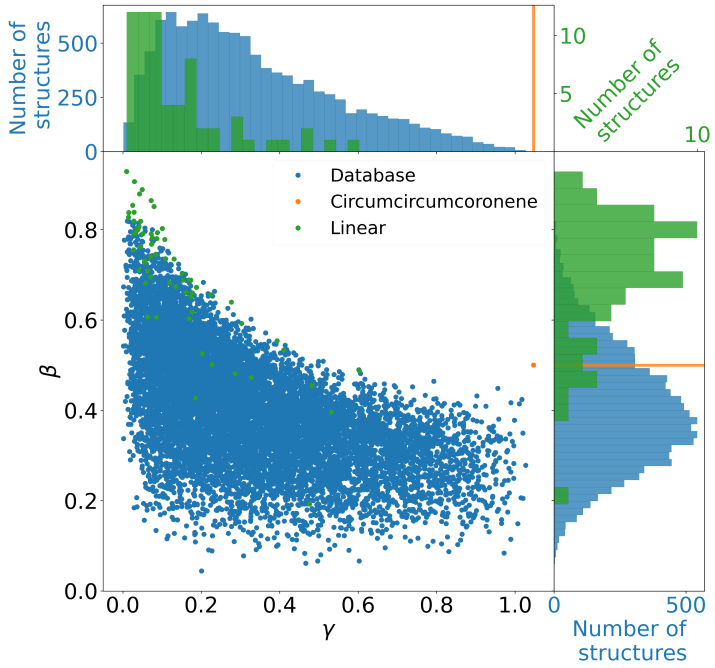


Figure 5.10: Hill-Wheeler parameters β and γ for the generated structures. The distribution of structures is shown on the right and top of the plot for β and γ. The scales employed for linear structures and those for database structures differ, given that the latter is a more numerous category. The circumcircumcoronene has one structure and is shown as an orange segment.

5.4.2 Structure-energy relationships

HOMO-LUMO gap distribution is illustrated in the Figure 5.11 according to different descriptors. Firstly, linear structures exhibit a HOMO-LUMO gap around 13 eV, whereas circumcircumcoronene is observed at 1.4 eV. The generated structures are found to be relatively close to the circumcircumcoronene, with some exhibiting a smaller HOMO-LUMO gap and a specific family appearing around 5 eV. As can be seen at the bottom right of the Fig. 5.11, this family has a maximum size of islands equal to one, indicating that this structure contains only single benzene spreads. Moreover, the number of aromatic units present in a given structure within this family ranges from three to five. When an island with a size of naphthalene or larger is present in the structure, the gap is less than 4 eV. Conversely, the gap roughly increases as the maximum size of islands decrease as can be seen at the bottom right of the Fig. 5.11.

London energy distribution is illustrated in the Figure 5.12. The circumcircumcoronene or the linear structure exhibit a lower degree of stabilization in terms of London energy than the structures from the database. The bottom right of the Figure 5.12 indicates that stabilization increases with the number of atoms for the database, showing

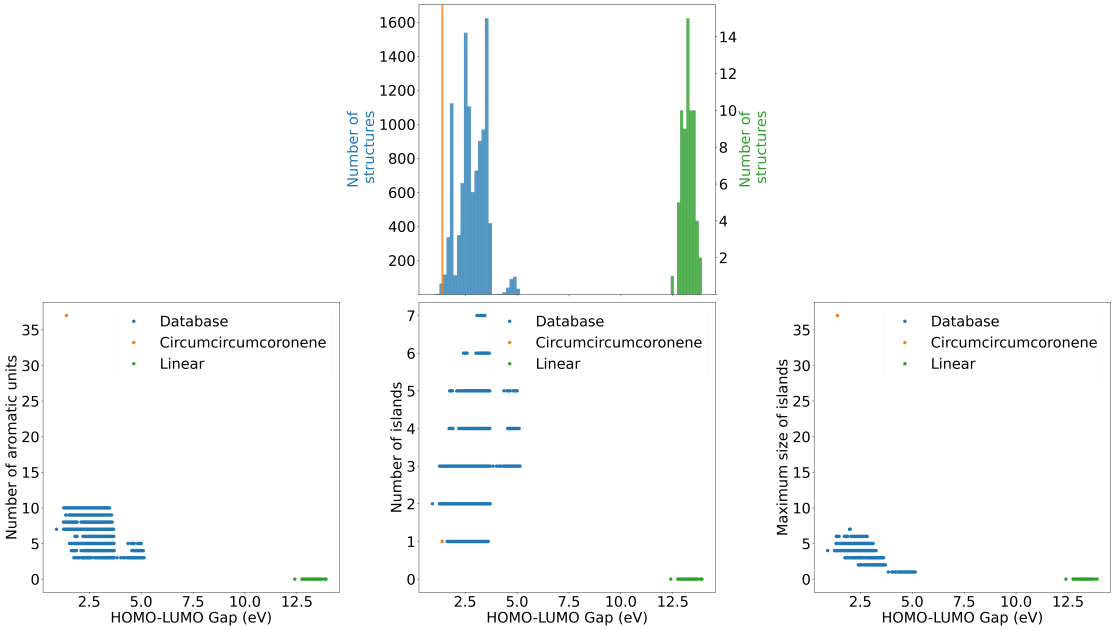


Figure 5.11: HOMO-LUMO gap of the database.

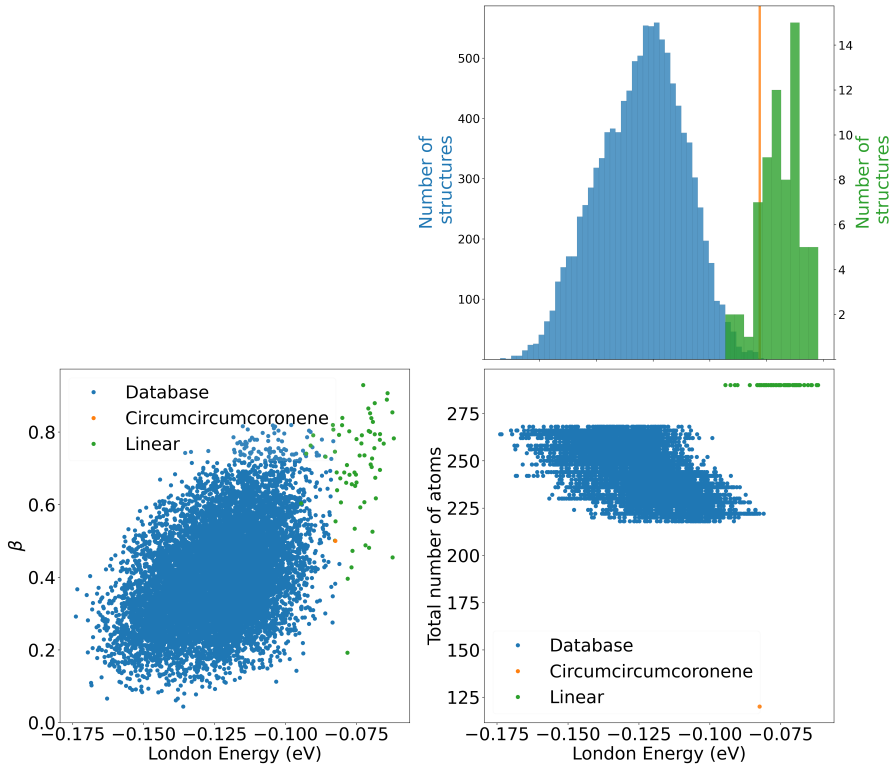
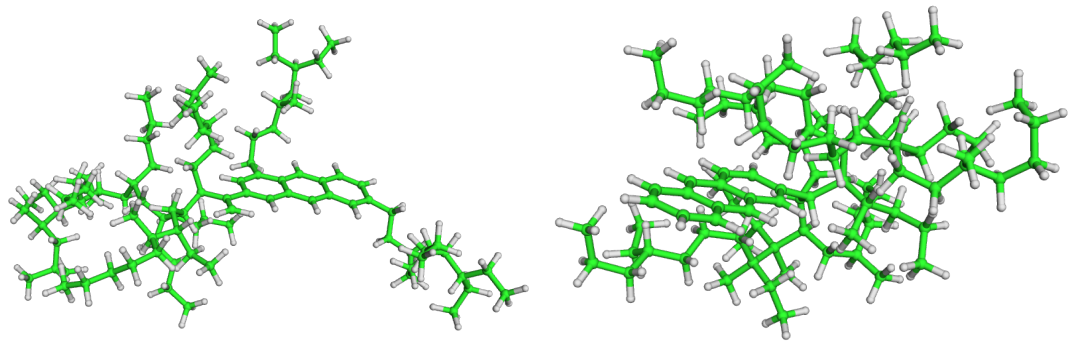


Figure 5.12: London energy of the database.



(a) Extended structure (large β) with a low stabilization based on London energy.

(b) Compact structure (small β) with a high stabilization based on London energy.

Figure 5.13: Illustration of the stabilization based on London energy.

that the larger the structures, the lower the London energy is (which is stabilizing). An expected result would be that stabilization would be greater for linear structures than for database structures, which is not the case. This phenomenon can be attributed to aromatic compounds facing hydrogens atoms for structures of the database (see Fig. 5.13). The structures with the most stabilizing London energy are compact and spherical, while those with less stabilizing London energy are more extended. This is evidenced by the bottom left of the Figure 5.12.

Ionisation energy is illustrated in the Figure 5.14. The ionization energy of circumcircumcoronene is 5.90 eV, while that of linear structures is approximately 8 eV. This suggests that the delocalization of charge in the π system is more readily achieved than in linear structures. The database encompasses a range of values from 5.5 to 7.35 eV. The loss of an electron is less unfavorable for a-C:H substructures than for linear structures, as charge delocalization occurs. The larger the island, the more favorable it is to lose electrons for the database. But this is not the only factor, otherwise the ionisation energy of the circumcircumcoronene should be under the database. Following observations, for a fixed number of the maximum size of islands, the ionization is favored for structures having multiple islands distributed in the structure. Structures with this island distribution could have better charge delocalization than the circumcircumcoronene.

Electronic affinity is illustrated in Figure 5.15. Circumcircumcoronene is situated around 2.4 eV, rendering it conducive to electron capture. In contrast, linear compounds exhibit a markedly negative electron affinity, which renders it not conducive to electron capture. It is observed that poor electron affinity is exhibited by linear structures and structures with a maximum island size of one, and that this improves as the maximum island size increases, as well as for circumcircumcoronene. The database encompasses a range of energies between -1 and 2 eV, exhibiting a behavior intermediate between the two other types of structures. In the database, capturing an electron therefore becomes a favorable phenomenon from a certain size of island.

Figure 5.16 allows for the observation of the shift in electronic affinity with respect

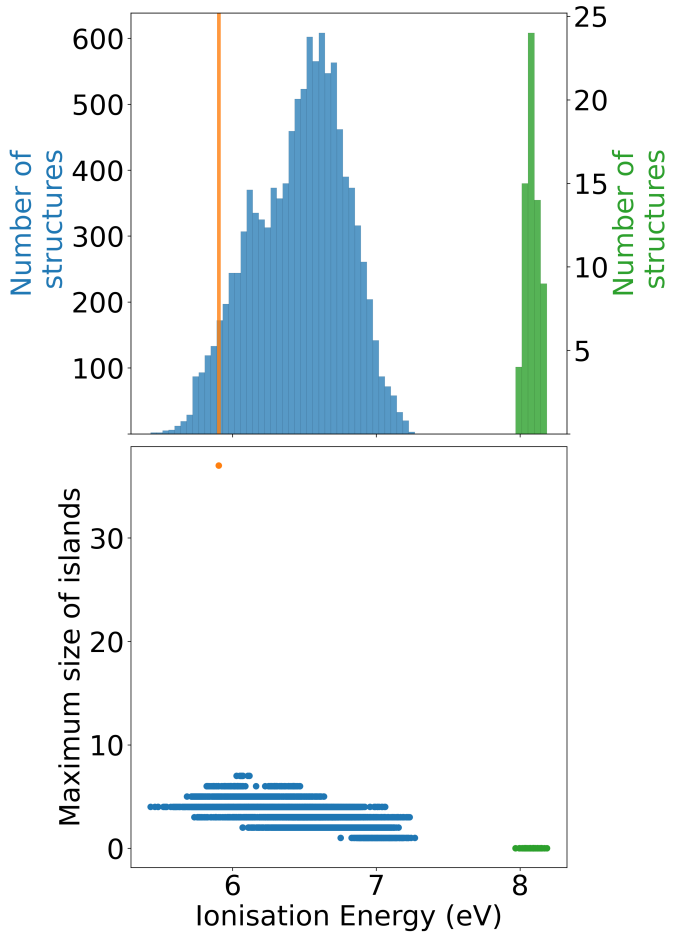


Figure 5.14: Ionisation energy of the database.

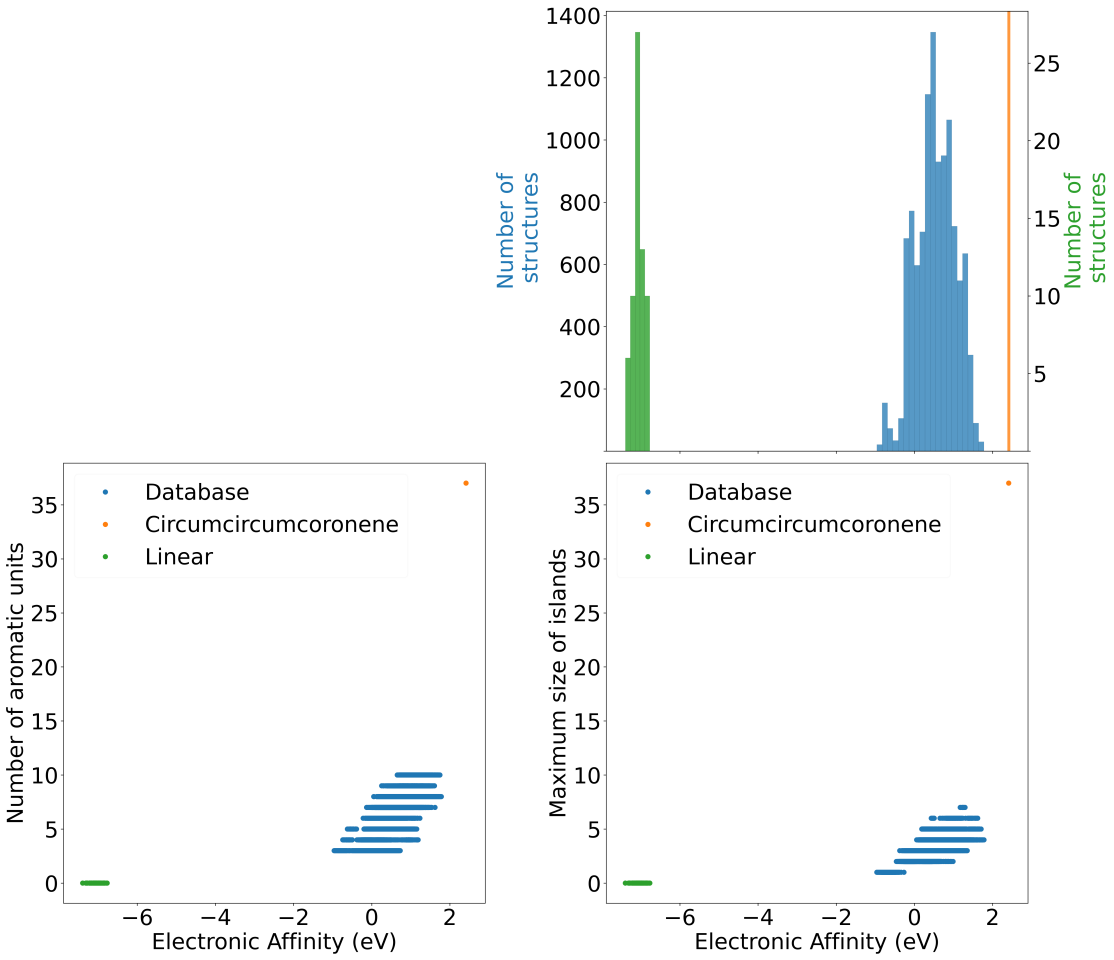
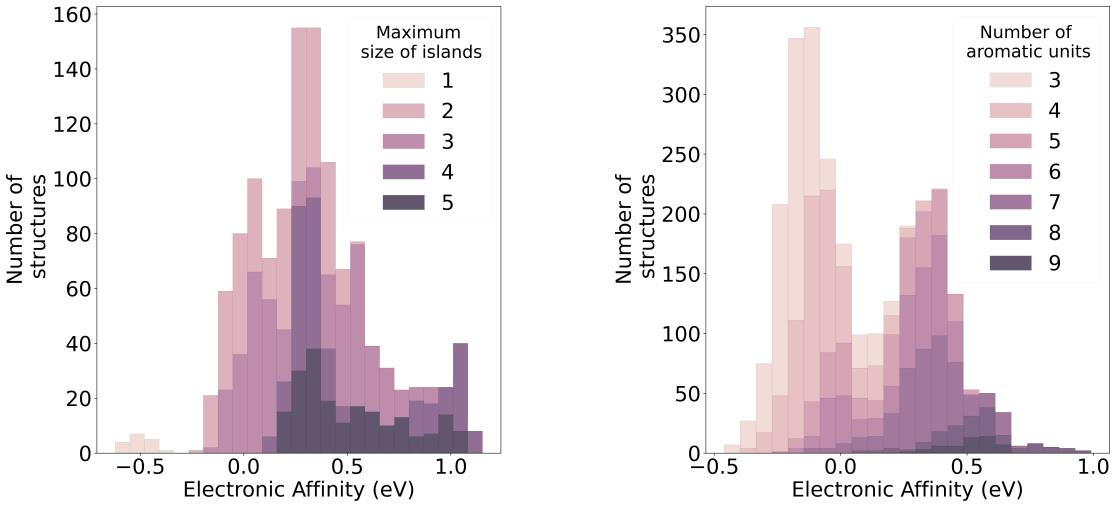


Figure 5.15: Electronic affinity of the database.



(a) Structures with a number of aromatic units equal to five.

(b) Structures with a maximum size of islands equal to two.

Figure 5.16: Illustration of the shift in electronic affinity with respect to the maximum size of islands and the number of aromatic units.

to the number of aromatic units and the maximum size of islands. In order to prevent a cumulative effect between the number of aromatic units and the maximum size of islands, the number of aromatic units for figure 5.16a and the maximum size of islands for figure 5.16b were fixed. The electronic affinity is more favorable for structures with a larger maximum island size, as evidenced by the shift in the distribution of electronic affinity. This is due to the fact that the larger the island, the less the structure is destabilized by electron capture. A similar trend is observed in structures with a larger number of aromatic units.

5.5 Conclusion

The analysis of the database generated in this study provides a comprehensive understanding of the relationship between the geometric and electronic descriptors of a-C:H structures. The Hill-Wheeler parameters β and γ were employed to analyze the shapes of the structures, revealing a wide range of shapes, from spherical to prolate. The electronic descriptors, including the HOMO-LUMO Gap, London energy, ionization energy, and electronic affinity, were employed to characterize the database. The results indicate that the generated structures exhibit a variety of electronic properties.
We observed that: (i) generated a-C:H structures are mostly spherical, (ii) a family with isolated rings exists in the database with a higher HOMO-LUMO gap than the rest of the structures, (iii) the stabilization from the London energy increases with the number of atoms, (iv) loss of an electron is less unfavorable for structures with bigger islands due to charge delocalization, (v) the capture of an electron becomes a stabilizing phenomenon when the the maximum size of islands increases.

The database represents a valuable resource for further research into the properties of a-C:H structures and their potential applications in the field of astrophysics. The database will be released to be accessible with the different descriptors for the community to be used. Moreover, the generation of the IR spectrum of each structures is still under development, and will be added to the database. These spectra may help to understand the type of structures present in differents studied sample. For example, it can help in identifying structures from the diffuse ISM spectra.
The algorithm used to generate the database can be adapted to other systems, thereby providing a versatile tool for the study of complex molecular structures. The constraints on the ratio could be altered to encompass different values, transformed into a different type of ratio, or simply removed to generate a database with a distinct distribution of structures. Future work could focus on the application of the database to other systems, such as polycyclic aromatic hydrocarbons (PAHs) or other carbon-based materials, in order to further explore the relationship between structure and properties. An extension to the structure based on other atoms could be considered.

Chapter 6

A stochastic approach for transition paths search

Contents

6.1 Introduction

Exploring the conformational space of a chemical system is a fundamental task in computational chemistry, for which many methods of PES exploration have been developed, as presented in Chapter 1. Among these methods, we presented in Chapter 2 a method, coupling the IGLOO global exploration algorithm and a DFTB potential, to obtain various local minima on a PES, and its application to phthalate molecules was described in Chapter 3. However, a missing piece of information remains, namely the displacements required to go from one state to another. Methods have been established to explore the transition paths, which allow to connect each state with each other. Some algorithms focus on finding the minimum energy path (MEP), i.e. the path between states where each point is located at a minimum for all directions tangent to the path. Other algorithms are stochastic, resulting in the identification of multiple potential paths from one state to another.

The potential energy used to evaluate the energy of conformations along a path is defined by models with varying degrees of accuracy, as presented in Chapter 1. Consequently, a bias may result in a path that is different from the experimental data. Obtaining multiple paths reduces the impact of potential energy biases, since a path favored by a given potential energy may differ from those favored by other potential energies.

From the multiple paths generated, patterns could be identified and a probability of observing a path could be calculated. To automatically determine these patterns and discretized classes, clustering algorithms are used.

This chapter develops a preliminary method to identify the different low energy paths between two conformations of a molecule using a robotic-inspired algorithm named Transition based Rapidly-exploring Random Trees (T-RRT). Due to the stochasticity of the algorithm, a method was implemented to differentiate and group trajectories. For this purpose, a review of measures and a presentation of a clustering method to group similar paths into a class of paths are made. Then each representative trajectory of a class is locally optimized using the Nudged Elastic Band (NEB) method to obtain a minimized path between the two conformations. Note that this last part is only presented in the method section, but not yet fully implemented.

6.1.1 Exploring transition path space

To describe the dynamics of conformational changes, the Transition State Theory (TST) [57, 89, 165] is used. This theory is central to understand how conformational changes progress from a state to another through a transition state (presented in Chapter 1). This chapter focuses on the conformational changes of a molecule between two states but the presented methods can also be used to characterize a reaction pathway.
In computational chemistry, algorithms that explore the transition path space can be broadly categorized into two types: those that explore transition paths while searching for transition states (requiring only a starting state) and those that are directly aimed at exploring transition paths (depending on the algorithms, some require only the starting state, others require both starting and final states). Note that in the latter category, a distinction must be made between deterministic algorithms, which focus most of the time on finding the MEP, and stochastic algorithms, which generate multiple potential paths.

