SCUOLA
NORMALE
SUPERIORE

Classe di Scienze

Corso di perfezionamento in
Metodi e Modelli per le Scienze Molecolari

XXXIV ciclo

Continuous Perception for Immersive Interaction and Computation in Molecular Sciences

Settore Scientifico Disciplinare CHIM/02

Candidato
dr. Federico Lazzari

Relatore

Prof. Vincenzo Barone

Anno accademico 2022/2023

Acknowledgments

This thesis is the product of many years of hard work and there are so many people that helped me and gave me support during my PhD studies. Trying to acknowledge all of them is really hard, and I will forget to mention someone for sure. I would like to start by acknowledging my family. During these years we went through many changes, not only related to the global pandemic situation but also personal changes. I would like to thank my father, Cesare Lazzari, that inspired me since I was little to pursue scientific studies. I hope he would be proud of me and my personal progresses. He truly believed in teaching and researching, he was a man of science but also a man of faith. Without his prompts, I wouldn't probably be here writing this thesis. Next, I would like to thank my mother, Loredana Bruno. These years have been really stressful and while we often clash with each other, she has always been very supportive and present even in the hardest moments. The material and emotional support she gave me is incommensurable, and I can't thank her enough. My brother, Virgilio Lazzari, and my sister, Annamaria Lazzari, have been a good supporting team and my sister in particular has given a lot of suggestions when it comes to my terrible English. I am grateful to all of my aunts and cousins for all those fun moments they helped me remembering there is so much more than work in life. I would also like to thank Daniele Giannecchini, he can handle my terrible personality and my sharp sense of humor.

Next, I would like to thank Prof. Vincenzo Barone for his scientific support and for the opportunity he gave me to continuously work in a prestigious research institution for almost ten years. I can't thank enough all of my colleagues and co-workers that made each day at work a fun day; Luigi Crisci, Marco Fusé, Monica Sanna, and Niccolò Albertini, just to name a few. Thanks to Giordano Mancini and Andrea Salvadori, they truly taught me the fundamentals of coding and they helped me navigate the world of academic research. Finally, I need to say a big "thank you" to Marco Mendolicchio and Silvia Di Grande, they are the best teammates you can hope for. Marco has always been very supportive and gave me a lot of help in writing this thesis. Silvia is always there whenever I need scientific or emotional help, she is an incredibly hard worker, a very smart and intelligent person but also extremely funny to joke with. I can't even count all of the moments I have felt as if everything was falling apart, and she always helped me to go back into focus pragmatically facing each difficulty separately. As she would say: *Daje*.

Acronyms

QM Quantum Mechanics. 6, 8, 9, 10, 13, 15, 17, 20, 31, 32, 60, 67, 77, 96, 101, 105, 134, 147, 152, 157, 158, 159, 160, 161, 163, 177

MCSCF Multi Configurational Self-Consistent Field. 6, 22, 23

CI Configuration Interaction. 6, 23, 24, 25, 30

CC Coupled-Cluster. 6, 24, 25

DFT Density Functional Theory. 6, 20, 25, 28, 29, 30, 37, 164, 173

FF Force Field. 6, 32, 33, 34, 35, 37, 89, 95, 96, 105, 129, 147, 149, 150, 151, 157, 158, 166

SVM Supporting Vector Machines. 6, 42, 45, 46, 47, 57

AI Artificial Intelligence. 9, 13, 14, 169, 178

ML Machine-Learning. 9, 10, 13, 14, 15, 20, 35, 36, 37, 38, 39, 40, 41, 42, 47, 53, 56, 57, 58, 59, 60, 61, 89, 109, 129, 147, 157, 163, 168, 169, 177, 178

NLP Natural-Language Processing. 9, 14

MP Molecular Perception. 9, 10, 15, 59, 60, 61, 62, 63, 71, 84, 96, 101, 129, 158, 177

TB Tight-Binding. 9, 15, 63, 64, 66, 67, 68, 72, 73, 74, 76

SE Semi-Experimental. 10, 133, 134, 136, 137, 139, 140, 141, 142, 144, 145, 147, 168, 178

PCS Pisa Composite Scheme. 10, 178

IVR Immersive Virtual Reality. 10, 15, 80, 172, 174, 176

TMA Templating Molecules Approach. 10, 129, 133, 135, 136, 137, 139, 140, 142, 144, 145, 147, 157, 171, 177, 178

LRA Linear Regression Analysis. 10, 133, 134, 135, 136, 137, 138, 139, 140, 144

MM Molecular Mechanics. 10, 15, 32, 37, 60, 96, 147, 158, 160, 161

CPU Central Processing Unit. 13

GPU Graphics Processing Unit. 13

GANN Generative Adversary Neural-Networks. 14

AR Augmented Reality. 14, 168, 169

VR Virtual Reality. 14, 129, 163, 168, 169, 170, 172, 174

QC Quantum Chemistry. 17, 37, 60, 96, 109, 132

MO Molecular Orbitals. 21, 22, 30

LCAO Linear Combination of Atomic Orbitals. 21, 22

HF Hartree-Fock. 21, 22, 23, 24, 29, 30

FCI Full Configuration Interaction. 23, 24

CASSCF Complete Active Space Self Consistent Field. 23

MP2 Second Order Møller-Plesset Perturbation Theory. 24, 29, 30, 31, 134, 135, 138

CCSD Coupled-Cluster Single and Double Excitations. 25, 30

KS Kohn-Sham. 27, 28, 29

LDA Local-Density Approximation. 28

LSDA Local-Spin Density Approximation. 28, 29

GGA Gradient-Corrected Approximation. 28, 29

B3LYP Becke, 3-parameter, Lee–Yang–Parr. 29, 65, 66, 67, 74, 134, 137, 139, 140, 141, 142, 143, 147, 151, 152, 153, 154, 155, 156, 157, 166

B2PLYP Becke, 2 order perturbation, Lee–Yang–Parr. 29, 134, 144, 173

LYP Lee–Yang–Parr. 29

revDSD rev-DSD-PBEP86-D3(BJ) [1–3]. 29, 134, 135, 136, 137, 139, 142, 144, 147, 158, 166, 178

CM5 Charge Model 5. 31, 32, 74, 75, 76, 77

ESP Electrostatic Potential. 31

RESP Restrained Electrostatic Potential. 31

EVB Empirical Valence Bond. 32

vdW van der Waals. 34, 35, 70, 73, 96, 97, 98, 101, 159

MD Molecular Dynamics. 34, 158

UFF Universal Force Field. 34

LJ Lennard-Jones. 34

FS Feature Space. 35, 40, 41, 42, 43, 46, 52, 53, 57, 58, 59, 62, 89, 91, 94, 96, 129, 131, 133, 136, 3, 145, 146, 147, 158, 163, 177

MSE Mean Square Error. 39

PAC Probably Approximately Correct. 39

QSAR Quantitative structure-activity relationship. 41

PCA Principal Component Analysis. 41, 42, 52, 131, 145, 146, 172

LR Linear Regression. 42, 46

NN Neural Networks. 42, 45, 54, 56, 57, 58, 147, 178

SGD Stochastic Gradient-Descent. 44, 57, 58

NC Naive Classifier. 45

DT Decision Tree. 47, 130, 131, 147, 178

RF Random Forest. 49

PAM Partition Around Medoids. 50

CH Calinski-Harabasz. 51, 52

DBSCAN Density-Based Spatial Clustering of Applications with Noise. 52

GA Genetic Algorithm. 54, 55, 56, 164, 178

SBX Simulated Binary Crossover. 55

PDB Protein Data Bank. 62, 84

SMILES Simplified Molecular-Input Line-Entry System. 62

LP Lone Pair. 62, 63, 67, 68, 69, 70, 71, 80

BFS Breadth-First Search. 64, 74, 81, 82

aug-cc-pVDZ Augmented Correlation Consistent Polarized Double Zeta basis set [4]. 65, 66, 74, 151, 152, 153, 154, 155, 156

jun-cc-pVDZ June Correlation Consistent Polarized Double Zeta basis set [5]. 67, 140, 141, 142, 143

RMSE Root-Mean-Square Error. 74, 76

CT Charge Transfer. 77, 78

TIP3P Transferable Intermolecular Potential with 3 Points. 78

IUPAC International Union of Pure and Applied Chemistry. 81

PIC Primitive Internal Coordinate. 105, 106

GIC Generalized Internal Coordinates. 106, 145

PES Potential Energy Surface. 107, 109, 129, 163, 164, 172, 173, 174, 175, 176

INC Internal-based Normal Coordinates. 108

ZPE Zero Point Energy. 110

MW Microwave. 110

GAFF General AMBER Force Field. 129, 130

LPCS Low-cost Pisa Composite Scheme. 133, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 147, 168, 178

VPT2 Second-Order Vibrational Perturbation Theory [6–8]. 133, 134, 139

jul-cc-pVDZ July Correlation Consistent Polarized Double Zeta basis set [5]. 134, 139, 144

jun-cc-pVTZ June Correlation Consistent Polarized Triple Zeta basis set [5]. 134, 136, 137, 158, 166

cc-pVTZ-F12 Peterson Correlation Consisted Polarized Triple Zeta basis set [9]. 135

CV Core Valence. 135

aug-cc-pVTZ Augmented Correlation Consistent Polarized Triple Zeta basis set [4]. 137, 173

RMSD Root-Mean-Square Deviation. 162, 166

GUI Graphical User Interface. 163, 164

LHS Latin Hypercube Sampling. 165

IM-EA Island Model - Evolutionary Algorithm. 166

CPCM Conductor-Like Polarizable Continuum Model. 167

NFT Non-fungible token. 168

TCP Transmission Control Protocol. 168

IP Internet Protocol. 168

MR Mixed Reality. 169, 171, 172

Contents

Abstract

Chemistry aims to understand the structure and reactions of molecules, which involve phenomena occurring at microscopic scales. However, scientists perceive the world at macroscopic scales, making it difficult to study complex molecular objects. Graphical representations, such as structural formulas, were developed to bridge this gap and aid in understanding. The advent of Quantum Mechanics further increased the complexity of the representation of microscopic objects. This dichotomy between conceptual representation and predictive quantification forms the foundation of Chemistry, now further explored with the rise of Artificial Intelligence. Recent advancements in computational sciences, increased computational power, and developments in Machine-Learning (ML) raise questions about the traditional scientific method. Computational scientists, who have relied on approximations based on fundamental rules, now face the possibility of accurately simulating nature without strictly adhering to its laws. This shift challenges the association between progress in understanding a phenomenon and the ability to predict it. Deep learning models can not only make predictions but also create new data. While these techniques find applications in fields like Natural-Language Processing, they suffer from limitations and lack true intelligence or awareness of physical laws. The thesis aims to create mathematical descriptors for atom types, bond types, and angle types in ML procedures, ensuring the retention of their chemical meaning. The goal is to make quantitative predictions while interpreting changes in descriptors as chemical changes. To achieve this, the thesis develops a software called Proxima for Molecular Perception, which automatically perceives features from molecules. Proxima treats strongly coupled electrons as covalent bonds and lone pairs, while delocalized electrons are modeled using a Tight-Binding model. The resulting Molecular Graph captures the weak interactions between these units. Overall, this thesis explores the intersection of computational chemistry and Machine-Learning to enhance our understanding and predictive capabilities in the field of Chemistry by building the so-called Virtual Laboratory, a virtual environment with automatic access to structural databases to test chemical ideas on the fly (pre-processing) and explore the output of computational software (post-processing).

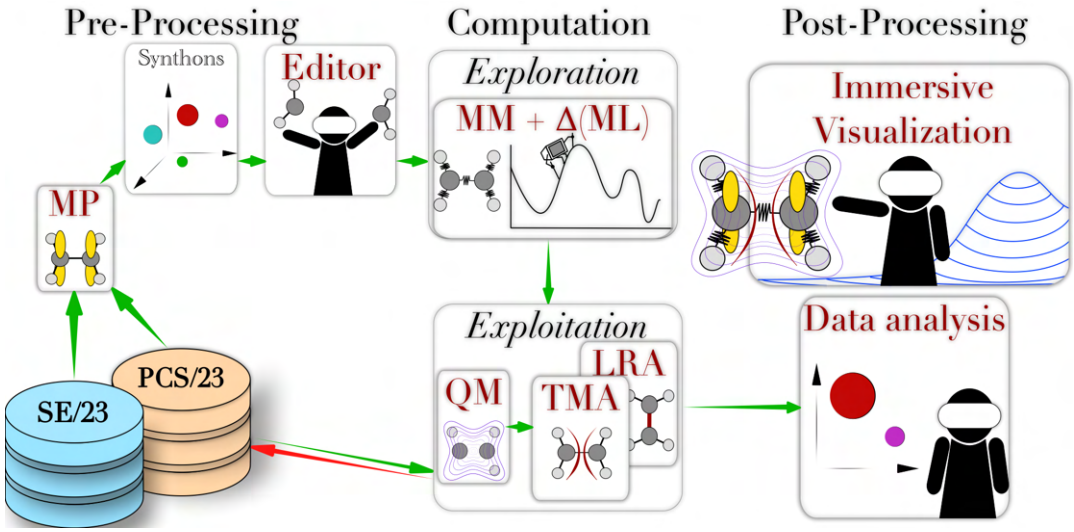


Figure 1: The flowchart of the Virtual Laboratory approach that underlies the work of this thesis. The pre-processing phase requires data to be retrieved from our SE/23 and PCS/23 databases. Thanks to Molecular Perception, the synthons are capable to detect available fragments and through a IVR editor it is possible to assemble them together in bigger molecules. Once the pre-processing phase is complete, the computation step allows to explore and exploit the chemical space by use of QM, TMA, LRA, MM, and ML techniques. In case the TMA procedure detects new fragments these can automatically be added to the PCS/23 database (the red arrow). The post-processing phase employs IVR tools to visualize data and explore the chemical space.

Publications

- Marta Martino, Andrea Salvadori, Federico Lazzari, Lorenzo Paoloni, Surajit Nandi, Giordano Mancini, Vincenzo Barone and Sergio Rampino. *Chemical promenades: Exploring potential-energy surfaces with immersive virtual reality*, Volume 41, Number 13, Pages 1310–1323, **2020**, J. Comp. Chem.

- Giordano Mancini, Marco Fusè, Federico Lazzari, Balasubramanian Chandramouli, and Vincenzo Barone. *Unsupervised search of low-lying conformers with spectroscopic accuracy: A two-step algorithm rooted into the island model evolutionary algorithm*, Volume 153, Number 12, Pages 124110, **2020**, J. Chem. Phys.

- Marta Martino, Federico Lazzari, Nicola Tasinato and Vincenzo Barone. *Virtual Reality bridge between Chemistry and Cultural Heritage: the "Sala degli Stemmi" Case Study*, Volume 949, Number 1, Pages 12020, **2020**, IOP Conference Series: Materials Science and Engineering.

- Federico Lazzari, Andrea Salvadori, Giordano Mancini and Vincenzo Barone. *Molecular Perception for Visualization and Computation: The Proxima Library*, Volume 60, Number 6, Pages 2668–2672, **2020**, J. Chem. Inf. Model.

- Giordano Mancini, Marco Fusè, Federico Lazzari and Vincenzo Barone. *Fast exploration of potential energy surfaces with a joint venture of quantum chemistry, evolutionary algorithms and unsupervised learning*, Volume 1, Number 6, Pages 790–805, **2022**, Digital Discovery, Published by the Royal Society of Chemistry.

- Emanuele Falbo, Marco Fusè, Federico Lazzari, Giordano Mancini and Vincenzo Barone. *Integration of Quantum Chemistry, Statistical Mechanics, and Artificial Intelligence for Computational Spectroscopy: The UV–Vis Spectrum of TEMPO Radical in Different Solvents*, Volume 18, Number 10, Pages 6203–6216, **2022**, J. Chem. Theory Comput.

- Niccolò Albertini, Jacopo Baldini, Andrea Dal Pino, Federico Lazzari, Stefano Legnaioli and Vincenzo Barone. *PROTEUS: an immersive tool for exploring the world of cultural heritage across space and time scales*, Volume 10, Number 1, Pages 71, **2022**, Heritage Science.

- Vincenzo Barone, Marco Fusè, Federico Lazzari, and Giordano Mancini. *Benchmark Structures and Conformational Landscapes of Amino Acids in the Gas Phase: a Joint Venture of Machine Learning, Quantum Chemistry, and Rotational Spectroscopy*, Volume 19, Number 4, Pages 1243–1260, **2023**, J. Chem. Theory Comput.

- Vincenzo Barone, Giorgia Ceselin, Federico Lazzari, and Nicola Tasinato. *Toward Spectroscopic Accuracy for the Structures of Large Molecules at DFT Cost: Refinement and Extension of the Nano-LEGO Approach.*, **2023**, J. Phys. Chem., Submitted

Chapter 1

Introduction

The goal of Chemistry is to study the structure of molecules and their reactions. A typical chemical bond is of the order of the Angstrom, which corresponds to 10^{-10} meters. Chemical bond breaking and formation, instead, is a phenomenon that can take place on the scale of picoseconds (10^{-12} seconds). However, the human scale of the scientist is the scale of meters and seconds. Thus, there is a fundamental difficulty in understanding and studying such complex objects that led to the creation of graphical representations, such as those now known as structural formulas. The need for graphical representations is a typical human necessity that encompasses several other topics, such as the development of music notation, Feynman diagrams, or even languages and phonetic symbols. The advent of Quantum Mechanics at the end of the 20th century further complicated the representation of such microscopic objects. In fact, if we think of a molecule as a set of atoms sparse in space described by Cartesian coordinates (the typical XYZ coordinates), quantum mechanics actually works in a complex Hilbert space. This dichotomy between conceptually representing a phenomenon so as to make it understandable by a human and the ability to predict and quantify its characteristics is at the foundation of Chemistry and it is nowadays more relevant than ever due to the advent of Artificial Intelligence [10].

The recent advancements in computational sciences, the ever-increasing power of CPU and GPU, the decreasing cost of computer memory, and the developments in the Machine-Learning field, raise new questions when it comes to our traditional view of the scientific method itself. The computational scientist, up to this point faithful to physics with layers of approximations on top of the fundamental rules of nature, now has to deal with the possibility of simulating nature with good accuracy without necessarily relying on its rules. This is a fundamental shift in the way of conceiving the scientific method. In fact, up to this point, making progress in "understanding" a physical/chemical phenomenon was often associated with a higher capability of predicting the phenomenon itself. The rise of deep-learning models broke such a relation; instead of having the scientist build approximations manually from fundamental laws, algorithms are capable of probabilistically finding such approximations by themselves. The disadvantage of such black-box methods is that the "understanding" of the phenomenon is lost and such ML models are not very flexible but excessively dependent on the case study (or the dataset) given as a problem to study. The lack of flexibility of such models is related to their lack of true intelligence or awareness of the

physical laws. However, in a world in need of even faster results (e.g. the recent rush to discover a vaccine for the COVID pandemic) these tools are more popular than ever in giving first guesses and suggestions to the scientists, who can later try to justify the physical meaning behind it. The most recent trends are the Generative Machine-Learning methods (e.g. Generative Adversary Neural-Networks or GANN [11], Stable Diffusion [12], etc.) where the algorithms are not only capable of making "predictions" on some input data but are capable of creating new data altogether. This new class of methods can already be seen in popular products such as the most recent ChatGPT implementations [13] in the Natural-Language Processing (NLP) landscape. Natural language text, images, and videos are obvious fields of applications of such techniques since these items are hardly described by any physical law, but are still solvable by us, humans. However, even in those fields the limits of such algorithms are well known as their capability of "hallucinating" making convincing predictions that are extremely wrong or creating wrong data. Moreover, there is also a long-term problem of training new models on datasets that can also be created by ML, building a circular path in the flow of data that raises new questions on the future of the field (such as the automatic creation of fake news, and the training of new models on such news). This shift towards "automated" algorithms that can process or create data is also happening in a historical moment when technology is more available than ever to almost anyone. The recent excitement on new forms of interactions with technology, not only from the mobile world but also in Augmented Reality and Virtual Reality, can help us identify a common denominator in recent technology, which is the subversion of reality, not much interest in understanding nature as itself but rather replicate it as convincingly as possible and eventually building it from scratch (whether is the so-called Artificial Intelligence, or the Metaverse). As a computational scientist, it is then important to keep awareness of the goal of science, which is first and foremost "understanding", with the use of fair graphical representations, over "predicting". However, it would be silly to ignore such progress and their prediction capabilities in cases of a well-designed model, and trying to identify those scenarios where it is possible to use some of these algorithms in speeding up the research process is of scientific relevance. As an example, the aforementioned Augmented Reality tools might increase in popularity and availability in the next 10 years or so. The scientist has a new opportunity then, which is to directly interact with its objects of study instead of having to rely on some intermediary (e.g. the computer screen, some log file, etc.). The fair use of AR won't be to create a new parallel reality to the one we are living, but instead to allow more natural intuitive interaction with complex physical/mathematical abstractions that still describe our reality (in a "What You See Is What You Get" fashion). The same application can be thought of for ML techniques, where a lot of procedures that the computational scientist has to perform manually can be streamlined by automation.

The main goal of the thesis is to create a new set of mathematical descriptors to describe atom types, bond types (synthons), or angle types to use in ML procedures while retaining the chemical meaning during the process. The goal is to make sure that the ability to make quantitative predictions of such descriptors goes hand in hand with the capability of interpreting changes in the descriptors themselves in terms of chemical change. In other words, it is a link between the

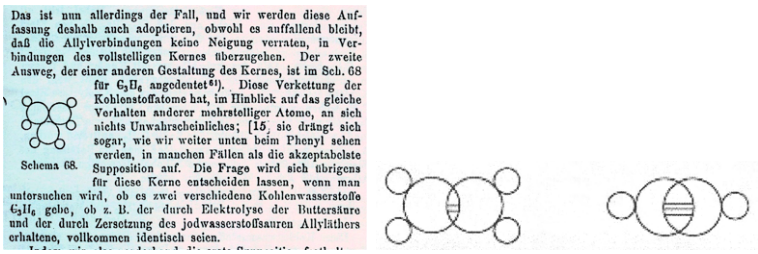
quantitative microscopic world of molecules and our traditional view of a molecule as a set of atoms connected by springs. To determine such descriptors, it was necessary to develop a set of tools that allow us to move from the traditional XYZ space to this new feature space. The way to do it was to develop a software, called Proxima, that automatically "perceives" such features from molecules (hence the name Molecular Perception). The Proxima's inner way of working is to treat ensembles of items deeply correlated with each other as single individual units. These units are weekly coupled together in a perturbative way. In practice, Proxima treats couples of strongly coupled electrons as covalent bonds and lone pairs, and treats units of delocalized electrons with a dedicated Tight-Binding model. The Molecular Graph is the resulting graph containing such units as nodes and the weak interactions between them as edges of the graph. The current version of Proxima is designed to work with fixed-topology systems, but the definition of our perception descriptors is well-designed to be continuous descriptors to be easily extended to changing topology systems (e.g. chemical reactions, transition states, etc.)

The context of the thesis is the one of the Virtual Laboratory as shown in Fig. 1. The Virtual-Laboratory is a workflow encompassing all phases of a computational study from the pre-processing phase dedicated to the construction of molecular systems from databases, through the computational phase that is dedicated to applying different Molecular Perception, Quantum Mechanics, Molecular Mechanics and Machine-Learning techniques for both exploitation and exploration, up to the post-processing phase employing either ML to perform data analysis or chemical intuition through the use of Immersive Virtual Reality tools. The thesis is organized as follows, after some initial chapters on the historical backgrounds of traditional computational methods and Machine-Learning techniques, we discuss our in-house software to perform Molecular Perception (required for the computation of charges and descriptors for ML). Then, a chapter is dedicated to the definition of continuous chemical descriptors. Finally, a chapter is dedicated to outlining some applications of our descriptors.

Chapter 2

Representation and Computation

The human desire to understand the mechanisms of complex phenomena is at the foundation of chemistry. By living in a world that is at the scale of the meter, we have difficulties understanding the complexity of a quantum-governed environment, making the problem of representing molecules a challenging task. Anticipating the work of Kekulé, in 1861 Loschmidt [14] studied cyclic compounds using a representation of the atoms in terms of spheres as shown in Fig. 2.1.


Das ist nun allerdings der Fall, und wir werden diese Auffassung deshalb auch adoptieren, obwohl es auffallend bleibt, daß die Allylverbindungen keine Neigung verraten, in Verbindungen des vollstelligen Kernes überzugehen. Der zweite Ausweg, der einer anderen Gestaltung des Kernes, ist im Sch. 68 für C_3H_6 angedeutet[61]). Diese Verkettung der Kohlenstoffatome hat, im Hinblick auf das gleiche Verhalten anderer mehrstelliger Atome, an sich nichts Unwahrscheinliches; [15] sie drängt sich sogar, wie wir weiter unten beim Phenyl sehen werden, in manchen Fällen als die akzeptabelste Supposition auf. Die Frage wird sich übrigens für diese Kerne entscheiden lassen, wenn man untersuchen wird, ob es zwei verschiedene Kohlenwasserstoffe C_3H_6 gebe, ob z. B. der durch Elektrolyse der Buttersäure und der durch Zersetzung des jodwasserstoffsauren Allyläthers erhaltene, vollkommen identisch seien.

Schema 68.

Figure 2.1: Joseph Loschmidt: Structural formulae, 1861.

With new discoveries and the increasing complexity of molecular structures discovered, new representations were required. In 1891, Emil Fischer introduced the Fischer projections to include three-dimensionality in the treatment of molecules. The awareness of the three-dimensionality of a molecule is what unlocked the study of stereo-chemistry. Thanks to discoveries in biochemistry and crystallography, the need to represent even bigger molecules (e.g., proteins) emerged. As a consequence, new concepts such as the ones of helix, beta-sheets, and secondary structures required new representations. Initially, the visualization of such complex structures employed physical models derived from accurate data measurements. John Kendrew [15] won the Nobel Prize in 1962 for his studies on Hemoglobin and Myoglobin with the use of X-Ray Crystallography. In Fig. 2.2, the original structure of Myoglobin resolved by Kendrew is shown.

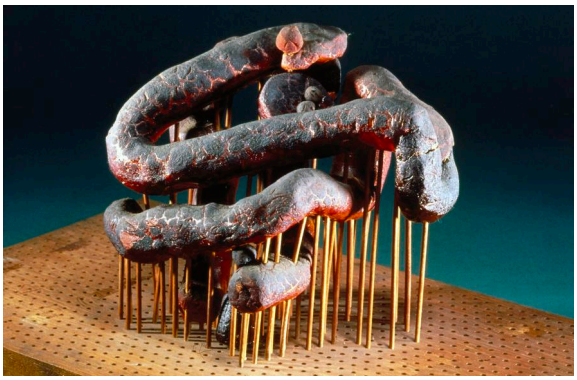


Figure 2.2: The original model of the Myoglobin molecule, (the first model of a protein molecule) built in plasticine in 1957 by Dr. J.C. (later Sir John) Kendrew.

The rationale behind such representations is that we are visual human beings, and we have a natural intuition to analyze visual data. The real breaking point that truly challenged molecular representation was the advent of Quantum Mechanics and how much chemistry can be explained in quantum terms. The traditional view of the world of molecules made of spherical balls connected together by springs was deeply challenged by the new discoveries in physics. In fact, abstract and complex mathematical concepts started to be used for describing the molecular structure, creating a gap between the quantitative nature of Quantum Chemistry and the qualitative interpretability of traditional organic and general chemistry.

2.1 The problem of QM Representation

At the dawn of the XXth century, physicists had to develop a new mechanical theory for the description of phenomena such as black-body radiation and the photoelectric effect. It became evident that objects behave differently at a microscopic scale. The development of quantum physics had a direct impact on chemistry since atoms could be described in quantum terms, revolutionizing the description of matter itself. In fact, the traditional concept of the locality of a particle (x, y, z, t), fails under the uncertainty principle $\sigma_x \sigma_p \geq \frac{\hbar}{2}$. In other words, it is not possible to know exactly both WHERE and WHEN something is. In particular, chemists were forced to rethink molecules and atoms, not as an assembly of particles orbiting one around the other (as planets do), but as delocalized clouds of probabilities described by wavefunctions, functions that tell us the "probability" of finding a particle in a given region of space:

$$\int_{x}^{x+dx} \left|\psi(t)\right|^2 dt = P_x^{x+dx} \tag{2.1}$$

Quantum Chemistry is focused on developing approximated quantum models not just for atoms but also for molecules. It is possible to argue that such an "invasion" of quantum physics into chemistry was not always well received. In fact, the split between the classical, easy-to-represent, but imprecise vision of molecules and the quantum, harder-to-represent, but precise description generated an equal division within the chemical community. History has strongly

proven that chemistry is not just a single-method field but is a multidisciplinary discipline that requires trying and confronting different approaches to the study of matter and molecules. It is very common for today's experimental studies to include computational analysis, and it is also important for many theoretical works to include "chemical" interpretations of results or experimental confirmations. In the following, we are going to focus on time-independent quantum theory by describing molecular structure not including the time evolution of systems. In this context, the quantum theory determines that the energy of a system (in our case an atom or a molecule) is discrete, not continuous, encapsulated in "quanta" of energy or energy levels obtained by the solution of the eigenvalue Schrödinger equation:

$$\hat{H}\psi = E\psi \tag{2.2}$$

And in general any "observable" (that is a measurable physical quantity) is obtained by the effect of an operator on the wavefunction:

$$x = \langle\psi|\hat{X}|\psi\rangle = \int \psi^* \hat{X}\psi \, dx \tag{2.3}$$

The importance of the wavefunction is now clear since it allows us to compute observables, but its representation is not so obvious. For single electron systems (e.g. the hydrogen atom) the Schrödinger equation has an analytical solution. The most common way to visualize these wavefunctions is through the concept of an "orbital". The idea is to use the square of the wavefunction which, as said before, represents the probability of finding a particle in a region of space, and by employing a cutoff value (e.g. 0.9) the function becomes an equation which, once solved, gives rise to the orbital surfaces (as shown in Fig. 2.3):

- $\hat{H}\psi_i = \epsilon_i\psi_i$
- $\rho_i(x) = \left|\psi_i^*(x)\psi_i(x)\right|^2$
- $\rho_i(x_{orb}) = 0.9$
- $x_{orb} = \rho_i^{-1}(0.9)$

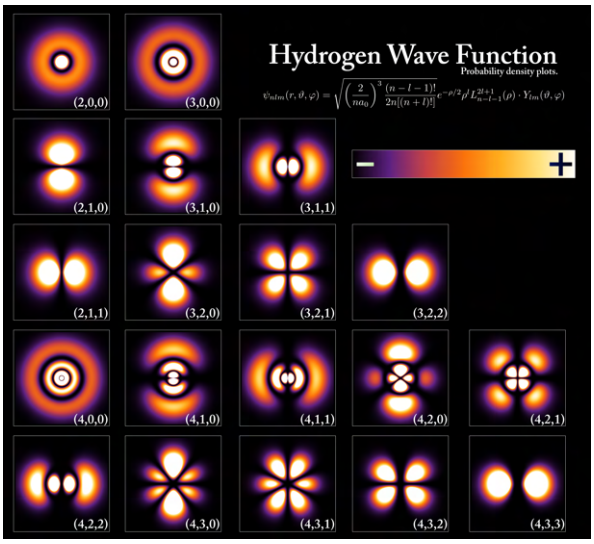


Figure 2.3: The atomic orbitals for the hydrogen atom.

The description of multi-electron systems is even more complicated due to the presence of a correlation between electrons. The correlation effect originates from the antisymmetry of the electronic wave-function (because of the fermionic nature of the electron), and also by the explicit potential repulsive interaction between electrons themselves. Electronic correlation is what determines the following inequality between the conditional probability of finding two electrons in a given region of space and the product of the individual probabilities:

$$\rho(r|r') \neq \rho(r)\rho(r') \tag{2.4}$$

The general assumption is that it is still possible to describe the wavefunction of a multi-electron system as a combination of single-electron wavefunctions. The way these orbitals are combined is through Slater determinants so as to maintain the antisymmetry property for fermionic wavefunctions. For example, in a wavefunction of N fermions:

$$\Psi(x_1, x_2, ..., x_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(x_1), \psi_2(x_1), ..., \psi_N(x_1) \\ \psi_1(x_2), \psi_2(x_2), ..., \psi_N(x_2) \\ ... \\ \psi_1(x_N), \psi_2(x_N), ..., \psi_N(x_N) \end{vmatrix} \tag{2.5}$$

It is obvious how such a description gets even more complicated and harder to represent. As a consequence, with the evolution of quantum theory, the "second quantization" formalism was introduced to simplify the notation. It is important to notice how the second quantization was, in some way, a different approach to the problem of "human representation" of quantum systems rather than a new physical theory since it does not provide new physical laws. In particular, in the second quantization approach, the idea is to build a formal mathematical language to describe quantum systems based on the human intuition of "placing electrons in orbitals" (though this is not physically exact since there is no orbital without electrons). As such, given N quantum levels, we can define a quantum state as a vector in the Fock space defined by its occupation numbers ($k_i = 1$ if the i-th spin-orbital is occupied, 0 otherwise):

$$\big|k\big\rangle = \big|k_1, k_2, ..., k_N\big\rangle \tag{2.6}$$

and the vacuum state occurs when all levels are empty $\big|vac\big\rangle = \big|0, 0, ..., 0\big\rangle$. In order to add an electron to a level a creation operator is employed:

$$a_P^\dagger \big|k_1, k_2, ..., 0_P, ..., k_N\big\rangle = \Gamma_P^k \big|k_1, k_2, ..., 1_P, ..., k_N\big\rangle \tag{2.7}$$

Where $\Gamma_P^k = \prod_{Q=1}^{P-1}(-1)^{k_Q}$ and $a_P^\dagger \big|k_1, k_2, ..., 1_P, ..., k_N\big\rangle = 0$. In order to remove an electron from a spin-orbital, instead, an annihilation operator is used:

$$a_P \big|k_1, k_2, ..., 1_P, ..., k_N\big\rangle = \Gamma_P^k \big|k_1, k_2, ..., 0_P, ..., k_N\big\rangle \tag{2.8}$$

With the same Γ_P^k and $a_P|k_1, k_2, ..., 0_P, ..., k_N\big\rangle = 0$.

Although the use of mathematics and abstraction seems to complicate the description of the system, it actually gets similar to the graphical intuition we have when discussing quantum levels by adding and removing electrons (such as the two configurations of Fig. 2.4 obtained by removing one electron from the bottom level and adding it to the top level).

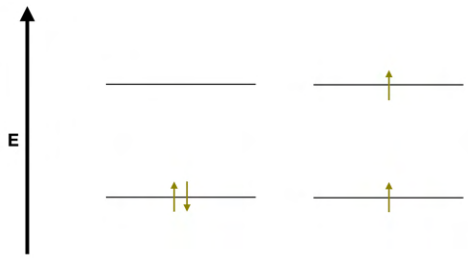


Figure 2.4: Two electronic configurations.

It is important to notice how the two main approaches usually employed in QM methods today are to either treat electrons as occupying single orbitals or to entirely delocalize the electrons working directly with the overall electron density (DFT methods). The new class of methods that try to describe a QM system as sub-sets of strongly correlated electrons weekly coupled together (F12 methods) represents an intermediary between traditional and DFT approaches. In the following chapter, we are going to discuss very briefly some of the most common quantum and classical methods for computing energy and molecular properties. It is not of course the main topic of the following thesis to give a full description of all of the computational methods in chemistry, far from that, but it is meant to provide a general discussion of the most common methodologies to then introduce the advent of ML and its different philosophy in the next chapter.

2.2 Electronic Structure Methods

The analytical solution of the Schrödinger equation is only found for mono-electronic systems such as the hydrogen atom or the $H_2{}^+$ molecule. Thus, throughout the years, multiple methods have been developed to numerically compute, with the least amount of approximations, molecular orbitals, and energies. The

general framework is the Born-Oppenheimer approximation (or adiabatic approximation) in which the movement of the nuclei of a molecule is decoupled from the movement of the electrons. Thus, the Schrödinger equation to solve is the one describing the electronic structure of a molecule taking a reference geometry with the molecular electronic Hamiltonian (in atomic units):

$$\hat{H} = \sum_{PQ} h_{PQ} a_P^\dagger a_Q + \frac{1}{2} \sum_{PQRS} g_{PQRS} a_P^\dagger a_R^\dagger a_S a_Q + h_{nuc} \tag{2.9}$$

Here we are using creation and annihilation operators $(a^\dagger, a)$ operating in the Fock space as described in the previous section. The one-electron term h_{PQ} describes the kinetic energy of an electron together with the nuclear potential energy field:

$$h_{PQ} = \int \phi_P^*(x) \left(-\frac{1}{2}\nabla^2 - \sum_l \frac{Z_l}{r_l} \right) \phi_Q(x) dx \tag{2.10}$$

Where Z_l are the nuclei atomic numbers, ϕ_P are the spin-orbitals defining the Fock space (the molecular spin-orbitals), and r_l is the distance between a generic point x and the l-th nucleus. The two-electron term describes the electron-electron interaction term (the hardest term to compute and accounts for the electron correlation):

$$g_{PQRS} = \int \int \frac{\phi_P^*(x_1)\phi_R^*(x_2)\phi_Q(x_1)\phi_S(x_2)}{r_{12}} dx_1 dx_2 \tag{2.11}$$

Where $r_{12} = ||x_2 - x_1||$. The last term, the h_{nuc}, is a number in the Born-Oppenheimer approximation and accounts for the nuclei-nuclei interaction energy:

$$h_{nuc} = \frac{1}{2} \sum_{I \neq J} \frac{Z_I Z_J}{R_{IJ}} \tag{2.12}$$

In general, the MO-LCAO (Linear Combination of Atomic Orbitals) approximation is taken into account to describe the molecular spin-orbitals (ϕ_P). In particular, the molecular orbital is described as a sum of functions centered on the nuclei such as the atomic orbitals. In general, the set of atom-centered functions employed is called the basis set and its choice hugely impacts the computation since it defines the shapes of the molecular orbitals.

2.2.1 Hartree-Fock

The Hartree-Fock [16] is the first method developed to compute molecular properties. The general idea of the HF method is to use a single configuration to describe the molecule (that is a single vector in the Fock space), and the energy is optimized by varying the coefficients of the LCAO spin-orbital base. As a consequence, the HF wave function can be described as:

$$\big|k\big\rangle = e^{-\hat{k}} \big|0\big\rangle \tag{2.13}$$

Where $|0\rangle$ is the reference configuration and $e^{-\hat{k}}$ is an operator that carries out unitary transformations within spin-orbitals. The other important characteristic of the HF method is that it simplifies the shape of the Hamiltonian by taking an "average field" for the electron-electron interaction (thus using a single variable function) instead of punctual values for each couple of electrons (that is a two-variable function). Thus, the "fock operator" is used in place of the Hamiltonian to find approximate solutions to energies and molecular orbitals:

$$\hat{f} = \hat{h} + \hat{V} \tag{2.14}$$

With $\hat{h}$ the true Hamiltonian kinetic energy and nuclei potential energy, and $\hat{V}$ the average Coulomb repulsion among electrons corrected for Fermi correlation:

$$\hat{V} = \sum_{PQ} V_{PQ} a_P^\dagger a_Q \tag{2.15}$$

$$V_{PQ} = \sum_{I} (g_{PQII} - g_{PIIQ}) \tag{2.16}$$

Where I runs over all occupied spin-orbitals while P and Q run over all spin-orbitals (occupied and unoccupied). The exchange correction term is required for the antisymmetry of the wave function. In this way, the HF method uncorrelates the electrons so that the "correlation energy" is generally defined using the HF method as a reference:

$$E_{corr} = E_{exact} - E_{HF} \tag{2.17}$$

As a consequence, many other methods were developed from the HF so as to introduce back some amount of correlation (thus called post-HF methods). The HF method is usually called Self-Consistent since the average electron-electron potential defining the Fock matrix (V_{PQ}) is obtained as the result of the diagonalization of the Fock matrix but is also required to build the Fock matrix thus generating an iterative procedure. The eigenvalues and the eigenvectors of the Fock matrix define both the energies and the molecular orbitals (the coefficients in the LCAO approximation).

2.2.2 MCSCF

The Multi Configurational Self-Consistent Field [17] is the natural evolution of the HF method since it takes into account multiple electronic configurations. In particular, MCSCF was proven successful in describing bond-breaking and molecular dissociations. In MCSCF theory, the wave function is written as a linear combination of determinants (configurations) whose expansion coefficients are optimized together with the MOs according to the variation principle. Thus, the MCSCF wave function can be written as:

$$|k, C\rangle = e^{-\hat{k}} \sum_i C_i |i\rangle \tag{2.18}$$

The same unitary transformation for the spin orbitals is applied ($e^{-\hat{k}}$), but this time instead of a single configuration we expand over multiple configurations

each one with its weight C_i. In this case, both the k and C terms are optimized by minimizing the energy through the variational principle. It is important to notice that the exact solution is only possible when all possible configurations are taken into account (FCI). The problem with the MCSCF is that it is only possible to treat relatively small numbers of configurations, one common approach being the CASSCF where the choice of the configurations is limited to a subset of orbitals (the "active" orbitals). Thus, MCSCF methods prove a flexible framework for treating "static correlation" (that is correlation arising from degenerate or nearly degenerate electronic configurations) but additional calculations need to be performed to get a good description of "dynamic correlation" (that is the correlation due to the movement of the electrons).

2.2.3 Configuration Interaction

In the CI method (Configuration Interaction) [18] the wave function is constructed as a linear combination of determinants:

$$\big|C\big\rangle = \sum_i C_i \big|i\big\rangle \tag{2.19}$$

This is in principle similar to the MCSCF method, with the difference being that only the configuration expansion is variationally optimized (that is the C_i), the orbitals are generated separately in a preceding HF or MCSCF calculation and are held fixed during the optimization of the configuration expansion. In theory, MCSCF wave functions would be more flexible than the CI wave functions, but in practice are limited to small configuration expansions. The CI successfully allows for the computation of the dynamic correlation energy.

2.2.4 Many-Body Perturbation Theory

The Møller-Plesset method [19] uses the perturbation theory instead of the variational approach to introduce correlation in many-body systems. In particular, the Hamiltonian is written as:

$$\hat{H} = \hat{f} + \hat{\Phi} + h_{nuc} \tag{2.20}$$

Where $\hat{f}$ is the Fock operator, and $\hat{\Phi}$ is the fluctuation potential which describes the difference between the true electron Coulomb potential $\hat{g}$ of the hamiltonian operator and the effective one-electron Fock potential $\hat{V}$ of the Fock operator:

$$\hat{\Phi} = \hat{g} - \hat{V} \tag{2.21}$$

Applying the standard machinery of the perturbation theory, we obtained the second order in the perturbation:

$$\begin{cases} E_{MP}^{(0)} = \langle HF | \hat{f} | HF \rangle = \sum_I \epsilon_I \\ E_{MP}^{(1)} = \langle HF | \hat{\Phi} | HF \rangle \\ E_{MP}^{(2)} = - \sum_{A>B,I>J} \frac{|g_{AIBJ} - g_{AJBI}|^2}{\epsilon_A + \epsilon_B - \epsilon_I - \epsilon_J} \end{cases} \tag{2.22}$$

Thus, the Hartree-Fock energy can be written as:

$$E_{HF} = E_{MP}^{(0)} + E_{MP}^{(1)} + h_{nuc} = \langle HF | \hat{H} | HF \rangle \tag{2.23}$$

And, by adding the second-order energy:

$$E_{MP2} = E_{HF} - \sum_{A>B,I>J} \frac{|g_{AIBJ} - g_{AJBI}|^2}{\epsilon_A + \epsilon_B - \epsilon_I - \epsilon_J} \tag{2.24}$$

In general, the Møller-Plesset method can be employed in its second-order formulation (MP2) or can be extended to higher orders (MP3, etc.).

2.2.5 Coupled-Cluster

The two most serious shortcomings of the CI approach are the lack of size-extensivity and the slow convergence to the FCI limit. These limits are overcome by the Coupled-Cluster theory [20]. The starting point is to rewrite the FCI wave function in terms of excitation operators. The excitation operator is an operator that promotes an electron from a spin-orbital to another in the Fock space:

$$\hat{X}_I^A | HF \rangle = C_I^A a_A^\dagger a_I | HF \rangle \tag{2.25}$$

Thus obtaining:

$$| FCI \rangle = \left(1 + \sum_{AI} \hat{X}_I^A + \sum_{A>B,I>J} \hat{X}_{IJ}^{AB} + ... \right) | HF \rangle \tag{2.26}$$

It is possible to recast the linear summation of excitation operators in the FCI in the form of a product wave function:

$$| CC \rangle = \left[\prod_{AI} (1 + \hat{X}_I^A) \right] \left[\prod_{A>B,I>J} (1 + \hat{X}_{IJ}^{AB}) \right] ... | HF \rangle \tag{2.27}$$

In order to simplify the algebraic manipulation of this product, we notice that since:

$$\hat{X}_{IJ}^{AB} \hat{X}_{IJ}^{AB} = 0 \tag{2.28}$$

we may write:

$$1 + \hat{X}_{IJ}^{AB} = 1 + \hat{X}_{IJ}^{AB} + \frac{1}{2} \hat{X}_{IJ}^{AB} \hat{X}_{IJ}^{AB} + ... = e^{\hat{X}_{IJ}^{AB}} \tag{2.29}$$

And similar for other excitations, thus obtaining:

$$\big|CC\big\rangle = \exp\left\{\sum_{AI} t_I^A a_A^\dagger a_I + \sum_{A>B,I>J} t_{IJ}^{AB} a_A^\dagger a_B^\dagger a_I a_J + ...\right\}\big|HF\big\rangle \tag{2.30}$$

Or

$$\big|CC\big\rangle = e^{\hat{T}}\big|HF\big\rangle \tag{2.31}$$

with

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + ... \tag{2.32}$$

being an operator containing single-excitations, double-excitations, etc. This exponential ansatz is at the foundation of the CC method. It is common practice to truncate the summation of the $\hat{T}$ operator so as reduce the number of computations (such as the CCSD, CC-Single and Double excitations), but even when truncating such operator we still get contributions from all determinants in contradiction with the CI methods. The other difference with the CI method is that in this case the energy is not obtained by means of a variational approach but instead by projecting onto a set of configurations $\big\langle\mu\big|$ that span the set of all states obtained by applying the truncated $\hat{T}$ operator.

$$e^{-\hat{T}}\hat{H}e^{\hat{T}}\big|HF\big\rangle = E_{CC}\big|HF\big\rangle \tag{2.33}$$

$$\big\langle\mu\big|e^{-\hat{T}}\hat{H}e^{\hat{T}}\big|HF\big\rangle = 0 \tag{2.34}$$

The energies are obtained by inverting the equations above.

2.2.6 Density Functional Theory

In this chapter, we have stressed how the biggest problem in representing the quantum world is having to deal with abstract quantities such as wave functions. The complexity of the wave function is not only in its nature but also in the number of coordinates that requires: $3n$ spatial coordinates and n spin coordinates. In fact, it is possible to say that the wave function contains more information than needed and is lacking direct physical significance. The real quantity that we can intuitively interpret is its square product that represents the electron density (i.e. the probability of finding the electron in a given region of space):

$$\rho(x,y,z) = N\sum_{s_1}...\sum_{s_N}\int dr_2...\int dr_N\big|\psi(r,s_1,r_2,s_2,...,r_N,s_N)\big|^2 \tag{2.35}$$

The real advantage of using the electron density is its reduced number of coordinates (x,y,z) with respect to the wave function. This is another example of how the problem of representation is directly linked to the complexity of the theoretical models for computation. The revolution happened in 1964 when Pierre Hohenberg and Walter Kohn proved two fundamental theorems that allowed the development of a Density Functional Theory [21].

The Hohenberg-Kohn theorem

The first theorem states that for molecules with a non-degenerate ground state, the ground-state molecular energy, wave function, and all other molecular electronic properties are uniquely determined by the ground-state electron probability density. In other words, the ground-state energy is a functional of the electron density:

$$E_0 = E_0[\rho_0] \tag{2.36}$$

The key is to prove that given an electron density ρ_0 for a ground-state non-degenerate system, this uniquely identifies an external potential $v(r_i)$ in the Hamiltonian (the external potential is essentially the coulomb interaction with the nuclei of the molecule). Being the external potential the only difference between two molecular Hamiltonians, it proves the uniqueness of the electron density. Let's suppose there are two different Hamiltonians $\hat{H}_a$ and $\hat{H}_b$ (thus two Hamiltonians that differ in their external potentials $v_a(r_i)$ and $v_b(r_i)$), the following relations must be valid:

$$\begin{cases} \hat{H}_a\psi_{0,a} = E_0^a\psi_{0,a} \\ \hat{H}_b\psi_{0,b} = E_0^b\psi_{0,b} \end{cases} \tag{2.37}$$

At this point, let's apply the variational theorem by using $\psi_{0,b}$ as a trial function for $\hat{H}_a$:

$$\begin{aligned} E_{0,a} \ &< \langle\psi_{0,b}|\hat{H}_a|\psi_{0,b}\rangle = \langle\psi_{0,b}|\hat{H}_a + \hat{H}_b - \hat{H}_b|\psi_{0,b}\rangle \\ &= \langle\psi_{0,b}|\hat{H}_a - \hat{H}_b|\psi_{0,b}\rangle + \langle\psi_{0,b}|\hat{H}_b|\psi_{0,b}\rangle \end{aligned} \tag{2.38}$$

Since the only difference between the two Hamiltonians is the external potential we get:

$$E_{0,a} < \left\langle \psi_{0,b}\middle| \sum_{i=1}^{n}[v_a(r_i) - v_b(r_i)]\middle|\psi_{0,b}\right\rangle + E_{0,b} \tag{2.39}$$

which in turn gives rise to the following two equations:

$$\begin{cases} E_{0,a} < \int \rho_{0,b}(r)[v_a(r) - v_b(r)]dr + E_{0,b} \\ E_{0,b} < \int \rho_{0,a}(r)[v_b(r) - v_a(r)]dr + E_{0,a} \end{cases} \tag{2.40}$$

It is clear that if the two-electron densities are identical despite the different external potentials, the summation of the two inequalities above gives rise to: $E_{0,a} + E_{0,b} < E_{0,a} + E_{0,b}$ which is clearly false. Thus, ρ_0 determines the molecular electronic Hamiltonian and so the ground-state wavefunctions, energy, and other properties. The second important theorem proves that the variational theorem can also be applied to the electron density, thus: the true ground-state electron density minimizes the energy functional $E[\rho_{tr}]$.

The Kohn-Sham Method

The Hohenberg-Kohn theorems don't tell us how to compute the energy from the electron density nor how to obtain the electron density without knowing the wavefunction first, they just prove that is theoretically possible. One strategy to practically solve the issue is the Kohn-Sham (KS) method developed in 1965 [22] that, in principle, is capable of yielding exact results although in practice requires an unknown operator that must be approximated. The starting point is to consider a fictional system of n non-interacting electrons each experiencing the same external potential $v_s(r_i)$ such as to make the ground-state electron probability density of the reference system equal to the exact ground-state electron density of the molecule ($\rho_s(r) = \rho_0(r)$). Since the theorems prove that the ground-state probability density function determines the external potential, the $v_s(r_i)$ is uniquely determined although we might not know how to practically compute it. In the reference system, electrons do not interact with each other, so the Hamiltonian of the reference system is simply:

$$\hat{H}_s = \sum_{i=1}^{n} \left[-\frac{1}{2}\nabla_i^2 + v_s(r_i) \right] = \sum_{i=1}^{n} \hat{h}_i^{KS} \tag{2.41}$$

Since the reference system is made of non-interacting particles, we can still write the ground-state wave function of the reference system as the antisymmetrized product (Slater determinant) of the lowest-energy KS spin-orbitals μ_i^{KS} of the reference system, where the spatial part is an eigenfunction of the one-electron KS operator: $\hat{h}_i^{KS}$. In order to quantify and approximate the external potential of the reference system, it is convenient to rewrite the energy as a functional of the electron density:

$$E_v[\rho] = \int \rho(r)v(r)dr + T_s[\rho] + \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + \Delta T[\rho] + \Delta V_{ee}[\rho] \tag{2.42}$$

Where T_s is the kinetic energy of the reference system made of non-interacting particles, easy to evaluate remembering that a single Slater determinant describes the non-interacting system:

$$T_s = -\frac{1}{2}\sum_{i=1}^{n} \langle \mu_i^{KS} | \nabla_i^2 | \mu_i^{KS} \rangle \tag{2.43}$$

The $\Delta T[\rho]$ functional is the difference between the kinetic energy of the molecule and the reference system, while ΔV_{ee} is the difference of the potential energy between the molecule and the reference system. These two terms get summed together in a single functional of the electron density called the exchange-correlation functional, thus obtaining:

$$E_0 = \int \rho(r)v(r)dr + T_s[\rho] + \frac{1}{2}\int\int \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + E_{xc}[\rho] \tag{2.44}$$

The variational principle on the electron density allows us to find the ground-state electron density by variationally changing the KS orbitals of the non-interacting system so as to minimize the energy functional. The exchange-correlation potential is obtained from the functional as follows:

$$v_{xc}(r) = \frac{\delta E_{xc}[\rho(r)]}{\delta \rho(r)} \tag{2.45}$$

The different strategies employed to approximate such functional differentiate the different DFT methods. In the following, we are going to show just the most relevant ones.

Local-Density Approximation

The simplest strategy to approximate the exchange-correlation functional is to treat the problem of a homogeneous gas of electrons, which is a decent approximation in case the electron density varies extremely slowly with position (thus Local-Density Approximation (LDA)). In the case of gas of electrons, the exchange-correlation functional is written as:

$$E_{xc}^{LDA}[\rho] = \int \rho(r)\epsilon_{xc}(\rho)dr \tag{2.46}$$

So that the $\epsilon_{xc}(\rho)$ term is the exchange-correlation energy per electron. It is possible to show that the exchange-correlation energy can be decoupled as the sum of the individual exchange and correlation contributions for which analytical solutions are provided in the literature. In the following, we are going to show just the exchange functional in the case of a homogeneous gas of electrons since the expression for the correlation component is far more complicated and easily available from literature [23].

$$\begin{cases} v_x^{LDA} = -[(3/\pi)\rho(r)]^{1/3} \\ E_x^{LDA} = \int \rho\epsilon_x dr = -\frac{3}{4}(\frac{3}{\pi})^{1/3} \int [\rho(r)]^{4/3} dr \end{cases} \tag{2.47}$$

For open-shell systems, the local-spin density approximation (LSDA) gives better results than the simple LDA, in which electrons with opposite spin have different spatial components in the KS orbitals.

Gradient-Corrected Functionals

The LDA models work best when the electron density does not vary rapidly with changes in position. In order to correct the behavior of the functional considering variations in electron density, Gradient-Corrected functionals (GGA) are introduced in which the gradients of the electron density are explicitly included:

$$E_{xc}^{GGA}[\rho] = \int f(\rho, \nabla\rho)dr \tag{2.48}$$

Approximate expressions are developed using theoretical considerations such as the behavior of the true (but unknown) functional. Often some empiricism is thrown in by choosing the values of parameters in the functional so as to get optimal results. Some common GGA functionals are Perdew and Wang's functionals PWx86, PWx91 [24–26] and Becke's functionals B88 or Bx88 [27].

Meta-GGA Functionals

The idea of Meta-GGA is to simply extend the treatment of the functional to second-order gradients in the form of:

$$E_{xc}^{MGGA}[\rho] = \int f(\rho, \nabla\rho, \nabla^2\rho, \tau)dr \tag{2.49}$$

With τ being the Kohn-Sham kinetic-energy density defined as:

$$\tau = \frac{1}{2}\left|\nabla\mu_i^{KS}\right|^2 \tag{2.50}$$

Meta-GGA functionals require a little more time than GGA but can give better results.

Hybrid Functionals

The exchange functional can be computed in terms of KS orbitals, in a similar fashion to HF, as:

$$E_x^{HF} = -\frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\langle \mu_i^{KS}(1)\mu_j^{KS}(2)\left|\frac{1}{r_{12}}\right|\mu_j^{KS}(1)\mu_i^{KS}(2)\right\rangle \tag{2.51}$$

The correlation energy can thus be identified with $E_c = E_{xc} - E_x$. The idea of hybrid functionals is to mix together such definitions of the exchange energy together with the LSDA, GGA, and Meta-GGA expressions. For example, the popular B3LYP functional [27, 28] (where 3 means is a 3 parameter functional) uses the following definition:

$$E_{xc}^{B3LYP} = (1-a_0-a_x)E_x^{LSDA} + a_0E_x^{HF} + a_xE_x^{B88} + (1-a_c)E_c^{LSDA} + a_cE_c^{LYP} \tag{2.52}$$

The parameters a_0, a_x, a_c are optimized on molecular atomization energies.

Double Hybrid Functionals

Grimme [29] proposed to employ the MP2 energy correction formula to improve DFT energies. To do so, a hybrid-GGA functional is defined as:

$$H_{xc}^{hybrid-GGA} = a_1E_x^{GGA} + (1-a_1)E_x^{HF} + a_2E_c^{GGA} \tag{2.53}$$

Then this functional is used to self-consistently solve for KS orbitals. Then, an improved value of the functional is computed as:

$$E_{xc} = E_{xc}^{hybrid-GGA} + (1-a_2)E_c^{KS-MP2} \tag{2.54}$$

Where E_c^{KS-MP2} is calculated from MP2 equations as a second-order perturbation. This new functional is then employed to compute ground-state properties. Grimme defined the B2PLYP [30] functional by taking E_x^{GGA} as the B88 exchange functional and E_c^{GGA} as the LYP correlation functional. Another functional of this kind is the rev-DSD-PBEP86-D3(BJ) [1–3] (revDSD).

2.2.7 F12 methods

As already discussed, the two main philosophical approaches to electronic structure calculations have been the treatment of electrons as units entirely localized in spin-orbitals (post-HF approaches), or fully delocalized in electron densities (DFT approaches). A different and recent way to approach the electronic structure calculation problem is to place electrons in spin-orbitals, but correlate portions of these together through a dedicated operator including an explicit radial dependency. This is the general approach of the so-called F12 methods [31]. It is not the goal of the following section to do a full review on the F12 but just to briefly introduce the concept since it is at the philosophical foundation of our software for Molecular Perception (Proxima), where independent portions of interacting electrons are weekly coupled together in a perturbative way. For simplicity, we are going to discuss the CCSD-F12 method [32], but the same F12 approach can be employed on different levels of theory such as the MP2. Recalling the CCSD method, the wave function is written as:

$$\big|\psi\big\rangle = \exp\left(\hat{T}_1 + \hat{T}_2\right)\big|\phi\big\rangle \tag{2.55}$$

In traditional CCSD, the $\hat{T}_1$ and $\hat{T}_2$ operators are written as:

$$\begin{cases} \hat{T}_1 = \sum_{a,i} t_a^i \hat{X}_i^a \\ \hat{T}_2 = \frac{1}{2}\sum_{ab,ij} t_{ab}^{ij} \hat{X}_{ij}^{ab} \end{cases} \tag{2.56}$$

With $\hat{X}$ being the excitation operators, $\{i, j, ...\}$ being occupied orbitals and $\{a, b, ...\}$ the virtual orbitals. In CCSD-F12 theory the $\hat{T}_2$ operator is modified as follows:

$$\hat{T}_2 = \frac{1}{2}\sum_{ab,ij} t_{ab}^{ij} \hat{X}_{ij}^{ab} + \frac{1}{2}\sum_{ab,ij} \tau_{\alpha\beta}^{ij} \hat{X}_{ij}^{\alpha\beta} \tag{2.57}$$

The additional amplitudes $\tau_{\alpha\beta}^{ij}$ are defined as follows:

$$\tau_{\alpha\beta}^{ij} = T_{mn}^{ij} \hat{F}_{\alpha\beta}^{mn} \tag{2.58}$$

With

$$\hat{F}_{\alpha\beta}^{mn} = \big\langle mn\big|F_{12}\hat{Q}_{12}\big|\alpha\beta\big\rangle \tag{2.59}$$

Where the $\hat{Q}_{12}$ operator is required to make the F12 configurations orthogonal to the configurations in the molecular orbital (MO) space. The explicitly correlated terms improve the wave function's description of electrons coming close to each other. They augment the conventional CI expansion by additional functions, in which the orbital products $\phi_i(r_1)\phi_j(r_2)$ have been replaced by short-range pair correlation functions

$$\big|u_{ij}(r_1, r_2)\big\rangle = T_{mn}^{ij}\hat{Q}_{12}F_{12}(r_{12})\big|\phi_m(r_1)\phi_n(r_2)\big\rangle \tag{2.60}$$

The terms $mn = ij$ and $mn = ji$ are the most important ones. The function $\left|u_{ij}\right\rangle$ represents a negative short-range hole in the orbital product $\left|ij\right\rangle$. If added to the reference function, it directly suppresses the probability of finding the two $\left|ij\right\rangle$ electrons in a spatial configuration where they are close to each other. In contrast to orbital products, the short-range correlation factor $\hat{Q}_{12}F_{12}\left|\phi_m\phi_n\right\rangle$ can describe the wave function cusp for $r_{12} \to 0$ correctly. Both aspects fix major deficiencies of conventional wave function expansions in terms of Slater determinants. As already stated, MP2-F12 methods exist in which the F_{12} operator is added as a perturbation. In practice, the Ansatz of these methods is to express the $F(r_{12})$ term as a combination of Gaussians to fit an exponential equation.

$$F(r_{12}) = -\frac{1}{\gamma}\exp(-\gamma r_{12}) \approx \sum_i c_i \exp(-\alpha_i r_{12}^2) \tag{2.61}$$

2.2.8 The atomic charges

In this section, we are going to briefly discuss the CM5 charges employed in QM computations since these are going to be of relevance when discussing Proxima's own perception algorithms. In general, the dipole moment, the electron density, and the other multipoles are the only physical observables, but the desire to represent a delocalized quantity such as the electron density in terms of partial charges localized on atoms is yet again an example of how methods and algorithms are developed in order to convert QM quantities into a traditional chemical representation of molecules. Different strategies can be employed to assign such charges and can be divided into 4 classes [33]:

- Class I. These charges are derived by using nonquantum mechanical approaches such as classical models of dipoles or by using a model to extract the charges directly from experimental data, e.g., from the experimental dipole moment of a diatomic molecule.

- Class II. These charges are based on a partitioning of the electron charge density obtained from a Quantum Mechanics calculation into atomic populations. For example, class II charges are those obtained using Hirshfeld population analysis [34], Mulliken population analysis, Löwdin population analysis, natural bond orbital population analysis, atomic polar tensor-based population analysis, etc. These charges obtained from population analysis may depend on the level of theory, for example, on the choice of density functional and/or basis set, and they may yield an unrealistic representation of the molecular dipole moment and higher-order multipole moments in complex molecules.

- Class III. These atomic charges are those fitted to reproduce a physical observable like a quantum-mechanically calculated electrostatic potential (ESP). In general, ESP-derived charges depend on the molecular coordinate system orientation and the choice of fitting points, and they sometimes exhibit an unphysical dependence on internal bond rotations, and their determination can also suffer from ill-conditioning for interior (or buried) atoms in molecules, especially larger ones. These deficiencies of class III charges can be mitigated by using restrained electrostatic potential (RESP)

fitting and introducing other refinements, for example, in the point selection algorithm.

- Class IV. These charges are defined through parametrization to reproduce accurately charge-dependent observables (such as dipole moments) obtained from experimental results or from high-level Quantum Mechanics calculations that are acceptably converged for the quantity under consideration.

In the following, we are going to consider the CMx family of Class IV charges. In particular, the CM5 charges since these are the charges used as a reference in training our own Proxima perception algorithms. In the CM5 model [33], atomic charges in a molecule (either neutral or ionic) are defined by the following equations:

$$q_k^{CM5} = q_k^{HPA} + \sum_{k' \neq k} T_{kk'} B_{kk'} \tag{2.62}$$

Where k and k' run over all the atoms in the molecule, q_k^{HPA} is the partial atomic charge obtained from Hirshfeld population analysis [34] and $T_{kk'} = -T_{k'k}$ is a model parameter to be determined. The term $B_{kk'}$ is defined as:

$$B_{kk'} = \exp\left[-\alpha(r_{kk'} - R_{Z_k} - R_{Z_{k'}})\right] \tag{2.63}$$

Here α is another parameter of the CM5 model and Z_i are the atomic numbers and R_{Z_i} are the covalent radii. The good agreement with experimental dipole moments is confirmed in the original work. Conceptually is as if the original Hirshfeld charges [34] were corrected by adding a Δ on the atomic charge obtained by regression with a radial function for each couple of atoms considered. The good agreement with experimental dipole moments is what prompted us to use such charges as a reference when developing Proxima's own algorithm, as shown in the corresponding chapter.

2.3 Force Field

As already discussed at the beginning of this chapter, the advent of Quantum Mechanics didn't stop the study of molecules as classical entities, and in practice, a lot of improvements have been done in recent years in increasing the accuracy of energy calculations with classical Force Field (FF). The goal of Molecular Mechanics is to use classical mechanics instead of quantum mechanics for the description of the energy of molecular systems. In particular, the goal is to find an analytical expression that computes the energy given the geometry of the molecule $E(x_1, y_1, z_1, ..., x_n, y_n, z_n)$. The Force Field is so-called since it allows us to compute energy, forces, gradient, and hessian analytically from the geometry of the molecular system. In order to correctly quantify these contributions, it is generally assumed that the molecule is in a geometry around its minimum (although recent Force Field have been developed taking reactivity into account, e.g. EVB [35], ReaxFF [36]). In this way, the energy can be reasonably approximated using an expansion around the reference geometry $\vec{X}_{eq}$. In fact, by expanding the energy expression around equilibrium geometry values, it is possible to think of

a molecule, in classical terms, as a set of atoms connected by springs. This mechanical model of a molecule goes back as far as 1930 [37]. In particular, Andrew and co-workers realized that a bond can behave like a spring obeying Hooke's law around the minimum energy distance, and by reasoning around the "strain" that atoms feel in cycles where the angles are different than the conventional tetrahedral geometry, they arrived at formalizing the concept of bending. Their first attempt at parameterizing a mechanical model for molecules was to take Raman spectra, calculate frequencies and associate the spring force constants to the intensity of the frequency by manipulating physical units by means of the formula:

$$\nu = \frac{1}{2\pi}\left(\frac{k}{m}\right)^{1/2} \tag{2.64}$$

Where ν is the frequency, m is the reduced mass and k is the force constant. From the Raman spectrum of ethane, they found a frequency of 990 wave number for the vibration of two carbon atoms that gave rise to a force constant of $4x10^5$ dynes per cm. In the original work, they went as far as trying to find a physical spring that would replicate such force constant obtaining the values shown in Fig. 2.5, building a physical model of a molecule to study its vibrations, an example is given in Fig. 2.6.

