UNIVERSITY OF GOTHENBURG

DEPARTMENT OF PHYSICS

Opto-vibrational coupling in molecular solar thermal storage materials: Electronic structure calculations and neural-network-based analysis

Giannis
Kostaras

Degree project for Master of Science (60 hec) with a major in Physics
2020, 60 HEC
Second Cycle

Opto-vibrational coupling in molecular solar thermal storage materials: Electronic structure calculations and neural-network-based analysis
Giannis Kostaras
Department of Physics
University of Gothenburg

Abstract

Molecular solar thermal storage materials are proposed as a clean, renewable energy solution for a world with ever increasing energy needs. Norbornadiene is an organic compound suitable for molecular solar thermal storage systems. Computational methods such as density functional theory offer solutions for improvement of norbornadiene-based molecular solar thermal storage systems via theoretical spectroscopy. Machine learning methods, such as artificial neural networks may offer useful insights to improve theoretical spectroscopy methods.

Keywords: molecular solar thermal storage, electronic structure calculations, density functional theory, machine learning, artificial neural networks

Acknowledgements

I would like to thank Prof. Paul Erhard, my supervisor, for his work and guidance. This thesis project is a result of his vision and expertise.

I would like to thank the members of the Computational Materials Group at Chalmers for their help and their colleagueship.

I would also like to thank Prof. Cathy Horellou, she has been extremely helpful throughout my studies at the University of Gothenburg as a director of my Master's program.

Lastly, I would like to thank my family. My parents, Georgios Kostaras and Nikoletta Christodoulopoulou-Kostara have always been tremendously supportive in my effort in every way possible.

Giannis Kostaras, Gothenburg, 2020

Contents

List of Figures

Glossary

DFT density functional theory. vii, xi, 2, 3, 8, 15

MBTR many-body tensor representation. v, 15, 39
MD molecular dynamics. 6
ML machine learning. v, 2, 3, 7, 8, 11, 14, 15
MOST molecular solar thermal. v, vii, 1–5, 8
MSE mean squared error. 33–37, 39–41

SOAP smooth overlap of atomic positions. v, 15, 39

TDDFT time-dependent DFT. vii, 2, 5, 6, 15

1
Introduction

The ever increasing need for energy is considered a major challenge for humanity during this century. One of the problems that science and technology need to solve within this scope is the demand for clean, renewable energy sources. Many proposed solutions employ solar energy as a viable solution to replace fossil fuel consumption, wherever it is possible to do so [6]. One solar energy-based renewable energy solution is the utilization of MOST systems, in which the solar energy is stored and released on demand as thermal energy [7]. The energy storage and conversion is achieved by the photoisomerization of a compound to a strained isomer and the reverse exoergic isomerization, usually by applying a heterogeneous catalyst. The final step in this process releases thermal energy, which can be used for various applications, such as heating water when not enough sunlight is available, heating houses, vehicles etc. [8]

One organic compound suitable for this application is norbornadienne. The MOST system based on the norbornadienne-quadricyclane photoisomerization cycle is widely studied. In this system, norbornadiene, when irradiated turns into quadricyclane, its strained isomer. Quadricyclane thus stores the irradiation energy and remains kinetically stable. The stored energy can be released as thermal energy on demand through the inverse reaction, usually by applying a catalyst or by electrochemical means. However, the incident light required to achieve the photoisomerization of norbornadiene to quadricyclane must have a wavelength in general less than 300 nm (about 4.1 eV), which renders it impractical since the spectral peak of sunlight at sea level occurs around 500 nm (about 2.5 eV) [9]. Together with the relatively low quantum yield of the absorption process, this makes the pure norbornadiene-quadricyclane system an unsuitable thermal photoswitch candidate to act as a MOST material.