6.1.1.1 Saddle point search algorithms

Algorithms that search for transition states are designed to locate the high-energy configurations along the reaction coordinate that act as a shifting point (detailed in the Chapter 1) between different states.

Dimer method The Dimer method [84] is used to locate transition states on a potential energy surface, thereby facilitating the study of reaction mechanisms. Developed to overcome certain limitations inherent in other transition state search methods, such as the NEB method (explained below), the dimer method provides a more direct approach to find saddle points without requiring an initial guess of the reaction path.

The core concept of the dimer method involves the use of a "dimer", which is essentially two points or configurations that are close to each other on the potential energy surface. This dimer is used to probe the curvature of the energy landscape by rotating and translating geometry in space to minimize the energy along one direction while maximizing it along another. This process is facilitated by the calculation of the Hessian

matrix, or an approximation thereof, which is less computationally costly compared to methods that require the full Hessian calculation.

The Dimer method efficiently find the highest curvature directions by orienting the dimer along low curvature directions and thus, indirectly probing the higher curvature directions which correspond to the reaction coordinate. This characteristic makes the dimer method particularly valuable for systems with complex energy landscapes, including those with multiple minima and saddle points.

However, the dimer method has its drawbacks. It can be sensitive to the choice of initial configurations and can converge slowly in the case of very flat landscapes. This sensitivity can lead to increased computational time compared to other methods if not properly managed. In addition, while the method is less dependent on a full Hessian calculation, the accuracy of the approximations used can affect the precision of the saddle point found.

Conjugated Peak Refinement (CPR) The Conjugated Peak Refinement (CPR) method by Fischer and Karplus [58] is a technique used to locate saddle points on potential energy surfaces. This method iteratively refines an initial guess of the saddle point by following a conjugate direction that effectively balances the need to ascend and descend the potential energy landscape. The process begins with an initial point, and then alternates between moving along the direction of the negative gradient and a conjugate direction that maintains orthogonality to previous search directions. By continuously adjusting these directions, the CPR method converges towards the saddle point.

Compared to the Dimer method, the primary difference between with CPR lies in their operational strategies. CPR utilizes conjugate directions to maintain orthogonality and efficiently converge to the saddle point, which can be more effective in complex landscapes with multiple minima and maxima. In contrast, the Dimer method's reliance on detecting negative curvature directions allows it to be particularly adept at finding saddle points directly associated with reaction pathways. Additionally, the Dimer method often requires fewer iterations to identify the saddle point due to its direct focus on negative curvature, whereas CPR might involve more extensive searches to refine the saddle point accurately. Both methods are valuable in their respective contexts and are chosen based on the specific characteristics of the potential energy surface being analyzed. The CPR method requires careful handling of the Hessian matrix and its eigenvalues, which can be computationally expensive and challenging in large systems.

6.1.1.2 Transition path search algorithms

The second category includes algorithms that aim to identify a transition path. Depending on stochastic or deterministic nature, these algorithms can generate multiple potential pathways connecting states or converge to the MEP.

Elastic Band Method (EB) The Elastic Band (EB) method, also known as the Plain Elastic-Band method, is a computational technique part of the deterministic methods,

and is used to identify MEP between two given states in a PES [49]. This method constructs a path by generating a sequence of images, or replicas, that link the initial and final states. These images are typically maintained at equidistant intervals along the path with a string force to ensure a smooth transition.

The fundamental objective of the EB method is to minimize a total energy of the path E, which is described by the following equation:

$$E(x_1, \ldots, x_{N-1}) = \sum_{i=1}^{N-1} U(x_i) + \frac{k\Delta\alpha}{2} \sum_{i=1}^{N} \frac{|x_i - x_{i-1}|^2}{\Delta\alpha^2} \tag{6.1}$$

$U(x_i)$ is the potential energy of image i and N represent the number of images. The last term define an elastic energy due to the virtual springs between consecutive images with k, being the spring constant and $\Delta\alpha$ representing the segment length between images.

The dynamics of image adjustments are governed by:

$$\dot{x}_i = -\frac{\partial E}{\partial x_i} = -\nabla U(x_i) + k\frac{x_{i+1} + x_{i-1} - 2x_i}{\Delta\alpha}, \quad i = 1, \ldots, N-1, \tag{6.2}$$

This equation describes how each image moves in response to the forces exerted by the potential energy and the springs. While this method is straightforward and intuitive, one common challenge it faces is the "corner-cutting" phenomenon. This occurs when the equidistant constraint on the images leads to the neglect of regions with sharp variations in the potential landscape, potentially omitting important transition states or details of the pathway.

Nudged Elastic Band Method (NEB) The NEB method [85, 12] follows a deterministic scheme as the EB method. It improves upon the traditional EB approach by selectively applying forces to guide the system along the minimum energy path. In the NEB method, only the normal component of the potential force and the tangential component of the spring force are considered, which helps in reducing unphysical artifacts like corner-cutting observed in the EB method. The dynamics of image adjustments are governed by:

$$\dot{x}_i = -[\nabla U(x_i)]^{\perp} + (F_i \cdot \hat{t}_i)\hat{t}_i, \quad i = 1, \ldots, N-1 \tag{6.3}$$

where $F_i = k(x_{i+1} + x_{i-1} - 2x_i)/\Delta\alpha$ and $\hat{t}_i$ denotes the tangent vector along the elastic band at x_i. This equation describes how each image moves in response to the forces exerted by the potential energy and the springs. These equations highlight the decomposition of forces into components parallel and perpendicular to the path, ensuring that the path smoothly transitions through the energy landscape without artificial distortions.

Despite its advantages, setting the appropriate spring constant (k) in NEB is often challenging. An overly high k-value results in a rigid path that necessitates smaller steps in the numerical solution of the ordinary differential equations (ODEs), while a too low k-value makes the chain overly flexible, potentially deviating significantly from critical saddle points, thus compromising the path's accuracy.

NEB variants To address these issues, variants like the Climbing Image-NEB (CI-NEB) have been developed. CI-NEB enhances the identification of saddle points by eliminating the spring force at the highest energy image and inverting the potential force to push the image towards the peak of the energy barrier [85]. This is done by selecting the image with the highest potential energy as the NEB calculation progresses.

Another variant, Energy Weighted-CI-NEB (EW-CI-NEB), modifies the spring constant based on the potential energy of each image to improve resolution in areas of high energy variation, as described by:

$$k_i = \begin{cases} (1-\alpha_i)k_u + \alpha_i k_l, & \text{if } E_i > E_{\text{ref}} \\ k_l, & \text{otherwise} \end{cases} \tag{6.4}$$

$$\alpha_i = \frac{E_{\max} - E_i}{E_{\max} - E_{\text{ref}}} \tag{6.5}$$

where k_u and k_l are the upper- and lower-bound value of the spring constant. E_i is the higher energy image of the pair of images connected by line segment i, $E_{\max}$ is the current estimation of the maximum potential energy along the path, and E_{ref} is a reference energy choosen to be equal to the energy of either the starting or goal energy minimum. This adaptive spring constant allows for increased resolution where needed, and reduced resolution in more stable regions of the path.

Finally, NEB-TS is a technique used in subsequent iterations to refine the search for the transition state (TS) after an initial NEB analysis. This method employs an eigenvalue-following technique known as the partitioned rational function optimization (P-RFO) to locate the TS by tracking the direction indicated by a selected eigenvalue of the Hessian matrix, ensuring that this eigenvalue is negative while all others remain positive, leading to a precise location of the first-order saddle point [14].

Transition Path Sampling Method (TPS) Transition Path Sampling (TPS) method [24, 42] is a stochastic computational method employed to explore transition pathways between two distinct states on a potential energy surface. This technique is particularly advantageous for systems where transitions are rare and involve crossing high energy barriers, such as in complex chemical reactions and biomolecular conformational changes.

TPS operates fundamentally by not just sampling states, but by sampling entire pathways that the system takes from an initial state A to a final state B. This method is initiated with an existing pathway, typically generated by a standard dynamical simulation, which provides a basic trajectory connecting the two states.

To refine and explore new pathways, TPS employs three primary Monte Carlo moves: shooting, shifting, and branching. Each move is designed to explore the space of possible transition paths thoroughly:

- Shooting is one of the key techniques in TPS. This move involves randomly selecting a point along the current path and slightly perturbing either the positions or the momenta at this point. Following this perturbation, the system's dynamics are

integrated both forward to state B and backward to state A, thereby generating a new candidate path. This new path is then subjected to a Metropolis acceptance criterion, which assesses its feasibility based on how well it represents the underlying dynamics of the system. This process ensures that only physically plausible paths are retained in the path ensemble.

- Shifting is another strategic move used in TPS. Instead of altering the path locally, as in shooting, shifting adjusts the entire path along the time axis. This temporal adjustment can be either forward or backward, allowing the exploration of different segments of the timeline where the transition might occur under varying conditions. This technique helps in discovering transition paths that are more probable or efficient but might have been overlooked in the initial path sampling.

- Branching, the third technique, expands the exploration by starting new trajectories from various points along the existing path. This approach effectively allows the system to explore new areas of the PES that might not be accessible from the original path. By branching out in different directions, TPS can uncover diverse transition mechanisms, providing a broader understanding of the possible ways the system can evolve from state A to state B.

Despite the effectiveness of TPS in uncovering the intricate dynamics of complex systems, it is computationally intensive. The need for detailed simulations across potentially high-dimensional landscapes requires significant computational resources.

The Zero-Temperature String Method The Zero-Temperature String Method [164] is a deterministic technique employed to identify the MEP. This method is particularly useful when the transition involves high energy barriers.

The string method conceptualizes the reaction path as a 'string' stretched between two energy minima, representing the initial and final states. This string is discretized into a series of points known as 'images' that depict configurations of the system along the path. The method's core objective is to relax this string into the lowest energy path connecting the two minima.

- Initialization: Initially, the string is placed manually between the minima and discretized into several images. These images are distributed evenly along the initial guess of the path.

- Relaxation Process: Each image is then relaxed independently to minimize its local energy, typically by using gradient descent techniques or other local optimization methods. The relaxation is performed orthogonally to the string to ensure that the images move towards the MEP without drifting along the path.

- Reparametrization: To maintain an even distribution of images along the path, the string is reparametrized periodically throughout the relaxation process. This reparametrization adjusts the positions of the images to keep them evenly spaced, ensuring that each segment of the string equally contributes to the depiction of the pathway.

The process is iteratively repeated until the changes in the images positions between successive iterations are minimal, indicating that the string has converged to the MEP. This convergence suggests that the string now represents the most probable, energetically favorable pathway for the transition between the two states.
The Zero-Temperature String Method is highly efficient as it focuses solely on finding the MEP without requiring the simulation of dynamic trajectories, making it faster than methods based on full dynamical simulations. Additionally, it provides a precise representation of the transition pathway, crucial for understanding complex reaction mechanisms. However, the method's success heavily depends on the initial placement of the string and requires significant computational resources for the local optimization of each image, which can be computationally intensive for large or complex systems.

Activation Relaxation Technique (ART) The Activation Relaxation Technique (ART), as described in studies by Barkema and Mousseau [17, 111] is depicted in Fig. 6.1. This stochastic method is achieved through a two-step process: moving towards a saddle point (activation), and ending at a new local minimum (relaxation).

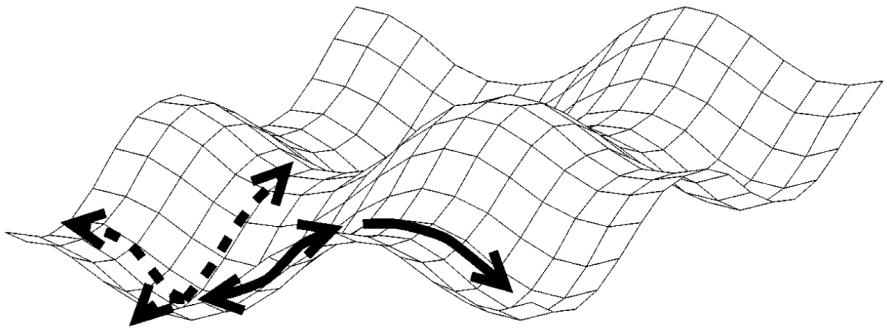


Figure 6.1: Illustrative scheme of ART from [17].

The activation phase is a critical component of the ART where the system is driven from a local energy minimum to an adjacent saddle point on the energy landscape. This phase is initiated by perturbing the system slightly to move it out of its current stable state. The goal is to induce a controlled ascent to a higher energy state, which represents a transition state between different configurations. In practical terms, the activation process begins by selecting an atom or a group of atoms and randomly displacing them from their equilibrium positions. The method employed to find this direction involves calculating the force acting on atoms and modifying it so that the system ascends energy gradients instead of descending them, which is the typical behavior in energy minimization techniques. This modified force vector propels the system upwards on the energy landscape, effectively inverting the usual dynamics that seek energy minima.

The relaxation phase follows the activation phase where, after reaching or approaching a saddle point, the system is allowed to relax towards directions where the energy decreases. This process helps the system to follow a downhill path that ideally leads to

a new energy minimum, thus confirming whether the explored path constitutes a valid transition between conformers.

ART's ability to explore high-energy regions of the energy landscape and locate transition states makes it a powerful tool for studying the kinetics and mechanisms of transformations and reactions in complex materials.

Transition-based Rapidly-exploring Random Trees (T-RRT) The stochastic technique, T-RRT algorithm, is a variant of the well-known RRT (discussed in Chapter 1) and permit the identification of transition paths [44, 43, 92, 93]. Unlike traditional RRT, T-RRT incorporates a transition test that guides the expansion of the tree toward energetically favorable regions. This transition test is based on the Metropolis criterion, which is crucial for accepting new nodes into the tree structure. The acceptance probability of a new node, represented by the following equation, reflects this criterion:

$$p_{ij} = \begin{cases} \exp\left(-\frac{E_j - E_i}{k_B T}\right), & \text{if } E_j > E_i \\ 1, & \text{otherwise} \end{cases} \tag{6.6}$$

Here, E_i and E_j denote the energies of the parent node and the newly proposed child node, respectively, k_B is the Boltzmann constant, and T symbolizes a temperature-like parameter used to control the acceptance rate of new nodes. This parameter is crucial as it allows the algorithm to occasionally accept higher-energy states, facilitating the exploration of energy barriers and avoiding traps in local minima. Notably, the temperature T is not an actual physical temperature but a tunable parameter within the algorithm. In the context of this implementation, adjustments to T occur only when E_j exceeds E_i. The adjustment is governed by:

$$T_{new} = \begin{cases} T \cdot 2^{-\frac{E_j - E_i}{\text{energyRange}}} & \text{if } p_{ij} > 0.5 \\ T \cdot 2^{T_{rate}} & \text{otherwise} \end{cases} \tag{6.7}$$

where T_{rate} is a parameter influencing the rate at which the temperature increases. $energyRange$ is equal to the maximum value between one and the difference between the threshold energy (an user parameter) and the lowest value found during the exploration. This dynamic adjustment of T enhances the algorithm's ability to adaptively explore the PES.

6.1.2 Trajectory comparison and clustering

Among the methods used to identify the possible paths, some techniques are stochastic, resulting in several pathways connecting a conformation to another. This necessitates the use of a clustering method to group similar trajectories and identify the classes of possible paths. Clustering is a fundamental technique in data analysis that groups similar data points together based on a defined metric or similarity measure. In the context of trajectory analysis, clustering methods are used to group similar trajectories together, allowing for the identification of common patterns. Most clustering methods

require a distance metric, or at least a similarity measure. Distances and clustering methods presented in this Chapter are detailed from the article "Review and Perspective for Distance-Based Clustering of Vehicle Trajectories" by Besse et al. [20].

6.1.2.1 Distance metrics or similarity measures

A trajectory can be define as a set of points in the configuration space, and the notion of time is not considered in the trajectories defined here. Indexes defined the order of the points in the trajectory, which can be noted as:

$$T^i = \{s_1^i, s_2^i, \ldots, s_n^i\} \tag{6.8}$$

where s_k^i is a k-th line segment between configurations p_{k-1}^i and p_k^i of the trajectory i. The distance between two trajectories T^i and T^j is noted as $D(T^i, T^j)$. Note that distances presented bellow do not respect all the properties of a metric: symmetry, triangle inequality or identity of indiscernibles.

Dynamic Time Warping (DTW) The Dynamic Time Warping (DTW) distance [19] is a widely used metric for comparing time series data, including trajectories and is part of warping based distances. This method aligns two sequences by stretching or compressing them in indexes to find the optimal match. A method to solve this problem is to define a $n_i \times n_j$ grid G. Each cell of the grid $g_{k,l}$ is defined by the pair (p_k^i, p_l^j). Then a warping path $W = w_1, w_2, \ldots, w_{|W|}$ crossing G is defined as:

$$\begin{aligned} &w_1 = g_{1,1}, \\ &w_{|W|} = g_{n^i,n^j}, \\ &\text{if } w_k = g_{k_i,k_j}, \text{then } w_{k+1} = (g_{k_{i+1},k_j} \text{ or } g_{k_i,k_{j+1}} \text{ or } g_{k_{i+1},k_{j+1}}) \end{aligned} \tag{6.9}$$

A warping distance is computed by minimizing a cost function between each pair of points defining the warping path. The DTW distance between two trajectories T^i and T^j is defined as:

$$D(T^i, T^j) = min_W \left[\sum_{k=1}^{|W|} \delta(w_k) \right] \tag{6.10}$$

where W is the warping path and $\delta(w_k)$ is the cost function which can be the Euclidean distance between two points p_k^i and p_k^j.

DTW is particularly useful for comparing trajectories with different lengths as it can account for variations in the index dimension. However, the computational complexity of DTW can be high, especially for large datasets, which may limit its applicability in certain scenarios.

Haussdorf distance The Haussdorf distance is a metric used to compare two sets of points in a metric space. It is defined as the maximum distance between a point in one set and its closest point in the other set (see Fig. 6.2) and is defined as:

$$Haus(X,Y) = \max\left\{\sup_{x\in X}\inf_{y\in Y} d(x,y), \sup_{y\in Y}\inf_{x\in X} d(x,y)\right\} \tag{6.11}$$

where X and Y are two spaces and $d(x,y)$ is the distance between two points x and y. The distance could have multiple definitions depending on the context. In the context

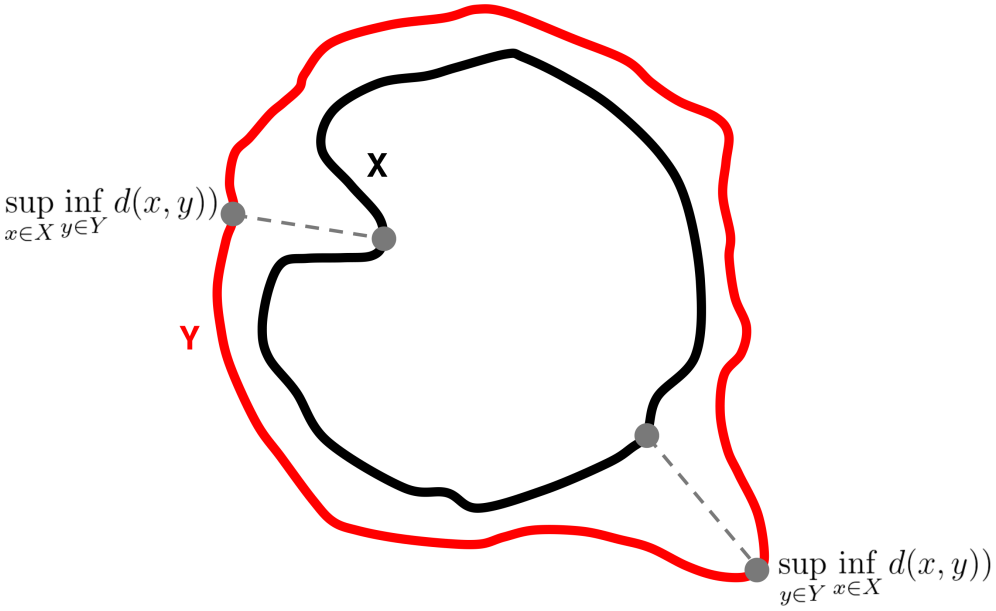


Figure 6.2: Haussdorf distance between two spaces. The distance between the two spaces is the maximum distance between a point in one space and its closest point in the other space.

of trajectories and considering the monotonicity of segments defining the trajectory, the distance between a point p_i^1 of the trajectory T^1 and a segment $s_{i_2}^2$ of the trajectory T^2 (see Fig. 6.3) is defined as:

$$D_{ps}(p_{i_1}^1, s_{i_2}^2) = \begin{cases} \|p_{i_1}^1 p_{i_1}^{1,\text{proj}}\|_2 & \text{if } p_{i_1}^{1,\text{proj}} \in s_{i_2}^2, \\ \min(\|p_{i_1}^1 p_{i_2}^2\|_2, \|p_{i_1}^1 p_{i_2+1}^2\|_2) & \text{otherwise.} \end{cases} \tag{6.12}$$

where $p_{i_1}^{1,\text{proj}}$ is the projection of $p_{i_1}^1$ on the segment $s_{i_2}^2$ and $p_{i_2}^2$ and $p_{i_2+1}^2$ are the two points defining the segment $s_{i_2}^2$. The Haussdorf distance between two trajectories T^1 and T^2 is defined as:

$$\begin{aligned} D_{\text{Hausdorff}}(T^1,T^2) &= \max\left\{Haus(T^1,T^2), Haus(T^2,T^1)\right\} \\ &= \max\left\{\begin{array}{c} \max\limits_{\substack{i_1\in[1..n_1]\\ j_2\in[1..n_2-1]}} D_{ps}(p_{i_1}^1, s_{j_2}^2), \\ \max\limits_{\substack{j_1\in[1..n_1-1]\\ i_2\in[1..n_2]}} D_{ps}(p_{i_2}^2, s_{j_1}^1) \end{array}\right\} \end{aligned} \tag{6.13}$$

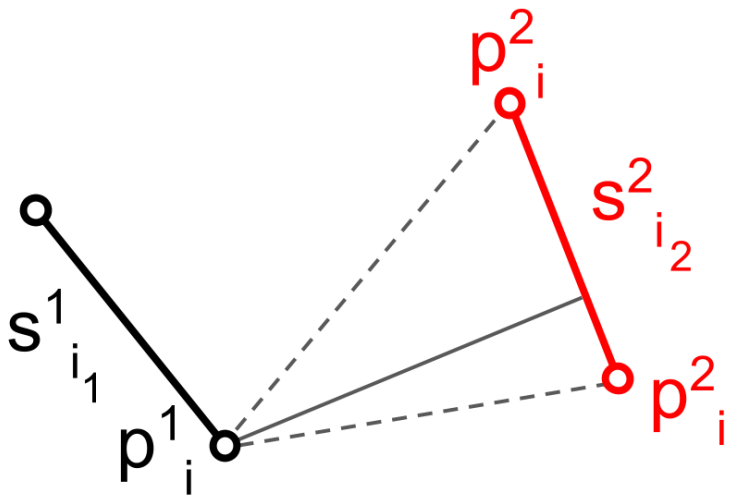


Figure 6.3: Distance between a point p_i^1 and a segment $s_{i_2}^2$.

where n_1 and n_2 are the number of points in the trajectories T^1 and T^2 respectively. The Haussdorf distance is particularly useful for comparing trajectories that may have different lengths or shapes, as it captures the maximum separation between the two sets of points. The main trouble enlighted by [20] is that the Haussdorf distance does not take into account the global shape of the trajectories, and is sensitive to the noise in the trajectories.