Mean Diameter of Coil	13/16 in.
Diameter of wire	#11 (0.120)
No. of active turns	10
Free length	2–1/2 in.
Force per inch deflection	53 lbs.

Figure 2.5: The parameters used for building the spring to represent oscillations in molecules [37].

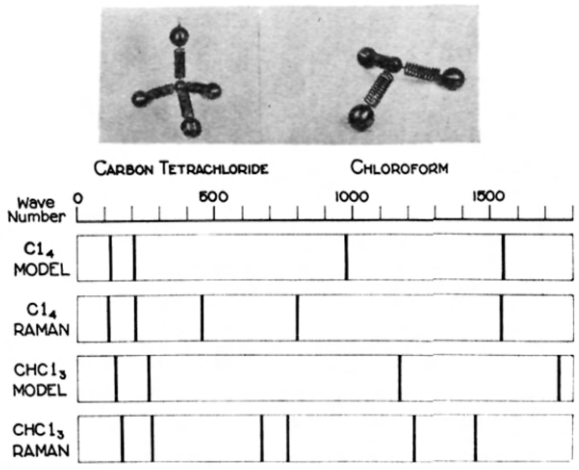


Figure 2.6: The physical models built for studying vibrations mechanically [37].

Since then, more Force Field were developed during the '50 and '60, with some of them still in use today such as MM and its variations MM2, MM3, etc.

[38–40]. In particular, MM was focused on hydrocarbons and defined the energy of a system as a summation over stretching, bending, torsions (for the staggered conformation in ethane), and van der Waals interactions.

$$E = E_{stretching} + E_{bending} + E_{torsion} + E_{vdw} \tag{2.65}$$

$$E_{stretching} = \sum \frac{k_s}{2}(l - l_0)^2 \tag{2.66}$$

$$E_{bending} = \sum \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{2.67}$$

$$E_{torsion} = \sum \left[\frac{V_1}{2}(1 - cos(\omega)) + \frac{V_2}{2}(1 - cos(2\omega)) + ...\right] \tag{2.68}$$

$$E_{vdw} = \epsilon \left[-c_1 \left(\frac{r*}{r}\right)^6 + c_2 e^{\left(-c3\frac{r}{r*}\right)}\right] \tag{2.69}$$

Noticeably, the electrostatic term was missing from the energy expression although they recognized its importance and they suggested using the Del-Re method to face the issue [41] (see Chap. 4 for details on the Del-Re method). MM later evolved in MM2 and MM3 and from MM the MMFF94 and MMFF94s [42, 43] were built. With the advancement in computers and computational power, it was now possible to perform Molecular Dynamics simulations using these Force Field instead of building models manually. At the same time, other Force Field arose with particular regard to the UFF force field (1992) [44] that tried to obtain a set of parameters valid for all the elements of the periodic table, up to AMBER (2002) [45] which is a family of Force Field that all share the same functional form, which allowed more research in trying to parameterize different molecules due to the easiness of just changing parameters in the software while keeping the same functional form. In fact, it is interesting to notice that, in general, most Force Field employ the same functional form which is not that different from the original equation of the MM force field:

$$\begin{aligned} E = \sum_{bonds} K_R(R - R_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 + \sum_{dihedral} \frac{V_n}{2}\left[1 + cos(n\phi - \gamma)\right] \\ + \sum_{i<j} \left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} + \frac{q_i q_j}{\epsilon R_{ij}}\right] \end{aligned} \tag{2.70}$$

There are variations in the choices of functional forms to use, in fact in some Force Field (e.g. UFF) the Morse equation can be employed instead of the simple harmonic one to account for anharmonicity. Moreover, the presence of "improper dihedral angles" can be taken into account treating them as angles in the harmonic equation but with a distinct set of parameters (e.g. the inversion of ammonia). The van der Waals equation is typically used in the Lennard-Jones potential form because of its computational advantages. Traditionally, a separated term for the formation of hydrogen bonds [46] was employed of the form:

$$E_{hbond} = \frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}}$$

In most recent Force Field, however, this term is ignored and the van der Waals and electrostatic interactions are parameterized so as to correctly take into account and describe the formation of hydrogen bonds.

2.4 Conclusions

In this chapter, we have discussed the problem of representation in Chemistry, the most common electronic structure methods, and just a brief introduction to Force Field and their history. In the next chapter, we are going to discuss Machine-Learning (ML) methods and procedures. In particular, if all of this chapter was based on physics as a foundational layer of rules on top of which approximations and models are built, the next chapter is going to illustrate how the problem of representation (that in this chapter was mostly about choosing the right combination of coordinates/formalism and a physical model) is going to be re-framed as the problem of finding a good Feature Space, while "learning" rules through probabilistic engines.

Chapter 3

Machine-Learning

In previous chapters, we stressed the importance of representation and its strong relation with the ability to carry out computations through physical laws. Moreover, we have also highlighted the Machine-Learning (ML) disconnection between the "prediction" of numerical data and the "understanding" of data. In this chapter, we are going to provide a general background on the basics of ML techniques. In Fig. 3.1, the traditional scheme for a non-adaptive scheme is shown, which is the typical scenario for physically based models.

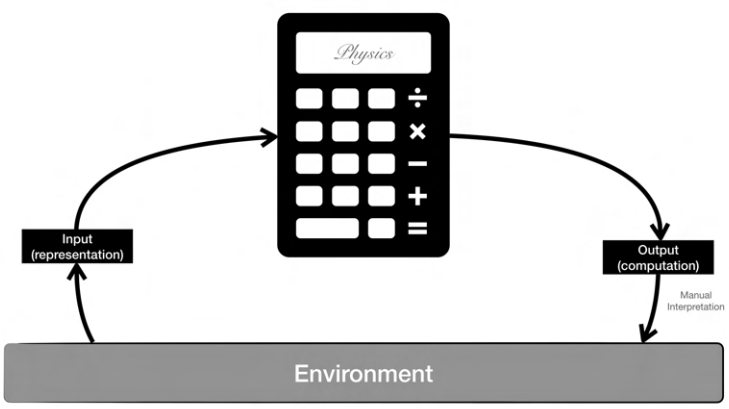


Figure 3.1: The traditional scheme for non-adaptive systems.

The advantage of using non-adaptive models, which are grounded in physical laws, is that they guarantee the validity of the output data since physical laws are hard-coded and immutable, and possible approximations are carried out manually giving more control to the computational expert on which method to choose. The disadvantages of such a non-adaptive scheme arise when:

- The physical computation requires too many approximations.
- The physical computation is too slow or complex.
- The problem is not easily solved by physical rules.
- The input representation has to be pre-processed to be used in the physical computation.

- The output results must be post-processed to be interpreted by the user.

The different approach in "Machine-Learning" (ML) is to use more flexible models that can receive feedback from the environment as they carry out computations so as to adapt themselves to get better results as they "learn" from the environment. In Fig. 3.2, the scheme for such adaptive systems is shown. The Machine-Learning field has deeply impacted our culture and society, especially in those fields where strict physical rules are not available (image recognition, voice dictation, natural language processing, etc.).

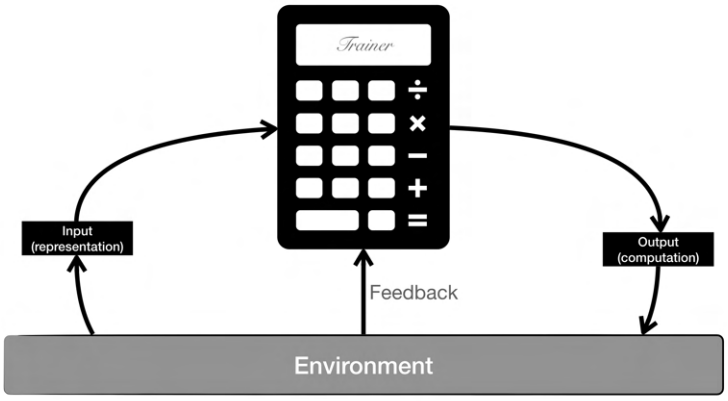


Figure 3.2: The scheme for an adaptive learning system.

The difficulty of applying ML techniques to chemistry arises from the lower degree of control the computational scientist has over the computational engine. In particular, the typical argument against the application of ML techniques to chemistry is the abandonment of physical rules for the computation of observables: physical rules by themselves don't need to receive feedback from the environment since these rules are supposed to describe reality as well as possible by themselves. However, we have already noticed in the previous chapter how, even with the development of Quantum Chemistry, there are many approximations that need to be carried out in order to get reliable results. In other words, having the exact solution "in principle" ($\hat{H}\psi = E\psi$) does not imply having the exact solution "in practice". These approximations may work for some systems and not for others and in general the choice of the right method and basis set has to be done manually by chemical intuition. Not only that but with the advent of Density-Functional Theory (DFT), the problem has shifted to the definition of the right exchange functionals that are essentially unknown, this is a field where ML techniques could help in determining these functionals while still using a reference quantum model (thus maintaining the physical nature of the problem). Moreover, the shift from strict physical rules to more heuristical models has already happened with the advent of Molecular Mechanics and methods based on Force Field whose definitions are often uncertain and based on intuition from the creators of the force field itself.

In general, ML methods are distinguished in:

- Supervised

- Unsupervised
- Semi-Supervised
- Reinforcement Learning

This classification puts emphasis on the presence of a teacher or supervisor that provides a precise measure of the error to the machine, this is usually provided as a training set of couples of inputs and expected outputs. The goal of the ML procedure, in a supervised scheme, is to minimize the difference between the expected outputs and the output its model is computing by means of a loss function. In a supervised scenario, the goal is to build a model that works with the training set but of course, should be robust enough to be extended to unknown inputs, thus it is important to avoid the problem of overfitting the data. There are also other methods in ML that do not necessarily require the presence of a supervisor, these are called unsupervised. There is also the set of "Semi-Supervised" algorithms used in situations when it is necessary to categorize a large amount of data with only a few complete (labeled) examples or when there is the need to put some constraints in the procedure. The last category is the set of "Reinforcement Learning" algorithms where, despite the absence of a supervisor, the feedback is also given by the environment although in a more qualitative and imprecise way, a feedback that is generally called "reward", and is especially used in non-deterministic environments.

In the following sections, a brief summary of the most common algorithms employed in ML techniques is shown without the goal to give an extensive description but just a general introduction.

3.1 The problem of Learnability

In Machine-Learning we often employ flexible parametric models, as described before, so as to be able to receive feedback from the environment (thus simulating the learning process). A parametric model can be split into two parts:

- A static structure
- A dynamic set of parameters

Typically the static structure of the model is immutable (except for algorithms that include a re-modeling phase) while the learning process affects the set of parameters which can vary. In particular, if we consider a set of n parameters we are defining an n-dimensional space, we can then define an "hypothesis" as a particular choice for each of these parameters:

$$H = \{\theta_2, \theta_2, ..., \theta_n\} \tag{3.1}$$

In general, when working in a supervised scenario, we define a custom non-negative error function (e_m) which takes the expected and predicted output values as arguments:

$$E_H = \sum_{i=1}^{N} e_m(\tilde{y}_i, y_i), e_m \geq 0, \forall \tilde{y}_i, y_i \tag{3.2}$$

The goal is to reduce the total error by searching for the best hypothesis H. The most common error function is the Mean Square Error (MSE) which is also called a loss function since it has to be minimized:

$$E_H = \frac{1}{N} \sum_{i=1}^{N} (\tilde{y}_i - y_i)^2 \tag{3.3}$$

Another useful loss function is the zero-one-loss which is useful for binary classifications:

$$L_{0/1H}(\tilde{y}_i, y_i) = \begin{cases} 0, \tilde{y}_i = y_i \\ 1, \tilde{y}_i \neq y_1 \end{cases} \tag{3.4}$$

A helpful interpretation of a generic and continuous loss function can be expressed in terms of potential energy:

$$Energy_H = \frac{1}{2} \sum_{i=1}^{N} e_m(\tilde{y}_i, y_i)^2 \tag{3.5}$$

Just like in the physical situation, the goal is to employ some algorithms to explore the potential energy surface to find the minima that give us the best hypothesis.

In practice, we are being a little naive right now by assuming that it is always possible to find a solution to such optimization problems, but in general, it has to be proven formally that is possible to determine the learnability of a concept given some conditions. In 1984, computer scientist L. Valiant proposed the PAC approach (Probably Approximately Correct) to determine whether a problem is learnable by a computer [47]. In order to simplify its description, let's assume we are dealing with a classification problem where algorithm A has to learn a set of concepts. In particular, a concept is a subset of input patterns X which determines the same output element (which means are classified the same). The learning process, or learning the concept, is the minimization of the loss function as described above. However, given a problem, we may have infinite hypotheses and a probabilistic trade-off is necessary, thus:

An algorithm A can learn the class C of all concepts (making them PAC learnable) if it's able to find a hypothesis H with a procedure $O(n^k)$ so that A, with a probability p, can classify all patterns correctly with a maximum allowed error m_e. This must be valid for all statistical distributions on X and for a number of training samples which must be greater or equal to a minimum value depending only on p and m_e.

3.2 Maximum-likelihood learning

The first attempt at ML comes, of course, from statistics and probabilities. In particular, given a dataset X and a hypothesis h, we can define the likelihood of the hypothesis as:

$$L(h|X) = P(X|h) \tag{3.6}$$

Where $P(X|h_i)$ is the "a posteriori" probability of the dataset X given the hypothesis h. In general multiple hypotheses should be considered:

$$\max_i L(h_i|X) = \min_i \frac{1}{L(h_i|X)} \tag{3.7}$$

And by using probabilities:

$$\min_i \frac{1}{\prod_i P(X|h_i)} \tag{3.8}$$

This can be turned into a simple expression by applying natural logarithms:

$$\max_i \log L(h_i|X) = \min_i -\log L(h_i|X) = \min_i \sum_i -\log P(X|h_i) \tag{3.9}$$

The last term is a summation that can be easily derived and used in most optimization algorithms [48].

3.3 The Feature Space

In the previous section, we talked about data in general terms, without specifying the type of representation of the data employed. In particular, when looking at our dataset X we may represent each point of the dataset as a vector $\vec{x}_i \in X$. These vectors are defined in a vectorial space that is called the Feature Space. The Feature Space represents the space of the problem studied. As a consequence, any change in the definition of the Feature Space brings a change in the numerical representation of each point of the dataset thus changing the quality of the ML procedure. In other words, choosing a Feature Space is the exact problem of choosing the "representation" for our data (see the first chapter for the problem of the representation). The Feature Space can be directly defined from the data if this has a simple numerical form (e.g. in representing images of a dataset we may define an image as an array of (width $\times$ height) numbers each one of which is the (r,g,b) value for the pixel), or we can use smarter descriptors which already encode useful information. The decision of the Feature Space is critical since is all the ML algorithm knows about data: if the Feature Space does not implicitly contain the phenomena investigated, there is no ML algorithm capable of retrieving it. The problem of choosing the correct Feature Space can be reformulated in terms of information theory, by defining "entropy" [49] as:

$$H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{3.10}$$

The definition of entropy relies both on dataset X and on a probability distribution p and is generally measured in bits (due to the logarithm). In general, higher entropies are preferred since it means that the given feature carries more information. For example, let's consider the problem of tossing a coin: our dataset is defined as $X = \{head, cross\}$ and each one of these appears with the same

probability of $p = \frac{1}{2}$, thus resulting in an entropy of $H(X) = 1$. However, if the number of possible outcomes grows (even with the same probability p), the entropy increases because of the summation in its definition. For general Gaussian distributions it is possible to prove that the entropy is proportional to the variance:

$$H(X) = \frac{1}{2}(1 + \ln(2\pi\sigma^2)) \tag{3.11}$$

In other words, a good feature should increase the variance in the dataset. The other typical problem when dealing with the definition of a Feature Space is the "independence" of the features. In principle, each feature should be independent from the others but depending on the situation this is not always the case. It is useful to introduce the "conditional entropy" as:

$$H(X|Y) = -\sum_{x\in X, y\in Y} p(x,y)\log_2 \frac{p(x,y)}{p(y)} \tag{3.12}$$

Thanks to the conditional entropy we can define the mutual information as the amount of information shared by both variables and therefore the reduction of uncertainty about X provided by the knowledge of Y:

$$I(X;Y) = H(X) - H(X|Y) \tag{3.13}$$

In principle, when X and Y are independent, they don't share any information and this is easily proved by taking $p(x,y) = p(x)p(y)$ for independent distributions, obtaining $H(X|Y) = H(X)$ thus $I(X;Y) = 0$. The opposite situation is in having close to 0 conditional entropies (which means that Y is able to describe X quite well), then the mutual information becomes $I(X;Y) = H(X)$.

The choice of a Feature Space is nontrivial, and many descriptors in cheminformatics have been introduced to describe molecular datasets (e.g. QSAR [50], etc.). The truth is that depending on the type of phenomena investigated, different Feature Spaces are required. Moreover, a common approach is to combine different ML algorithms in multiple steps so as to let the machine decide which Feature Space best represents the dataset. As an example, starting from an initial numerical dataset an ML model computes a suitable Feature Space and another ML model computes the desired quantities on this Feature Space. After a Feature Space has been selected, this is usually normalized and the dataset is split into a Training set and a Test set. In this way, it is possible to get an estimate of the correctness of the ML algorithm by checking the prediction accuracy on the Test set while training on the Training set. Of course, both sets must reflect the original data distribution.

3.3.1 PCA

In the previous section, we stressed the importance of identifying features that have low entropy (do not provide meaningful information) or features that share too much information thus being redundant. It is generally good practice before starting the training process to filter out those features that do not provide useful information and a common way to do it is by means of the Principal Component Analysis (PCA) [51]. The PCA is also useful in those situations when we have a

very high number of variables and we want to reduce them. The general idea of the PCA is to assess how much information is brought by each component, and the correlation between them, by building a covariance matrix:

$$\boldsymbol{C} = \begin{bmatrix} \sigma_1^2 ... \sigma_{1m} \\ ..... \\ \sigma_{m1} ... \sigma_m^2 \end{bmatrix} \tag{3.14}$$

Where:

$$\sigma_{ij} = \frac{1}{m} \sum_k (x_{ki} - E[X_i])(x_{kj} - E[X_j]) \tag{3.15}$$

C is symmetric and positive semidefinite so all the eigenvalues are non-negative. The interpretation of each eigenvalue can be the "weight" that the relative feature has in describing the dataset. Thus, by ordering the eigenvalues and selecting only the first above a certain threshold it is possible to rebuild, using the corresponding eigenvectors, a sub-space whose dimension is lower than the dimension of the original one. In this way, we are not only reducing the dimensionality of the problem but we are also projecting the dataset in a new sub-space of features.

3.3.2 Linear-Regression

It is time to talk about ML procedures by starting with the regression models [48], in particular with the Linear Regression model due to its simplicity and historical relevance. The problem of regression is to obtain continuous values ($Y = \{y_1, y_2, ..., y_n\}, y_i \in \mathbf{R}$) from the dataset ($X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}, \vec{X}_i \in \mathbf{R}^m$). The simplicity of the linear model is to use a hyperplane to describe the behavior of the target quantity:

$$y = a_0 + \sum_{i=1}^{n} a_i x_i, \;\; A = \{a_0, a_1, ..., a_n\} \tag{3.16}$$

The advantage of Linear Regression models is that they allow to treat also some non-linear situations such as the polynomial regression, where the trick is to increase the dimensionality of the Feature Space by including nonlinear terms:

$$\bar{x} = (x_1, x_2) \longrightarrow \bar{x}_t = (x_1, x_2, x_1^2, x_2^2, x_1 x_2) \tag{3.17}$$

This is yet another example of how the right definition of a Feature Space simplifies the ML model. More complex models (such as Supporting Vector Machines or Neural Networks) should be considered when the underlying phenomena investigated is intrinsically non-linear. Some variations of the Linear Regression are shown below.

Ridge

The Ridge regression [52] imposes an additional shrinkage penalty to the ordinary least squares loss function:

$$L(\bar{w}) = ||X\bar{w} - \bar{y}||_2^2 + \alpha ||\bar{w}||_2^2 \tag{3.18}$$

The introduction of a weight vector has to be kept under control by the additional term multiplied by the coefficient α that avoids the uncontrolled growth of the weights.

Lasso

The Lasso [53] is conceptually similar to the Ridge but imposes an L1 norm on the weights:

$$L(\bar{w}) = \frac{1}{2n}||X\bar{w} - \bar{y}||_2^2 + \alpha||\bar{w}||_1 \tag{3.19}$$

The shift to the L1 norm is to allow a potentially higher number of null coefficients.

Elastic-Net

The Elastic-Net [54] model tries to combine together the advantages of the Ridge and Lasso by including both L2 and L1 norms, thus resulting in a model sparse like a pure Lasso but with the same regularization ability as provided by Ridge:

$$L(\bar{w}) = \frac{1}{2n}||X\bar{w} - \bar{y}||_2^2 + \alpha\beta||\bar{w}||_1 + \frac{\alpha(1-\beta)}{2}||\bar{w}||_2^2 \tag{3.20}$$

3.4 Classification

The problem of classification in Machine-Learning is to train the machine to classify data in classes. As an example, if we define two classes A and B, the machine should tell whether some data is of class A or class B. In molecular sciences a good example of this problem could be the automatic assignment of atom types to atoms, providing a good Feature Space, so as to classify each atom to its corresponding type.

3.4.1 Linear-Classification

The first method we discuss is a linear method so, given two classes A and B for simplicity, it tries to find the optimal hyperplane that separates the two classes. In multi-class problems, the reasoning remains identical. So given our dataset:

$$X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}, \vec{x}_i \in \mathbf{R}^m \tag{3.21}$$

We have a target set for the classification:

$$Y = \{y_1, y_2, ..., y_n\}, y_i \in \{0, 1\} \tag{3.22}$$

By defining a weight vector:

$$W = \{w_1, w_2, ..., w_m\}, w_i \in \mathbf{R} \tag{3.23}$$

We can define the quantity z:

$$z = \vec{x}\vec{w} \tag{3.24}$$

So that if $\vec{x}$ is variable, z is the value determined by the hyperplane equation. Therefore, the optimization procedure regards the w coefficients so that the classification holds on:

$$sign(z) = \begin{cases} +1, x \in A \\ -1, x \in B \end{cases} \tag{3.25}$$

3.4.2 Logistic Regression

The Logistic Regression approach, although it's called regression, is a classification method based on the "probability" of a sample belonging to a class. It can be used, as an example, in determining whether to recommend cesarean delivery [55]. The starting point is still the Linear Classification but instead of using the sign of the z value directly, we use a sigmoid function that we can interpret as the probability for the sample to belong to a class:

$$\sigma(\vec{x}; \vec{w}) = \frac{1}{1 + e^{-\vec{x}\vec{w}}} \tag{3.26}$$

At this point, finding the optimal parameters means maximizing the log-likelihood as shown in Sec. 3.2:

$$L(\vec{w}; y) = \sum_i \log P(y_i | \vec{x}_i \vec{w}) \tag{3.27}$$

Therefore we need to minimize the loss function:

$$j(\vec{w}) = -L(\vec{w}; y) = -\sum_i (y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i))) \tag{3.28}$$

Stochastic Gradient descent algorithms

It is just worth noting that the optimization of the weights $\vec{w}$ for the classification problem can be done with many other algorithms. As an example, the idea behind the stochastic gradient-descent (SGD) is to iterate over the weights so as to move in the opposite direction of the gradient of the loss function:

$$\vec{w}(k+1) = \vec{w}(k) - \gamma \nabla L(\vec{w}) \tag{3.29}$$

The procedure is applied to batches randomly extracted from the overall dataset.

3.4.3 Naive Classifier

The term naive is not because these algorithms are limited or less effective, but is due to an assumption that we are going to discuss later. The starting point for such classifiers is the Bayes theorem of conditional probabilities defined as:

$$\begin{cases} P(A \cap B) = P(A|B)P(B) \\ P(B \cap A) = P(B|A)P(A) \end{cases} \tag{3.30}$$

Thus the Bayes theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.31}$$

Let's suppose we are testing whether a "feature" is accurate enough in classifying correctly some data (e.g., assuming that if an email has a number of characters below 50 it can be classified as Spam). The A-Posteriori probability ($P(Spam|text < 50\ char)$) behaves like a likelihood for our theory (see previous sections). The denominator is usually less important since our goal is to maximize/minimize functions so:

$$P(A|B) = \alpha P(B|A)P(A) \tag{3.32}$$

The problem arises when there are multiple concurrent conditions:

$$P(A|C_1 \cap C_2 \cap ... \cap C_n) \tag{3.33}$$

This makes things more complicated but the assumption of the Naive Classifier (hence the name naive) is to assume conditional independence of causes that is:

$$P(A|C_1 \cap ... \cap C_n) = \alpha P(C_1|A)...P(C_n|A)P(A) \tag{3.34}$$

In general, the conditional independence of causes is rarely true (for example, if an email has a number of characters below 50 it can increase the probability of finding an image thus these two conditions are not independent). However, it usually behaves well even when the naive condition is violated [56].

In order to classify an input vector $\vec{x}$ into one of the classes y_i, the a-posteriori probability is computed for each class and the higher determines the assignment:

$$P(y_i|x_1, x_2, ..., x_n) = \alpha P(y_i) \prod_i P(x_i|y_i) \tag{3.35}$$

The probabilities are obtained by frequency counting.

3.4.4 Supporting Vector Machines

Together with Neural Networks, Support Vector Machines (SVM) [57] are usually the best choice when a linear hyperplane is not possible to be found in a classification task. The starting point for discussing SVM is with the usual Hyperplane in the linear case for simplicity. In reality, for a normalized set of data is possible to define two boundary hyperplanes containing only a few elements (the support vectors) as shown in Fig. 3.3.

The goal is to maximize the distance between these two hyperplanes so as to avoid misclassification due to more overlap between the two classes. In fact, there are multiple choices for a hyperplane to divide the two classes, but just one optimal solution that maximizes the distance between the two boundaries. In this case, the two boundaries are parallel hence the distance between them is a multiple of the hyperplane $\vec{w}$ vector:

$$\vec{x}_2 - \vec{x}_1 = t\vec{w} \tag{3.36}$$

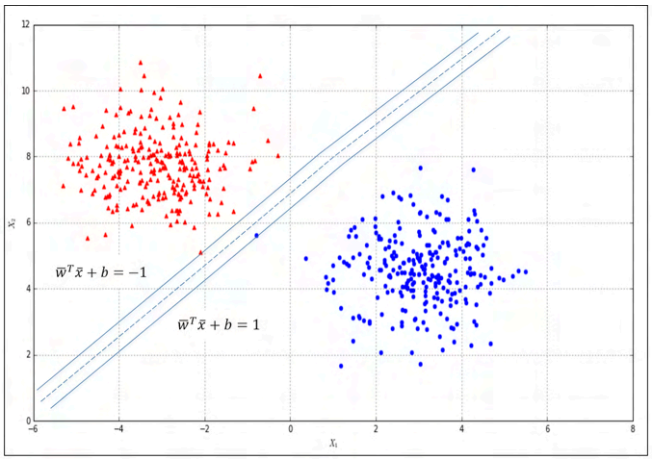


Figure 3.3: The two boundary hyperplanes for a normalized set of data.

by connecting two points between each boundary. Now, considering the boundary hyperplane equation we get:

$$\vec{w}^T\vec{x}_2 + b = \vec{w}^T(\vec{x}_1 + t\vec{w}) + b = (\vec{w}^T\vec{x}_1 + b) + t||\vec{w}||^2 = 1 \tag{3.37}$$

The first term of the last part is equal to -1, thus by solving for t we get:

$$t = \frac{2}{||\vec{w}||^2} \tag{3.38}$$

Thus the distance between $\vec{x}_1$ and $\vec{x}_2$ is:

$$d(\vec{x}_1, \vec{x}_2) = \frac{2}{||\vec{w}||} \tag{3.39}$$

Moreover, by imposing $\{-1, 1\}$ as labels for the two classes, we can write the following constraint for each point of the dataset:

$$y_i(\vec{w}^T\vec{x}_i + b) \geq 1, \forall(\vec{x}_i, y_i) \tag{3.40}$$

The Kernel Trick

In discussing non-linear problems the general approach is the same as discussed for the Linear Regression model where the Feature Space has increased in dimensions including non-linear terms. In the case of SVM, however, further considerations must be done. The SVM worked by optimizing the following two equations:

$$\begin{cases} \min \frac{1}{2}||\vec{w}|| \\ y_i(\vec{w}^T\vec{x}_i + b) \geq 1 \end{cases} \tag{3.41}$$

By applying Lagrange Multipliers, and taking $||\vec{w}||^2$ instead of its square root:

$$L(\vec{w}, b, \alpha) = \frac{1}{2}\vec{w}^T\vec{w} + \sum_i \alpha_i(1 - y_i(\vec{w}^T\vec{x}_i + b)) \tag{3.42}$$

And by taking the corresponding derivatives we discover that the set of parameters $\vec{w}$ is now dependent on the new set of Lagrange multipliers α_i, thus having to optimize the following:

$$\begin{cases} \max\left(\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \vec{x}_i^T \vec{x}_j\right) \\ \sum_i \alpha_i y_i = 0 \end{cases} \tag{3.43}$$

As said before, the trick to go from linear to non-linear space is to add dimensions ($\vec{x}_i \longrightarrow \Phi(\vec{x}_i)$) but this comes at the cost of expensive computations (the number of dimensions rises considerably and even the dot product $\Phi(x)^T\Phi(x)$ becomes troublesome). However, the advantage of the SVM is to use the Kernel trick to express its dot products. In particular, there are special functions (called kernels [48, 58]) that have the nice property:

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i)^T \Phi(\vec{x}_j) \tag{3.44}$$

These Kernels reduce the complexity and make the SVM a good candidate for non-linear problems. Here are some examples of Kernel functions, such as the radial basis function:

$$K(\vec{x}_i, \vec{x}_j) = e^{-\gamma|\vec{x}_i - \vec{x}_j|^2} \tag{3.45}$$

The polynomial kernel:

$$K(\vec{x}_i, \vec{x}_j) = (\gamma \vec{x}_i^T \vec{x}_j + r)^c \tag{3.46}$$

And the sigmoid Kernel:

$$K(\vec{x}_i, \vec{x}_j) = \frac{1 - e^{-2(\gamma \vec{x}_i^T \vec{x}_j + r)}}{1 + e^{-2(\gamma \vec{x}_i^T \vec{x}_j + r)}} \tag{3.47}$$

3.4.5 Decision trees

The last category of ML methods for classification tasks that we are going to discuss is Decision Tree [59]. Even if these methods are not used a lot in complex classification tasks, they have the great advantage of giving an easy representation of the "chain of thoughts" of an ML algorithm since, as the name implies, they represent a chain of decisions done sequentially. In particular, given an input dataset X:

$$X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}, \vec{x}_i \in \mathbf{R}^m \tag{3.48}$$

We have m features for each point in the dataset. The binary decision process is simply defined by a threshold for each feature: if the feature is below the threshold a choice is made, otherwise, a different choice is done as shown in Fig. 3.4.

The learning process impacts the thresholds t at every binary decision step thus changing the "chain of thoughts" of the algorithm. The leaf nodes are the classes that we want to assign. In defining the "structure" of a binary decision

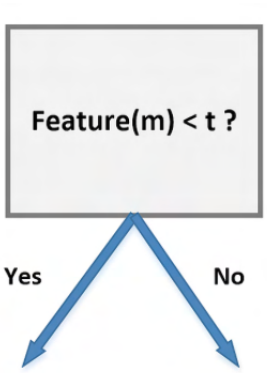


Figure 3.4: The binary decision process

tree it is important to make sure that each child node (the result of a decision) contains less information than the parent node (the decision): in other words, the entropy should decrease. In fact, let's suppose we have two categories to assign to each point in the dataset: (A, B). In this case, at the starting root node, we will have higher uncertainty about whether the specific point is either A or B. As we start to make decisions, we reduce the uncertainty until we reach the leaf node A or B.

As an example, in Fig. 3.5 two choices for an atom-type classification decision tree are shown.

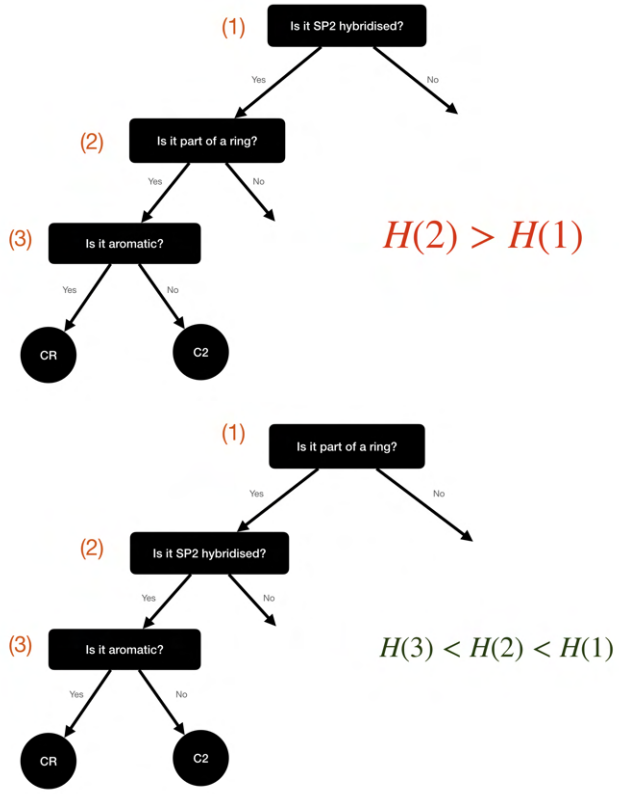


Figure 3.5: Two examples of binary trees for atom type classification, the first resulting in an increase of entropy while the second resulting in a decrease of entropy.

In the first choice, we get an increase in entropy since the set of atoms that are part of a ring contains both SP3 and SP2 atoms: thus, is a bigger set than simply the set of all SP2 atoms. A better choice for the decision tree is the latter where the entropy is always decreasing. This concept can be further formalized in terms of "impurity". To do so, let's consider that each selection node defines a subset of elements that satisfy its condition. For example, the set of all atoms that are part of a ring in the first tree of Fig. 3.5. Each node is defined by the tuple $\sigma = \langle i, t_k \rangle$ where i is the index of the feature (is it in a ring or not) while t_k is the selection threshold (for general continuous features). We can now define the total impurity for a selected node as:

$$I(D, \sigma) = \frac{N_{left}}{N_D} I(D_{left}) + \frac{N_{right}}{N_D} I(D_{right}) \tag{3.49}$$

Where D is the whole dataset entering the selected node, D_{left} and D_{right} are the resulting datasets after applying our decision. Of course, there are many possible definitions of impurity indices that satisfy the above relation. Here we are listing just some of them such as the Gini Impurity Index:

$$I_{Gini}(j) = \sum_{i}^{C} p(i|j)(1 - p(i|j)) \tag{3.50}$$

Where C is the total number of classes and $p(i|j)$ is the ratio between the total number of samples belonging to class i and the total number of samples of the selected node j. Of course, as said before, the most common interpretation of impurity is in terms of entropy in information theory thus introducing the Cross-Entropy Impurity Index:

$$I_{Cross-entropy} = -\sum_{i}^{C} p(i|j) \log p(i|j) \tag{3.51}$$

The impurity is not only important in deciding the "shape" of a tree but also in deciding the "importance" of a feature in describing a decision process. The definition of the importance of a feature is the following:

$$Importance(x_i) = \sum_{k} \frac{N_k}{N} \Delta I_{x_i} \tag{3.52}$$

The sum is extended to all nodes that use the feature x_i and N_k is the number of samples that reach that node while quantifying the change in impurity due to feature x_i.

Random Forests

The Random Forest algorithm [60] is an extension of the single binary tree algorithm to optimize in parallel different binary trees defined on different subsets of features. In order to perform the classification, a voting approach is most commonly used where the most voted class by the trees is considered. The importance of a feature becomes:

$$Importance(x_i) = \frac{1}{N_{Trees}} \sum_{t} \sum_{k} \frac{N_k}{N} \Delta I_{x_i} \tag{3.53}$$

3.5 Clustering

The problem with Classification was to let the machine learn a pattern to associate points to certain labels (it was a supervised approach since we were responsible for giving the right reference labels for the training data to the machine). The Clustering approach [48], however, is unsupervised and tries to automatically group together points that are similar to each other using some sort of similarity metric. More formally, if we have to deal with the same dataset X:

$$X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}, \vec{x}_i \in \mathbf{R}^m \tag{3.54}$$

We assume it is possible to find a nonunique criterion so that each sample can be associated with a specific group:

$$g_k = G(\vec{x}_i), k = \{0, 1, ..., t\} \tag{3.55}$$

Each group is called a cluster and the process of finding G is called clustering. In general, hard-clustering techniques are by far the most common and require that each point is assigned to one cluster only. The opposite of that, one point can be part of two different clusters with different weights, is called soft/fuzzy clustering, and is not treated in the following.

3.5.1 K-means

The K-means method [61, 62] is the most common method employed in clustering procedures. However, it requires the number of clusters to be given as an input. The algorithm works by initially assigning k initial centroids in random positions (the k-means++ variant uses a complicated mathematical formulation that selects the initial centroids so that they are statistically close to the final ones):

$$K^{(0)} = \{\vec{\mu}_1^{(0)}, \vec{\mu}_2^{(0)}, ..., \vec{\mu}_k^{(0)}\} \tag{3.56}$$

Then, the inertia of the i-th centroid is defined as follows:

$$SS_{w_i} = \sum_t ||\vec{x}_t - \vec{\mu}_i||^2, \forall i \in (1, k) \tag{3.57}$$

At the start, each point of the dataset is assigned to the cluster of the closest centroid. Then, at the next iteration, the centroid is recalculated with the new members of the cluster. With the new centroids recomputed, the data points are assigned again, and so on until we reach convergence (that is, the centroids don't change significantly anymore). This is analogous to saying that we are minimizing the inertia of each centroid. A variant of the K-Means algorithms is the PAM (Partition Around Medoids) [63] instead of getting the mean point between the data in order to compute the centroid, this is taken as a medoid or the closest point of the dataset to the mean. The biggest limitation of the K-Means method is its use of the Euclidean radial distance that works well with convex data (that is data that has a tendency to form "blobs"), but in general, if the data show some different behavior other algorithms should be used. In most applications, K-Means is still the best option. The only real-life difficulty in employing the K-Means method is having to choose in advance the number of

clusters to use. In the following, several scores and indices are introduced that can help in evaluating the optimal number of clusters.

Inertia score

The first score that is reasonable to look for is the inertia score that we have already introduced since the goal of the K-Means is to minimize the inertia. The problem with the inertia score, however, is that it only reaches 0 when each point of the dataset forms a cluster on its own which is not optimal. This score is often interpreted with the so-called "elbow rule" [48], which essentially means looking for the number of clusters that have drastically reduced the inertia with respect to the previous one while not changing a lot in subsequent clusters.

Silhouette score

The idea of the Silhouette [64] is based on the principle of increasing the internal cohesion between the points of a cluster and increasing the distance between the clusters themselves. After defining a distance metric (Euclidean) we can compute the average intracluster distance for each element:

$$\alpha(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i, j) \tag{3.58}$$

We can also define the average nearest-cluster distance (that is the lowest intercluster distance):

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j) \tag{3.59}$$

Then the silhouette is defined as:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{3.60}$$

It is a value defined between -1 and 1, 1 is optimal, and 0 means there is a cluster overlap. A value close to -1 means that the sample has been assigned to the wrong cluster.

Calinski-Harabasz index

The Calinski-Harabasz index (CH) [65] is also based on the concept of dense and well-separated clusters. First, we define the inter-cluster dispersion matrix as:

$$\boldsymbol{B} = \sum_t n_t (\vec{\mu} - \vec{\mu}_t)(\vec{\mu} - \vec{\mu}_t)^T \tag{3.61}$$

Where n_t is the number of points belonging to cluster t, $\vec{\mu}$ is the total centroid and $\vec{\mu}_t$ is the centroid of the t-th cluster. The intracluster dispersion matrix instead can be written as:

$$\boldsymbol{X} = \sum_t \sum_{x \in C_t} (\vec{x} - \vec{\mu}_t)(\vec{x} - \vec{\mu}_t)^T \tag{3.62}$$

For each datapoint of the cluster x. Then the CH index is defined as:

$$CH(k) = \frac{N-k}{k-1}\frac{tr\{\boldsymbol{B}\}}{tr\{\boldsymbol{X}\}} \tag{3.63}$$

Also, in this case, the goal is to maximize such an index so as to maximize the inter-cluster dispersion and minimize the intracluster dispersion.

3.5.2 DBSCAN

In those situations where the dataset shows a non-convex behavior, K-Means fails. A common alternative is the Density-Based Spatial Clustering of Applications with Noise (also called, DBSCAN [66]). The idea is actually simple: a cluster is defined as a high-density area (with no restrictions on its shape) surrounded by a low-density area. The procedure starts by analyzing a small area (formally a point surrounded by a minimum number of other samples). If the density is high enough, this point is considered as part of a cluster. At this point, his neighbors are taken into account and if they also are in a region of high local density they are merged with the first area. If they don't have an equally high density, they determine a topological separation. When all areas have been scanned, the clusters are automatically assigned because they are islands surrounded by empty space. It is important to notice that in the DBSCAN approach, the number of clusters is not required in advance, since these are automatically detected by checking the density.

3.5.3 Spectral clustering

A more sophisticated approach consists of building a symmetric affinity matrix A whose elements a_{ij} determine the "affinity" between two samples. The choice of the Kernel function in this case is the choice for the affinity measurement (usually radial basis functions). The matrix is diagonalized and the clustering procedure is applied to a subset of eigenvectors (each spectral clustering variant has its own procedure). This is conceptually similar to the PCA approach but works with affinity matrices instead of covariance matrices.

3.5.4 Hierarchical Clustering

The idea of hierarchical clustering approaches is to find a hierarchy of partial sub-clusters that can be assembled together in bigger clusters. The agglomerating cluster approaches use a bottom-up approach in building up bigger and bigger clusters, while the divisive clustering approaches use a top-down approach by splitting big clusters into smaller pieces. In general, the agglomerating clustering approach is preferred to the divisive one for better performance.

Agglomerating clustering

The agglomerating clustering approach requires the definition of a metric in the Feature Space which defines automatically an affinity between data points. Once an affinity is defined, the next step is defining a linkage that is a criterion to aggregate different clusters. There are many possibilities [67]:

- Complete Linkage, where for each pair of clusters the algorithm computes and merges them to minimize the maximum distance between the clusters (the distance of the farthest elements).
- Average Linkage, where instead of the maximum distance between the clusters the average is used.
- Ward's linkage, that computes the sum of the squared distances within the clusters.

A common approach to visualize the agglomeration process in action is through dendrograms, special plots that show the agglomeration process starting from each individual data point up to the final number of clusters.

3.6 Deep Learning

The most recent evolution in Machine-Learning methods is generally indicated by the term "Deep" Learning, which is usually associated with the increased number of data available and the increased complexity of models. Deep Learning is just a different flavor of traditional Machine-Learning. In previous sections, we stressed the importance of the Representation of data, which is critical both in developing a physical model of phenomena and in performing ML algorithms. In particular, in the case of ML algorithms, we said that the definition of the Feature Space must correctly represent the data and the underlying phenomena of interest since that is all the machine knows about the problem. Deep Learning techniques usually go a step further by taking a different approach to the problem of the representation by developing models that not only learn from numerical data but can also optimize their representation (the Feature Space) to get even better results. The clear advantage of such methods is that we don't have to get extremely accurate features since the model tries to also optimize them, and this has led to a wider application of ML to more complex problems (e.g. language processing, vision, sound recognition, artificial intelligence, etc.). The clear disadvantage of deep learning approaches is in scientific research, where the desire is not only to get the right numbers but also an "interpretation" for their computations (a human-understandable representation). In practice, though, there are many applications of deep learning in science due to the practical limitations of human resources. Imagine, for example, having to search for all possible conformers of a molecule. In theory, a human scientist could try to guess them manually, optimizing the structure for each of those conformers hoping to find local minima of the potential energy function. In small molecules, this is generally not a huge problem, but with bigger molecules, the search becomes unfeasible for a human researcher. The truth is: machines are good at big numbers. In the case of conformer research, for example, employing a deep learning model can decrease research time by automatically searching for minima structure while still allowing the physical interpretation of a conformer that at that point can be manually checked. The field of deep learning is yet another gigantic field of research that requires a separate book to be fully described. In the following, we are just giving some basic description of two commonly used deep-learning algorithms: the

genetic algorithms (GA), and the neural networks (NN). It is interesting to notice how many deep-learning algorithms take inspiration from biology, trying to replicate biological phenomena to imitate intelligent behavior at a deeper level (evolution in the case of the genetic algorithms, the human brain in the case of neural networks).

3.6.1 Genetic Algorithms

John Holland developed genetic algorithms (GAs) [68, 69] in the 1960s. They are algorithms based on natural selection and natural laws of genetics, which aim to solve optimization problems. These algorithms have the following iterative process to find the optimal solution:

- Properly represent the encoding of the problem. Most of the problems use binary encoding.
- Evaluate each individual with a fitness function or target function, which determines the value or performance of each solution.
- Choose a configuration selection strategy, which will be in charge of the construction of the new population (new generation).
- Choose a mechanism to implement the genetic crossover operator.
- Build a mechanism to implement the genetic mutation operator.

It is obvious from the previous points how biology inspired the development of GAs: each individual of a population is evaluated towards a target function that represents its biological fitness (higher is better, and the evolution process goes to higher fitness). At each step of the evolution process, a new population is created by the previous one by performing genetic crossover and mutation over each individual and selecting, through a certain strategy, those individuals who can survive. Each individual in the population is described by a set of chromosomes:

$$I = (c_1, c_2, ..., c_n) \tag{3.64}$$

In general, each chromosome is a set of genes, each one of which can have different values defined as alleles. Each individual in this population is evaluated against a fitness function $f(I)$, which is then normalized. Normalization means dividing the fitness value of each individual by the sum of all fitness values so that the sum of all resulting fitness values equals 1. Then, Selection is performed by employing different methods dependent on the value of these normalized features (e.g. Roulette Wheel Selection, Rank Selection, etc.). At the end of the Selection phase, a sub-set of individuals remains. To generate a new population, genetic operators are applied. The two genetic operations are:

- Crossover: This operator swaps the genetic information of two parents to reproduce an offspring. It is performed on parent pairs that are selected randomly to generate a child population of equal size to the parent population.

- Mutation: This operator adds new genetic information to the new child population. This is achieved by flipping some bits in the chromosome. Mutation solves the problem of local minimum and enhances diversification.

The new population is then evaluated again in an iterative process until a given number of evolutionary iterations are performed or some heuristic of convergence is satisfied. The example of the conformer search, given in the introduction to deep learning section, has been done through a genetic algorithm procedure where the genes are the set of torsion angles for a molecule and the fitness is for the resulting molecule to describe a local minima (that is a conformer). The most important choices that must be made when applying a GA are the types of selection + crossover and mutation. As an example, in the case of the conformer search application, we employed the tournament selection (with a tournament size of 2) to ensure a balance between the diversity and fitness of parents and then switch to elitism for the last 5 percent of planned generations if the search has not yet stopped. For the former choice, one possibility is to interpolate the alleles with the simulated binary crossover (SBX) approach [70], which employs the so-called β factor, defined in terms of a uniformly distributed random number μ and a spreading factor η (the latter is proportional to how much offspring alleles will resemble those of the parents):

$$\beta = \begin{cases} \dfrac{1}{2\mu^{\eta+1}}, \mu \in [0, 0.5] \\ \dfrac{1}{2}(1-\mu)^{\frac{1}{\eta+1}}, \mu \in [0.5, 1] \end{cases} \tag{3.65}$$

In a second step β is employed to interpolate the parent's coordinates:

$$\begin{cases} C_1 = 0.5[(1+\beta)P_1 - (1-\beta)P_2] \\ C_2 = 0.5[(1+\beta)P_2 + (1-\beta)P_1] \end{cases} \tag{3.66}$$

Here, P1 and P2 (i.e. parent 1 and parent 2) are the actual specimens mating (that is, P1 and P2 coordinates will be always mixed), whereas C1 (child 1) and C2 (child 2) are the corresponding offspring. A simple constant probability method is used to check if a specimen was to be mutated and then to uniformly select a gene.

The $(\lambda + \mu)$ model

In the $(\lambda + \mu)$ evolutionary algorithm (EA) [71], at each generation μ parents generate λ offspring; then survival occurs and the population size is reduced back to μ. In the implementation of the model for a conformer search application (see Chap. 8 for details), the selection rate (s) parameter was introduced, i.e., the number of new offspring that will be created at each generation; $\mu/2$ pairs of existing specimens always generate $\lambda/2$ pairs of different offspring i.e. a unitary λ/μ ratio was employed and $\lambda = s \cdot P$ where s is the selection rate and P the population size. In other words, the population size P becomes $(1+s) \cdot P$ when offspring are generated, and it is shrunk back to P when the worst s specimens (parents and offspring) are eliminated. The rationale behind the choice of this specific method is related to the high cost of evaluating the fitness of a new

individual, which in the case of the conformer search implies a (costly) electronic structure calculation: high-fitness individuals are then worth being preserved in the population until some really improved individual is found.

The island model

The island model [72] is another variant of a GA in which the operators (competition, selection, survival, and reproduction) act separately on suitable subpopulations (islands), which are mixed only at predefined intervals by a dedicated operator (migration). The underlying idea is that for flexible systems the positions of atoms belonging to different moieties can to some extent be relaxed separately (in the GA language these would correspond to low-order nonrelated schemata).

The "hall of fame"

In some cases, the fitness evaluation of an individual is a time-consuming step of the evolution, and the cost connected to the disruption of a promising specimen is high. For this reason, a new feature, known as "hall of fame" [73], is introduced which transmits a fraction of the best individuals' $h \cdot P$ to new generations inhibiting any mutation. The new population size is then $(1 + S) \cdot P + h \cdot P$ (S is the selection pressure, P is the population size, and h is the hall of fame size) before survivor selection, when it is shrunk to P.

3.6.2 Neural Networks

Although Neural-Networks (NNS) are still state-of-the-art ML algorithms, they are not a very recent invention. In fact, the first perceptron was invented in 1943 by McCulloch and Pitts [74] (Fig. 3.6).

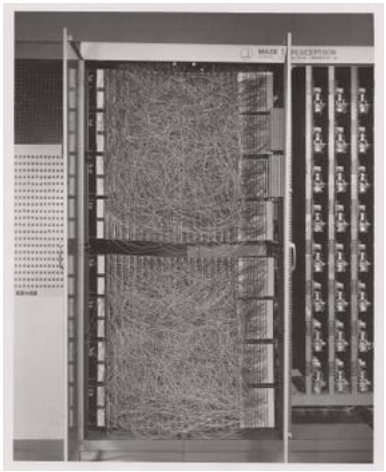


Figure 3.6: The first perceptron (1943) [75]

Their implementation of the perceptron was not in software but in hardware: This machine was designed for image recognition. It had an array of 400 photocells, randomly connected to the "neurons". Weights were encoded in potentiometers, and weight updates during learning were performed by electric motors.

Today, the modern representation of a perceptron is a mathematical one and is shown in Fig. 3.7, this is then encoded in software and not in hardware.

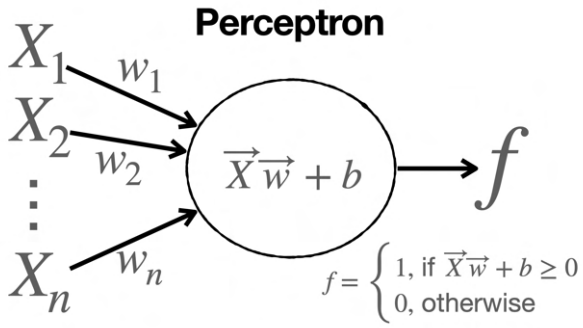


Figure 3.7: The modern representation of a perceptron.

In general, it acts as a binary classifier where the output can either be 0 or 1 depending on the set of weights: (w_i, b) and the input values x_i. Thus, a perceptron model acts as a supervised learning model where a training set of data is given to the machine (x_i) together with the supervised classification data, and the weights are optimized with a minimization procedure (most commonly Stochastic Gradient-Descent) so as to better fit the data. This minimization step is what is usually referred to as the learning step since the algorithm learns the correct weights. The historical inspiration for the perceptron model came from biology, in particular the neuron. In biology, a neuron is a particular cell capable of taking some input signals and if the sum of the electric signals goes above a given "activation value" then the neuron fires a signal out. The activation value in the perceptron model is represented by the b parameter. The desire of ML scientists to replicate in software and math what the human brain does so as to create "real" intelligent behaviors is not really the focus of the research in ML anymore. Although there are still many research fields (especially in biology and psychology) that try to understand more about the way the human brain works through the use of ML and NNS, the focus of computer scientists is on improving the prediction skills of such algorithms even if that means departing from a strictly biological model. In the end, the human brain acted as an inspiration for NNS as much as evolution acted as an inspiration for genetic algorithms. In recent years, huge neural networks have been developed with a number of neurons that is approximately the one of the brain of a mouse, also thanks to the advancement in hardware. The critical insight to understand is that "intelligence" is not just about the number of neurons or the size of the brain, but is a complex behavior that arises from the quality of the connections between units. Unfortunately, very little is known in biology about how these neurons form a network in the brain, and even if we know there are areas of the brain dedicated to specific tasks we know very little about the neuron-by-neuron connections required to simulate a brain. In our case, our interest in neural networks really comes from a mathematical problem: solving non-linear regression. In previous sections we talked about linear regression and how can be used to describe even non-linearity by employing some Kernel function (SVM) or enlarging the Feature Space with non-linear terms. NNS provides an elegant solution to the non-linear regression problem. In fact, although the single perceptron performs a linear combination

of parameters and values ($\vec{X}\vec{w} + b$), the real magic comes in combining multiple layers of neurons. In Fig. 3.8, an example of a simple neural network and a "deep" network is shown.

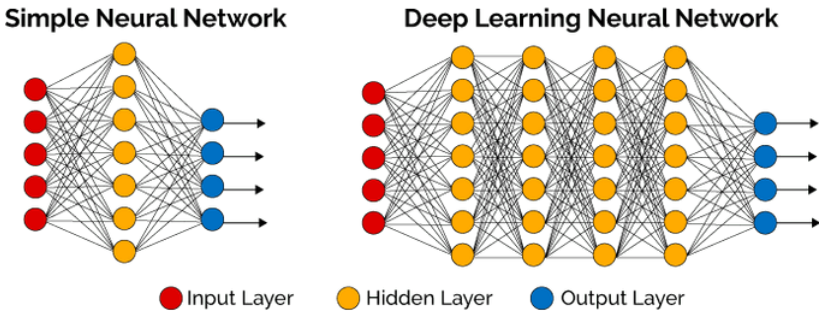


Figure 3.8: A simple and a "deep" neural network.

Neurons are grouped together in layers, the first layer is the input layer that receives the initial data $\vec{x}$, and the last layer is the output layer that provides continuous outputs. Neurons in the middle are said to be placed in hidden layers and now the definition of "deep learning" should appear clearer since is related to the complexity of the model and the number of layers. In the introduction to deep learning, we also said that the Feature Space is automatically optimized by the model itself and that is still true in NNS. In fact, each hidden layer extrapolates different information from the input data thus representing the data internally in a way that allows him to replicate the correct outputs. The perceptron model was a binary classifier, to allow NNS to describe continuous outputs which are important for regression a sigmoid function is usually introduced:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.67}$$

The source of non-linearity thus comes from the output of a neuron that is usually: $\sigma(\vec{X}\vec{w} + b)$. The sigma function is the most commonly used for its nice behavior (it ranges from 0 to 1) and the nice property of its derivatives which simplify computations (especially with gradient descent optimizations):

$$\frac{\delta\sigma}{\delta x} = \sigma(1 - \sigma) \tag{3.68}$$

The gradient required to perform Stochastic Gradient-Descent is computed with a backprogagation algorithm [76] based on the chain rules of derivatives. In fact, since each neuron takes the output of the previous layer as an input, is like combining multiple functions in a single one thus creating a chain of functions:

$$output = f_n(f_n - 1(f_n - 2(....f_3(f_2(f_1(\vec{X})))...))) \tag{3.69}$$

In conclusion, NNS is a great tool to perform non-linear regression whenever the source of non-linearity is unknown or an explicit equation is not available.

3.7 Conclusions

In this chapter, we discussed the main Machine-Learning (ML) methods and procedures giving a general introduction to the field. In this context, the importance

of building the right Feature Space is crucial for ML algorithms as much as it is crucial for the representation and visualization of data. As a consequence, in the next chapter, we are going to introduce the field of Molecular Perception which tries to derive meaningful chemical descriptors for molecular systems starting from the minimum amount of information available.

Chapter 4

Molecular Perception

The terrific improvements in hardware and software in the last decades have transformed Quantum Mechanics from a specialist domain to a general tool employed by researchers to complement experimental studies of systems and processes of increasing complexity from material science to drug design. At the same time, the chemistry language is now well established and formulated using a fragment-based discrete model employing a relatively simple and highly effective vocabulary. In this approach, not only couples of highly correlated electrons are considered as individual components (the covalent bonds), but even groups of atoms bonded together are considered independent units (e. g. aromatic moieties or functional groups). Unfortunately, the mainstream quantum chemical models are not directly compatible with such vocabulary; this gave rise to a historical dichotomy between qualitative concepts and quantitative computations which is still present nowadays. It is possible however, to use a chemical-like functional form for the largest contributions to the molecular energy based on a revival of the relatively simple models employed in the early stages of Quantum Chemistry, thereby reconciling the fragment-based vocabulary with first principles, providing additional insight on the phenomena studied. As a matter of fact, the renaissance of the valence-bond model and the increasing use of explicitly correlated electron pairs (geminals, F12 approaches [31, 32], etc.) show that models based on loosely coupled groups of strongly correlated electrons have still much to offer in theoretical chemistry.

This is the general context of our work on Molecular Perception (MP) [77], which is the set of rules and techniques that derive additional information and give chemical meaning to an initial set of raw data. Traditional chemical perception is performed starting from an initial set of atoms distributed in space. Suitable heuristics are then applied to derive chemical quantities such as covalent bonds, hydrogen bonds, charges, etc. Being able to directly identify the most relevant chemical properties of a molecule with minimal computational complexity allows atom types assignment for Molecular Mechanics simulations. It is also the basis of chemical visualization. Thus, Molecular Perception aims at combining the benefits of a human easy-to-interpret representation of molecules with a quantitative analysis of the molecular properties, using Machine-Learning as a bridge between the two fields. The Molecular Perception (MP) [77–79] algorithms and heuristics here discussed have been implemented in a custom C++ software library called Proxima [77], available to users in the Python language through Cython bindings

[80]. The Python language is handy for its huge amount of resources when it comes to ML [81–83].

The initial data for MP procedures is usually the set of atoms scattered in space, thus:

$$\begin{cases} Z_i, \forall i \\ \vec{p_i} = [x_i, y_i, z_i], \forall i \end{cases} \tag{4.1}$$

However, the entry point for a perception procedure can also be the bond order matrix that defines the two-body interaction between each pair of atoms in the system:

$$\begin{cases} Z_i, \forall i \\ BO_{ij}, \forall i, j \end{cases} \tag{4.2}$$

In case Cartesian coordinates are given as inputs, a bond order matrix is computed as the first step so that the subsequent MP procedures can be executed using the same matrix as a reference irrespective of the source input. The task is to determine a topology while keeping an internal continuous representation of the system through the bond order matrix. The topology is generally identified with a molecular graph [84], which is a graph whose vertices are the atoms ($\{Z_i\}$), and the edges are the two-pair bonds connecting each pair of atoms ($\{BO_{ij}\}$). Although a bond order is defined between each couple of atoms, such representation is only kept internally by the software and only a partial molecular graph is provided to the user, that is the molecular graph containing only those edges whose bond order is above a given threshold (typically 0.5). In this way, the molecular graph coincides with the traditional representation of a molecule every chemist is used to. It is important to notice that the molecular graph is just a particular case of a more general chemical graph where each vertex is an (almost) independent unit in the molecule and each edge is the intensity of a connection between these units in the molecule itself. As an example, an entire aromatic group in a molecule can be considered almost independent because of its nature (a group of highly coupled atoms) and thus can be treated as a single vertex in a more general graph. However, since the goal is to simplify the representation to the user, this information is only kept internally along the continuous bond order matrix. Thus, in computing the topology for the system, the following operations are performed:

- Computation of σ bond orders (BO^{σ})
- Computation of π bond orders (BO^{π})
- Computation of non-covalent bond orders (e.g., BO^{hbond})

Once the topology is computed and the bond order matrix is assigned, MP algorithms can further explore the properties of the system (e.g. perception of charges, perception of rings, etc.) or help in the pre-processing and post-processing of the data (e.g. solvation procedures). In the following, the perception of topology and these other perception algorithms are treated in detail.

4.1 Topology Perception

As mentioned in the introduction, Molecular Perception aims at describing complex systems starting from the minimum amount of information available [77]. In the present context, it is sufficient to know how atoms are placed in real space with their 3D coordinates, or how atoms are bonded to each other. In both cases, an automatic perception fills the missing information. Proxima supports PDB [85] and XYZ files for explicit encoding of three-dimensional coordinates. It is also possible to provide simple SMILES [86] or connectivity matrices in order to encode the topology information directly.

From a purely physical point, there is no distinction between "Intra-Molecular" and "Inter-Molecular" phenomena; the fragment-based approach of chemistry must, therefore, be based on some kind of approximation. To this end, we will start by building a skeleton of covalent (possibly delocalized) bonds and then proceed to add weaker interactions between bonds, Lone Pairs, and holes. Then, inter-molecular interactions come into place when interactions between disjoint fragments not linked by the covalent skeleton are considered.

In traditional chemical perception atoms are the basic bricks. However, atoms in molecules are too different from spherical atoms to allow a simple description of chemical phenomena. We decided, therefore, to go back to the old concept of the valence state of an atom, i.e. an atom with a formal charge (given by the user) in the electronic state with the maximum possible spin multiplicity. Therefore, the basic entities of the Feature Space employed in our perception are electrons, Lone Pairs, and holes lying in properly hybridized orbitals, which are then employed to build the molecular framework. After defining an initial discrete connectivity matrix, all the following operations are not performed in the real space, but in the 'Feature Space' of charges and bond orders, thus allowing seamless topology modifications, which are beyond the capabilities of the most widespread Molecular Perception packages.

4.1.1 Covalent Bonds

The perception of covalent bonds has traditionally relied on the relative distances between the atoms [87]. Here, the standard formulation of covalent bond perception (that checks whether the relative distances are below the sum of covalent radii plus a threshold) is enhanced through the explicit inclusion of electronegativity [88, 89], namely

$$r_{min} \leq d_{ij} \leq r_{cov,i} + r_{cov,j} - 0.07\Delta\chi^2 + tolerance \tag{4.3}$$

where r_{min} is a threshold, which allows avoiding clashes issuing from poor initial geometries, d_{ij} is the relative distance between the i and j centers, $r_{cov,i}$ and $r_{cov,j}$ are the relative covalent radii, $\Delta\chi$ is the electronegativity difference and the tolerance is generally 0.4 Angstroms [87]. Noted is that pairs of atoms at distances shorter than r_{min} are not connected, but are explicitly signaled by warnings in the output.

The problem with the given formulation, however, is that it just classifies two atoms as either bonded or not-bonded, without providing a "strength" (a bond

order) for the pair considered. In the following, we decided to employ this general expression for two-body interactions [88, 89]:

$$BO^{\sigma}(r) = \begin{cases} 1, \text{ if } r < r_m \\ e^{\alpha(r-r_m)^2}, \text{ otherwise} \end{cases} \tag{4.4}$$

Where r is the distance between a couple of atoms considered and r_m is the theoretical bond length as described above ($r_{cov,i} + r_{cov,j} - 0.07\Delta\chi^2$). In order to find the value for the exponential α we apply the additional constraint that the value of the bond order at a distance of $r_m \cdot 1.33$ angstroms should be equal to 0.5. Thus, in perceiving covalent bonds, only bond orders whose values are above 0.5 are perceived. The choice of the Gaussian in describing the covalent bond interaction allows us to cut the interaction to 1 around r_m, otherwise an exponential would introduce a non-derivable point.

4.1.2 Delocalized π systems

Once covalent bonds are perceived, electrons are considered. In Fig. 4.1, an example of how the perception procedure works in considering electronic structure is given.

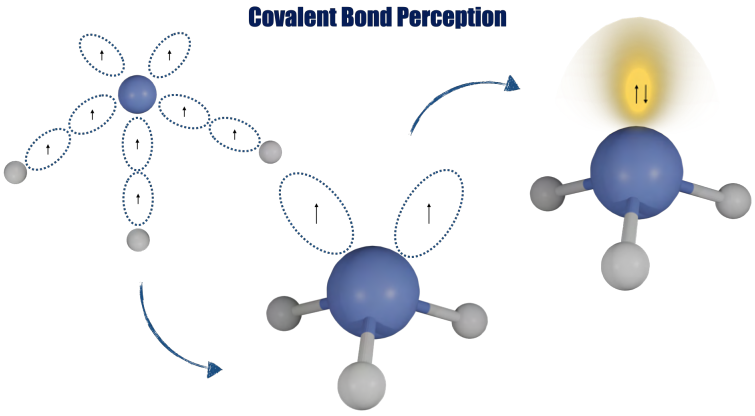


Figure 4.1: The Molecular Perception of Covalent Bonds in NH_3.

In particular, for each covalent bond perceived, a couple of electrons are localized into the bond itself. As a consequence, starting from an initial configuration for the atom of completely uncoupled electrons, these are coupled in covalent bonds, and the remaining electrons are either coupled together in Lone Pairs (if in even number) or radical species. In case the covalent bond perception steps give rise to a number of covalent bonds higher than the number of electrons available, these bonds are sorted from the longest to the shortest and are removed in this order until there is consistency between the number of electrons and the number of covalent bonds (valence constraint). Of course, formal charges given by the user are used to correctly assign the number of electrons to each atomic species.