This could lead the research to a conclusion that it is not possible for such a molecular system to be a good MOST candidate, but ongoing research intends to counter these problems. A practical photoswitch needs to be able to undergo the transformation from the low to high strain energy compound by absorbing light in the visible spectrum. The sunlight exhibits its spectral peak on the surface of the earth in these wavelengths (500 nm or about 2.5 eV as mentioned). This means that the onset of light absorption by a chemical compound should be at least in this region instead of ultraviolet to achieve the efficiency required for MOST applications. This issue is solved by substituted norbornadienne compounds [10]. Substitution and addition of chemical groups to the original norbornadienne molecule can cause a red shift of the onset of absorption, which enables the isomerization of norbornadiene to quadricyclane when exposed to sunlight [10, 11].

Several substituted compounds in the family of norbornadienne have been investigated, both experimentally and computationally. Substituted norbornadiene-based MOST systems have been shown to perform better in visible wavelengths than unsubstituted norbornadiene [10, 11]. Ab initio computational methods reproduce these results and also offer some insight to the quantum mechanical aspects of light absorption [7]. Specifically, research can show which molecular side groups and structures are more suitable. This can lead to better choices of substituted molecules [12].

One challenge that ab initio calculations pose is the computational complexity which makes them time and resource consuming. It has been noted [13] that molecules that have similar features will show correlated results when quantum mechanical computations are performed. This is the reason why ML has been proposed as an alternative to computer time intensive computations such as DFT. Given enough output, a ML model should be able to pick up these correlations and predict spectral features of different molecules, thus accelerating the effort to engineer better compounds with the desirable properties needed for a successful MOST system.

In the case of norbornadiene, ML models can be trained with already existing computational spectroscopy data from DFT/TDDFT calculations performed for different molecular variants [3]. There are many publications by the Computational Materials research group at Chalmers University of Technology which exhibit the results of DFT/TDDFT calculations for the development of theoretical spectroscopic models of norbornadiene variants, alongside experimental spectroscopy performed at the Department of Applied Chemistry at Chalmers. The absorption spectra of various substituted systems are well studied, providing us with insights about the choice of donor/acceptor groups. Furthermore, different parameters have been used in DFT calculations, such as exchange-correlation functionals and their results have been studied in comparison to the experimental spectroscopy. The suitability of these methods has been established, but increasing accuracy can be time demanding [7]. These studies have given a better understanding of the physical mechanisms and properties that affect the absorption spectra of molecules in the norbornadiene family. With the amount of computational data available, we can try to train ML models to predict the spectral features, especially the red shift and correlate it with specific molecular traits, such as the orientation of an aryl group etc. Instead of relying on computationally expensive and time-consuming algorithms, a trained model could be sufficiently accurate in guiding the research to the best possible optimizations of the molecular systems studied here. If it can predict which substituted norbornadiene variants exhibit the most desirable red shift and quantum yield, it can be applied as a first test for new substituted molecules that are going to be tested in the future.

2
Background and motivation

In this chapter, the basic theory behind this thesis is discussed. Further information on the norbornadiene-quadricyclane molecular solar thermal (MOST) systems is presented. Theoretical spectroscopy with computational methods, specifically DFT is described. Finally, the reasoning behind the use of ML and specifically artificial neural networks is discussed, alongside with background theory and previous research of ML on quantum chemistry.

2.1 MOST systems

The ever increasing energy needs of our society have led to an increase in usage of fossil fuels. By 2019, the amount of coal, gas and oil consumed as fuel reached a rate of 11743 Mt/year. As time progresses, the need for energy production increases. Today, it is obvious that the sustainable economic growth and prosperity of humanity cannot be based on the combustion of fossil fuels because they are (1) a limited resource and (2) CO_2 production happens at a rate that results in its accumulation in the Earth's atmosphere. These are the main reasons behind the search for alternative energy sources, especially those that are renewable, like solar energy.

Solar energy can be converted to electric energy, but it can also be stored as heat. An important part of our everyday energy needs is dedicated to temperature regulation (in the E.U. countries for example, about half of the produced energy is consumed for temperature regulation [14]). This is an important motivation behind the development of MOST systems.