Frechet distance The Frechet distance [66] is a metric used to compare two curves in a metric space. It can be defined informally as "the minimum length of a leash that allows a dog and its owner to traverse their respective paths simultaneously, with the dog on one curve and the owner on the other" (see Fig. 6.4). Using the monotonous

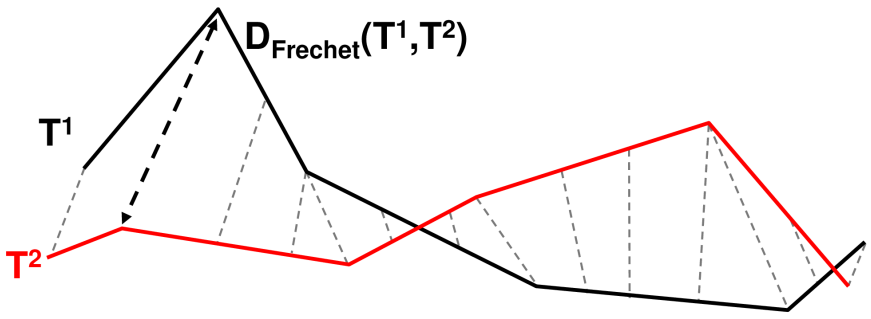


Figure 6.4: Frechet distance between two trajectories T^1 and T^2.

property of a segment, the Frechet distance between two segments is defined as:

$$\begin{aligned} D_{\text{Frechet}}(s^1_{i_1}, s^2_{i_2}) &= \max \left\{ \begin{array}{l} D_{ps}(p^1_{i_1}, s^2_{i_2}), D_{ps}(p^1_{i_1+1}, s^2_{i_2}), \\ D_{ps}(p^2_{i_2}, s^1_{i_1}), D_{ps}(p^2_{i_2+1}, s^1_{i_1}) \end{array} \right\} \\ &= \varepsilon_{i_1,i_2}. \end{aligned} \tag{6.14}$$

Then the Frechet distance between two trajectories T^1 and T^2 is computed by looking at every pairs of segments between T^1 and T^2, and finding the minimum value of ε.

One Way Distance (OWD) The One Way Distance (OWD) [110] between a trajectory T^i and a trajectory T^j is defined as the integral of the distance from points of the piece wise linear representation of T^i_{pl} to T^j_{pl} divided by the length of the trajectory T^i_{pl}:

$$D_{\text{OWD}}(T^i, T^j) = \frac{1}{n^i_{pl}} \int_{p^i \in T^i_{pl}} D_{\text{point}}(p^i, T^j) dp^i \tag{6.15}$$

where $D_{\text{point}}(p^i, T^j)$ is the distance between a point p^i and the trajectory T^j so that:

$$D_{\text{point}}(p, T) = \min_{q \in T_{pl}} \|pq\|_2. \tag{6.16}$$

The OWD is not symmetric, but can be symmetrized by taking the mean of the OWD between T^i and T^j and the OWD between T^j and T^i:

$$D_{\text{SOWD}}(T^i, T^j) = \frac{1}{2} \left(D_{\text{OWD}}(T^i, T^j) + D_{\text{OWD}}(T^j, T^i) \right). \tag{6.17}$$

This distance still does not satisfy the triangle inequality, and is time consuming. The Symmetrized OWD is particularly useful for comparing trajectories with different lengths and shapes.

Symmetrized Segment-Path Distance (SSPD) The Symmetrized Segment-Path Distance (SSPD) [20] is a similarity measure that consider the entire shape of the trajectories. The definition of the distance between a point and a segment is the same as the equation 6.12. Then the distance between a point $p^1_{i_1}$ of the trajectory T^1 and the trajectory T^2 is defined as:

$$D_{pT}(p^1_{i_1}, T^2) = \min_{i_2 \in [0,\ldots,n_2-1]} D_{ps}(p^1_{i_1}, s^2_{i_2}) \tag{6.18}$$

where n_2 is the number of points in the trajectory T^2. With this definition, the Segment-path distance (see Fig. 6.5) from a trajectory T^1 to a trajectory T^2 can be defined as the mean of all distances from points of the trajectory T^1 to the trajectory T^2:

$$D_{\text{SPD}}(T^1, T^2) = \frac{1}{n_1} \sum_{i_1=1}^{n_1} D_{pT}(p^1_{i_1}, T^2) \tag{6.19}$$

where n_1 is the number of points in the trajectory T^1. The SPD is not symmetric, but can be symmetrized by taking the mean of the SPD between T^1 and T^2 and the SPD between T^2 and T^1:

$$D_{\text{SSPD}}(T^1, T^2) = \frac{1}{2}\left(D_{\text{SPD}}(T^1, T^2) + D_{\text{SPD}}(T^2, T^1)\right). \tag{6.20}$$

In the article of Besse et al. [20], the SSPD is presented as the most efficient distances to compare trajectories, based on criteria such as comparaison of the entire shape of trajectories, less sensibility of the noise or acceptable computational time.

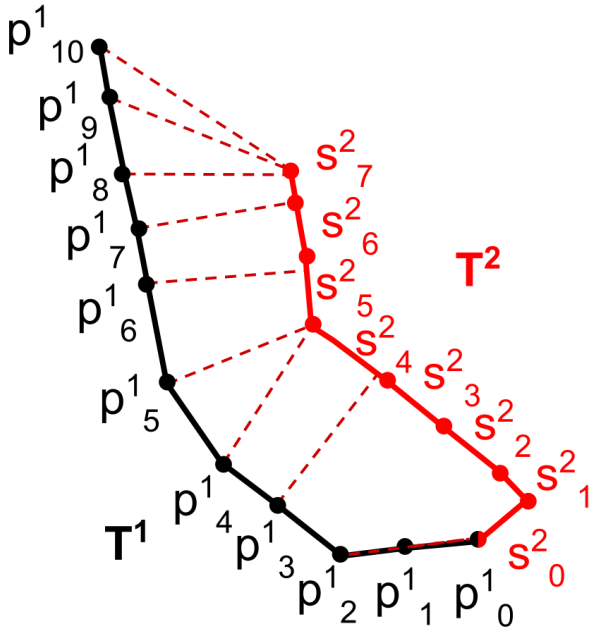
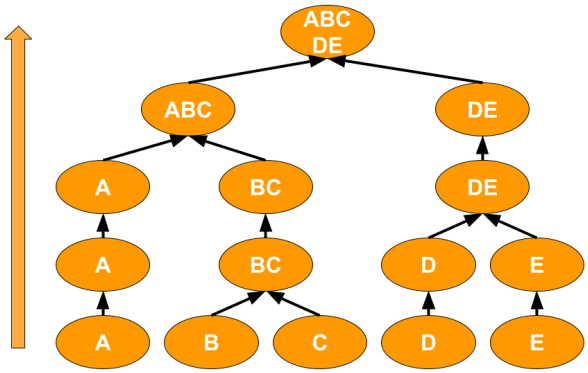


Figure 6.5: Segment-Path Distance (SPD) between two trajectories T^1 and T^2.

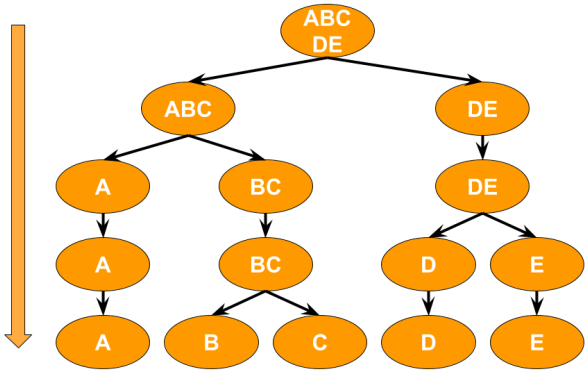
6.1.2.2 Clustering method

The clustering methods [3] has to be chosen according to the distance used to compare the trajectories. Some of the presented distances are not metric, so the clustering methods has to be compatible with this kind of distance. For example, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [32] is based on the nearest neighbor algorithm and necessitate a metric. For the clustering of trajectories using the SSPD as a similarity measure, the Hierarchical Clustering analysis (HCA) is a suitable method. HCA is a method designed to hierarchize elements based on their distance. Two strategies exist to construct the hierarchy of clusters: agglomerative and divisive.

In the agglomerative strategy (Fig. 6.6a), each element starts in its own cluster and is successively merged with the closest cluster. In the divisive strategy (Fig. 6.6b), all elements start in the same cluster and are progressively separated.



(a) HCA agglomerative strategy.



(b) HCA divisive strategy.

Figure 6.6: Illustration of Hierarchical Clustering Analysis.

The distance between trajectories and clusters should be distinguished. For example, in the Figure 6.6a, a distance defined by SSPD for example could be defined between elements B and C. But a method has to be established to have the distance between the cluster formed by A and the cluster formed by B and C. Several algorithms were designed to perform this task, including single linkage, complete linkage, average linkage, centroid linkage and Ward's method [119]. These algorithms differ in how they calculate the distance between clusters and how they merge clusters during the clustering process. For example, the single linkage algorithm calculates the distance between two clusters as the minimum distance between any two elements in the two clusters, while the complete linkage algorithm calculates the distance as the maximum distance between any two elements in the two clusters. The choice of the clustering algorithm depends on the nature of the data and the desired outcome of the clustering process.

6.2 Methods

The identification of low-energy conformations, between which transition paths will be searched, is achieved through the IGLOO/DFTB coupling, as detailed in Chapter 2. From these minima, a method is proposed to generate transition paths between them as presented in Fig. 6.7.

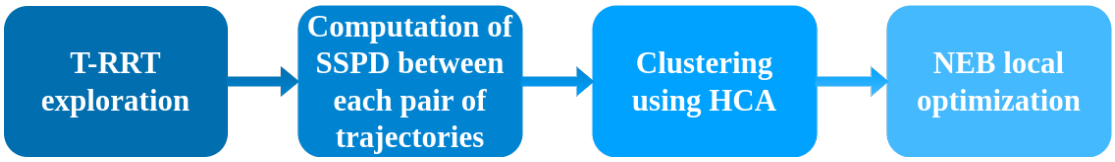


Figure 6.7: Workflow of the method.

T-RRT exploration: Firstly, the T-RRT algorithm is employed to identify potential transition pathways between two low-energy conformations. As explained in the transition path search algorithms section, the T-RRT will generate several trajectories connecting the minima. Parameters such as the temperature (which will be specified in the results section) are set to obtain paths following as much as possible the low energy regions of the PES. The exploration algorithm used is an Anytime version of the T-RRT who continue to optimize the graph until a stopping criterion is reached. The time of exploration is the criterion stopping this step. The result of this step is not a tree, which give a direct solution between the initial and final states but a graph with several paths connecting the two states. A path connecting differents states has to be extracted using Dijkstra's algorithm [22]. This algorithm focuses on finding a path in the graph that minimizes a cost function. The trajectories obtained are then differentiated using a similarity measurement method.

Similarity measure with SSPD: Subsequently, a similarity measure is required to differentiate the trajectories. The SSPD is the chosen similarity measure as it strikes an optimal balance between the global shape of the trajectories and local differences, while also offering a better performance in terms of computation time compared to other presented distances. It should be noted that conformations are defined by their dihedral angle values. As detailed in Equation 6.12, a distance between points of the trajectories must be established (a point here is a conformation). In order to achieve this, the root mean square deviation (RMSD) is employed to compute the distance between two points, with the following definition:

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_i^1 - \theta_i^2\right)^2} \tag{6.21}$$

where θ_i^1 and θ_i^2 are the dihedral angles values of the conformations 1 and 2 respectively and N is the total number of dihedral angles in the conformations.

As detailed in equation 6.12, a projection from a point (p_i^1) of the trajectory 1 to a segment (s_i^2) of the trajectory 2 is required. For this purpose, an interpolation between the two points defining the s_i^2 is performed to obtain the projection.

Clustering using HCA: A HCA is performed on the generated trajectories to obtain the class of trajectories from the exploration. The HCA can be performed using a variety of algorithms, as previously described. The available algorithms in the Python scikit-learn library used, include single linkage, complete linkage, average linkage, centroid linkage, and Ward's method [119]. In order to determine the optimal number of clusters and to analyze the quality of clustering for each algorithm, it is necessary to define both intra and extra-cluster variances. However, these variances are not readily available for trajectories defined in this manner. As developed in the article [20], a Between-Like and the Within-Like criteria can be defined as:

$$\begin{aligned} BC &= \sum_{k=1}^{K} D(T^{ex}, T^{ex}_{\mathcal{C}_k}) \\ WC &= \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{\mathcal{T}_i \in \mathcal{C}_k} D(T^{ex}_{\mathcal{C}_k}, T^i) \end{aligned} \tag{6.22}$$

where $|\mathcal{C}_k|$ is the cardinality of the cluster $\mathcal{C}_k$ (the number of elements of the cluster), BC defined the Between-Like criterion and WC the Within-Like criterion. The BC criterion is defined to characterize the distribution between each cluster; thus, the objective in this study is to have the most distinct clusters. The WC criterion is defined to characterize the distribution within a cluster. The objective of this study is to define clusters having the most similar elements inside. T^{ex} represent the examplary trajectory, which is defined as the closest trajectory of a set of trajectories. As an example, for the cluster $\mathcal{C}_k$ in the set of trajectories $\mathcal{T}$, an examplary trajectory could be defined as:

$$T^{ex}_{\mathcal{C}_k} = \min_{\substack{T^i_{\mathcal{C}_k} \\ i \in [0 \ldots n^{\mathcal{C}_k}]}} \left\{ \sum_{\substack{j=1 \\ j \neq i}}^{n^{\mathcal{C}_k}} D(T^i, T^j) \right\} \tag{6.23}$$

The application of these criteria allows for the determination of the number of clusters and the selection of the most appropriate linkage method, by comparing their performance on these variances. Subsequently, the class of clusters is identified.

NEB local optimization: The exemplary trajectory of each cluster is used to initialize a NEB calculation. This method is implemented in the deMonNano code [131]. The NEB calculation is used to refine the transition path between the two low energy conformations. Using the exemplary trajectory as input has the advantage of providing a good first guess for the NEB calculation, which may facilitate convergence. Note that this part is still under development and will not be discussed in the results section.

6.3 Results

As a first demonstration of the proposed methodology, the approach outlined in the preceeding section was employed to analyze the alanine dipeptide, which was previously discussed in the Chapter 2. A T-RRT exploration on two low energy conformations was conducted with an initial temperature of 20 K and a temperature rate of 0.1. These parameters were empirically defined in order to ensure optimal growth of the tree in a reasonable time frame and to converge to paths with the lowest energy possible. The SSPD distance was calculated for each of the 100 paths generated by the T-RRT algorithm. This number of paths was selected to have an acceptable computational time and to ensure that every possible low-energy path has been represented. Based on these trajectories, a HCA was performed.
The BC and WC criteria were computed and yielded the following results in Fig. 6.8. Usually, the sum of the intra and inter-cluster variance is constant. But the sum of the described criteria BC and WC is not constant, which is why each of these criteria is presented.

As can be observed, the single linkage algorithm exhibits a faster convergence than

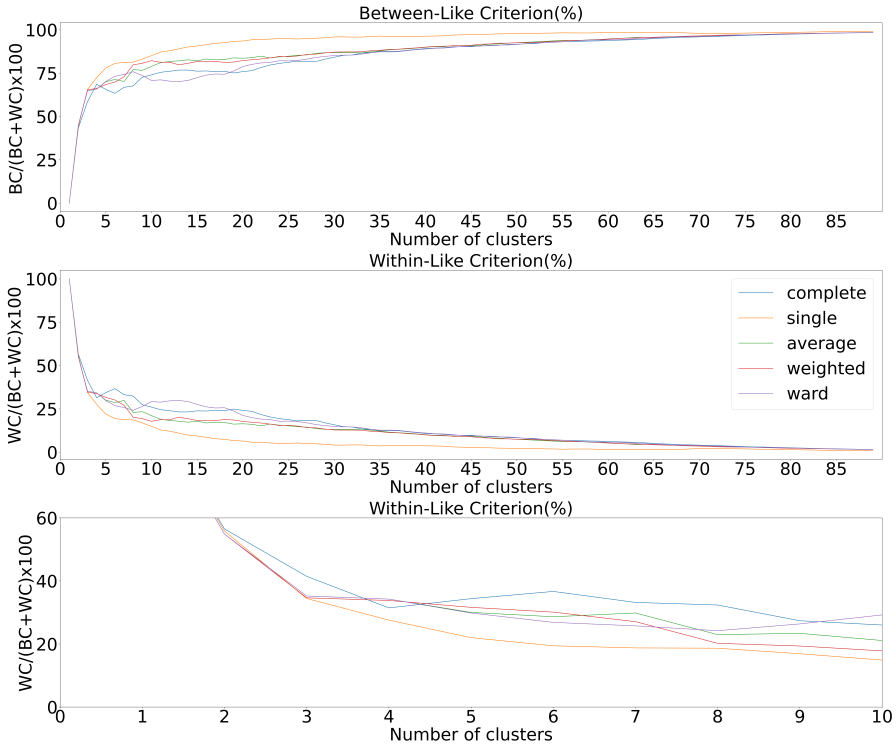


Figure 6.8: Quality evaluation of the clustering method. The legend detailed the algorithm used to perform the HCA.

other algorithms to a low Within-Like criterion, making it the better algorithm for performing the HCA. This difference is not particularly pronounced, and results may vary with larger chemical systems. A plateau is observed around 4 clusters, as can be seen

in the bottom of Figure 6.8. After this number of clusters, the Within-Like criterion does not exhibit a significant decrease. This analysis indicates that the clustering on the T-RRT paths for the alanine dipeptide will be based on four clusters using the single algorithm.

The HCA using a single linkage algorithm is performed and clustering results are presented in Fig. 6.9. The method was successful in grouping the paths into four distinct

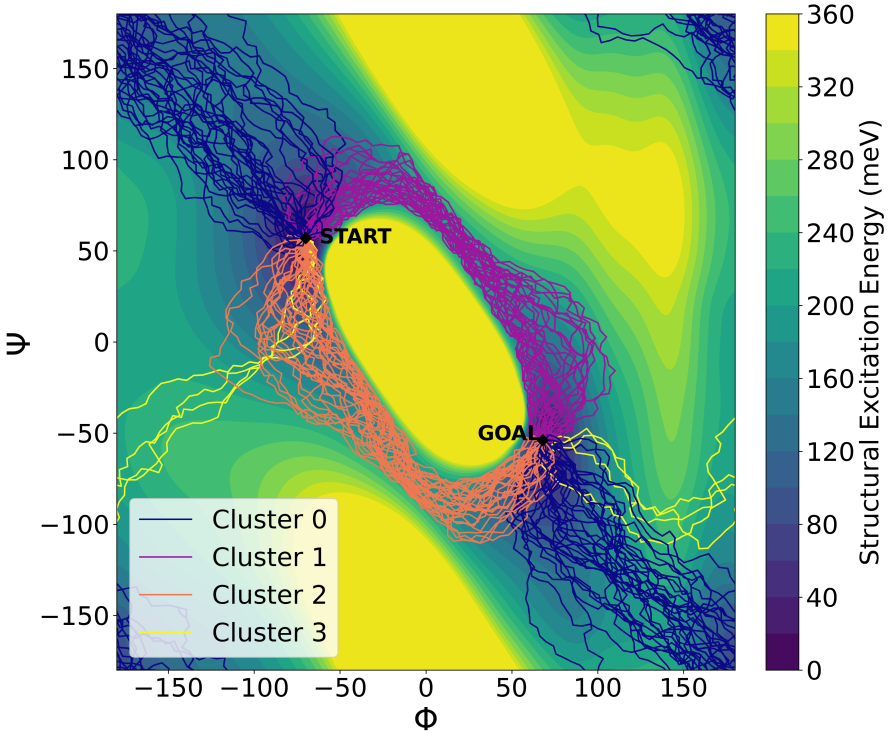


Figure 6.9: Transition paths of alanine dipeptide using a single linkage.

clusters, each representing a different transition pathway between the two low-energy conformations. Every cluster passes through a low-energy region to connect each minimum, with the exception of cluster 3, which appears to only traverse a relatively high-energy region. This path could be readily avoided by reducing the temperature of the T-RRT algorithm if necessary. This class is of interest as a test case for the clustering procedure, which is designed to split these types of paths into different clusters.

For each cluster, the examplary trajectory (define in the method section) is presented in the Figure 6.10. As the number of points/conformations are not the same between each trajectory, an interpolation is made for this figure. Although all the paths have different shapes, each seems to reach a maximum energy level with a close value. As discussed above, the trajectory for the cluster 3 passes only through a high-energy region, which is not ideal. The trajectory from the cluster 0 passes through two energy barriers and and go through a local minimum, which is identified in most studies of the alanine dipeptide PES [82]. Trajectories from cluster 1 and 2 exhibit a similar shape, with a local minimum close to the start or close to the end for cluster 1 and 2, respectively.

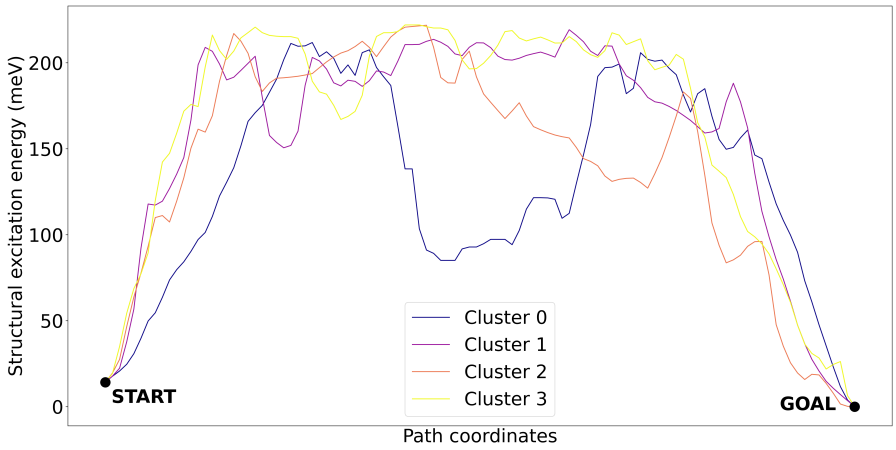


Figure 6.10: Exemplary trajectory of each cluster.