In order to compute good-quality π bond orders we have employed a semi-empirical Tight-Binding approach. The simple TB model used in the following always uses a single orbital per atom and is based on the original formulation of

the Huckel-Del Re model [90–92] both for localized σ and delocalized π moieties. The π portions of the system are perceived by performing a Breadth-First Search (BFS) [93], which groups together nearby atoms having unpaired electrons, electron pairs, and holes. The wave function for each ensemble is expressed in terms of effective orthogonal orbitals centered at the involved atoms. In the approximation that only atoms linked by a The simple TB model used in the following always uses a single orbital per atom and is based on the original formulation of the Huckel-Del Re model [90–92] both for localized σ and delocalized π moieties. The π portions of the system are perceived by performing a Breadth-First Search (BFS) [93], which groups together nearby atoms having unpaired electrons, electron pairs, and holes. The wave function for each ensemble is expressed in terms of effective orthogonal orbitals centered at the involved atoms. In the approximation that only atoms linked by a sigma bond are coupled, the effective TB Hamiltonian can be always recast in a tri-diagonal form, whose eigenvectors convey all the information needed to obtain atomic charges and bond orders: bond are coupled, the effective TB Hamiltonian can be always recast in a tri-diagonal form, whose eigenvectors convey all the information needed to obtain atomic charges and bond orders:

$$TB_{ij} = \begin{cases} \alpha_{ii}, i = j \\ \beta_{ij}, j = i \pm 1 \\ 0, otherwise \end{cases} \tag{4.5}$$

In the case of delocalized systems, the Tight-Binding matrix (4.5) is diagonalized and the set of eigenvectors is sorted from the lowest eigenvalue to the highest one. Then bond orders can be computed with the following expression:

$$BO_{ij}^{\pi} = \sum_{k} n_k c_i^k c_j^k \tag{4.6}$$

In general, the traditional Huckel model [90–92] works well for simple π systems where each atom contributes with one orbital to the Tight-Binding scheme. However, in considering *sp* systems an atom can contribute with multiple π orbitals such as in the case of Fig. 4.2b.

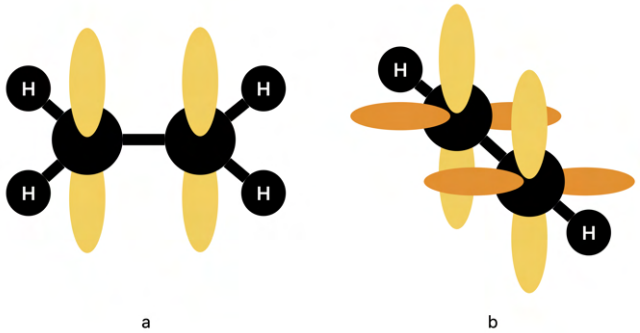


Figure 4.2: a) The single-occupied p orbitals for the ethylene molecule b) The single-occupied p orbitals for the ethyne molecule (two orbitals for each atom).

In the case shown in such a picture, the yellow orbitals are coupled together and the overlap with the orange orbitals goes to zero since these are orthogonal.

In considering such systems, it is important to identify each element of the matrix TB_{ij} with an orbital of an atom instead of directly with the atom itself. Each atom a will be identified by an ensemble of coefficients $C(a)$ that correspond to each orbital that the atom itself offers to the delocalized system considered. In this way, the previous equation used for computing the π bond orders must be modified in the following way:

$$BO_{ij}^{\pi} = \sum_{l} n_e^{(l)} \sum_{k_1 \in C(i), k_2 \in C(j)} c_{k_1}^{(l)} c_{k_2}^{(l)} \tag{4.7}$$

Where l identifies the energy level and $n_e^{(l)}$ is the number of electrons in the l level. In our implementation we employed the following definition for the β_{ij} term for a couple of orbitals:

$$\beta_{ij} = -\frac{1}{2}(\alpha_i + \alpha_j) \cdot \cos^2\theta \cdot \left\{ e^{\gamma_{ij}\left(1-\frac{r}{r_m}\right)} - 2e^{\frac{\gamma_{ij}}{2}\left(1-\frac{r}{r_m}\right)} \right\} \tag{4.8}$$

In this case, $cos^2\theta$ accounts for the angular dependency of the overlapping term, while r is the actual bond length and r_m is the equilibrium bond length computed using reference covalent radii (for double and triple bonds accordingly). The functional form of the radial component for the β term is a Morse function [94]. The maximum absolute value of the off-diagonal term (β) is taken as the average of the diagonal terms (α): as a consequence, the only parameters required are the diagonal α and the radial decay γ. In order to obtain such parameters, the Wiberg bond orders [95] were taken as references computed at the B3LYP[27]/aug-cc-pVDZ level. In this case, the parameters were chosen so as to obtain the best fit of Wiberg bond orders [95] with a second-order polynomial in the range of double bonds, and a linear expression in the range of triple bonds. The need for such interpolation arises from the natural tendency of the Huckel-like models toward over-delocalization: for instance, in the case of benzene, the analytical solution of a Huckel model gives a π bond order of $\frac{2}{3}$ [96] to be added to a σ bond order of 1 and compared to a Wiberg bond order [95] of 1.4. In optimizing parameters we also took care of correctly describing the hyperconjugation effect [97, 98] by assigning different parameters to those carbon atoms connected to other sp^3 carbon atoms with hydrogens (positive hyperconjugation: the σ C-H orbital interacts) and to those carbon atoms connected to other sp^3 carbon atoms bonded to halogens (negative hyperconjugations: the σ^* C-X orbital interacts). In Table 4.1, the resulting parameters are shown.

	α	γ
C1	-68.96	10.39
C2	-67.05	16.27
NSP	-40.39	4.52
N1	-33.65	68.68
N2	-100.00	39.84
O1	-53.44	0.01
O2	-100.00	49.74
P1	-68.41	20.03
P2	-99.99	100.00
S1	-37.20	0.02
S2	-100.00	38.83
pHyp1	-44.43	22.58
nHyp1	-90.22	3.67
pHyp2	-62.48	73.82
nHyp2	-59.38	72.99

Table 4.1: The Tight-Binding parameters in relative units.

The best fit of Wiberg bond orders [95] is obtained with the following corrections:

$$\begin{cases} BO_{corr}(BO_o) = 1.0007 \cdot BO_o^2 - 1.9892 \cdot BO_o + 1.9554, \text{ if } BO_o < 2.1 \\ BO_{corr}(BO_o) = 2.5942 \cdot BO_o - 4.8860, \text{ otherwise} \end{cases} \tag{4.9}$$

Except for bonds involving *sp* carbons in negative or positive hyperconjugation, for which a better fit with the Wiberg bond order [95] is given by the following equation:

$$BO_{corr}(BO_o) = -0.8548 \cdot BO_\pi + 5.8665 \tag{4.10}$$

In Fig. 4.3, the result of the correlation between the bond orders for a large set of organic molecules is shown. This has been obtained by optimizing the α and β parameters.

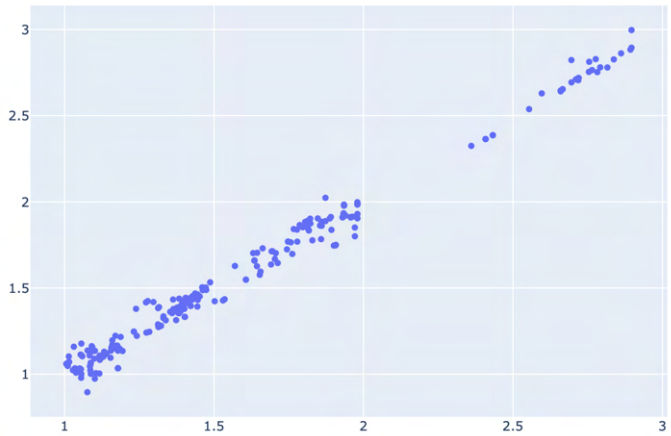


Figure 4.3: The correlation between the Proxima bond order on the horizontal axis and the Wiberg B3LYP[27]/aug-cc-pVDZ bond orders [95] for the vertical axis. The maximum difference is 0.15.

It is of general interest to notice how the use of the $cos^2(\theta)$ contribution in the β term of the Tight-Binding matrix correctly describes the behavior of the bond orders with changing angle values. An example is the Biphenyl molecule observing how the bond order changes when rotating the two rings one with respect to the other. In Fig. 4.4, the rotation around the bridge bond of the two biphenyl groups is shown.

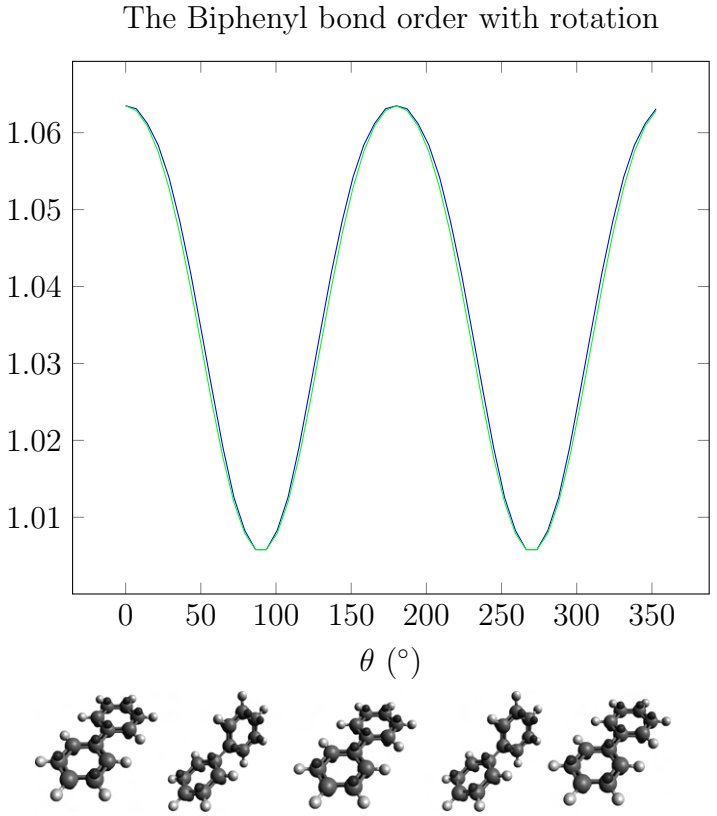


Figure 4.4: The change in the bond order of the bridge bond in the biphenyl molecule as the two rings rotate in one respect to the other. The blue line is the QM Wiberg bond order [95] computed at the B3LYP[27]/jun-cc-pVDZ level, while the green line is the bond order computed by Proxima. On the x-axis, the torsional angle is shown in degrees.

It is interesting to notice how the correct behavior gets replicated by Proxima, and also how the bond order lowers as the two rings approach 90 degrees breaking conjugation.

4.1.3 Non-Covalent Bonds

Hydrogen bonds will be considered in detail for purposes of illustration, but the same procedure can be employed for different kinds of interactions (e.g., halogen bonds, etc.). To check for the presence of a hydrogen bond, each possible acceptor atom (such as nitrogen, oxygen, phosphorous, or sulfur) is checked for the presence of Lone Pairs. If this is the case, nearby atoms are checked for the presence of hydrogen atoms connected to donor atoms. The Lone Pairs of the acceptor atom are actively used for the perception of the bond so as to avoid finding hydrogen

bonds in the so-called σ hole, the region inside the cone created by the rotation of the two Lone Pairs. With respect to more traditional expressions, we don't need any more specific angular parameters since the best orientation of the hydrogen atom, the Lone Pair involved, and the acceptor atom is always linear provided that Lone Pairs are properly oriented. To this end, we enforce angles of 109.47° or 120° for couples of Lone Pairs at sp^3 or sp^2 centers, respectively. Single Lone Pairs are, instead, placed by minimizing the difference of the angles they form with the other substituents. In the following, we are going to discuss the computation of the hydrogen bond intensity for just a single hydrogen bond defined by the triplet of atoms: (A,H,D). In this context, hydrogen bonds are considered independent thus a complete description of non-independent interactions (such as bifurcated bonds) is left for future works. The absolute value of the intensity of a hydrogen bond is computed employing a radial expression such as a modified Morse recently introduced in a work about noncovalent interactions [99], the same expression of the β term in the TB model, in this case, the reference distance considered is between the Donor and Acceptor atoms:

$$I(r) = -1 \cdot \exp\left\{ \left(\alpha\left(1-\frac{r}{r_m}\right)\right) - \left[\left(\frac{r}{r_m}\right)^4 - 2\left(\frac{r}{r_m}\right)^2 + 3\right] \exp\left(\frac{\alpha}{2}\left(1-\frac{r}{r_m}\right)\right)\right\} \tag{4.11}$$

The r_m parameter for the computation of the global intensity is taken as the one introduced in a study about hydrogen bond interactions [100, 101]. The α parameter is chosen so that the intensity is 0.5 when $r = r_m + 0.4$. The difference between a hydrogen bond and a σ bond, however, is the presence of a "directionality" due to the presence of a third atom: the bridge hydrogen atom. In order to account for such directionality in quantifying the strength of a hydrogen bond, Lone Pairs are considered. For example, in Fig. 4.5, the computation of the intensity of a hydrogen bond is shown for a single Lone Pair configuration.

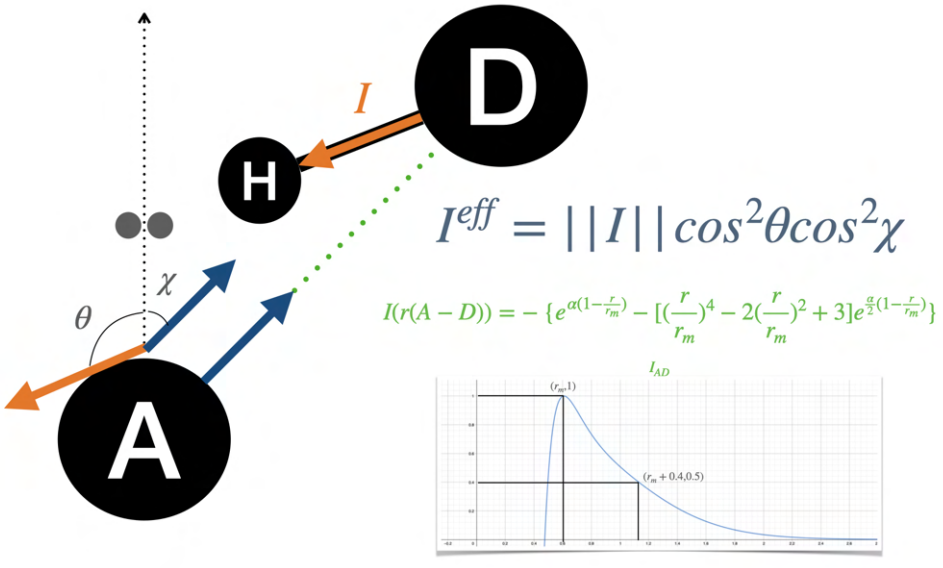


Figure 4.5: The strength of a hydrogen bond in a single Lone Pair configuration.

In this case, only the relative strength along the direction of the Lone Pair is the effective strength of the hydrogen bond, because of its directional nature. In order to quantify the relative strength along the Lone Pair direction, the traditional dot product between vectors should be used as $||I||cos(\theta)$. However, in cases where the hydrogen atom is placed orthogonal to the Lone Pair (with θ higher than 90°), we want to cut the interaction to 0. The problem with cutting the interaction to 0 above 90° is that we introduce a non-derivable point since the cosine has a maximum derivative around 90° angles. To solve this problem, we compute the projection of the intensity along a given direction through a modified operator defined as $||I||cos^2(\theta)$. By employing the square of the cosine, instead of the single cosine, we require the derivative to go to 0 at around 90°, so we can cut the interaction with no further problem. It is also important to remember that these quantities are based on heuristics rather than physics and the only important characteristic that they should respect is to describe correctly the phenomena in relative terms. Precise values can then be obtained by other regression methods. In practice, by treating the cos^2 term as a scaling factor, we not only need to make sure that the orientation of the D-H is correct but also that the A-D orientation is best aligned. To do so, we simply consider the additional angle χ as shown in Fig. 4.5:

$$I_{eff} = ||I||cos^2\theta cos^2\chi \tag{4.12}$$

If two Lone Pairs are considered, as it is in the case of carbonyl oxygen, we have two different sources of directionality that we must take into account. In Fig. 4.6, the computation for the effective strength of a hydrogen bond in such a Lone Pair configuration is shown.

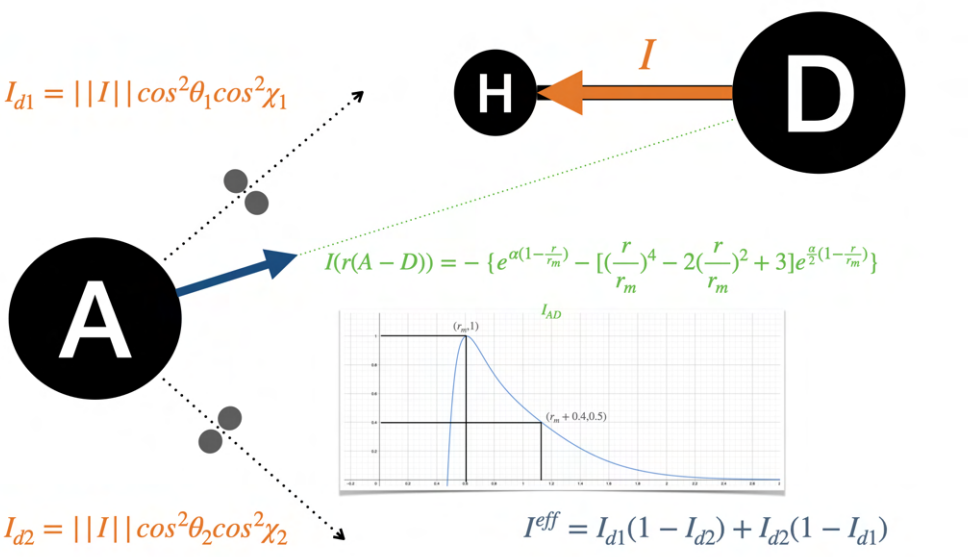


Figure 4.6: The strength of a hydrogen bond in a double Lone Pair configuration.

The same consideration made above on "cutting the interaction around 0" is still valid so, in our modified form:

$$\begin{cases} I_{d1} = ||I||cos^2(\theta_1)cos^2(\chi_1) \\ I_{d2} = ||I||cos^2(\theta_2)cos^2(\chi_2) \end{cases} \tag{4.13}$$

Moreover, in this case, we get two different directional intensities: I_{d1}, I_{d2}. In order to combine them into a single number we compute the effective intensity as follows:

$$I_{eff} = I_{d1}(1 - I_{d2}) + I_{d2}(1 - I_{d1}) \tag{4.14}$$

This can be interpreted as: the effective intensity has maximum value if the hydrogen atom is directed along d1 AND NOT along d2, OR if it is directed along d2 AND NOT along d1. It is clear how in the general case of n Lone Pairs placed on the acceptor atom, we can extend the treatment by computing our modified intensities along each Lone Pair ($\{I_{di}\}$), and then combining them in an effective intensity:

$$I_{eff} = \sum_{i}^{n} I_{di} \prod_{j \neq i} (1 - I_{dj}) \tag{4.15}$$

As an example, in Fig. 4.7, the hydrogen bond profile is given for a donor group rotating around an acceptor.

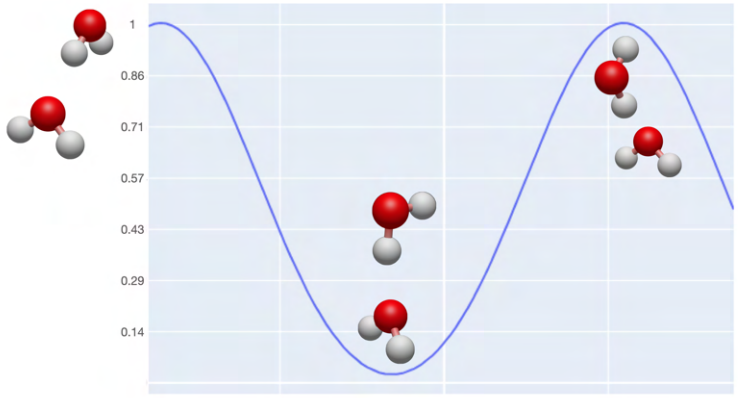


Figure 4.7: An example of hydrogen bond intensity computed by rotation of the donor group around the acceptor.

It is important to notice the correct representation of the σ hole as an unfavorable configuration for the formation of the hydrogen bond. Once hydrogen bonds are detected, they can be added to the list of bonds. As a consequence, both polarization and charge transfer across hydrogen bonds can be taken into proper account. As mentioned above, other kinds of non-covalent interactions can be treated in the same way whenever customary van der Waals and electrostatic interactions are not sufficiently accurate. It is interesting to notice that, in principle, it would be possible to check for the presence of holes in the neighborhood of atoms with electron pairs for the detection of, more generally, dative bonds in which hydrogen bonds are a special case. It is important to keep in mind that a hole is assigned to a hydrogen atom by default in our procedure.

4.2 Empty Valence

Once the topology has been assigned, MP algorithms are used to extract further information from the system or even to manipulate the system itself for pre-processing or post-processing operations. One example is the detection of missing hydrogen atoms. The hydrogen atom is the simplest element and is the most common substituent in filling the empty valence of heavier atoms. Moreover, widely employed experimental techniques such as X-Ray crystallography are not able to detect unambiguously most hydrogen atoms due to their low charge density. As a consequence, hydrogen atoms are very often missing from primary sources of information and must be added by ad hoc procedures. The general approach employed here is based on the perception of the hybridization of the involved atoms, which allows us to know the number of connected atoms with their reference geometry. As an example, if a neutral carbon atom with two connected atoms and sp^2 hybridization is detected, then a third hydrogen atom is added in the corresponding spot. There are some edge cases though that deserve a more careful treatment, like, e.g., terminal and isolated atoms. In the case of terminal atoms, there is a single bond thus there are no angles to check for hybridization. However, the perception of valence electrons permits the detection of the number of unpaired electrons. Thus, by performing π bond perception, a hydrogen atom is added to each σ unpaired electron so as to fill valences. The geometry used to decide the hydrogen positions takes the original bond as a reference. In the case of isolated atoms, i.e., atoms without bonds, the valence electrons are all unpaired except for the natural Lone Pairs. As an example, an isolated carbon atom will have four unpaired electrons and an isolated oxygen will have two unpaired electrons instead. Thus, an isolated carbon atom is interpreted as methane, and an isolated oxygen atom as a water molecule. Another edge case is aromatic rings involving nitrogen atoms. In fact, aromatic nitrogen atoms can be bonded to a hydrogen atom or not. In this case, aromaticity can be perceived by checking the planarity of the ring itself and the number of hydrogens added to each nitrogen atom so that the total number of electrons delocalized in the ring respects the aromaticity rules. The real problem with isolated atoms is that the lack of reference bonds introduces some ambiguity in the preferred orientation of the hydrogen atoms to be added. The general rule of thumb, developed with the water molecule in mind, is to check for the two nearby atoms between the ones with higher electronegativities that are closer to the central atom, as shown in Fig. 4.8.

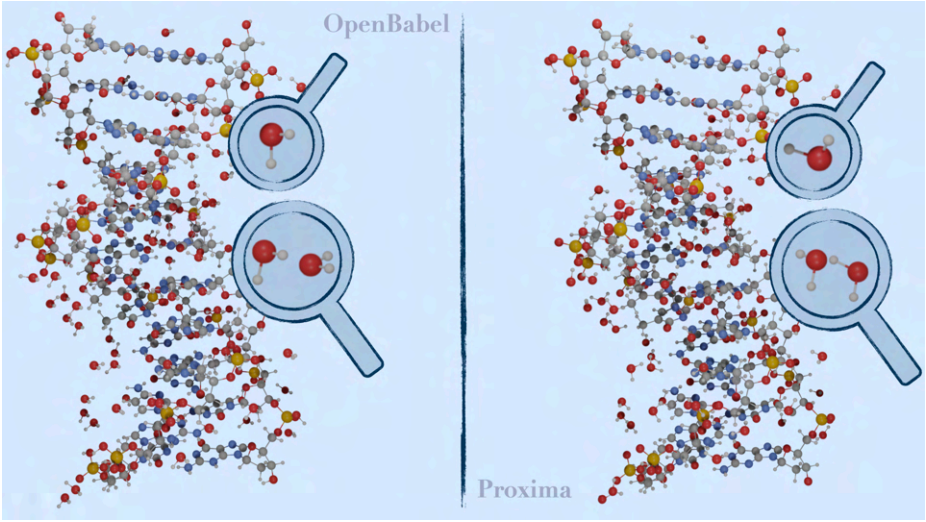


Figure 4.8: Automatic addition of implicit hydrogens. Notice the change in the orientation of the hydrogen atoms in the case of water molecules, where Proxima uses the most electronegative nearby atoms to define hydrogen orientation.

Thus, we can define a reference plane in which to place the first two hydrogens (as in the case of the water molecule) and additional out-of-plane hydrogens (as in the case of the methane molecule).

4.3 Charge Perception

The other fundamental perception step, once the topology is computed, is the perception of charges. Let's take a two-body system (such as a σ bond), the TB matrix can be written as follows:

$$\boldsymbol{TB} = \alpha^0 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \beta^0 \begin{bmatrix} \delta_A & \epsilon_{AB} \\ \epsilon_{BA} & \delta_B \end{bmatrix} \tag{4.16}$$

where:

$$\alpha_A = \alpha^0 + \delta_A \beta^0 \tag{4.17}$$

$$\beta_{AB} = \epsilon_{AB} \beta^0 \tag{4.18}$$

And the following is satisfied for every atom [92]:

$$\delta_A = \delta_A^0 + \sum_{\text{B connected to A}} \gamma_A(B) \delta_B \tag{4.19}$$

If the starting geometry of the system is only approximate, the off-diagonal terms are given fixed reference values, but if sufficiently accurate geometries are available (or during geometry optimizations) an explicit r_{AB} dependence for the β term can be employed as follows:

$$\epsilon_{AB} = f_{AB}(r) + \frac{\alpha^0}{\beta^0} + \bar{\delta} \tag{4.20}$$

With:

$$\bar{\delta} = \frac{\delta_A + \delta_B}{2} \tag{4.21}$$

And:

$$\begin{aligned} f_{AB}(r) =& \frac{\Delta E_{AB}}{\beta^0} \Bigg\{ \exp\left(\alpha\left(1 - \frac{r}{r_m}\right)\right) \\ & - \left[\left(\frac{r}{r_m}\right)^{2n} - 2\left(\frac{r}{r_m}\right)^{n} + 3\right] \exp\left(\frac{\alpha}{2}\left(1 - \frac{r}{r_m}\right)\right) \Bigg\} \end{aligned} \tag{4.22}$$

This expression is the same used for the β term of the TB model [99]. In the case of covalent interactions, instead, the $n = 0$ expression is taken obtaining a traditional Morse. Here ΔE is the dissociation energy and the r_m parameters are either covalent or vdW radii. Diagonalization of the TB matrix provides the eigenvalues of each two-body system

$$E = \alpha^0 + \beta^0 \left(\bar{\delta} \pm \sqrt{\Delta\delta^2 + \epsilon_{AB}^2}\right) \tag{4.23}$$

Where:

$$\Delta\delta = \frac{\delta_A - \delta_B}{2} \tag{4.24}$$

Thus, by truncation to the first order (using the ϵ_{AB} expression above):

$$E = -\beta^0 \cdot f_{AB}(r) \tag{4.25}$$

Proving that the choice of the β off-diagonal parameter is the one determining the shape of the overall energy of the system. At this point, charges can be directly computed with the following expressions for a σ bond:

$$Q_A = \sum Q'_{A(B)} \tag{4.26}$$

$$Q'_{A(B)} = \frac{\delta_B - \delta_A}{2\epsilon_{AB}} \tag{4.27}$$

In order to improve the computation of charges, we decided to simplify the original formulation thus deriving a custom set of new parameters. The original Del Re expression for the computation of the σ charge is the following [92]:

$$Q_i = \sum_j \frac{\delta_j - \delta_i}{2\epsilon_{ij}} \tag{4.28}$$

This equation (where for each atom the condition $\delta_i = \delta_i^0 + \sum_j \gamma_{ij}\delta_j$ is satisfied) is employed in its first-order variation (where $\delta_i = \delta_i^0 + \sum_j \gamma_{ij}\delta_j^0$). Such approximation is generally valid for systems in which standard electronegativities are not that different from in situ electronegativities. In general, the ϵ_{ij} and γ_{ij} parameters can be further expressed as dependent on single atom parameters. The simplest combination rule that we have employed in the following is simply the average:

$$\epsilon_{ij} = \frac{\epsilon_i^0 + \epsilon_j^0}{2} \tag{4.29}$$

$$\gamma_{ij} = \frac{\gamma_i^0 + \gamma_j^0}{2} \tag{4.30}$$

Moreover, we decided to explicitly include the total $(\sigma + \pi)$ bond order into the equation so to account for delocalization. In addition to the total bond order, since we want to correctly scale charges depending on the strength of the The simple TB model used in the following always uses a single orbital per atom and is based on the original formulation of the Huckel-Del Re model [90–92] both for localized σ and delocalized π moieties. The π portions of the system are perceived by performing a Breadth-First Search (BFS) [93], which groups together nearby atoms having unpaired electrons, electron pairs, and holes. The wave function for each ensemble is expressed in terms of effective orthogonal orbitals centered at the involved atoms. In the approximation that only atoms linked by a sigma bond are coupled, the effective TB Hamiltonian can be always recast in a tri-diagonal form, whose eigenvectors convey all the information needed to obtain atomic charges and bond orders: bond, each term $\delta_j - \delta_i$ is further multiplied by the The simple TB model used in the following always uses a single orbital per atom and is based on the original formulation of the Huckel-Del Re model [90–92] both for localized σ and delocalized π moieties. The π portions of the system are perceived by performing a Breadth-First Search (BFS) [93], which groups together nearby atoms having unpaired electrons, electron pairs, and holes. The wave function for each ensemble is expressed in terms of effective orthogonal orbitals centered at the involved atoms. In the approximation that only atoms linked by a sigma bond are coupled, the effective TB Hamiltonian can be always recast in a tri-diagonal form, whose eigenvectors convey all the information needed to obtain atomic charges and bond orders: bond order ($BO^\sigma \in [0, 1]$) so as the γ_{ij} term. The σ charge then is:

$$Q_i = \sum_j \frac{(\delta_j - \delta_i) BO_{ij}^\sigma}{2\epsilon_{ij} BO_{ij}^{\sigma+\pi}} \tag{4.31}$$

By taking CM5 charges computed at the B3LYP[27]/aug-cc-pVDZ level of theory, we optimized the δ_i, ϵ_i and γ parameters using a simple differential evolution algorithm minimizing the RMSE value between the computed and predicted charges. The resulting optimized charges for sp^3 systems are shown in Fig. 4.9.

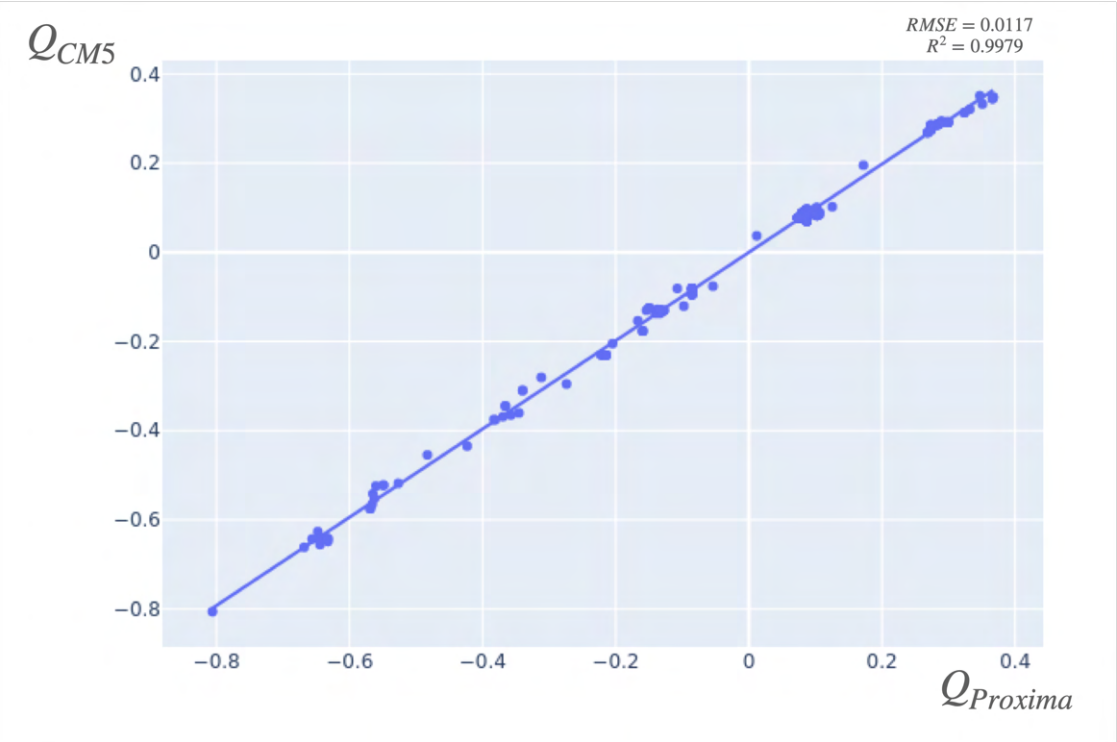


Figure 4.9: The correlation between the charges computed by Proxima and the CM5 total charges for C, N, and O atoms in the case of sp^3 systems for 228 atoms from 50 molecules.

In Fig. 4.10, instead, charges are shown for sp^2/sp systems.

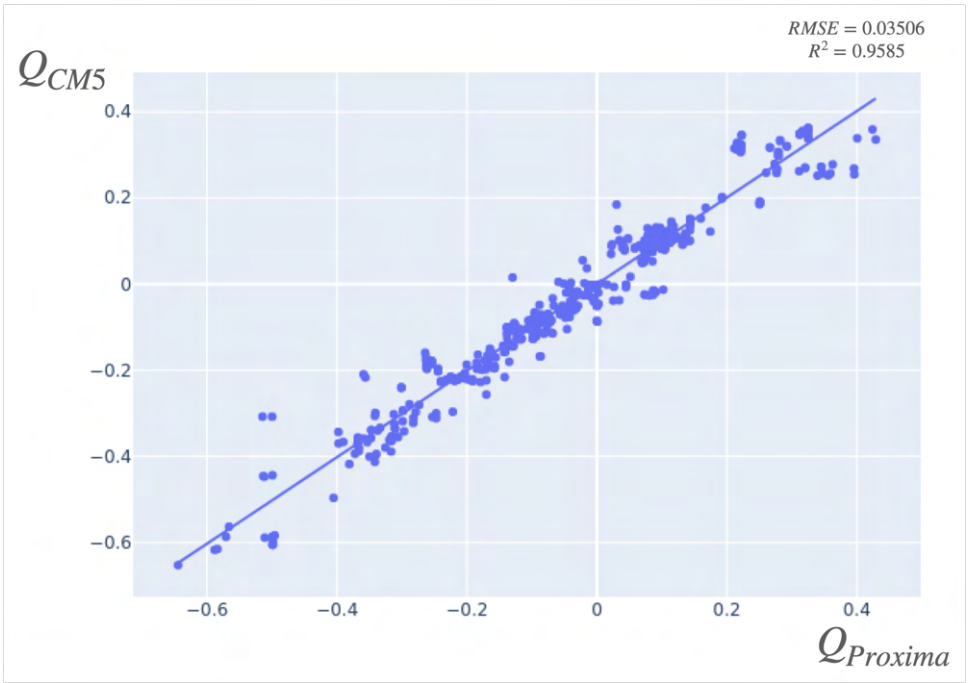


Figure 4.10: The correlation between the charges computed by Proxima and the CM5 charges on a dataset of 933 atoms from 164 molecules.

The overall parameters are reported on Tab. 4.2.

	δ	ϵ	γ
H	-0.028	0.221	0.053
CSP3	0.060	0.907	0.079
C1	-5.060	-61.401	0.127
C1p	-3.585	-56.672	-0.098
C1n	-2.047	-53.626	-0.099
C2	-8.476	-63.927	0.215
C2p	0.057	-38.541	0.380
C2n	0.054	-28.526	0.318
NSP3	0.290	0.891	0.056
NSP	0.848	7.035	-0.514
N1	-0.263	-35.300	5.043
N1(amide)	-11.733	-58.479	-0.086
N2	-8.969	-33.169	-0.093
OSP3	0.351	0.843	0.131
O1	-0.871	47.050	5.554
O2	-10.505	-33.311	-0.094
PSP3	0.225	12.642	0.271
P1	-0.965	-81.853	2.098
P2	-3.234	-41.186	-0.389
S(SP3)	0.138	1.151	0.135
S1	-14.770	31.327	0.438
S2	-4.703	-45.660	-0.094
F	-0.180	-2.018	-0.480
CL	-0.087	-5.091	-0.617
BR	0.049	9.737	-7.187

Table 4.2: The parameters used for charge computation.

The higher variation of charges in sp^2 systems with respect to the sp^3 ones are explained by the approximation of the original Del-Re model to its first iteration only which determines a lack of delocalization in describing the overall atomic charge (although is partially taken into account due to the presence of the π bond order at the denominator). In fact, the biggest deviations are seen in small delocalized environments with different heteroatoms bonded together. These maximum differences are the nitrogens of urea (0.16 charge difference in absolute value), the carbon atom of acetone (0.16 charge difference), and the nitrogen of ethyl carbamate (0.16 charge difference). Then, all the other charges have a difference with the CM5 that is below 0.15 with an RMSE value of 0.036. In fact, the good news is that the charge computed for systems of our interest such as peptides and simple amides is good enough. In future developments, we are going to study a multi-goal strategy to simultaneously optimize the Tight-Binding parameters for charges and bond orders, although the Huckel TB method has shown a tendency to over-delocalize (such as in the case of bond orders) while CM5 charges show smaller variations. However, despite all these limitations, the agreement between CM5 and Proxima charges is still good enough to justify its use in perception procedures. In cases the user wants to reach higher precision in a region of the molecule, it is still possible to compute CM5 charges with traditional

QM methods and correct them point by point. In fact, Proxima can partition the system into custom "fragments". The advantage of such an approach is that it is also possible to do the opposite: instead of dividing a molecular system into fragments, it is possible to assemble fragments together forming a new molecular system. In fact, by having access to a database of different fragments with charges computed at the quantum level, we can assemble together these fragments in bigger systems by just recomputing charges where necessary. In Fig. 4.11, the mathematical equations for computed charges from disconnected fragments are summarized. In the picture, the x/y atoms are disconnected from the A/B atoms and then A and B are connected together. In general, the charge can always be written as:

$$Q_{A'} = Q_A + \frac{\delta_B - \delta_A}{2\epsilon_{AB}} + K_{A,B,X,Y} \tag{4.32}$$

Depending on whether the X atom is a terminal atom, and whether x = y, the $K_{A,B,X,Y}$ changes as shown in Fig. 4.11.

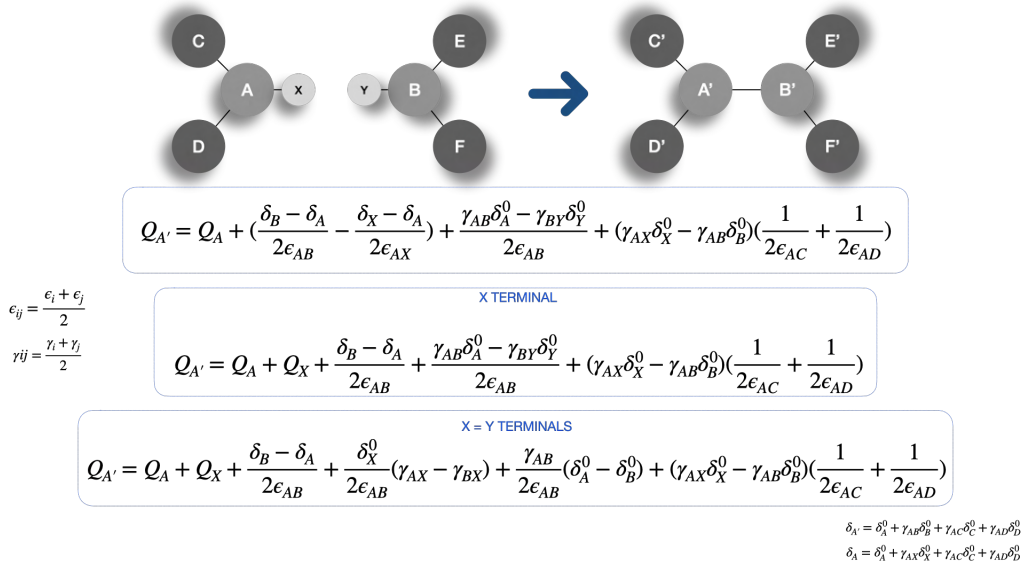


Figure 4.11: The process of taking two different fragments, disconnecting the x/y atoms from the A/B atoms, and then creating a new bond between A and B.

4.3.1 The Hydrogen Bond Charge contribution

The formation of a Hydrogen Bond introduces a Δ on the theoretical charges. In general, it is a contribution of Charge Transfer (CT) and polarization. In general, we focused on inter-molecular hydrogen bond formation such as the ones involving water molecules. As a consequence, the polarization effect was taken into account requiring it to replicate the charge of water in bulk. In order to quantify the Charge Transfer effect, a set of clusters of water molecules and ammonia has been taken into account. For each molecule, the Δ_i has been computed as the difference between the CM5 charge of the i-th atom in the cluster and the same atom in the isolated molecule. The Charge Transfer is then quantified as the summation of these deltas onto the acceptor group.

$$\left|CT\right| = \left| \sum_{i\in\{a_i\}} \Delta_i \right| = -\left| \sum_{i\in\{d_i\}} \Delta_i \right| \tag{4.33}$$

Where $\{a_i\}$ is the ensemble of atoms connected to the acceptor atom and $\{d_i\}$ is the ensemble of atoms connected to the donor atom. In general, for a given hydrogen bond with an intensity I, the Charge Transfer is expressed as:

$$\left|CT\right| = \left|CT\right|_{opt} \cdot I \tag{4.34}$$

As a consequence, the effect of the Charge Transfer on the charges is the following:

$$Q_{hb}(i) = Q^0(i) + \%(i) \cdot \sum_{hbnds} CT_{hbnd} \tag{4.35}$$

Where % is a function of the "percentage" of Charge Transfer that each atom involved can take upon itself (so we have the constraint of $\sum \%_i = 1$), and:

$$CT = \begin{cases} -\left|CT\right|, & \text{Donor group,} \\ \left|CT\right|, & \text{Acceptor group} \end{cases} \tag{4.36}$$

The general values employed for the % function is 0.6 for the donor/acceptor heteroatom and $\frac{1-0.6}{n_{bonded}}$ for each atom directly bonded to the donor/acceptor. The Charge Transfer absolute value for the couples of N and O atoms is shown in Tab. 4.3.

		Acceptor	
Donor		O	N
	O	0.086844	0.125037
	N	0.125037	0.105360

Table 4.3: The Charge Transfer parameters employed between acceptor and donor atoms.

In addition to the Charge Transfer, the effect of polarization must be taken into account, since our goal is to replicate the TIP3P [102] charges of the water molecule in bulk. The effect of Charge Transfer vanishes since there are an equal amount of in and out-hydrogen bonds, so they compensate. In order to avoid this effect, we add an internal polarization value that compensates itself within the molecule thus not affecting the overall Charge Transfer. For each D-H donor bond involved in a hydrogen bond, a value of 0.104 is added to the hydrogen atom and a value of -0.104 is added to the donor atom. The polarization thus happens on the single bond and thus does not affect the Charge Transfer. In Fig. 4.12 an example of the formation of a cluster of water molecules is shown.

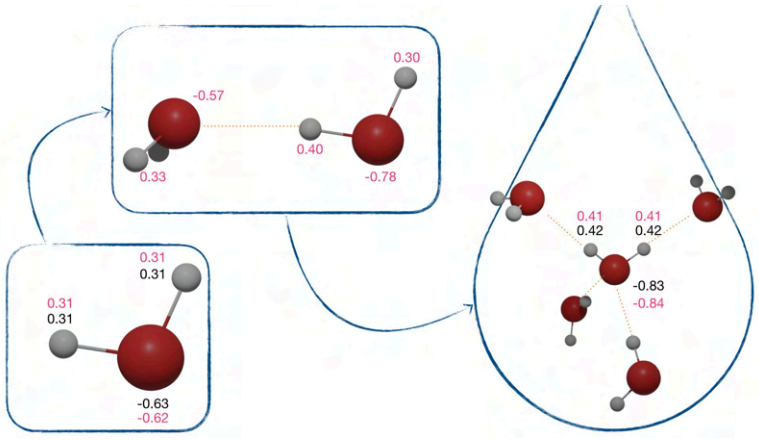


Figure 4.12: The formation of a water cluster

4.4 Ring Perception

The first version of Proxima was already capable of perceiving chemical rings in complex structures using Horton's algorithm. [77, 103, 104] In the current version, chemical rings can also be classified as aromatic and nonaromatic by checking the average torsion angle for planarity and also considering the number of π electrons involved in the cycle. Moreover, Proxima capabilities were extended in order to describe the conformational space associated with ring flexibility. The widely employed ring puckering coordinates [105–107] were added to the tools box, allowing the classification and the ring reconstruction upon modification of the ring coordinates. Cremer Pople coordinates allow for a complete description of ring puckering motions in flexible ring molecules with a set of $N_{ring} - 3$ coordinates, where N_{ring} is the number of atoms belonging to the ring structure (e.g., in the case of the cyclopentane molecule only the five carbon atoms, therefore $N_{ring} = 5$). After rotating and translating the Cartesian framework in which the coordinates of the molecular system under investigation are given according to the following prescriptions:

$$\begin{aligned} \sum_{j=1}^{N_{ring}} r_j &= 0 \\ r' &= \sum_{j=1}^{N_{ring}} r_j sin\left(\frac{2\pi(j-1)}{N_{ring}}\right) \\ r'' &= \sum_{j=1}^{N_{ring}} r_j cos\left(\frac{2\pi(j-1)}{N_{ring}}\right) \\ \hat{z} &= \frac{r' \times r''}{|r' \times r''|} \end{aligned} \tag{4.37}$$

where the position vectors r_j specify the positions of the atoms involved in the ring structure, the Cremer Pople coordinates are defined as follows:

$$q_m = (\frac{2}{N_{ring}})^{\frac{1}{2}} \left\{ \left[\sum_{j=1}^{N_{ring}} z_j cos\left(\frac{2\pi m(j-1)}{N_{ring}}\right) \right]^2 + \left[\sum_{j=1}^{N_{ring}} z_j sin\left(\frac{2\pi m(j-1)}{N_{ring}}\right) \right]^2 \right\}^{\frac{1}{2}}$$

$$\theta_m = \begin{cases} \arctan\left[-\frac{\sum_{j=1}^{N_{ring}} z_j sin\left(\frac{2\pi m(j-1)}{N_{ring}}\right)}{\sum_{j=1}^{N_{ring}} z_j cos\left(\frac{2\pi m(j-1)}{N_{ring}}\right)} \right], \text{ if } cos(\theta_m) > 0 \\ \arctan\left[-\frac{\sum_{j=1}^{N_{ring}} z_j sin\left(\frac{2\pi m(j-1)}{N_{ring}}\right)}{\sum_{j=1}^{N_{ring}} z_j cos\left(\frac{2\pi m(j-1)}{N_{ring}}\right)} \right] + \pi, \text{ if } cos(\theta_m) < 0 \\ sgn\left[sin(\theta_m)\right] \frac{\pi}{2}, \text{ if } cos(\theta_m) = 0 \end{cases} \tag{4.38}$$

The main advantage of ring coordinates is that a cycle can be described with only two coordinates: q_m, θ_m. Thus, a reduced dimensionality PES suitable for analysis in an IVR environment can be obtained by plotting the energy as a function of the two polar coordinates minimized with respect to all the remaining coordinates. This application is shown in Chap. 8.

4.5 Tautomers

Tautomers play an important role in chemistry and biology; the general properties of a compound are the result of the equilibrium between these different structures. Thus, tools that allow the detection of all possible tautomers are much needed especially in computational chemistry and drug discovery [108]. The perception of tautomers in Proxima is focused on keto-enol and imin-amin equilibria, given a specific geometry. This is possible thanks to our perception of electron pairs located on the atoms. In particular, these types of tautomerisms are of the 1,3 type, which means that the atoms involved are the atoms at distances 1,2, and 3 with respect to the hydrogen atom (distances in number of bonds). Proxima is capable of detecting the weight of a given tautomeric form (in a given input geometry) by checking the relative distance between the virtual site (the Lone Pair) of the acceptor atom (the atom that is forming a bond with the hydrogen during the tautomeric equilibrium) and the hydrogen atom. The maximum number of tautomers for such equilibria can be computed by:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{4.39}$$

With n the number of hydrogens that can be shifted and k the number of Lone Pairs that are available. This number gets reduced once Proxima has assigned the weight to each tautomer. Moreover, Proxima also distinguishes between a neutral tautomeric form and one with charge separation presenting the second one only if the molecule is considered in solvent and not in the gas phase. In Fig. 4.13, the tautomers of cycloserine detected by proxima are shown both for the gas phase and for the bulk (with charge separation).

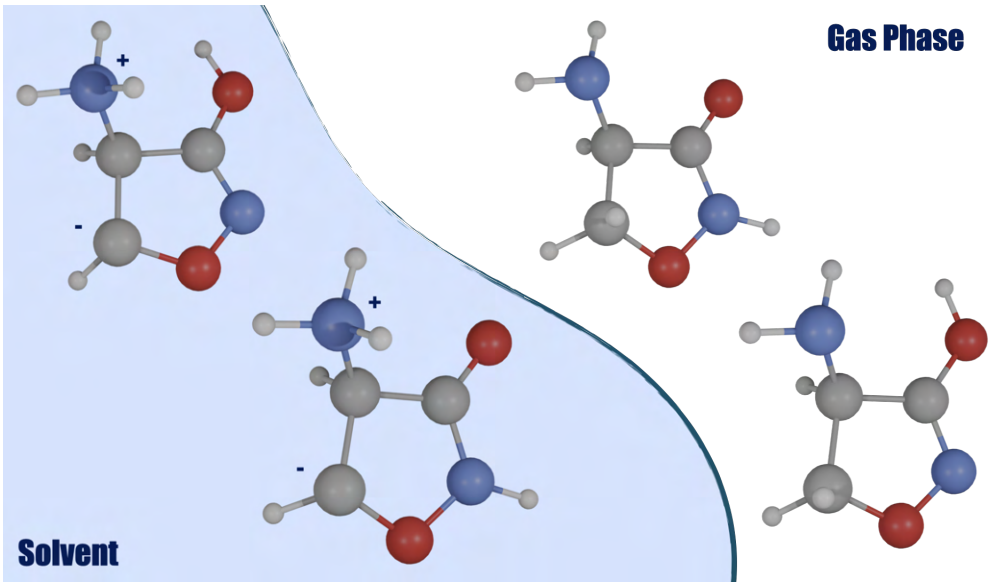


Figure 4.13: The tautomers of cycloserine detected in vacuum and in solvent with charge separation.

Proxima, by detecting the ring in these tautomeric forms, can also be integrated with one of our tools for the automatic assignment of ring coordinates.

4.6 Chiral centers

Chiral atoms are detected automatically by an approach reminiscent of the Molassembler model [109], which is based on IUPAC rules and involves three main steps, namely score assignment to the substituents of the investigated atom, ordering of those substituents and determination of the chirality.

4.6.1 Score assignment

Traditional priority rules are based on atomic numbers. An atom with a high atomic number has a higher priority. However, in comparing two atoms with the same atomic number, we have to look for the atoms bonded for differences in atomic numbers. In fact, the procedure is iterated until a difference between two substituents is found. The first step is to perform a BFS (Breadth-First Search) [93] on each substituent so as to subdivide it into levels and assign a score to each level. As an example, in Fig. 4.14, a molecule is shown with its subdivision in levels.

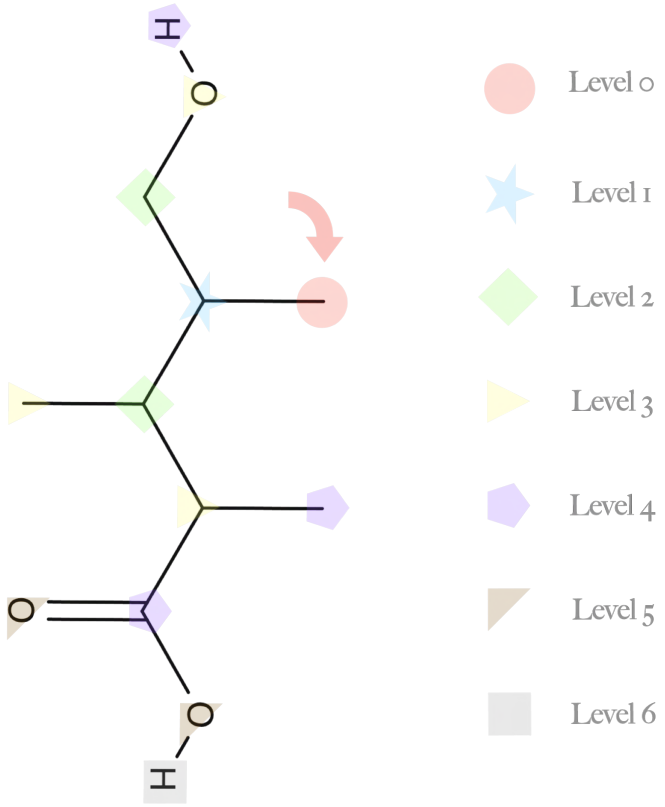


Figure 4.14: The subdivision in levels through the BFS on an example molecule. The starting point of the BFS is the atom pointed by the red arrow.

The BFS has a natural tendency to subdivide graphs into levels since is a "child-first" exploration of the molecular graph. The score of each level is a weighted sum of the atomic numbers involved with weights corresponding to the products of the bond orders of the bonds traveled from the origin atom to each of the atoms. This is coherent with the standard priority rules. In fact, an atom in a double bond should be considered twice and an atom in a triple bond should instead be considered three times. As mentioned above, continuous bond orders are employed in place of discrete ones, but the logic of the assignments remains the same. The score is assigned up to level 3 of each substituent, the maximum number of levels to consider for scoring can also be given as input by the user.

4.6.2 Order assignment

At the end of the previous step, we are left with a set of scores for each level of each substituent. In order to "sort" substituents, we start by looking at the score of the first level of all the substituents (that is the atomic number of the atom directly bonded to the chiral center). The substituents are sorted on the basis of this score. If there are subsets of substituents that have the same score, at the current level, the same procedure is repeated on each subset but using the scores of the subsequent level. In fact, this is repeated, going down on levels, up until there are substituents with equal scores (or we have reached the last level). If we reach the last level and there are still substituents with equal scores, we consider

the atom achiral. If not, an R/S assignment is done in the next step.

4.6.3 Chirality assignment

Once substituents are ordered, it is possible to verify the three-dimensional disposition so as to assign an R/S chiral state. In Fig. 4.15, a general chiral atom (C) bonded to its four substituents is shown with two configurations: R, S. Each substituent is identified by a label G_i, where i is the relative ordering of the substituent as determined in the previous step on the basis of the level scores.

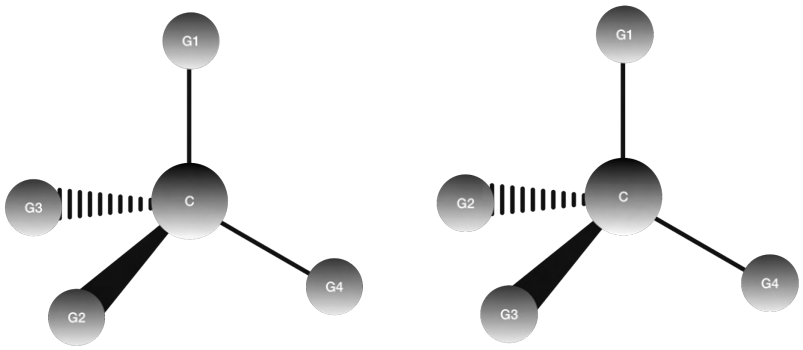


Figure 4.15: A central chiral atom (C) bonded to four substituents (G_1, G_2, G_3, G_4) in an R configuration (on the left) and an S configuration (on the right).

In this example, the ordering is the following: $G_1 > G_2 > G_3 > G_4$. Thus, G_4 is the substituent with the lowest score and G_1 is the substituent with the highest score. In the following, each substituent is identified by the position of the atom directly bonded to the central chiral atom. Thus, the $\vec{G}_1$ vector points towards the atom of the G_1 group bonded to the central C atom. The central C atom, of course, has its own $\vec{C}$ position. The assignment is performed as described below:

- R is assigned if:
  $((\vec{G}_4 - \vec{C}) \times (\vec{G}_1 - \vec{C})) \cdot (\vec{G}_2 - \vec{C}) \geq 0$ AND
  $((\vec{G}_4 - \vec{C}) \times (\vec{G}_1 - \vec{C})) \cdot (\vec{G}_3 - \vec{C}) \leq 0$

- S is assigned if:
  $((\vec{G}_4 - \vec{C}) \times (\vec{G}_1 - \vec{C})) \cdot (\vec{G}_2 - \vec{C}) \leq 0$ AND
  $((\vec{G}_4 - \vec{C}) \times (\vec{G}_1 - \vec{C})) \cdot (\vec{G}_3 - \vec{C}) \geq 0$

In other words, the G_2 group should be on the same side of the vector product between G_4 and G_1 in an R configuration, and on the other side in an S configuration.

4.6.4 Alkene and Allene Stereochemistry

The perception of alkene stereochemistry follows the same priority rules highlighted before on the four substituents of the alkene (two for each sp^2 atom). In this case, we check whether the two substituents on each atom that has the higher weight are on the same or opposite sides of the bond. In the first case, we assign a Z stereochemistry, E otherwise. The same approach is applied to the imine bond where the E/Z stereochemistry can also be interpreted as syn/anti stereochemistry. The perception of allene stereochemistry follows the same rules, but in this case, the traditional R/S geometric criterion is applied to the four substituents of the external bond of the allene.

4.7 Solvation procedures

The Proxima Molecular Perception library [77] is employed for the automatic solvation of molecular structures. The motivation for developing a new solvation tool, despite the presence of many others in the field (e.g. gromacs [110]), is that we needed a more custom option. Indeed, many solvation procedures have their own internal database of solvent structures. In our case, the solvent is loaded from a file explicitly given by the user and there are no limits on the type of structure required. In fact, it is also possible to use solid-state crystal structures as "solvent" (e.g. in studying the intercalation of a molecule in a periodic solid). The only requirement is to use PDB files containing the solvent structure in a cell. Moreover, the biggest advantage of having solvation procedures in Proxima is that are easily portable in different tools since Proxima is a multilanguage library (C++, Python). In the following, the procedure is illustrated both in theory and code.

4.7.1 Periodic Solvation

The general solvation procedure is available in the C++ version of Proxima and also in the Python version of Proxima (PyProxima). The starting point of these procedures is to instantiate a new SolventGenerator object and load a solute molecular system from a file. The code for these two steps is illustrated below:

```
C++
SolventGeneratorSP sgen(new SolventGenerator());
MolecularSystemSP solute = Parser::readPDB("solute.pdb");

PyProxima
sgen = pyproxima.PySolventGenerator()
solute = pyproxima.PyParser().readPDB(b"solute.pdb")
```

Only C++ code will be shown from this point. To load the solvent molecular structure (currently, only PDB files are supported) and add the solute to the SolventGenerator object, the *addCustomSolvent* and the *addSolute* methods have to be invoked.

```
sgen->setCustomSolvent("solvent.pdb");
sgen->addSolute(solute);
```

At this point, it is possible to replicate the solvent molecules so as to cover the space surrounding the solute up to a maximum distance r_{max}. The method is the following:

```
sgen->generateSolvent(r_max);
```

The operations this method performs are described in the next two subsections.

4.7.2 Sphere Generation

The portion of space that is required to be solvated is defined by a sphere with a radius of r_{max} surrounding the solute in the center of its bounding box. The center of the bounding box is preferred over the geometric center since it is not affected by the internal disposition of atoms but it rather accounts only for the overall shape of the molecule.

4.7.3 Cell Generation

Once this sphere is computed, we need to find an extension of the solvent cell that circumscribes the solute sphere. The solvent cell is defined by an origin vector and three generally non-orthogonal vectors oriented along the sides of the cell (see Fig. 4.16).

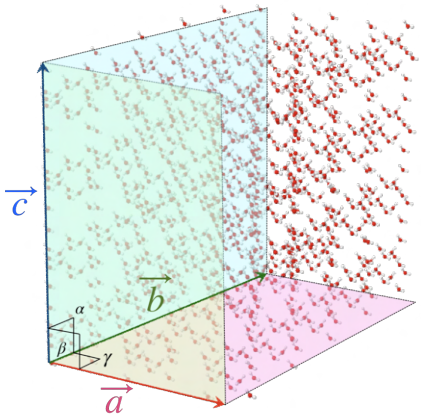


Figure 4.16: The $\vec{a}$, $\vec{b}$, $\vec{c}$ vectors define together the solvent cell. These are generally non orthogonal and their relative angles are $\alpha = \frac{\vec{b}\cdot\vec{c}}{||\vec{b}||\cdot||\vec{c}||}$, $\beta = \frac{\vec{a}\cdot\vec{c}}{||\vec{a}||\cdot||\vec{c}||}$, $\gamma = \frac{\vec{a}\cdot\vec{b}}{||\vec{a}||\cdot||\vec{b}||}$

It is possible to replicate the solvent system along its three cell axis vectors. Thus, there are two problems to solve:

- Finding the origin of the new cell that surrounds the sphere.
- Finding the number of replicas of the unit cell, along the axis vectors, so as to cover the entire sphere.

The first problem is solved by considering the three planes defined by the couples of vectors $(a, b), (b, c), (c, a)$ respectively (see Fig. 4.16 for the naming of the axes). Each one of these planes defines two other tangent planes to the sphere by simple translation (See Fig. 4.17).

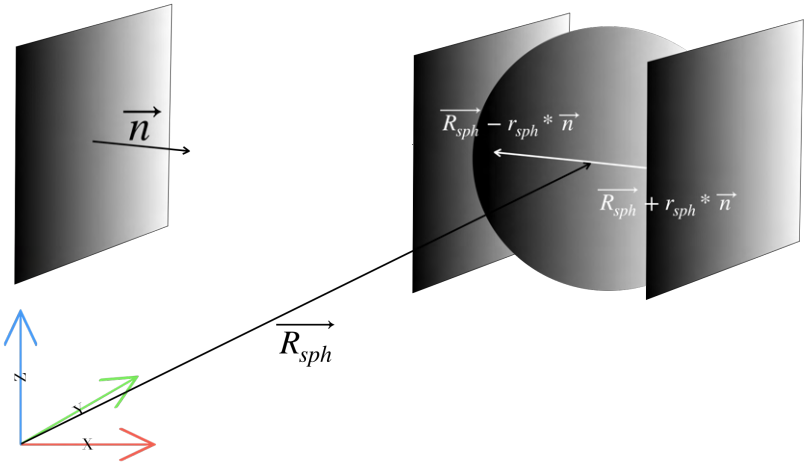


Figure 4.17: One plane defines two tangent planes to a sphere by simple translation. In the figure, $\vec{n}$ is the orthogonal vector to the plane that defines its orientation, $\vec{R}_{sph}$ is the center of the sphere, The tangent points are $\vec{R}_{sph} \pm r_{sph} \cdot \vec{n}$, where r_{sph} is the radius of the sphere.

To describe a plane, a vector perpendicular to the plane and an application point is needed. The perpendicular vectors are obtained by performing cross product between the cell vectors (so $\vec{a} \times \vec{b}, \vec{b} \times \vec{c}, \vec{c} \times \vec{a}$). The application points are instead obtained by summing together the center of the sphere with a vector having the radius of the sphere as length and oriented as the plane vector (see Fig. 4.17). Thus, since there are three cell planes and one sphere, there is a total of $3 \cdot 2 = 6$ tangent planes to the sphere. The interest is in finding the origin point of the new cell, this is obtained by the intersection of the three planes having the following application points: $\vec{R}_{sph} - \frac{\vec{a} \times \vec{b}}{||\vec{a} \times \vec{b}||} r_{sph}$, $\vec{R_{sph}} - \frac{\vec{b} \times \vec{c}}{||\vec{b} \times \vec{c}||} r_{sph}$, $\vec{R}_{sph} - \frac{\vec{c} \times \vec{a}}{||\vec{c} \times \vec{a}||} r_{sphere}$. This origin point is the $\vec{O}$ vector in Fig. 4.18.

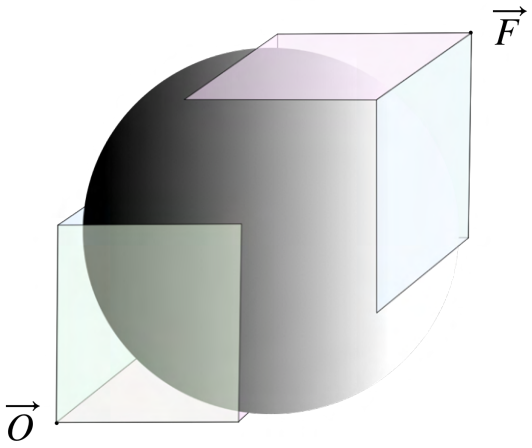


Figure 4.18: The $\vec{O}$ point is the origin of the new cell, the $\vec{F}$ point instead is the furthest away point from the origin of the cell. These points are obtained by the simple intersection of the tangent planes to the sphere.