6.4 Conclusion

This work presents methods for exploring the PES and identifying transition paths between low-energy conformations, as well as for identifying saddle points on a PES. Some of the algorithms presented for identifying transition paths are stochastic, and can produce a large number of potential pathways. To analyze the results from multiple runs of these algorithms, a similarity measure is required to compare the paths, and several distances are presented. Then, a clustering method must be used to define the class of trajectories. The Hierarchical Clustering Analysis was presented as a suitable method for defining classes of trajectories. A methodology was proposed combining T-RRT exploration, SSPD as a similarity measure and HCA as clustering method to identify low energy paths between two conformations on the alanine dipeptide PES. Following the successful testing of this methodology on a test case, further research will be conducted to test more complex systems, specifically the phthalate molecule presented in Chapter 3. Another objective is to refine the representative paths of each trajectory class through the use of NEB calculations.

Chapter 7

General Conclusions and Perspectives

Contents

7.1 General Conclusions

Several topics have been discussed in this thesis. All of them are connected by a main theme, the exploration of high-dimensional spaces for molecular study and modeling. In Chapters 2, 4, and 6, three distinct algorithms have been presented for exploring spaces. These are the conformational space, which represents the different geometries accessible for a molecule; the molecular space, which represents the diversity of molecules available in a family; and the transition path space, which represents the paths connecting different states. These algorithms serve distinct purposes, but they are linked to each other, as the output of one can be used as input for another. In Chapters 3, 5, and 6, these algorithms have been applied to diverse molecules, including a benchmark molecule (alanine dipeptide), environmental pollutants (phthalates), and structures that can be found in the ISM (substructure units of hydrogenated amorphous carbon polymer).

7.1.1 Global exploration of energy landscapes

The combination of IGLOO and DFTB allows for a non-redundant exploration of PES at a relatively high level of theory, representing a promising approach to investigate the conformational space of molecules. The method was shown to be able to explore

the PES of the alanine dipeptide, and of phthalate molecules. The well-known main minima were found for the alanine dipeptide, used here as a simple example to test the method. Then, the coupling was used to explore the conformational PES of three molecules representative of the phthalate family: BBP, DBP and DEHP. The selection of these molecules was driven by their notable impacts on human health. Our findings indicate distinct conformational landscape characteristics for these molecules, despite their structural similarities and comparable molecular sizes. The analysis of structural excitation energy spectra revealed distinct distributions, indicating that despite similarities in their chemical composition, these molecules exhibit notable differences. DBP demonstrated distinct peaks within its spectrum, suggesting a ordered set of low energy conformations. In contrast, DEHP displayed a continuous spectrum, indicating a close-energy state arrangement, while BBP exhibited intermediate characteristics, blending traits of both DBP and DEHP.

To rationalize these differences, we employed descriptors based on specific molecular distances and dihedral angles. The structural stability of DBP's lower-energy configurations appears to be predominantly influenced by oxygen-oxygen Coulomb interactions. For BBP, conformations were observed where the positively charged hydrogen atoms of the butyl side-chain are oriented towards the negatively charged aromatic carbon atoms, enhancing Coulomb interaction stability. DEHP's extended and branched side-chains introduce steric hindrance and dispersive interactions, leading to a competitive landscape among numerous isomers.

These interactions shape the geometric characteristics of the phthalate molecules, resulting in distinct peaks for DBP and BBP, or a broader spectral feature for DEHP in their respective O-O and C-O distances distribution plots. Furthermore, a significant correlation was observed between the two dihedral angles defining the side-chain orientations across all three molecules, underlining the complex interplay of forces shaping their energy landscapes.

7.1.2 Large-scale generation of atomistic models of aromatic hydrocarbons

A novel algorithm has been developed for the large-scale generation of atomistic models, focusing on hydrogenated amorphous carbon (a-C:H) structures. Using the SMILES formalism, the approach provides flexibility and efficiency in generating molecular structures that adhere to specific chemical constraints.

The algorithm ensures structural integrity (by removing 3D structures with incorrect connectivity) and optimizes geometry to minimize atomic collisions. The comprehensive database of a-C:H structures, characterized by their geometric and electronic descriptors, provides deep insight into the relationship between molecular geometry and electronic properties. The structures of the database exhibit a wide range of shapes and specific families. For instance, a family could be those with isolated rings and a higher HOMO-LUMO gap. These results illustrate the database potential for further scientific studies.

7.1.3 Transition paths sampling

A new strategy was proposed to generate and analyze multiple potential pathways connecting different minima on the PES. It relies on T-RRT for the transition path sampling. This particular approach uses the SSPD as a similarity measure to effectively manage and categorize the complex array of trajectories generated during the exploration. HCA is then applied to classify these trajectories, ensuring a systematic approach to understanding pathway similarities and differences.

This integrated methodology has been successfully applied to identify low-energy paths between conformational states of the alanine dipeptide on its PES, demonstrating its ability to accurately delineate transition paths. An automatic scheme has been developed to discriminate between different generated paths for this system. As observed, these trajectories have differents shapes, but a maximum energy state at a similar level. Moreover, these trajectories have yet to be further refined through the application of a NEB method.

7.2 Perspectives

Perspectives remain to enhance the exploration and develop new ideas. The perspectives for this thesis are both in terms of application and methodology.

7.2.1 Extending the global exploration: environmental effects and new chemical systems

The global exploration scheme based on the IGLOO/DFTB coupling has been applied to phthalate molecules in the gas phase. To obtain properties comparable to experimental data on environmental problems, a study has to be performed in a solvent such as water, and possibly with multiple structures. A QM-MM explicit solvent raises the question of how to organize it in the simulation box to have relevant properties. On the other hand, an implicit solvent on the DFTB potential raises the question of the quality of the added potential. Both options could be considered to address this problem. For the multistructure problem, an exploration scheme has to be defined to handle not only intermolecular but also intramolecular displacements.
Another perspective is the extension of the method to other chemical systems. It is important to note that this method can be applied to several molecules, although few have been tested that are particularly relevant. The coupling was used to molecules from the azine family, but the results were not conclusive, due to non-physics behavior with the DFTB potential of the carbon-nitrogen bond. Another possible application is on the Tamoxifene, which is a drug used to treat breast cancer. The tools developed in this thesis can allow for a fast selection of the most relevant structures and their electronic and structural properties can be used to build chemical databases of polluting molecules.

7.2.2 Exploring the a-C:H substructures and extending to other molecular compositions

After having obtained both geometric and electronic properties of a-C:H substructures units, infrared spectra is the next step, but the computational time to generate many spectra is long. This work is still under development. Furthermore, the variations between the spectra of each structure will be examined using previously defined descriptors to understand the impact of geometric or electronic changes on the IR spectra. In addition, the algorithm and the database need to be published so that the community can use them.
Future work could be done on two different parts: the generation algorithms and the properties obtained from the database. The generation algorithm could be extended to other molecules by adding more fragments to the algorithm and changing the functional group based constraint. This change would be easy to make, but required a testing phase to ensure that the connectivity of each atom is respected. This topic is interesting for other fields that require a database of structures, such as protein or ligand databases. The database may be utilized as a training set for machine learning algorithms with the objective of predicting the properties of a-C:H substructures. For example, the IR spectrum could be employed as a training set for a model capable of predicting the IR spectrum of novel a-C:H structures.

7.2.3 Developments of the transition paths method

Based on the methodology already developed for the generation of transition paths, two options are now possible: (i) the application to more complex systems such as the minima of the phthalate family, and (ii) the use of the generated paths as input to an NEB calculation as a better input than the "straight line" between the minima. This last point could be interesting as the stochastic nature of the T-RRT could lead to different paths that NEB wouldn't be able to find if it is simply initialized from a linear interpolation between the initial and final states (as usually done).

List of Figures

Bibliography

[1] S. Abb et al. “Carbohydrate Self-Assembly at Surfaces: STM Imaging of Sucrose Conformation and Ordering on Cu(100)”. In: *Angewandte Chemie* 58 (2019), pp. 8336–8340 (Cited on pages 18, 25, 161).

[2] S. Abb et al. “Polymorphism in carbohydrate self-assembly at surfaces: STM imaging and theoretical modelling of trehalose on Cu(100)”. In: *RSC Advances* 9 (2019), pp. 35813–35819 (Cited on pages 18, 25, 161).

[3] C. Aggarwal and C. Reddy. *Data Clustering: Algorithms and Applications*. Aug. 2013 (Cited on page 111).

[4] A. Aktürk and A. Sebetci. “BH-DFTB/DFT calculations for iron clusters”. In: *AIP Advances* 6.5 (May 2016), p. 055103 (Cited on page 23).

[5] L. J. Allamandola, A. G. G. M. Tielens, and J. R. Barker. “Polycyclic aromatic hydrocarbons and the unidentified infrared emission bands: auto exhaust along the milky way.” In: *The Astrophysical Journal* 290 (Mar. 1, 1985), pp. L25–L28. (Visited on 05/21/2024) (Cited on page 78).

[6] M. P Allen and D. J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017 (Cited on pages 6, 149).

[7] C. Almeras et al. *Projet PERSAN Les Phtalates*. Tech. rep. 2010 (Cited on page 35).

[8] M. S. Apaydin et al. “Capturing molecular energy landscapes with probabilistic conformational roadmaps”. In: *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*. Vol. 1. 2001, 932–939 vol.1 (Cited on pages 17, 160).

[9] M. S. Apaydin et al. “Stochastic Conformational Roadmaps for Computing Ensemble Properties of Molecular Motion”. In: *Algorithmic Foundations of Robotics V* 7 (2004), pp. 131–147. (Visited on 04/11/2024) (Cited on pages 17, 160).

[10] M. S. Apaydin et al. “Stochastic roadmap simulation for the study of ligand-protein interactions”. In: *Bioinformatics* 1.1 (2002), pp. 1–8 (Cited on pages 17, 160).

[11] M. S. Apaydin et al. “Stochastic roadmap simulation: an efficient representation and algorithm for analyzing molecular motion”. In: *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 10.3 (2003), pp. 257–281 (Cited on pages 17, 160).

[12] V. Ásgeirsson et al. “Nudged Elastic Band Method for Molecular Reactions Using Energy-Weighted Springs Combined with Eigenvector Following”. In: *Journal of Chemical Theory and Computation* 17.8 (2021), pp. 4929–4945. (Visited on 06/06/2023) (Cited on page 102).

[13] P. W. Atkins and R. S. Friedman. *Molecular quantum mechanics*. Oxford University Press, USA, 2011 (Cited on pages 5, 147).

[14] J. Baker. "An algorithm for the location of transition states". In: *Journal of Computational Chemistry* 7.4 (Aug. 1986), pp. 385–395. (Visited on 07/31/2023) (Cited on page 103).

[15] C. Bannwarth, S. Ehlert, and S. Grimme. "GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions". In: *Journal of Chemical Theory and Computation* 15.3 (Mar. 12, 2019), pp. 1652–1671. (Visited on 03/11/2024) (Cited on page 62).

[16] A. Barducci, M. Bonomi, and M. Parrinello. "Metadynamics". In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5 (2011), pp. 826–843 (Cited on pages 13, 23, 156).

[17] G. T. Barkema and N. Mousseau. "Event-Based Relaxation of Continuous Disordered Systems". In: *Physical Review Letters* 77.21 (Nov. 18, 1996), pp. 4358–4361. (Visited on 07/30/2023) (Cited on page 105).

[18] M. Bendikov et al. "Oligoacenes: Theoretical Prediction of Open-Shell Singlet Diradical Ground States". In: *Journal of the American Chemical Society* 126.24 (June 23, 2004), pp. 7416–7417. (Visited on 05/22/2024) (Cited on page 62).

[19] D. J. Berndt and J. Clifford. "Using dynamic time warping to find patterns in time series". In: *Proceedings of the 3rd international conference on knowledge discovery and data mining*. 1994, pp. 359–370 (Cited on page 107).

[20] Philippe C. Besse et al. "Review and Perspective for Distance-Based Clustering of Vehicle Trajectories". In: *IEEE Transactions on Intelligent Transportation Systems* 17.11 (2016), pp. 3306–3317 (Cited on pages 107, 109–111, 114).

[21] A. Blanco et al. "Hydrogenated amorphous carbon grains and the 2175 A interstellar hump". In: *Astrophysical Journal, Part 2-Letters* 382 (1991), pp. L97–L99 (Cited on page 79).

[22] I. Al-Bluwi, T. Siméon, and J. Cortés. "Motion planning algorithms for molecular simulations: A survey". In: *Computer Science Review* 6.4 (2012), pp. 125–143 (Cited on pages 16, 23, 113, 159).

[23] H. Bockhorn. *Soot formation in combustion: mechanisms and models*. Vol. 59. Springer Science & Business Media, 2013 (Cited on page 62).

[24] P. G. Bolhuis et al. "Transition path sampling: Throwing ropes over rough mountain passes, in the dark". In: *Annual Review of Physical Chemistry* 53.1 (2002), pp. 291–318 (Cited on page 103).

[25] T. C. Bond et al. "Bounding the role of black carbon in the climate system: A scientific assessment". In: *Journal of Geophysical Research: Atmospheres* 118.11 (2013), pp. 5380–5552. (Visited on 03/11/2024) (Cited on page 61).

[26] R. Bonneau and D. Baker. "Ab initio protein structure prediction: progress and prospects". In: *Annual Review of Biophysics and Biomolecular Structure* 30 (2001), pp. 173–189 (Cited on pages 16, 159).

[27] L. O. P. Borbón and L. O. Paz Borbón. "Theoretical Background and Methodology". In: *Computational Studies of Transition Metal Nanoalloys* (2011), pp. 15–31 (Cited on page 23).

[28] M. Born and R. Oppenheimer. "Zur Quantentheorie der Molekeln". In: *Annalen der Physik* 389.20 (1927), pp. 457–484 (Cited on pages 3, 146).

[29] "CaGe – a Virtual Environment for Studying Some Special Classes of Plane Graphs – an Update". In: *MATCH – Communications in Mathematical and in Computer Chemistry* (July 18, 2009) (Cited on page 62).

[30] F. Calvo. "All-exchanges parallel tempering". In: *The Journal of Chemical Physics* 123.12 (Sept. 28, 2005), p. 124106. (Visited on 04/11/2024) (Cited on pages 12, 155).

[31] F. Calvo et al. "Three-dimensional global optimization of Na n + sodium clusters in the range n <~ 40". In: *Physical Review B* 62.15 (Oct. 15, 2000), pp. 10394–10404. (Visited on 05/02/2024) (Cited on page 86).

[32] R. JGB Campello, D. Moulavi, and J. Sander. "Density-based clustering based on hierarchical density estimates". In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer. 2013, pp. 160–172 (Cited on page 111).

[33] T. Casalini. "Fundamentals and application of modeling in support of spinal cord injury repair strategies". In: *Spinal Cord Injury (SCI) Repair Strategies*. Elsevier, 2020, pp. 279–306 (Cited on pages 8, 150).

[34] X. Chen et al. "Recent progresses of global minimum searches of nanoclusters with a constrained Basin-Hopping algorithm in the TGMin program". In: *Computational and Theoretical Chemistry* 1107 (2017), pp. 57–65 (Cited on page 23).

[35] T. Chiang et al. "Predicting experimental quantities in protein folding kinetics using stochastic roadmap simulation". In: *Research in Computational Molecular Biology: 10th Annual International Conference, RECOMB 2006, Venice, Italy, April 2-5, 2006. Proceedings 10*. 2006, pp. 410–424 (Cited on pages 17, 160).

[36] T. Chiang et al. "Using Stochastic Roadmap Simulation to Predict Experimental Quantities in Protein Folding Kinetics: Folding Rates and Phi-Values". In: *Journal of Computational Biology* 14.5 (June 2007), pp. 578–593. (Visited on 04/11/2024) (Cited on pages 17, 160).

[37] T. Choi. "Simulation of the (H2O) 8 cluster with the SCC-DFTB electronic structure method". In: *Chemical Physics Letters* 543 (2012), pp. 45–49 (Cited on page 23).

[38] T. Choi et al. "Application of the SCC-DFTB method to hydroxide water clusters and aqueous hydroxide solutions". In: *The Journal of Physical Chemistry B* 117.17 (2013), pp. 5165–5179 (Cited on page 23).

[39] I. Colón et al. “Identification of phthalate esters in the serum of young Puerto Rican girls with premature breast development.” In: *Environmental health perspectives* 108.9 (2000), pp. 895–900 (Cited on page 36).

[40] J. Cortés et al. “A path planning approach for computing large-amplitude motions of flexible molecules”. In: *Bioinformatics* (2005) (Cited on page 25).

[41] E. Dartois et al. “Ultraviolet photoproduction of ISM dust - Laboratory characterisation and astrophysical relevance”. In: *Astronomy & Astrophysics* 432.3 (Mar. 1, 2005), pp. 895–908. (Visited on 10/26/2023) (Cited on pages 65, 67, 78–80, 82, 83).

[42] C. Dellago, P.G. Bolhuis, and P.L. Geissler. “Transition Path Sampling Methods”. In: *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*. Berlin, Heidelberg: Springer, 2006, pp. 349–391. (Visited on 06/07/2024) (Cited on page 103).

[43] D. Devaurs, T. Siméon, and J. Cortés. “A multi-tree extension of the transition-based RRT: Application to ordering-and-pathfinding problems in continuous cost spaces”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2014, pp. 2991–2996 (Cited on page 106).

[44] D. Devaurs et al. “Characterizing energy landscapes of peptides using a combination of stochastic algorithms”. In: *IEEE Transactions on Nanobioscience* 14.5 (2015), pp. 545–552 (Cited on pages 23, 106).

[45] B. T. Draine. *Physics of the Interstellar and Intergalactic Medium*. Jan. 1, 2011. (Visited on 05/21/2024) (Cited on page 78).

[46] WW Duley. “Infrared absorption due to hydrogenated amorphous carbon in the diffuse interstellar medium”. In: *The Astrophysical Journal* 430 (1994) (Cited on page 79).

[47] WW. Duley et al. “Integrated absorbances in the 3.4 μm CHn band in hydrogenated amorphous carbon”. In: *The Astrophysical Journal* 503.2 (1998), p. L183 (Cited on page 79).

[48] E. Dwek. “The Evolution of the Elemental Abundances in the Gas and Dust Phases of the Galaxy”. In: *The Astrophysical Journal* 501.2 (July 10, 1998), pp. 643–665. (Visited on 05/21/2024) (Cited on page 79).

[49] W. E and E. Vanden-Eijnden. “Transition-Path Theory and Path-Finding Algorithms for the Study of Rare Events”. In: *Annual Review of Physical Chemistry* 61.1 (2010), pp. 391–420. (Visited on 06/06/2023) (Cited on page 102).

[50] D. J. Earl and M. W. Deem. “Parallel tempering: Theory, applications, and new perspectives”. In: *Physical Chemistry Chemical Physics* 7.23 (Nov. 16, 2005), pp. 3910–3916. (Visited on 04/11/2024) (Cited on pages 12, 155).

[51] R. Eisenschitz and F. London. “Über das Verhältnis der van der Waalsschen Kräfte zu den homöopolaren Bindungskräften”. In: *Zeitschrift für Physik* 60.7 (July 1, 1930), pp. 491–527. (Visited on 05/02/2024) (Cited on page 89).

[52] M. Elstner. “The SCC-DFTB method and its application to biological systems”. In: *Theoretical Chemistry Accounts* 116.1 (July 31, 2006), pp. 316–325. (Visited on 09/19/2019) (Cited on page 27).

[53] M. Elstner and G. Seifert. “Density functional tight binding”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 372.2011 (2014), p. 20120483 (Cited on page 27).

[54] M. Elstner et al. “A Self-Consistent Charge Density-Functional Based Tight-Binding Scheme for Large Biomolecules”. In: *Physica Status Solidi (b)* 217.1 (Jan. 2000), pp. 357–376. (Visited on 09/19/2019) (Cited on page 27).

[55] M. Elstner et al. “Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties”. In: *Physical Review B* 58.11 (1998), p. 7260 (Cited on pages 6, 23, 26, 148).

[56] A. Estaña et al. “Hybrid parallelization of a multi-tree path search algorithm: Application to highly-flexible biomolecules”. In: *Parallel Computing* 77 (2018), pp. 84–100 (Cited on page 23).

[57] H. Eyring. “The Activated Complex and the Absolute Rate of Chemical Reactions.” In: *Chemical Reviews* 17.1 (Aug. 1, 1935), pp. 65–77. (Visited on 07/31/2023) (Cited on page 100).

[58] S. Fischer and M. Karplus. “Conjugate peak refinement: an algorithm for finding reaction paths and accurate transition states in systems with many degrees of freedom”. In: *Chemical Physics Letters* 194.3 (June 1992), pp. 252–261. (Visited on 06/07/2023) (Cited on page 101).

[59] N. Flagey et al. “Spitzer/IRAC and ISOCAM/CVF insights on the origin of the near to mid-IR Galactic diffuse emission”. In: *Astronomy & Astrophysics* 453.3 (July 1, 2006), pp. 969–978. (Visited on 03/11/2024) (Cited on page 78).

[60] V. Fock. “„Selfconsistent field“ mit Austausch für Natrium”. In: *Zeitschrift für Physik* 62.11 (Nov. 1, 1930), pp. 795–805. (Visited on 05/27/2024) (Cited on pages 5, 147).

[61] V. Fock. “Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems”. In: *Zeitschrift für Physik* 61.1 (Jan. 1, 1930), pp. 126–148. (Visited on 04/12/2024) (Cited on pages 5, 147).

[62] G. C. Fonger. “Hazardous substances data bank (HSDB) as a source of environmental fate information on chemicals”. In: *Toxicology* 103.2 (Nov. 30, 1995), pp. 137–145. (Visited on 03/11/2024) (Cited on page 61).

[63] L. Fortunato. “Solutions of the Bohr hamiltonian, a compendium”. In: *The European Physical Journal A* 26 (S1 Oct. 2005), pp. 1–30. (Visited on 05/02/2024) (Cited on page 88).

[64] W. M. C. Foulkes and R. Haydock. “Tight-binding models and density-functional theory”. In: *Physical Review B* 39.17 (1989), p. 12520 (Cited on page 25).

[65] T. Frauenheim et al. “Atomistic simulations of complex materials: ground-state and excited-state properties”. In: *Journal of Physics: Condensed Matter* 14.11 (2002), p. 3015 (Cited on page 27).