To solve the second problem instead, we first need to determine the point of the new cell that is the furthest away from the origin (the $\vec{F}$ point in Fig. 4.18). It has been already said that each plane defines two tangent planes to the sphere by simple translation. In order to find $\vec{F}$, we need to find the intersection between the planes with the following application points: $\vec{R}_{sph}+\frac{\vec{a}\times\vec{b}}{||\vec{a}\times\vec{b}||}r_{sph}$, $\vec{R}_{sph}+\frac{\vec{b}\times\vec{c}}{||\vec{b}\times\vec{c}||}r_{sph}$, $\vec{R}_{sph}+\frac{\vec{c}\times\vec{a}}{||\vec{c}\times\vec{a}||}r_{sph}$. The problem of finding the number of replicas of the cell along each axis so as to include the $\vec{F}$ point, in an orthogonal cell, would be simply solved by dividing the distance of the point from the origin by the length of each side of the cell. The extension to nonorthogonal reference axes requires the definition of a matrix having the cell vector axes as columns:

$$\boldsymbol{M} = \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \tag{4.40}$$

Then we define the $\vec{r}$ vector as $\vec{F}-\vec{O}$. To find the number of replicas of the cell along each axis (l, m, n) all that is required is to multiply the inverse matrix of $\boldsymbol{M}$ by $\vec{r}$.

$$\begin{bmatrix} l \\ m \\ n \end{bmatrix} = \boldsymbol{M}^{-1}\vec{r} \tag{4.41}$$

Once a solvent molecular system is loaded, all the atoms are translated so as to match the origin of the coordinates with the origin of the new cell ($\vec{O}$). Then, each solvent atom is duplicated $l \cdot m \cdot n$ times so as to cover all the space around the solute.

4.7.4 Ellipsoid Cell

It is also possible to solvate the solute in a non-spherical manner, by rotating the solute along its inertia axes. Then, the bounding box and the whole procedure are

the same as before (just that now we are in a rotated system along the principal axes of inertia). At this point, the last difference is the criterium used for selecting solvent molecules using the analytical expression for the ellipsoid instead of the sphere for the check of the centroid.

4.7.5 Further Refiniments

At this point, the solute is solvated with the solvent but there are still some refinements that have to be performed:

- Since the solvent cell does not have spherical symmetry, there are extra solvent molecules not required.
- Some solvent molecules might collide, or be too close, to the solute.

The first problem is easy to solve, since all is required is to erase solvent molecules whose centroid lies above a given threshold from the solute. In this way, we are spherically symmetrizing the added solvent. In code, this is done by calling the following method:

```
sgen->removeExtraSolvent(d);
```

d is the distance threshold in Angstrom above which solvent molecules are removed. The second problem is solved in a similar way. In this case, we check that each solvent molecule is not "too close" (below a given threshold) to the solute. However, we are not using the centroid of each molecule to test this condition but rather using each atom's position since we have to be really sure there are no atomic collisions in the system. This operation is performed with the following method:

```
sgen->removeSolventCloseToSolute(d);
```

d is the distance threshold in Angstrom below which solvent molecules are removed. To retrieve the final solvated molecular system we can finally call the $getMolecularSystem$ method:

```
MolecularSystemSP ms = sgen->getMolecularSystem();
```

Chapter 5

The Chemical Feature Space

In previous chapters, we stressed the importance of computing main topological features such as bond orders (both σ and π) and charges, and we already anticipated how such quantities can be employed in defining custom Feature Spaces such as the one for defining an atom type. The atom type is a classic example of an entity that a chemist innately knows, which is hard to define formally. In fact, very many definitions of atom types have been given depending on the choice of the Force Field [42, 44, 45, 111]. There is a general consensus to identify the atom type with the "local environment" of a given atom in a molecule, assigning the same type to different atoms placed in the same local environment. However, a clear definition of what is this "local environment" is far from reached. Some Force Fields hugely rely on the hybridization status of the atom in a molecule (e.g. UFF [44]), while others rely on its presence in functional groups (e.g. AMBER [45]). It was of our interest to find a unique global definition of continuous atom types; the advantage of finding a more general definition is not only in simplifying the discussions around the theme but it also opens the possibility of applying Machine-Learning techniques to automate complex tasks that would otherwise be done by hand [10]. In building a continuous atom-type definition, we wanted to make sure to describe all possible variations of what is known as the "local environment" for an atom. The reason is that we don't want to introduce a new extra layer of chaos by adding a new definition of atom type, but instead, we want to generalize the already existing atom types to a single continuous definition. In other words, our continuous atom types should also be capable of describing old discrete atom types, as will be validated in Chap. 8.

5.1 The Atom Type

The definition of a "local environment" for an atom clearly has to do with its connections in a molecule. Given a set of atoms described by atomic numbers and charges ($\{Z_i, q_i\}$), and a bond order matrix defining the relative interactions between couples of atoms (BO_{ij}), we can then define the atom type as a continuous 4-dimensional quantity:

$$\vec{AT}_i = \begin{bmatrix} \prod_j (1 + BO_{ij}) \\ \sum_j \sqrt{Z_i Z_j} BO_{ij} \\ q_i \\ \sum_{j,k} BO_{ij} BO_{ik} BO_{jk} \end{bmatrix} \tag{5.1}$$

The Bond Order matrix can be derived both from electronic structure computations or by Proxima and, in general, the "locality" of the environment comes from the fact that a well-behaved bond order goes to 0 as two atoms are not interacting. Each element of this continuous atom type is called a feature, namely:

- The delocalization feature
- The coordination feature
- The charge feature
- The rigidity feature

5.1.1 The delocalization feature

The first element $\vec{AT}_{i1}$ is called the delocalization feature and is defined as follows:

$$\prod_j (1 + BO_{ij}) \tag{5.2}$$

The product contains the bond orders with each other atom j in the system scaled by 1 so that when two atoms are not interacting it doesn't vanishes, thus avoiding singularity problems in the product (it would always be 0). The reason for its name is related to its strong correlation with the hybridization of the atom and in particular with its delocalization. In order to show some properties of such delocalization feature, and the origin of its name, let's consider the simplified molecular graph G(V,E) obtained by Proxima after performing topology perception, where each node in the graph is an atom and each edge is a bond. As a consequence, instead of having to compute BO_{ij} for every pair of atoms, we just need the value for those couples which are connected in the molecular graph after performing perception. It is important to notice how the relations we are going to show are formally valid only in the case of acyclic graphs, so if we use a continuous BO_{ij} matrix issued from quantum chemical methods these relations do not hold because if we represent each vertex of the graph with an atom, it has a connection with all the others thus creating a cyclic graph by definition. However, such properties are still useful to discuss so as to motivate the reasoning behind its name "delocalization". An interesting mathematical property is that in principle this feature depends on the overall connectivity of the graph. In fact, it is possible to express the feature of a given vertex as a function of the features of all the others (thus only N-1 features are required to be known).

Theorem 5.1.1. *Given a connected acyclic graph G(V,E),*

$$\vec{AT}_{i1} = \prod_{j \in B(i)} (1 + BO_{ij}), \forall i \in V \implies \vec{AT}_{i1} = \prod_j \vec{AT}_{j1}^{(-1)^{d_{ij}+1}} \tag{5.3}$$

Here, d_{ij} is the minimum distance from vertices i and j in number of edges, and B(i) is the set of vertices directly connected to i.

Proof. Knowing that $\vec{AT}_{i1} = \prod_{j \in B(i)}(1 + BO_{ij}), \forall i \in V$ we can express each edge value $1 + BO_{ij}$ as dependent from the values of the connected vertices

$$1 + BO_{ij} = \frac{\vec{AT}_{j1}}{\prod_{i' \neq i}(1 + BO_{i'j})} \tag{5.4}$$

By iterating such relation for each i' vertex, moving along the graph we obtain:

$$1 + BO_{ij} = \frac{\vec{AT}_{j1}}{\prod_{i' \in \{C_j(i)\}} \vec{AT}_{i'1}^{(-1)^{d_{ji'}+1}}} \tag{5.5}$$

Where $C_j(i)$ is the set of vertices that are reachable from j by moving through the molecular graph obtained by disconnecting the i atom from j, and $d_{ji'}$ is the minimum distance between the j and the i' vertices measured in number of edges. By substituting this relation for each $1 + BO_{ij}$ in the definition of $\vec{AT}_{i1}$ we have proven such relation. □

In fact, it is also possible to prove an explicit relation between the features of connected atoms:

Corollary 5.1.1.1. *Given a connected acyclic graph G(V,E),*

$$\vec{AT}_{i1} = \prod_{j \in B(i)} (1 + BO_{ij}), \forall i \in V \implies \frac{\vec{AT}_{i1}}{\vec{AT}_{j1}} = \frac{\prod_{j' \in \{C_i(j)\}} \vec{AT}_{j'1}^{(-1)^{d_{ij'}+1}}}{\prod_{i' \in \{C_j(i)\}} \vec{AT}_{i'1}^{(-1)^{d_{ji'}+1}}}, \forall \{i,j\} \in E \tag{5.6}$$

Proof. By isolating the edge value of the edge connecting the i and j vertices we get:

$$\begin{cases} \vec{AT}_{i1} = (1 + BO_{ij}) \prod_{j' != j} (1 + BO_{ij'}) \\ \vec{AT}_{j1} = (1 + BO_{ij}) \prod_{i' != i} (1 + BO_{ji'}) \end{cases} \tag{5.7}$$

By equating $1 + BO_{ij}$ from both equations and rearranging we have proven the relation. □

We have thus shown how to move from an edge function space $\{1 + BO_{ij}\}$ to a vertex function space $\{\vec{AT}_{i1}\}$, however, it is interesting to notice how in principle it is also possible to move from the Feature Space to the bond orders if some criteria are respected by the vertex features. In particular:

Theorem 5.1.2. *Given a connected acyclic graph G(V,E) and a vertex function $\vec{AT}_{i1}, \forall i \in V$,*

$$\vec{AT}_{i1} = \prod_{j} \vec{AT}_{j1}^{(-1)^{d_{ij}+1}}, \forall i \implies \vec{AT}_{i1} = \prod_{j \in B(i)} (1 + BO_{ij}) \tag{5.8}$$

Where $1 + BO_{ij}$ is an edge function ($\{i,j\} \in E$) and B(i) is the set of vertices directly connected to i by some edge.

Proof. Knowing that $\vec{AT}_{i1} = \prod_j \vec{AT}_{j1}^{(-1)^{d_{ij}+1}}$ we can separate the product by separating the features of the vertices directly connected to the central i vertex (that we can define as the B(i) set of vertices) as follows:

$$\vec{AT}_{i1} = \prod_{j \in B(i)} \vec{AT}_{j1} \prod_{j' \in B(j) \setminus i} \vec{AT}_{j'1}^{(-1)^{d_{jj'}+1}} \tag{5.9}$$

Notice how $\vec{AT}_{j1} \prod_{j' \in B(j)i} \vec{AT}_{j'1}^{(-1)^{d_{jj'}+1}}$ only depends on j and i (since the space of j' is $B(j) \setminus i$) thus can be expressed as an edge function proofing the above relation. □

Moreover, it is also possible to prove that if the relation above is valid for just one vertex of the graph, then it is valid for all of them for positive defined features (such as the ones used in molecular graphs).

Theorem 5.1.3. *Given a connected acyclic graph G(V,E) and a vertex function* $\vec{AT}_{i1} | \vec{AT}_{i1} > 0, \forall i \in V$ *,*

$$\exists i \in V | \vec{AT}_{i1} = \prod_j \vec{AT}_{j1}^{(-1)^{d_{ij}+1}} \implies \vec{AT}_{i1} = \prod_j \vec{AT}_{j1}^{(-1)^{d_{ij}+1}}, \forall i \in V \tag{5.10}$$

Proof. If exists at least one vertex $i \in V$ so that $\vec{AT}_{i1} = \prod_j \vec{AT}_{j1}^{(-1)^{d_{ij}+1}}$, we can use this expression to isolate the $\vec{AT}_{k1}$ value for a general vertex $k \in V$, so that $d_{ij} = d_{ik} + d_{jk}$, as follows:

$$\vec{AT}_{i1} = \vec{AT}_{k1}^{(-1)^{d_{ik}+1}} \prod_{j \neq i,k} \vec{AT}_{j1}^{(-1)^{d_{ij}+1}} \tag{5.11}$$

By inverting such a relation we get:

$$\vec{AT}_{k1}^{(-1)^{d_i k+1}} = \frac{\vec{AT}_{i1}}{\prod_{j \neq i,k} \vec{AT}_{j1}^{(-1)^{d_{ij}+1}}} = \vec{AT}_{i1} \prod_{j \neq k,i} \vec{AT}_{j1}^{(-1)^{d_{ij}}} \tag{5.12}$$

Since $d_{ii} = 0$, we can include back $\vec{AT}_{i1}$ in the product:

$$\vec{AT}_{k1}^{(-1)^{d_i k+1}} = \prod_{j \neq k} \vec{AT}_{j1}^{(-1)^{d_{ij}}} \tag{5.13}$$

Since we have a positively defined vertex function, we can take the root of the relation above by isolating $\vec{AT}_{k1}$:

$$\vec{AT}_{k1} = \prod_{j \neq k} \vec{AT}_{j1}^{(-1)^{d_{ij}-d_{ik}+1}} \tag{5.14}$$

Since $d_{ij} = d_{ik} + d_{jk}$, we get the following relation for a generic $k \in V$ which proves the thesis.

$$\vec{AT}_{k1} = \prod_{j \neq k} \vec{AT}_{j1}^{(-1)^{d_{jk}+1}} \tag{5.15}$$

□

In Fig. 5.1, an example is shown with a simplified graph.

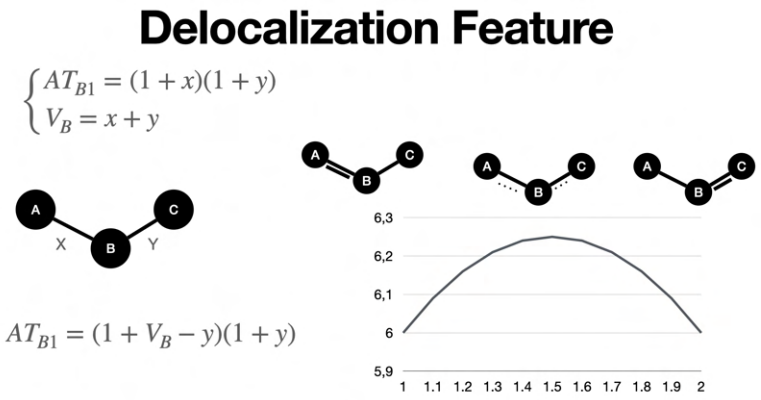


Figure 5.1: An example of the delocalization feature for a central node connected to other two nodes in a graph given an edge function BO_{ij}.

In that case, a valence constraint to the central atom is introduced ($V_i = \sum_j BO_{ij}$), proving that the delocalization feature actually represents the delocalization around the atom.

5.1.2 The Coordination Feature

The second component in the continuous atom type definition is the coordination feature defined as:

$$\sum_j \sqrt{Z_i Z_j} BO_{ij} \tag{5.16}$$

This feature can be thought of as the coordination around the atom scaled by atomic numbers, a sort of average of the square difference in atomic numbers, so as to distinguish the same atom that is connected to different numbers of heteroatoms.

5.1.3 The Charge Feature

The charge can be written as the summation of a partial charge (q_i^δ) and the formal charge (q_i^+):

$$q_i = q_i^\delta + q_i^+ \tag{5.17}$$

In this way, the same atom but in different ionic or polarization states can be distinguished. The partial charge employed in our applications is the one of Proxima discussed in the previous chapter.

5.1.4 The Rigidity Feature

The rigidity feature is defined as:

$$\sum_{j,k} BO_{ij} BO_{ik} BO_{jk} \tag{5.18}$$

The idea behind the rigidity feature is to distinguish atoms placed in very rigid structures such as the short cycles cyclopropane or cyclobutane. In Fig. 5.2, the formation of cyclopropane from propane is shown so as to emphasize the increase in strength in the BO_{AC} intensity.

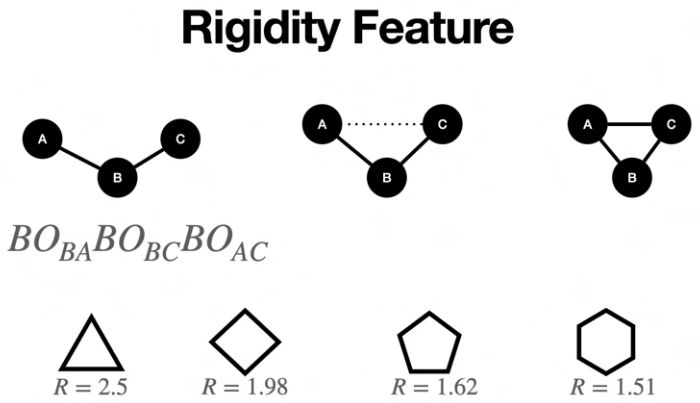


Figure 5.2: The rigidity feature.

In computing the feature, all possible triplets of atoms must be considered, but in the case of a simplified molecular graph such as the one computed by Proxima through perception, it is possible to impose that j and k are geminal atoms which are atoms connected to the same atom after perception occurred.

5.2 The synthon

The Feature Space introduced in describing atom types can be extended to ensembles of atoms. In fact, the simplest set of atoms coupled together is the covalent bond. In order to describe the covalent bond as a single vector in a Feature Space, 4 features per atom are required plus an additional feature that can be the bond order between the atoms connected. It is important, however, to define an ordering between the atoms so as to avoid ambiguities in defining such a vector. The natural way is to apply traditional stereochemistry rules to determine a priority between the atoms involved in a bond, following such nomenclature we can then define the synthon as the bond type vector in the Feature Space between atom I and J as:

$$\vec{S} = \begin{bmatrix} D_I \\ C_I \\ Q_I \\ R_I \\ D_J \\ C_J \\ Q_J \\ R_J \\ BO_{IJ} \end{bmatrix} \tag{5.19}$$

The same treatment can be easily extended to multiples ensembles of atoms, following the same logic of defining a common ordering of the atoms within the group and using their features, together with some geometrical descriptors of the ensemble if explicitly needed by the task studied, to get the best description of the system. In the case of the valence angle, the bond order between the geminal atoms (the BO_{jk}) can be taken as an explicit feature in the synthon definition. The advantage of our features is that, in principle, they can be used in excited or transition states since such a relation must be conveyed to the descriptors by the input bond order matrix and charges. In this thesis, the goal is to work on fixed-topology structures but these descriptors were designed to be as flexible as possible for future studies. In the next chapter, the analytical equations of some energy profiles are discussed to provide the foundation for the development of a future Force Field.

Chapter 6

The Energy Profiles

The definition of an atom type in the previous chapter prompted us to think about a possible application in the Molecular Mechanics field by prototyping a tool that automatically creates a custom Force Field for a given molecule or set of molecules. Once these molecules are provided by the user, the Proxima library should be able to cluster atoms that are similar to the others by means of the atom type as defined in the previous chapter. The same applies to bond and angle types. Then, by computing hessian and gradients by Quantum Chemistry methods, we can subtract the non-bond van der Waals and electrostatic energies (computed with the Molecular Perception algorithms shown in previous chapters) from the QM values obtaining the relevant energy profiles for the desired variable. In principle, if the molecule is small enough, it is not necessary to cluster atoms and bonds and the hessian and gradient can be computed for the overall molecule. If the molecule is big, however, the Proxima library should extract a subset of fragments of interest for each characteristic bond and angle so as to compute derivatives on those smaller fragments thus reducing the computational cost. In choosing the functional form of energy, our guess is that it is possible to express most energy contributions with the same formal expression. In particular, the idea is to shift from the physical space of cartesian coordinates ($r_i = [x_i, y_i, z_i]$) to our Feature Space of bond orders and charges. In fact, by performing a suitable coordinate mapping between cartesian coordinates and bond orders it is possible to move from one space to the other. This bond order mapping is motivated by the correct asymptotic behavior of bond orders that go to zero as distances increase. Although Proxima [77] computes bond orders as a combination of a σ and a π component, because of the imprecise nature of the input geometries in perception procedures that might not be a trustworthy, most common mapping between these two spaces rely on the radial distances between atoms. Most commons have an exponential form such as the one originally proposed by Pauling [112]:

$$BO_k = e^{-\alpha(r_k - r_e)} \tag{6.1}$$

This expression is well-behaved in the case of asymmetric potential expressions. However, we can expand this definition of the bond order by defining a custom mapping as:

$$BO_k = e^{-\alpha(r_k - r_e)^n} \tag{6.2}$$

With $n = 1$ we obtain the traditional bond order but with $n = 2$ we obtain

a Gaussian, thus accounting for symmetric environments. In general, the energy profile of a coordinate can be expressed as a polynomial of the bond order:

$$E_k = f(BO_k) = P_m\left((r_k - r_e)^l - BO_k\right) = P_m\left((r_k - r_e)^l - e^{-\alpha(r_k - r_e)^n}\right) \quad (6.3)$$

In this case, with $l = 0$, $n = 2$, and $m = 1$ we obtain a Gaussian-like equation that is well-suited for symmetric potentials. In the case of $l = 0$, $n = 1$, and $m = 2$, instead, we obtain the traditional Morse equation that is well suited for asymmetric potentials. The explicit inclusion of radial coordinates r_k is necessary to correctly describe dispersions, such as in the case of non-covalent interactions (e.g. hydrogen bond), and double well potentials, such as in tautomerism. In general, the n degree is related to the symmetry state of the energy profile and it is possible to expand the polynomial P_m to higher degrees. However, in the following, we are going to focus on simple energy profiles, and further developments of this bond order mapping are left for future works.

6.1 Single Well Potential

The single well potential is the typical potential for stretchings and bendings around 180° such as in the case of the HCN molecule. It is also the case of van der Waals and hydrogen bond interactions where the known 6-12 and 10-12 expressions are taken as a reference. Here we are going to discuss the functional form in the case of a symmetric and asymmetric potential.

6.1.1 Symmetric

In the case of a symmetric single well potential, a simple formula is employed using a Gaussian ($l = 0$, $n = 2$, $m = 1$):

$$E_k = \epsilon\left(1 - e^{\alpha(x_k - x_m)^2}\right) \quad (6.4)$$

In this case, the use of a Gaussian correctly describes the behavior around the minimum x_m by forcing the third derivative to go to zero. In order to determine the two parameters α and ϵ, two different strategies can be employed:

- Computing the second and fourth derivative values in the minimum forcing the energy profile to assume the same values.
- Computing the second derivative around the minimum while also forcing the functional form to pass through a chosen point ($(E(x_i) = E_{ref})$).

In general, two conditions are required to compute the two parameters. For completeness, we report the required derivatives:

$$\begin{cases} \dfrac{\partial E_k}{\partial x_k} = -2\epsilon\alpha(x_k - x_m)e^{\alpha(x_k-x_m)^2} \\ \dfrac{\partial^2 E_k}{\partial x_k^2} = -2\epsilon\alpha e^{\alpha(x_k-x_m)^2} - 4\epsilon\alpha^2(x_k - x_m)^2 e^{\alpha(x_k-x_m)^2} \\ \dfrac{\partial^3 E_k}{\partial x_k^3} = -12\epsilon\alpha^2(x_k - x_m)e^{\alpha(x_k-x_m)^2} - 8\epsilon\alpha^3(x_k - x_m)^3 e^{\alpha(x_k-x_m)^2} \\ \dfrac{\partial^4 E_k}{\partial x_k^4} = -12\epsilon\alpha^2 e^{\alpha(x_k-x_m)^2} - 48\epsilon\alpha^3(x_k - x_m)^2 e^{\alpha(x_k-x_m)^2} - 16\epsilon\alpha^4(x_k - x_m)^4 e^{\alpha(x_k-x_m)^2} \end{cases} \tag{6.5}$$

The third and the first derivative around x_m go to zero as intended, while the second derivative goes to $-2\epsilon\alpha$ and the fourth goes to $-12\epsilon\alpha^2$.

6.1.2 Asymmetric

In the case of an asymmetric single well potential, the expression is taken with $l = 0$, $n = 1$ and $m = 2$, obtaining the traditional Morse equation:

$$E_k = \epsilon\left(e^{\alpha\left(1-\frac{x_k}{x_m}\right)} - 2e^{\frac{\alpha}{2}\left(1-\frac{x_k}{x_m}\right)}\right) \tag{6.6}$$

In a recent work about van der Waals molecular energies it has been proven that a more general expression of the Morse equation including a polynomial provides better results [99]. The motivation for introducing a polynomial multiplying the exponential is the presence of long-range interactions due to dispersion forces that scale as r^{-6} and r^{-8}:

$$E_k = \epsilon\left(e^{\alpha\left(1-\frac{x_k}{x_m}\right)} - \left[\left(\frac{x_k}{x_m}\right)^{2n} - 2\left(\frac{x_k}{x_m}\right)^{n} + 3\right] e^{\frac{\alpha}{2}\left(1-\frac{x_k}{x_m}\right)}\right) \tag{6.7}$$

Although the expression seems far more complicated than the simple Gaussian-like for the symmetric potential, it is worth noting that the only parameter to be determined is the α exponential parameter since the depth of the well ϵ is known given the asymptotic behavior of the energy that goes to zero as the variable approaches infinity. The only condition worth applying is imposing the correct second derivative around the minimum x_m, here for simplicity we just show the second-order derivatives around the minimum:

$$\begin{cases} \left(\dfrac{\partial^2 E_k}{\partial x_k^2}\right)_{x_m}^{n=2} = \alpha^2\dfrac{\epsilon}{2x_m^2} - \dfrac{8\epsilon}{x_m^2} \\ \left(\dfrac{\partial^2 E_k}{\partial x_k^2}\right)_{x_m}^{n=0} = 2\epsilon\alpha^2 \end{cases} \tag{6.8}$$

6.2 Double Well Potential

The double well potential is typical of bending interactions where the reference angle is not 180° (e.g. H_2O). However, even some proton transfer phenomena

show such double-well behaviors [113, 114]. The reference angle is not the minimum energy angle, but the angle at which we have the energy barrier going from one well to the other. As an example, taking the H_2O molecule, the reference angle (if we consider the H-O-H angle) is at 180° with one well located at (180° - 104.7° = 75.3°) and the other at (180° + 75.3° = 255.3°). The behavior of the third derivative of the potential around the reference angle determines the asymmetry of the potential. In general, as we show, the logic is to use the summation of the power of the variable x_k with a Gaussian, creating the asymmetry by shifting the two expressions one with respect to the other. In principle, a double well potential can still be described in terms of Morse equations summed together and the choice of which expression to use relies on the shape of the potential as shown in the next section.

6.2.1 Symmetric

Taking the reference angle as the origin of our coordinate system, we can define a symmetric double well potential as [115]:

$$E_k = ax_k^{2n} + be^{-\alpha x_k^2} - b \tag{6.9}$$

The nice property of centering both power and Gaussian in the reference angle is that the third derivatives of both the power and the Gaussian go to zero. In this case, the number of parameters to be determined is equal to three (a, b, α). As a consequence, three conditions are imposed:

$$\begin{cases} E(x_e) = \epsilon \\ \left(\dfrac{\partial E}{\partial x_k}\right)_{x_e} = 0 \\ \left(\dfrac{\partial^2 E}{\partial x_k^2}\right)_{x_e} = K \end{cases} \tag{6.10}$$

Thanks to the symmetry of the problem, we only have to worry about one of the two wells (x_e with respect to the location of the barrier) to get the correct behavior. It is important to remember that the depth of the well, ϵ is a negative number since the zero of the energy is taken as the barrier. The nice consequence of this is that we can get a global profile of the potential only considering the local properties around the minimum. By imposing the three conditions above we arrive at the following transcendental equation:

$$\epsilon e^{-\alpha x_e^2}\left[\frac{2n(2n-2)}{n}\alpha + 4\alpha^2 x_e^2\right] = K\left[e^{-\alpha x_e^2} + \frac{\alpha x_e^2}{n}e^{-\alpha x_e^2} - 1\right] \tag{6.11}$$

Solving numerically such an equation gives us the α value for the given second derivative K. Once α is determined, it is possible to obtain back the b parameter as:

$$b = \frac{\epsilon}{e^{-\alpha x_e^2} + \frac{\alpha x_e^2}{n}e^{-\alpha x_e^2} - 1} \tag{6.12}$$

And the a parameter as:

$$a = \frac{\alpha x_e^{2-2n}}{n} b e^{-\alpha x_e^2} \tag{6.13}$$

Transcendental equations might not always have a solution, and in order to find the conditions in which they can be solved it is convenient to rewrite the transcendental equation isolating the α variable on the left side [115]:

$$\alpha = \frac{K}{\epsilon} \frac{n}{2n(2n-2)} e^{\alpha x_e^2} \left[\frac{\alpha x_e^2}{n} e^{-\alpha x_e^2} + e^{-\alpha x_e^2} - 1 \right] - \frac{2\alpha^2 x_e^2}{2n-2} \tag{6.14}$$

The left side is a straight line with an angular coefficient equal to 1 ($y = \alpha$). The right side, for $\alpha \to \infty$ diverges and goes to infinite positive (remember that the depth of the well ϵ is a negative number). Moreover, for $\alpha \to 0$, the right side of the equation goes to zero. Thus, in order for the transcendental equation to have a solution, the right side should intersect with the left side and to do so its first derivative for $\alpha \to 0$ must be below 1 (the first derivative of the left side). In case $n = 1$, this equation has a solution if the following condition is satisfied:

$$\frac{K}{32|\epsilon|} D^2 < 1 \tag{6.15}$$

In the general case, for $n > 1$, this other condition must be satisfied:

$$\frac{K}{16|\epsilon|} \frac{D^2}{n} < 1 \tag{6.16}$$

Where D is the inter-minima distance ($D = 2x_e$). The presence of the degree of the polynomial n in these conditions allows us to employ this equation for many different types of wells since it is just necessary to change the degree of the polynomial to find a suitable potential. In fact, given a well-defined by its x_e, ϵ, K values, it is always possible to find a degree n for which this potential has a solution since this condition changes as $1/n$ which tends to 0 (which is below 1) for $n \to \infty$. In cases this equation has no solutions for the desired n, in addition to changing the n itself, it is also possible to express a double-well potential with the summation of two Morse equations [116]:

$$E(\zeta) = \nu_0 - [2\nu_0 e^{-\zeta_0}] cosh(\zeta) + [\nu_0 e^{-2\zeta_0}] cosh(2\zeta) \tag{6.17}$$

where $\zeta = \beta x$ and $\zeta_0 = \beta x_0$. The parameters ν_0, β, x_0 can be uniquely determined from ϵ, D, K. The equations are:

$$\begin{aligned} cosh(\zeta_k) &= \frac{1}{2} e^{\zeta_0} \\ \zeta_k &= \pm \beta D/2 \\ \epsilon &= \frac{1}{2} \nu_0 [1 - 2e^{-\zeta_0}]^2 \\ K &= \beta^2 \nu_0 [1 - 4e^{-2\zeta_0}] \end{aligned} \tag{6.18}$$

From these equations, we obtain a transcendental equation in β:

$$\beta = \left(\frac{K}{2\epsilon} \right)^{\frac{1}{2}} \left(\frac{1 - 2e^{-\zeta_0}}{1 + 2e^{-\zeta_0}} \right)^{\frac{1}{2}} \tag{6.19}$$

The limit of such an approach though is that it admits real solutions if, and only if,

$$\frac{1}{4}D\sqrt{\frac{K}{2\epsilon}} > 1 \tag{6.20}$$

Which is a strict condition on the shape of the well, thus not flexible enough.

6.2.2 Asymmetric

In case asymmetry is introduced, it is still possible to use the same expression by shifting the Gaussian with respect to the center of the parabola, so as to create an asymmetry by changing the location of the energy barrier and the depths of the two wells. As an example, we can center the power at S and the Gaussian at C:

$$E_k = a(x_k - S)^{2n} + be^{-\alpha(x_k - C)^2} \tag{6.21}$$

The problem with such an expression is that we lose every nice symmetry property and the energy barrier does not coincide either with S or C, in order to determine the energy barrier the first derivative must be computed and set to zero. Having now 5 parameters to select (a, S, b, α, C), it becomes harder to decouple the problem as done before by obtaining a global energy profile just from local derivatives around the minimum (obtaining at least 6 conditions, 3 for well, for 5 parameters). As a consequence, a more general fitting procedure is employed by means of differential evolution algorithms that try to mimic the profile of the energy going from one well to the other. This is the only case where a scan of the energy profile is required to compute the correct parameters. The differential evolution algorithms work in an iterative fashion by shifting power and Gaussian, while changing the other parameters, until it does not converge minimizing the absolute value of the maximum difference between the energy and the predicted value in the scan region.

6.3 Electrostatic

The electrostatic energy term, together with the van der Waals, comes directly from Molecular Perception algorithms, and new QM computations are not required to compute this term when building the force field for a molecule.

$$E(r) = 332.0636\frac{Q_i Q_j}{r_{ij}} \text{ (kcal/mol)} \tag{6.22}$$

r_{ij} is the Angstrom distance between the pair of atoms considered and Q_i are the partial charges derived from the perception algorithms. The electrostatic energy does not have a well and is only considered for atoms that are more than three covalent bonds away from each other (as for van der Waals).

6.4 Cartesian Gradient and Hessian

It is worth noticing that all of the radial potential expressions we have employed are of the form:

$$\sum_{j>i} V(r_{ij}) \tag{6.23}$$

It is of general interest to build tools that automatically compute Gradient and Hessian in cartesian coordinates based on a general radial pair function $V(r_{ij})$. Thus, in the following, we are going to show such mathematical equations for completeness. These will be employed in the test case of the glycine dipeptide analog in the relative chapter. To start, let's take the main expression that relates radial coordinates (r_{ij}) to cartesian coordinates (x, y, z):

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{6.24}$$

The first important equation to derive that will be useful later is the first derivative of r_{ij} with respect to a generic variable x, y, or z of atom i or j:

$$\begin{cases} \dfrac{\partial r_{ij}}{\partial x_i} = \dfrac{(x_i - x_j)}{r_{ij}} \\ \dfrac{\partial r_{ij}}{\partial x_j} = \dfrac{(x_j - x_i)}{r_{ij}} \\ \dfrac{\partial r_{ij}}{\partial y_i} = \dfrac{(y_i - y_j)}{r_{ij}} \\ \dfrac{\partial r_{ij}}{\partial y_j} = \dfrac{(y_j - y_i)}{r_{ij}} \\ \dfrac{\partial r_{ij}}{\partial z_i} = \dfrac{(z_i - z_j)}{r_{ij}} \\ \dfrac{\partial r_{ij}}{\partial z_j} = \dfrac{(z_j - z_i)}{r_{ij}} \end{cases} \tag{6.25}$$

The first derivative is easy to compute, in the case of the second derivative instead we need to distinguish whether the two derivatives act on the same atom, just on different XYZ variables, or on different atoms. In the case the second derivative acts on two different atoms we get the following second derivatives:

$$\begin{cases} \dfrac{\partial^2 r_{ij}}{\partial x_j \partial x_i} = \dfrac{(x_j - x_i)^2}{r_{ij}^3} - \dfrac{1}{r_{ij}} \\ \dfrac{\partial^2 r_{ij}}{\partial y_j \partial y_i} = \dfrac{(y_j - y_i)^2}{r_{ij}^3} - \dfrac{1}{r_{ij}} \\ \dfrac{\partial^2 r_{ij}}{\partial z_j \partial z_i} = \dfrac{(z_j - z_i)^2}{r_{ij}^3} - \dfrac{1}{r_{ij}} \\ \dfrac{\partial^2 r_{ij}}{\partial x_j \partial y_i} = \dfrac{(x_j - x_i)(y_j - y_i)}{r_{ij}^3} \\ \dfrac{\partial^2 r_{ij}}{\partial x_j \partial z_i} = \dfrac{(x_j - x_i)(z_j - z_i)}{r_{ij}^3} \\ \dfrac{\partial^2 r_{ij}}{\partial y_j \partial z_i} = \dfrac{(y_j - y_i)(z_j - z_i)}{r_{ij}^3} \end{cases} \tag{6.26}$$

In case the derivatives act on the same atom we obtain the following second derivatives:

$$\begin{cases} \dfrac{\partial^2 r_{ij}}{\partial x_j^2} = \dfrac{1}{r_{ij}} - \dfrac{(x_j - x_i)^2}{r_{ij}^2} \\ \dfrac{\partial^2 r_{ij}}{\partial y_j^2} = \dfrac{1}{r_{ij}} - \dfrac{(y_j - y_i)^2}{r_{ij}^2} \\ \dfrac{\partial^2 r_{ij}}{\partial z_j^2} = \dfrac{1}{r_{ij}} - \dfrac{(z_j - z_i)^2}{r_{ij}^2} \\ \dfrac{\partial^2 r_{ij}}{\partial x_j \partial y_j} = -\dfrac{(x_j - x_i)(y_j - y_i)}{r_{ij}^3} \\ \dfrac{\partial^2 r_{ij}}{\partial x_j \partial z_j} = -\dfrac{(x_j - x_i)(z_j - z_i)}{r_{ij}^3} \\ \dfrac{\partial^2 r_{ij}}{\partial y_j \partial z_j} = -\dfrac{(y_j - y_i)(z_j - z_i)}{r_{ij}^3} \end{cases} \tag{6.27}$$

Thanks to these derivatives, we can now compute the Gradient and the Hessian of any radial function of the form of Eq. 6.23 as a function of the first and second derivatives of the radial component $V(r_{ij})$:

$$\begin{cases} G(r_{ij}) = \dfrac{\partial V(r_{ij})}{\partial r_{ij}} \\ H(r_{ij}) = \dfrac{\partial^2 V(r_{ij})}{\partial r_{ij}^2} \end{cases} \tag{6.28}$$

6.4.1 The Gradient

In the case of the gradient, we want to compute a vector of the form:

$$\begin{bmatrix} \frac{\partial \sum_{j>i} V(r_{ij})}{\partial x_1} \\ \frac{\partial \sum_{j>i} V(r_{ij})}{\partial y_1} \\ \frac{\partial \sum_{j>i} V(r_{ij})}{\partial z_1} \\ \cdots \\ \frac{\partial \sum_{j>i} V(r_{ij})}{\partial x_k} \\ \frac{\partial \sum_{j>i} V(r_{ij})}{\partial y_k} \\ \frac{\partial \sum_{j>i} V(r_{ij})}{\partial z_k} \\ \cdots \end{bmatrix} \tag{6.29}$$

let's compute the single element derivative with respect to x_k:

$$\frac{\partial \sum_{j>i} V(r_{ij})}{\partial x_k} = \sum_{j>i} \frac{\partial}{\partial x_k} V(r_{ij}) = \sum_{j \neq k} \frac{\partial}{\partial x_k} V(r_{kj}) = \sum_{j \neq k} G(r_{kj}) \frac{\partial r_{kj}}{\partial x_k} \tag{6.30}$$

As can be seen, each entry in the vector can be expressed as a summation over the radial gradients G multiplied by the radial derivatives in cartesian coordinates that we have derived previously. The trick was to express the partial derivative with respect to x_k in terms of radial partial derivatives thanks to the chain rule:

$$\frac{\partial}{\partial x_k} = \frac{\partial r_{ki}}{\partial x_k} \frac{\partial}{\partial r_{kj}} \tag{6.31}$$

6.4.2 The Hessian

The Hessian is a $\mathbf{R}^{3N \times 3N}$ matrix defined as:

$$\begin{bmatrix} \frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial^2 x_1} & \frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial x_1 \partial y_1} & \frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial x_1 \partial z_1} & \cdots \\ \frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial y_1 \partial x_1} & \frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial^2 y_1} & \frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial y_1 \partial z_1} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \tag{6.32}$$

The number of rows and columns is 3N since we have x, y, and z for each atom. In this case, we are going to compute the second derivative with respect to two variables x and y:

$$\frac{\partial \sum_{j>i} V(r_{ij})}{\partial x_l \partial y_k} = \frac{\partial}{\partial x_l} \sum_{j \neq k} G(r_{kj}) \frac{\partial r_{kj}}{\partial y_k} = \sum_{j \neq k} \frac{\partial G(r_{kj})}{\partial x_l} \frac{\partial r_{kj}}{\partial y_k} + G(r_{kj}) \frac{\partial}{\partial x_l} \frac{\partial}{\partial y_k} r_{kj} \tag{6.33}$$

In this case, in order to proceed forward, we need to employ the same equations derived before about the second-order derivatives of r_{ij}. As a consequence, we have to make the same distinction whether the second derivative applies to the same atom of the first derivative or not. In case it is the second derivative on the same atom ($l = k$), we can write:

$$\frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial x_l \partial y_l} = \sum_{j \neq l} \frac{\partial G(r_{lj})}{\partial x_l} \frac{\partial r_{lj}}{\partial y_l} + G(r_{lj}) \frac{\partial}{\partial x_l} \frac{\partial}{\partial y_l} r_{lj} \tag{6.34}$$

By remembering the chain rule and the definition of $H(r_{ij})$ we can finally write:

$$\frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial x_l \partial y_l} = \sum_{j \neq l} H(r_{lj}) \frac{\partial r_{lj}}{\partial x_l} \frac{\partial r_{lj}}{\partial y_l} + G(r_{lj}) \frac{\partial^2 r_{lj}}{\partial x_l \partial y_l} \tag{6.35}$$

In the case the atoms are different ($l \neq k$), instead, the summation disappears since the only term that survives is the one involving the r distance between the two atoms:

$$\frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial x_l \partial y_k} = \frac{\partial G(r_{kl})}{\partial x_l} \frac{\partial r_{kl}}{\partial y_k} + G(r_{kl}) \frac{\partial^2 r_{kl}}{\partial x_l \partial y_k} \tag{6.36}$$

Again, remembering the chain rule and the definition of H we finally get:

$$\frac{\partial^2 \sum_{j>i} V(r_{ij})}{\partial x_l \partial y_k} = H(r_{kl}) \frac{\partial r_{kl}}{\partial x_l} \frac{\partial r_{kl}}{\partial y_k} + G(r_{kl}) \frac{\partial^2 r_{kl}}{\partial x_l \partial y_k} \tag{6.37}$$

6.5 From Cartesian to Internal coordinates

In this chapter, the focus shifted to energy expressions that are functions of a single variable. However, the energy of a polyatomic molecule composed of N_a atoms is dependent on all internal degrees of freedom, whose number is $3N_a - 6$ for non-linear systems and $3N_a - 5$ for linear systems. While most of the QM computations are based on a Cartesian description of nuclear motions, internal curvilinear coordinates present different advantages in treating molecular vibrations. First of all, they are curvilinear as nuclear motions. Secondly, internal coordinates present a versatility such that they can in principle reproduce any kind of vibration and are independent of the overall translations and rotations. Finally, the use of this kind of coordinates leads to a better decoupling of different vibrational degrees of freedom, this last aspect being of fundamental relevance in the parametrization of Force Fields, as well as in spectroscopic and kinetic studies. In the next sections, the main sets of internal coordinates will be shortly addressed, followed by a discussion of the harmonic theory of vibrations both in terms of Cartesian and internal coordinates.

6.5.1 Definition of Internal Coordinates

The simplest set of internal coordinates is represented by the so-called primitive internal coordinates (PICs), represented by the full list of bond lengths, bond angles and dihedrals. Let us consider a set of M internal coordinates $\boldsymbol{s} = \{s_1, ..., s_M\}$, which can be expanded through a Taylor series around the equilibrium geometry,

$$\begin{aligned} s_i - s_i^{eq} = & \sum_{j=1}^{3N} \left(\frac{\partial s_i}{\partial x_j} \right)_{eq} \left(x_j - x_j^{eq} \right) \\ & + \frac{1}{2} \sum_{j=1}^{3N} \sum_{k=1}^{3N} \left(\frac{\partial^2 s_i}{\partial x_j \partial x_k} \right) \left(x_j - x_j^{eq} \right) \left(x_k - x_k^{eq} \right) + O\left(|x|^2 \right) \end{aligned} \tag{6.38}$$

where $\boldsymbol{x} = \{x_1, ..., x_{3N_a}\}$ is the vector collecting all nuclear Cartesian coordinates, B_{ij} is an element of the Wilson $\boldsymbol{B}$ matrix [117] and $\boldsymbol{B'}$ is the tensor representing its Cartesian derivative:

$$B_{ij} = \left(\frac{\partial s_i}{\partial x_j}\right)_{eq} \tag{6.39}$$

$$B'_{ijk} = \left(\frac{\partial^2 s_i}{\partial x_j \partial x_k}\right)_{eq} \tag{6.40}$$

In general, internal coordinates can be classified as redundant or non-redundant depending on their number. Redundant coordinates are characterized by a set of coordinates larger than the number of normal modes, so linear dependencies exist among the internal coordinates. Conversely, non-redundant internal coordinates are deprived of such dependencies. Non-redundant internal coordinates can be obtained through a linear transformation from the Primitive Internal Coordinates or any redundant set. It is noteworthy that PICs can be further extended based on the so-called Generalized Internal Coordinates (GICs), which augment the redundant set with particular coordinates like ring puckering, to build coordinates able to properly describe the vibrations associated with the most important structural deformations. The conversion from internal coordinates to Cartesian coordinates is not straightforward and is non-linear so many algorithms and procedures have been developed and discussed in the literature, generally employing an iterative procedure [118–120].

6.5.2 Harmonic theory of molecular vibrations

The calculation of harmonic frequencies and normal coordinates can be strongly affected by the choice of reference coordinates. In the cartesian-based framework, the first step is calculating the mass-weighted force constants matrix $\boldsymbol{H_m}$, defined as:

$$\boldsymbol{H_m} = \boldsymbol{M^{-1/2} H_x M^{-1/2}} \tag{6.41}$$

where $\boldsymbol{M}$ is the diagonal matrix of nuclear masses, while $\boldsymbol{H_x}$ is the Cartesian Hessian matrix.

$$\boldsymbol{H_{ij}} = \left(\frac{\partial^2 V}{\partial x_i \partial x_j}\right)_{eq} \tag{6.42}$$

The $\boldsymbol{H_m}$ matrix is then diagonalized,

$$\boldsymbol{H_m L} = \boldsymbol{L\Lambda} \tag{6.43}$$

Where Λ is the diagonal matrix of the squared harmonic frequencies, while $\boldsymbol{L}$ collects the eigenvectors. The eigenvectors corresponding to non-null eigenvalues are used to build the normal coordinates $\boldsymbol{Q}$:

$$\boldsymbol{Q} = \boldsymbol{L}^T \boldsymbol{M}^{1/2} \boldsymbol{\Delta X} \tag{6.44}$$

where $\Delta x_i = x_i - x_i^{eq}$. When the formulation is based on a generic set of internal coordinates $\boldsymbol{s}$, the kinetic energy operator [121, 122]

$$T = -\frac{\hbar^2}{2}\sum_{i=1}^{M}\sum_{j=1}^{M} det(\boldsymbol{G})^{1/4}\frac{\partial}{\partial s_i} det(\boldsymbol{G})^{-1/2} G_{ij}\frac{\partial}{\partial s_j} det(\boldsymbol{G})^{1/4} \tag{6.45}$$

is not diagonalized anymore. In the above expression, the Wilson $\boldsymbol{G}$ matrix has been introduced.

$$\boldsymbol{G} = \boldsymbol{B}\boldsymbol{M}^{-1}\boldsymbol{B}^{\boldsymbol{T}} \tag{6.46}$$

At the harmonic level, the $\boldsymbol{G}$ matrix can be approximated with its reference geometry value and its dependence on the coordinates can be neglected:

$$T = -\frac{\hbar^2}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{ij}^{eq}\frac{\partial^2}{\partial s_i \partial s_j} \tag{6.47}$$

Moreover, the potential energy term of the Hamiltonian can be automatically written in terms of internal coordinates as a Taylor-series expansion around the equilibrium value:

$$V = V^{eq} + \sum_{i=1}^{N}\left(\frac{\partial V}{\partial s_i}\right)_{eq} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial^2 V}{\partial s_i \partial s_j}\right)_{eq} s_i s_j + O(|s|^2) \tag{6.48}$$

By assuming the condition of stationary point and shifting the origin of the Potential Energy Surface (PES) to zero, the above equation truncated to the second order can be written as

$$V = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} F_{ij} s_i s_j \tag{6.49}$$

where $\boldsymbol{F}$ is the Hessian matrix of the potential energy with respect to the internal coordinates. The $\boldsymbol{F}$ matrix can be calculated from the Cartesian Hessian matrix ($\boldsymbol{H_x}$) [120] as

$$\boldsymbol{F} = \{\boldsymbol{B}^{\dagger}\}^T[\boldsymbol{H_x} - \boldsymbol{g}_s\boldsymbol{B'}]\boldsymbol{B}^{\dagger} \tag{6.50}$$

This expression is generally valid for non-equilibrium geometries also because of the inclusion of the gradient $\boldsymbol{g_s}$ with respect to the internal coordinates. To compute such a gradient starting from the one in cartesian coordinates we can write

$$\boldsymbol{g_s} = \{\boldsymbol{B}^{\dagger}\}^T\boldsymbol{g_x} \tag{6.51}$$

In this equations $\boldsymbol{B}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\boldsymbol{B}$ [123] defined as

$$\boldsymbol{B}^{\dagger} = (\boldsymbol{B}\boldsymbol{U}\boldsymbol{B}^T)^{-1}\boldsymbol{B}^T\boldsymbol{U} \tag{6.52}$$

Where $\boldsymbol{U}$ is a general (3M×3M) arbitrary matrix. In general, the contribution of translations and rotations can be factored out through the application of a projection matrix $\bar{\boldsymbol{P}} = \boldsymbol{B}^{\dagger}\boldsymbol{B}$. More specifically, the gradient and the Hessian in internal coordinates can be obtained as:

$$\boldsymbol{g_s} = \{\boldsymbol{B}^\dagger\}^T \bar{\boldsymbol{P}} \boldsymbol{g_x} \tag{6.53}$$

$$\boldsymbol{F} = \{\boldsymbol{B}^\dagger\}^T \bar{\boldsymbol{P}} [\boldsymbol{H_x} - \boldsymbol{g_s} \boldsymbol{B'}] \bar{\boldsymbol{P}} \boldsymbol{B}^\dagger \tag{6.54}$$

These are the general expressions for moving from the space of Cartesian Coordinates to the space of Internal Coordinates. Once this procedure has been carried out, the harmonic vibrational Hamiltonian can be defined as

$$H^{(0)} = \frac{1}{2} \sum_{i=1}^{M} \sum_{j=1}^{M} \left(G_{ij} P_i^s P_j^s + F_{ij} s_i s_j \right) \tag{6.55}$$

Where $P_i^s = -i\hbar\partial/\partial s_i$. As previously anticipated, the kinetic energy is not diagonal, with the different coordinates being coupled by the $\boldsymbol{G}$ matrix. As a result, a set of coordinates that leads to the diagonalization of not only $\boldsymbol{F}$ but also $\boldsymbol{G}$, must be defined. For this purpose, a set of normal coordinates $\boldsymbol{Q}$ can be introduced:

$$\boldsymbol{s} = \boldsymbol{L}\boldsymbol{Q} \tag{6.56}$$

By applying this new change of coordinates, it is possible to rewrite the Hamiltonian in the following form:

$$H^{(0)} = \frac{1}{2} \left[\boldsymbol{P}^T \boldsymbol{L}^{-1} \boldsymbol{G} (\boldsymbol{L}^{-1})^T \boldsymbol{P} + \boldsymbol{Q}^T \boldsymbol{L}^T \boldsymbol{F} \boldsymbol{L} \boldsymbol{Q} \right] \tag{6.57}$$

$\boldsymbol{L}$ is defined so that in this basis the $\boldsymbol{F}$ matrix is diagonal while $\boldsymbol{G}$ is equal to the identity matrix. This corresponds to the resolution of the following equation [117]:

$$\boldsymbol{G}\boldsymbol{F}\boldsymbol{L} = \boldsymbol{L}\boldsymbol{\Lambda} \tag{6.58}$$

In other words, the harmonic frequencies and Internal-based Normal Coordinates (INCs) can be calculated through the diagonalization of the $\boldsymbol{GF}$ matrix product. It should be noted that, differently from the Cartesian-based formulation, the matrix to be diagonalized is not symmetric, implying that the normal coordinates do not form an orthogonal basis. However, as demonstrated by Myazawa [124] the equation above can be recast in a symmetric form characterized by the same eigenvalues

$$(\boldsymbol{G}^{1/2} \boldsymbol{F} \boldsymbol{G}^{1/2})(\boldsymbol{G}^{-1/2} \boldsymbol{L}) = (\boldsymbol{G}^{-1/2} \boldsymbol{L}) \boldsymbol{\Lambda} \tag{6.59}$$

where the columns of the matrix $\boldsymbol{G}^{-1/2}\boldsymbol{L}$ are the eigenvectors of the symmetric matrix $\boldsymbol{G}^{1/2}\boldsymbol{F}\boldsymbol{G}^{1/2}$, and thus are orthogonal.

Chapter 7

Current state-of-the-art applications

For flexible molecules such as those of interest to our studies, a significant challenge is related to a large number of conformers and the fast relaxation of some of them to more stable counterparts due to the presence of low interconversion energy barriers. An inaccurate account of the relaxation processes can bias any direct thermochemical interpretation of the results provided by rotational spectroscopy experiments [125, 126]. Quantum-chemical (QC) computations can help to tackle this challenge, especially because the gas phase is their most natural playground [127, 128]. Unfortunately, for medium-sized systems, the usual dichotomy between accuracy and feasibility, which is the quest for accurate yet feasible predictions, comes into place [129]. State-of-the-art QC approaches can rival the experimental counterparts for small semi-rigid systems in the gas phase [129–131], but they are characterized by a very unfavorable scaling with the dimension of the system to be investigated. This already prevents their brute-force application to biomolecule building blocks containing more than a dozen of atoms and characterized by several low-energy minima. Furthermore, the powerful local optimization techniques developed for semi-rigid systems are ineffective for flexible systems, which require exploring rugged potential energy surfaces (PESs) [132, 133].

For the reasons mentioned above, the accurate characterization needed by high-resolution spectroscopy requires an integrated computational approach that employs QC models of increasing accuracy in the different steps of an exploration/exploitation strategy guided by machine learning (ML) tools. As already mentioned in the introduction of this thesis, the main steps of this strategy [73, 132] can be summarized as follows:

1. Unsupervised perception of the molecular system to identify hard and soft degrees of freedom [77].

2. Knowledge-based selection and constrained geometry optimizations of a limited number of conformers employing a fast semi-empirical method [133].

3. Exploration of the PES governed by soft degrees of freedom using the same semi-empirical method of the previous step, guided by a purposely tailored evolutionary algorithm with the aim of finding other low-lying minima [132].

4. Refinement of the most stable structures by hybrid and then double-hybrid density functionals [73].

5. Analysis of relaxation paths between pairs of adjacent energy minima [134].

6. Evaluation of accurate electronic energies for the final panel of low-energy minima [135].

7. Computation of Zero Point Energy (ZPE) and thermal contributions to enthalpies and entropies [6, 134, 136].

8. Computation of spectroscopic parameters for the energy minima with non-negligible populations [134].

The present thesis's main aim is to improve some key steps of that approach further. However, in parallel, several challenging studies have been performed employing a stable version of the tool [137]. For example, the systematic study of prototypical amino acids was completed recently (Ref. [138]) and here reported. These compounds represent a particularly appealing playground because their rich conformational landscape is tuned by the competition among different types of intra-molecular non-covalent interactions involving, together with the amino and carboxylic acid moieties of the backbone, also side-chain groups. At the same time, results from Microwave (MW) experiments are available for several conformers of most natural α-amino acids, and provide accurate data for benchmarking theory.

JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

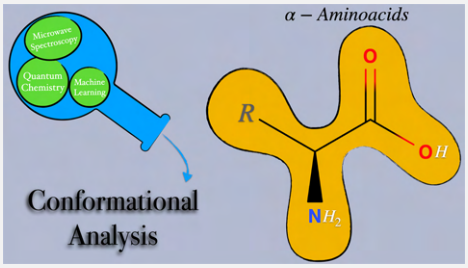
Benchmark Structures and Conformational Landscapes of Amino Acids in the Gas Phase: A Joint Venture of Machine Learning, Quantum Chemistry, and Rotational Spectroscopy

Vincenzo Barone,* Marco Fusè, Federico Lazzari, and Giordano Mancini

Cite This: J. Chem. Theory Comput. 2023, 19, 1243−1260 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The accurate characterization of prototypical bricks of life can strongly benefit from the integration of high resolution spectroscopy and quantum mechanical computations. We have selected a number of representative amino acids (glycine, alanine, serine, cysteine, threonine, aspartic acid and asparagine) to validate a new computational setup rooted in quantum-chemical computations of increasing accuracy guided by machine learning tools. Together with low-lying energy minima, the barriers ruling their interconversion are evaluated in order to unravel possible fast relaxation paths. Vibrational and thermal effects are also included in order to estimate relative free energies at the temperature of interest in the experiment. The spectroscopic parameters of all the most stable conformers predicted by this computational strategy, which do not have low-energy relaxation paths available, closely match those of the species detected in microwave experiments. Together with their intrinsic interest, these accurate results represent ideal benchmarks for more approximate methods.



1. INTRODUCTION

Thanks to its high resolution and noninvasivity, gas-phase molecular spectroscopy has become the method of choice to investigate the role of intrinsic stereoelectronic effects in tuning the physical-chemical properties of biomolecule building blocks.[1,2] In particular, the supersonic-jet expansion technique[3] coupled to laser ablation[4] is allowing the recording of gas-phase microwave (MW) spectra for these thermolabile compounds, which are usually characterized by high melting points.[5] However, the fast relaxation of some structures to more stable counterparts in the presence of low energy barriers can bias any direct thermochemical interpretation of the results provided by this technique.[6,7]

Accurate quantum chemical (QC) computations can help to solve this kind of problem,[8,9] but the effective exploration of flat potential energy surfaces (PESs) and the characterization of their stationary points for medium- to large-size flexible systems are still challenging for at least two different reasons. From the one side, the size of the systems prevents a brute force approach employing very accurate but very expensive state-of-the-art QC methodologies.[10−12] From the other side, the very powerful local optimization techniques developed for semirigid systems are not effective for the exploration of rugged potential energy surfaces (PES) characterized by a huge number of energy minima possibly separated by low-energy barriers.[13,14]

This situation calls for an integrated computational approach employing QC models of increasing accuracy in the different steps of an exploration/exploitation strategy guided by machine learning (ML) tools.[13,15−17] The effective strategy of this kind we have been developing in the past few years starts from a knowledge-based selection and constrained geometry optimizations of a limited number of conformers employing a fast semiempirical method.[14,18] Next, an effective exploration of the whole conformational PES is performed by the same semiempirical method guided by a purposely tailored evolutionary algorithm with the aim of finding other low-lying minima.[13] The results of this step are refined by hybrid and then double-hybrid density functionals,[19,20] and possible relaxation paths between pairs of adjacent energy minima are identified.[16] Once a panel of low-energy minima has been defined, accurate relative energies are computed by reduced-scaling composite methods.[21−26] These results are integrated by zero point energies (ZPE) and thermal contributions to enthalpies and entropies employing anharmonic approaches rooted in the second order vibrational perturbation theory (VPT2)[27−34] and proper treatment of hindered rotations.[35,36] Finally, accurate spectroscopic parameters of the energy minima with nonnegligible populations under the experimental

Received: November 15, 2022
Published: February 2, 2023

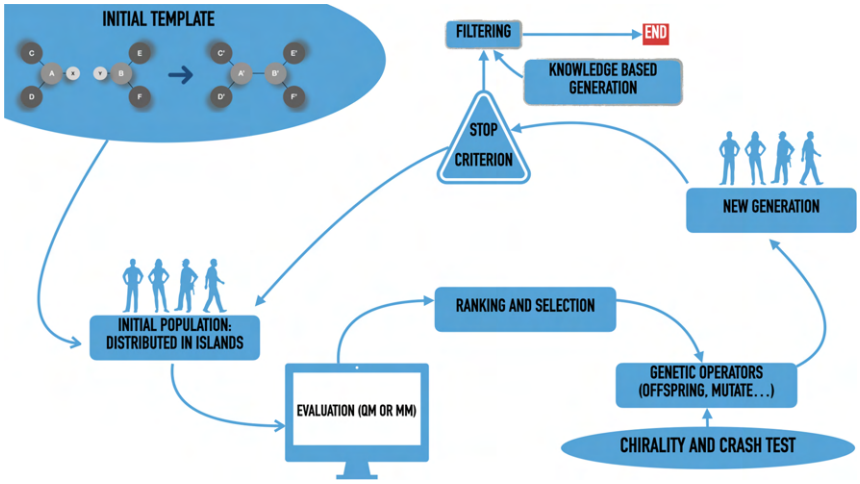


Figure 1. Flowchart of PES exploration. See main text for further details.

conditions of interest are computed.[37] In the specific case of rotational spectroscopy, improved equilibrium rotational constants are obtained by refining the optimized geometries by a linear regression approach.[20,38]

Among the main biomolecule building blocks, natural α-amino acids, which exist exclusively in neutral form in the gas phase, represent a particularly appealing playground because their rich conformational landscape is tuned by the competition among different kinds of intramolecular hydrogen bonds. At the same time, MW results are available for several conformers of most natural α-amino acids,[39−50] which represent very demanding benchmarks for the *a priori* prediction of structural and spectroscopic parameters. We have therefore selected the glycine and alanine prototypes together with a panel of α-amino acids with polar side chains (serine, threonine, cysteine, aspartic acid, and asparagine) with the aim of providing benchmark results allowing unbiased comparisons with experimental results. In fact, the current standards for the computation of MW parameters of biomolecule building blocks in the gas phase (see, e.g., refs 1, 6, 7, 39, and 51) employ QC methods of limited accuracy, pay marginal attention to the geometrical parameters, and neglect vibrational corrections. However, these limitations hamper any *a priori* prediction of the spectroscopic outcome, allowing at most its *a posteriori* interpretation in terms of the agreement between experimental and computed spectroscopic parameters for a predefined number of conformers.

Based on these premises, the goal of the present study is to improve and validate a general strategy able to find all the conformers detectable in supersonic jet expansions taking also into account fast relaxation processes possibly leading to the disappearance of some low-lying species. Unbiased comparison with spectroscopic results is made possible by the accuracy of the computational results, which will be shown to provide mean unsigned errors (MUEs) within 20 MHz for rotational constants and 10 cm^{-1} for both relative energies and vibrational frequencies (entering zero point energies and thermal contributions to thermodynamic functions). Together with their intrinsic interest of the studied molecules, these results will provide also a reference set for more approximate methods and/or search techniques.

2. PES EXPLORATION

The general strategy for the exploration of conformational PESs is based on a continuous perception of molecular structures performed by the PROXIMA software,[52] which is able to detect characteristic structural motifs and to separate soft (in the present context dihedral angles) and hard degrees of freedom. Then, a knowledge-based systematic search of soft degrees of freedom[14] can be optionally performed, which produces a panel of guess structures (e.g., the 3^n staggered conformers generated by rotations around n nonterminal single bonds, which are not a part of cycles). The geometries of these candidates are next optimized using the fast GFN2-XTB semiempirical method,[18] which has been selected because it tends to underestimate energy differences (i.e., to produce a too large set of candidates), which allows a safer use of energy thresholds for further processing. Next, a custom implementation of the island model evolutionary algorithm (IM-EA)[53] is employed to produce other candidates starting from an initial population (P_0) generated by the so-called Latin Hypercube stratified sampling[54] in order to maximize the diversity of soft degrees of freedom. The chemical descriptor (fitness) of each structure is the relative electronic energy obtained by GFN2-XTB geometry optimizations of the stiff degrees of freedom. Improved populations are then built iteratively for a given number of cycles by applying, with predetermined probability, different genetic operators, namely, crossover (interpolation of the features of different related structures for creating new ones), mutation (change of one or more soft degrees of freedom with some stochastic rule), and selection (high chance for high fitness structures of propagating their features in the next cycles). In the IM-EA, the different operators act separately on disjoint regions of the conformational landscape (islands), which are mixed only at predefined intervals by a dedicated operator (migration). Furthermore, some of the best structures found in each cycle are directly transferred to the next cycle (the so-called Hall of Fame).[55] All those choices are dictated by the high cost of evaluating the fitness of a new structure by constrained geometry optimizations. As a consequence, high fitness structures are worth being preserved in the population until some significantly improved structure is found. Typical values of the initial population, maximum

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

number of cycles, and number of islands are 100, 50, and 4, which result in about 1000 constrained geometry optimizations for each run of the algorithm. In order to further increase the coverage of the conformational space, 4 runs with different initial populations are performed for each molecular system. The full set of parameters employed in the IM-EA algorithm is given in Table S1 of the Supporting Information (SI), while further details are given in refs 13 and 17. Figure 1 shows a schematic flowchart of the current version of the whole algorithm, which is available under the GPL3 license at https://github.com/tuthmose/IM_EA.

At the end of the whole exploration, low-energy conformers within a predefined energy range are selected from the panel of structures issued from IM-EA and, possibly, knowledge-based steps by eliminating too similar structures (in terms of rotational constants and root-mean-square deviations of heavy atom positions) and then performing single point energy evaluations at the B3LYP/jun-cc-pVDZ level,[56,57] also including Grimme's D3BJ dispersion corrections.[58] In the following, this computational model will be referred to simply as B3. The choice of the specific functional is not critical in this step because it is used only for the selection of an initial panel of structures to be next refined at higher levels. The B3 model has been selected because it is routinely employed in the interpretation of MW studies and, more importantly, provides reasonable anharmonic corrections (vide infra).

In the next step, structures lying within a smaller energy range are optimized at the same level, and the surviving ones define the panel of candidates for the final structural refinement, which is performed employing the revDSD-PBEP86-D3BJ/jun-cc-pv(T+d)Z model[59−61] (hereafter rDSD) for both geometry optimization and evaluation of harmonic force fields.[62] The rDSD functional has been selected because several studies have shown that it provides excellent geometrical structures,[38] dipole moments,[63] spectroscopic parameters,[37] noncovalent intermolecular interactions,[23,64] and conformational landscapes.[10,65,66]

This composite strategy allows for strongly reducing the number of expensive geometry optimizations by hybrid and, especially, double-hybrid functionals without any loss of accuracy in the final results. The different energy thresholds depend on the system and the spectroscopic technique of interest. For the specific case of rotational spectroscopy, a conservative limit for the relative stability of detectable conformers is around 900 cm^{-1} (which corresponds to a relative population of about 1% at room temperature, where kT/hc = 207 cm^{-1}).[1,16] As a consequence, the typical thresholds for the acceptance of semiempirical structures, B3 geometry optimizations, and final rDSD refinement are 2500, 1500, and 1000 cm^{-1}, respectively. These choices lead to about 100 B3 computations (including both single point and geometry optimizations) and no more than 20 rDSD geometry optimizations for each molecular system.