[66] M. M. Fréchet. “Sur quelques points du calcul fonctionnel”. In: *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 22.1 (Dec. 1, 1906), pp. 1–72. (Visited on 06/11/2024) (Cited on page 109).

[67] M. Frenklach and E. D. Feigelson. “Formation of Polycyclic Aromatic Hydrocarbons in Circumstellar Envelopes”. In: *The Astrophysical Journal* 341 (June 1, 1989), p. 372. (Visited on 03/11/2024) (Cited on page 61).

[68] M. J. Frisch et al. *Gaussian 16 Revision C.01.* 2016 (Cited on page 37).

[69] D. G. Furton, J. W. Laiho, and A. N Witt. “The amount of interstellar carbon locked in solid hydrogenated amorphous carbon”. In: *The Astrophysical Journal* 526.2 (1999), p. 752 (Cited on page 79).

[70] M. Gaus, Q. Cui, and M. Elstner. “Density functional tight binding: application to organic and biological molecules”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 4.1 (2014), pp. 49–61 (Cited on page 27).

[71] M. Gaus, Q. Cui, and M. Elstner. “DFTB3: Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method (SCC-DFTB)”. In: *Journal of chemical theory and computation* 7.4 (Apr. 2011), pp. 931–948 (Cited on pages 27, 37).

[72] M. Gaus, A. Goez, and M. Elstner. “Parametrization and benchmark of DFTB3 for organic molecules”. In: *Journal of Chemical Theory and Computation* 9.1 (2013), pp. 338–354 (Cited on page 37).

[73] B. Gipson et al. “Computational models of protein kinematics and dynamics: Beyond simulation”. In: *Annual Review of Analytical Chemistry* 5.1 (2012), pp. 273–291 (Cited on page 23).

[74] M. Godard. “Les carbones amorphes hydrogénés : observations, synthèse et caractérisation en laboratoire de poussières interstellaires”. PhD thesis. Université Paris Sud - Paris XI, Sept. 22, 2011. (Visited on 05/07/2024) (Cited on page 77).

[75] R. Goldman et al. “Smoking Increases Carcinogenic Polycyclic Aromatic Hydrocarbons in Human Lung Tissue1”. In: *Cancer Research* 61.17 (Sept. 1, 2001), pp. 6367–6371 (Cited on page 61).

[76] CM. Goringe, DR. Bowler, and E. Hernández. “Tight-binding modelling of materials”. In: *Reports on Progress in Physics* 60.12 (1997), p. 1447 (Cited on page 26).

[77] VI. Grishko and WW Duley. “Infrared absorption and emission spectra of hydrogenated amorphous carbon prepared in the presence of oxygen, nitrogen, ammonia, and carbon monoxide”. In: *The Astrophysical Journal* 568.1 (2002), p. 448 (Cited on page 79).

[78] B. L. Hammond, W. A. Lester, and P. J. Reynolds. *Monte Carlo Methods in Ab Initio Quantum Chemistry.* World Scientific, 1994 (Cited on page 23).

[79] T. Hansson, C. Oostenbrink, and W. van Gunsteren. "Molecular dynamics simulations". In: *Current opinion in structural biology* 12.2 (2002), pp. 190–196 (Cited on page 23).

[80] J. A Harrison et al. "Review of force fields and intermolecular potentials used in atomistic computational materials research". In: *Applied Physics Reviews* 5.3 (2018) (Cited on page 23).

[81] D. R. Hartree. "The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part II. Some Results and Discussion". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.1 (Jan. 1928), pp. 111–132. (Visited on 05/27/2024) (Cited on pages 5, 147).

[82] B. Hashemian, D. Millán, and M. Arroyo. "Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables". In: *The Journal of Chemical Physics* 139.21 (Dec. 2, 2013), p. 214101. (Visited on 07/13/2024) (Cited on page 116).

[83] R. Hauser and AM. Calafat. "Phthalates and human health". In: *Occupational and Environmental Medicine* 62.11 (2005), pp. 806–818 (Cited on page 36).

[84] G. Henkelman and H. Jónsson. "A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives". In: *The Journal of Chemical Physics* 111.15 (Oct. 15, 1999), pp. 7010–7022. (Visited on 06/07/2023) (Cited on page 100).

[85] G. Henkelman and H. Jónsson. "Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points". In: *The Journal of Chemical Physics* 113.22 (Dec. 8, 2000), pp. 9978–9985. (Visited on 06/06/2023) (Cited on pages 102, 103).

[86] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". In: *Physical Review* 136.3 (Nov. 9, 1964), B864–B871. (Visited on 04/12/2024) (Cited on pages 6, 148).

[87] J. H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. The MIT Press, Apr. 29, 1992. (Visited on 04/11/2024) (Cited on pages 15, 158).

[88] D. J. Hollenbach and A. G. G. M. Tielens. "Dense Photodissociation Regions (PDRs)". In: *Annual Review of Astronomy and Astrophysics* 35 (Volume 35, 1997 Sept. 1, 1997), pp. 179–215. (Visited on 05/21/2024) (Cited on page 78).

[89] J. Horiuti. "On the Statistical Mechanical Treatment of the Absolute Rate of Chemical Reaction". In: *Bulletin of the Chemical Society of Japan* 13.1 (1938), pp. 210–216 (Cited on page 100).

[90] G. Hou, X. Zhu, and Q. Cui. "An Implicit Solvent Model for SCC-DFTB with Charge-Dependent Radii". In: *Journal of Chemical Theory and Computation* 6.8 (Aug. 2010), pp. 2303–2314 (Cited on page 58).

[91] M. Imanishi. "The 3.4-μ m absorption feature towards three obscured active galactic nuclei". In: *Monthly Notices of the Royal Astronomical Society* 319.1 (2000), pp. 331–336 (Cited on page 79).

[92] L. Jaillet, J. Cortés, and T. Siméon. "Sampling-Based Path Planning on Configuration-Space Costmaps". In: *IEEE Transactions on Robotics* 26.4 (2010), pp. 635–646 (Cited on pages 18, 106, 160).

[93] L. Jaillet et al. "Randomized tree construction algorithm to explore energy landscapes". In: *Journal of Computational Chemistry* 32.16 (2011), pp. 3464–3474 (Cited on pages 18, 23, 25, 33, 106, 160).

[94] AP. Jones, WW. Duley, and DA. Williams. "The structure and evolution of hydrogenated amorphous carbon grains and mantles in the interstellar medium". In: *Royal Astronomical Society, Quarterly Journal* 31 (1990), pp. 567–582 (Cited on page 79).

[95] R. Kavlock et al. "NTP center for the evaluation of risks to human reproduction: phthalates expert panel report on the reproductive and developmental toxicity of di (2-ethylhexyl) phthalate". In: *Reproductive toxicology (Elmsford, NY)* 16.5 (2002), pp. 529–653 (Cited on page 36).

[96] L.E. Kavraki et al. "Probabilistic roadmaps for path planning in high-dimensional configuration spaces". In: *IEEE Transactions on Robotics and Automation* 12.4 (Aug. 1996), pp. 566–580. (Visited on 05/31/2024) (Cited on pages 16, 159).

[97] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. "Optimization by Simulated Annealing". In: *Science* 220.4598 (May 13, 1983), pp. 671–680. (Visited on 04/11/2024) (Cited on pages 14, 157).

[98] S. Kmiecik et al. "Coarse-Grained Protein Models and Their Applications". In: *Chemical Reviews* 116.14 (July 27, 2016), pp. 7898–7936. (Visited on 05/27/2024) (Cited on pages 8, 150).

[99] S. Knippenberg et al. "The low-lying excited states of neutral polyacenes and their radical cations: a quantum chemical study employing the algebraic diagrammatic construction scheme of second order". In: *Molecular Physics* (Oct. 10, 2010). (Visited on 05/22/2024) (Cited on page 62).

[100] T. Krüger et al. "Validation of the density-functional based tight-binding approximation method for the calculation of reaction energies and other data". In: *The Journal of Chemical physics* 122.11 (2019/10/04 2005), p. 114110 (Cited on page 27).

[101] S. Kumar et al. "The Weighted Histogram Analysis method for free-energy calculations on biomolecules". In: *Journal of Computational Chemistry* 13.8 (1992), pp. 1011–1021 (Cited on pages 13, 156).

[102] A. Laio and M. Parrinello. "Escaping free-energy minima". In: *Proceedings of the national academy of sciences* 99.20 (2002), pp. 12562–12566 (Cited on pages 13, 156).

[103] S. LaValle. “Rapidly-exploring random trees: A new tool for path planning”. In: *Research Report 9811* (1998) (Cited on pages 17, 160).

[104] S. LaValle and J. J. Kuffner Jr. “Randomized kinodynamic planning”. In: *The International Journal of Robotics Research* 20.5 (2001), pp. 378–400 (Cited on pages 23–25).

[105] A. R. Leach. *Molecular modelling: principles and applications*. Pearson education, 2001 (Cited on pages 6, 149).

[106] J. Leguy et al. “EvoMol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation”. In: *Journal of Cheminformatics* 12.1 (Sept. 16, 2020), p. 55. (Visited on 11/29/2023) (Cited on pages 62, 64).

[107] T. Lengauer and M. Rarey. “Computational methods for biomolecular docking”. In: *Current Opinion in Structural Biology* 6.3 (June 1996), pp. 402–406 (Cited on pages 16, 159).

[108] *Les Phtalates : Sources d’exposition et impregnation humaine*. Tech. rep. Paris, France, 2011 (Cited on page 36).

[109] Z. Li and H. A. Scheraga. “Monte Carlo-minimization approach to the multiple-minima problem in protein folding.” In: *Proceedings of the National Academy of Sciences* 84.19 (1987), pp. 6611–6615 (Cited on pages 14, 23, 25, 156).

[110] B. Lin and J. Su. “Shapes based trajectory queries for moving objects”. In: *Proceedings of the 13th annual ACM international workshop on Geographic information systems*. 2005, pp. 21–30 (Cited on page 110).

[111] R. Malek and N. Mousseau. “Dynamics of Lennard-Jones clusters: A characterization of the activation-relaxation technique”. In: *Physical Review E* 62.6 (Dec. 1, 2000), pp. 7723–7728. (Visited on 07/30/2023) (Cited on page 105).

[112] W. Margerit et al. “IGLOO: An Iterative Global Exploration and Local Optimization Algorithm to Find Diverse Low-Energy Conformations of Flexible Molecules”. In: *Algorithms* 16.10 (2023) (Cited on pages 16, 18, 23, 25, 28, 30, 161).

[113] R. Martín-Doménech, E. Dartois, and G. M. M. Caro. “Vacuum ultraviolet photolysis of hydrogenated amorphous carbons - III. Diffusion of photo-produced H2 as a function of temperature”. In: *Astronomy & Astrophysics* 591 (July 1, 2016), A107. (Visited on 05/06/2024) (Cited on page 79).

[114] *matsci-0-3 CC*. (Visited on 07/04/2024) (Cited on page 33).

[115] E. Mayo Yanes, S. Chakraborty, and R. Gershoni-Poranne. “COMPAS-2: a dataset of cata-condensed hetero-polycyclic aromatic systems”. In: *Scientific Data* 11.1 (Jan. 19, 2024), p. 97. (Visited on 03/11/2024) (Cited on page 62).

[116] V Mennella et al. “Activation of the 3.4 micron band in carbon grains by exposure to atomic hydrogen”. In: *The Astrophysical Journal* 524.1 (1999), p. L71 (Cited on page 79).

[117] N. Metropolis and S. Ulam. "The Monte Carlo method". In: *Journal of the American Statistical Association* 44.247 (Sept. 1949), pp. 335–341 (Cited on pages 10, 152).

[118] N. Metropolis et al. "Equation of State Calculations by Fast Computing Machines". In: *The Journal of Chemical Physics* 21.6 (2020/04/07 1953), pp. 1087–1092 (Cited on pages 10, 17, 152, 160).

[119] D. Müllner. "Modern hierarchical, agglomerative clustering algorithms". In: *arXiv preprint arXiv:1109.2378* (2011) (Cited on pages 112, 114).

[120] F. Neese. "Software update: the ORCA program system, version 4.0". In: *WIREs Computational Molecular Science* 8.1 (2018), e1327. (Visited on 03/11/2024) (Cited on page 62).

[121] F. Neese. "The ORCA program system". In: *WIREs Computational Molecular Science* 2.1 (2012), pp. 73–78. (Visited on 03/11/2024) (Cited on page 62).

[122] T. A. Niehaus et al. "Tight-binding approach to time-dependent density-functional response theory". In: *Physical Review B* 63.8 (2001), p. 085108 (Cited on page 27).

[123] A. F. Oliveira et al. "Density-functional based tight-binding: an approximate DFT method". In: *Journal of the Brazilian Chemical Society* 20 (2009), pp. 1193–1205 (Cited on page 27).

[124] R. H. Pain. *Mechanisms of Protein Folding*. Oxford: Oxford University Press, U.S.A., May 12, 1994. 284 pp. (Cited on pages 16, 159).

[125] R. G. Parr. "Density functional theory". In: *Electron Distributions and the Chemical Bond*. Springer, 1982, pp. 95–100 (Cited on page 23).

[126] Y. J. Pendleton and L. J. Allamandola. "The Organic Refractory Material in the Diffuse Interstellar Medium: Mid-Infrared Spectroscopic Constraints". In: *The Astrophysical Journal Supplement Series* 138.1 (Jan. 2002), p. 75 (Cited on page 80).

[127] C. A. Pope Iii et al. "Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution". In: *Jama* 287.9 (2002), pp. 1132–1141 (Cited on page 62).

[128] D. Porezag et al. "Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon". In: *Physical Review B* 51.19 (1995), p. 12947 (Cited on pages 23, 25).

[129] G. A. Poskrebyshev. "The standard thermochemical properties of the p-benzylphenol and dimethyl phthalate, and their temperature dependencies". In: *Computational and Theoretical Chemistry* 1197 (2021), p. 113146 (Cited on page 37).

[130] V. Ramanathan and G. Carmichael. "Global and regional climate changes due to black carbon". In: *Nature Geoscience* 1.4 (Apr. 2008), pp. 221–227. (Visited on 03/11/2024) (Cited on page 62).

[131] M. Rapacioli et al. 2023 (Cited on pages 74, 114).

[132] M. Rapacioli et al. "Extensions of DFTB to investigate molecular complexes and clusters". In: *Physica Status Solidi (b)* 249.2 (2012), pp. 245–258 (Cited on page 27).

[133] M. Rapacioli et al. "Formation and destruction of polycyclic aromatic hydrocarbon clusters in the interstellar medium". In: *Astronomy & Astrophysics* 460.2 (Dec. 1, 2006), pp. 519–531. (Visited on 03/11/2024) (Cited on page 78).

[134] *RESEARCH - NASA Ames Research Center - NAI Team*. Feb. 27, 2014. (Visited on 05/21/2024) (Cited on page 77).

[135] CA. Roth et al. "Hybridizing rapidly exploring random trees and basin hopping yields an improved exploration of energy landscapes". In: *Journal of Computational Chemistry* 37.8 (2016), pp. 739–752 (Cited on page 23).

[136] SA. Sandford et al. "The interstellar CH stretching band near 3.4 microns-Constraints on the composition of organic material in the diffuse interstellar medium". In: *Astrophysical Journal, Part 1* 371 (1991), pp. 607–620 (Cited on page 79).

[137] M. G. Saunders and G. A. Voth. "Coarse-Graining Methods for Computational Biology". In: *Annual Review of Biophysics* 42 (Volume 42, 2013 May 6, 2013), pp. 73–93. (Visited on 05/27/2024) (Cited on pages 8, 150).

[138] H. B. Schlegel. "Exploring potential energy surfaces for chemical reactions: An overview of some practical methods". In: *Journal of Computational Chemistry* 24.12 (2003), pp. 1514–1527 (Cited on page 23).

[139] E. Schrödinger. "Quantisierung als Eigenwertproblem". In: *Annalen der Physik* 384.4 (1926), pp. 361–376 (Cited on pages 2, 144).

[140] AD Scott, WW Duley, and HR Jahani. "Infrared emission spectra from hydrogenated amorphous carbon". In: *The Astrophysical Journal* 490.2 (1997), p. L175 (Cited on page 79).

[141] G. Seifert, D. Porezag, and Th. Frauenheim. "Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme". In: *International Journal of Quantum Chemistry* 58.2 (1996), pp. 185–192 (Cited on pages 23, 25).

[142] D. Selli, G. Fazio, and C. Di Valentin. "Modelling realistic TiO2 nanospheres: A benchmark study of SCC-DFTB against hybrid DFT". In: *The Journal of Chemical Physics* 147.16 (2019/09/16 2017), p. 164701 (Cited on page 23).

[143] A. Shehu and E. Plaku. "A survey of computational treatments of biomolecules by robotics-inspired methods modeling equilibrium structure and dynamic". In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 509–572 (Cited on page 23).

[144] D. S. Sholl and J.e A. Steckel. *Density Functional Theory: A Practical Introduction*. John Wiley & Sons, Mar. 2009 (Cited on pages 5, 148).

[145] A. P. Singh, JC. Latombe, and D. L. Brutlag. "A motion planning approach to flexible ligand binding." In: *ISMB*. 1999, pp. 252–261 (Cited on pages 16, 159).

[146] J. C. Slater. "Note on Hartree's Method". In: *Physical Review* 35.2 (Jan. 15, 1930), pp. 210–211. (Visited on 05/27/2024) (Cited on pages 5, 147).

[147] F. Spiegelman et al. "Density-functional tight-binding: basic concepts and applications to molecules and clusters". In: *Advances in Physics: X* 5.1 (2020), p. 1710252 (Cited on pages 23, 27).

[148] HWW. Spoon et al. "Ice features in the mid-IR spectra of galactic nuclei". In: *Astronomy & Astrophysics* 385.3 (2002), pp. 1022–1041 (Cited on page 79).

[149] C. A. Staples et al. "The environmental fate of phthalate esters: A literature review". In: *Chemosphere* 35.4 (1997), pp. 667–749 (Cited on page 35).

[150] Y. Sugita and Y. Okamoto. "Replica-exchange molecular dynamics method for protein folding". In: *Chemical Physics Letters* 314.1 (Nov. 26, 1999), pp. 141–151. (Visited on 04/11/2024) (Cited on pages 12, 23, 155).

[151] Y. Sugita and Y.o Okamoto. "Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape". In: *Chemical Physics Letters* 329.3 (Oct. 20, 2000), pp. 261–270. (Visited on 04/11/2024) (Cited on pages 12, 23, 155).

[152] R. H. Swendsen and J. Wang. "Replica Monte Carlo Simulation of Spin-Glasses". In: *Physical Review Letters* 57.21 (Nov. 24, 1986), pp. 2607–2609. (Visited on 04/11/2024) (Cited on pages 12, 23, 155).

[153] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Courier Corporation, July 2, 1996. 484 pp. (Cited on pages 5, 148).

[154] A.G.G.M. Tielens. "Interstellar Polycyclic Aromatic Hydrocarbon Molecules". In: *Annual Review of Astronomy and Astrophysics* 46.1 (2008), pp. 289–337. (Visited on 03/11/2024) (Cited on page 78).

[155] C. Tönshoff and H. F. Bettinger. "Pushing the Limits of Acene Chemistry: The Recent Surge of Large Acenes". In: *Chemistry – A European Journal* 27.10 (2021), pp. 3193–3212. (Visited on 05/22/2024) (Cited on page 62).

[156] G. M. Torrie and J. P. Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling". In: *Journal of Computational Physics* 23.2 (Feb. 1, 1977), pp. 187–199. (Visited on 04/11/2024) (Cited on pages 13, 155).

[157] P. J. M. Van Laarhoven and E. H. L. Aarts. *Simulated Annealing: Theory and Applications*. Dordrecht: Springer Netherlands, 1987. (Visited on 04/11/2024) (Cited on pages 14, 157).

[158] L. Verstraete. "The role of PAHs in the physics of the interstellar medium". In: *EAS Publications Series* 46 (2011), pp. 415–426. (Visited on 05/21/2024) (Cited on page 78).

[159] A. Wahab and R. Gershoni-Poranne. "COMPAS-3: a dataset of peri-condensed polybenzenoid hydrocarbons". In: *Physical Chemistry Chemical Physics* 26.21 (2024), pp. 15344–15357 (Cited on page 62).

[160] A. Wahab et al. "The compas project: A computational database of polycyclic aromatic systems. phase 1: cata-condensed polybenzenoid hydrocarbons". In: *Journal of Chemical Information and Modeling* 62.16 (2022), pp. 3704–3713 (Cited on pages 62, 63).

[161] V. Wakelam et al. "H2 formation on interstellar dust grains: The viewpoints of theory, experiments, models and observations". In: *Molecular Astrophysics* 9 (Dec. 1, 2017), pp. 1–36. (Visited on 05/21/2024) (Cited on page 79).

[162] D. J. Wales and J. PK. Doye. "Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms". In: *The Journal of Physical Chemistry A* 101.28 (1997), pp. 5111–5116 (Cited on pages 14, 23, 25, 156).

[163] *Webpage of Mathias Rapacioli.* (Visited on 04/29/2024) (Cited on page 30).

[164] E. Weinan, W. Ren, and E. Vanden-Eijnden. "String method for the study of rare events". In: *Physical Review B* 66.5 (2002), p. 052301 (Cited on page 104).

[165] E. Wigner. "The transition state method". In: *Transactions of the Faraday Society* 34.0 (Jan. 1, 1938), pp. 29–41. (Visited on 06/05/2024) (Cited on page 100).

[166] M. J. Wornat et al. "Polycyclic aromatic hydrocarbons identified in soot extracts from domestic coal-burning stoves of Henan Province, China". In: *Environmental Science & Technology* 35.10 (May 15, 2001), pp. 1943–1952 (Cited on page 61).

[167] M. Xue and C. Zhu. "The Socket Programming and Software Design for Communication Based on Client/Server". In: *2009 Pacific-Asia Conference on Circuits, Communications and Systems*. 2009, pp. 775–777 (Cited on page 28).

[168] M. Yang et al. "Combine umbrella sampling with integrated tempering method for efficient and accurate calculation of free energy changes of complex energy surface". In: *The Journal of Chemical Physics* 141.4 (2014) (Cited on page 23).

[169] Y. Yang et al. "Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction". In: *The Journal of Physical Chemistry A* 111.42 (2007), pp. 10861–10873 (Cited on page 27).

[170] T. W. Yen and S. K. Lai. "Use of Density Functional Theory Method to Calculate Structures of Neutral Carbon Clusters Cn ($3 \leq n \leq 24$) and Study their Variability of Structural Forms". In: *The Journal of chemical physics* 142.8 (2015), p. 084313 (Cited on page 23).

[171] TW. Yen et al. "Studying the Varied Shapes of Gold Clusters by an Elegant Optimization Algorithm that Hybridizes the Density Functional Tight-Binding Theory and the Density Functional Theory". In: *Computer Physics Communications* 220 (2017), pp. 143–149 (Cited on page 23).

[172] M. Yusef Buey, T. Mineva, and M. Rapacioli. "Coupling density functional based tight binding with class 1 force fields in a hybrid QM/MM scheme". In: *Theoretical Chemistry Accounts* 141.3 (2022), p. 16 (Cited on page 58).

[173] 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". en. In: *Theoretical Chemistry Accounts* 120.1 (May 2008), pp. 215–241. (Visited on 10/02/2023) (Cited on page 37).

[174] L. Zhechkov et al. "An Efficient a Posteriori Treatment for Dispersion Interaction in Density-Functional-Based Tight Binding". In: *Journal of Chemical Theory and Computation* 1.5 (2005), pp. 841–847 (Cited on page 37).

Annexe A

Introduction en français

Contents

A.1 Introduction

Un défi de taille dans le domaine des systèmes atomiques et moléculaires est d'acquérir une compréhension plus approfondie de leurs propriétés fondamentales. Ce défi est d'autant plus important que les systèmes étudiés deviennent de plus en plus complexes et que la nécessité d'une exploration efficace de leur paysage énergétique pour prédire leur comportement dans divers environnements physiques et biologiques augmente. Ce chapitre présentera les principaux concepts développés dans cette thèse afin de répondre au problème posé et de contextualiser les méthodes de l'état de l'art. Un cadre théorique sera développé pour introduire le concept de structures électroniques et définir les équations de Schrödinger et l'approximation de Born-Oppenheimer. Ensuite, les méthodes d'énergie potentielle seront présentées, chaque méthode traitant de problèmes différents et variant principalement en termes de précision et d'efficacité. Enfin, des méthodes d'exploration globale seront introduites pour illustrer la diversité des techniques disponibles pour étudier la surface d'énergie potentielle (SEP).

A.2 Modélisation atomistique/moléculaire et structure électronique

Les atomes sont composés d'un noyau contenant des protons et des neutrons, entouré d'électrons. La disposition des électrons dans les orbitales détermine les propriétés chimiques de l'atome et ses interactions avec d'autres atomes. Les propriétés des atomes et des molécules sont régies par les lois de la mécanique quantique, qui décrivent le comportement des particules aux niveaux atomique et subatomique. Les principes de la mécanique quantique fournissent un cadre pour comprendre la structure des atomes et des molécules, la nature des liaisons chimiques et les interactions entre les molécules.

Les molécules sont composées de deux ou plusieurs atomes reliés entre eux par des liaisons chimiques. L'arrangement spécifique des atomes au sein d'une molécule détermine sa forme et ses propriétés. Les molécules présentent un large éventail de propriétés et de comportements, qui dépendent de leur composition, de leur structure et de leurs interactions. La compréhension des caractéristiques des molécules est essentielle pour prédire leur comportement et leurs propriétés dans divers processus chimiques.

A.2.1 Équation de Schrödinger

En 1926, Erwin Schrödinger, physicien autrichien, a introduit une fonction d'onde qui décrit comment l'état quantique d'un système physique change au fil du temps [139]. La formulation de l'équation de Schrödinger est l'une des principales réalisations dans le domaine de la chimie quantique. Cette équation est l'un des postulats les plus importants de la mécanique quantique et a joué un rôle crucial dans notre compréhension du monde subatomique.

L'équation de Schrödinger dépendante du temps s'écrit :

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{R},\mathbf{r},t) = \hat{H}\Psi(\mathbf{R},\mathbf{r},t) \tag{A.1}$$

où $\hbar$ est la constante de Planck réduite, $\hbar = \frac{h}{2\pi}$, h est la constante de Planck égale à $6.62607015.10^{-34}$ J.s., i est l'unité imaginaire, $\Psi(\mathbf{R},\mathbf{r},t)$ est la fonction d'onde du système, qui contient des informations sur la position du noyau $\mathbf{R}$, des électrons $\mathbf{r}$ et du temps t, et $\hat{H}$ est l'opérateur hamiltonien, associé à l'énergie totale du système. Dans le cas particulier des systèmes en état stationnaire, c'est-à-dire ceux dont les propriétés ne varient pas dans le temps, les variables temporelles et spatiales sont séparées. La fonction d'onde d'état (fonction propre) est définie comme :

$$\Psi(\mathbf{R},\mathbf{r},t) = e^{-i\frac{Et}{\hbar}}\psi(\mathbf{R},\mathbf{r}) \tag{A.2}$$

$\psi(\mathbf{R},\mathbf{r})$ représentent la contribution spatiale à la fonction d'onde et peuvent être obtenus en résolvant l'équation de Schrödinger indépendante du temps. Cette équation s'écrit:

$$\hat{H}\psi(\mathbf{R},\mathbf{r}) = E\psi(\mathbf{R},\mathbf{r}) \tag{A.3}$$

L'opérateur hamiltonien joue un rôle central dans l'équation, dictant la dynamique du système en définissant le paysage énergétique dans lequel le système évolue.

Sous cette forme, l'opérateur hamiltonien appliqué à la fonction d'onde $\psi(\mathbf{r})$ est égal à l'énergie E du système multipliée par la fonction d'onde. L'opérateur hamiltonien est la somme de l'opérateur cinétique et des termes d'énergie potentielle, et est défini comme:

$$\hat{H} = \hat{T}_e + \hat{T}_n + \hat{V}_{ee} + \hat{V}_{en} + \hat{V}_{nn} \tag{A.4}$$

où e et n désignent respectivement les composantes électronique et nucléaire, et les indices indiquent le type d'interaction (électron-électron, électron-noyau et noyau-noyau). Les termes d'énergie cinétique $\hat{T}_e$ et $\hat{T}_n$ sont les opérateurs associés aux énergies cinétiques des électrons et du noyau. Les termes d'énergie potentielle $\hat{V}_{ee}$ représentent l'interaction de Coulomb répulsive entre les électrons, $\hat{V}_{en}$ l'interaction de Coulomb attractive entre le noyau et les électrons et $\hat{V}_{nn}$ l'interaction de Coulomb répulsive entre les noyaux. En définissant un système de particules avec N électrons et M noyaux, les cinq termes de l'hamiltonien (Eq. A.4) en unités atomiques ($\hbar = m_e = e = c = 1$) peuvent être écrits :

$$\hat{T}_e = -\frac{1}{2}\sum_{i=1}^{N} \nabla_i^2 \tag{A.5}$$

$$\hat{T}_n = -\frac{1}{2}\sum_{A=1}^{M} \frac{1}{M_A}\nabla_A^2 \tag{A.6}$$

$$\hat{V}_{en} = -\sum_{i=1}^{N}\sum_{A=1}^{M} \frac{Z_A}{r_{iA}} \tag{A.7}$$

$$\hat{V}_{ee} = \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{1}{r_{ij}} \tag{A.8}$$

$$\hat{V}_{nn} = \sum_{A=1}^{M}\sum_{B>A}^{M} \frac{Z_A Z_B}{R_{AB}} \tag{A.9}$$

où r_{iA} est la distance entre le i-ième électron et le A-ième noyau, r_{ij} est la distance entre le i-ième et le j-ième électron, R_{AB} est la distance entre le A-ième et le B-ième noyau, Z_A est le numéro atomique du A-ième noyau et M_A est la masse du A-ième noyau. Enfin, l'opérateur laplacien ∇^2 est défini en coordonnées cartésiennes comme $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$.
En pratique, il est impossible de résoudre analytiquement l'équation de Schrödinger pour les systèmes comportant plus d'un électron. C'est pourquoi des méthodes numériques combinées à des approximations sont utilisées pour résoudre cette équation.

A.2.2 Approximation de Born-Oppenheimer

En 1927, Max Born et Robert Oppenheimer ont introduit l'approximation de Born-Oppenheimer, qui simplifie l'équation de Schrödinger en traitant les mouvements élec-

troniques et nucléaires comme des variables indépendantes [28]. La masse du noyau est beaucoup plus importante que la masse des électrons ($m_p \approx 1836 m_e$), ce qui rend le mouvement électronique beaucoup plus rapide que le mouvement du noyau, ce qui conduit à une séparation des variables pour décrire les mouvements électroniques et nucléaires. En conséquence, la fonction d'onde électronique peut s'ajuster instantanément aux changements de positions du noyau, et les électrons se déplacent dans un champ potentiel généré par le noyau. La fonction d'onde totale peut être écrite comme un produit des fonctions d'onde électronique et nucléaire :

$$\Psi(\mathbf{r}, \mathbf{R}) = \psi_n(\mathbf{R})\psi_e(\mathbf{r}; \mathbf{R}) \tag{A.10}$$

où $\psi_n(\mathbf{R})$ est la fonction d'onde nucléaire, et $\psi_e(\mathbf{r}; \mathbf{R})$ est la fonction d'onde électronique qui dépend paramétriquement des coordonnées nucléaires. Le problème électronique peut être résolu indépendamment du mouvement nucléaire. À une position donnée des noyaux, le problème électronique peut être obtenu en résolvant une équation de Schrödinger indépendante du temps pour les électrons uniquement :

$$\hat{H}_e\psi_e(\mathbf{r}; \mathbf{R}) = E_e(\mathbf{R})\psi_e(\mathbf{r}; \mathbf{R}) \tag{A.11}$$

où E_e est l'énergie électronique et l'opérateur hamiltonien électronique est donné comme suit :

$$\hat{H}_e = \hat{T}_e + \hat{V}_{ee} + \hat{V}_{en} \tag{A.12}$$

Notez que la résolution de l'équation A.11 conduit à plusieurs solutions correspondant à différents états électroniques. Dans de nombreux cas, et en particulier dans la suite de cette thèse, seule la solution correspondant à la valeur propre la plus basse en énergie est considérée, également appelée état fondamental électronique. La dynamique des noyaux est gouvernée par l'énergie potentielle obtenue en ajoutant la répulsion nucléaire à l'énergie électronique :

$$E(\mathbf{R}) = E_e(\mathbf{R}) + \hat{V}_{nn} \tag{A.13}$$

Selon l'équation A.13, l'énergie potentielle du système peut être calculée pour une configuration nucléaire donnée. Cette approximation simplifie l'équation de Schrödinger en réduisant le nombre de variables et permet de définir une surface d'énergie potentielle qui décrit le paysage énergétique du système chimique. En plus de la première approximation, les noyaux sont souvent considérés comme des particules classiques (particules ponctuelles). Ils peuvent donc être traités à partir de la loi de Newton en utilisant les SEP définis dans l'équation A.13.

A.3 Énergie potentielle

Pour calculer l'énergie potentielle d'un système, plusieurs méthodes ont été développées dans le domaine de la chimie computationnelle, variant en complexité et en précision comme présenté dans la figure A.1. Le choix de la taille de système atteignable avec la méthode et le coût de calcul sont corrélés, de sorte que seul le niveau de précision requis

et le coût de calcul doivent être déterminé. La section suivante présente les principales méthodes utilisées pour le calcul de l'énergie potentielle d'un système, y compris celles basées sur les fonctions d'onde, ainsi que la Density functional theory (DFT), la Density Functional Tight Binding (DFTB), les champs de force (FF) et le coarse-graining (CG).

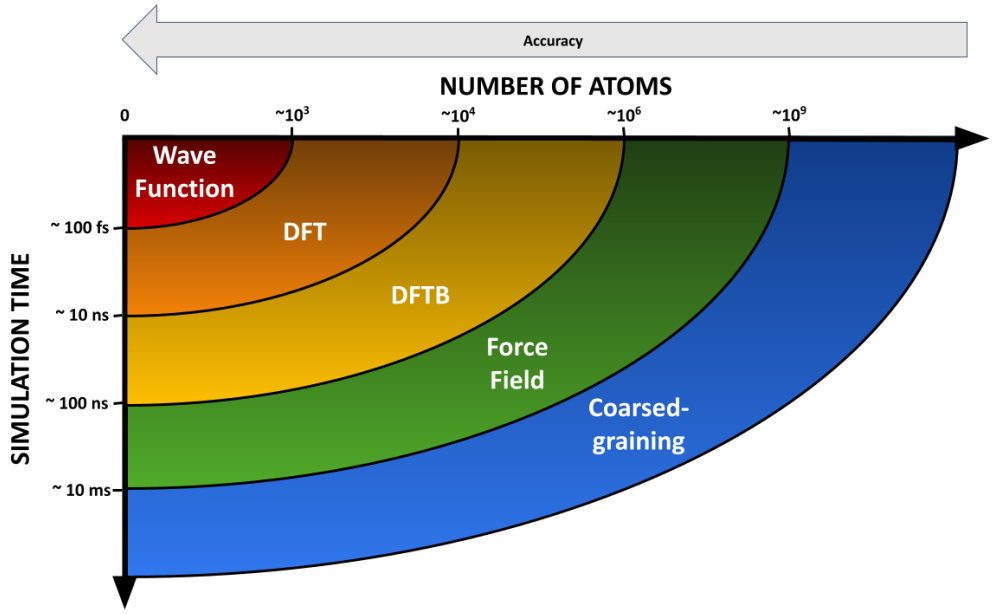


Figure A.1: Échelle d'énergie potentielle.

A.3.1 Méthodes basées sur la fonction d'onde

La méthode Hartree-Fock (HF) [13, 81, 146, 61, 60] est une approche de champ moyen dans laquelle les électrons sont supposés évoluer indépendamment dans un potentiel effectif façonné à la fois par les noyaux et l'ensemble des électrons. La méthode approxime la fonction d'onde multi-électronique totale du système comme un produit de fonctions d'onde mono-électroniques. Le schéma de calcul implique la résolution d'un ensemble de N équations de Schrödinger à une seule particule de manière autoconsistante.

Néanmoins, Fock a identifié une lacune importante dans la formulation de Hartree, à savoir sa non-conformité avec le principe d'exclusion de Pauli. Il en résulte une fonction d'onde qui n'est pas antisymétrique en ce qui concerne l'échange de particules. Pour y remédier, Fock a reformulé la fonction d'onde sous la forme d'un déterminant de Slater de fonctions à un seul électron, en incorporant les caractéristiques fermioniques des électrons et en introduisant le terme d'énergie d'échange dans l'hamiltonien. C'est ainsi que la méthode originale a évolué vers ce qui est aujourd'hui largement reconnu comme la méthode Hartree-Fock.

Malgré ses contributions importantes, la méthode Hartree-Fock ne tient pas compte de l'énergie de corrélation, c'est-à-dire de l'écart entre l'énergie mécanique quantique exacte et l'énergie estimée par les calculs Hartree-Fock.

A.3.2 Density functional theory (DFT)

La Density functional theory (DFT) [144, 153] a été introduite par Hohenberg et Kohn en 1964 et développée par Kohn et Sham en 1965. La DFT est une théorie de mécanique quantique utilisée pour étudier la structure électronique des systèmes à corps multiples, principalement les atomes, les molécules et les phases condensées. Contrairement aux méthodes basées directement sur la fonction d'onde, la DFT décrit un système en termes de densité électronique plutôt que de fonction d'onde.

La DFT repose sur deux théorèmes de Hohenberg-Kohn (HK) [86] :

1. Le premier théorème HK stipule que les propriétés de l'état fondamental d'un système à plusieurs électrons sont déterminées de manière unique par sa densité électronique $\rho(\mathbf{r})$. Cela implique que toutes les propriétés observables du système sont des fonctionnelles de la densité électronique.

2. Le deuxième théorème HK fournit un principe variationnel pour la densité électronique. Il stipule que la fonctionnelle de l'énergie totale $E[\rho]$ a sa valeur minimale à l'état fondamental réel de la densité électronique du système.

Sur la base de ces théorèmes, Kohn et Sham ont développé un schéma pratique connu sous le nom d'équations de Kohn-Sham (KS) :

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}), \tag{A.14}$$

où $\psi_i(\mathbf{r})$ sont les orbitales de Kohn-Sham, ε_i sont leurs valeurs propres correspondantes, et $V_{\text{eff}}(\mathbf{r})$ est le potentiel effectif qui comprend le potentiel externe, le potentiel de Hartree et le potentiel d'échange-corrélation. Le potentiel effectif s'exprime comme suit :

$$V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_{\text{Hartree}}[\rho(\mathbf{r})] + V_{\text{xc}}[\rho(\mathbf{r})]. \tag{A.15}$$

Le potentiel d'échange-corrélation $V_{\text{xc}}[\rho(\mathbf{r})]$ est le composant le plus critique dans les calculs DFT et incorpore tous les effets de nombreux corps. La détermination d'une fonctionnelle précise pour V_{textxc} est un domaine de recherche majeur en DFT.

La DFT est considérée comme plus précise et plus efficace que la méthode Hartree-Fock (HF), car elle inclut intrinsèquement les effets de corrélation électronique. Elle est largement utilisée pour calculer la structure électronique des molécules et prédire leurs propriétés.

A.3.3 Density Functional Tight Binding (DFTB)

La Density Functional Tight Binding (DFTB) est une méthode semi-empirique qui approxime la structure électronique d'un système à l'aide d'un ensemble de base minimal. La méthode DFTB a été introduite par Elstner en 1998 [55]. Elle est basée sur l'approximation de la liaison forte, qui simplifie la structure électronique d'un système. La méthode DFTB est un outil de chimie computationnelle populaire pour calculer la structure électronique et prédire les propriétés des molécules. Elle est particulièrement

efficace pour les systèmes de grande taille et intègre les effets de corrélation électronique. Le chapitre 2 fournira une présentation plus détaillée de la méthode DFTB.

A.3.4 Force Field (FF)

Les méthodes de champ de force (FF) [105, 6] sont des approches classiques utilisées en chimie computationnelle pour estimer l'énergie potentielle d'un système. Ces méthodes, qui trouvent leurs racines dans la mécanique classique, utilisent les équations classiques du mouvement pour décrire la manière dont les positions et les vitesses des particules évoluent au fil du temps. Une méthode FF approxime l'énergie potentielle d'un système sur la base des positions de ses atomes. Elles sont particulièrement efficaces pour simuler de grands systèmes moléculaires en raison de leur faible temps de calcul par rapport aux méthodes de la mécanique quantique. Cette augmentation de l'efficacité s'accompagne d'une diminution concomitante de la précision des calculs. L'énergie potentielle U d'un système dans la méthode du champ de force est généralement exprimée comme une somme de contributions provenant d'interactions liées et non liées :

$$U = U_{\text{bond}} + U_{\text{angle}} + U_{\text{dihedral}} + U_{\text{non-bonded}}, \tag{A.16}$$

chaque composante étant définie comme suit :

$$U_{\text{bond}} = \sum_{\text{bonds}} k_i^b (r - r_0)^2, \tag{A.17}$$

$$U_{\text{angle}} = \sum_{\text{angles}} k_i^\theta (\theta - \theta_0)^2, \tag{A.18}$$

$$U_{\text{dihedral}} = \sum_{\text{dihedrals}} k_i^\phi [1 + \cos(n\phi - \delta)], \tag{A.19}$$

$$U_{\text{non-bonded}} = \sum_{\text{non-bonded pairs}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right]. \tag{A.20}$$

Ici, k_i^b, k_i^θ, and k_i^ϕ sont les constantes de force pour les longueurs de liaison, les angles de liaison, et l'angle dièdre respectivement ; r_0 et θ_0 sont les valeurs d'équilibre pour les longueurs de liaison et les angles de liaison. ϕ est l'angle dièdre, δ est la phase et n définit le nombre de minima ou de maxima entre 0 and 2π. A_{ij} et B_{ij} sont les paramètres du potentiel de Lennard-Jones décrivant les interactions de van der Waals, tandis que q_i et q_j sont les charges sur les atomes i et j, et r_{ij} est la distance qui les sépare. ε_0 est la permittivité du vide.

L'efficacité des méthodes FF permet de simuler des milliers voire des dizaines de milliers d'atomes en simplifiant les interactions entre atomes par l'utilisation de ressorts. Ces méthodes sont donc indispensables pour l'étude des grandes biomolécules telles que les protéines et les acides nucléiques. De plus, l'incorporation d'interactions non liées, telles que les forces de van der Waals et les forces électrostatiques, permet une représentation plus précise de la dynamique et des propriétés moléculaires.

A.3.5 Coarsed-graining (CG)

La méthode Coarsed-graining (CG) est une approche utilisée pour approximer l'énergie potentielle d'un système en réduisant le nombre de degrés de liberté. L'approche CG est basée sur le concept d'interactions effectives, qui simplifie le paysage énergétique du système en regroupant les particules en gros grains. Cette méthode est largement utilisée en chimie informatique pour calculer l'énergie potentielle des molécules et prédire leurs propriétés. La CG est plus efficace que les méthodes atomistiques en termes de taille maximale du système qui peut être simulé dans un délai convenable pour les grands systèmes. Cet avantage est à mettre en perspective avec la perte de précision.

Le principal avantage de la méthode de Coarsed-graining est sa capacité à capturer les propriétés physiques essentielles d'un système tout en omettant les détails fins qui n'affectent pas de manière significative le comportement global. En rationalisant le modèle de calcul, les méthodes CG peuvent accélérer considérablement les calculs, ce qui permet de simuler des phénomènes macroscopiques et d'explorer les comportements des systèmes à des échelles impossibles à atteindre avec les approches atomistiques conventionnelles.

L'élaboration d'un modèle à gros grains implique la sélection des sites à gros grains appropriés. Les paramètres sont souvent dérivés de données expérimentales ou de simulations atomistiques de haut niveau et doivent être ajustés pour garantir que le modèle CG reproduise les propriétés spécifiques souhaitées, telles que le comportement des phases ou les coefficients de diffusion. Une fois le modèle développé, il est essentiel de le valider et de l'affiner en comparant ses prédictions avec des résultats expérimentaux ou des simulations plus détaillées, en procédant aux ajustements nécessaires pour améliorer la précision et la fiabilité.

Le coarse-graining est largement appliqué à l'étude des macromolécules biologiques telles que les protéines et les acides nucléiques [98, 137, 33], permettant aux chercheurs d'étudier les changements de conformation à grande échelle et les interactions complexes sur des périodes prolongées. Il s'agit également d'un outil essentiel dans la science des matériaux, en particulier dans l'étude des polymères et des matériaux mous, où la compréhension de la structure et de la dynamique à grande échelle est vitale.

Malgré ses nombreux avantages, le coarse-graining pose également certains problèmes, notamment la perte d'informations détaillées au niveau atomique, qui peuvent être cruciales pour comprendre des propriétés spécifiques telles que la cinétique de réaction ou les informations électroniques détaillées. En outre, le succès d'un modèle à gros grains dépend de l'équilibre entre les détails qui sont conservés et ceux qui sont éliminés, ce qui nécessite une connaissance approfondie du système et des techniques de modélisation.