As mentioned in the Introduction, conformational relaxation can take place under the experimental conditions whenever the energy barriers ruling the interconversion are sufficiently low, with an upper limit of about 400 cm^{-1} being usually employed for discriminating in rotational spectroscopy of amino acids and related compounds.[6,7,67] With the aim of unraveling fast conformational relaxations, we always perform relaxed torsional scans at the rDSD level in order to obtain preliminary information on low-energy interconversion paths. Next, after precise location of transition states (TSs) by full geometry optimizations, their nature is checked by computing Hessian matrices.

3. RELATIVE STABILITIES AND SPECTROSCOPIC PARAMETERS

The typical MUEs of rDSD bond lengths (0.003 Å) and valence angles (0.003 radians, i.e., 0.15°) observed in the large SE100 database[38] are largely sufficient to obtain accurate relative electronic energies of different conformers by single-point energy evaluations using composite methods rooted in the coupled cluster (CC) ansatz.[68] In particular, the CC model including single, double, and perturbative estimate of triple excitations (CCSD(T))[69] is considered the gold standard for this kind of computations provided that complete basis set (CBS) extrapolation and core valence (CV) correlation are taken into the proper account. The key idea of the reduced cost Cheap scheme (ChS) is that, starting from frozen core (fc) CCSD(T) computations in conjunction with the cc-pVTZ basis set,[57] CBS and CV terms can be computed with good accuracy and negligible additional cost employing second order Møller−Plesset perturbation theory (MP2).[70] Several benchmarks[22,23] have shown that improved noncovalent interactions can be obtained employing partially augmented (jun-cc-pV(n+d)Z) basis sets,[61,71] and the corresponding model is labeled junChS. Replacement of conventional methods with the explicitly correlated (F12) variants leads to the junChSF12 model, which is even more accurate without any excessive additional cost. In detail, the starting point is the frozen-core (fc) CCSD(T)-F12b(3C/FIX) method[72−74] again in conjunction with the jun-cc-pV(T+d)Z basis set.[61,75] The corresponding auxiliary basis sets are also employed for resolution of the identity and density fitting, and the geminal exponent (γ) was fixed to 1.0 a_0^{-1}.[75,76] CBS extrapolation is carried out with the standard n^{-3} two-point formula[77] employing MP2F12/jun-cc-pV(X+d)Z energies with X = T and Q. The CV contribution is then incorporated as the difference between all-electron (ae) and fc MP2F12 calculations, both with the cc-pCVW(T+d)Z basis set.[78] A systematic study of noncovalent intermolecular interactions[23] showed that the junChSF12 approach is affected by small basis set superposition errors (BSSE), which would be difficult to take into account for intramolecular interactions. Furthermore, comparison with the most accurate results available for a panel of representative noncovalent complexes provided an average absolute error smaller than 10 cm^{-1}.[22,23]

To determine the relative stability of different low-energy minima, one has to move from electronic energy differences to the corresponding relative enthalpies at 0K (ΔH_0°) or free energies (ΔG°) at a temperature depending on the experimental conditions. The vibrational contributions to thermodynamic functions are usually computed by the harmonic oscillator (HO) model, which shows the largest errors in the high frequency (overestimated contributions to zero point energies) and low frequency (overestimated contributions to entropies) regions.

The first issue is solved in the present work by estimating anharmonic contributions in the framework of second order vibrational perturbation theory (VPT2), which provides analytical and resonance free expressions for the ZPEs.[79] Harmonic (rDSD) and anharmonic (B3) contributions are employed in this connection, since a recent benchmark study has shown that for semirigid molecules the average absolute error of zero point energies with respect to accurate

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

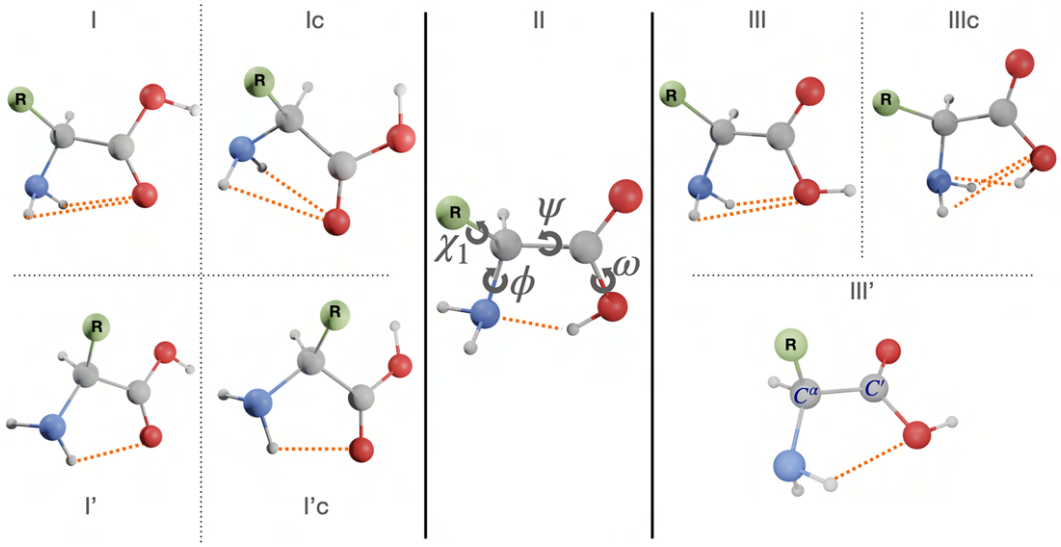


Figure 2. Structures of low-lying backbone conformers of α-amino acids.

experimental results is reduced from 53 to 17 cm^{-1} when going from the HO to the VPT2 anharmonic model.[80]

The treatment of low frequency contributions (typically less than 100 cm^{-1}) is more involved because different modes (e.g., torsions, inversions, etc.) need be identified, characterized, and treated by proper variational anharmonic computations.[36,81] In the same benchmark study mentioned above in connection with anharmonic ZPEs,[80] it has been shown that the simple one-dimensional hindered rotor model proposed by Ayala[82] in conjunction with the VPT2 model for the other vibrational modes leads to remarkably accurate vibrational entropies for both semirigid and flexible molecules for which accurate experimental results are available. In particular, an average absolute error of 3 cm^{-1} is obtained for the $T\Delta S$ contribution to free energies at room temperature. In the present context, test computations showed that the unbiased detection of hindered rotations becomes ambiguous for some conformers, so that we prefer to resort to the much simpler and black-box quasi-harmonic (QH) approximation.[35,83] In the QH approach, below a given cutoff value, entropic terms are obtained from the free-rotor model, and a damping function is used to interpolate between free-rotor and harmonic oscillator expressions close to the cutoff frequency.

The leading terms of MW spectra are the rotational constants of the vibrational ground-state (B_0^i, where i refers to the inertial axes a, b, c), which include vibrational corrections (ΔB_{vib}^i) in addition to equilibrium rotational constants (B_e^i).[84] In the framework of the VPT2 approximation,[85] the ground-state rotational constants can be expressed as

$$B_0^i = B_e^i + \Delta B_{vib}^i = B_e^i - \frac{1}{2}\sum_r \alpha_r^i \quad (1)$$

where the α_r's are the vibration−rotation interaction constants and the sum runs over all r vibrational modes. Noted is that the evaluation of the α_r's implies anharmonic force field calculations and that the sum appearing in eq 1 (contrary to individual terms) does not involve any resonance issue at the VPT2 level (for details, see, e.g., refs 11, 86, 87). ΔB_{vib}^i being a small fraction of the corresponding B_e^i (typically 0.5%),[88] it can be determined at an affordable level of theory (B3 in the present context) without significantly affecting the accuracy of the resulting vibrational ground-state rotational constant.[11,89] At the same time, inclusion of vibrational corrections is not warranted if the errors on the computed rotational constants are not much lower than 1% (50 MHz for a constant of 5000 MHz). Therefore, equilibrium rotational constants require very accurate geometrical parameters, which can be obtained only with state-of-the-art composite methods incorporating high excitation orders in the correlation treatment. These methods are able to deliver errors on equilibrium rotational constants as low as 0.1% (5 MHz for a rotational constant of 5000 MHz).[90] The reduced cost junChSF12 composite method delivers typical relative errors of 0.2%,[11,80,91] which are still sufficient for the unequivocal prediction and assignment of different conformers in the MW spectra of flexible molecules. Higher relative errors (typically 0.4−0.5%) are obtained at the rDSD level. However, the systematic nature of the errors permits geometrical parameters to be obtained and, thus, equilibrium rotational constants, rivaling the accuracy of the jun-ChSF12 counterparts by the linear regression approach (LRA). In this model, the computed geometrical parameters (r_{comp}) are corrected for systematic errors by means of scaling factors (a) and offset values (b) depending on the nature of the involved atoms and determined once for ever from a large database of accurate semiexperimental (SE) equilibrium geometries:[38,92]

$$r_{LRA} = (1 + a) \times r_{comp} + b \quad (2)$$

The a and b values for different bonds and valence angles are taken from ref 38. Noted is that the intrinsic accuracy of the rDSD model leads in most case to b = 0.0 together with very small a values for bond lengths and that, among valence angles, only OCO and HCH need be corrected. Several studies have confirmed that very accurate molecular structures can be obtained employing this approach (referred to in the following as rDSD-LRA).[16,38,92,93]

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 1. Relative Electronic Energies (ΔE), Enthalpies at 0 K ($\Delta H_0^\circ = \Delta$(E+ZPE)), and Free Energies at Room Temperature (ΔG°) (all in cm^{-1}; 1 kJ/mol = 83.59 cm^{-1}) for the Glycine Conformers

Conformer	Label	ΔE_{best} [a]	ΔE_{ChS} [b]	ΔE_{rDSD}	ΔH°_{0H} [c]	ΔG°_H [d]	ΔZPE [e]	$\Delta(T\Delta S)$ [f]	ΔG° [g]
ttt	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ccc	II*	223.8	236.5	214.8	345.9	468.8	−38.5	−29.4	400.9
gtt	I′*	433.7	431.5	447.8	406.6	482.6	−23.2	−39.5	419.9
tct	III	605.1	605.8	583.5	630.1	239.6	6.7	215.3	461.6
gct	III′*	926.8	935.5	918.9	937.5	969.4	−9.2	1.8	962.0
ttc	Ic	1678.8	1688.7	1675.4	1616.5	1659.6	1.9	−20.4	1641.1
tcc	IIIc	2042.5	2051.5	2071.3	2131.9	2027.3	−10.8	−15.2	2001.3
gtc	I′c*	2119.4	2118.9	2140.6	2012.3	2085.7	−28.8	−35.6	2021.3

[a]CBS+CV+ft+fq+rel. from ref 101. [b]JunChSF12 at rDSD geometries. [c]JunChSF12 electronic energies with rDSD harmonic ZPE. [d]JunChSF12 electronic energies with rDSD harmonic ZPE and thermal contributions. [e]Difference between anharmonic and harmonic ZPEs at the B3 level. [f]Difference between quasi-harmonic and harmonic $T\Delta S$ (see text for details). [g]Sum of columns 7, 8, and 9.

Additional parameters of particular relevance for MW spectroscopy are the nuclear quadrupole coupling constants (χ_{ii}, i referring to the inertia axis a, b, or c).[94] Nuclear quadrupole coupling is the interaction between the quadrupole moment of a nucleus with nuclear spin $I \geq 1$) and the electric gradient at the nucleus itself.[86] Since at least one ^{14}N quadrupolar nucleus is present in all amino acids, nuclear quadrupole coupling constants are important for accurate predictions of rotational spectra because they determine a splitting of the rotational transitions, which generates the so-called hypefine structure. Since a systematic study of rDSD quadrupole coupling constants has not yet been performed, the comparison with the experimental values for several conformers of different amino acids represents per se an interesting benchmark. We anticipate that vibrational effects on nuclear quadrupole coupling constants are usually smaller than the uncertainty affecting the computed equilibrium values, and thus they have not been considered in this work.

Finally, the components of dipole moments determine the intensities of rotational transitions and, as already mentioned, rDSD is expected to provide reliable values.[63]

Concerning technical details, the Gaussian package[95] has been used for all calculations except the junChSF12 and QH ones, which have been performed with the help of the Molpro[76] and GoodVibes[83] software, respectively.

4. STRUCTURE AND SOFT DEGREES OF FREEDOM

The conformation of isolated amino acids is determined by both backbone (ϕ = H−N−C^α−C′, ψ = N−C^α−C′−O(H), and ω = C^α−C′−O−H) and side chain (χ, defined more precisely in the following) torsional angles, as shown in the central panel of Figure 2. However, the nonplanarity of the NH_2 moiety suggests replacing the customary ϕ dihedral angle (H−N−C^α−C′) with ϕ' = LP−N−C^α−C′ = ϕ + 120°, where LP is the nitrogen lone-pair perpendicular to the plane defined by the two aminic hydrogens and the C^α atom.

The most stable backbone structures involve the formation of hydrogen bonds (see Figure 2), which can be classified as I (bifurcated NH_2···O═C, $\phi' \approx 180°$, $\psi \approx 180°$, $\omega \approx 180°$), II (N···HO, $\phi' \approx 0°$, $\psi \approx 0°$, $\omega \approx 0°$), or III (bifurcated NH_2···OH, $\phi' \approx 180°$, $\psi \approx 0°$, $\omega \approx 180°$).[4] Higher energy minima can be classified as type I′ (single HNH···O═C hydrogen bond, $\phi' \approx 90°$, $\psi \approx 180°$, $\omega \approx 180°$) or type III′ (single HNH···OH hydrogen bond, $\phi' \approx 180°$, $\psi \approx 90°$, $\omega \approx 180°$). Furthermore, conformers of type I, I′, and III have higher energy counterparts for $\omega \approx 0°$, labeled in the following as Ic, I′c, and IIIc, respectively. The customary c, g, s, and t labels are used to indicate the cis, gauche, skew, and trans conformations for each dihedral angle in the order ϕ',ψ, ω/χ_1, ..., χ_n.

For purposes of consistency with the original experimental studies, capital letters L, M, N, ... are used in some cases to label conformers of amino acids with polar side chains in order of decreasing relative populations estimated from MW spectra.[39,45,46,96]

5. RESULTS AND DISCUSSION

5.1. The Smallest Prototypes: Glycine and Alanine.

Glycine has been extensively characterized from both experimental and computational points of view (see refs 97−101 and references therein). Its limited size allowed the exploitation of state-of-the-art composite schemes including, together with CBS and CV contributions evaluated at the CCSD(T) level, also full account of triple excitations, perturbative inclusion of quadruple excitations, and relativistic contributions (CBS+CV+fT+pQ+rel).[101] All the eight conformers mentioned above (I, II, III, I′, III′, Ic, IIIc, and I′c) have been characterized with four of them (I, III, Ic, and IIIc) having a planar backbone (C_s point group) and the other four (labeled with an asterisk to signal the presence of two equivalent nonplanar backbones) lacking any symmetry[98,99] (see Table S2 of the SI). Concerning relative stabilities, the junChSF12 model performs remarkably well with an average absolute error of 6 cm^{-1} from the most accurate available results[101] (see Table 1). The largest discrepancy (13 cm^{-1}) is observed for the II conformer, which is slightly stabilized by triple and quadruple excitations. Also the accuracy of the rDSD model (maximum error (MAX) and MUE of 29 and 15 cm^{-1} with respect to the most accurate available results) is largely sufficient for most purposes and gives further support to the use of this computational level for geometry optimizations and harmonic frequency evaluations.

Zero point and thermal contributions have a nonnegligible effect, leading to a significant destabilization of structure II and a strong stabilization of structure III (see Table 1). Inclusion of anharmonic contributions in ZPEs is needed for obtaining quantitative results but does not alter the stability order of the different conformers. Finally, the main effect of the QH corrections is to reduce the overstabilization of structure III produced by the harmonic oscillator model (see Table 1).

A shorter N···O distance in the II form with respect to I parallels the greater strength of the OH···N hydrogen bond with respect to its NH···O counterpart. Despite these relative hydrogen-bond strengths, the I conformer is more stable than II by about 230 cm^{-1} due to the more favorable (ω = 180°

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 2. Rotational Constants (MHz), Quadrupole Coupling Constants (χ in MHz), and Dipole Moment Components (μ in debye) of the Detected Conformers of Glycine

Parameter	Iexp[a]	Icalc[b]	IIexp[a]	II calc[b]
A_0	10341.5279(49)	10311.35	10130.1521(57)	10144.00
B_0	3876.1806(23)	3865.70	4071.5120(17)	4059.68
C_0	2912.3518(16)	2904.74	3007.4852(14)	2999.51
χ_{aa}	−1.208(9)	−1.336	1.773(2)	1.922
χ_{bb}	−0.343(8)	−0.448	−3.194(4)	−3.344
χ_{cc}	1.551(9)	1.785	1.421(4)	1.422
μ_a	0.911(3)	1.01	5.372(34)	5.39
μ_b	0.607(5)	0.66	0.93(1)	0.83
μ_c	0.0	0.0	0.0	0.03

[a]From ref 40. Standard errors are given in parenthesis in units of the last digit. [b]rDSD-LRA equilibrium geometries, rDSD equilibrium properties, and B3 vibrational corrections (only for rotational constants).

Table 3. Relative Electronic Energies (ΔE), Enthalpies at 0 K (ΔH_0° = Δ(E+ZPE)), and Free Energies at Room Temperature (ΔG°) for the Alanine Conformers[a]

Conformer	Label	ΔE_{ChS}[b]	ΔE_{rDSD}	ΔH_{0H}°[c]	ΔG_H°[d]	ΔZPE[e]	$\Delta(T\Delta S)$[f]	ΔG°[g]
ttt	I	0.0	0.0	0.0	0.0	0.0	0.0	0.0
cg$^-$c	II$^-$	35.6	29.5	172.5	316.5	−28.1	−46.8	241.6
cgc	II	103.1	106.4	215.6	321.4	−26.7	−20.4	274.3
tg$^-$t	III$^-$	432.6	412.8	443.0	386.4	−12.5	42.1	416.0
tgt	III	436.0	410.1	452.1	329.9	*h*	−81.0	*h*
ga$^-$t	I′	396.5	406.7	389.7	448.0	6.5	−25.7	428.8
gat	I′$^-$	446.6	480.5	425.9	490.4	−8.1	−19.8	462.5
ggt	III′	613.5	655.6	592.4	631.7	4.0	−0.8	634.9
gg$^-$t	III′$^-$	789.7	782.7	774.3	804.6	1.3	−5.7	800.2
tsc	Ic	1736.0	1730.3	1681.6	1696.1	−0.7	−16.0	1679.4
ts$^-$c	III$^-$c	1980.5	1968.5	1928.2	2006.6	−54.3	−12.5	1839.8
g$^-$tc	I′c	2116.5	2123.5	2052.8	2105.8	−28.1	−24.8	2052.9
gtc	I′$^-$c	2154.9	2165.3	2043.0	1983.5	27.3	25.2	2036.0

[a]All the data are in cm^{-1}. [b]JunChSF12 at rDSD geometries. [c]JunChSF12 electronic energies with rDSD harmonic ZPE. [d]JunChSF12 electronic energies with rDSD harmonic ZPE and thermal contributions. [e]Difference between anharmonic and harmonic ZPE at the B3 level. [f]Difference between quasi-harmonic and harmonic $T\Delta S$ (see text for details). [g]Sum of columns 6, 7, and 8. [h]No minimum at the B3 level.

versus ω = 0°) arrangement of the carboxylic group in the I form. The role of the arrangement of the carboxylic group is confirmed by the nearly constant destabilization of the Ic and I′c forms with respect to their I and I′ counterparts (1690 cm^{-1} for Ic vs I and 1687 cm^{-1} for I′c vs I′). At the same time, the reduced stability of the III form with respect to I (about 600 cm^{-1}) is related to the lower strength of the bifurcated $NH_2\cdots O(H)$ hydrogen bond with respect to its $NH_2\cdots O(=C)$ counterpart for identical arrangements of the carboxylic moiety. Finally conformers I′ and III′ are less stable than their I and III counterparts (by 430 and 330 cm^{-1}, respectively) because a bifurcated hydrogen bond is replaced by a more conventional single hydrogen bond. This trend could change in the presence of polar side chains because it allows the formation of additional backbone (side chain) hydrogen bonds (vide infra).

Computation of energy barriers ruling the interconversion between pairs of adjacent conformers shows that structures III and I′ relax easily to structure I (with energy barriers of about 250 and 70 cm^{-1}, respectively), whereas structure I′c relaxes to structure Ic (with an energy barrier of about 25 cm^{-1}). Furthermore, the relative stability of structures III′ (927 cm^{-1}), Ic (1679 cm^{-1}), and IIIc (2043 cm^{-1}) are too low to permit their unequivocal characterization by MW spectroscopy. We are thus left with only two conformers (I and II), which could be (and have actually been) detected in MW experiments.[40]

The availability of the experimental rotational constants for several isotopic species allowed the determination of very accurate semiexperimental equilibrium structures.[102] For the I conformer, the MAX and MUE of rDSD geometrical parameters with respect to their semiexperimental counterparts are 0.0049 and 0.0019 Å for bond lengths and 0.46 and 0.15° for valence angles. The rDSD-LRA model does not change the situation for valence angles but reduces the errors of bond lengths by about five times (0.0008 and 0.0004), reaching the accuracy of state-of-the-art composite methods.[11,102] More generally, all the computed spectroscopic parameters of the I and II conformers are in remarkable agreement with their experimental counterparts[40] (see Table 2), with MAX and MUE of 30.2 and 13.6 MHz for rotational constants, 0.23 and 0.13 MHz for quadrupole coupling constants, and 0.1 and 0.05 D for dipole moment components. The errors for rotational constants and quadrupole coupling constants are close to those delivered by the ChS composite method (MAX and MUE of and 60.8 and 16.5 MHz for rotational constants 0.19 and 0.10 for quadrupole coupling constants).[99] These results confirm that junChSF12 relative energies, rDSD-LRA structural parameters, and rDSD spectroscopic parameters can be confidently used for the comparison with experiments and represent reliable benchmarks for less refined quantum chemical methods.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Moving to alanine,[41,103−108] the two sides of the average backbone plane are no longer equivalent, with two nearly isoenergetic minima (corresponding to positive or negative values of the ψ dihedral angle) being expected at least for structures of II, I′, III′, and I′c type. The number of conformers thus increases to 12, but unconstrained geometry optimizations lead also to a splitting of structure III into III and III$^-$, although the energy difference is so tiny that an effective planar structure is expected. In all the energy minima the methyl group is found in a staggered position with respect to the substituents at C^α with rotational barriers of about 1200 cm^{-1}, close to the value of 1140 cm^{-1} obtained for ethane at a comparable computational level.[109]

The MAX and MUE of rDSD computations with respect to the junChSF12 reference (42.1 and 15.7 cm^{-1}) are more than five times smaller than the corresponding B3 values (222.2 and 91.6 cm^{-1}) and less than half the corresponding MP2 values (96.5 and 28.8 cm^{-1}). What is even more important, junChSF12 and rDSD provide the same stability order, whereas B3 and MP2 computations overestimate the stability of type II conformers (see Table S3 of the SI).

As already mentioned, the comparison with experiment requires the computation of the relative free energies for the different conformers at the temperature of the carrier gas (in order to evaluate their population) and of transition states ruling their interconversion. The results collected in Table 3 show that all the conformers involving ω values around 0° (Ic, III$^-$c, I′$^-$c, and I′c) are too unstable to permit their unequivocal detection in MW experiments. Furthermore, relaxation of I′ and III′ conformers to their more stable I and III counterparts is ruled by low energy barriers, which are easily overcome in the typical conditions of supersonic-jet expansion. Low energy barriers govern also the relaxation of III to I and II to II$^-$ conformers. As a consequence, only the I and II$^-$ conformers could be detected in MW studies, with the former collecting the populations of I, III$^-$, III, I′, I′$^-$, III′, and III′$^-$ conformers and the latter those of the II and II$^-$ conformers. It is remarkable that the relative population of conformer I computed at room temperature from the free energies collected in Table 1 (76%) is in good agreement with the experimental estimate (80%),[41] whereas a significantly lower relative population (54%) would have been predicted neglecting zero point and thermal effects.

Table 4 collects the experimental and computed rotational parameters for the I and II$^-$ conformers. A remarkable agreement is noted with the MAX and MUE of rDSD-LRA/B3 rotational constants (36.1 and 10.5 MHz) being even better than those (50.5 and 12.2 MHz) obtained at the much more expensive CCSD(T)/cc-pVTZ level.[106] It is noteworthy that for both conformers of alanine the error on the B_0 rotational constant is much higher than those affecting the other two rotational constants, whereas in both the observed conformers of glycine the largest error was found for A_0.

Table 4. Rotational Constants and Quadrupole Coupling Constants (χ) in MHz of the Detected Conformers of Alanine

Parameter	Iexp[a]	Icalc[b]	IIAexp[a]	IIAcalc[b]
A_0	5066.1455(7)	5061.61	4973.0546(35)	4972.03
B_0	3100.9507(5)	3070.85	3228.3375(56)	3192.26
C_0	2264.0131(4)	2273.39	2307.8090(42)	2326.00
χ_{aa}	−3.2567(11)	−3.4864	0.4515(17)	0.8298
χ_{bb}	2.0093(16)	1.9918	0.3267(21)	0.4207
χ_{cc}	1.2474(16)	1.4946	−0.7782(21)	−1.2505

[a]From ref 41. Standard errors are given in parenthesis in units of the last digit. [b]rDSD-LRA equilibrium geometries, rDSD properties, and B3 vibrational corrections.

The geometrical parameter most sensitive to conformational changes is the $NC^\alpha C'$ valence angle, which decreases by about 3.5° when going from the I to the II$^-$ conformer, consistent with the trans-angle rule of hyperconjugative and steric effects.[110] At the same time, the C═O bond length shows the expected lengthening by about 0.002−0.003 Å when going from free (structure II$^-$) to hydrogen-bonded (structure I) forms.

The only significant differences between the geometrical parameters of glycine and those of alanine concerns the C^α−C′ bond length (shorter in glycine by about 0.007 Å for both conformers) and the $NC^\alpha C'$ valence angle (narrower in glycine by about 2° for both conformers). Therefore, the main structural differences between glycine and alanine are highly localized at the C^α. As already mentioned, the ψ torsional angle characterizes the backbone deviation from planarity (see Tables S2 and S3 of the SI). For I conformers, it is exactly equal to 180° in glycine, whereas the lack of any symmetry induces a change of more than 15° in alanine. On the other hand, comparable ψ values are observed for the II forms of glycine and alanine (12° and 15°, respectively).

5.2. Amino Acids with Polar Side Chains. Systematic investigations have revealed that, in analogy with alanine, the natural amino acids containing simple nonpolar side chains (valine,[42] isoleucine,[43] and leucine[44]) present two dominant conformers of types I and II, respectively. On the other hand, the conformational landscape of natural amino acids with polar side chains is much richer due to the synergy or competition between intrabackbone and backbone (side chain) hydrogen bonds.

Let us start our discussion from serine (Ser), which has two soft degrees of freedom in its CH_2OH side chain (χ_1 = N−C^α−C^β−O and χ_2 = C^α−C^β−O−H), with the OH moiety able to act either as donor or acceptor in quite strong intramolecular hydrogen bonds.[111] The increased number of soft degrees of freedom (from 3 to 5) makes this system suitable for applying the PES exploration strategy introduced in the previous sections, which produces 12 low-energy conformers (see Table 5).

However, the IIg$^-$g$^-$ conformer relaxes to the more stable IIg$^-$t form through rotation around χ_2; IIIg$^-$g relaxes to Ig$^-$g through rotation around ψ; the less stable IIg$^-$t conformer relaxes to its more stable counterpart through a planar structure (invert ϕ', ψ, and ω); Igt relaxes to I′gg$^-$ through rotation around χ_2, and Igg relaxes to III′gg through rotation around ψ. We are thus left with seven conformers possibly detectable in MW experiments: three of type II, two of type III′, and one each for types I and I′ (see Figure 3).

All the most stable conformers are stabilized by both intrabackbone and backbone (side chain) hydrogen bonds (see Figure 3). Furthermore, contrary to III conformers, III′ structures are locked in sufficiently deep wells to become detectable by one HNH···OH (III′gg) or OH···O═C (III′tg$^-$) hydrogen bond between the backbone and the side chain in addition to the intrabackbone HNH···OH hydrogen bond.

ZPEs and thermal contributions alter the ordering of the four most stable conformers stabilizing, as usual, structures of

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 5. rDSD Relative Electronic Energies, Harmonic Zero Point Energies, Thermal Contributions, and Quasi-harmonic Corrections, together with Difference with JunChSF12 Electronic Energies and B3 Anharmonic Corrections (all in cm^{-1}) for the Low-Lying Conformers of Serine[a]

Label	ΔE_{rDSD}	ΔChS	ΔZPE_H	ΔTh_H	$\Delta ZPE_{(anh-H)}$	$T\Delta S_{(QH-H)}$	ΔG°[b]	ϕ'	ψ	ω	χ_1	χ_2
IIgg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−33.7	21.9	−6.2	59.4	79.3
Ig^-g	161.8	11.2	−121.8	−113.1	21.3	7.7	−32.9	159.3	166.5	177.3	−55.9	44.1
$IItg^-$	222.2	11.4	24.3	54.7	−10.0	−39.7	262.9	−31.5	19.2	−4.5	−171.9	−54.2
$I'gg^-$	337.7	−43.1	−122.4	−34.9	9.3	−11.4	135.2	95.1	−173.4	−180.0	57.1	−46.9
III′gg	531.9	−0.4	−60.7	−54.6	10.9	9.6	436.7	−168.9	67.2	−177.1	59.4	68.6
IIg^-t	602.4	32.2	−71.9	−28.3	15.9	−39.5	510.8	30.3	−14.2	2.4	−60.0	178.3
$III'tg^-$	792.7	8.2	−32.6	−192.3	−2.7	11.4	584.7	178.0	64.9	−178.3	−178.4	−70.6
IIg^-g^-[c]	607.9	43.0	−46.5	−9.6	−12.6	−11.0	571.2	29.8	−15.7	3.4	−58.6	−76.9
$III'g^-g$[d]	731.9	34.1	−120.7	−197.7	5.3	61.9	514.8	165.8	−27.4	−176.7	−56.4	43.0
IIg^-t[e]	759.4	35.0	−129.3	−81.6	5.0	−14.0	574.5	−34.9	18.5	−3.6	−58.9	−174.3
Igt^-[f]	853.1	4.5	−225.8	−131.5	6.5	5.3	512.1	−169.6	−179.6	−179.1	65.1	−175.6
Igg[g]	869.3	−6.9	−188.3	−185.0	9.6	11.4	510.1	−164.3	−165.8	−176.5	66.4	83.3

[a]Best estimates of relative free energies at room temperature (ΔG° in cm^{-1}) and dihedral angles optimized at the rDSD level (ϕ', ψ, ω, χ_1 = N−C^{α}−C^{β}−O and χ_2 = −C^{α}−C^{β}−O−H in degrees) are also given. See main text for details. [b]Sum of columns 2, 3, 4, 5, 6, and 7. [c]Relaxes to IIg^-t. [d]Relaxes to Ig^-g. [e]Relaxes to the other IIg^-t form. [f]Relaxes to Igg^-. [g]Relaxes to III′gg.

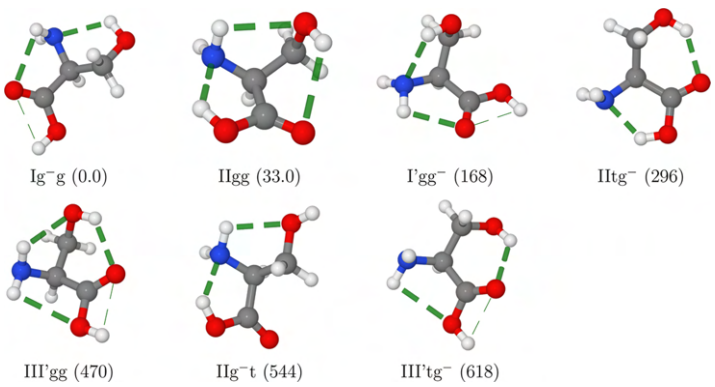


Figure 3. Representations of the seven serine conformers detected in MW spectra with the computed relative free energies at room temperature (in cm^{-1}) given in parentheses. H-bonds are highlighted by dashed lines.

Table 6. Ground-State Rotational Constants (A_0, B_0, and C_0 in MHz), ^{14}N-Nuclear Quadrupole Coupling Constants (χ in MHz), and Electric Dipole Moment Components (μ in debye) of the Seven Most Stable Serine Conformers[a]

Calc.[b]	Ig^-g	IIgg	$I'gg^-$	$IItg^-$	III′gg	IIg^-t	$III'tg^-$
A_0^c	4461.34	3549.33	3505.74	3630.86	3950.32	4508.13	3464.84
B_0^c	1823.01	2372.38	2305.21	2382.52	2222.91	1843.00	2304.68
C_0^c	1441.95	1734.67	1803.62	1515.28	1657.03	1462.05	1604.74
χ_{aa}	−4.5535	−3.6696	−0.9235	−3.8114	−0.6094	−0.3660	−1.0975
χ_{bb}	2.8681	2.1341	2.5528	2.1268	−0.6702	2.0569	−0.6582
χ_{cc}	1.6854	1.5355	−1.6293	1.6847	1.2796	−1.6909	1.7557
μ_a	1.8574	2.1328	−0.4050	−0.7709	−2.5568	4.0962	−2.8253
μ_b	−0.2255	−3.1566	−0.7361	4.8433	−0.2893	−1.7795	−0.5939
μ_c	0.7853	−1.4660	−2.7540	−0.1467	−0.5279	0.2594	0.5548
ΔG^0	0.0	32.9	168.1	295.8	469.2	543.7	617.6
Exp.[c]	L	M	N	O	P	R	Q
A_0	4479.0320(12)	3557.20088(35)	3524.38806(41)	3638.05784(38)	3931.7548(76)	4517.473(17)	3510.4015(35)
B_0	830.16170(25)	2380.37208(40)	2307.76826(70)	2387.89651(99)	2242.76701(70)	1846.99360(30)	2321.90829(24)
C_0	1443.79545(28)	1740.92458(10)	1805.20788(60)	1519.18716(36)	1664.53012(57)	1463.79646(31)	1584.38608(32)
χ_{aa}	−4.3023(27)	−3.4616(19)	−1.1343(35)	−3.6257(57)	−0.6733(67)	−0.6066(55)	−1.0486(55)
χ_{bb}	2.82359(63)	2.07974(93)	2.5043(50)	2.06213(26)	−0.456(16)	2.0723(82)	−0.5637(53)
χ_{cc}	1.4788(46)	1.3819(47)	−1.3701(50)	1 0.5906(50)	1.129(16)	−1.466(30)	1.612(21)

[a]Relative free energies at room temperature (ΔG^0 in cm^{-1}) are also reported. [b]Computed data are at the rDSD level (including LRA corrections for equilibrium rotational constants) except for electronic energies (junChSF12) and vibrational corrections to equilibrium rotational constants (B3). [c]Standard errors are shown in parentheses in units of the last digits.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 7. rDSD Relative Electronic Energies, Harmonic Zero Point Energies, Thermal Contributions, and Quasi-harmonic Corrections, together with Difference with JunChSF12 Electronic Energies and B3 Anharmonic Corrections (all in cm^{-1}) for the Low-Lying Conformers of Threonine[a]

Label	ΔE_{rDSD}	ΔChS	ΔZPE_H	ΔTh_H	$\Delta ZPE_{(anh-H)}$	$T\Delta S_{(QH-H)}$	$\Delta G^{\circ b}$	ϕ'	ψ	ω	χ_1	χ_2
IIgg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−33.7	21.8	−6.1	60.0	77.1
Ig$^-$g	218.5	34.5	−85.9	−83.0	27.7	20.4	132.2	162.5	143.7	177.8	−55.4	42.4
IItg$^-$	371.5	2.8	24.4	58.6	20.1	−30.5	446.9	−25.4	13.1	−2.3	−168.8	−53.7
I′gg$^-$	459.8	−36.4	−93.7	13.4	21.3	−14.7	349.7	99.6	−175.8	−179.2	56.9	−47.3
III′g$^-$g	574.6	45.8	−119.6	−135.4	44.4	45.4	455.2	168.2	−51.1	−179.4	−56.1	42.1
III′gg	624.2	10.8	−76.8	−62.6	−37.4	4.4	462.6	−170.6	72.6	−176.4	57.8	65.8
IIg$^-$t	711.2	9.2	−68.8	−25.5	7.5	−8.1	625.5	35.1	−21.0	5.4	−54.6	−177.8
IIgt[c]	586.1	6.9	−137.1	−104.9	−48.5	9.5	312.0	−26.2	11.5	−2.6	50.1	161.2
IIg$^-$g$^-$	725.9	24.4	−5.8	−9.0	0.1	−7.5	728.1	34.1	−21.8	6.1	−51.8	−84.1
Igt[d]	962.3	8.8	−242.2	−137.4	4.0	8.4	603.9	−172.7	178.6	−179.4	64.0	179.4

[a]Best estimates of relative free energies at room temperature (ΔG° in cm^{-1}) and dihedral angles optimized at the rDSD level (ϕ', ψ, ω, χ_1 = N−C^α−C^β−O, and χ_2 = C^α−C^β−O−H in degrees) are also given. See main text for details. [b]Sum of columns 2, 3, 4, 5, 6, and 7. [c]Relaxes to IIgg. [d]Relaxes to Ig$^-$g.

type I with respect to their type II counterparts. The stability order provided by the computed free energies at room temperature matches perfectly the estimate based on the relative intensities of the MW signals.[45] According to both theory and experiments, the first four conformers (one of type I, one of type I′, and two of type II) are significantly more stable than the two conformers of type III′ and a further conformer of type II, which have, in turn, comparable stability (see Table 6).

The rotational constants of the two most stable conformers have been recently computed by geometry optimizations at the ChS level, reaching MAX and MUE of 28.7 and 10.6 MHz, respectively.[23,112] It is noteworthy that even smaller MAX and MUE (17.7 and 8.1 MHz, respectively) are obtained at the rDSD-LRA level, whose strongly reduced cost has allowed us to compute the spectroscopic parameters of all the other low-energy conformers. The remarkable agreement between computed and experimental results for all the detected conformers of serine confirms the accuracy of our computational strategy.

The next studied system is threonine (Thr),[113] in which a methyl group replaces one of the hydrogen atoms bonded to C^β, leading to the $CHCH_3OH$ side chain which has again two soft degrees of freedom (χ_1 = N−C^α−C^β−O and χ_2 = C^α−C^β−O−H) since the terminal methyl group is frozen in a staggered conformation with an estimated rotation barrier of 1400 cm^{-1}. There is now a second chiral center in addition to the C^α atom, with the natural amino acid being 2*S*,3*R*-threonine. The conformational landscape of threonine has been investigated in two different studies,[113,114] which obtained 71 and 56 conformers, respectively, in a range of about 4000 cm^{-1}, but the final set of conformers was the same up to a relative energy of 1600 cm^{-1}. The knowledge-based step of our conformational exploration started from the 12 low-energy conformers of serine collected in Table 5, each of them being then split into two nonequivalent structures. Next, the IM-EA algorithm was used to generate additional low-energy minima. At the end of these two steps and the subsequent filtering/refinement we are left with the 10 low-energy conformers (within an energy range of 1000 cm^{-1}) collected in Table 7. It is noteworthy that this finding is in full agreement with ref 114.

The predicted population of conformer IIg$^-$g$^-$ is too low to allow its detection in MW experiments, and conformers IIgt and Igt relax easily to conformers IIgg and Ig$^-$g, respectively. We are thus left with the same number (seven) and backbone conformation (three conformers of type II, two of type III′, and one each for types I and I′) of the structures discussed above for serine, which should be (and have actually been[46]) detected in MW experiments. However, the presence of the β methyl group increases the energy barrier governing relaxation of the III′g$^-$g conformer to its Ig$^-$g counterpart from about 200 to about 800 cm^{-1} when going from serine to threonine. As a consequence, the III′g$^-$g conformer is observed in threonine in place of the less stable III′tg$^-$ conformer observed in serine (see Figure 4). At the same time, a general destabilization of all conformers with respect to IIgg accompanies the substitution of a β hydrogen atom with a methyl group (see Figure 5).

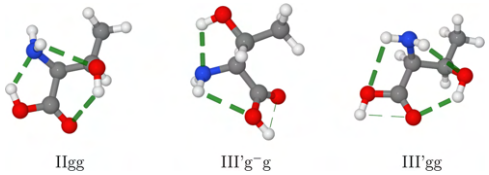


Figure 4. Absolute energy minimum and low-lying III′ conformers of threonine. The H-bonds are highlighted by dashed lines.

The two most stable (IIgg and Ig$^-$g) and the three least stable (III′g$^-$g, III′gg, and IIg$^-$t) conformers are the same in terms of electronic energies, enthalpies, or free energies. The relative ordering of the two intermediate conformers is, instead, altered by both ZPE and thermal contributions.

All the spectroscopic parameters of the seven low-energy conformers of threonine detected in a recent microwave study[46] show a remarkable agreement with those computed for the most stable conformers predicted by our computations (see Table 8). The relative stability order estimated from the experimental results is Ig$^-$g > IIgg > I′gg$^-$ > IIg$^-$t ≈ III′g$^-$g ≈ IItg$^-$ ≈ III′gg, which is in general agreement with the computed relative free energies except for the inversion between Ig$^-$g and IIgg conformers and the position of the IIg$^-$t structure.

Replacement of the oxygen atom in the side chain of serine by a sulfur produces cysteine (Cys), whose CH_2SH side chain has again two soft degrees of freedom (χ_1 = N−C^α−C^β−S and χ_2 = C^α−C^β−S−H). One might think that the same

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

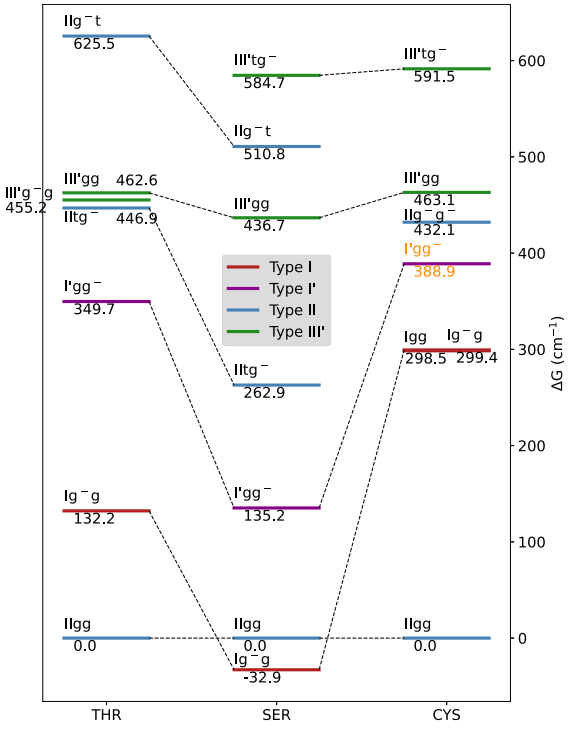


Figure 5. Observed conformers of threonine, serine, and cysteine. The relative free energies at room temperature (ΔG in cm^{-1}, see text for details) are given for each amino acid with respect to its IIgg conformer. The relations between the observed conformers of the three amino acids are highlighted with dashed lines. The conformer I′gg⁻ of cysteine, which has not been detected in MW studies, is reported with orange labels.

conformers should be detected for cysteine and serine. However, the strengths of the H-bonds possibly formed by the thiol group are weaker than those of its alcohol counterpart. Therefore, it is expected that the barriers separating low-lying conformers decrease and in some instances may even disappear. In ref 115, a systematic scan of the conformational PES at the MP2/cc-pVTZ level led to the identification of 71 unique conformers, thus defining a reference data set. The knowledge-based step of our PES exploration involved the 12 low-energy conformers found for serine and integration of these structures with those issued from the IM-EA exploration employing sufficiently high energy thresholds allowed us to retrieve all the structures of the reference data set.[115] Then, refinement of the results by the usual energy tresholds led to the 9 conformers collected in Table 9. The rDSD results are once again in very good agreement with their junChSF12 counterparts (MAX and MUE of 44 and 24 cm^{-1}, respectively).

Among those nine conformers, the two least stable ones have too low populations to allow their unequivocal experimental characterization and the conformer I′gg⁻ relaxes easily to its Ig⁻g counterpart, which has a similar shape. Therefore, the number of detectable conformers reduces to 6: two each for types I, II, and III′ (see Figure 6). The backbone structure of the most stable conformer and the general trends are similar to those discussed above for serine and threonine (see Figure 5), but the conformers Igg and IIg⁻g⁻ replace the I′gg⁻ and IIg⁻t counterparts observed in both serine and threonine.

The spectroscopic parameters computed at the rDSD level are in remarkable agreement with their experimental counterparts[39] with MUEs of 11.7, 5.7, and 3.1 MHz for the A_0, B_0, and C_0 rotational constants, respectively (Table 10). The errors on B_0 and C_0 are quite low already at the rDSD level (see Table S9 of the SI), whereas errors as large as 40 MHz are obtained for the A_0 rotational constant. For most conformers,

Table 8. Ground-State Rotational Constants (A_0, B_0, and C_0 in MHz), ^{14}N-Nuclear Quadrupole Coupling Constants (χ in MHz), and Electric Dipole Moment Components (μ in debye) of the Seven Most Stable Conformers of Threonine[a]

	IIgg	Ig⁻g	I′gg⁻	IItg⁻	III′g⁻g	III′gg	IIg⁻t
				Computed[b]			
A_0	3223.67	2864.48	3141.58	2671.88	2885.67	3375.76	2907.62
B_0	1528.34	1602.22	1501.39	1774.76	1564.86	1474.72	1656.40
C_0	1265.11	1214.77	1313.14	1376.76	1243.85	1234.69	1187.62
χ_{aa}	−3.5846	−4.3527	−0.2035	−4.1467	−4.2988	−2.2144	−0.3702
χ_{bb}	1.7308	2.6918	2.9006	2.4728	2.5948	−0.2748	2.6214
χ_{cc}	1.8538	1.6609	−2.6971	1.6739	1.7040	2.4892	−2.2513
μ_a	2.85	−2.06	−0.19	0.04	−1.86	−2.23	3.78
μ_b	2.95	0.01	−0.34	4.91	1.58	0.98	2.00
μ_c	−0.94	0.97	−2.90	−0.16	1.33	−0.87	0.19
ΔG^0	0.0	132.2	349.7	446.9	455.2	462.6	625.5
				Experimental[c]			
A_0	3232.4827(12)	2872.77049(48)	3148.59247(32)	2670.72096(53)	2889.93352(45)	3379.841(14)	2912.6227(20)
B_0	1533.71801(32)	1608.95699(26)	1506.27679(37)	1784.66894(60)	1572.32152(50)	1482.04984(21)	1660.21807(34)
C_0	1267.88615(34)	1211.39762(38)	1316.33575(44)	1383.75384(51)	1241.83423(47)	1237.59121(22)	1189.31443(34)
χ_{aa}	−3.4971(21)	−4.1859(25)	−0.7403(21)	−3.7652(73)	−4.1529(32)	−2.201(14)	−0.544(11)
χ_{bb}	1.7519(27)	2.661(42)	2.8781(28)	2.4258(75)	2.5682(46)	−0.157(50)	2.582(16)
χ_{cc}	1.7452(60)	1.5248(17)	−2.1378(70)	1.3394(20)	1.5846(46)	2.358(64)	−2.038(50)

[a]The computed relative free energies at room temperature (ΔG^0 in cm^{-1}) are also reported. [b]Computed data are at the rDSD level (including LRA corrections for equilibrium rotational constants) except for electronic energies (junChSF12) and vibrational corrections to equilibrium rotational constants (B3). [c]Standard errors are shown in parentheses in units of the last digits.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 9. rDSD Relative Electronic Energies, Harmonic Zero Point Energies, Thermal Contributions, and Quasi-harmonic Corrections, together with Difference with JunChSF12 Electronic Energies and B3 Anharmonic Corrections (all in cm^{-1}) for the Low-Lying Conformers of Cysteine[a]

Label	ΔE_{rDSD}	ΔChS	ΔZPE_H	ΔTh_H	$\Delta ZPE_{(anh-H)}$	$T\Delta S_{(QH-H)}$	ΔG°[b]	ϕ'	ψ	ω	χ_1	χ_2
IIgg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−32.7	18.6	−4.8	57.1	71.8
IIg$^-$g$^-$	501.1	36.3	−24.6	−53.3	−7.6	−19.8	432.1	34.4	−18.0	4.0	−60.9	−65.4
Igg	571.1	−9.3	−177.6	−196.9	47.8	63.4	298.5	−171.3	−175.8	−177.4	63.7	74.7
Ig$^-$g	630.7	8.9	−180.6	−198.4	10.0	28.8	299.4	162.9	162.6	177.5	−65.2	51.0
III′gg	706.7	−18.5	−123.9	−168.2	1.6	65.4	463.1	−172.3	34.5	177.4	61.6	76.1
III′tg$^-$	873.8	39.0	−153.0	−185.5	−22.5	39.7	591.5	175.6	85.6	−175.8	−175.4	−75.8
I′gg$^-$[c]	722.3	−44.1	−201.0	−85.4	1.2	−4.2	388.8	98.9	−173.0	179.8	64.3	−52.5
III′gg$^-$	950.0	41.2	−79.1	−164.4	−15.2	19.3	751.8	114.0	79.9	−65.7	88.3	16.7
IIgt	1056.5	−0.4	−38.8	−24.8	17.0	−13.8	995.7	152.9	1.8	−27.2	100.7	27.0

[a]Best estimates of relative free energies at room temperature (ΔG° in cm^{-1}) and dihedral angles optimized at the rDSD level (ϕ', ψ, ω, χ_1 = N−C$^{\alpha}$−C$^{\beta}$−S, and χ_2 = C$^{\alpha}$−C$^{\beta}$−S−H in degrees) are also given. See main text for details. [b]Sum of columns 2, 3, 4, 5, 6, and 7. [c]Relaxes to Ig$^-$g.

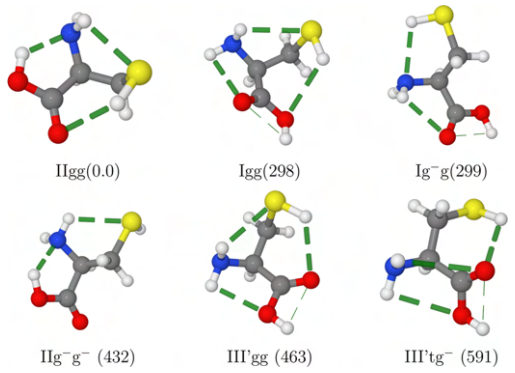


Figure 6. Cysteine conformers detected in MW experiments with the computed relative free energies at room temperature (in cm^{-1}) given in parentheses. H-bonds are highlighted by dashed lines.

the C−S bond is nearly perpendicular to the average backbone direction (see Figure 6) and is, in turn, roughly aligned with the *a* axis. As a consequence, any overestimation of the C−S bond length results in a nonnegligible underestimation of the A_0 rotational constant. In this connection, the LRA correction brings the computed values in remarkable agreement with experiment (the maximum error is obtained for the Ig$^-$g conformer and amounts to 18 MHz, i.e., 0.4%).

Let us now analyze aspartic acid, the simplest amino acid containing two carboxylic groups. The CH_2COOH side chain has three dihedral angles (χ_1 = N−C$^{\alpha}$−C$^{\beta}$−C$^{\gamma}$, χ_2 = C$^{\alpha}$−C$^{\beta}$−C$^{\gamma}$−O(H) and χ_3 = C$^{\beta}$−C$^{\gamma}$−O−H). However, χ_3 is frozen in trans (favored and not explicitly labeled in the following) or cis (labeled by c in the following) conformations. A recent systematic analysis of the conformational landscape[116] identified 19 energy minima in a range of 3500 cm^{-1}, and we were able to locate all those minima by our general exploration strategy with enlarged energy thresholds. Within this panel of candidates, only 9 conformers have electronic energies lying within 1000 cm^{-1} above the absolute energy minimum (see Table 11). Once again, a good quantitative agreement is observed between junChSF12 and rDSD results

Table 10. Ground-State Rotational Constants (A_0, B_0, and C_0 in MHz), ^{14}N-Nuclear Quadrupole Coupling Constants (χ in MHz), and Electric Dipole Moment Components (μ in debye) of the Six Most Stable Energy Minima of Cysteine[a]

Calc.[b]	IIgg	Igg	Ig$^-$g	III′gg	IIg$^-$g$^-$	III′tg$^-$
A_0	3063.27	2874.44	4217.57	3223.13	4352.34	2989.53
B_0	1600.59	1615.60	1181.79	1563.71	1173.71	1524.30
C_0	1327.34	1366.95	1000.82	1267.50	1012.74	1210.12
χ_{aa}	−3.3302	−0.0280	−4.5456	0.0509	−0.1942	0.5818
χ_{bb}	2.5198	0.3553	2.8019	−0.5218	2.2497	−2.1507
χ_{cc}	0.8104	−0.3273	1.7437	0.4708	−2.0555	1.5689
μ_a	1.40	−1.02	−1.81	2.86	2.33	−2.12
μ_b	3.98	−1.43	0.37	−2.42	−0.18	0.31
μ_c	−1.53	−1.39	0.57	1.36	−0.20	−0.02
ΔG^0	0.0	187.3	260.6	396.1	459.5	574.3
Exp.[c]	O	N	L	P	M	Q
A_0	3071.1437(15)	2889.44652(93)	4235.63210(58)	3216.218(26)	4359.22320(77)	3004.1689(90)
B_0	1606.53664(36)	1622.99829(32)	1187.27897(20)	1572.74943(63)	1178.27610(13)	1527.40718(53)
C_0	1331.80185(34)	1367.83448(26)	1003.10663(23)	1276.79135(55)	1015.27433(13)	1210.70722(46)
χ_{aa}	−3.1200(53)	−0.1465(36)	−4.263(11)	0.0	−0.4060(9)	0.505(10)
χ_{bb}	2.4418(61)	0.4419(43)	2.776(11)	−0.449(25)	2.2314(43)	−1.991(20)
χ_{cc}	0.6782(61)	−0.2954(43)	1.488(11)	0.449(25)	−1.8254(43)	1.486(20)

[a]The computed relative free energies at room temperature (ΔG^0 in cm^{-1}) are also reported. [b]Computed data are at the rDSD level (including LRA corrections for equilibrium rotational constants) except for electronic energies (junChSF12) and vibrational corrections to equilibrium rotational constants (B3). [c]Standard errors are shown in parentheses in units of the last digits.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 11. rDSD Relative Electronic Energies, Harmonic Zero Point Energies, Thermal Contributions and Quasi-harmonic Corrections, together with Difference with JunChSF12 Electronic Energies and B3 Anharmonic Corrections (all in cm^{-1}) for the Low-Lying Conformers of Aspartic Acid[a]

Label	ΔE_{rDSD}	ΔChS	ΔZPE_H	ΔTh_H	$\Delta ZPE_{(anh-H)}$	$T\Delta S_{(QH-H)}$	$\Delta G^{\circ b}$	ϕ'	ψ	ω	χ_1	χ_2
IIgt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	36.3	20.7	−6.3	61.7	168.8
IIg⁻t	133.5	7.5	−57.1	−60.8	−13.1	−9.7	0.3	−35.7	20.0	−4.3	−65.5	174.7
Igt	288.7	−4.3	−143.6	−80.1	−17.3	7.0	50.4	179.6	−164.5	−177.8	67.8	−177.2
Ig⁻gc	341.0	−4.1	7.3	−27.4	−38.8	−30.9	247.1	164.6	162.4	177.2	−63.0	38.8
III′gt	350.5	83.5	−114.2	−99.3	−13.7	27.3	234.1	177.0	24.4	178.1	65.9	−179.7
I′g⁻t	478.9	−40.6	−178.2	−121.6	−24.5	9.1	123.1	86.6	−167.5	−177.0	−63.9	169.8
I′gg⁻c	682.8	74.2	−4.4	18.0	−21.9	3.0	751.7	85.9	−179.1	−165.3	62.5	−36.9
IIItt	777.4	9.7	−86.0	−25.3	−12.7	−65.0	598.1	169.6	5.1	167.2	−158.3	171.1
I′tt	1136.3	−17.7	−202.4	−193.0	−17.3	51.1	757.0	62.7	−179.5	58.6	−173.7	−160.1

[b]Sum of columns 2, 3, 4, 5, 6, and 7. [a]Best estimates of relative free energies at room temperature (ΔG° in cm^{-1}) and dihedral angles optimized at the rDSD level (ϕ', ψ, ω, $\chi_1 = N-C^{\alpha}-C^{\beta}-C^{\gamma}$, and $\chi_2 = C^{\alpha}-C^{\beta}-C^{\gamma}-O(H)$ in degrees) are also given. The χ_3 angle ($C^{\beta}-C^{\gamma}-O-H$) is always close to 180° (not explicitly indicated) or 0° (evidenced by the last "c" in the conformer label). See main text for details.

with the MAX and MUE between the two methods being 83.5 and 30.2 cm^{-1} without any inversion in the relative stability order. Inclusion of zero point and thermal effects produces significant changes in the trend issued from relative electronic energies with the most striking effect being, as usual, the destabilization of all the conformers showing type II hydrogen bridges (see Table 11). The six most populated conformers shown in Figure 7 are significantly more stable than the next 3 ones, and exactly six species were detected in MW experiments.[96]

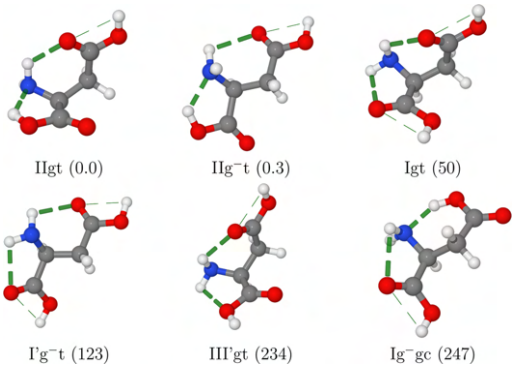


Figure 7. Conformers of aspartic acid detected in MW experiments with the computed relative free energies at room temperature (in cm^{-1}) given in parentheses. H-bonds are highlighted by dashed lines.

Both the I′g⁻t and III′gt conformers are more stable than their I and III counterparts due to the replacement of an intrabackbone bifurcated $NH_2\cdots O{=}C$ or $NH_2\cdots OH$ hydrogen bond by a single $HNH\cdots O{=}C$ or $HNH\cdots OH$ hydrogen bond plus a single $HNH\cdots O{=}C$ backbone (side chain) hydrogen bond. The increased stability explains also the absence of low-barrier relaxation paths from these conformers to I structures.

The spectroscopic parameters collected in Table 12 show a remarkable agreement between theory and experiment. It is noteworthy that previous MP2/6-311++G(d,p)[96] computations forecasted that one or two different conformers should be experimentally detected and that the spectroscopic constants obtained at that level show MAX and MUE with respect to experiment (29.2 and 10.6 MHz) more than three times larger than their rDSD-LRA counterparts (8.2 and 3.1 MHz). The rDSD MUE (smaller than 0.2%) approaches again the accuracy of state-of-the-art composite methods for small semirigid molecules[117] and permits the unbiased assignment of MW spectra.[118] The stability order of the six most populated conformers is, however, quite different between theory and experimental estimates with the strongest discrepancy concerning the inversion of the relative stability of I and II species. Although the experimental populations take into account also possible relaxation of higher-energy structures to the most stable conformers, according to the computed free energies the initial populations of all the species outside the six most stable ones are too low to alter the computed relative populations. From another point of view, the experimental estimates are based on a number of assumptions, which might not be fulfilled in the present case. Also taking these considerations in mind, the agreement between theory and experiment concerning the nature and spectroscopic parameters of all the observable species remains remarkable.

The last system considered in this study is asparagine, which is the only proteinogenic α-amino acid, together with glutamine,[119] containing an amide group. The soft degrees of freedom of the asparagine side chain (CH_2CONH_2) include two dihedral angles ($\chi_1 = N-C^{\alpha}-C^{\beta}-C^{\gamma}$, $\chi_2 = C^{\alpha}-C^{\beta}-C^{\gamma}-N$) because the coupled rotation/inversion displacements of the NH_2 amide moiety from the planar reference structure can be safely added to the panel of stiff degrees of freedom. The amide moiety can act either as a proton donor or as a proton acceptor, with this increasing the number of possible backbone (side chain) intramolecular hydrogen bonds. Asparagine in the gas-phase has been widely studied by both computational[47,120] and experimental[47,121] points of view, but a comprehensive characterization of its structure and conformational landscape has not yet been performed by state-of-the-art quantum chemical methods.

The usual exploration/refinement strategy provides 5 conformers with rDSD electronic energies within a little more than 1000 cm^{-1} above the absolute energy minimum (see Table 13). At this level only the most stable IIgg conformer (see Figure 8) should be detectable in MW experiments. The situation is thus very different from that found in the case of aspartic acid because the presence of the NH_2 amidic moiety in the side chain permits the compensation of the weak hydrogen bond in the carboxylic moiety (lacking in II structures of aspartic acid with respect to their I

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 12. Ground-State Rotational Constants (A_0, B_0, and C_0 in MHz), ^{14}N-Nuclear Quadrupole Coupling Constants (χ in MHz), and Electric Dipole Moment Components (μ in debye) of the Six Most Stable Energy Minima of Aspartic Acid[a]

Conformer	IIgt	IIg⁻t	Igt	I′g⁻t	III′gt	Ig⁻gc
			Computed[b]			
A_0	2607.9	3412.3	2546.8	3372.8	2643.8	3192.2
B_0	1188.9	900.4	1202.1	904.2	1182.9	943.8
C_0	1057.1	762.5	1067.2	778.1	1055.9	781.4
χ_{aa}	−3.7322	−3.4040	−0.2050	1.1611	−0.2629	−4.1388
χ_{bb}	2.7326	1.4552	−0.2987	2.7491	−0.3570	2.5722
χ_{cc}	0.9996	1.9488	0.5037	−3.9102	0.6199	1.5665
μ_a	2.3532	3.6076	1.0967	0.5375	0.3702	−5.2042
μ_b	4.1392	2.1025	1.2332	−1.8804	0.5037	1.1751
μ_c	−2.1974	1.4410	1.7069	−0.7507	0.2090	−0.6972
ΔG^0	0.0	0.3	50.4	123.1	234.1	247.1
Exp[c]	P	N	M	L	Q	O
A_0	2612.20878(26)	3416.43489(66)	2553.85523(70)	3378.20873(26)	2651.953(31)	3198.861(19)
B_0	1191.01132(17)	902.904474(79)	1205.08478(10)	907.373507(28)	1183.51697(30)	945.84803(7)
C_0	1057.33169(16)	764.631177(96)	1069.14318(10)	780.042139(32)	1054.98929(34)	781.75139(18)
χ_{aa}	−3.5601(63)	−3.3602(87)	−0.2774(35)	0.9560(35)	−0.295(27)	−3.995(19)
χ_{bb}	2.6538(54)	1.4823(73)	−0.2640(35)	2.7296(23)	−0.350(45)	2.524(32)
χ_{cc}	0.9064(54)	1.8778(73)	0.5414(35)	−3.6856(23)	0.645(45)	1.470(32)

[a]The computed relative free energies at room temperature (ΔG^0 in cm^{-1}) are also reported. [b]Computed data are at the rDSD level (including LRA corrections for equilibrium rotational constants) except for electronic energies (junChSF12) and vibrational corrections to equilibrium rotational constants (B3). [c]Standard errors are shown in parentheses in units of the last digits.

Table 13. rDSD Relative Electronic Energies, Harmonic Zero Point Energies, Thermal Contributions, and Quasi-harmonic Corrections, together with Difference with JunChSF12 Electronic Energies and B3 Anharmonic Corrections (all in cm^{-1}) for the Low-Lying Conformers of Asparagine[a]

Label	ΔE_{rDSD}	ΔChS	ΔZPE_H	ΔTh_H	$\Delta ZPE_{(anh-H)}$	$T\Delta S_{(QH-H)}$	$\Delta G^{\circ b}$	ϕ'	ψ	ω	χ_1	χ_2
IIgg	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−23.3	15.4	−4.9	58.5	101.0
IIg⁻t	727.6	−24.2	−222.8	−193.4	16.6	44.8	348.6	−36.9	20.5	−4.3	−65.6	177.0
Ig⁻g	826.9	27.3	−212.0	−260.1	81.8	82.5	546.4	172.4	161.0	177.3	−69.6	34.9
I′gg⁻	1016.6	8.8	−198.6	−62.1	61.6	24.1	850.4	80.6	−164.8	−178.6	69.8	−29.5
Igt	1072.6	−36.1	−367.2	−271.5	154.4	66.3	568.5	−179.7	−164.6	−178.0	67.1	−173.1

[a]Best estimates of Gibbs free energies (ΔG° in cm^{-1}) and dihedral angles optimized at the rDSD level (ϕ', ψ, ω, χ_1 = N−C$^\alpha$−C$^\beta$−C$^\gamma$, and χ_2 = C$^\alpha$−C$^\beta$−C$^\gamma$−N in degrees) are also given. The χ_3 angle (C$^\beta$−C$^\gamma$−N−H) is always close to 0°. See main text for details. [b]Sum of columns 2, 3, 4, 5, 6, and 7.

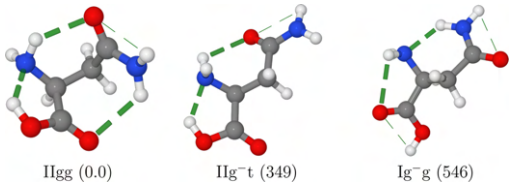


Figure 8. Most stable conformers of asparagine. The computed relative free energies at room temperature (in cm^{-1}) are given in parentheses. H-bonds are highlighted by dashed lines.

counterparts) by a backbone/side chain OH···NH$_2$ hydrogen bond without reducing the local stability of the amide moiety. In fact, an analogous situation would involve a 180° rotation of the OH moiety in the carboxylic group of the side chain in aspartic acid away from its most stable arrangement.