A.4 Méthodes d'exploration globale

L'exploration de la Surface d'Énergie Potentielle est essentielle pour obtenir des informations sur les configurations les plus stables, les états singuliers ou les propriétés thermodynamiques d'un système chimique. La SEP ou hypersurface énergétique $E(\mathbf{R})$

est une représentation de l'énergie potentielle d'un système en fonction de sa géométrie, qui peut être définie selon les positions atomiques ou les coordonnées internes. La SEP, également appelée paysage énergétique, montre les configurations stables et les régions de transition entre les différentes configurations. La SEP peut être utilisée pour déterminer la géométrie d'équilibre d'une molécule, prédire les propriétés des molécules, simuler des réactions chimiques et l'énergie d'activation requise pour que les réactions chimiques se produisent. Pour illustrer le concept, cette surface peut être représentée graphiquement sous la forme d'une surface bidimensionnelle, où l'énergie est tracée en fonction de la géométrie du système chimique (Fig. A.2). Il convient de noter qu'en réalité, le SEP est définie par le nombre de coordonnées du système étudié.

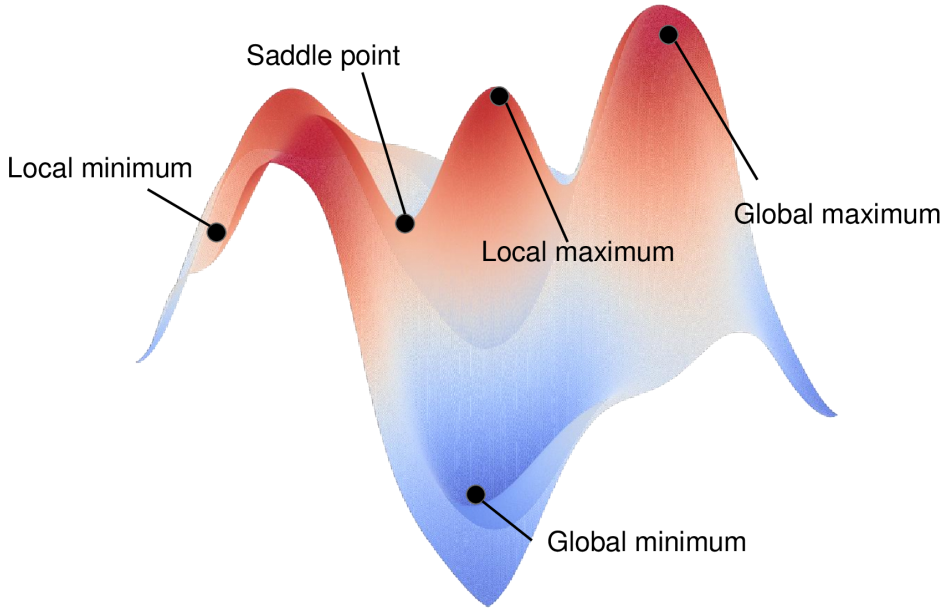


Figure A.2: Surface d'énergie potentielle d'une molécule.

Des états caractéristiques peuvent être définis sur la SEP, tels que les minima locaux. Pour un tel état, chaque dérivée première de l'énergie par rapport aux coordonnées géométriques est égale à zéro et chaque dérivée seconde est positive. Le minimum global est la configuration la plus stable du système.

Un point de selle sur une SEP est un point critique où le gradient (dérivée première) de l'énergie par rapport à toutes les coordonnées est nul, mais où la matrice Hessienne (dérivée seconde) présente une signature mixte et contient à la fois des valeurs propres positives et négatives. La présence d'au moins une valeur propre négative indique une direction d'instabilité, ce qui distingue un point de selle d'un minimum local.

Les points selles sont classés en fonction de leur *index*, qui est le nombre de valeurs propres négatives de la matrice hessienne en ce point. Cet indice détermine l'ordre du point selle :

- Un point selle de *premier ordre* , souvent appelé simplement point selle, possède

exactement une valeur propre négative. Ce type de point selle représente généralement un état de transition le long d'un chemin réactionnel.

- Les points de selle d'ordre supérieur ont plus d'une valeur propre négative et représentent des états de transition plus complexes qui peuvent impliquer des changements simultanés dans plusieurs directions.

La compréhension de l'ordre des points de selle est cruciale pour l'analyse de la voie et du mécanisme des réactions chimiques. Par exemple, l'identification des points de selle du premier ordre est essentielle pour localiser les états de transition, qui sont cruciaux pour le calcul des énergies d'activation et des vitesses de réaction.

Le calcul de l'énergie potentielle le long d'un seul degré de liberté permet de comprendre les propriétés d'une molécule pour une étude simple d'un système H_2 par exemple. Cependant, pour d'autres systèmes, la SEP devient plus complexes et de nombreux bassins dégénérés peuvent exister. Dans ce contexte, l'exploration de la SEP devient une tâche difficile car elle implique la recherche des configurations les plus stables de la molécule et des états de transition entre les différentes configurations. Plusieurs méthodes ont été mises au point pour explorer une SEP. Ces méthodes englobent un large éventail de techniques, y compris celles basées sur des méthodes Monte Carlo et de dynamique moléculaire, des techniques d'optimisation globale, ainsi que des approches inspirées de la robotique et dérivées de la planification de trajectoire. Les algorithmes axés sur la recherche de chemins de transition entre les bassins d'énergie, qui sont essentiels pour comprendre la dynamique des réactions chimiques et prédire les taux de réaction/transition, seront examinés au chapitre 6.

A.4.1 Méthodes d'échantillonnage standard

Cette section présente les techniques d'échantillonnage les plus utilisées pour étudier la SEP des molécules. La dynamique moléculaire et la méthode de Monte Carlo sont efficaces afin d'obtenir des propriétés thermodynamiques pour les deux et des propriétés cinétiques pour la dynamique moléculaire. Ces deux méthodes utilisent l'ensemble canonique comme procédure d'échantillonnage.

A.4.1.1 Monte Carlo (MC)

La méthode de Monte Carlo (MC) a été développée par Metropolis et Ulam dans les années 1940 dans le but de calculer des intégrales multidimensionnelles [117]. La méthode de Monte Carlo (MC) est une technique stochastique fondamentale utilisée pour explorer la surface d'énergie potentielle des systèmes chimiques en échantillonnant des configurations de manière aléatoire. La méthode de Monte Carlo la plus courante est celle de Metropolis (MMC). La MMC proposée par Metropolis et al. [118] est une méthode MC largement utilisée qui génère une séquence de configurations en acceptant ou en rejetant les mouvements proposés sur la base du critère de Metropolis. Le critère de Metropolis est basé sur la distribution de Boltzmann, qui stipule que la probabilité qu'un système se trouve dans un état particulier dépend exponentiellement de son énergie. Le critère

de Metropolis est donné par la formule suivante:

$$P_{\text{accept}} = \min\left(1, \exp\left(-\frac{\Delta E}{k_B T}\right)\right) \tag{A.21}$$

où ΔE est le différence d'énergie entre deux états du système, k_B est la constante de Boltzmann égale à 1.380649 $10^{-23} m^2 kg s^{-2} K^{-1}$, et T est la température du système. Le critère de Metropolis garantit que le système se déplace vers des états d'énergie plus faibles, correspondant aux configurations les plus stables de la molécule. La méthode MMC est largement utilisée en chimie informatique pour explorer les SEP, optimiser les structures moléculaires et simuler les réactions chimiques. La puissance de cette méthode réside dans sa simplicité et sa polyvalence, car elle nécessite un minimum d'hypothèses sur le système étudié. Toutefois, cette méthode peut être inefficace si l'échantillonnage aléatoire ne couvre pas efficacement les régions significatives des SEP. Cette limitation est souvent atténuée par des variantes plus sophistiquées des méthodes de Monte Carlo. L'efficacité de la technique de Monte Carlo dépend fortement du nombre d'échantillons et de la distribution à partir de laquelle ces échantillons sont tirés. Il est donc essentiel d'assurer une couverture large et représentative de l'espace d'état pour obtenir des résultats précis.

A.4.1.2 Dynamique Moléculaire (DM)

La Dynamique moléculaire (DM) est un outil puissant pour étudier le comportement dynamique des molécules en résolvant les équations classiques du mouvement des atomes dans la molécule. Les simulations de DM sont basées sur les lois du mouvement de Newton (Eq. A.22), qui décrivent comment les positions et les vitesses des particules changent au fil du temps. Les équations du mouvement sont intégrées numériquement pour simuler le mouvement des atomes dans la molécule. Les simulations de DM peuvent être utilisées pour explorer les SEP, optimiser les structures moléculaires et simuler des réactions chimiques.

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\frac{\partial E}{\partial \mathbf{r}_i} = F_i \tag{A.22}$$

où m_i est la masse du i-ème atome, $\mathbf{r}_i$ est la position du i-ème atome, E est l'énergie du système, et F_i est la force agissant sur le i-ème atome. La force agissant sur les atomes est calculée à partir du gradient de l'énergie potentielle.
Pour modéliser le système, l'équation du mouvement doit être intégrée à l'aide de divers algorithmes, dont la plupart sont basés sur l'expansion des séries de Taylor. L'algorithme de Verlet est l'algorithme le plus couramment utilisé et s'exprime comme suit :

$$\begin{aligned} \mathbf{r}_i(t + \Delta t) &= \mathbf{r}_i(t) + \mathbf{v}_i(t)\Delta t + \mathbf{a}_i(t)\frac{\Delta t^2}{2} \\ \mathbf{v}_i(t + \Delta t) &= \mathbf{v}_i(t) + \frac{\Delta t}{2}\left(\mathbf{a}_i(t) + \mathbf{a}_i(t + \Delta t)\right) \end{aligned} \tag{A.23}$$

où $\mathbf{r}_i$ est la position du i-ème atome, $\mathbf{v}_i$ est la vitesse du i-ème atome, $\mathbf{a}_i$ est l'accélération du i-ème atome, et Δt est le pas de temps. L'algorithme de Verlet est un intégrateur symplectique qui conserve l'énergie et la quantité de mouvement, ce qui le rend adapté

aux simulations à long terme des systèmes moléculaires. L'algorithme de Verlet à un pas est exprimé comme suit :

1. Initialisation du pas Δt, durée totale de la simulation $\mathcal{T}$.
2. Initialisation des conditions initiales : $t = 0$, $\mathbf{r}_i(0)$, $\mathbf{v}_i(0)$
3. Définition de la fonction $\mathbf{a}_i$
4. Tant que $t < \mathcal{T}$:
   (a) Calcul de $\mathbf{a}_i(t)$
   (b) Calcul de $\mathbf{r}_i(t + \Delta t)$
   (c) Calcul de $\mathbf{a}_i(t + \Delta t)$
   (d) Calcul de $\mathbf{v}_i(t + \Delta t)$
   (e) $t = t + \Delta t$

Chaque simulation se déroule dans des ensembles statistiques qui définissent les quantités thermodynamiques et leurs relations. Les ensembles les plus couramment utilisés sont les ensembles micro-canonique (NVE), canonique (NVT) et isotherme-isobare (NPT). Dans ces ensembles, les paramètres sont maintenus constants selon la nomenclature suivante : E pour l'énergie, N pour le nombre d'atomes, P pour la pression, T pour la température et V pour le volume. Par exemple, une simulation NVT aura un nombre d'atomes, un volume et une température constants. Des thermostats et barostats appropriés, tels que Nosé-Hoover et Andersen, sont utilisés pour maintenir la pression et la température. Toutefois, cette méthode peut s'avérer inefficace pour explorer les SEP de systèmes complexes avec des barrières énergétiques élevées, car le système peut se retrouver piégé dans des minima locaux. Pour surmonter cette limitation, des méthodes d'échantillonnage améliorées ont été développées pour accélérer l'exploration des SEP et améliorer la précision des résultats.

A.4.2 Méthodes d'échantillonnage avancées

Les méthodes d'échantillonnage avancées sont des techniques conçues pour améliorer l'efficacité de l'exploration des SEP en surmontant les limites des méthodes d'échantillonnage standard. Ces méthodes sont particulièrement avantageuses dans les systèmes présentant des paysages énergétiques accidentés, où la présence de barrières énergétiques élevées peut entraver la convergence des simulations. L'objectif des méthodes d'échantillonnage avancées est d'accélérer l'exploration des SEP, d'améliorer l'échantillonnage des événements rares et d'améliorer la précision des résultats. Ces méthodes utilisent des potentiels biaisés, des schémas de repondération ou des algorithmes avancés pour guider la simulation vers des régions importantes du paysage énergétique. Les approches susmentionnées ne permettent pas l'acquisition directe des propriétés thermodynamiques du système ; cependant, certaines méthodologies ont été développées pour identifier ces propriétés par l'analyse et la repondération des résultats de la simulation.

A.4.2.1 Parallel Tempering (PT)

Le Parallel Tempering (PT), également connu sous le nom de*méthode d'échange de répliques*, est une technique avancée conçue pour améliorer les problèmes d'ergodicité et de convergence dans les simulations MC et DM. Initialement introduite par Swendsen et Wang pour les simulations MC [152], elle a ensuite été adaptée aux simulations de DM par Sugita et Okamoto [150], le PT est largement utilisé dans diverses études [50, 30, 151].

Le PT consiste à effectuer plusieurs simulations simultanées à différentes températures. Cette méthode permet aux systèmes d'échanger des configurations à intervalles réguliers, ce qui favorise l'exploration des surfaces d'énergie potentielle en permettant aux systèmes de surmonter des barrières énergétiques élevées qui, autrement, entraveraient la convergence des simulations. Ces échanges sont régis par un critère de Metropolis-Hastings soigneusement conçu, qui garantit que l'équilibre thermodynamique n'est pas brisé. Le critère d'acceptation d'un échange entre les répliques i et j avec des températures T_i et T_j est donné par :

$$p = \min\left(1, \exp\left(-\Delta E\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_j}\right)\right)\right), \tag{A.24}$$

où $\Delta E = (E_j - E_i)$ et E_i et E_j sont les énergies des répliques i et j, respectivement.

Le PT s'avère particulièrement efficace dans les systèmes aux paysages énergétiques accidentés, où de nombreux minima locaux sont séparés par des barrières élevées. En permettant aux répliques à basse température d'échanger des informations avec les répliques à température plus élevée, le PT facilite le franchissement de barrières énergétiques qui seraient insurmontables aux seules températures plus basses. Ce mécanisme améliore considérablement la capacité des simulations à trouver le minimum global et à échantillonner avec précision les SEP.

A.4.2.2 Umbrella Sampling (US)

L'Umbrella Sampling (US) est une technique de calcul sophistiquée mise au point pour calculer le profil d'énergie libre le long d'une coordonnée de réaction spécifiée. Introduite par Torrie and Valleau [156], cette méthode améliore la capacité à explorer efficacement les SEP, en particulier dans les régions qu'il est généralement difficile d'échantillonner en raison de barrières énergétiques élevées ou de faibles probabilités d'occurrence.

Dans l'Umbrella Sampling, l'énergie potentielle du système est délibérément biaisée le long de la coordonnée de réaction. Ce biais est obtenu par l'introduction d'un potentiel supplémentaire, appelé *umbrella potential*, qui est conçu pour rendre les états moins probables plus accessibles. En modifiant le paysage des SEP, l'US permet un échantillonnage plus approfondi dans les régions d'intérêt, telles que les états de transition ou les états intermédiaires dans une réaction chimique.

Le processus consiste à effectuer une série de simulations, chacune avec un potentiel biaisé légèrement différent appliqué à un segment particulier du chemin de réaction. Les données recueillies lors de ces simulations sont ensuite intégrées à l'aide de techniques

telles que la Weighted Histogram Analysis Method (WHAM) [101] afin de reconstruire le profil d'énergie libre non biaisé.

Cette méthode est largement utilisée en chimie informatique pour étudier les changements d'énergie libre dans les réactions chimiques, prédire les propriétés moléculaires et comprendre les voies biochimiques complexes.

A.4.2.3 Metadynamics

La Metadynamics est une méthode de calcul puissante conçue pour améliorer l'exploration des SEP et faciliter le calcul des profils d'énergie libre. Introduit par Laio and Parrinello [102], la Metadynamics utilise un potentiel de biais dépendant de l'historique d'exploration pour éviter que le système ne soit piégé dans des minima locaux, un défi courant dans les simulations de dynamique moléculaire.

Le mécanisme central de la Metadynamics [16] implique l'ajout périodique de potentiels gaussiens à la position de l'état actuel du système dans une coordonnée de réaction sélectionnée. Cette stratégie rend défavorable de revisiter les états précédemment échantillonnés en créant une mémoire répulsive de ces états. Chaque potentiel gaussien est caractérisé par sa largeur et sa hauteur, qui sont essentielles pour garantir une exploration adéquate des SEP sans perdre en résolution des caractéristiques importantes.

Au fur et à mesure que la simulation progresse, ces potentiels gaussiens s'accumulent, créant un biais qui pousse le système à explorer de nouvelles régions. La tendance du système à revenir sur certains états diminue, ce qui permet une exploration complète des SEP.

La Metadynamics est devenue une technique largement utilisée en chimie informatique pour étudier les réactions chimiques complexes et prédire les propriétés moléculaires. Elle est particulièrement utile pour cartographier les paysages d'énergie libre des systèmes moléculaires et explorer les états de transition.

A.4.3 Méthodes d'optimisation globale

Les méthodes d'optimisation globale sont des techniques employées pour échantillonner efficacement les SEP et identifier les états les plus stables d'un système chimique. Ces méthodes ne sont pas conçues pour obtenir les propriétés thermodynamiques du système ; elles sont plutôt employées pour converger efficacement vers les bassins de basse énergie des SEP. Ces méthodes ne sont pas seulement applicables en chimie pour optimiser les structures moléculaires, mais aussi dans divers domaines tels que la physique, l'économie et la recherche opérationnelle, où il est essentiel de naviguer dans des fonctions complexes pour trouver des solutions optimales.

A.4.3.1 Basin Hopping (BH)

Le Basin Hopping (BH) proposé par Li and Scheraga [109] et Wales and Doye [162] est une méthode MMC qui incorpore une étape d'optimisation locale. Plus précisément, le BH (voir Fig. A.3) génère une séquence de configurations en effectuant des étapes d'optimisation locale suivies de perturbations aléatoires du système. L'étape

d'optimisation locale minimise l'énergie du système en ajustant les positions atomiques pour atteindre un minimum local sur la SEP. Le BH s'appuie sur un critère de Metropolis pour accepter ou rejeter une configuration obtenue après l'étape d'optimisation locale. Les perturbations aléatoires introduisent du bruit dans le système, ce qui lui permet de s'échapper des minima locaux et d'explorer différentes régions de la SEP. La méthode BH est efficace pour explorer les SEP, identifier les configurations les plus stables de la molécule.

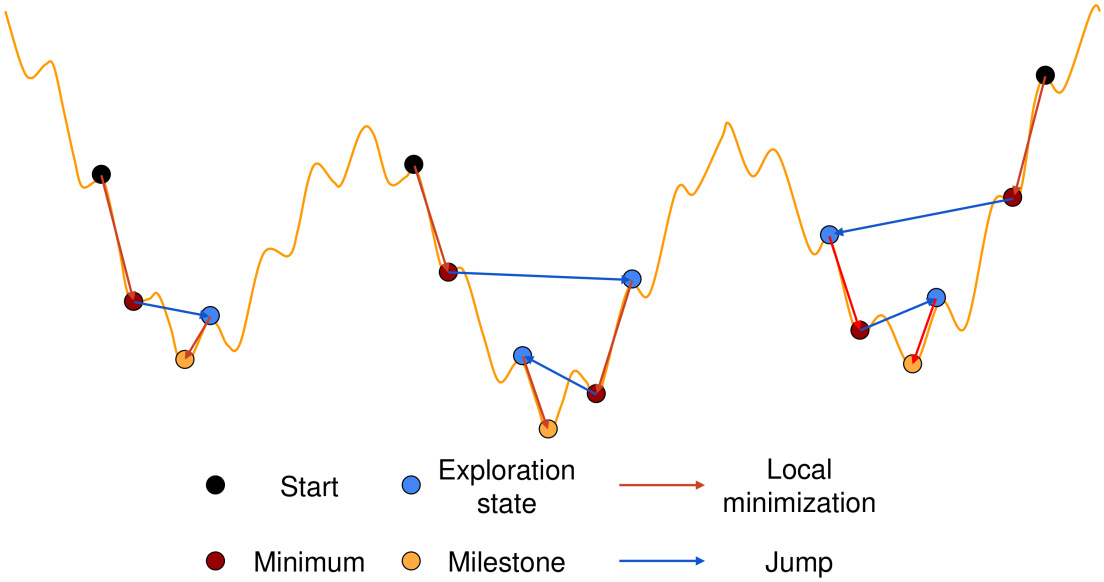


Figure A.3: Méthode de Basin Hopping pour explorer les SEP.

A.4.3.2 Recuit Simulé (RS)

Le Recuit Simulé (RS) est une technique d'optimisation stochastique inspirée du processus de recuit en métallurgie, où les matériaux sont chauffés puis progressivement refroidis pour minimiser leurs défauts et augmenter leur ductilité. Cette méthode, conceptualisée par Kirkpatrick, Gelatt, and Vecchi [97], est conçue pour trouver le minimum global d'une fonction sur un large espace de recherche, ce qui la rend idéale pour les problèmes d'optimisation complexes tels que l'optimisation de la structure moléculaire.

L'algorithme RS [157] démarre avec une température initiale élevée pour permettre une exploration approfondie des SEP. Cette phase aide le système à passer les barrières et à s'échapper des minima locaux au début du processus d'optimisation. À mesure que la température diminue, l'algorithme réduit l'échelle d'exploration, en affinant la solution à mesure qu'elle s'approche des états d'énergie inférieurs. La réduction de la température doit être soigneusement contrôlée par un programme de refroidissement, qui influence de manière critique l'équilibre entre l'exploration et l'exploitation.

La clé de la méthode RS est l'acceptation de nouveaux états au cours du processus de recherche, qui est régi par le critère de Metropolis (Eqn. A.21). Ce critère permet à

l'algorithme d'accepter non seulement les mouvements qui réduisent l'énergie, mais aussi certains qui l'augmentent, évitant ainsi le piège des minima locaux dans les premières étapes.

Le processus se poursuit en modifiant cycliquement la configuration du système et en abaissant progressivement la température jusqu'à ce qu'une température minimale de refroidissement soit atteinte ou que d'autres critères d'arrêt soient satisfaits.