ZPE and thermal contributions strongly stabilize all conformers with respect to the most stable IIgg strucure, so that the IIg⁻t form (see Figure 8) might become accessible to experimental characterization. As a matter of fact, several searches of transition states connecting IIg⁻t and IIgg conformers gave quite high energy barriers preventing any effective relaxation path. As a consequence, there is a disagreement between theory and experiment[47] about the number of low-lying conformers of asparagine. However, comparison between computed and experimental spectroscopic parameters for the single conformer detected in the MW study of ref 47 shows the usual remarkable agreement (see Table 14) with MAX and MUE as low as 12.6 and 4.7 MHz for rotational constants and 0.19 and 0.08 MHz for quadrupole coupling constants.

5.3. Trends of Intramolecular Interactions. The accurate results obtained for several amino acids permit the strengths of the main interactions governing the conformational landscapes of these flexible systems to be estimated. In particular, approximate values for the strengths of different hydrogen bonds can be computed from prototypical systems and used to rationalize energy differences among the conformers of different amino acids in terms of sums of stabilizations from near-atom interactions. Based on the energy difference between Ic and I or I′c and I′ conformers of glycine, for each carboxyl group ω = 180° is more stable than ω = 0° by about 1700 cm^{-1} and the same applies to χ_3 in the case of aspartic acid. Concerning other situations, the hydrogen bond donors can be ranked in the order O−H > N−H > S−H, and the hydrogen bond acceptors in the order *N* > O > S. As a

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Table 14. Experimental[47] and Computed Ground State Rotational Constants (A_0, B_0, C_0 in MHz) and Quadrupole Coupling Constants (χ in MHz) for the IIgg Conformer of Asparagine[a]

	Experimental			Computed		
	A_0	B_0	C_0	A_0	B_0	C_0
	2270.85145(85)	1387.80238(41)	1102.63540(41)	2258.22	1387.0	1101.81
	Experimental			Computed		
	χ_{aa}	χ_{bb}	χ_{cc}	χ_{aa}	χ_{bb}	χ_{cc}
N^{α}	−2.0313(50)	2.5720(57)	−0.5408(57)	−2.2164	2.6624	−0.4459
N^{β}	−1.4649(63)	1.5518(76)	−0.0870(76)	−1.5141	1.5607	−0.0466

[a]The values in parentheses are the experimental standard errors in units of the last digit.

consequence, the strongest hydrogen bond is $H_2N\cdots H-O$ (with an estimated strength of 3200 cm^{-1}), which involves the best donor and the best acceptor, followed by $H_2-N\cdots HNH$ (with an estimated strength of 2400 cm^{-1}). Those values, together with the energy differences among conformers I, II, and III of glycine, permit strengths of about of 1700 and 1100 cm^{-1} to be estimated for the bifurcated $NH_2\cdots O{=}C$ and $NH_2\cdots O-H$ hydrogen bonds. Furthermore, the difference between the pairs I, I′ and III, III′ leads to hydrogen bond strengths of about 1300 and 800 cm^{-1} for the more conventional single $H-N-H\cdots O{=}C$ and $HNH\cdots O-H$ hydrogen bonds. Finally, a comparable strength of about 800 cm^{-1} is estimated for the $H_2N\cdots H-S$, $H-S\cdots H-N-H$, and $S-H\cdots O{=}C$ hydrogen bonds. It is then quite straightforward to understand why conformer I is more stable than its II counterpart in the absence of backbone (side chain) hydrogen bonds (e.g., in alanine): in fact, the sum of $NH_2\cdots O{=}C$ and favorable carboxyl conformation exceeds by about 200 cm^{-1} the stronger $H_2N\cdots H-O$ hydrogen bond but with an unfavorable conformation of the carboxylic moiety. On the other hand, in serine and threonine, conformer II becomes more stable due to the extra stabilization related to an $O-H\cdots O{=}C$ hydrogen bond involving the backbone and the side chain. The same occurs in cysteine, where the 800 cm^{-1} gained from the $S-H\cdots O{=}C$ hydrogen bond makes the IIgg conformer more stable than the Igg counterpart by about 600 cm^{-1}. An analogous situation is found in aspartic acid, where the amine moiety is involved at the same time in an $OH\cdots N$ hydrogen bond within the backbone and a $HNH\cdots O{=}C$ hydrogen bond with the side chain. Finally, IIgg is by far the most stable conformer in asparagine because the presence of an amide group allows the formation of two additional backbone (side chain) hydrogen bonds. Type III conformers are intrinsically less stable than their I counterparts (due to the lower strength of $NH_2\cdots O-H$ with respect to $NH_2\cdots O{=}C$ hydrogen bond), and moreover, they can easily relax to I forms through rotation around ψ when not locked by additional interactions. However, III′ conformers featuring a single $H-N-H\cdots O-H$ hydrogen bond can be stabilized and locked into sufficiently deep energy wells upon involvement of the released N−H bond into additional hydrogen bonds with the side chain. This is the case, for instance, of the III′gg conformer in serine, threonine ($H-O\cdots H-N-H\cdots O-H$), and cysteine ($H-O\cdots H-N-H\cdots S-H$).

Hydrogen bonding is surely the driving force ruling the general trends of structures and relative stabilities, but the detailed geometry and energy changes between conformers depend strongly on other stereoelectronic effects like, e.g., hyperconjugation or steric repulsion. For instance, any additive picture based on individual hydrogen bond strengths is tuned by the preference of bulky vicinal substituents for trans or gauche conformations, which, in turn, depends on the balance between electrostatic, steric, and hyperconjugative effects. Furthermore, vibrational effects (affecting both ZPEs and entropic contributions) alter the stability order provided by relative electronic energies and must be taken into the proper account.

While the reader is referred to studies of specific systems for more detailed analyses along these lines,[46,47,96,111,115] we point out that only the availability of accurate results including all the stereoelectronic and vibrational effects (like those reported in the present paper) can provide an unbiased reference for building more realistic models (e.g., force fields including non additive terms) for the study of flexible biomolecules.

6. CONCLUDING REMARKS

In this paper, a general strategy aimed at the unbiased disentanglement of the conformational bath of flexible biomolecule building blocks in the gas phase has been further improved and validated for the specific case of representative natural α-amino acids. The use of curvilinear internal coordinates permits the separation between stiff and soft degrees of freedom. Then, effective exploration of the soft variables can be performed by purposely tailored evolutionary algorithms, whose fitness scores are obtained by constrained geometry optimizations of the stiff degrees of freedom employing a fast semiempirical method. Refinement of the energies and structures by a hybrid and then a last-generation double-hybrid functional allows very reliable results to be obtained minimizing the number of expensive computations. Application of the procedure to supersonic jet experiments requires also the location of transition states ruling the interconversion between pairs of adjacent energy minima and the identification of fast relaxation processes. Improved structures and relative energies are obtained by the rDSD-LRA approach and the junChSF12 composite method, respectively. Finally, the spectroscopic parameters of sufficiently populated conformers can be safely computed at the rDSD level.

The results obtained for glycine, alanine, and, especially, different natural α-amino acids with polar side chains are in full agreement with the available spectroscopic data and permit their unbiased interpretation in terms of the cooperation or competition between intrabackbone and backbone (side chain) hydrogen bonds.

Together with the intrinsic interest of the studied molecules, the results of the present investigation show that highly reliable analysis of the conformational landscape is today possible for flexible building blocks of biomolecules in the gas phase. Furthermore, we provide benchmark results for the validation

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

of cheaper quantum chemical methods, which become unavoidable for large biomolecules.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c01143.

Full set of parameters of the IM-EA algorithm (Table S1) and different contributions to the relative energies and rotational constants of low-energy conformers for glycine (Table S2), alanine (Table S3), serine (Tables S4 and S5), threonine (Tables S6 and S7), cysteine (Tables S8 and S9), aspartic acid (Tables S10 and S11) and asparagine (Tables S12 and S13) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Vincenzo Barone − *Scuola Normale Superiore di Pisa, 56126 Pisa, Italy;* orcid.org/0000-0001-6420-4107; Email: vincenzo.barone@sns.it

Authors

Marco Fusè − *DMMT-sede Europa, Universitá di Brescia, 25121 Brescia, Italy;* orcid.org/0000-0003-0130-5175

Federico Lazzari − *Scuola Normale Superiore di Pisa, 56126 Pisa, Italy;* orcid.org/0000-0003-4506-3200

Giordano Mancini − *Scuola Normale Superiore di Pisa, 56126 Pisa, Italy;* orcid.org/0000-0002-1327-7303

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c01143

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Funding from the Italian Ministry of University and Research (MUR, Grant 2017A4XRCA) and Italian Space Agency (ASI, "Life in Space" Project No. 2019-3-U.0) is gratefully acknowledged.

■ REFERENCES

(1) Alonso, E. R.; León, I.; Alonso, J. L. *Intra- and Intermolecular Interactions Between Non-Covalently Bonded Species*; Elsevier: 2020; pp 93−141.

(2) Barone, V.; Alessandrini, S.; Biczysko, M.; Cheeseman, J. R.; Clary, D. C.; McCoy, A. B.; DiRisio, R. J.; Neese, F.; Melosso, M.; Puzzarini, C. Computational molecular spectroscopy. *Nat. Rev. Methods Primers* **2021**, *1*, 38.

(3) Schols, G. *Atomic and Molecular Beam Methods*; Oxford University Press: 1988.

(4) Alonso, J. L.; López, J. C. *Gas-Phase IR Spectroscopy and Structure of Biological Molecules*; Springer: 2015; pp 335−401.

(5) Lesarri, A.; Mata, S.; López, J. C.; Alonso, J. L. A laser-ablation molecular-beam Fourier-transform microwave spectrometer: The rotational spectrum of organic solids. *Rev. Scient. Instr.* **2003**, *74*, 4799−4804.

(6) Godfrey, P. D.; Brown, R. D. Proportions of Species Observed in Jet Spectroscopy-Vibrational Energy Effects: Histamine Tautomers and Conformers. *J. Am. Chem. Soc.* **1998**, *120*, 10724−10732.

(7) Florio, G. M.; Christie, R. A.; Jordan, K. D.; Zwier, T. S. Conformational Preferences of Jet-Cooled Melatonin: Probing trans- and cis-Amide Regions of the Potential Energy Surface. *J. Am. Chem. Soc.* **2002**, *124*, 10236−10247.

(8) Helgaker, T.; Klopper, W.; Tew, D. P. Quantitative quantum chemistry. *Mol. Phys.* **2008**, *106*, 2107−2143.

(9) Karton, A. A computational chemistas guide to accurate thermochemistry for organic molecules. *WIRES, Comp. Mol. Sci.* **2016**, *6*, 292−310.

(10) Kesharwani, M. K.; Karton, A.; Martin, J. M. Benchmark ab initio conformational energies for the proteinogenic amino acids through explicitly correlated methods. *J. Chem. Theory Comput.* **2016**, *12*, 444−454.

(11) Puzzarini, C.; Bloino, J.; Tasinato, N.; Barone, V. Accuracy and interpretability: The devil and the holy grail. New routes across old boundaries in computational spectroscopy. *Chem. Rev.* **2019**, *119*, 8131−8191.

(12) Wang, P.; Shu, c.; Ye, H.; Biczysko, M. Structural and energetic properties of amino acids and peptides benchmarked by accurate theoretical and experimental data. *J. Phys. Chem. A* **2021**, *125*, 9826−9837.

(13) Mancini, G.; Fusé, M.; Lazzari, F.; Chandramouli, B.; Barone, V. Unsupervised search of low-lying conformers with spectroscopic accuracy: A two-step algorithm rooted into the island model evolutionary algorithm. *J. Chem. Phys.* **2020**, *153*, 124110.

(14) Ferro-Costas, D.; Mosquera-Lois, I.; Fernandez-Ramos, A. Torsiflex: an automatic generator of torsional conformers. application to the twenty proteinogenic amino acids. *J. Cheminf.* **2021**, *13*, 100.

(15) Barone, V.; Puzzarini, C.; Mancini, G. Integration of theory, simulation, artificial intelligence and virtual reality: a four-pillar approach for reconciling accuracy and interpretability in computational spectroscopy. *Phys. Chem. Chem. Phys.* **2021**, *23*, 17079−17096.

(16) León, I.; Fusé, M.; Alonso, E. R.; Mata, S.; Mancini, G.; Puzzarini, C.; Alonso, J. l.; Barone, V. Unbiased disentanglement of conformational baths with the help of microwave spectroscopy, quantum chemistry, and artificial intelligence: The puzzling case of homocysteine. *J. Chem. Phys.* **2022**, *157*, 074107.

(17) Mancini, G.; Fusé, M.; Lazzari, F.; Barone, V. Fast exploration of potential energy surfaces with a joint venture of quantum chemistry, evolutionary algorithms and unsupervised learning. *Digital Discovery* **2022**, *1*, 790−805.

(18) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB, an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.

(19) Fornaro, T.; Burini, D.; Biczysko, M.; Barone, V. Hydrogen-bonding effects on infrared spectra from anharmonic computations: uracil-water complexes and uracil dimers. *J. Phys. Chem. A* **2015**, *119*, 4224−4236.

(20) Penocchio, E.; Piccardo, M.; Barone, V. Semiexperimental Equilibrium Structures for Building Blocks of Organic and Biological Molecules: The B2PLYP Route. *J. Chem. Theory Comput.* **2015**, *11*, 4689−4707.

(21) Puzzarini, C.; Biczysko, M.; Barone, V.; Largo, L.; Pena, I.; Cabezas, C.; Alonso, J. L. Accurate characterization of the peptide linkage in the gas phase: a joint quantum-chemistry and rotational spectroscopy study of the glycine dipeptide analogue. *J. Phys. Chem. Lett.* **2014**, *5*, 534−540.

(22) Alessandrini, S.; Barone, V.; Puzzarini, C. Extension of the "Cheap" Composite Approach to Noncovalent Interactions: The jun-ChS Scheme. *J. Chem. Theory Comput.* **2020**, *16*, 988−1006.

(23) Lupi, J.; Alessandrini, S.; Barone, V.; Puzzarini, C. junChS and junChS-F12 Models: Parameter-free Efficient yet Accurate Composite Schemes for Energies and Structures of Noncovalent Complexes. *J. Chem. Theory Comput.* **2021**, *17*, 6974−6992.

(24) Gyevi-Nagy, L.; Kallay, M.; Nagy, P. R. Accurate reduced-cost CCSD(T) energies: parallel implementation, benchmarks and large-scale applications. *J. Chem. Theory Comput.* **2021**, *17*, 860−878.

(25) Kallay, M.; Horvath, R. A.; Gyevi-Nagy, L.; Nagy, P. R. Size-consisten explicitly correlated triple excitation correction. *J. Chem. Phys.* **2021**, *155*, 034107.

(26) Nagy, P. R.; Gyevi-Nagy, L.; Lorincz, B. D.; Kallay, M. Pursuing the basis set limit of CCSD(T) non-covalent interaction energies for medium-sized complexes: case study on the S66 compilation. *Mol. Phys.* **2022**, *120*, e2109526.
(27) Papousek, D.; Aliev, M. R. *Molecular vibrational-rotational spectra*; Elsevier Scientific Publishing Company: 1982.
(28) Gaw, F.; Willetts, A.; Handy, N.; Green, W. In *Advances in Molecular Vibrations and Collision Dynamics*; Bowman, J. M., Ed.; JAI Press: 1992; Vol. *1*, pp 186−195.
(29) Clabo, D. A., Jr.; Allen, W. D.; Remington, R. B.; Yamaguchi, Y.; Schaefer, H. F., III A systematic study of molecular vibrational anharmonicity and vibration-rotation interaction by self-consistent-fied higher-derivative methods. Asymmetric top molecules. *Chem. Phys.* **1988**, *123*, 187−239.
(30) Burcl, R.; Carter, S.; Handy, N. C. On the representation of potential energy surfaces of polyatomic molecules in normal coordinates: II. Parameterisation of the force field. *Chem. Phys. Lett.* **2003**, *373*, 357−365.
(31) Barone, V. Anharmonic vibrational properties by a fully automated second order perturbative approach. *J. Chem. Phys.* **2005**, *122*, 014108.
(32) Rosnik, A. M.; Polik, W. F. VPT2+K spectroscopic constants and matrix elements of the transformed vibrational Hamiltonian of a polyatomic molecule with resonances using Van Vleck perturbation theory. *Mol. Phys.* **2014**, *112*, 261−300.
(33) Franke, P. R.; Stanton, J. F.; Douberly, G. E. How to VPT2: Accurate and Intuitive Simulations of CH Stretching Infrared Spectra Using VPT2+ K with Large Effective Hamiltonian Resonance Treatments. *J. Phys. Chem. A* **2021**, *125*, 1301−1324.
(34) Mendolicchio, M.; Bloino, J.; Barone, V. Perturb-then-diagonalize vibrational engine exploiting curvilinear internal coordinates. *J. Chem. Theory Comput* **2022**, *18*, 7603.
(35) Grimme, S. Supramolecular Binding Thermodynamics by Dispersion-Corrected Density Functional Theory. *Chem. - A Eur. J.* **2012**, *18*, 9955−9964.
(36) Li, S.-C.; Lin, Y.-C.; Li, Y.-P. Comparative analysis of uncoupled mode approximations for molecular thermochemistry and kinetics. *J. Chem. Theory Comput* **2022**, *18*, 6866.
(37) Barone, V.; Ceselin, G.; Fusé, M.; Tasinato, N. Accuracy meets interpretability for computational spectroscopy by means of hybrid and double-hybrid functionals. *Front. Chem.* **2020**, *8*, 584203.
(38) Ceselin, G.; Barone, V.; Tasinato, N. Accurate Biomolecular Structures by the Nano-LEGO Approach: Pick the Bricks and Build Your Geometry. *J. Chem. Theory Comput.* **2021**, *17*, 7290−7311.
(39) Sanz, M. E.; Blanco, S.; López, J. C.; Alonso, J. L. Rotational Probes of Six Conformers of Neutral Cysteine. *Angew. Chem., Int. Ed.* **2008**, *47*, 6216−6220.
(40) Lovas, F. J.; Kawashima, Y.; Grabow, J. U.; Suenram, R. D.; Fraser, G. T.; Hirota, E. Microwave Spectra, Hyperfine Structure, and Electric Dipole Moments for Conformers I and II of Glycine. *Astrophys. J.* **1995**, *455*, L201−L204.
(41) Blanco, S.; Lesarri, A.; López, J. C.; Alonso, J. L. The gas-phase structure of alanine. *J. Am. Chem. Soc.* **2004**, *126*, 11675−11683.
(42) Lesarri, E. J.; Cocinero, J. C.; López, J. C.; Alonso, J. L. The Shape of Neutral Valine. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 605−610.
(43) Lesarri, E. J.; Sanchez, R.; Cocinero, C. J.; López, J. C.; Alonso, J. L. Coded Amino Acids in Gas Phase: the Shape of Isoleucine. *J. Am. Chem. Soc.* **2005**, *127*, 12952−12956.
(44) Cocinero, J. C.; Lesarri, E. J.; Grabow, J. U.; López, J. C.; Alonso, J. L. The Shape of Leucine in the Gas Phase. *Chem. Phys. Chem.* **2007**, *8*, 599−604.
(45) Blanco, S.; Sanz, M. E.; López, J. C.; Alonso, J. L. Revealing the multiple structures of serine. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 20183−20188.
(46) Alonso, J. L.; Pérez, C.; Eugenia Sanz, M.; López, J. C.; Blanco, S. Seven conformers of l-threonine in the gas phase: a LA-MB-FTMW study. *Phys. Chem. Chem. Phys.* **2009**, *11*, 617−627.
(47) Cabezas, C.; Varela, M.; Peña, I.; Mata, S.; López, J. C.; Alonso, J. L. The conformational locking of asparagine. *Chem. Commun.* **2012**, *48*, 5934−5936.
(48) Sanz, M. E.; Cabezas, C.; Mata, S.; Alonso, J. L. Rotational spectrum of tryptophan. *J. Chem. Phys.* **2014**, *140*, 204308.
(49) Perez, C.; Mata, S.; Blanco, S.; Lopez, J. C.; Alonso, J. L. Jet-cooled rotational spectrum of laser-ablated phenylalanine. *J. Phys. Chem. A* **2011**, *115*, 9653−9657.
(50) Perez, C.; Mata, S.; Cabezas, C.; Lopez, J. C.; Alonso, J. L. The rotational spectrum of tyrosine. *J. Phys. Chem. A* **2015**, *119*, 3731−3735.
(51) Nguyen, H. V. L.; Kleiner, I. Understanding (coupled) large amplitude motions: the interplay of microwave spectroscopy, spectral modeling, and quantum chemistry. *Phys. Sci. Rev.* **2022**, *7*, 679.
(52) Lazzari, F.; Salvadori, A.; Mancini, G.; Barone, V. Molecular Perception for Visualization and Computation: The Proxima Library. *J. Chem. Inf. Model.* **2020**, *60*, 2668−2672.
(53) Fogel, D. B., Bäck, T., Michalewicz, Z., Eds. *Evolutionary computation*; Institute of Physics Publishing: Bristol, 2000; OCLC: ocm44807816.
(54) Olsson, A.; Sandberg, G.; Dahlblom, O. *Struct. Saf.* **2003**, *25*, 47−68.
(55) Wirsansky, E. *Hands-On Genetic Algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*; Packt Publishing Ltd.: Birmingham, 2020.
(56) Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **1988**, *38*, 3098−3100.
(57) Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007−1023.
(58) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
(59) Santra, G.; Sylvetsky, N.; Martin, J. M. L. Minimally Empirical Double-Hybrid Functionals Trained against the GMTKN55 Database: revDSD-PBEP86-D4, revDOD-PBE-D4, and DOD-SCAN-D4. *J. Phys. Chem. A* **2019**, *123*, 5129−5143.
(60) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations. X. The atoms aluminum through argon revisited. *J. Chem. Phys.* **2001**, *114*, 9244−9253.
(61) Papajak, E.; Zheng, J.; Xu, X.; Leverentz, H. R.; Truhlar, D. G. Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions. *J. Chem. Theory Comput.* **2011**, *7*, 3027−3034.
(62) Biczysko, M.; Panek, P.; Scalmani, G.; Bloino, J.; Barone, V. Harmonic and Anharmonic Vibrational Frequency Calculations with the Double-Hybrid B2PLYP Method: Analytic Second Derivatives and Benchmark Studies. *J. Chem. Theory Comput.* **2010**, *6*, 2115−2125.
(63) Hait, D.; Head-Gordon, M. How accurate is density functional theory at predicting dipole moments? An assessment using a new database of 200 benchmark values. *J. Chem. Theory Comput.* **2018**, *14*, 1969−1981.
(64) Kriz, K.; Novacek, M.; Rezac, J. Non-covalent interactions atlas benchmark data sets 3: repulsive contacts. *J. Chem. Theory Comput.* **2021**, *17*, 1548−1561.
(65) Kang, Y. K.; Park, H. S. Assessment of CCSD(T), MP2, DFT-D, CBS-QB3, and G4(MP2) Methods for Conformational Study of Alanine and Proline Dipeptides. *Chem. Phys. Lett.* **2014**, *600*, 112−117.
(66) Kang, Y. K.; Park, H. S. Exploring Conformational Preferences of Alanine Tetrapeptide by CCSD(T), MP2, and Dispersion-Corrected DFT Methods. *Chem. Phys. Lett.* **2018**, *702*, 69−75.
(67) Ruoff, R. S.; Klots, T. D.; Emilsson, T.; Gutowsky, H. S. Relaxation of conformers and isomers in seeded supersonic jets of inert gases. *J. Chem. Phys.* **1990**, *93*, 3142−3150.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

(68) Shavitt, I.; Bartlett, R. J. *Many-body methods in chemistry and physics: MBPT and coupled-cluster theory*; Cambridge University Press: 2009.
(69) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989**, *157*, 479−483.
(70) Møller, C.; Plesset, M. S. Note on an Approximation Treatment for Many-Electron Systems. *Phys. Rev.* **1934**, *46*, 618−622.
(71) Peterson, K. A.; Dunning, T. H. Accurate correlation consistent basis sets for molecular core-valence correlation effects: The second row atoms Al-Ar, and the first row atoms B-Ne revisited. *J. Chem. Phys.* **2002**, *117*, 10548−10560.
(72) Adler, T. B.; Knizia, G.; Werner, H.-J. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.* **2007**, *127*, 221106.
(73) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD(T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.
(74) Werner, H.-J.; Knizia, G.; Manby, F. R. Explicitly correlated coupled cluster methods with pair-specific geminals. *Mol. Phys.* **2011**, *109*, 407−417.
(75) Peterson, K. A.; Adler, T. B.; Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B-Ne, and Al−Ar. *J. Chem. Phys.* **2008**, *128*, 084102.
(76) Werner, H.-J.; Knowles, P. J.; Manby, F. R.; Black, J. A.; Doll, K.; Heßelmann, A.; Kats, D.; Köhn, A.; Korona, T.; Kreplin, D. A.; Ma, Q.; Miller, T. F.; Mitrushchenkov, A.; Peterson, K. A.; Polyak, I.; Rauhut, G.; Sibaev, M. The Molpro quantum chemistry package. *J. Chem. Phys.* **2020**, *152*, 144107.
(77) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639−9646.
(78) Hill, J. G.; Mazumder, S.; Peterson, K. A. Correlation consistent basis sets for molecular core-valence effects with explicitly correlated wave functions: The atoms B−Ne and Al−Ar. *J. Chem. Phys.* **2010**, *132*, 054108.
(79) Schuurman, M. S.; Allen, W. D.; Schaefer, H. F. The ab initio limit quartic force field of BH3. *J. Comput. Chem.* **2005**, *26*, 1106−1112.
(80) Barone, V.; Lupi, J.; Salta, Z.; Tasinato, N. Development and validation of a parameter-free model chemistry for the computation of reliable reaction rates. *J. Chem. Theory Comput.* **2021**, *17*, 4913−4928.
(81) Dzib, E.; Merino, G. The hindered rotor theory: a review. *WIRES Comp. Mol. Sci.* **2022**, *12*, e1583.
(82) Ayala, P. Y.; Schlegel, H. B. Identification and treatment of internal rotation in normal mode vibrational analysis. *J. Chem. Phys.* **1998**, *108*, 2314−2325.
(83) Luchini, G.; Alegre-Requena, J.; Funes-Ardoiz, I.; Paton, R. GoodVibes: automated thermochemistry for heterogeneous computational chemistry data [version 1; peer review: 2 approved with reservations]. *F1000Research* **2020**, *9*, 291.
(84) Pulay, P.; Meyer, W.; Boggs, J. E. Cubic force constants and equilibrium geometry of methane from Hartree−Fock and correlated wavefunctions. *J. Chem. Phys.* **1978**, *68*, 5077−5085.
(85) Mills, I. M. In *Molecular Spectroscopy: Modern Research*; Rao, K. N., Matthews, C. W., Eds.; Academic Press: 1972; Vol. *1*, pp 115−140.
(86) Puzzarini, C.; Stanton, J. F.; Gauss, J. Quantum-chemical calculation of spectroscopic parameters for rotational spectroscopy. *Int. Rev. Phys. Chem.* **2010**, *29*, 273−367.
(87) Liévin, J.; Demaison, J.; Herman, M.; Fayt, A.; Puzzarini, C. Comparison of the experimental, semi-experimental and ab initio equilibrium structures of acetylene: Influence of relativisitic effects and of the diagonal Born−Oppenheimer corrections. *J. Chem. Phys.* **2011**, *134*, 064119.
(88) Puzzarini, C.; Stanton, J. F. Connections between the accuracy of rotational constants and equilibrium molecular structures. *Phys. Chem. Chem. Phys.* **2023**, *25*, 1421.
(89) Piccardo, M.; Penocchio, E.; Puzzarini, C.; Biczysko, M.; Barone, V. Semi-Experimental Equilibrium Structure Determinations by Employing B3LYP/SNSD Anharmonic Force Fields: Validation and Application to Semirigid Organic Molecules. *J. Phys. Chem. A* **2015**, *119*, 2058−2082.
(90) Puzzarini, C.; Heckert, M.; Gauss, J. The accuracy of rotational constants predicted by high-level quantum-chemical calculations. I. Molecules containing first-row atoms. *J. Chem. Phys.* **2008**, *128*, 194108.
(91) Puzzarini, C.; Barone, V. Extending the Molecular Size in Accurate Quantum-Chemical Calculations: The Equilibrium Structure and Spectroscopic Properties of Uracil. *Phys. Chem. Chem. Phys.* **2011**, *13*, 7189−7197.
(92) Melli, A.; Tonolo, F.; Barone, V.; Puzzarini, C. Extending the Applicability of the Semi-experimental Approach by Means of Template Molecule and Linear Regression Models on Top of DFT Computations. *J. Phys. Chem. A* **2021**, *125*, 9904−9916.
(93) Alonso, E. R.; Fusé, M.; León, I.; Puzzarini, C.; Alonso, J. L.; Barone, V. Exploring the Maze of Cycloserine Conformers in the Gas Phase Guided by Microwave Spectroscopy and Quantum Chemistry. *J. Phys. Chem. A* **2021**, *125*, 2121−2129.
(94) Gordy, W.; Cook, R. L.; Weissberger, A. *Microwave molecular spectra*; Wiley: New York, 1984; Vol. *18*.
(95) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Broth-ers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16*, Revision C.01; Gaussian Inc.: Wallingford, CT, 2016.
(96) Sanz, M. E.; López, J. C.; Alonso, J. L. Six conformers of neutral aspartic acid identified in the gas phase. *Phys. Chem. Chem. Phys.* **2010**, *12*, 3573−3578.
(97) Bazso, G.; Magyarfalvi, G.; Tarczay, G. Tunneling lifetime of the TTc/VIp conformer of glycine in low-temperature matrices. *J. Phys. Chem. A* **2012**, *116*, 10539.
(98) Barone, V.; Biczysko, M.; Bloino, J.; Puzzarini, C. Glycine conformers: a never-ending story? *Phys. Chem. Chem. Phys.* **2013**, *15*, 1358−1363.
(99) Barone, V.; Biczysko, M.; Bloino, J.; Puzzarini, C. Accurate structure, thermochemistry and spectroscopic parameters from CC and CC/DFT schemes: the challenge of the conformational equilibrium in glycine. *Phys. Chem. Chem. Phys.* **2013**, *15*, 10094−10111.
(100) Chandramouli, B.; Del Galdo, S.; Fusé, M.; Barone, V.; Mancini, G. Two-level stochastic search of low-energy conformers for molecular spectroscopy: implementation and validation of MM and QM models. *Phys. Chem. Chem. Phys.* **2019**, *21*, 19921−19934.
(101) Nacsa, A. B.; Czakó, G. Benchmark ab initio proton affinity of glycine. *Phys. Chem. Chem. Phys.* **2021**, *23*, 9663−9671.
(102) Puzzarini, C.; Barone, V. Diving for accurate structures in the ocean of molecular systems with the help of spectroscopy and quantum chemistry. *Acc. Chem. Res.* **2018**, *51*, 548−556.
(103) Linder, R.; Nispel, M.; Haber, T.; Kleinermanns, K. Gas-phase FT-IR-spectra of natural amino acids. *Chem. Phys. Lett.* **2005**, *409*, 260−264.
(104) Linder, R.; Seefeld, K.; Vavra, A.; Kleinermanns, K. Gas-phase infrared spectra of nonaromatic amino acids. *Chem. Phys. Lett.* **2008**, *453*, 1−6.
(105) Balabin, R. M. The identification of the two missing conformers of gas-phase alanine: a jet-cooled Raman spectroscopy study. *Phys. Chem. Chem. Phys.* **2010**, *12*, 5980−5982.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

(106) Jaeger, H. M.; Schaefer, H. F., III; Demaison, J.; Császár, A.; Allen, W. D. Lowest-lying conformers of alanine: pushing theory to ascertain precise energetics and semi experimental R_e structure. *J. Chem. Theory Comput.* **2010**, *6*, 3066−3078.
(107) Császár, A. Conformers of gaseous alanine. *J. Phys. Chem.* **1996**, *100*, 3541−3551.
(108) Meinert, C.; Garcia, A. D.; Topin, J.; Jones, N. C.; Diekmann, M.; Berger, R.; Nahon, L.; Hoffmann, S. V.; Meierhenrich, U. J. Amino acid gas phase circular dichroism and implications for the origin of biomolecular asymmetry. *Nature Commun.* **2022**, *13*, 502.
(109) Csaszar, A.; Allen, W. D.; Schaefer, H. F., III In pursuit of the ab-initio limit for conformational energy prototypes. *J. Chem. Phys.* **1998**, *108*, 9751.
(110) Rasanen, M.; Aspiala, A.; Homanen, L.; Murto, J. IR-induced photorotamerization of 2-aminoethanol in low-temperature matrices. AB initio optimized geometries of conformers. *J. Mol. Struct.* **1983**, *96*, 81−100.
(111) He, K.; Allen, W. D. Conformers of gaseous serine. *J. Chem. Theory Comput.* **2016**, *12*, 3571−3582.
(112) Sheng, M.; Silvestrini, F.; Biczysko, M.; Puzzarini, C. Structural and vibrational properties of amino acids from composite schemes and double-hybrid DFT: hydrogen bonding in serine as a test case. *J. Phys. Chem. A* **2021**, *125*, 9099−9114.
(113) Szidarovszky, T.; Czakó, G.; Császár, A. G. Conformers of gaseous threonine. *Mol. Phys.* **2009**, *107*, 761−775.
(114) Zhang, M.; Lin, Z. Ab initio studies of the conformers and conformational distribution of the gaseous hydroxyamino acid threonine. *THEOCHEM* **2006**, *760*, 159−166.
(115) Wilke, J. J.; Lind, M. C.; Schaefer, H. F. I.; Csaszar, A. G.; Allen, W. D. Conformers of gaseous cysteine. *J. Chem. Theory Comput.* **2009**, *5*, 1511−1523.
(116) Comitani, F.; Rossi, K.; Ceriotti, M.; Sanz, M. E.; Molteni, C. Mapping the conformational free energy of aspartic acid in the gas phase and in aqueous solution. *J. Chem. Phys.* **2017**, *146*, 145102.
(117) Watrous, A. G.; Westbrook, B. R.; Fortenberry, R. F12-TZ-cCR: a methodology for faster and still highly accurate quartic force fields. *J. Phys. Chem. A* **2021**, *125*, 10532−10540.
(118) Xie, F.; Fusé, M.; Hazrah, A. S.; Jaeger, W.; Barone, V.; Xu, Y. Discovering the elusive global minimum in a ternary chiral cluster: rotational spectra of propylene oxide trimer. *Angew. Chem., Int. Ed.* **2020**, *132*, 22613−22616.
(119) León, I.; Alonso, E. R.; Mata, S.; Cabezas, C.; Alonso, J. L. Unveiling the Neutral Forms of Glutamine. *Angew. Chem., Int. Ed. Engl.* **2019**, *58*, 16002−16007.
(120) Boeckx, B.; Maes, G. The conformational behavior and H-bond structure of asparagine: a theoretical and experimental matrix-isolation FT-IR study. *Biophys. Chem.* **2012**, *165*−*166*, 62−73.
(121) Chen, M.; Huang, Z.; Lin, Z. Ab initio studies of gas phase asparagine conformers. *THEOCHEM* **2005**, *719*, 153−158.

https://doi.org/10.1021/acs.jctc.2c01143
J. Chem. Theory Comput. 2023, 19, 1243−1260

Chapter 8

Validation and applications of new features

This chapter aims to validate the Proxima Molecular Perception software and the accompanying definition of the Chemical Feature Space before applying them systematically to relevant chemical problems. The chapter is subdivided as follows: The first section, Chemical Conditions, presents some validation examples of chemical descriptors for the "classification" and "clustering" of atom types and the application of the synthon approach to the TMA scheme. The second section, Physical Conditions, proves the effectiveness of some new potential energy functions for describing bendings and stretchings, finally discussing the couplings between these different energy contributions taking the Ip conformer of the glycine molecule as an example. Finally, the exploration of Potential Energy Surfaces with the help of ML algorithms and Virtual Reality tools is discussed in the last section.

8.1 Chemical Conditions

A continuous description of atom types permits the application of many of the different Machine-Learning techniques already discussed in previous chapters. There are three main paths that we can follow; the first choice is trying to replicate known discrete atom types from continuous atom types, the second is trying to create new sets of atom types that depend on the molecular system investigated and lastly, the third possibility is of keeping the atom type as a continuous quantity while obtaining geometrical parameters from the continuous description of atoms in molecules.

8.1.1 Atom type classification

In order to validate the effectiveness of describing atom types of traditional Force Field from continuous features, we took as a reference the Carbon atom types from the General AMBER Force Field Force Field [111] shown in Tab. 8.1.

Atom Type	Description
c	sp^2 in C=O, C=S
c1	sp
c2	sp^2 aliphatic
c3	sp^3
ca	sp^2 aromatic
cc	inner sp^2 in conjugated ring systems
ce	inner sp^2 C in conjugated chain systems
cp	bridge aromatic
cu	sp^2 in three-membered rings
cv	sp^2 in four-membered rings
cx	sp^2 in three-membered rings
cy	sp^3 in four-membered rings

Table 8.1: The General AMBER Force Field [111] carbon atom types.

A small training set of just 147 atoms and a test set of 50 atoms have been employed. A simple Decision Tree algorithm (see Chap. 3), optimized with the scikit-learn python library [81], reached an accuracy of 94% on this small dataset obtaining the tree of Fig. 8.1.

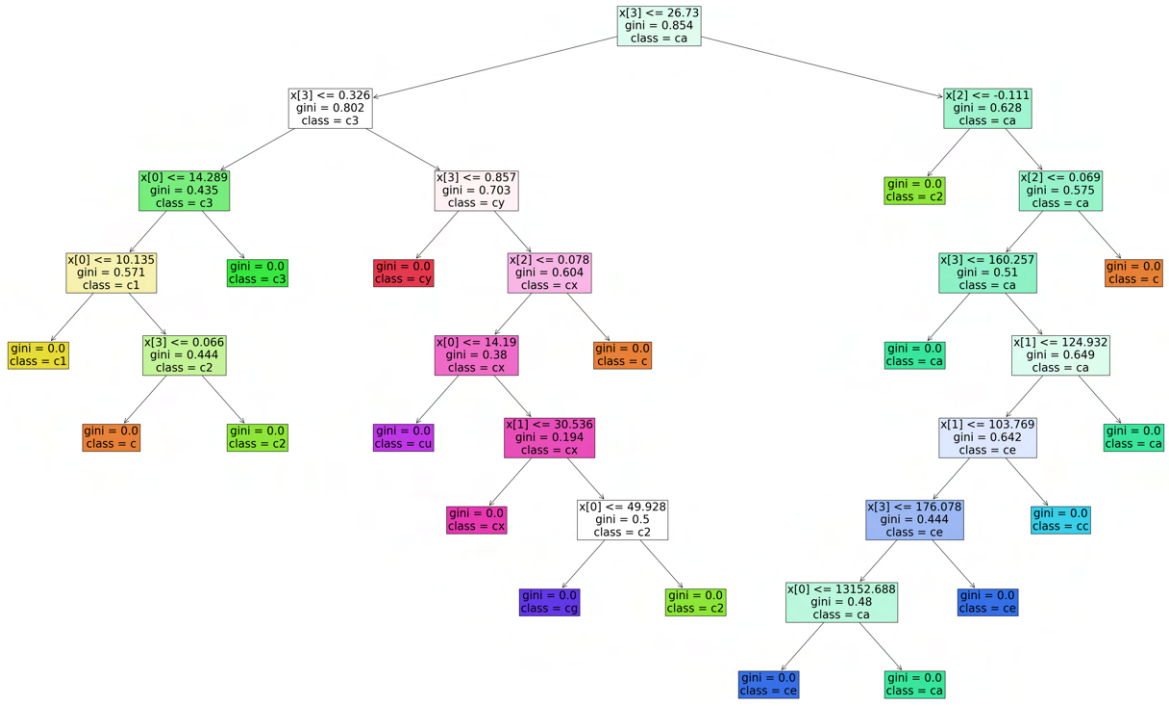


Figure 8.1: The Decision Tree for General AMBER Force Field atom types.

The measured importance in the decision for each feature is:

- Delocalization: 16.29%
- Coordination: 14.88%
- Charge: 19.00%

- Rigidity: 49.83%

See Chap. 3 to get details on the way each importance is measured. The high importance of the rigidity shows the good design of the Feature Space since many atom types of Tab. 8.1 are distinguished mostly on the basis of their presence in small rings and the size of the ring itself. Of course, the data set is really small, and more data points would be necessary to better train the model. Moreover, the Decision Tree is not the most common classification algorithm used because of its fluctuations. However, other algorithms such as the random forest do not provide a simple Decision Tree to be shown graphically and the goal of this section was just to prove the good design of the features in representing traditional atom types.

8.1.2 Discrete dynamic atom types

The second application of the Feature Space is to find a discrete set of "dynamic atom types" that are specific to the molecules considered. To do so, the 4-dimensional Feature Space is computed, normalized so that each feature is centered on its average value with a standard deviation equal to one, and a PCA (Principal Component Analysis, see Chap. 3) is performed to reduce the number of dimensions to 3 since there might be dependencies between features that we try to remove. The Silhouette coefficient is then computed on this new Feature Space for an increasing number of clusters determined with the K-Means algorithm. The optimal number of clusters is then automatically assigned in the first maximum of the Silhouette. By performing clustering operations with the scikit-learn python library [81], we are building a new set of discrete atom types that are specific to the molecule considered. In Fig. 8.2, these dynamic atom types are shown for three different molecules.

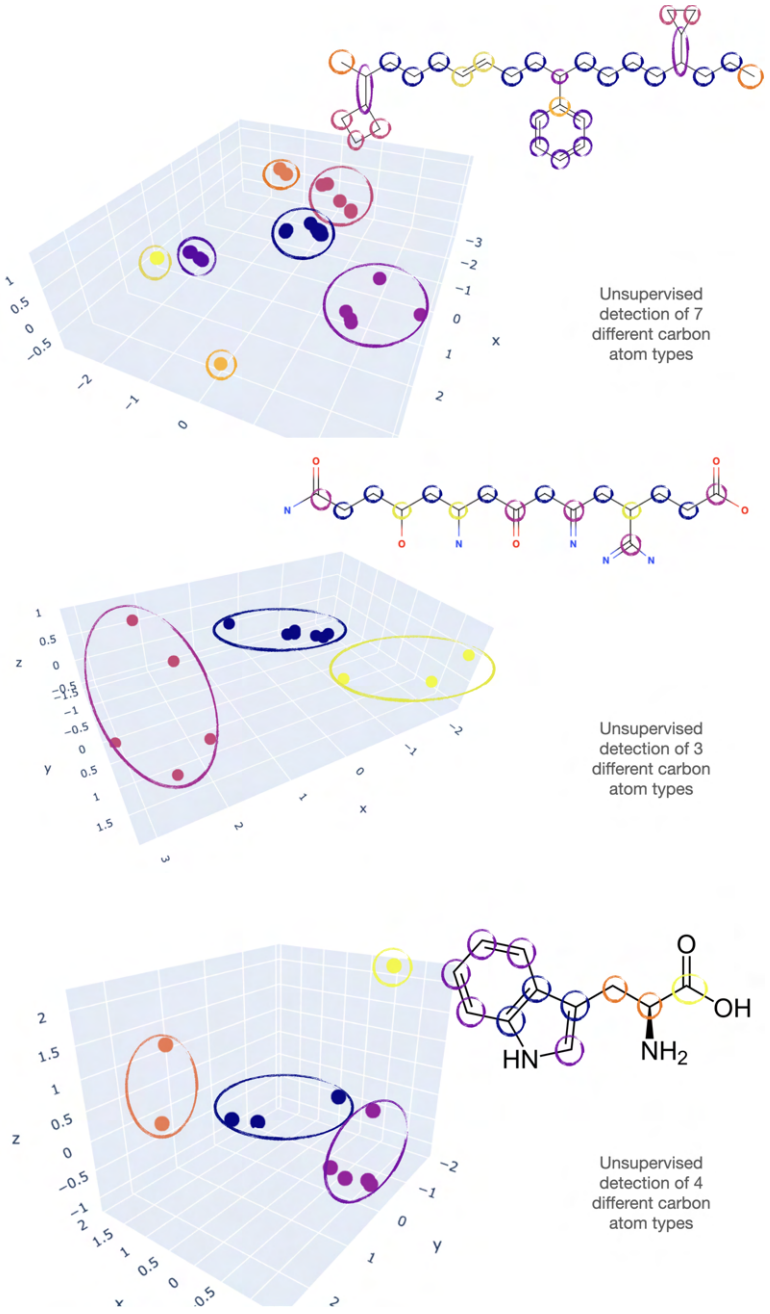


Figure 8.2: The clustering procedure applied to three different molecular systems.

In this case, the bond orders and the charges of Proxima have been used in a molecular graph computed with the topology perception algorithms discussed in the previous chapter. It is important to notice, however, that in principle we could use a bond order matrix derived from Quantum Chemistry computations in computing the same feature and applying the same clustering method, thus allowing us to study the formation of atom types in less obvious situations such as in the presence of metal complexes. Notice how the clustering correctly accounts for the rigidity of the atom in the molecule (e.g. the cyclopropane and cyclobutane rings of the first molecule), for the different delocalization states of the atom (such as the double bond CSP2-CSP2 and the aromatic CSP2 in rings), and for the presence of heteroatoms. It is important to notice that this type of clustering depends on the percentage with which an atom appears in the dataset, areas of

the Feature Space that are more dense tend to form more clusters than sparse areas.

8.1.3 The synthon and the Fragment Databases

The synthon was introduced as the evolution of the continuous atom type to the description of covalent bonds. The last application we are showing is the detection of sub-fragments in molecules. The need for partitioning a molecule into smaller fragments comes from the need to speed up computations on larger molecules. In a recent work [139], a new tool has been developed for the accurate refinement of molecular structures. The Nano-Lego tool, as the name implies, is designed to be a user-friendly tool to assemble fragments of molecules into bigger molecules and to refine geometries by awareness of sub-fragments. These two operations, assembly and refinement, are the Templating Molecules Approach (TMA) and the Linear Regression Analysis approach (LRA) respectively. In the following, we are going to illustrate the LRA approach, its recent evolution in the LPCS approach, the TMA scheme, and the use of synthons in automatically performing TMA.

LRA

The general idea of the Linear Regression Analysis (LRA) is to compute quantities at two different levels of theory (X and Y) performing a linear regression between the two to correct new X values as close as possible to the ones computed with Y. In general, the linear assumption works well when dealing with properties computed with two methods that are similar in accuracy. An example has been the application of LRA to the determination of Semi-Experimental molecular structures. The analysis of high-resolution spectra (with rotational spectroscopy being the technique of reference in the present context) provides the spectroscopic parameters for the vibrationally averaged structure of one or more vibrational states. The direct experimental outcomes are the rotational constants, which are proportional to the inverse of the inertia moments in the Eckart frame [139]. Since they depend on both the coordinates and the masses of the atoms in the molecule, measurements performed for a sufficient number of isotopologues provide the information needed for determining all the averaged geometrical parameters of the corresponding vibrational states. To move to the equilibrium configuration, however, vibrational contributions need to be considered and the rotational constants of the equilibrium geometry have to be employed in the fitting procedure. In the majority of cases, therefore, a pure experimental route is not practicable. In this connection, the so-called Semi-Experimental (SE) approach represents the best method for obtaining accurate equilibrium structures for all but the smallest (two, or three atoms) molecules. To exploit this method for semi-rigid molecules, second-order vibrational perturbation theory (VPT2) [6–8] comes into play, providing explicit expressions of the vibrational corrections to the rotational constants in terms of second and semi-diagonal third-order derivatives of the potential energy with respect to normal modes. The situation is particularly favorable because the sum of the corrections issuing from the different normal modes (contrary to the individual terms) is devoid of any

possible resonance and, thanks to a fortuitous but very general error compensation, can be computed with remarkable accuracy by not too-sophisticated QM approaches (e.g., MP2, hybrid, or double-hybrid density functionals). Additional electronic contributions can also be taken into account, but their role is negligible except in very peculiar cases. In a recent work [140], it has been shown that last-generation hybrid (PW6B95) and double-hybrid (rev-DSD-PBEP86-D3(BJ) [1–3] [1]) functionals in conjunction with partially augmented basis sets (jul-cc-pVDZ and jun-cc-pVTZ [5], respectively) provide improved results for several spectroscopic properties with respect to the B3LYP [27] and B2PLYP [30] models used in the previous compilation. The general approach is to subtract vibrational contributions (obtained by VPT2 applied to semi-diagonal cubic force constants evaluated at the rev-DSD-PBEP86-D3(BJ) [1–3]/jun-cc-pV(T+d)Z/jul-cc-pVDZ level of theory [1, 4, 5]) from the corresponding ground-state rotational constants measured experimentally:

$$B_{\alpha}^{SE} = B_{\alpha}^{0} - \Delta B_{\alpha}^{vib} \tag{8.1}$$

The MSR software [141] has been developed to find the geometrical structure, considering all isotopologues, that best fit such Semi-Experimental rotational constant. Thus, this geometry is called the Semi-Experimental structure. In the LRA approach, the difference between geometrical parameters obtained from Semi-Experimental structures (r_e) and the ones computed through geometry optimization of the X level of theory (r_X) are fitted through a linear equation of the type:

$$r_e = (1 + A)r_X + B \tag{8.2}$$

As discussed at the beginning, the problem with such linear behavior is that it requires the two geometries (in this case X= rev-DSD-PBEP86-D3(BJ) [1–3] and Y=SE) to be in the same range of accuracy. In general, when we use lower-level computations as our X method we observe a loss in linearity and much noise in the relation between geometrical parameters. The bond parameters are shown in Tab. 8.2.

	A	B
CC	-0.00184	0
CH	-0.00239	0
CO	-0.00297	0
CN	-0.00234	0
CS	-0.01222	0.01672
CF	-0.00307	0
CCl	-0.0043	0
NH	-0.00216	0
OH	0.24674	-0.24091

Table 8.2: The LRA parameters for correcting revDSD geometries for the original Nano-Lego tool [139].

The next evolution of the LRA approach might be to employ intrinsically non-linear engines (such as Neural Networks) to correlate geometrical parameters

computed with many different levels of theory, thus providing accurate geometries starting from low-resolution structures.

LPCS

A variant of the LRA approach that we are recently developing is the LPCS (Low-cost Pisa Composite Scheme). The idea is to start from a geometry optimized at the rev-DSD-PBEP86-D3(BJ) [1–3]/cc-pVTZ-F12 (revDSD/3F12 from now on) [9], which is already a high-quality structure obtained employing an F12 basis set. To obtain the LPCS structure, an additional contribution must be taken into account, namely the Core Valence (CV) correlation. To quantify such contribution, the difference between an all-electron and a frozen core MP2 computation can be added requiring three different computations to be performed. However, we have noticed that by employing an LRA-like approach, we can then quantify such CV contribution directly onto the geometrical parameters. The key observation is that the effect of CV correlation on covalent bonds is almost constant and shrinks the bond. In particular, by writing the corrected bond length as:

$$r = (1 - a)r_{revDSD/3F12} - b \tag{8.3}$$

we can then express the CV correlation contribution as:

$$\begin{cases} a = 0 \\ b = 0.0012(n_i + n_j - 2) \end{cases} \tag{8.4}$$

In this equation, units are Angstroms, and n_i, n_j are the principal quantum numbers for the atoms considered (1 for first-row atoms, 2 for second-row atoms, etc.). This corresponds to saying that the effect of CV is a constant shrinking of the bond lengths, with a perfectly linear dependency with a coefficient of 1 between the revDSD /3F12 [9] distances and the LPCS ones. Another alternative formulation might be setting the b parameter to 0 while changing the angular coefficient of the line obtaining:

$$\begin{cases} a = 0.0011\sqrt{n_i n_j - 1} \\ b = 0 \end{cases} \tag{8.5}$$

This should give almost the same contribution to distances with the advantage of allowing us to employ such a formulation, in future works, to correct force constants in a similar fashion. However, in the following, we are going to employ the formulation with the constant b since the latter is still requiring much more testing.

TMA

The TMA strategy is not in contrast to the LRA or LPCS ones but can instead be integrated with these other methods to refine geometries. The TMA can be thought of as a trick to scale the refinement to bigger molecules for which multiple fragments can be processed in parallel. The idea is quite simple and is to detect sub-fragments of a given molecule for which higher resolution structures are already available thus using that information to correct the overall geometry.

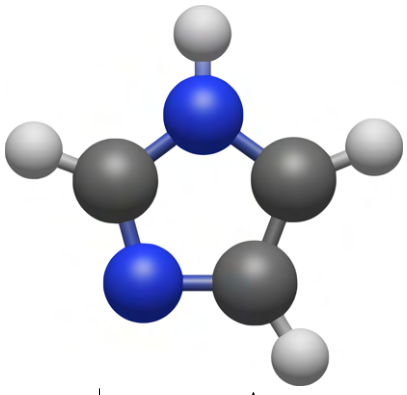
In particular, let us assume that for a given molecule computed at the X level of theory, a given sub-fragment is already available with both the geometry at the X level of theory and the best high-res structure. Then, the TMA model assumes that the geometry variation between X and high-res remains the same independent of whether or not we are talking about the sub-fragment or the overall molecule. Thus, we can correct each geometrical parameter with:

$$r^{e}_{mol} = r^{X}_{mol} + r^{best}_{frag} - r^{X}_{frag} \tag{8.6}$$

It is important to remember that the same assumption of the LRA method is still in place: there is a linear scaling between geometrical parameters computed at the X level of theory and the ones of the best structure available. In the current context, we are going to assume such behavior for geometries computed at the revDSD level but non-linearities will be taken into account in future works. The TMA approach requires:

- A fragment-detection strategy
- A correction strategy of geometrical parameters for which sub-fragments are not available

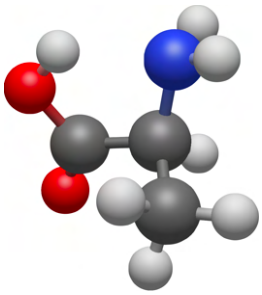
In general, the "best" structure we can have is the Semi-Experimental one. However, SE structures are hard to obtain. Thus, when SE structures are missing from our internal database we can rely on a second database of LPCS geometries, which are easier to obtain. To detect sub-fragments, the synthon approach has been employed for bonds and angles. We observed that a good detection strategy employing synthons consists in checking whether the computed synthon for X-level geometrical parameter (bond or angle) is within 50 units of Euclidean distance from the X-level synthon of the sub-fragment bond in Feature Space units. If so, the sub-fragment synthon is assigned to the molecule synthon and the correction is performed with the equation written above. At the junction between fragments, a correction strategy must be employed. The simplest one is to keep the parameter at the X level of theory, which has been proven good enough for revDSD geometries. Otherwise, a common LRA strategy can still be employed for correcting such geometrical parameters. In the following, several examples of molecules computed with the revDSD/jun-cc-pVTZ [5] method and corrected with several strategies are shown. In Fig. 8.3, the rotational constants for the imidazole molecule are reported.



MHz	A	B	C
EXP a)	9725.32600	9374.01100	4771.92800
Δ_{vib} b)	-79.86500	-77.22700	-40.98200
Δ_{el} c)	0.52400	0.55900	-0.15800
SE (EXP - Δ_{vib} - Δ_{el})	9804.66700	9450.67900	4813.06800
REV	9755.64061	9406.72360	4789.00274
LRA	9801.42460	9447.50460	4810.60550
LPCS	9805.00090	9455.17980	4813.45670
SE/TMA	9805.65773	9452.08673	4812.81323
ERR % (REV)	-0.50003	-0.46510	-0.50000
ERR % (LRA)	-0.03307	-0.03359	-0.05116
ERR % (LPCS)	0.00341	0.04762	0.00808
ERR % (SE/TMA)	0.01010	0.01490	-0.00529

Figure 8.3: The imidazole rotational constants. a) Values are taken from Ref. [142] b) Values computed at the revDSD/jun-cc-pVTZ [2, 3] level of theory c) Values are taken from Ref. [143] and are computed at the B3LYP/aug-cc-pVTZ.

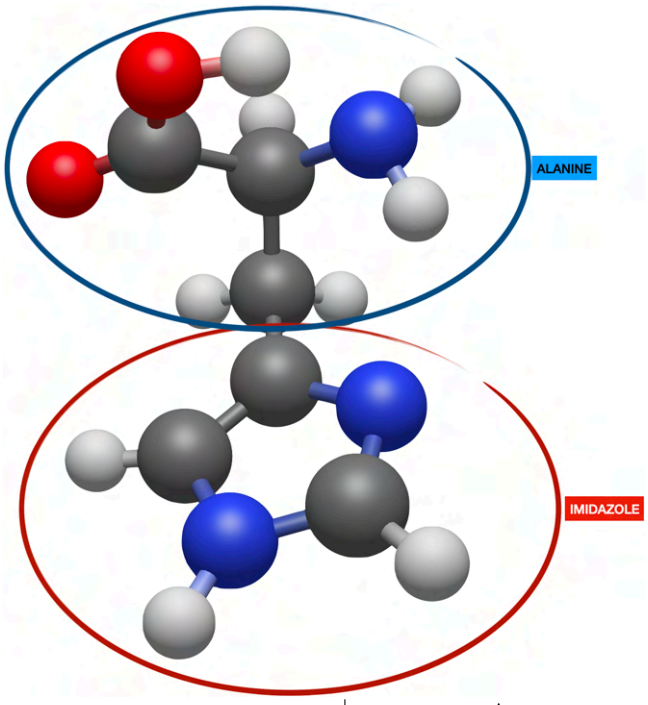
In this case, the Semi-Experimental structure is available from the SE23 [139] database, and the TMA is employed just to validate the use of the synthon in detecting fragments (thus not distinguishing the TMA from the SE structure). As can be seen, the percentage relative error on the rotational constants of the revDSD structure is higher than any other. The LRA improves the accuracy by reducing the error, but the Semi-Experimental structure is, of course, the highest quality structure although the LPCS method comes very close with an order of magnitude improvement over the A rotational constant (although slightly worsening the B and C rotational constants). In this case, the TMA is identical to its SE counterpart since the synthon correctly assigned the entire molecule to the imidazole fragment as a validation. To employ the TMA, fragments are required. In the following, we are going to study four amino acids: tyrosine, phenylalanine, tryptophan, and histidine. To apply the TMA, a fragment for the backbone is required. To start, we are going to use the II conformer of the alanine molecule as shown in Fig. 8.4.



MHz	A	B	C
EXP a)	4973.0558	3228.3379	2307.8090
Δ_{vib} b)	-52.0713	-16.5586	-27.4565
SE (EXP - Δ_{vib})	5025.1271	3244.8965	2335.2655
REV	5005.8975	3205.3552	2334.2973
LPCS	5032.1071	3223.9208	2343.5498
LRA	5030.1472	3219.2802	2344.8370
SE	5026.18306	3246.90140	2335.95969
ERR % (REV)	-0.3827	-1.2186	-0.0415
ERR % (LPCS)	0.1389	-0.6464	0.3548
ERR % (LRA)	0.0999	-0.7894	0.4099
ERR % (SE)	0.02101	0.06179	0.02973

Figure 8.4: The rotational constants for the II conformer of the alanine molecule using different strategies. a) Values are taken from Ref. [144] b) Values are obtained from Ref. [144] computed at the MP2/6-31G(d) [145–152] level of theory.

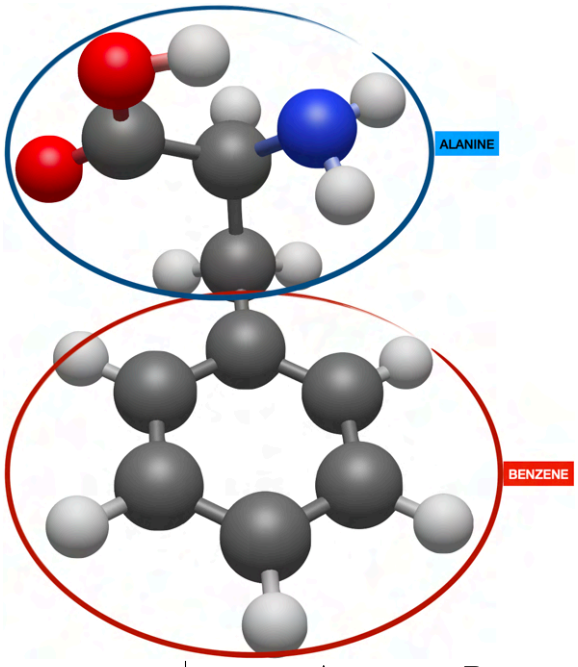
Here, we show the difference in accuracy using the LRA structure and the LPCS structure. In this case, the LPCS is in the same range of accuracy as the LRA which is a common trend observed. Having the imidazole and the alanine fragments, it is possible to validate the application of LRA at the junction between the two. Starting from the parent histidine molecule, shown in Fig. 8.5, we computed rotational constants employing two different strategies at the junction between fragments.



MHz	A	B	C
EXP a)	1847.53472	831.71551	745.94445
Δ_{vib} b)	-14.7198	-2.2784	-2.6382
SE (EXP - Δ_{vib})	1862.2545	833.9939	748.5827
REV	1846.3236	833.9905	748.0919
LRA	1854.7568	837.4040	751.1480
TMA (TMA with REV at junction)	1861.9812	833.2180	748.6539
TMA (TMA with LRA at junction)	1862.1523	834.0210	749.2956
ERR % (REV)	-0.8555	-0.0004	-0.0656
ERR % (LRA)	-0.4026	0.4089	0.3427
ERR % (TMA with REV at junction)	-0.0147	-0.0930	0.0095
ERR % (TMA with LRA at junction)	-0.0055	0.0033	0.0952

Figure 8.5: The rotational constants for the parent histidine molecule using different strategies. a) Values are taken from Ref. [153] b) Values computed using the framework of VPT2 at the B3LYP–D3(BJ)/jul-cc-pVDZ [2, 3, 5, 27, 28] level of theory.

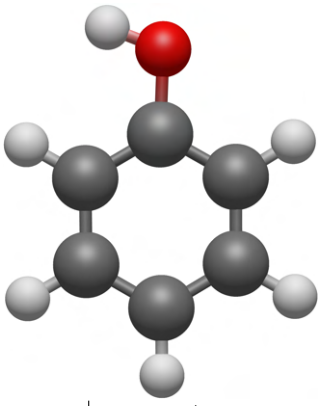
In this case, we are going to omit the LPCS structure since SE structures are available for every fragment involved. The first validation is that the software correctly employs the synthons to detect the fragments to unite: the imidazole and alanine fragments. The TMA approaches outperform the simple LRA strategy if enough fragments are provided as reference SE molecules, by going below 1% relative error in the rotational constant. Employing the LRA correction at the junction between fragments shows a small improvement over the pure-revDSD strategy (that is keeping non-available parameters as the revDSD geometry). Still, the advantage of the pure-revDSD strategy is that there is no need for pre-compute LRA parameters to correct the geometry, thus being a parameter-free model. In future examples, the pure-revDSD strategy is employed unless it is specified otherwise. The success of such a strategy can be further observed with the IIgg conformer of the phenylalanine molecule shown in Fig. 8.6.



MHz	A	B	C
EXP a)	1666.0436	638.5631	568.7684
Δ_{vib} b)	-14.5240	-3.3160	-2.6140
SE (EXP - Δ_{vib})	1680.5676	641.8791	571.3824
REV	1665.6639	642.0987	571.5029
LRA	1672.8395	644.5957	573.6966
TMA	1677.7566	641.3864	571.6095
ERR % (REV)	-0.8868	0.0342	0.0211
ERR % (LRA)	-0.4599	0.4232	0.4050
ERR % (TMA)	-0.1673	-0.0768	0.0397

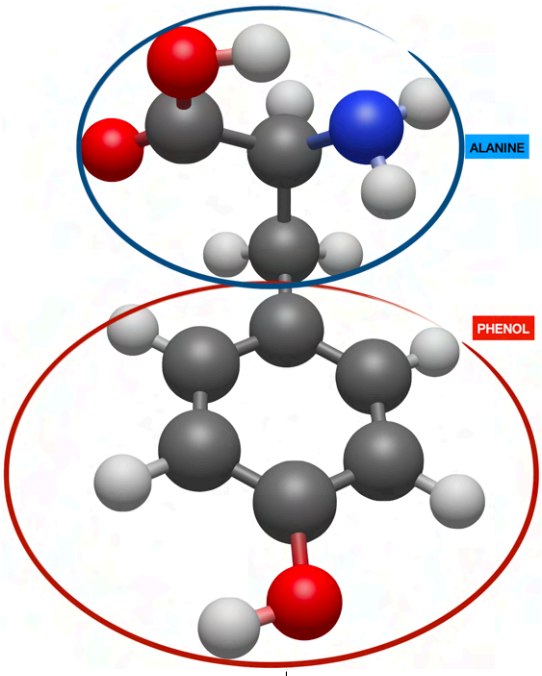
Figure 8.6: The rotational constants for the IIgg conformer of the phenylalanine molecule using different strategies. a) Values are taken from Ref. [154] b) Values computed at the B3LYP -D3(BJ)/jun-cc-pVDZ [2, 3, 27, 28] level of theory.

In this case, the benzene fragment required to perform TMA is taken from the SE23 [139] database. The advantage of our synthon approach is that when a new molecule is provided by the user, the software can automatically detect which portions of the molecule are missing from the internal database thus warning the user to provide new fragments. The user can then decide to provide the best structure he has for the fragment requested (by performing LPCS). In this case, the TMA structure has a rotational constant that is considerably better than the LRA one (from -0.5% error on the A rotational constant computed from LRA geometry, to -0.2% of the TMA geometry). The same improvement can also be seen in the B and C rotational constants. Up to this point, we validated the use of SE fragments in performing TMA. To further validate the use of LPCS geometries, we took the IICgg conformer of tyrosine for which the phenol and alanine-IIg fragments are both available from the SE23 [139] database. We also took the Phenol geometry computed with the LPCS scheme (as shown in Fig. 8.7) and we compared the results of the TMA using both types of fragments as shown in Fig. 8.8.