A.4.3.3 Algorithmes Génétiques (AG)

Les algorithmes génétiques (AG) sont une classe de méthodes d'optimisation stochastiques qui imitent le processus de sélection naturelle et d'évolution, tel qu'il a été décrit par Darwin. Cette approche a été formalisée par Holland [87] et est particulièrement utile en chimie informatique pour optimiser les structures moléculaires. Les AG fonctionnent en générant une population variée de solutions candidates, chacune représentant une configuration possible de la molécule étudiée.

Le point central de la méthode AG réside dans son processus itératif où la population évolue sur plusieurs générations vers une solution optimale. Cette évolution est pilotée par des opérateurs génétiques : la sélection, le croisement et la mutation. La sélection imite les pressions naturelles de survie en privilégiant les individus ayant une meilleure condition physique, ce qui leur permet de transmettre leurs gènes à la génération suivante. Le croisement, ou recombinaison, est un processus par lequel des paires d'individus échangent des segments de leur matériel génétique pour produire de nouveaux variants, en combinant les caractéristiques bénéfiques des deux parents. La mutation introduit des changements aléatoires dans les gènes individuels, fournissant de nouvelles variations génétiques et aidant la population à éviter les minima locaux en explorant de nouvelles zones de l'espace de solution.

A chaque itération, l'algorithme évalue l'aptitude de tous les individus de la population, généralement mesurée en fonction de leur capacité à résoudre le problème d'optimisation ou à répondre aux critères souhaités. Les opérateurs génétiques sont ensuite appliqués pour créer une nouvelle génération, idéalement avec une aptitude moyenne plus élevée que la précédente. Au fil des générations, la population converge vers une solution optimale, imitant ainsi le processus évolutif d'adaptation.

La flexibilité et l'efficacité des algorithmes génétiques les rendent particulièrement adaptés aux problèmes pour lesquels les techniques d'optimisation traditionnelles peinent à donner de bons résultats en raison de la complexité du paysage impliquant de nombreux optima locaux.

A.4.4 Méthodes inspirées par la robotique

La planification des mouvements est un problème fondamental en robotique qui implique l'identification d'une trajectoire sans collision pour un robot qui se déplace d'une configuration initiale à une configuration cible. Ce problème peut également s'appliquer à un bras robotique doté d'un nombre limité d'articulations, qui doit par exemple ramasser un objet. Divers algorithmes ont été développés pour atteindre ces objectifs.

Ces algorithmes ont évolué au-delà de leur champ d'application initial et ont été utilisés dans divers domaines, notamment la fabrication industrielle, l'animation par ordinateur et la bioinformatique. Par exemple, ils ont été utilisés dans le contexte du repliement des protéines et de l'optimisation des structures moléculaires [26, 107, 124, 22]. En chimie computationnelle, ces algorithmes ont été employés pour explorer efficacement les SEP. Contrairement aux méthodes d'optimisation globale, ces algorithmes sont d'abord conçus pour explorer efficacement un espace à haute dimension, mais ne visent pas directement à trouver le minimum global. Néanmoins, la dernière méthode présentée, appelée Iterative Global Exploration and Local Optimization (IGLOO), est une méthode qui combine à la fois l'algorithme de planification de mouvement et l'optimisation locale. Certains de ces algorithmes sont capables d'identifier des états de faible énergie, ainsi que de les relier entre eux afin d'identifier les chemins de transition entre ces derniers. Cet aspect sera développé plus en détail au chapitre 6.

A.4.4.1 Probabilistic Roadmap (PRM)

La méthode Probabilistic Roadmap (PRM), introduite par Kavraki et al. [96], est utilisée pour résoudre les problèmes de planification de mouvements en haute dimension. La PRM fonctionne par échantillonnage itératif d'une configuration de l'espace de configuration. Si la configuration est exempte de collision, elle est ajoutée au graphe en tant que nœud. Le nouveau nœud est connecté au graphe en trouvant ses voisins les plus proches. Si le chemin entre le nouveau nœud et les voisins les plus proches est sans collision, il est ajouté au graphe sous la forme d'une ligne droite. Ces étapes sont répétées jusqu'à ce qu'un critère d'arrêt soit atteint. Le graphe peut alors être utilisée pour trouver un chemin entre les nœuds à l'aide d'algorithmes de recherche de graphes tels que Dijkstra ou A* [22]. Des extensions de PRM impliquant des calculs d'énergie, présentées ensuite, ont été proposées pour explorer les SEP.

A.4.4.2 Probabilistic Conformational Roadmaps (PCR)

La méthode Probabilistic Conformational Roadmaps (PCR) proposée par Singh, Latombe, and Brutlag [145] est une méthode basée sur le PRM qui génère un graphe en acceptant ou en rejetant de nouveaux nœuds à l'aide d'une fonction de probabilité favorisant les conformations à faible énergie. La probabilité d'acceptation est évaluée comme suit :

$$P(\text{accept}, q) = \begin{cases} 1 & \text{if } E_q < E_{\min} \\ \dfrac{E_{\max} - E_q}{E_{\max} - E_{\min}} & \text{if } E_{\min} \leq E_q \leq E_{\max} \\ 0 & \text{if } E_q > E_{\max} \end{cases} \tag{A.25}$$

où E_q est l'énergie de la conformation q, $E_{\min}$ et $E_{\max}$ sont des valeurs seuils fixées pour le système. A chaque arête e_{ij} est associé un poids représentant la probabilité de la transition entre les conformations connectées. Une série de conformations intermédiaires est générée le long du chemin $\{q_i = c_0, c_1, .., c_n = q_j\}$ reliant les deux nœuds q_i et q_j

(le nombre de conformations intermédiaires est un paramètre). Le poids de l'arête est calculé comme suit :

$$\begin{aligned} w(e_{ij}) &= -\sum_{i=0}^{n-1} \log(P_i) \\ P_i &= \frac{e^{-\frac{(E_{i+1}-E_i)}{KT}}}{e^{-\frac{(E_{i+1}-E_i)}{KT}} + e^{-\frac{(E_{i-1}-E_i)}{KT}}} \end{aligned} \tag{A.26}$$

où E_i est l'énergie de la conformation c_i, n le nombre d'images, K la constante de Boltzmann et T la température. La PCR a été appliqué pour trouver les mouvements énergétiquement favorables de biomolécules [8].

A.4.4.3 Stochastic Roadmap Simulation (SRS)

La méthode Stochastic Roadmap Simulation (SRS) [10, 11, 35, 36, 9] est une amélioration de la PCR. La différence réside dans la fonction de probabilité, qui est conforme au critère de Metropolis [118]. La fonction de probabilité est évaluée comme suit :

$$P_{ij} = \begin{cases} \frac{1}{n_i} \exp(-\frac{\Delta E_{ij}}{KT}) & \text{if } \Delta E_{ij} > 0 \\ \frac{1}{n_i} & \text{otherwise} \end{cases} \tag{A.27}$$

$$P_{ii} = 1 - \sum_{j \neq i} P_{ij} \tag{A.28}$$

où n_i est le nombre de voisins du nœud q_i, ΔE_{ij} est la différence d'énergie entre les nœuds q_i et q_j, K est la constante de Boltzmann et T est la température. La SRS a été utilisé pour prédire les interactions ligand-protéine [8].

A.4.4.4 Rapidly-exploring Random Tree (RRT)

La méthode Rapidly-exploring Random Tree (RRT), introduite par LaValle [103], est utilisée pour résoudre les problèmes de planification de mouvement en haute dimension. La RRT fonctionne par échantillonnage itératif d'une configuration de l'espace de configuration. Si la configuration est exempte de collision, elle est ajoutée à l'arbre en tant que nœud. Le nouveau nœud est connecté à l'arbre en trouvant son voisin le plus proche. L'arbre est étendu en ajoutant de nouveaux nœuds dans la direction de l'échantillon généré aléatoirement. La principale différence entre cette méthode et la PRM est que le nouveau nœud est relié au voisin le plus proche et non à tous les voisins pour lesquels il existe un chemin sans collision. La méthode RRT sera examinée plus en détail au chapitre 2.

A.4.4.5 Transition-RRT (T-RRT)

La méthode Transition-RRT (T-RRT) proposée par Jaillet, Cortés, and Siméon [92, 93] est une méthode basée sur la RRT avec un test de transition pour favoriser l'exploration des régions de faible énergie des SEP. Le test de transition est basé sur le critère de Metropolis inspiré des méthodes MC et est utilisé pour accepter ou rejeter de nouveaux nœuds en fonction du changement d'énergie et de la température du système. Contraire-

ment à la méthode MC, la température est auto-adaptative pendant l'exploration afin d'ajuster dynamiquement l'exploration des SEP. L'algorithme tient compte de chaque rejet et acceptation du test de transition pour ajuster la température. La méthode T-RRT sera examinée plus en détail au chapitre 6.

A.4.4.6 Iterative Global Exploration and Local Optimization (IGLOO)

La méthode Iterative Global Exploration and Local Optimization (IGLOO) [112] combine l'exploration des SEP avec une optimisation locale. La méthode IGLOO est un algorithme itératif composé de trois étapes principales : une étape d'exploration, une étape d'optimisation locale et une étape de filtrage. L'étape d'exploration est réalisée à l'aide d'une méthode basée sur la RRT pour explorer les SEP. L'étape d'optimisation locale est réalisée à l'aide d'une méthode d'optimisation locale afin de minimiser l'énergie potentielle des états explorés. L'étape de filtrage est utilisée pour supprimer les états redondants et améliorer l'efficacité de l'exploration à l'itération suivante. IGLOO a été appliquée avec succès pour prédire la structure des molécules de disaccharide sur des surfaces métalliques [1, 2]. IGLOO sera discuté plus en détail dans le chapitre 2.

A.5 Résumé des contributions

La thèse contient plusieurs contributions au domaine de la chimie computationnelle.

Un aperçu des différents chapitres est donné ici.

Chapitre 2: Ce chapitre présente le couplage des méthodes IGLOO et DFTB pour l'exploration de l'espace conformationnel des molécules. IGLOO s'inspire de la planification des mouvements en robotique, tandis que DFTB est une méthode de chimie quantique. La partie développement détail l'interface entre les deux logiciels développés dans nos laboratoires. La méthode IGLOO, mise en œuvre dans le logicielle MoMA, est couplée à la méthode DFTB, mise en œuvre dans le code deMonNano. Comme première application, l'approche a été appliquée au dipeptide d'alanine, un petit peptide. L'exploration a permis d'identifier les conformations énergétiquement favorables, démontrant ainsi l'efficacité du couplage dans la réduction des coûts de calcul tout en maintenant une description précise du système chimique.

Chapitre 3: Ce chapitre montre l'exploration des SEP de molécules de la famille des phtalates en utilisant le couplage IGLOODFTB introduit dans le chapitre précédent. Les phtalates sont une famille de composésc chimiques largement utilisés dans les produits de consommation. Il est important de comprendre leur comportement conformationnel étant donné les impacts de ces composés sur l'environnement et la santé. Le chapitre commence par une introduction aux phtalates, soulignant l'importance et la nécessité d'une exploration détaillée du paysage énergétique. Le couplage est initialisé IGLOO avec une multitude d'états initiaux pour assurer une couverture complète et plusieurs exécutions indépendantes sont effectuées pour tenir compte de la nature stochastique

de la méthode. Cette approche a révélé de nombreux bassins de basse énergie et a facilité l'identification de conformations stables dans une gamme variée de molécules de phtalates. Parmi les résultats significatifs, citons l'identification de divers minima conformationnels, qui ont été analysés à l'aide de descripteurs énergétiques et structurels. Les descripteurs susmentionnés ont facilité la compréhension des interactions au sein des molécules de phtalates, y compris les effets des arrangements des chaînes latérales sur la stabilité moléculaire. En outre, le chapitre compare les calculs DFTB avec des calculs DFT pour valider la précision du potentiel sur ces molécules. Les résultats illustrent l'efficacité du couplage IGLOODFTB dans l'exploration des paysages d'énergie potentielle complexes des phtalates, fournissant des indications précieuses sur leur dynamique conformationnelle.

Chapitre 4: Ce chapitre présente un algorithme innovant pour la génération de modèles atomistiques d'hydrocarbures aromatiques de grande taille. L'accent est mis sur l'intégration de techniques de génération basées sur les graphes moléculaires avec des ajouts d'atomes et de fragments, avec un accent particulier sur le maintien de contraintes prédéfinies. L'introduction donne un aperçu de l'importance des hydrocarbures aromatiques dans divers domaines scientifiques, notamment l'astrophysique et les sciences de l'environnement. Elle souligne la nécessité de disposer de modèles précis pour simuler et comprendre leur comportement dans différents environnements. La méthodologie comprend deux composantes principales : le SMILES Generator et le Structure generator. Le SMILES Generator est conçu pour produire une série de SMILES qui respectent les contraintes spécifiées sur les types et les ratios de liaisons et d'atomes. Pour ce faire, il utilise un processus précis qui comprend la sélection des types de fragments, la sélection des atomes dans le graphe moléculaire et l'ajout de fragments afin de construire la structure moléculaire de manière incrémentale. Ensuite, l'algorithme du Structure Generator permet de générer des structures tridimensionnelles. Ce processus implique la génération de structures initiales non optimisées à partir des SMILES. Ces structures sont ensuite optimisées par une série d'étapes visant à minimiser l'auto-collision et à garantir la validité de la structure.

Chapitre 5: Ce chapitre met en avant l'application des algorithmes précédemment développés sur des sous-structures de polymères de carbone amorphe hydrogéné. Le chapitre commence par un aperçu des connaissances actuelles sur les sous-structures des polymères de carbone amorphe hydrogéné dans le milieu interstellaire (ISM), en mettant l'accent sur leur détection par les bandes d'absorption infrarouge et sur leur rôle central dans divers processus physico-chimiques dans l'espace. La section méthodologie décrit la production et l'analyse des polymères de carbone amorphe hydrogéné. Ensuite, des paramètres d'évaluation des structures générées sont définis, en se concentrant sur des descripteurs géométriques et électroniques. Un descripteur géométrique tel que les paramètres de Hill-Wheeler, qui évaluent la déformation de la forme par rapport à une sphère parfaite, est défini, et des descripteurs électroniques tels que l'écart HOMO-LUMO et l'énergie de London sont calculés, donnant un aperçu des propriétés électron-

iques des structures. L'évaluation de ces descripteurs a montré des variations significatives dans les formes et les propriétés électroniques des sous-structures de polymères de carbone amorphe hydrogéné.

Chapitre 6: Ce chapitre traite des chemins de transition entre les conformations de faible énergie dans les systèmes moléculaires, en présentant diverses techniques de calcul pour identifier ces chemins. Le chapitre commence par discuter de la base théorique défini par la Transition State Theory, en soulignant l'importance de l'identification du chemin d'énergie minimale (MEP) qui représente la voie la plus favorable pour une réaction. Diverses méthodes de calcul sont examinées pour identifier et analyser ces chemins de transition. Celles-ci incluent la Dimer method pour localiser les points de selle sur la surface d'énergie potentielle, et des méthodologies avancées telles que la méthode Nudged Elastic Band (NEB), qui affine le chemin pour minimiser l'énergie le long de la coordonnée de réaction. Une méthodologie préliminaire pour explorer la diversité des chemins de transition entre les conformations de faible énergie est présentée. L'exploration de ces chemins est réalisée à l'aide de l'algorithme stochastique T-RRT, qui génère de nombreux chemins. Une mesure de similarité est ensuite appliquée pour différencier ces chemins, et une méthode de regroupement est ensuite utilisée pour identifier les chemins identiques. Ensuite, un chemin représentatif de chaque groupe est sélectionné pour être localement optimisé. La méthodologie est appliquée sur la dipeptide d'alanine et des résultats préliminaires sont présentés.

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Université de Toulouse

Titre : Couplage de modèles de chimie quantique et d'algorithmes haute performance pour l'exploration globale du paysage énergétique de systèmes atomiques et moléculaires

Mots clés : Modélisation atomique et moléculaire, Surfaces d'énergie potentielle, Exploration du paysage conformationnel, Algorithmes inspirés de la robotique, Optimisation globale, Identification de chemins de transition

Résumé : L'objectif principal de cette thèse est de développer des méthodes efficaces pour caractériser les conformations des molécules à un niveau quantique. Différentes méthodes dédiées au calcul de l'énergie potentielle d'une molécule sont examinées, ainsi que les schémas d'exploration globale des surfaces d'énergie potentielle (SEP) les plus populaires sont présentés. Une contribution clé de cette thèse est le couplage de la méthode IGLOO (Iterative Global exploration and LOcal Optimization), inspirée de la robotique, mise en œuvre dans le logiciel MoMA, avec le potentiel basé sur la "Density-Functional based Tight-Binding" (DFTB), implémenté dans le logiciel deMonNano. IGLOO intègre l'algorithme de planification de mouvement "Rapidly-exploring Random Trees" (RRT) avec des optimisations locales de l'énergie et un filtrage des structures. Une preuve de concept a été réalisée par l'identification des conformations de basse énergie de la molécule de d'alanine dipeptide.
Le couplage IGLOO/DFTB a été appliqué à la cartographie des SEP de trois molécules de taille proche de la famille des phtalates (dibutyl phtalate DBP, benzyl butyl phtalate BBP et di-2-éthylhexyl phtalate DEHP), donnant un aperçu détaillé de leurs différents paysages conformationnels. Divers descripteurs géométriques ont été utilisés pour analyser leurs relations structure-énergie. Les interactions de Coulomb, l'encombrement stérique et les interactions dispersives sont à l'origine des propriétés géométriques et une forte corrélation a été mise en évidence entre les deux angles diédraux décrivant l'orientation des chaînes latérales des molécules de phtalate.
En complément, un algorithme innovant pour la génération à grande échelle de molécules, incluant une variété de conformations, est présenté. Il combine la génération de graphes de molécules avec des techniques d'ajout d'atomes ou de fragments. Il est appliqué pour fournir une vaste base de données de structures 3D de molécules de carbone amorphe hydrogéné (a-CH). L'analyse de la base de données générée dans cette étude permet de comprendre la relation entre les descripteurs géométriques et électroniques des structures a-C:H. Ces propriétés sont comparées à celles des hydrocarbures aromatiques polycycliques (HAP) compacts et des chaînes linéaires, qui représentent des cas limites.
Enfin, une revue des méthodes visant à identifier les points de selle et les chemins de transition entre les conformations de faible énergie sur la SEP est présentée. Une première étape pour l'identification des chemins de transition entre les conformations de faible énergie à l'aide d'un algorithme de planification de mouvement, connu sous le nom de Transition-based RRT (T-RRT), est présentée. Une mesure de similarité, désignée sous le nom de Symmetrized Segment-Path Distance (SSPD), est utilisée pour comparer les trajectoires générées. Ensuite, une technique de regroupement, à savoir Analyse de regroupement hiérarchique (HCA), est employée pour regrouper les trajectoires afin d'identifier les classes de chemin donnant la dynamique des changements de conformation. La méthodologie a été appliquée avec succès à l'identification de chemins à faible énergie entre deux minima de la SEP de l'alanine dipeptide.
Dans l'ensemble, les travaux présentent des avancées significatives dans l'exploration de SEP de molécules complexes au niveau quantique, y compris (i) le couplage IGLOO/DFTB (ii) un nouvel algorithme pour la génération de structures 3D de molécules à grande échelle et (iii) un schéma original permettant l'identification de multiples chemins de transition. Des corrélations entre les propriétés structurelles, énergétiques et électroniques ont été mises en évidence pour les molécules polluantes de la famille des phtalates ainsi que pour les a-CH ayant une importance du point de vue astrophysique. Ces contributions ouvrent la voie à de futures recherches visant à étendre ces méthodes à des systèmes plus grands et plus complexes.

Title: Coupling of quantum chemistry models and high-performance algorithms for the global exploration of the energy landscape of atomic and molecular systems

Key words: Atomic and molecular modeling, Potential energy surfaces, Conformational exploration, Robotics-inspired algorithms, Global optimization, Transition path identification

Abstract: The primary aim of this thesis is to develop efficient methods for characterizing molecular conformations at a quantum level. Various methods devoted to the computation of molecular potential energy are reviewed, as well as the most popular potential energy surfaces (PES) global exploration schemes. In this context, a key contribution of this thesis is the coupling of the robotics-inspired Iterative Global exploration and LOcal Optimization (IGLOO) method, implemented in the MoMA software, with the quantum Density-Functional based Tight-Binding (DFTB) potential, implemented in the deMonNano software. The IGLOO algorithm integrates the motion planning Rapidly-exploring Random Trees (RRT) algorithm with local optimization and structural filtering. A proof of concept has been done through the identification of low-energy conformations of the alanine dipeptide.
The IGLOO/DFTB coupling has been applied to the mapping of the PES of three close-sized molecules of the phthalate family (dibutyl phthalate DBP, benzyl butyl phthalate BBP and di-2-ethylhexyl phthalate DEHP), providing detailed insights into their different conformational landscapes. Various geometrical descriptors have been used to analyze their structure-energy relationships. Coulomb interactions, steric hindrance, and dispersive interactions have been found to drive the geometric properties and a strong correlation has been evidenced between the two dihedral angles describing the side-chains orientation of the phthalate molecules. The results demonstrate the method's capability to identify low-energy minima without prior knowledge of the PES.
Furthermore, an innovative algorithm for the large-scale generation of molecular structures, including a conformational variety, is presented. It combines molecular graph generation with atom or fragment addition techniques. It is applied to provide an extensive database of 3D structures of hydrogenated amorphous carbon (a-CH) molecules. The analysis of the database generated in this study provides a comprehensive understanding of the relationship between the geometrical and electronic descriptors of a-C:H structures. These properties are compared with those of compact Polycyclic Aromatic Hydrocarbons and linear chains, representing limit cases.
Finally, a review is given on methods aiming at identifying saddle points and transition paths between low-energy conformations on the PES. A first step toward the identification of transition paths between low-energy conformations using a motion planning algorithm, known as Transition-based Rapidly-exploring Random Trees (T-RRT), is presented. A similarity measure, designated as the Symmetrized Segment-Path Distance (SSPD), is used to compare the generated trajectories. Subsequently, a clustering technique, namely the Hierarchical Clustering Analysis (HCA), is employed to group similar trajectories in order to identify the common pathways, thereby providing valuable insights into the dynamics of conformational changes. The methodology has been successfully applied to the identification of low-energy paths between two minima of the alanine dipeptide PES.
Overall, the research presents significant advancements in the exploration of complex molecular PES at a quantum level including (i) the IGLOO/DFTB coupling (ii) a novel algorithm for 3D structure generation of large-scale molecules and (iii) an original scheme allowing for the identification of multiple transition paths. Correlations between the structural, energetic and electronic properties have been evidenced for the polluting phthalate molecules and astrophysically relevant hydrogenated amorphous carbon (a-CH) molecules. These contributions pave the way for future research, aiming to extend these methods to larger and more complex systems.