MHz	A	B	C
EXP a)	5650.46	2619.20	1789.84
Δ_{vib} b)	-41.4	-17.1	-12.1
SE (EXP - Δ_{vib})	5691.86	2636.3	1801.94
SE	5693.33	2637.30	1802.39
LPCS	5699.72	2637.59	1803.16
ERR % (SE)	0.0258	0.0381	0.0249
ERR % (LPCS)	0.1382	0.0488	0.0678

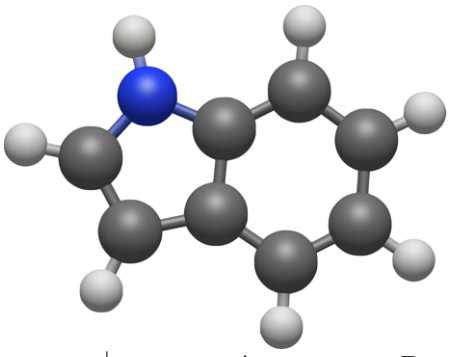
Figure 8.7: The rotational constants for the phenol fragment obtained from LPCS and the SE structure. a) Values are taken from Ref. [155] b) Values computed at the B3LYP -D3(BJ)/jun-cc-pVDZ [2, 3, 27, 28] level of theory.



MHz	A	B	C
EXP a)	1525.25430	465.48173	427.31023
Δ_{vib} b)	-13.74000	-1.86000	-1.55000
SE (EXP - Δ_{vib})	1538.99430	467.34173	428.86023
REV	1524.69038	467.69054	429.28287
TMA from SE fragment	1535.59640	467.55780	429.57850
TMA from LPCS fragment	1535.52620	468.13810	430.08630
ERR % (REV)	-0.92943	0.07464	0.09855
ERR % (TMA from SE fragment)	-0.22079	0.04623	0.16748
ERR % (TMA from LPCS fragment)	-0.22535	0.17040	0.28589

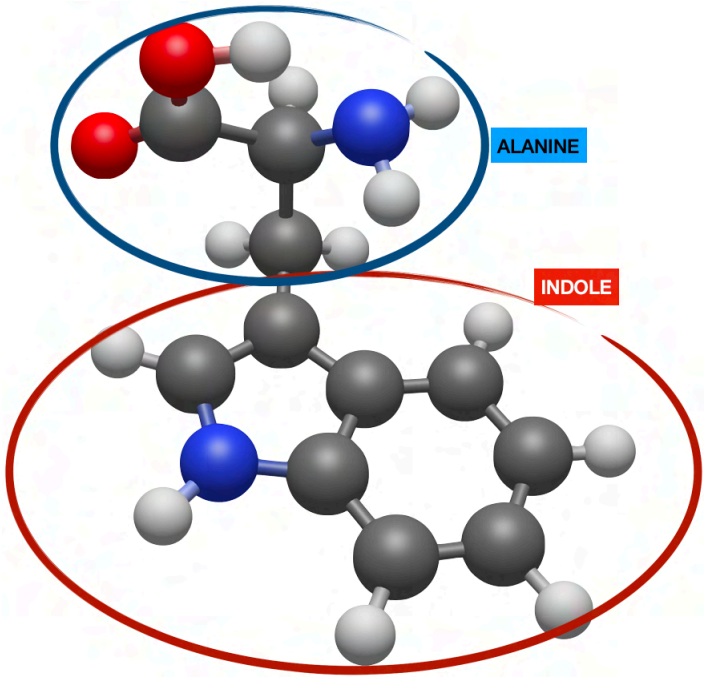
Figure 8.8: The rotational constants for the IICgg tyrosine conformer using different strategies. a) Values are taken from [154] b) Values computed at the B3LYP-D3(BJ)/jun-cc-pVDZ [3, 27, 28] level of theory.

It is apparent that the relative percentage error of the TMA/SE approach is comparable to the TMA/LPCS approach thus allowing us to employ LPCS structures in those cases where the SE fragment structure is missing. The advantage of the LPCS scheme is that it is relatively easy to obtain accurate geometries of molecules without having to rely on the availability of experimental data as in the case of SE structures. The last example is the tryptophane molecule for which the Semi-Experimental structure is not available. In this case, we computed the LPCS molecule starting from the revDSD geometry as shown in Fig. 8.9. Here, the indole LPCS molecule is taken as the fragment for the side chain. In Fig. 8.10, the results are summarized.



MHz	A	B	C
EXP a)	3877.8396	1636.0462	1150.9006
Δ_{vib} b)	-29.9	-10	-7.3
SE (EXP - Δ_{vib})	3907.7396	1646.0462	1158.2006
REV	3890.9041	1639.0569	1153.2474
LPCS	3908.9439	1646.6559	1158.5942
ERR % (REV)	-0.4308	-0.4246	-0.4277
ERR % (LPCS)	0.0308	0.0370	0.0340

Figure 8.9: The rotational constants for the indole molecule using LPCS. a) Values are taken from Ref. [156] b) Values computed at the B3LYP-D3(BJ)/jun-cc-pVDZ [2, 3, 27, 28] level of theory.



MHz	A	B	C
EXP a)	1243.5844	392.4841	346.8847
Δ_{vib} b)	-9.2	-2.1	-1.7
SE (EXP - Δ_{vib})	1252.7844	394.58409	348.58467
REV	1240.9716	394.9716	348.8415
LRA	1246.1897	396.5053	350.1888
TMA	1245.8874	394.4764	348.6075
ERR % (REV)	-0.9429	0.0982	0.0737
ERR % (LRA)	-0.5264	0.4869	0.4602
ERR % (TMA)	-0.5505	-0.0273	0.0065

Figure 8.10: The rotational constants for the IIb+ conformer of the tryptophane molecule using the SE fragment for alanine and the LPCS for indole. a) Values are taken from Ref. [157] b) Values computed at the B2PLYP-D3(BJ)/jul-cc-pVDZ [2, 3, 5, 158] level of theory.

It is worth noticing that the TMA approach outperforms the LRA by drastically reducing the relative percentage error on the B and C rotational constants while being comparable to LRA in the A rotational constant. This proves the TMA approach's efficacy even when non-SE fragments are employed. However, the interesting aspect of the tryptophane molecule is that it confirms a trend that we already saw with the tyrosine molecule. The A rotational constant has a relative percentage error of 0.5% for tryptophane and 0.2% for tyrosine, which are worse than the previous errors obtained for the initial molecules reported (phenylalanine and histidine). A possible explanation for this is the formation of a non-covalent interaction between the amine group of the alanine fragment and the π system of the lateral chains of histidine and tryptophane. In fact, in this TMA implementation torsions are not considered and are kept fixed at their revDSD [5] values (the pure-revDSD strategy). In future works, it will be interesting to study and quantify the effect of correcting torsion angles on the overall rotational constant.

In order to summarize the entire procedure, the TMA workflow relies on two databases (SE23 and LPCS23) and can be summarized as follows:

- Perception of synthons (bonds and angles) on the input geometry
- Detection of fragments from the SE23 database
- In cases SE fragments are missing, detection of LPCS structure from the LPCS23 database (or generation of input files to compute the required LPCS geometries).
- Correction of the input molecular geometry (with a given junction strategy)

At the current stage of development, the correction of the geometry is performed by employing the Generalized Internal Coordinates (GIC) syntax of the Gaussian software [118] (see Chap. 6 for details). Still, the future goal is to work with Cartesian coordinates for the geometry correction step directly. To get a better insight into the synthon detection mechanism for the TMA, which automatically detects sub-fragments in molecules, we plotted the 2-component PCA representation of the Feature Space in Fig. 8.11 for the C-C bonds of the fragments considered in the previous computations (alanine IIg, imidazole, indole, benzene, and phenol).

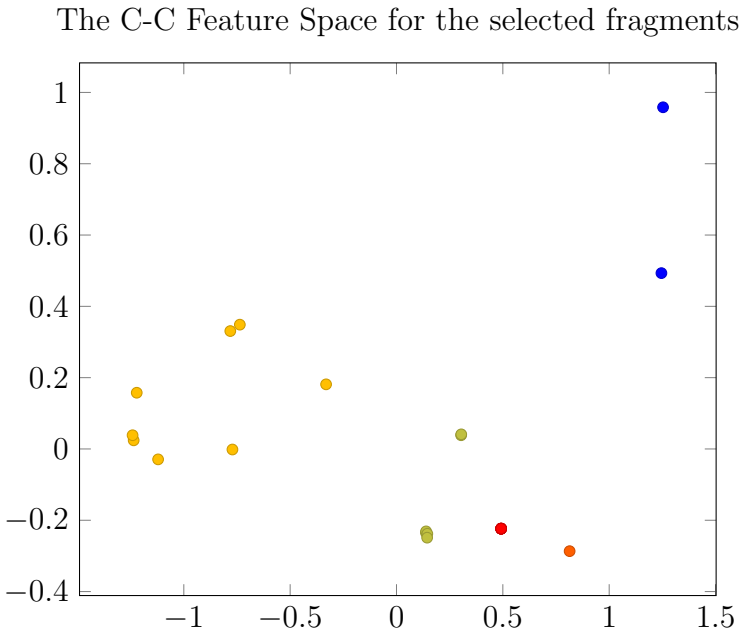


Figure 8.11: The Feature Space of the C-C synthons for the fragments considered. Each point is the 2-component PCA representation of the relative synthon which is a bond in the dataset of fragments employed. Blue: alanine-IIg conformer, Green: phenol, Light Orange: indole, Dark Orange: imidazole, Red: benzene.

Each color identifies a different molecule, while each point is an individual bond (the synthon of the bond) projected in PCA space. Each time the software has to search for a synthon representing a bond in the dataset, it checks the Euclidean distances between the point to detect and each point in the synthon

Feature Space. The assignment is performed with the closest one whether the Euclidean distance is below 50 units (in the non-normalized full Feature Space). In the case of the C-C synthons, the benzene has 6 different C-C bonds that perfectly overlap in Feature Space because of symmetry (the red points in Fig. 8.11). The imidazole only contains a single C-C bond thus giving rise to a single point in the plot (the dark orange one). The phenol has two bonds that are significantly different from the other four; the ones directly connected to the carbon atom bonded to the OH group. This can be seen in Feature Space, by looking at the green points of phenol and observing that two of them are fairly distant from the others (the two ones higher in the PCA plot). The alanine molecule contains two chemically different C-C bonds: the first is the one with the residue, while the second is with the carbon atom of the carboxyl group. The indole is a complex fragment, and each bond is extremely different from all the others so in the PCA representation is hard to get a qualitative interpretation of the way points are scattered. It is interesting to notice, however, how there are two couples of points similar to each other in the light orange portion of the Feature Space. The first couple on the left portion of the plot is the almost-symmetric bonds of the hexagonal ring (the one on top and the one on the bottom of the hexagonal ring), while a second couple of fairly similar bonds is placed in the top region of the indole Feature Space and represents two C-C bonds that are connected to the bridge bond with the pentagonal portion of the molecule. In Fig. 8.12, we plotted the Feature Space for the C-N synthons instead.

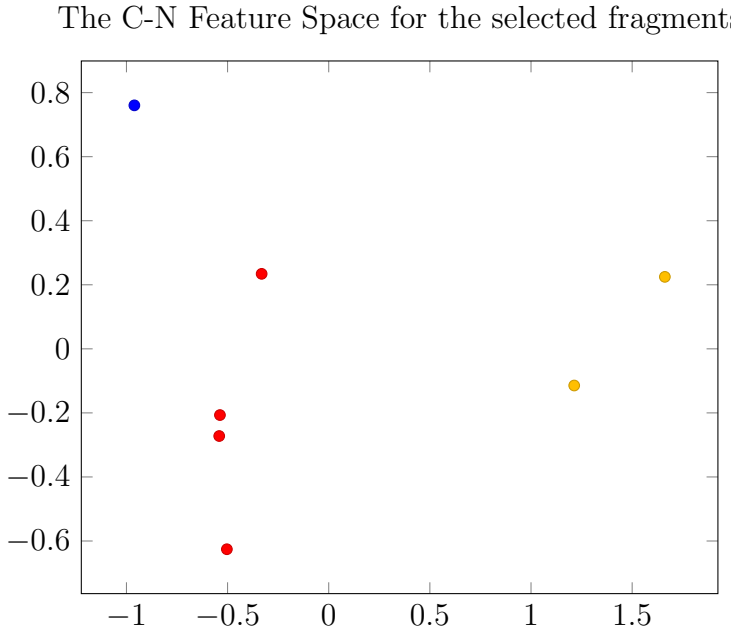


Figure 8.12: The Feature Space of the C-N synthons for the fragments considered. Each point is the 2-component PCA representation of the relative synthon which is a bond in the dataset of fragments employed. Blue: alanine-IIg conformer, Red: imidazole, Orange: indole.

In this case, the benzene and phenol fragments are not appearing since they don't contain C-N bonds. The alanine molecule contains a single C-N bond involving the amine group which is the blue point in the plot. The imidazole

molecule, instead, contains 4 different C-N bonds each one of which is different from the others because of the complete lack of symmetry in the molecule (due to the presence of an -(NH)- atom on one side, and a -N- atom on the other side of the planar ring). Moving from the top to the bottom of the plot, the first red dot is the C-N bond on the top right of the molecule shown in Fig. 8.3. Then, there is the one on the top left of the molecule, we then encounter the central bottom C-C(H)-N-C bond, and lastly, there is the N-C(H)-N-C bond on the bottom left. The indole contributes with only two bonds which are in the right portion of the Feature Space.

In this first section, we just showed three basic examples of the application of synthons and Feature Space to relevant chemical problems. The potential such Feature Space unlocks is huge since it allows an easy-to-interpret way to apply a vast majority of ML algorithms in solving chemical problems. In the case of TMA structures, we already underlined how the assumption of linearity is at the foundation of the current strategy and is reasonable for high-quality structures such as the revDSD ones. In future developments of the TMA, our goal will be to test the non-linear behavior of the geometrical parameters computed at lower levels of theory (e.g. XTB [159], B3LYP [27], etc.) with the SE/LPCS parameters and to use intrinsically non-linear engines (e.g. Neural Networks (NN)) to perform the regression. The Feature Space is a key ingredient in using NN since it provides information that the network can use in determining the non-linear behavior of the geometrical parameters and that we can interpret to understand what patterns arise in such regressions, in a similar fashion to how we interpreted the rigidity relevance of the atoms in the Decision Tree example at the beginning of the section.

8.2 Physical Conditions

If atom types are based on chemical intuition and heuristics, physics describes the energy of molecular systems. In this section, we will validate the shape of the single/double well symmetric/asymmetric potentials introduced in Chap. 6, and we will discuss the mixing of QM and MM energy contributions in the case of the Ip conformer of the glycine molecule. These validations should act as a foundation on top of which we will be able to build, in the future, our workflow to automatically parametrize Force Fields for generic molecules around energy minima (fixed topology) as described at the end of this section.

8.2.1 Van der Waals

To start, we are going to discuss the treatment of non-covalent interactions in particular van der Waals interactions. The reference expression used is the Amber [45] formulation of the 6-12 Lennard-Jones (LJ) potential:

$$LJ_{i,j}(r) = \epsilon_{i,j}\left(\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6}\right) \tag{8.7}$$

In our context, this function behaves like an asymmetric single well potential. To derive the correct set of parameters, we need to impose the condition that the

second derivative of our potential should be equal to the second derivative of the Lennard-Jones potential at equilibrium distance r_m:

$$\left(\frac{\partial^2 U(r)}{\partial r^2}\right)_{r=r_m} = \left(\frac{\partial^2 LJ^{6-12}}{\partial r^2}\right)_{r=r_m} \tag{8.8}$$

The second derivative of the asymmetric single well potential in $r = r_m$ (for $n = 2$ as discussed in Chap. 6) is equal to:

$$\left(\frac{\partial^2 U(r)}{\partial r^2}\right)_{r=r_m} = \alpha^2 \frac{\epsilon}{2r_m^2} - \frac{8\epsilon}{r_m^2} \tag{8.9}$$

While the second derivative of the Lennard-Jones potential in $r = r_m$ is equal to:

$$\left(\frac{\partial^2 LJ^{6-12}}{\partial r^2}\right)_{r=r_m} = \frac{72\epsilon}{r_m^2} \tag{8.10}$$

Thus we get:

$$\alpha = \sqrt{160} \tag{8.11}$$

Which is interestingly independent of the couple of atoms considered. In Fig. 8.13, three examples show the superposition of the two energy curves the Lennard-Jones 6-12 and the modified morse. It is interesting to notice that at short distances, the modified Morse potential is softer than its Lennard-Jones counterpart, whereas there is a very good agreement in the neighborhood of the r_m equilibrium distance.

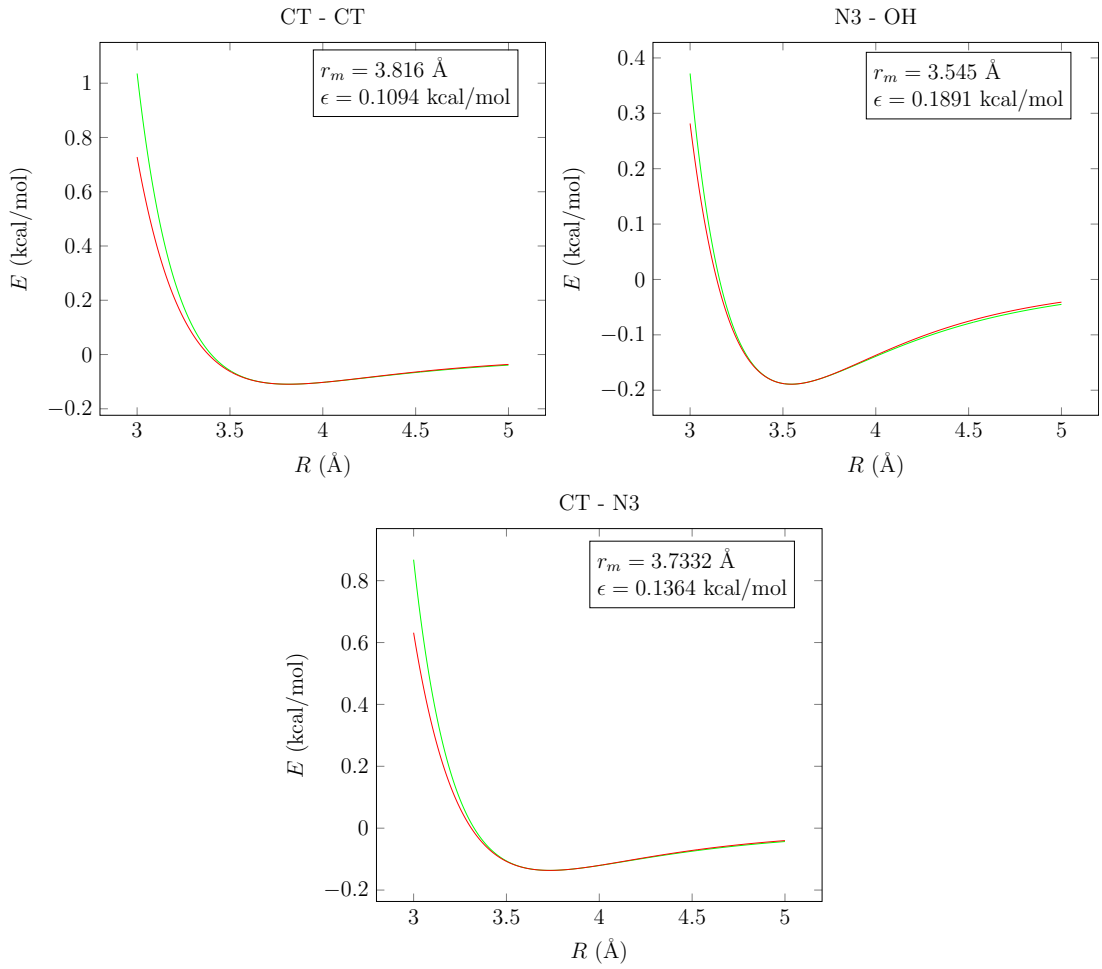


Figure 8.13: The difference between the Lennard-jones 6-12 potential (green line) and the asymmetric single well potential (red line) is given for three different couples of AMBER [45] atom types: CT - CT, N3 - OH, CT - N3. Parameters were taken from the parm99 file of AmberTools [160].

8.2.2 The Hydrogen Bond

In the case of the Hydrogen Bond interaction, we take the 10-12 formulation of the energy as a reference:

$$E_{hbond} = \frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \tag{8.12}$$

This formulation is the one usually employed within Force Fields [46] and is another example of an asymmetric single well potential. The reference distance for the well and the depth of the well itself can be computed as follows:

$$\begin{cases} r_m = \sqrt{\dfrac{12C}{10D}} \\ \epsilon = [(\dfrac{10}{12})^6 - (\dfrac{10}{12})^5]\dfrac{D^6}{C^5} \end{cases} \tag{8.13}$$

The second derivative of such an equation around the distance r_m is the following:

$$\left(\frac{\partial^2 E}{\partial r^2}\right)_{r=r_m} = \frac{240 \cdot 10^6}{12^7}\frac{D^7}{C^6} \tag{8.14}$$

By imposing the equivalence with the second-order derivative of our potential, we get the α parameter desired as:

$$\alpha = \sqrt{16 + \frac{480 \cdot 10^6}{12^7}\frac{r_m^2}{\epsilon}\frac{D^7}{C^6}} \tag{8.15}$$

In Fig. 8.14, three examples of hydrogen bond energy profiles are shown using parameters from parm91X data of AmberTools [160]. These curves are extremely similar proving the effectiveness of our model.

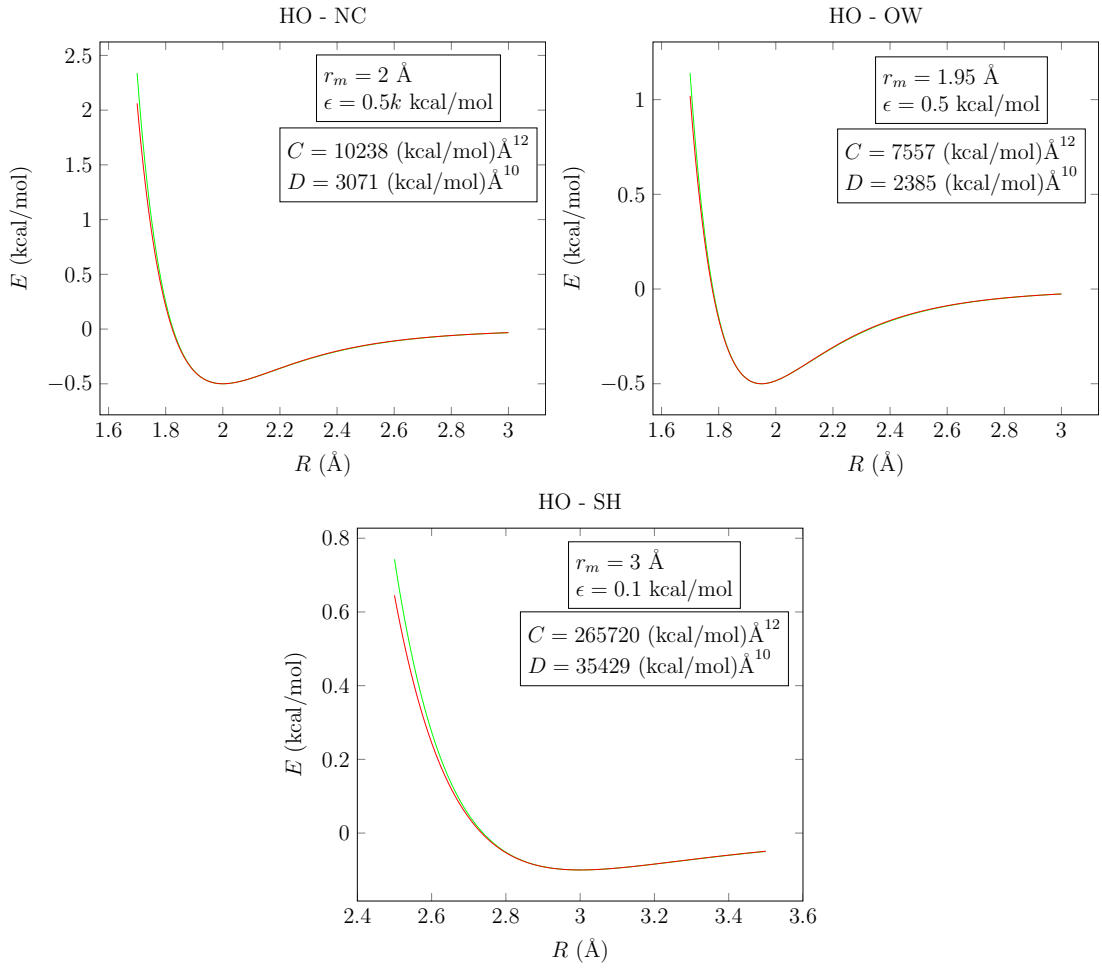


Figure 8.14: The difference between the 10-12 potential (green line) and the asymmetric single well potential (red line) for three different couples of AMBER [45] atom types: HO - NC, HO - OW, HO - SH. Parameters were taken from the Parm91X file of AmberTools [160].

8.2.3 The Stretching

In a recent contribution [94], Morse bonding potentials have been extensively parameterized for the atom types in the MM3 Force Field using high-level CCSD(T) (F12*) energies. To show the flexibility of the asymmetric single well potential,

we decided to use these dissociation energies as the ϵ energy and the second-order derivative of the energy as the harmonic force constant from the AMBER [45] Force Field (parm99). In the case,of $n = 2$ and $m = 0$ (see Chap. 6), we obtain the following condition for the α parameter:

$$\alpha = \sqrt{\frac{2kr_m^2}{\epsilon}} \tag{8.16}$$

In Fig. 8.15, three examples are shown where the harmonic AMBER potential is shown together with our potential.

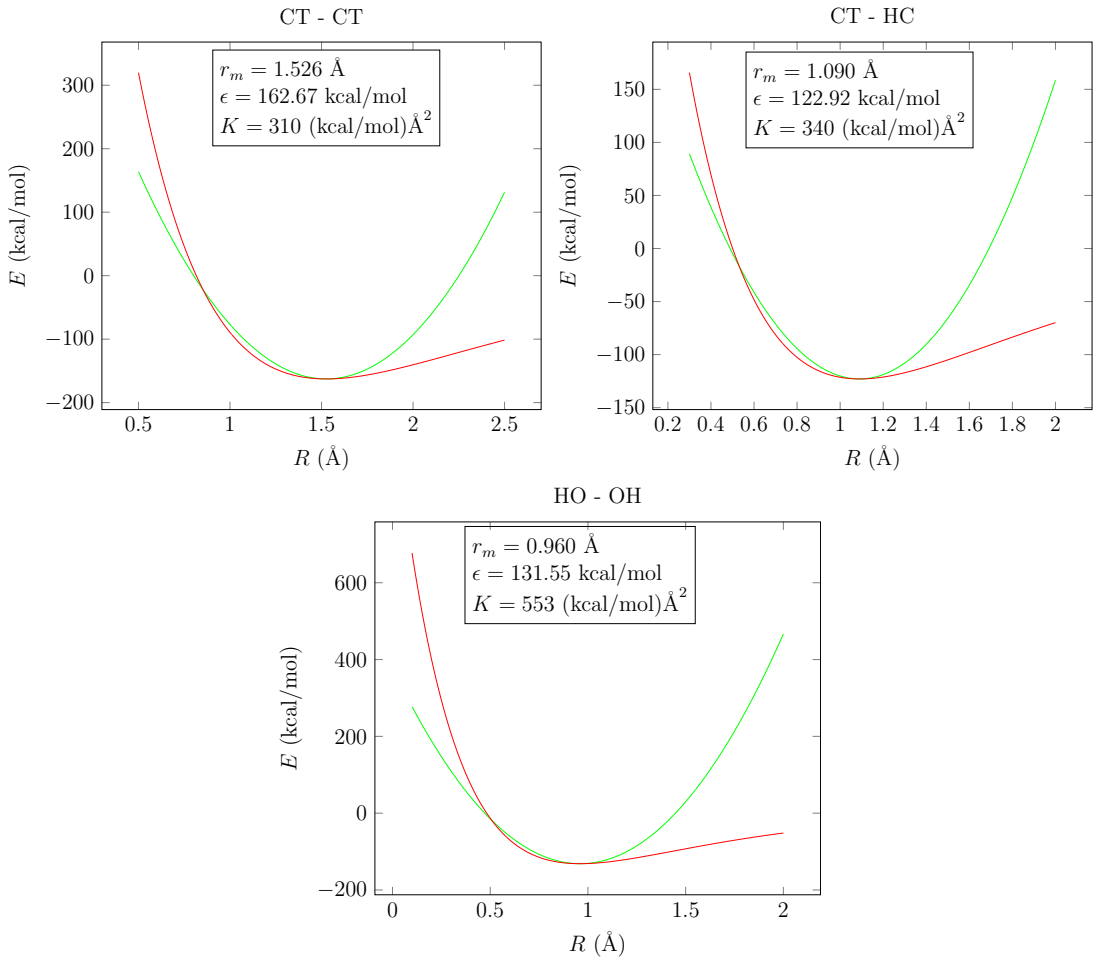


Figure 8.15: The difference between the harmonic AMBER potential (green line) and the asymmetric single well potential (red line) is given for different couples of AMBER atom types: CT - CT, CT - HC, and HO - OH. Second derivatives were taken from the Parm91X file of AmberTools [160], and energies from MM3 work [94].

8.2.4 The Bending

In the case of Bending interactions, we performed some scans at the B3LYP/aug-cc-pVDZ level of theory. The first case we are going to treat is the symmetric single well bending around 180° of the HCN molecule shown in Fig. 8.16.

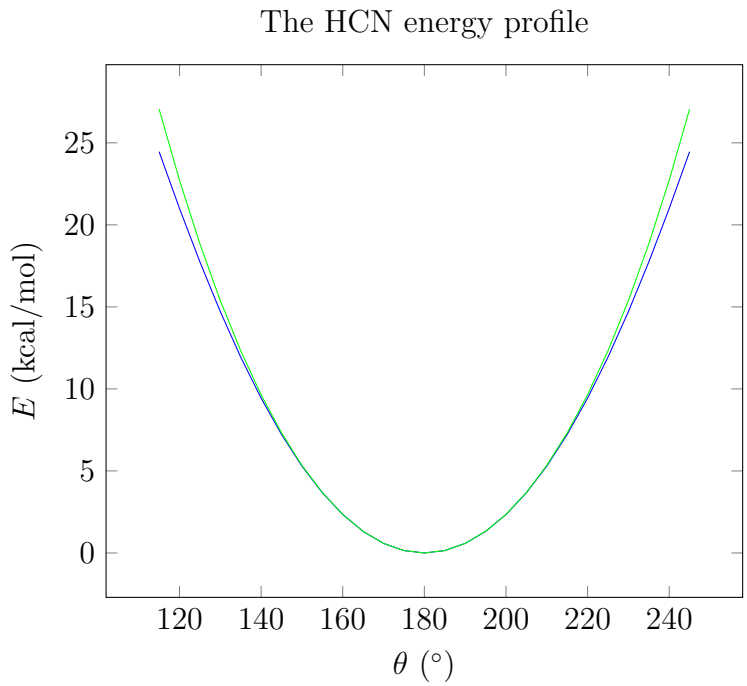


Figure 8.16: The bending profile for the H-C-N angle. The blue line represents the B3LYP/aug-cc-pVDZ scan and the green line is the Gaussian symmetric potential. Energies are kcal/mol and angles are radians.

Here, the analytical expression employed is the Gaussian:

$$E_k = \epsilon \left(1 - e^{\alpha(x_k - x_m)^2}\right) \tag{8.17}$$

and, by imposing the equivalence of the second derivative between the two profiles, we obtained the following values for the α and ϵ parameters:

$$\begin{cases} \alpha = 0.148 \text{ [1/rad}^2\text{]} \\ \epsilon = -128.734 \text{ [kcal/mol]} \end{cases} \tag{8.18}$$

It is interesting to notice how the pure Gaussian profile tends to rise more than the QM profile moving to the extremes, but interpreting the interaction as "pure bending", at that point, is quite hard since there is an overlap between the atoms at the extreme (H and N). The behavior around the minimum θ_e is instead well replicated. Another example of a symmetric bending is the double well profile of the H-O-H angle with an equilibrium angle of 104.5°in the water molecule as shown in Fig. 8.17.

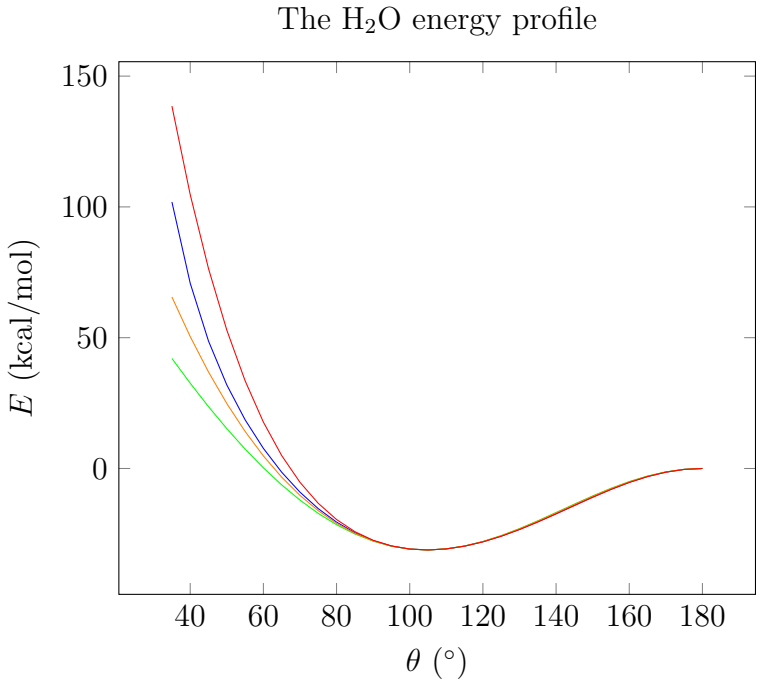


Figure 8.17: The bending profile for the H-O-N angle. The blue line represents the B3LYP/aug-cc-pVDZ scan and the green line is the summation of a Gaussian with a power of x^{2n} with $n = 1$, the orange line is with $n = 2$, and the red line is with $n = 3$. Energies are kcal/mol and angles are in radians

In this case, we had to solve the equation in n as described in Chap. 6 when talking about double well symmetric potentials.

$$E_k = \epsilon \left(e^{\alpha\left(1-\frac{x_k}{x_m}\right)} - \left[\left(\frac{x_k}{x_m}\right)^{2n} - 2\left(\frac{x_k}{x_m}\right)^{n} + 3 \right] e^{\frac{\alpha}{2}\left(1-\frac{x_k}{x_m}\right)} \right) \tag{8.19}$$

We solved it for two different n powers; the harmonic-like and a fourth-order power of x. As shown in the picture, the asymptotic behavior outside of the well region gets increasingly better represented as n increases. The parameters obtained for $n = 1$ are the following:

$$\begin{cases} \alpha = 0.609 \text{ [1/rad}^2\text{]} \\ b = 109.97 \text{ [kcal/mol]} \\ a = 23.39 \text{ [kcal/(mol} \cdot \text{rad}^2\text{)]} \end{cases} \tag{8.20}$$

In the case of $n = 2$, instead, the parameters are the following:

$$\begin{cases} \alpha = 0.878 \text{ [1/rad}^2\text{]} \\ b = 50.73 \text{ [kcal/mol]} \\ a = 2.83 \text{ [kcal/(mol} \cdot \text{rad}^4\text{)]} \end{cases} \tag{8.21}$$

In the case of a symmetric double well potential, such as the inversion of ammonia where the angle considered is in common with the three hydrogens and involves the lone pair directed towards the central axis passing between the three

hydrogens, as shown in Fig. 8.18, the expression to use is the sum of a polynomial and a Gaussian.

$$E_k = ax_k^{2n} + be^{-\alpha x_k^2} - b \tag{8.22}$$

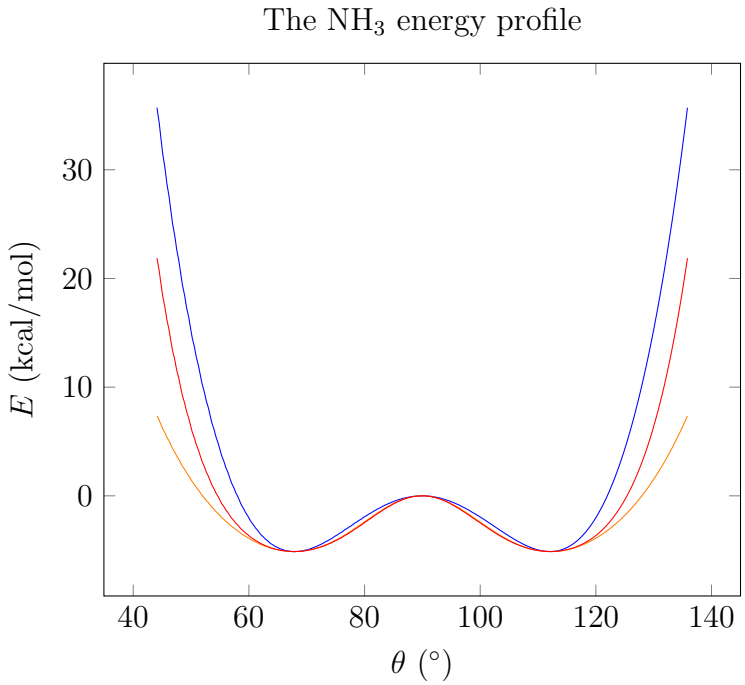


Figure 8.18: The bending profile for the LP-N-H angle. The blue line represents the B3LYP/aug-cc-pVDZ scan, the orange line is the summation of a Gaussian with a power of x^{2n} with $n = 2$ and the red one with$n = 3$.

In this case, the parameters were obtained directly for the$n = 2$ potential form:

$$\begin{cases} \alpha = 15.31 \text{ [1/rad}^2\text{]} \\ b = 6.55 \text{ [kcal/mol]} \\ a = 33.95 \text{ [kcal/(mol} \cdot \text{rad}^4\text{)]} \end{cases} \tag{8.23}$$

For $n = 3$ instead:

$$\begin{cases} \alpha = 18.09 \text{ [1/rad}^2\text{]} \\ b = 5.87 \text{ [kcal/mol]} \\ a = 105.78 \text{ [kcal/(mol} \cdot \text{rad}^6\text{)]} \end{cases} \tag{8.24}$$

Again, the region between the two wells involves a barrier and gets represented correctly by our potential. In this case, a transcendental equation must be solved to find the correct profile. The resolution of the transcendental equation is the limiting step in finding such energy profiles but can be easily automated through Python scripts that search for numerical solutions of such equation (the fsolve function of the scipy python package [82]).

8.2.5 Large amplitude motions

In the following, the two molecules of Fig. 8.19 are studied.

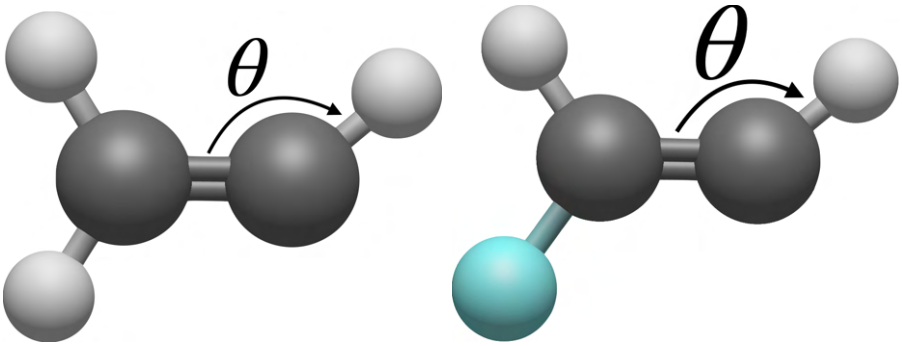


Figure 8.19: The C-C-H angle.

These two radicals have a double well potential when describing the proton moving from one side of the molecule to the other, exchanging the proton with the unpaired electron (in an sp^2-like geometry). The angle considered is the C-C-H angle in both cases, the presence of the fluorine atom (the blue one) in the second case creates a small asymmetry in the energy profile allowing us to show both the symmetric and asymmetric double well curves. In Fig. 8.20, the energy profile for the $CH_2CH^\bullet$ molecule is shown to have a symmetric double well profile.

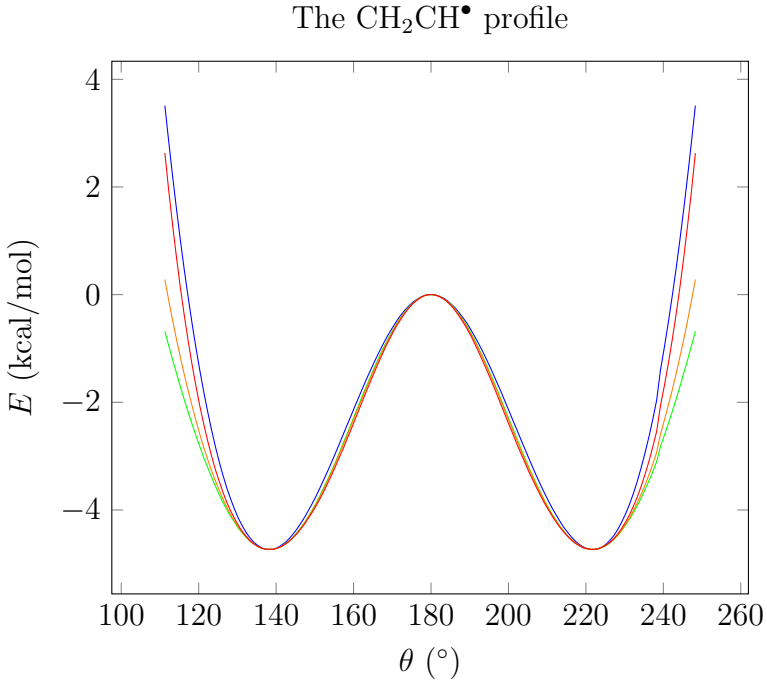


Figure 8.20: The energy profile for the $CH_2CH^\bullet$ molecule with a variation of the C-C-H angle. The blue line represents the B3LYP/aug-cc-pVDZ scan and the green line is the summation of a Gaussian with a power of x^{2n} with $n = 1$. The orange line uses $n = 2$, and the red line is $n = 3$.

In this case, the well is located around 138.37° with a depth of -4.735 kcal/mol. The parameters describing the energy profiles are the following, for$n = 1$:

$$\begin{cases} \alpha = 2.755 \text{ [1/rad}^2\text{]} \\ b = 11.03 \text{ [kcal/mol]} \\ a = 7.05 \text{ [kcal/(mol} \cdot \text{rad}^2\text{)]} \end{cases} \tag{8.25}$$

for$n = 2$:

$$\begin{cases} \alpha = 3.617 \text{ [1/rad}^2\text{]} \\ b = 6.65 \text{ [kcal/mol]} \\ a = 3.33 \text{ [kcal/(mol} \cdot \text{rad}^4\text{)]} \end{cases} \tag{8.26}$$

And for$n = 3$:

$$\begin{cases} \alpha = 4.495 \text{ [1/rad}^2\text{]} \\ b = 5.67 \text{ [kcal/mol]} \\ a = 2.78 \text{ [kcal/(mol} \cdot \text{rad}^6\text{)]} \end{cases} \tag{8.27}$$

The situation is different when asymmetry is introduced through the Fluorine atom, as shown in Fig. 8.21.

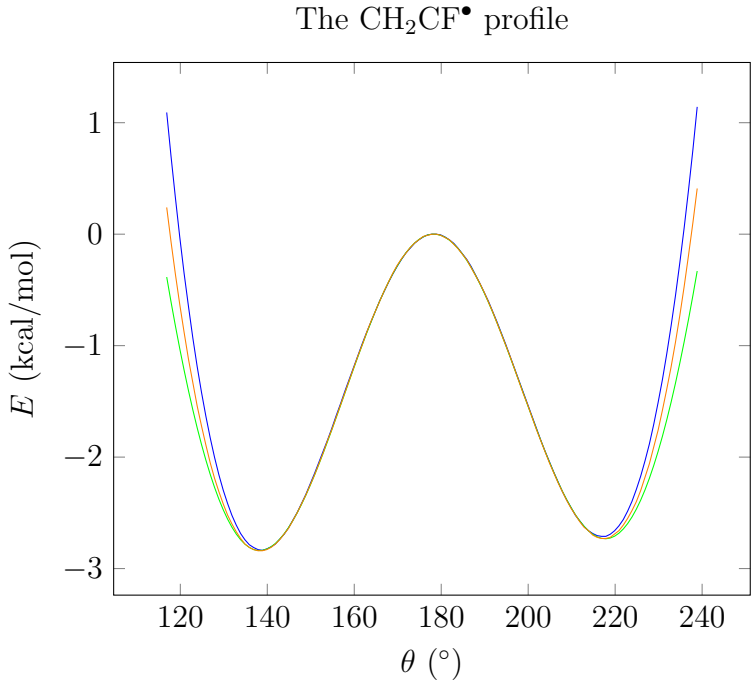


Figure 8.21: The energy profile for the $CH_2CF^\bullet$ molecule with a variation of the C-C-H angle. The blue line represents the B3LYP/aug-cc-pVDZ scan and the green line is the summation of a Gaussian with a power of x^{2n} with $n = 1$, the orange line with$n = 2$ instead.

In this case, we have a slight asymmetry with two wells located at 138.87° and around 217° respectively. The first well has a depth of -2.836 kcal/mol and the second of -2.711 kcal/mol. The barrier between the two wells is not located at 180° anymore due to the asymmetry but is shifted towards the left at around 178°. It is a small asymmetry but enough to force us to find the energy profile through a differential evolution algorithm (implemented in the scipy python package [82]).

The Gaussian and the parabola (here we solved just for$n = 1$) will be shifted one concerning the other and will have different centers.

$$E_k = a(x_k - S)^{2n} + be^{-\alpha(x_k - C)^2} \tag{8.28}$$

Thus, getting the following five parameters after optimization for$n = 1$:

$$\begin{cases} \alpha = 2.016 \text{ [1/rad}^2\text{]} \\ b = 10.994 \text{ [kcal/mol]} \\ C = 178.11^\circ/3.109 \text{ rad} \\ a = 8.396 \text{ [kcal/(mol} \cdot \text{rad}^2\text{)]} \\ S = 177.85^\circ/3.104 \text{ rad} \end{cases} \tag{8.29}$$

Here, S is the center of the parabola and C is the center of the Gaussian and it can be easily seen that they are both shifted towards the first well concerning 180°. In the case of$n = 2$, the following parameters instead:

$$\begin{cases} \alpha = 2.654 \text{ [1/rad}^2\text{]} \\ b = 0.00825 \text{ [kcal/mol]} \\ C = 178.27^\circ/3.111 rad \\ a = 0.0065 \text{ [kcal/(mol} \cdot \text{rad}^4\text{)]} \\ S = 177.644^\circ/3.100 \text{ rad} \end{cases} \tag{8.30}$$

8.2.6 The mixing of QM and Perception

Until now, we performed individual B3LYP scans along the coordinates of interest to obtain the desired energy profiles. However, as we already emphasize at the end of Chap. 6, the energy is a function of all internal coordinates. The general assumption is that non-covalent energy terms can be decoupled from the covalent ones (stretching, bending, torsions, etc.). However, the mixing of non-covalent energy terms derived from Molecular Perception with energy terms derived from QM computations must be validated. The final application would be a tool able to automatically parametrize custom Force Fields for molecules, whose overall workflow has the following logic:

- Geometry Assembly. Building a tool to automatically generate initial geometries and structures in an immersive virtual environment starting from scratch or assembling fragments (TMA). More information is provided in the next section when discussing the Virtual Laboratory.
- synthon detection. Automatically detecting, using ML, the relevant "atom types" and parameters required to describe the energy of the molecules given as input.
- Molecular Perception. Computing Electrostatic and van der Waals non-covalent interactions based on the previous synthon detection stage.
- QM computation. Performing the minimum amount of QM computations needed to obtain the energy profiles for covalent energy terms (for the computed set of atom types), removing non-covalent contributions from the QM energy.

- MM computation. Perform MD/MM simulations on the molecule by combining the energy profiles computed at the previous step together with the non-covalent perceived interactions.

This workflow is still a work in progress, but the main ingredients are already developed and validated. In this context, we already tested the atom and bond Feature Space, developed perception algorithms to compute charges and non-covalent interactions, and derived conditions and equations to get energy profiles depending on the energy minima and their derivatives (for de-coupled variables). In order to prove our workflow, a discussion of the mixing energy terms deriving from Molecular Perception (the non-covalent interactions) and the energy profiles obtained from QM computations is required. Moreover, it is of relevance to discuss the presence of couplings between different coordinates and the eventual error introduced when dealing with only a diagonal Force Field. To do so, an example system is going to be taken into account which is the Ip conformer of the glycine molecule (Fig. 8.22) [161, 162].

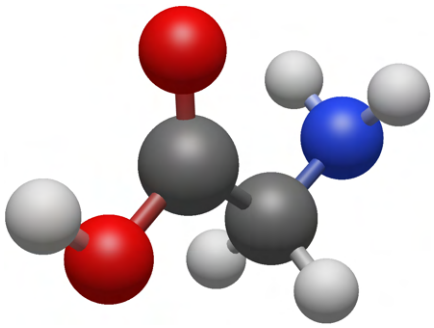


Figure 8.22: The Ip conformer of glycine.

In Fig. 8.23, the plot of the normalized $\bar{\boldsymbol{F}}$ matrix computed at the rev-DSD-PBEP86-D3(BJ) [1–3]/jun-cc-pVTZ [5] level of theory is shown in grayscale with respect to the 36 redundant coordinates of the molecule.

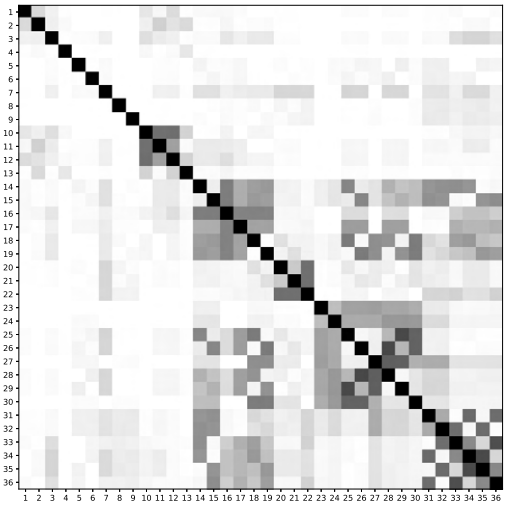


Figure 8.23: The normalized QM $\bar{\boldsymbol{F}}$ matrix plot in grayscale. The $\bar{\boldsymbol{F}}$ is computed with respect to the 36 redundant coordinates.

The normalized $\bar{\boldsymbol{F}}$ is defined from $\boldsymbol{F}$ (see Chap. 6 for details on the computation of such matrix) as:

$$\bar{F}_{ij} = \frac{F_{ij}}{\sqrt{F_{ii}F_{jj}}} \tag{8.31}$$

From now on we are going to assume that the $\boldsymbol{F}$ matrix is normalized. This plot allows for simple detection of the critical feature of the relation between different energy profiles. In particular, the first 9 entries in the matrix are the stretching energy terms reasonably decoupled from the remaining coordinates being nearly diagonal. To quantify the effect of the non-covalent components of the energy, we computed the cartesian Hessian $H_x^{NonBond}$ as described in Chap. 6. The energy contributions considered are the electrostatic energy computed using the charges perceived by the Proxima software, and the van der Waals amber contribution. We then subtracted such Hessian from the total QM Hessian H_x^{QM} in cartesian coordinates, and we converted the resulting hybrid Hessian in internal coordinates through the application of the $\boldsymbol{B}$ matrix as described in Eq. 6.54 in Chap. 6. Then, by diagonalization of the $\boldsymbol{GF}$ matrix, as described in Chap.6, we obtained the frequencies of vibration by taking the square roots of the resulting eigenvalues ($\omega_{s_i} = \sqrt{\Lambda_{ii}}$). In Fig. 8.24, the frequencies of vibration are shown for each normal mode both the one resulting from the QM $\boldsymbol{F}^{QM}$ matrix and the hybrid $\boldsymbol{F}^{hybr}$ matrix.

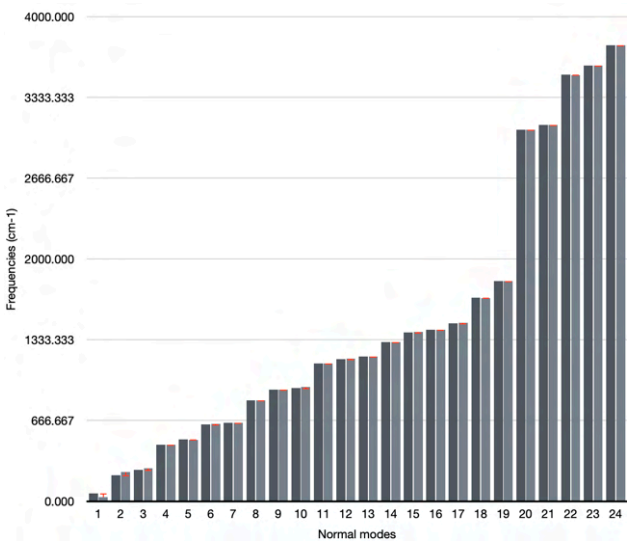


Figure 8.24: The distribution in normal modes of the QM harmonic frequencies (cm^{-1}) (dark gray) and the hybrid = QM - NonBond harmonic frequencies (light gray). The red error bar represents the difference between the QM and hybrid values. These frequencies are computed using the full $\bar{\boldsymbol{F}}$ matrix.

It can be observed that the removal of non-covalent interactions from the QM Hessian does not alter the frequencies of vibration. The biggest variation is for the lowest frequencies which are associated with large amplitude motions, the first one having the highest deviation from the QM value of 34 cm^{-1}. However, the average difference between the QM and the hybrid frequencies is -0.51 cm^{-1}, which is extremely reasonable. It is interesting to notice how these differences become minimal (the red error bar in the plot) as we reach the stretching region

at around 3000 cm^{-1} (the normal modes at the end of the plot). The other interesting aspect to study is whether reducing the number of terms considered in the $\bar{\boldsymbol{F}}$ matrix without significantly impacting the harmonic frequencies is possible. To do so, we tried two different strategies:

- Completely removing non-diagonal terms from the normalized $\bar{\boldsymbol{F}}$ matrix remaining with only the diagonal terms.
- Only removing those non-diagonal terms in the normalized $\bar{\boldsymbol{F}}$ matrix whose values are below 2% of the maximum.

The effect of removing all non-diagonal terms can be observed in Fig. 8.25. Here, the situation is drastically different since these new frequencies (the ones

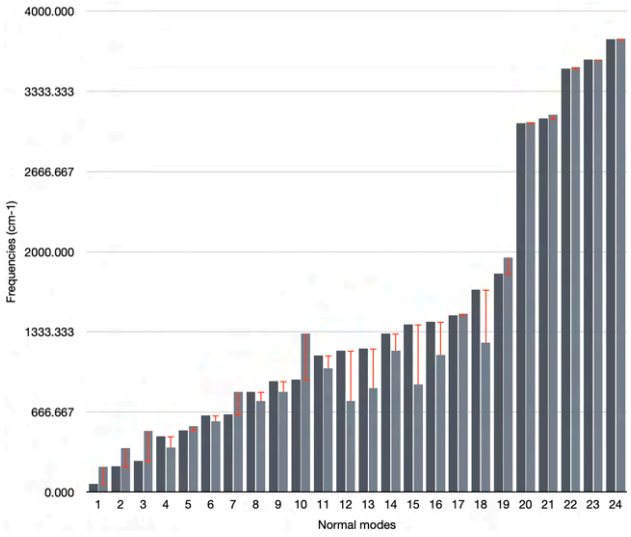


Figure 8.25: The distribution in normal modes of the QM harmonic frequencies (cm^{-1}) (dark gray) and the QM-MM harmonic frequencies (light gray). The red error bar represents the difference between the QM and QM-Hybrid values. These frequencies are computed using only the diagonal elements of the normalized $\bar{\boldsymbol{F}}$ matrix.

computed with such hybrid normalized $\bar{\boldsymbol{F}}$ matrix) differ substantially from the pure QM ones. Here the average difference is 48.383 cm^{-1} which is not acceptable. In the high-frequency region of the spectrum, we can observe that stretching vibrations are the least affected by such deviations. This is compatible with what we already observed with the plot of the normalized $\bar{\boldsymbol{F}}$ matrix in Fig. 8.23 where stretching terms are almost fully diagonal. It is clear that to find a balance between accuracy and a reduced number of terms to consider a threshold-based strategy must be employed when evaluating how many terms of the normalized $\bar{\boldsymbol{F}}$ to retain. In this context, we found that a general threshold of around 2% is good enough as shown in Fig.8.26.

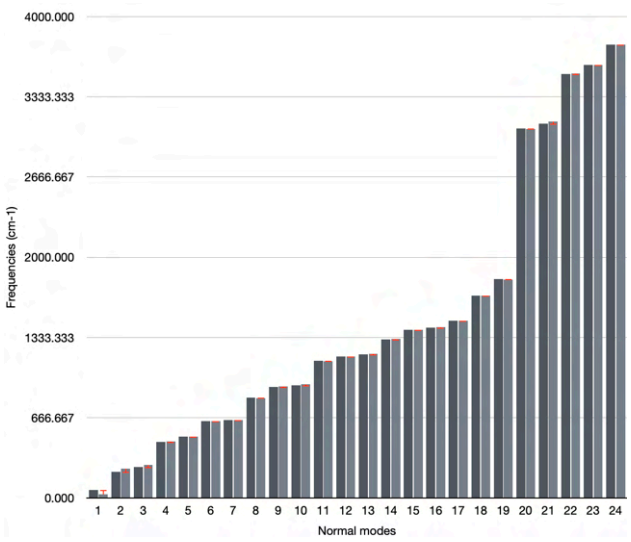


Figure 8.26: The distribution in normal modes of the QM harmonic frequencies (cm^{-1}) (dark gray) and the QM-MM harmonic frequencies (light gray). The red error bar represents the difference between the QM and QM-MM values. These frequencies are computed using the diagonal elements of the $\boldsymbol{F}$ matrix while keeping the off-diagonal terms that are below 2%.

The total number of internal coordinates is 36, thus giving rise to a total of $36 \cdot \frac{35}{2} = 630$ couplings (the off-diagonal terms of the $\bar{\boldsymbol{F}}$ matrix). By removing those terms whose value is below 0.02 of the normalized $\bar{\boldsymbol{F}}$ (the 2% threshold) we removed a total of 273 terms which is almost half the total number of couplings. This is a significant reduction in the number of terms to consider, and it does not substantially affect the frequencies as shown in Fig. 8.26. Here the average difference between the QM frequencies and such Hybrid (2%) frequencies is about -1.174 cm^{-1}. This is slightly bigger than the average difference obtained with the pure QM $\bar{\boldsymbol{F}}$ matrix but much better than the simple diagonal approach of removing every coupling between normal modes. These 273 couplings removed are also shown in Fig. 8.27.

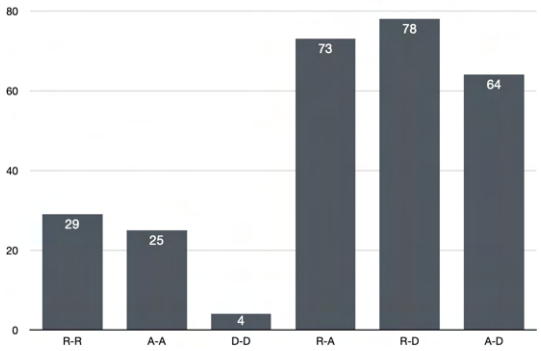


Figure 8.27: The 273 coupling terms removed, these are separated in couplings between stretching (R), bending (A), and dihedral terms (D).

Here, it is possible to observe how the vast majority of couplings involve stretching with torsions and angles. It is interesting to notice how dihedrals are fairly decoupled between themselves and the same applies to stretching and

angles. The threshold for cutting non-diagonal terms in the $\bar{\boldsymbol{F}}$ matrix has been taken as a rule of thumb but to get a better sense of how such threshold impacts the number of couplings removed we performed a scan from the 2% threshold up to a 20% threshold showing the results in Fig. 8.28.

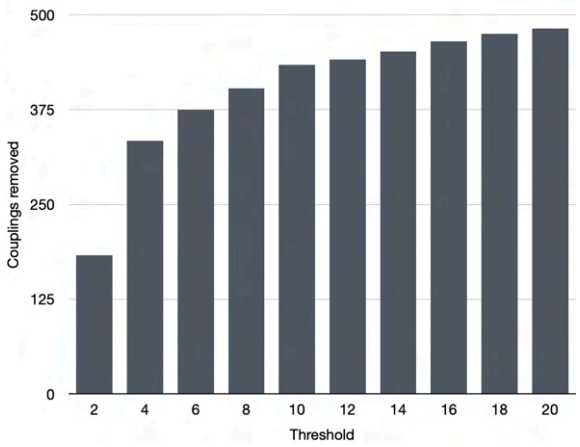


Figure 8.28: The number of couplings removed depending on the cutoff threshold (%).

Here it is possible to observe that the number of couplings increases, although the biggest effect is seen after the aforementioned 2% threshold. In addition to the number of couplings removed, in Fig. 8.29 we show the RMSD between the frequencies computed with the cutoff couplings and the original ones depending on the threshold value.

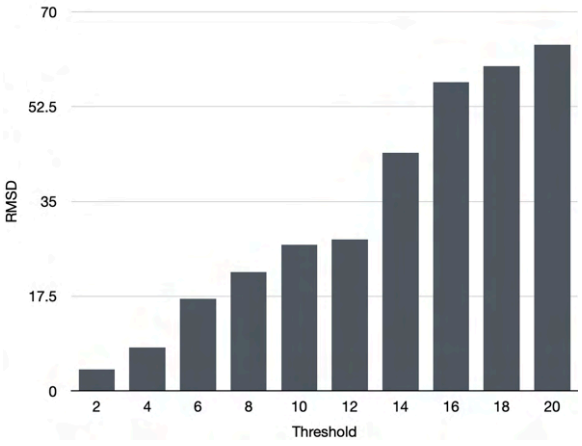


Figure 8.29: The RMSD of the frequencies in cm^{-1} depending on the cutoff threshold (%).

Here, it is possible to observe that after a threshold of 12% the RMSD drastically increases. The combination of the two plots allows us to validate the use of a 2% threshold in the removal of the coupling interaction terms, obtaining the new normalized $\bar{\boldsymbol{F}}$ shown in Fig. 8.30.

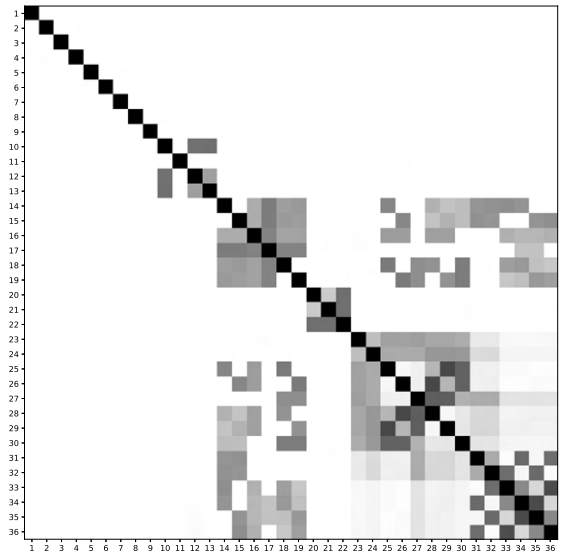


Figure 8.30: The normalized QM $\bar{\boldsymbol{F}}$ matrix plot in grayscale removing the couplings below 20%. The $\bar{\boldsymbol{F}}$ is computed with respect to the 36 redundant coordinates.

This test case on the Ip conformer of glycine helps us understand the conditions through which we can decouple different energy terms to obtain individual energy profiles for each contribution. In the case of diagonal contributions to the energy (stretching and bending), we already proved the efficacy of our analytical expressions for single/double symmetric/asymmetric wells. The next step in the development of such workflow is to derive flexible analytical expressions for the description of torsions and the needed coupling terms (the ones above 2%) to describe the total energy of the system. The goal will be to build an automatic tool that automatically performs all these steps.

8.3 Exploration

Having combined chemical intuition, with the development of a good Feature Space, together with physical calculations, and having defined a workflow to sample energy profiles and compute energies quickly and efficiently, we now have the tools to explore the chemical space of multiple species. This final dynamic exploration step can either be driven by human intuition, through the development of proper Graphical User Interfaces or Virtual Reality experiences, or by software by means of dedicated Machine-Learning algorithms. In this chapter we already discussed the structure of our pipeline: (perception → chemical descriptors → Feature Space → physical computations → dynamic exploration). Here, the dynamic exploration step is further analyzed by showing some use of Genetic Algorithms for the problem of the search of conformers, and some Virtual Reality applications for the editing of structures and PES exploration.

8.3.1 Conformer Search

The exploration of the conformational space for molecules of medium size is a tricky problem that can be tackled by means of both empirical research grounded

in chemical intuition, both automatic tools that employ some algorithm to explore in the most efficient way such space (e.g. CONFAB [163], CONFLEX [164], CREST [165], etc.). In the development of automatic tools that perform the exploration, the idea was to employ Genetic Algorithms to help the search for stable conformers. More details on the theory behind Genetic Algorithms are shown in Chap. 3. The software that drives the exploration was developed with user experience in mind, building a Graphical User Interface (GUI) to help the user prepare the inputs (Fig. 8.31).

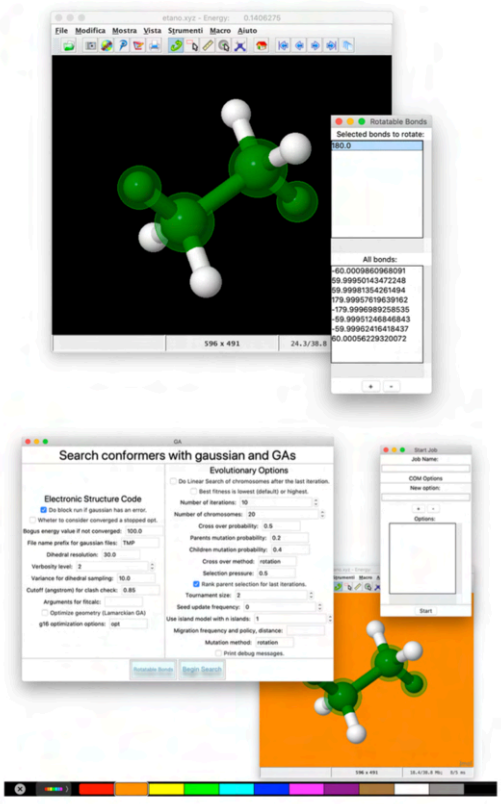


Figure 8.31: The Graphical User Interface (GUI) for the conformer search software.

In this context, a specimen in a GA is a molecular structure whose genes are the set of coordinates being used in the search and the alleles are the specific values of those coordinates, which identify a structure in the PES together with its fitness (here its Semi-Empirical or DFT energy). Hence, crossover implies mixing the coordinates of two-parent structures to generate new ones, while mutation changes the value of one coordinate moving the structure to a new region of the PES. The best specimens are those with the lowest absolute energy. Obviously, the manipulation of structures must avoid the generation of atomic clashes or unphysical structures. For intra-molecular conformational searches crossover works in the following way: (i) starting from the first gene (a dihedral angle value) the mean value of the parent's alleles is calculated; (ii) a stepwise rotation is performed around the selected dihedral angle towards each parent (since two offspring are generated) until no clashes are present (up to using the parent allele); the step size depends on the number of allowed attempts (default = 20).

Mutation works similarly: after a gene and new allele have been generated, the dihedral angle is rotated from the new value towards the old value until clashes are solved. Proxima has been employed during the search to preserve the stereochemistry of the molecules investigated and avoid the clashing of atoms. In order to generate the initial population, alleles are generated from a Latin Hypercube Sampling [166] (LHS hereafter), which is a form of stratified sampling used to generate controlled random ensembles. In a one-dimensional LHS if we have to extract N samples from a distribution we divide it into N evenly spaced regions and then pick a value from each region with uniform probability; in other words, we get one ensemble of N points. Scaling to two variables we divide the space of each variable into N intervals and thus we get an N by N squared grid from which we can get one set of N points (with the requirement that they will not be neighbors or touch at a vertex). With m variables the procedure is similar and there will be just one sampling point for each m-dimensional interval. This procedure is repeated for each specimen that must be generated in the initial population

Aspartic Acid

The first case study is the gas phase conformational landscape of aspartic acid [73] (Fig. 8.32), which is the smallest proteinogenic α-amino acid involving a carboxylic group in the side chain.

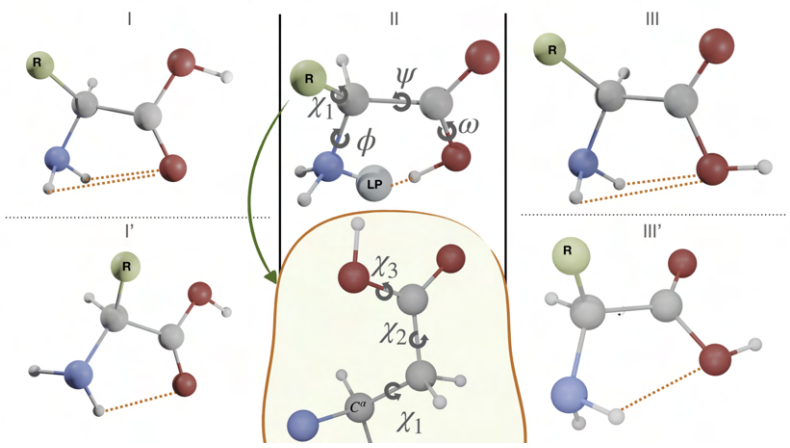


Figure 8.32: The dihedral angles of the aspartic acid with its main conformers.

Its conformational behavior is ruled by the six dihedral angles shown in Fig. 8.32. Three of them belong to the backbone (ϕ, ψ and ω) and the other three to the side-chain (χ_i, i = 1...3). The conventional i labels c, g^-, g and t are used to indicate cis, gauche, or trans conformations of each dihedral angle, whereas the non-planarity of the NH_2 moiety suggests replacing the customary ϕ dihedral angle (HNCC) by $\phi' = LP - C - C - C = \phi + 120°$ (LP is the nitrogen lone-pair). The only conformers observed experimentally for amino acids are stabilized by hydrogen bonds between the amine and carboxyl moieties of the backbone, which can be either bifurcated (e.g., type I, $NH_2 \cdots O{=}C$, $\phi' = 180°, \psi = 180°$ and $\omega = 180°$), or conventional (e.g., type II, $N \cdots H(O)$, $\phi' = 0°, \psi = 0°$ and $\omega = 0°$). Additional conformers are observed when polar side chains are present,

which involve both intra-backbone and backbone-side chain hydrogen bonds. In particular, starting from type I structures, rotation of the NH_2 moiety by about 90° allows its involvement in two different H-bonds (I' conformer, $\phi' = 90°, \psi = 180°$ and $\omega = 180°$). Conformers involving the backbone OH oxygen as the acceptor and the NH_2 moiety as the donor (type III, bifurcated $\phi' = 180°, \psi = 0°$ and $\omega = 180°$, or type III', single, $\phi' = 180°, \psi = 90°$ and $\omega = 180°$) have also been observed in some cases, but they are always the least populated. The IM-EA software has been employed using the GFN2-xTB [159] Force Field. To this end, starting from the 4000 candidates found in each replica, a first reduction to about 1000 structures is obtained by applying a threshold of 25 kJ/mol with respect to the absolute energy minimum. These candidates were compared with each other in terms of the root-mean-square deviations of heavy atom positions and the rotational constant. The 300 structures remaining after this selection are further reduced to about 30 by clustering procedures and subsequent full geometry optimization at the B3LYP/6-311G++(d,p) level leads to 12 conformers lying within 16 kJ/mol. The structures of this final panel of candidates were finally refined at the rev-DSD-PBEP86-D3(BJ) [1–3]/jun-cc-pVTZ level [1, 5]. This composite strategy allows the number of costly geometry optimizations by using hybrid and, especially, double-hybrid functionals to be strongly reduced and to end up with 10 conformers lying within 12 kJ/mol above the absolute energy minimum. In Fig. 8.33, the rotational constants for the six most stable conformers of aspartic acid resulting from such exploration are shown.

Conformer	IIgtt	IIg⁻tt	Igtt	III′gtt	I′g⁻tt	Ig⁻gc
Experimental						
A_0^c	2612.20878(26)	3416.43489(66)	2553.85523(70)	2651.953(31)	3378.20873(26)	3198.861(19)
B_0^c	1191.01132(17)	902.904474(79)	1205.08478(10)	1183.51697(30)	907.373507(28)	945.84803(7)
C_0^c	1057.33169(16)	764.631177(96)	1069.14318(10)	1054.98929(34)	780.042139(32)	781.75139(18)
Computed						
A_0^c	2607.9	3412.3	2546.8	2643.8	3372.8	3192.2
B_0^c	1188.9	900.4	1202.1	1182.9	904.2	943.8
C_0^c	1057.1	762.5	1067.2	1055.9	778.1	781.4

Figure 8.33: Rotational constants (MHz) of the six most stable conformers of aspartic acid issued from the experiment [167] or rev-DSD-PBEP86-D3(BJ) [1–3]/jun-cc-pVTZ [1, 5] computations. Vibrational corrections to rev-DSD-PBEP86-D3(BJ) [1–3]/jun-cc-pVTZ [1, 5] equilibrium rotational constants have been computed at the B3LYP/6-311G++(d,p) [27, 28, 168, 169] level. [73]

Threonine

In the second case, we studied the Threonine molecule [132]. Initial searches were performed in gas-phase using a population of 28 chromosomes either with a single population or using the island model. These runs produced 1394 and 532 structures, respectively, before stalling or reaching the programmed maximum number of generations. Comparison of the sampled structures against the reference dataset [170] using RMSD (cutoff of 0.2 Å) shows that both evolutionary algorithms missed 8 structures out of 56 but with significantly different convergence rates: while the single population runs had similar behavior, slowly improving until the last few generations, the run performed with the island model was able to converge in just 25 generations or about 550 QC calculations. Any-

way, even the worst outcome represents a significant improvement with respect to stochastic methods like Monte Carlo (less than 1400 calculations vs 3000). After the extensive exploration of threonine in the gas phase, we proceeded to characterize its charged forms in solution, employing the Conductor-Like Polarizable Continuum Model (CPCM) [171] to take into account bulk solvent effects. In this case, run-time topology checks are critical since proton transfers may take place during the search. As is well known, in aqueous solutions at neutral pH, the zwitterionic form of amino acids is more stable than its neutral counterpart. PH changes then lead to either protonation of the carboxylate group or deprotonation of the NH_3 moiety. Thus, an extensive exploration of the charged forms of threonine is pivotal to analyzing the relationships among the low-lying conformers of the differently charged species and identifying the preferred paths for protonation or deprotonation. In Fig. 8.34, the geometries and relative free energies of the low-energy conformers (within 12 kJ/mol above the global energy minimum of each form) for anionic, zwitterionic, and cationic forms of threonine are reported, with orange lines connecting closely related structures.

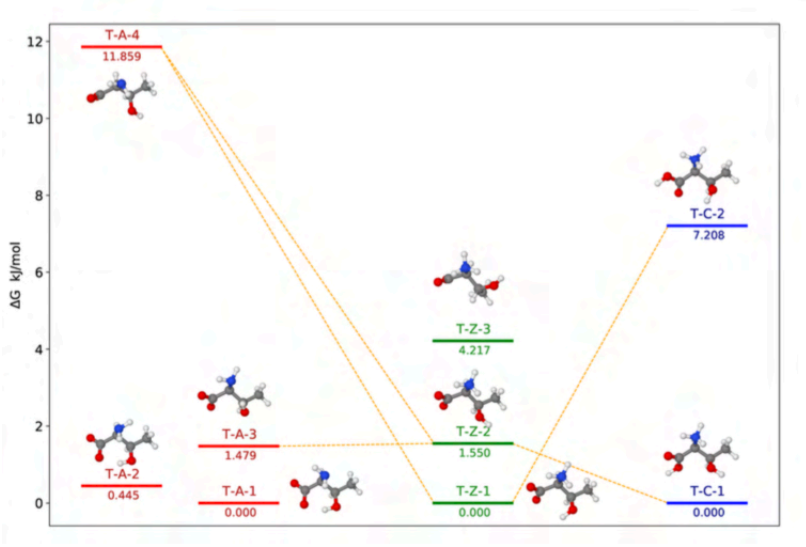


Figure 8.34: The most stable conformer of L-threonine [132].

The most stable conformer of the cationic form (the blue energy level in the figure) is characterized by hydrogen bonds of the positively charged NH_3 group with both the carboxylic and hydroxylic oxygens. This structure is closely related to the second low-energy conformer of the zwitterionic form, which is only 1.5 kJ/mol above the global energy minimum of this form. Interconversion between the two conformers is ruled by the rotation of the hydroxyl hydrogen atom. Only slight structural rearrangements occur during the deprotonation of the carboxylic group. In the case of the anionic form, after deprotonation of the ammonium group, the strongest hydrogen bond is formed between one carboxylic oxygen and the hydrogen of the hydroxyl group (rather than with aminic hydrogen). The interaction between the NH_2 and OH groups is retained in the less stable T-A-3 and T-A-4 conformers, which represent the possible connections between the zwitterionic and the anionic forms upon NH_3 deprotonation.

8.3.2 The Virtual Laboratory

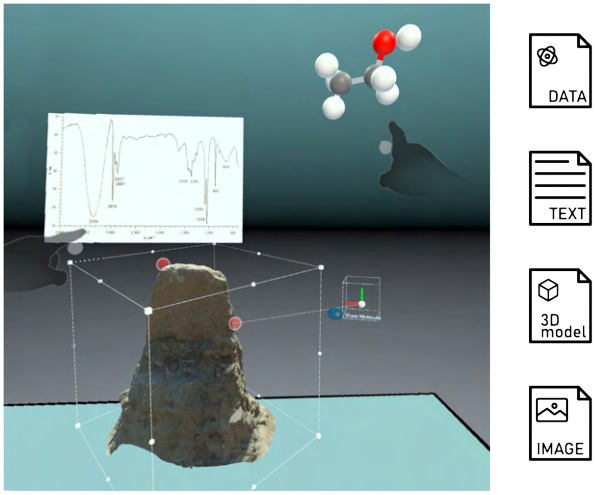


Figure 8.35: The Virtual Laboratory [172] is a virtual collaborative environment connected to data sources to quickly analyze and elaborate data, acting as a bridge between the macroscopic and the microscopic world.

The steps of a computational workflow (pre - processing, computation, post - processing) are traditionally handled separately in different software and platforms. The idea of the Virtual Laboratory is to build a unified environment where it is possible to perform all these different tasks with a focus on collaboration and data sharing between institutions. At the present moment, the Virtual Laboratory is an in-development prototype that uses advanced VR and AR technologies together with traditional desktop environments to link together different applications (e.g. Gaussian for computation, Proxima for Molecular Perception, python ML frameworks for the analysis of data, etc.). However, the long-term goal is to build a unified platform that easily integrates multiple databases, such as the aforementioned ones (SE23, LPCS23). In other words: the goal is to explore the chemical space through human intuition. This idea of building an environment that encompasses both the physical and digital world is a concept that gets proposed frequently in many different technology fields (e.g. gaming, social media, etc.) and that in recent years has been often mentioned as the metaverse.

The term "metaverse" originated in 1992 from a science-fiction novel called Snow Crash by Neal Stephenson [173]. Recently, the term has been widely used to indicate advancements in web technologies (e.g. Web3 [174], NFT [175], etc.) and the concept of a decentralized structure for digital identities. In other words, the idea of the metaverse is to have a unique protocol to associate each physical individual to a digital identity (similar to how protocols of TCP/IP were established as the foundation for the internet, together with internet domains) so that each person is uniquely identified on the web. The decentralized nature of the metaverse is so that each "digital good" is not linked to a specific application (e.g. a game) but instead is linked to a specific identity that can be theoretically shared across applications. A common example is a digital avatar that can be custom created and used in the same way in multiple applications since it is linked to our identity. Moreover, the metaverse enthusiasts hope to build complete digital

markets where users can buy digital goods and use them in every application, as much as you can buy a physical good from the web today. The idea of the metaverse is so deeply linked to Virtual Reality and Augmented Reality since it allows the ability to use these digital goods also in real-life experiences thus blending even more the barrier between the digital and the physical world.

There is, however, a general sense of skepticism towards the concept of the metaverse: The first problem arises from the nature of the metaverse itself and its concept of unique digital identity which poses not few privacy problems. Even with current technologies, there are a lot of issues and discussions about who should be in charge of detaining information about the users of a service (the government, the company that builds the service, etc.) and although encryption is a common strategy employed, there is a lack of clear international policy that ensures that something like the metaverse can be completely safe as an environment. Most importantly, there is the general question of whether something like the metaverse is even needed. Although we tend to get excited when it comes to new technologies, it is important to develop the attitude of distinguishing whether it's a problem in search of a technology to solve it, or it's a technology in search of a problem to solve. Moreover, the metaverse is slowly becoming a "buzzword" used by companies to boost credibility among investors and the general audience. In a sort of analogy, the "Metaverse" is related to Virtual Reality as much as "Artificial Intelligence" is related to Machine-Learning, the latter are technologies whereas the first are philosophies. In the Machine-Learning chapter we discussed the underlying mathematics and algorithmic nature of most common ML practices, the "Artificial Intelligence " philosophy is to employ them to create generative digital assistance tools that are indistinguishable from humans in behavior, thus surpassing the Turing test. The algorithms themselves, however, can have much wider ranges of applications. In the case of Mixed Reality, the usefulness comes from just the ease of visualization and representation, as a technology, without necessarily using it to build an entirely digital world parallel to ours (the metaverse). It is as if we would look at the Internet (the technology) just as a synonym for social platforms (a philosophy on how to use the web).

Independently of whether the metaverse concept will be successful or not, there is a huge opportunity to take inspiration from it in scientific collaboration. As discussed previously, data repositories of scientific information are fundamental to modern research, and often these databases are freely accessed by the scientific community (e.g. the protein data bank [85]). In a sort of way, we are already living in a scientific metaverse, where all scientists are identified in their field by their publications and academic research. However, there is a disconnection between the discovery phase, which is based on dialog, meetings, and human interactions, and the research phase which requires using advanced coding to do even some simple analysis on repositories of data. Simple tests and on-the-fly analysis can provide good insights to then later employ more advanced coding for accurate measurements. Repositories can be accessed directly every time, even during a meeting, and with advanced visualization techniques (e.g. AR, VR, etc.) can be easily processed on the fly in a collaborative environment. This is the idea of the Virtual Laboratory, to become a useful practical tool where new data can be produced on-the-fly starting from stored data, and shared between scientists in a connected environment.

In the next sections, we are going to illustrate the two main applications we developed following the philosophy of the Virtual Laboratory.

The molecular editor

The first application we developed is a molecular viewer that can be used with the Meta Quest 2 [176] hardware. The viewer is a module that can be loaded in a generic Unity [177] project to visualize molecular data in different macroscopic environments as shown in Fig. 8.36.

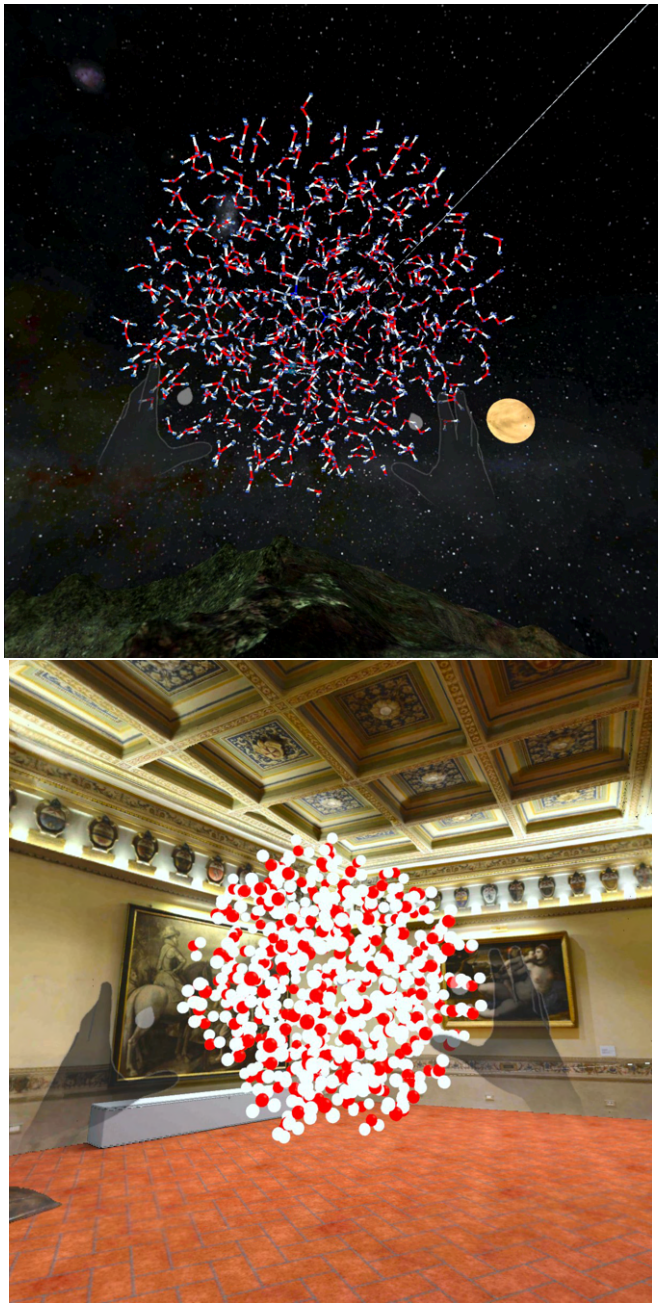


Figure 8.36: The Proxima spherical solvent generation procedure implemented in Virtual Reality [178].

Moreover, the software allows for the editing of molecular structures. In par-

ticular, the user can employ their physical hands to place atoms in space and, with a dedicated button, optimize the geometry. The optimization step is required since the placement of atoms by hand is far from precise but still far more immediate to build molecules faster. The optimization is performed by connecting the headset to a server running the XTB [159] software. The geometry can then be further refined utilizing the TMA tools described in previous sections. The resulting geometry is, as a consequence, accurate enough although being built literally by hand. This application employs all of the Proxima molecular perception algorithms. As a consequence, atomic charges can be directly computed on the headset and the electrostatic potential can be easily visualized employing a custom plane shown in Fig. 8.37.

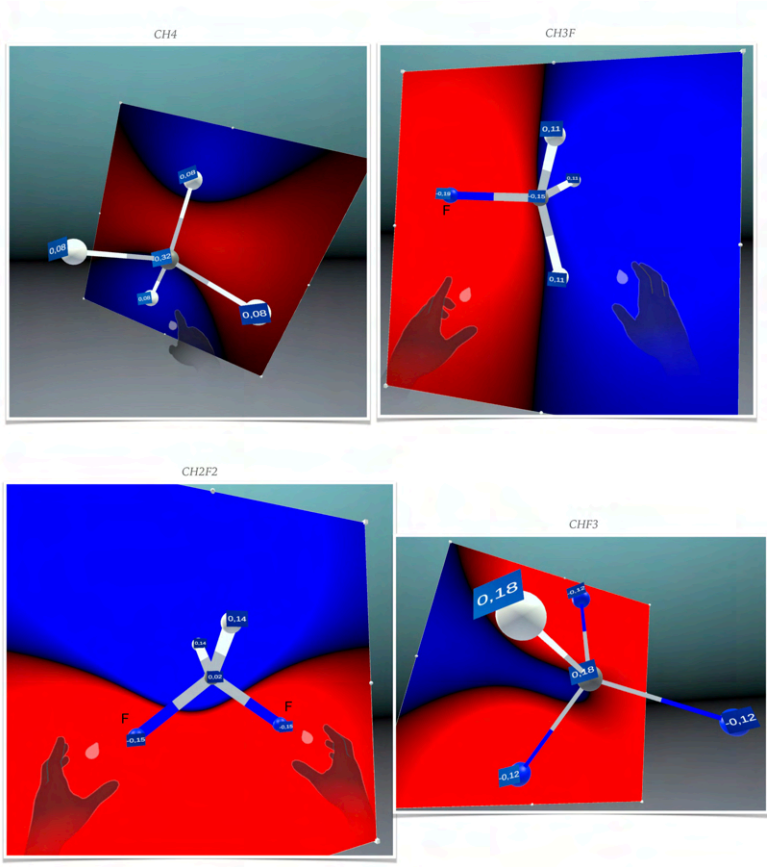


Figure 8.37: The shader developed to visualize the electrostatic potential on a moveable plane.

Here, the plane can be moved in space, and through a custom shader it colors itself depending on the value of the electrostatic potential in that point. Another feature of proxima employed in the software is the automatic generation of spherically solvated environments (see Chap 4). In Fig. 8.36, two spherically solvated systems are shown in two different macroscopic environments (an asteroid and the model of Sala Stemmi, from Scuola Normale Superiore, obtained from photogrammetry). The possibility to move from macroscopic environments such as the ones of Fig. 8.36 and the microscopic world of molecules is one of the advantages of these Mixed Reality applications and allows for the easy implementation of advanced algorithms in simple applications making them available

to non-experts in the education field or to experimentalists.

The PES explorer

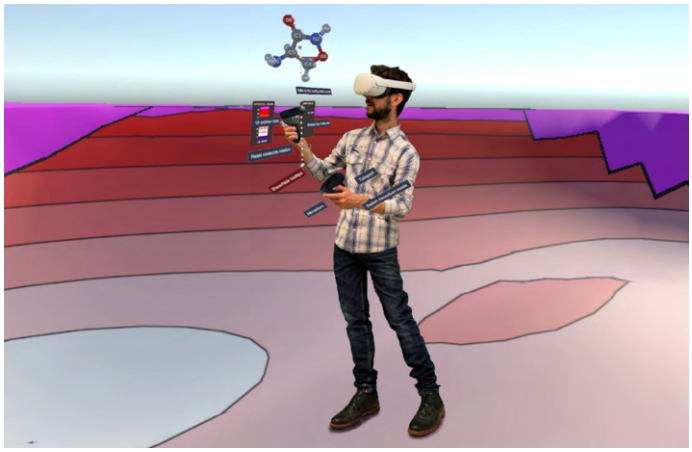


Figure 8.38: The Potential Energy Surface (PES) exploration in Virtual Reality.

As already stressed in the introduction to this chapter, the exploration step can either be driven by automatic tools (machine intelligence) or by human intuition. In this context, the exploration of a Potential Energy Surface (PES) is a complex task that can be extremely simplified by Mixed Reality (MR) technologies. The PES of an N-atom molecular system is an inherently multidimensional function, or hypersurface, depending on the related $3N$-6 generalized coordinates. In the context of human exploration, the software should help the user (the scientist) to physically explore the PES searching for points of interest. The goal is to map the multidimensional PES to a three-dimensional representation that can be explored in Virtual Reality while retaining the amount of chemical information needed to describe the phenomenon investigated. In order to achieve such a goal, we have developed AVATAR (Advanced Virtual Approach to Topological Analysis of Reactivity) [179], an IVR application based on head-mounted displays and handheld controllers that take advantage of IVR for the specific task of immersive visual analysis of PES s based on the following two key concepts: (a) the reduction of the dimensionality of the PES to two process-tailored, physically meaningful generalized coordinates, and (b) the analogy between the evolution of a chemical process and a pathway through valleys (potential wells) and mountain passes (saddle points) of the associated potential energy landscape. As mentioned earlier, the description of the relative assembly of an N-atom system is achieved by using $3N$-6 independent geometrical coordinates. In order to represent these PES s in 3D IVR environments, this number has to be opportunely reduced to two dimensions (given the obvious limitation of human perception). In fact, unless a nongeometric encoding is used for the energy, in three-dimensional space there are only two dimensions available for describing the geometry of the system, as the third dimension has then to be used to represent the potential energy values. In these studies, Principal Component Analysis (PCA), the most widespread feature extraction technique, is used to reduce the dimensionality of the problem. However, PCA-based approaches suffer from many drawbacks, including their

inability to capture any nonlinear nature of data and also to characterize strong overlapping data. While all these techniques can be used in the preprocessing stage, a useful alternative approach is that of adopting specialized sets of ad hoc defined coordinates. In particular, the recipe that we exploit in AVATAR is as follows: (a) perform a change of coordinates by opportunely combining the generalized coordinates thus obtaining a new set of coordinates, two of which have a high descriptive value for the process under investigation (so-called process coordinates); (b) for each combination of the values of these two coordinates, relax the potential energy with respect to all other coordinates to get a so-called relaxed 3D representation of the energy landscape as a function of the two process coordinates. In this context, it is useful to remember the definition of the ring's coordinates as discussed in Chap. 4. In this case, it is possible to define an entire five-membered ring conformation by means of two variables, by relaxing the other degrees of freedom. In this context, the relaxation paths are shown in Fig. 8.39 for five-membered rings.

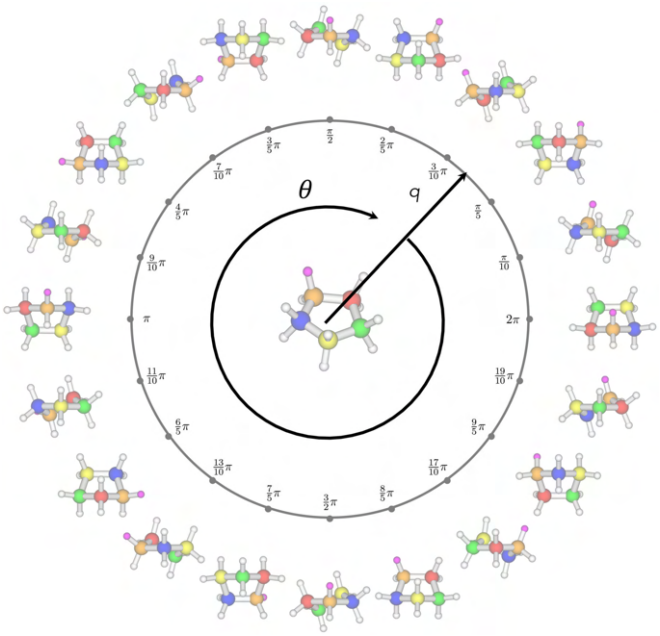


Figure 8.39: Circular relaxed plots of the PES for ring-puckering in five-term rings by means of the puckering amplitude q and the pseudorotation angle θ [179].

The first example shown is the interconversion between conformers in ring-puckering motions of silacyclopentane, a molecule for which experimental far-infrared [180, 181], microwave [182] and Raman [183] spectra are available in the literature. A new bidimensional PES for this system has been calculated explicitly with density-functional theory (DFT) by employing B2PLYP [30] as an exchange-correlation functional combined with Grimme's D3(BJ) dispersion [2, 3, 184] and aug-cc-pVTZ basis set introduced by Truhlar and coworkers [185, 186]. For the sampling of the bi-dimensional PES a 70X70 uniformly spaced rectangular grid in $qcos\theta$ and $qsin\theta$ was used. The number of calculations to be performed was reduced by exploiting symmetry relations between portions of the circular domain. Each sampled point of the bi-dimensional PES was calculated independently through constrained optimization. Each input was constructed

using both Cartesian and primitive internal coordinates in the same Z-matrix. The PES for the silacyclopentane that can be explored in Virtual Reality is shown in Fig. 8.40.

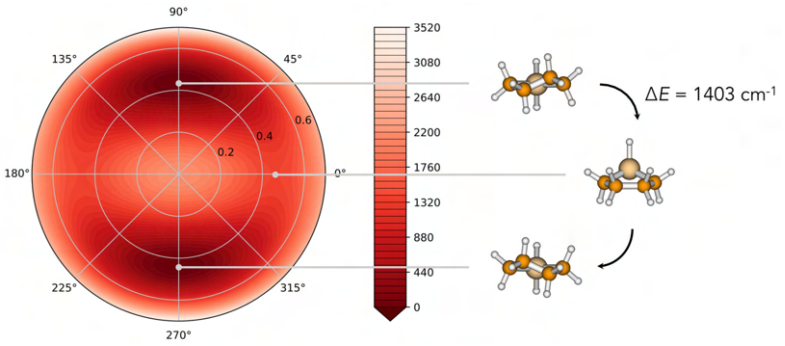


Figure 8.40: Circular relaxed plot of silacyclopentane as a function of ring-puckering coordinates q and θ (energy values are in cm^{-1}) [179].

In the IVR session with AVATAR, the user starts on the mountain peak at the origin (second-order saddle point), corresponding to the ring planar conformation and being higher in the energy of 2193 cm^{-1} with respect to the most stable conformer. This is reached by descending the peak either along the $\phi = 90$ line or along the $\phi = 270$ line, leading to two symmetric and energetically identical wells corresponding to a twisted structure of C_2 symmetry ($q = 0.436$). These equivalent minima are associated with two enantiomeric structures and are connected by two equivalent transition states of C_s symmetry along a circular path with q comprised between 0.4 and 0.5, and accordingly featuring two mountain passes. Such qualitative description is confirmed by experimental results [180–183]. The barrier to the pseudorotation (the height of the mountain passes above the two potential wells) calculated in this work is 1403 cm^{-1}, which can be compared with the extrapolations of the barrier obtained from experimental results: the values available in the literature are 3.89 kcal/mol (about 1360 cm^{-1}) [180] and 1414 cm^{-1} [183], both in good agreement with the calculated value reported here. The calculated energy of the second-order saddle point associated with the planar conformation of the silacyclopentane ring molecule is 2193 cm^{-1}, to be compared with a lower value of 1559 cm^{-1} already available in the literature [183]. Another example studied in the context of the VR exploration of potential energy surfaces has been the case of Atom-Diatom reactions ($A + BC \longrightarrow AB + C$). Though the simplicity of such a process makes it look like a mere abstract model far from the complexity of the real world, these kinds of reactive collisions can be experimentally reproduced and characterized through so-called crossed-molecular-beam experiments [187] and are of prominent relevance in astrochemistry due to the extreme conditions of low temperatures and pressure of the interstellar medium [188]. This kind of reaction involves three atoms ($N = 3$), and the associated PES depends on $3N$-6 = 3 generalized coordinates. A set of generalized coordinates commonly adopted for the description of the process is shown on top of Fig. 8.41, consisting of two internuclear distances (those of the breaking and forming bonds) and the angle between them.

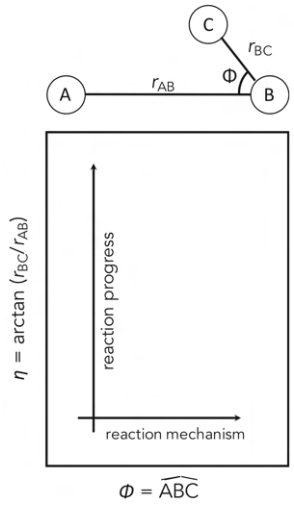


Figure 8.41: Rectangular relaxed plot of the PES for A + BC → AB + C reactions by means of the "reaction-progress" and "reaction-mechanism" coordinates [179].

Useful two-dimensional PES representations may be obtained by retaining the two distances and either fixing the angle or minimizing the energy with respect to it. While these plots are widely adopted and provide a useful description of fixed-angle or minimum-energy reaction paths, information on possible competing reaction mechanisms involving different atom–diatom approaching or scattering angles is undoubtedly lost. A more informative representation is the so-called rectangular relaxed plot [189, 190] where the PES for a given atom–diatom exchange reaction is represented as a function of a "reaction-progress" and a "reaction-mechanism" coordinate based on the following coordinate change:

$$\begin{cases} \eta = arctan(r_{BC}/r_{AB}) \\ \Phi = A\hat{B}C \\ \rho = \sqrt{r_{AB}^2 + r_{BC}^2} \end{cases} \tag{8.32}$$

The first of these coordinates, η, is a reaction-progress coordinate in that it measures the ratio between the breaking over the forming bond distance. The second coordinate accounts for the detailed mechanism by which the reaction occurs, as it relates to the approaching angle of A toward BC and the scattering angle of C from the newly formed AB diatom. The third coordinate is an "overall-size" coordinate with less informative content. A rectangular relaxed plot (see Figure 8.41) is obtained by plotting for each couple (η, Φ) the value of the potential energy minimized with respect to ρ, that is: $min_\rho V(\eta, \phi)$. The name is rectangular and derives from the shape of the domain of points usually adopted in such representation, where reaction progress is emphasized by using the long side of the rectangle for coordinate η and the short side for coordinate Φ. Moving along the horizontal axis of the rectangle in Figure 8.41, the Φ angle changes providing information about the detailed mechanism of the reaction, while moving along the vertical axis the η angle changes quantifying the progress of the reaction. In Fig. 8.42 the potential energy surface obtained for a chemical reaction of astrochemical interest [191] ($C + CH^+ \longrightarrow C_2{}^+ + H$) is shown.

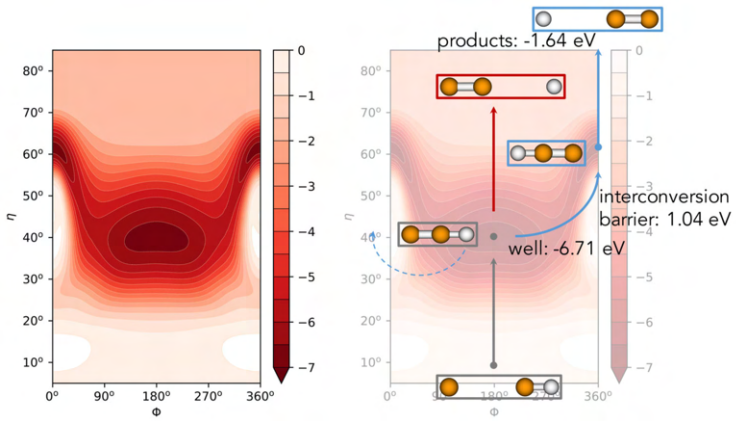


Figure 8.42: Left-hand side: rectangular relaxed representation of the PES for the $C + CH^+ \longrightarrow C_2{}^+ + H$ (energy values are in eV). Right-hand side: competing reaction mechanisms superimposed to the PES representation [179].

The energy zero is usually set at the reactants channel. In a typical IVR session, the user starts in the reactants region (bottom edge of the rectangle) and faces a deep well approximately mid-way on the route to products, which are located at the top edge of the rectangle. Such a potential well, as deep as 6.71 eV, corresponds to the barrierless formation of a collinear reaction intermediate C_2H^+ with the C-H bond shorter than the C-C bond. As the user moves toward the well, he/she will see (floating above the right-hand controller) the C_A atom approaching the C_BH^+ diatom from the carbon side (where labels A and B have been adopted to distinguish between the two carbon atoms), the angle of approach depending on his/her position along the Φ axis. Once the $C_AC_BH^+$ intermediate is reached, the user will realize that among the infinite alternative paths leading to products, there are two that are more interesting than others. The user can either move on toward products by staying in the middle of the Φ axis (red path in Figure 8.42: dissociation into products with the H atom leaving $C_2{}^+$ from the C_B side in a collinear fashion) or explore a second, identical potential well due to rotation of the hydrogen atom about the carbon–carbon bond (blue path in Figure 8.42). This last path involves overcoming a mountain pass between the two potential valleys (rotational barrier of 1.04 eV) before forming the linear triatomic $C_AC_BH^+$ and further proceeding to products by the departure of H from the C_A side. Once in the product region, by looking at the color and at the energy display or simply looking back to the (higher) reactant region, the user will realize that the reaction is exoergic by 1.64 eV.

Chapter 9

Conclusions

The present thesis aimed to develop and validate a new set of Chemical Descriptors derived from Molecular Perception to represent the feature space of atoms and bonds (the synthon). In the first chapter, we discussed the problem of Representation in Chemistry and how it is deeply linked to the problem of Computation. Then, in the second chapter, we noticed how the paradigm of Computation has recently shifted due to the advent of Machine-Learning and how such a shift required deep thought into the way chemical data are represented numerically. We highlighted how the application of ML algorithms to chemical problems requires the creation of dedicated Feature Spaces. To build such Feature Spaces, basic descriptors for the molecular properties must be provided. Thus, in the third chapter, we discussed the importance of Molecular Perception as the field that tries to compute molecular properties (e.g. covalent bonds, charges, etc.) starting from the minimum amount of information possible (e.g. XYZ coordinates). The Molecular Perception algorithms are all developed in a C++ library with Python bindings called Proxima. The fourth chapter was specifically dedicated to the definition of a new chemical feature space for atoms and bonds based on the bond order and charge matrix computed by Proxima. Then, we discussed how a possible use of our feature space could be the automatic assignment of atom types and the automatic parameterization of energy profiles with correct boundaries for force fields. Thus, in Chapter 6, we discussed the analytical expressions for single/double well symmetric/asymmetric potentials, how to compute the gradient and the hessian for radial energies in cartesian coordinates, and how to move from cartesian coordinates to a generic set of internal coordinates. Finally, in the last chapter, we presented applications of our feature space to the problem of assigning traditional atom types, clustering similar atoms, and performing fragment detection on a database to refine structures (the TMA approach). We also showed the effectiveness of the single/double well analytical expressions in describing common non-covalent, stretching, and bending energy terms. We concluded the thesis by discussing the Virtual Laboratory as the final outcome of all these different areas (traditional QM computations, Machine-Learning, Molecular Perception, etc.) in building a collaborative environment with shared access to structural databases. The overall workflow of this thesis has already been summarized in Fig. 1, where the flow of data is shown in the pre-processing phase from our databases up to the computational step that exploits and explores chemical spaces up to the immersive visualization of the

computational outcomes.

The next step will be to better consolidate the Virtual Laboratory workflow by building user interfaces to access our SE/23 and PCS/23 databases with the ability to automatically perform TMA. The application of synthons and the atomic feature space has been limited to simple ML procedures since the main goal was to validate and test the efficacy of our definition. However, the next step will be to employ the feature space in treating intrinsically nonlinear problems through advanced ML algorithms. An example is the same TMA approach where we assumed linearity when studying change in geometrical parameters of the refined fragment structures (SE or LPCS) with respect to the original ones (in this case revDSD). In general, such an assumption might fail when dealing with lower levels of theory and intrinsically non-linear models might be employed (e.g. Neural Networks). Moreover, up to this point, the refinement has been performed on geometrical parameters but the same refinement approach could be applied to other physical-chemical properties (e.g. dipole moments, nuclear quadrupole couplings, force constants, etc.). It is important to remember that the computation of energy profiles, as described in Chapter 6 and validated in the final application chapter, can also be extended to use ML models to predict the well parameters. As an example, by extending the synthon to the treatment of angles, it could be possible to train a network to automatically detect the depth and position of the bending energy wells. The other possible path for future research is to employ our feature space not only as a tool for analysis and refinement but as a mean to generate new data thanks to the recent advancements in generative AI models. In particular, by having a numerical interpretable description of atoms and bonds, we could study the application of generative models in building molecules that respect certain atomic, bond, or fragment requirements. This could in principle help in the automatic generation of molecular candidates that satisfy certain needs (e.g. drug discovery, conformational analysis, etc.) not just based on randomness (such as in the case of Genetic Algorithm) or pure database patterns (e.g. TMA), but using guided Artificial Intelligence.

It is important to remark how much human expertise is still at the core of scientific research and it is now more important than ever to build user interfaces and graphical tools that allow the expert to guide and understand such black-box engines. Command lines and spreadsheets work very well when dealing with tabular numerical data computed with pre-defined deterministic methods, and they speed up research by processing a lot of data with short text prompts such as single terminal commands. Black-box methods, however, still require to be monitored and their behavior must be interpreted. We already saw an example in the application chapter when visualizing the Decision Tree or the clustering results for atom types or even the feature space of TMA fragments. I would like to conclude this thesis by quoting the words of the famous scientist August Kekulé that claimed that he had visions helping him understand the structure of benzene. We now know that Loschmidt anticipated Kekulé in determining the structure of benzene. However, scientific credit is also given to a large degree based on how well the claim becomes known, and how widely it is communicated and disseminated:

> I was sitting and writing my textbook, but the work did not progress; my thoughts were elsewhere. I turned my chair to the fire and dozed.

> Again the atoms were gamboling before my eyes. This time the smaller groups kept modestly in the background. My mental eye, rendered more acute by the repeated visions of the kind, could now distinguish larger structures of manifold conformation; long rows sometimes more closely fitted together all twining and twisting in snake-like motion. But look! What was that? One of the snakes had seized hold of its own tail, and the form whirled mockingly before my eyes. As if by a flash of lightning I awoke; and this time also I spent the rest of the night working out the consequences of the hypothesis. Let us learn to dream, gentlemen, and then perhaps we shall learn the truth ... but let us beware of publishing our dreams before they have been put to the proof by the waking understanding.

Kekulé famous dream of the benzene structure, as quoted in A Life of Magic Chemistry: Autobiographical Reflections of a Nobel Prize Winner (2001) by George A. Olah, p. 54 [192]

Bibliography

(1) Santra, G.; Sylvetsky, N.; Martin, J. M. *The Journal of Physical Chemistry A* **2019**, *123*, 5129–5143.

(2) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *The Journal of Chemical Physics* **2010**, *132*, 154104.

(3) Grimme, S.; Ehrlich, S.; Goerigk, L. *Journal of Computational Chemistry* **2011**, *32*, 1456–1465.

(4) Dunning, T. H.; Peterson, K. A.; Wilson, A. K. *The Journal of Chemical Physics* **2001**, *114*, 9244–9253.

(5) Papajak, E.; Zheng, J.; Xu, X.; Leverentz H., R.; Truhlar D., G. *Journal of Chemical Theory and Computation* **2011**, *7*, 3027–3034.

(6) Barone, V. *The Journal of Chemical Physics* **2005**, *122*, 014108.

(7) Bloino, J.; Biczysko, M.; Barone, V. *Journal of Chemical Theory and Computation* **2012**, *8*, 1015–1036.

(8) Bloino, J.; Biczysko, M.; Barone, V. *The Journal of Physical Chemistry A* **2015**, *119*, 11862–11874.

(9) Peterson, K. A.; Dunning Jr, T. H. *The Journal of Chemical Physics* **2002**, *117*, 10548–10560.

(10) Ess, D. H.; Jelfs, K. E.; Kulik, H. J. *The Journal of Chemical Physics* **2022**, *157*, 120401.

(11) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. *Communications of the ACM* **2020**, *63*, 139–144.

(12) Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp 10684–10695.

(13) OpenAI ChatGPT, https://chat.openai.com/, [Online; accessed 11-May-2023], 2023.

(14) Loschmidt, J., *Chemische Studien. Reprinted by the Aldrich Chemical Company, Milwaukee, WI, Cat. No. Z-l 8576-0, 1989.* 1861.

(15) De Chadarevian, S. *Protein Science* **2018**.

(16) Hartree, D. R.; Hartree, W. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences* **1935**, *150*, 9–33.

(17) Wahl, A. C.; Das, G. In *Methods of Electronic Structure Theory*, Schaefer, H. F., Ed.; Springer US: Boston, MA, 1977, pp 51–78.

(18) David Sherrill, C.; Schaefer, H. F. In Löwdin, P.-O., Sabin, J. R., Zerner, M. C., Brändas, E., Eds.; Advances in Quantum Chemistry, Vol. 34; Academic Press: 1999, pp 143–269.

(19) Møller, C.; Plesset, M. S. *Physical Reviews* **1934**, *46*, 618–622.

(20) Paldus, J. In *Theory and Applications of Computational Chemistry*, Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005, pp 115–147.

(21) Hohenberg, P.; Kohn, W. *Physical Reviews* **1964**, *136*, B864–B871.

(22) Kohn, W.; Sham, L. J. *Physical Reviews* **1965**, *140*, A1133–A1138.

(23) Vosko, S. H.; Wilk, L.; Nusair, M. *Canadian Journal of Physics* **1980**, *58*, 1200–1211.

(24) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Physical Reviews B* **1992**, *46*, 6671–6687.

(25) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Physical Reviews B* **1993**, *48*, 4978–4978.

(26) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Reviews Letters* **1996**, *77*, 3865–3868.

(27) Becke, A. D. *Physical Reviews A* **1988**, *38*, 3098–3100.

(28) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

(29) Grimme, S.; Neese, F. *The Journal of Chemical Physics* **2007**, *127*, 154116, DOI: 10.1063/1.2772854.

(30) Grimme, S. *The Journal of Chemical Physics* **2006**, *124*, 034108.

(31) Kong, L.; Bischoff, F. A.; Valeev, E. F. *Chemical Reviews* **2012**, *112*, PMID: 22176553, 75–107.

(32) Knizia, G.; Adler, T. B.; Werner, H.-J. *The Journal of Chemical Physics* **2009**, *130*, 054104, DOI: 10.1063/1.3054300.

(33) Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2012**, *8*, PMID: 26596602, 527–541.

(34) Hirshfeld, F. L. *Theoretica chimica acta* **1977**, *44*, 129–138.

(35) Warshel, A.; Florián, J. In *Encyclopedia of Computational Chemistry*; John Wiley & Sons, Ltd: 2004.

(36) Van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. *The Journal of Physical Chemistry A* **2001**, *105*, 9396–9409.

(37) Kettering, C. F.; Shutts, L. W.; Andrews, D. H. *Physical Reviews* **1930**, *36*, 531–543.

(38) Allinger, N. L. *Advances in Physical Organic Chemistry* **1976**, *13*, 1–82.

(39) Allinger, N. L. *Journal of the American Chemical Society* **1977**, *99*, 8127–8134.

(40) Allinger, N. L.; Yuh, Y. H.; Lii, J. H. *Journal of the American Chemical Society* **1989**, *111*, 8551–8566.

(41) Del Re, G. *Journal of the Chemical Society* **1958**, 4031–4040.

(42) Halgren, T. A. *Journal of Computational Chemistry* **1996**, *17*, 490–519.

(43) Halgren, T. A. *Journal of Computational Chemistry* **1999**, *20*, 720–729.

(44) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A. I.; Skiff, W. M. *Journal of the American Chemical Society* **1992**, *114*, 10024–10035.

(45) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr., K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *Journal of Computational Chemistry* **2005**, *26*, 1668–1688.

(46) Ferguson, D. M.; Kollman, P. A. *Journal of Computational Chemistry* **1991**, *12*, 620–626.

(47) Valiant, L. G. *Communications of the ACM* **1984**, *27*, 1134–1142.

(48) Bonaccorso, G., *Machine learning algorithms*; Packt Publishing Ltd: 2017.

(49) Goodfellow, I.; Bengio, Y.; Courville, A., *Deep learning*; MIT press: 2016.

(50) Tropsha, A. *Molecular Informatics* **2010**, *29*, 476–488.

(51) F.R.S., K. P. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2*, 559–572.

(52) Hoerl, A. E.; Kennard, R. W. *Technometrics* **1970**, *12*, 55–67.

(53) Tibshirani, R. *Journal of the Royal Statistical Society. Series B (Methodological)* **1996**, *58*, 267–288.

(54) Zou, H.; Hastie, T. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320.

(55) Mor-Yosef, S.; Samueloff, A.; Modan, B.; Navot, D.; Schenker, J. G. *Obstetrics & Gynecology* **1990**, *75*.

(56) Zhang, H. *The Florida AI Research Society* **2004**, *1*, 3.

(57) Cortes, C.; Vapnik, V. *Machine Learning* **1995**, *20*, 273–297.

(58) Schölkopf, B.; Williamson, R. C.; Smola, A.; Shawe-Taylor, J.; Platt, J. In *Advances in Neural Information Processing Systems*, ed. by Solla, S.; Leen, T.; Müller, K., MIT Press: 1999; Vol. 12.

(59) Buntine, W. *Statistics and Computing* **1992**, *2*, 63–73.

(60) Ho, T. K. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, 1995; Vol. 1, 278–282 vol.1.

(61) MacQueen, J In *5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp 281–297.

(62) Lloyd, S. *IEEE transactions on information theory* **1982**, *28*, 129–137.

(63) Kaufman, L., *Partitioning around medoids (program pam)*; John Wiley & Sons, Inc.: 1990; Vol. 344, pp 68–125.

(64) Rousseeuw, P. J. *Journal of computational and applied mathematics* **1987**, *20*, 53–65.

(65) Caliński, T.; Harabasz, J. *Communications in Statistics-theory and Methods* **1974**, *3*, 1–27.

(66) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X., et al. In *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996; Vol. 96, pp 226–231.

(67) Sharma, S.; Batra, N., et al. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp 568–573.

(68) Jh, H., *Adaptation in natural and artificial systems*; The University of Michigan Press: 1975.

(69) Mitchell, M. In *Complexity*, 1995; Vol. 1, pp 31–39.

(70) Llanio-Trujillo, J.; Marques, J.; Pereira, F. *The Journal of Physical Chemistry A* **2011**, *115*, 2130–2138.

(71) Bäck, T.; Schwefel, H.-P. *Evolutionary computation* **1993**, *1*, 1–23.

(72) Whitley, D.; Rana, S.; Heckendorn, R. B. *Journal of computing and information technology* **1999**, *7*, 33–47.

(73) Mancini, G.; Fusè, M.; Lazzari, F.; Barone, V. *Digital Discovery* **2022**, *1*, 790–805.

(74) McCulloch, W. S.; Pitts, W. *The bulletin of mathematical biophysics* **1943**, *5*, 115–133.

(75) Cornell University News Service records. Division of Rare and Manuscript Collections, Cornell University Library.

(76) Rojas, R.; Rojas, R. *Neural networks: a systematic introduction* **1996**, 149–182.

(77) Lazzari, F.; Salvadori, A.; Mancini, G.; Barone, V. *Journal of Chemical Information and Modeling* **2020**, *60*, 2668–2672.

(78) Labute, P. *Journal of Chemical Information and Modeling* **2005**, *45*, 215–221.

(79) Zhao, Y.; Cheng, T.; Wang, R. *Journal of Chemical Information and Modeling* **2007**, *47*, 1379–1385.

(80) Behnel, S.; Bradshaw, R.; Citro, C.; Dalcin, L.; Seljebotn, D. S.; Smith, K. *Computing in Science & Engineering* **2010**, *13*, 31–39.

(81) Pedregosa, F. et al. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(82) Virtanen, P. et al. *Nature Methods* **2020**, *17*, 261–272.

(83) Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M., et al. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 2016; Vol. 16, pp 265–283.

(84) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. *Journal of computer-aided molecular design* **2016**, *30*, 595–608.

(85) Bank, P. D. *Nature New Biology* **1971**, *233*, 223.

(86) Weininger, D. *Journal of chemical information and computer sciences* **1988**, *28*, 31–36.

(87) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*, 1–14.

(88) Schomaker, V.; Stevenson, D. *Journal of the American Chemical Society* **1941**, *63*, 37–40.

(89) Porterfield, W. W., *Inorganic chemistry*; Academic press: 2013.

(90) Houser, J. J.; Klopman, G. *Journal of Computational Chemistry* **1988**, *9*, 893–904.

(91) Nash, H.; Grossman, S.; Bradley, D. *Nature* **1968**, *219*, 370–370.

(92) Del Re, G. *Journal of the Chemical Society (Resumed)* **1958**, 4031–4040.

(93) Bundy, A.; Wallen, L. *Catalogue of artificial intelligence tools* **1984**, 13–13.

(94) Shannon, R.; Hornung, B; Tew, D.; Glowacki, D. *The Journal of Physical Chemistry A* **2019**, *123*, 2991–2999.

(95) Wiberg, K. B. *Tetrahedron* **1968**, *24*, 1083–1096.

(96) Rauk, A., *Orbital interaction theory of organic chemistry*; John Wiley & Sons: 2004.

(97) Mo, Y.; Wu, W.; Song, L.; Lin, M.; Zhang, Q.; Gao, J. *Angewandte Chemie* **2004**, *116*, 2020–2024.

(98) Mullins, J. *Journal of Chemical Education* **2012**, *89*, 834–836.

(99) Yang, L.; Sun, L.; Deng, W.-Q. *The Journal of Physical Chemistry A* **2020**, *124*, 2102–2107.

(100) Pagliai, M.; Cardini, G.; Righini, R.; Schettino, V. *The Journal of Chemical Physics* **2003**, *119*, 6655–6662.

(101) Pagliai, M.; Muniz-Miranda, F.; Cardini, G.; Righini, R.; Schettino, V. *The Journal of Physical Chemistry Letters* **2010**, *1*, 2951–2955.

(102) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *The Journal of Chemical Physics* **1983**, *79*, 926–935.

(103) Berger, F.; Gritzmann, P.; de Vries, S. *Algorithmica* **2004**, *40*, 51–62.

(104) Horton, J. D. *SIAM Journal on Computing* **1987**, *16*, 358–366.

(105) Cremer, D. t.; Pople, J. *Journal of the American Chemical Society* **1975**, *97*, 1354–1358.

(106) Paoloni, L.; Rampino, S.; Barone, V. *Journal of Chemical Theory and Computation* **2019**, *15*, 4280–4294.

(107) Zou, W.; Izotov, D.; Cremer, D. *The Journal of Physical Chemistry A* **2011**, *115*, 8731–8742.

(108) Kochev, N. T.; Paskaleva, V. H.; Jeliazkova, N. *Molecular informatics* **2013**, *32*, 481–504.

(109) Sobez, J.-G.; Reiher, M. *Journal of Chemical Information and Modeling* **2020**, *60*, 3884–3900.

(110) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *Journal of Computational Chemistry* **2005**, *26*, 1701–1718.

(111) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(112) Pauling, L. *Journal of the American Chemical Society* **1947**, *69*, 542–553.

(113) De la Vega, J. R.; Busch, J. H.; Schauble, J. H.; Kunze, K. L.; Haggert, B. E. *Journal of the American Chemical Society* **1982**, *104*, 3295–3299.

(114) Flanigan, M. C.; de la Vega, J. R. *The Journal of Chemical Physics* **2003**, *61*, 1882–1891.

(115) Fluder Jr, E. M.; de La Vega, J. R. *Chemical Physics Letters* **1978**, *59*, 454–456.

(116) Downing, C. *Journal of Mathematical Physics* **2013**, *54*, 072101.

(117) Wilson Jr, E. B. *The Journal of Chemical Physics* **1939**, *7*, 1047–1052.

(118) Frisch, M. J. et al. Gaussian˜16 Revision C.01, Gaussian Inc. Wallingford CT, 2016.

(119) Pulay, P.; Fogarasi, G.; Pang, F.; Boggs, J. E. *Journal of the American Chemical Society* **1979**, *101*, 2550–2560.

(120) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *Journal of Computational Chemistry* **1996**, *17*, 49–56.

(121) Lauvergnat, D.; Nauts, A. *The Journal of Chemical Physics* **2002**, *116*, 8560–8570.

(122) Schaad, L.; Hu, J. *Journal of Molecular Structure* **1989**, *185*, 203–215.

(123) Brandhorst, K.; Grunenberg, J. *The Journal of Chemical Physics* **2010**, *132*, 184101.

(124) Miyazawa, T. *The Journal of Chemical Physics* **1958**, *29*, 246–246.

(125) Godfrey, P. D.; Rodgers, F. M.; Brown, R. D. *Journal of the American Chemical Society* **1997**, *119*, 2232–2239.

(126) Florio, G. M.; Christie, R. A.; Jordan, K. D.; Zwier, T. S. *Journal of the American Chemical Society* **2002**, *124*, 10236–10247.

(127) Helgaker, T.; Klopper, W.; Tew, D. P. *Molecular Physics* **2008**, *106*, 2107–2143.

(128) Karton, A. *WIREs, Computational Molecular Science* **2016**, *6*, 292–310.

(129) Puzzarini, C.; Bloino, J.; Tasinato, N.; Barone, V. *Chemical reviews* **2019**, *119*, 8131–8191.

(130) Kesharwani, M. K.; Karton, A.; Martin, J. M. *Journal of Chemical Theory and Computation* **2016**, *12*, 444–454.

(131) Wang, P.; Shu, c.; Ye, H.; Biczysko, M. *Journal of Physical Chemistry A* **2021**, *125*, 9826–9837.

(132) Mancini, G.; Fusè, M.; Lazzari, F.; Chandramouli, B.; Barone, V. *The Journal of Chemical Physics* **2020**, *153*, 124110.

(133) Ferro-Costas, D.; Mosquera-Lois, I.; Fernandez-Ramos, A. *Journal of Cheminformatics* **2021**, *13*, 100.

(134) León, I; Fusè, M.; Alonso, E. R.; Mata, S.; Mancini, G.; Puzzarini, C.; Alonso, J. L.; Barone, V. *Journal of Chemical Physics* **2022**, *157*, 074107.

(135) Lupi, J.; Alessandrini, S.; Barone, V.; Puzzarini, C. *Journal of Chemical Theory and Computation* **2021**, *17*, 6974–6992.

(136) Grimme, S. *Chemistry - A European Journal* **2012**, *18*, 9955–9964.

(137) Falbo, E.; Fusè, M.; Lazzari, F.; Mancini, G.; Barone, V. *Journal of Chemical Theory and Computation* **2022**, *18*, 6203–6216.

(138) Barone, V.; Fusè, M.; Lazzari, F.; Mancini, G. *Journal of Chemical Theory and Computation* **2023**, *19*, 1243–1260.

(139) Ceselin, G.; Barone, V.; Tasinato, N. *Journal of Chemical Theory and Computation* **2021**, *17*, 7290–7311.

(140) Barone, V.; Ceselin, G.; Fusè, M.; Tasinato, N. *Frontiers in Chemistry* **2020**, *8*, 584203.

(141) Mendolicchio, M.; Penocchio, E.; Licari, D.; Tasinato, N.; Barone, V. *Journal of Chemical Theory and Computation* **2017**, *13*, 3060–3075.

(142) Christen, D.; Griffiths, J. H.; Sheridan, J. *Zeitschrift für Naturforschung A* **1981**, *36*, 1378–1385.

(143) Piccardo, M.; Penocchio, E.; Puzzarini, C.; Biczysko, M.; Barone, V. *The Journal of Physical Chemistry A* **2015**, *119*, 2058–2082.

(144) Jaeger, H. M.; Schaefer III, H. F.; Demaison, J.; Császár, A. G.; Allen, W. D. *Journal of Chemical Theory and Computation* **2010**, *6*, 3066–3078.

(145) Blaudeau, J.-P.; McGrath, M. P.; Curtiss, L. A.; Radom, L. *The Journal of Chemical Physics* **1997**, *107*, 5016–5021.

(146) Hariharan, P.; Pople, J. A. *Molecular Physics* **1974**, *27*, 209–214.

(147) Rassolov, V. A.; Pople, J. A.; Ratner, M. A.; Windus, T. L. *The Journal of Chemical Physics* **1998**, *109*, 1223–1229.

(148) Binning Jr, R.; Curtiss, L. *Journal of Computational Chemistry* **1990**, *11*, 1206–1216.

(149) Francl, M. M.; Pietro, W. J.; Hehre, W. J.; Binkley, J. S.; Gordon, M. S.; DeFrees, D. J.; Pople, J. A. *The Journal of Chemical Physics* **1982**, *77*, 3654–3665.

(150) Hariharan, P. C.; Pople, J. A. *Theoretica chimica acta* **1973**, *28*, 213–222.

(151) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *The Journal of Chemical Physics* **1972**, *56*, 2257–2261.

(152) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *The Journal of Chemical Physics* **1971**, *54*, 724–728.

(153) Bermúdez, C.; Mata, S.; Cabezas, C.; Alonso, J. L. *Angewandte Chemie* **2014**, *126*, 11195–11198.

(154) Barone, V.; Fusè, M. *The Journal of Physical Chemistry A* **2023**.

(155) Kojima, T. *Journal of the Physical Society of Japan* **1960**, *15*, 284–287.

(156) Nesvadba, R.; Studeckỳ, T.; Uhlíková, T.; Urban, Š. *Journal of Molecular Spectroscopy* **2017**, *339*, 6–11.

(157) Sanz, M. E.; Cabezas, C.; Mata, S.; Alonso, J. L. *The Journal of Chemical Physics* **2014**, *140*, 05B619_1.

(158) Goerigk, L.; Grimme, S. *The Journal of Physical Chemistry A* **2009**, *113*, 767–776.

(159) Bannwarth, C.; Ehlert, S.; Grimme, S. *Journal of Chemical Theory and Computation* **2019**, *15*, 1652–1671.

(160) https://ambermd.org/AmberTools.php.

(161) Barone, V.; Biczysko, M.; Bloino, J.; Puzzarini, C. *Physical Chemistry Chemical Physics* **2013**, *15*, 1358–1363.

(162) Csaszar, A. G. *Journal of the American Chemical Society* **1992**, *114*, 9568–9575.

(163) O'Boyle, N. M.; Vandermeersch, T.; Flynn, C. J.; Maguire, A. R.; Hutchison, G. R. *Journal of Cheminformatics* **2011**, *3*, 1–9.

(164) Goto, H *TechConnect Briefs* **2003**, 32–35.

(165) Pracht, P.; Bohle, F.; Grimme, S. *Physical Chemistry Chemical Physics* **2020**, *22*, 7169–7192.

(166) Olsson, A.; Sandberg, G.; Dahlblom, O. *Structural safety* **2003**, *25*, 47–68.

(167) Sanz, M. E.; López, J. C.; Alonso, J. L. *Physical Chemistry Chemical Physics* **2010**, *12*, 3573–3578.

(168) McLean, A.; Chandler, G. *The Journal of Chemical Physics* **1980**, *72*, 5639–5648.

(169) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *The Journal of Chemical Physics* **1980**, *72*, 650–654.

(170) Szidarovszky, T.; Czakó, G.; Császár, A. G. *Molecular Physics* **2009**, *107*, 761–775.

(171) Cossi, M.; Rega, N.; Scalmani, G.; Barone, V. *Journal of Computational Chemistry* **2003**, *24*, 669–681.

(172) Albertini, N.; Baldini, J.; Dal Pino, A.; Lazzari, F.; Legnaioli, S.; Barone, V. *Heritage Science* **2022**, *10*, 71.

(173) Stephenson, N., *Snow crash: A novel*; Spectra: 2003.

(174) Nath, K.; Dhar, S.; Basishtha, S. In *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*, 2014, pp 86–89.

(175) Wang, Q.; Li, R.; Wang, Q.; Chen, S. *arXiv preprint arXiv:2105.07447* **2021**.

(176) Meta Quest 2, https://www.meta.com/it/quest/products/quest-2/.

(177) Unity Engine, https://unity.com.

(178) Martino, M; Lazzari, F; Tasinato, N; Barone, V In *IOP Conference Series: Materials Science and Engineering*, 2020; Vol. 949, p 012020.

(179) Martino, M.; Salvadori, A.; Lazzari, F.; Paoloni, L.; Nandi, S.; Mancini, G.; Barone, V.; Rampino, S. *Journal of Computational Chemistry* **2020**, *41*, 1310–1323.

(180) Laane, J. *The Journal of Chemical Physics* **1969**, *50*, 1946–1951.

(181) Durig, J.; Willis Jr, J. *Journal of Molecular Spectroscopy* **1969**, *32*, 320–342.

(182) Durig, J.; Lafferty, W.; Kalasinsky, V. *The Journal of Physical Chemistry* **1976**, *80*, 1199–1202.

(183) Durig, J.; Natter, W.; Kalasinsky, V. *The Journal of Chemical Physics* **1977**, *67*, 4756–4759.

(184) Goerigk, L.; Grimme, S. *Journal of Chemical Theory and Computation* **2011**, *7*, 291–309.

(185) Papajak, E.; Leverentz, H. R.; Zheng, J.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2009**, *5*, 1197–1202.

(186) Papajak, E.; Truhlar, D. G. *Journal of Chemical Theory and Computation* **2010**, *6*, 597–601.

(187) Casavecchia, P. *Reports on Progress in Physics* **2000**, *63*, 355.

(188) Wiesenfeld, L.; Thi, W.-F.; Caselli, P.; Faure, A.; Bizzocchi, L.; Brandão, J.; Duflot, D.; Herbst, E.; Klippenstein, S. J.; Komatsuzaki, T., et al. *arXiv preprint arXiv:1610.00438* **2016**.

(189) Rampino, S.; Skouteris, D.; Laganà, A. *Theoretical Chemistry Accounts* **2009**, *123*, 249–256.

(190) Rampino, S.; Skouteris, D.; Laganà, A. *International Journal of Quantum Chemistry* **2010**, *110*, 358–367.

(191) Douglas, A.; Herzberg, G *The Astrophysical Journal* **1941**, *94*, 381.

(192) Olah, G. A., *A life of magic chemistry: Autobiographical reflections of a nobel prize winner*; John Wiley & Sons: 2002.