These systems act as molecular photoswitches by absorbing sunlight and can be converted to their original form by emitting the stored energy as heat. This heat can then be used for applications like heating the interior of buildings or heating water. They can be irradiated by sunlight so that they undergo a reversible transformation process and store the absorbed solar energy. Then, they can be stored in liquid form inside a storage tank (Figure 2.1).

A typical chemical process in molecules that act as photoswitches is photoisomerization. Examples of chemical compounds that can be used in MOST systems are azobenzene, dihydroazulene/vinylheptafulvene and the norbornadiene/quadricyclane system which is studied here. [15]

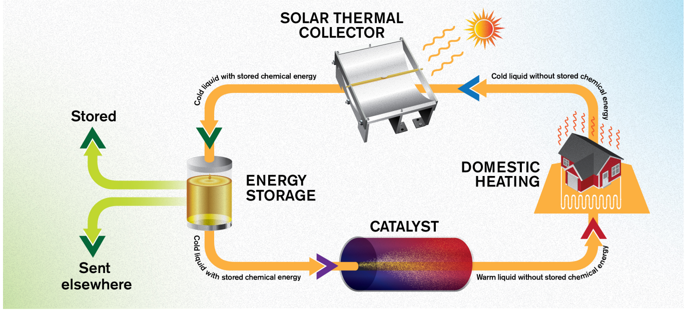
SOLAR THERMAL COLLECTOR
Cold liquid with stored chemical energy
Cold liquid without stored chemical energy
Stored
ENERGY STORAGE
DOMESTIC HEATING
CATALYST
Sent elsewhere
Cold liquid with stored chemical energy
Warm liquid without stored chemical energy

Figure 2.1: Application of a MOST system for domestic use (image taken from Chalmers' Chemistry and Chemical Engineering news [1]).

2.1.1 The norbornadiene–quadricyclane system and its variants

A well-studied example of a MOST system is the norbornadiene–quadricyclane system (N-Q), with quadricyclane as the high energy isomer. The parent molecule (norbornadiene) was named after the island of Borneo, because several compounds that occur naturally there have similar structure as norbornadiene. Norbornadiene, however, has not been found to occur in the natural environment but it was artificially synthesized for the first time in the early 1950s. Today, norbornadiene is produced with the Diels-Alder reactions of acetylenes and cyclopentadienes. The photoisomerization process (Figure 2.2) results in the highly strained molecule of quadricyclane [16].

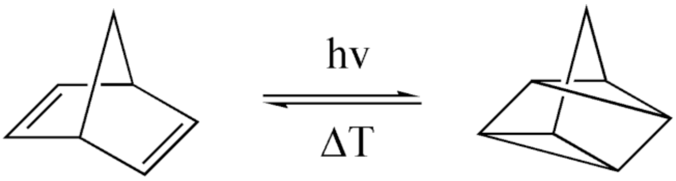

Figure 2.2: Photoisomerization of unsubstituted norbornadiene (left) to quadricyclane (right) [2].

Measurements have shown that such a system is able to store energy up to 1000 kJ kg^{-1}. This energy is a result of the high strain of quadricyclane [16]. For the reverse process of isomerization of quadricyclane to norbornadiene in the liquid state in benzene solution, the enthalpy was measured to be ~ -89 kJ mol^{-1} [17]. This energy can be released not only via a catalyst, but electrochemically as well [18] [11]. The first step of this process is the [2+2] photocycloaddition of unsubstituted norbornadiene to its valence isomer quadricyclane. The absorption edge of this process lies at 267 nm (4.64 eV). Considering the applications for which MOST systems are designed, norbornadiene is unsuitable since sunlight spectrum at sea level exhibits very low intensity for wavelengths below 300 nm [19]. Substituted norbornadiene derivatives exhibit a red-shift in the absorption onsets and maxima, which makes them better candidates for a MOST system.

Considering the better response of the compound to the sunlight spectrum, many different norbornadiene substitutes have been examined both experimentally and with the aid of computational methods, in order to classify their physico-chemical properties such as absorption spectra and find the compounds with the most desirable properties for application in MOST systems.

2.2 Theoretical spectroscopy

Improvements in MOST systems require means of improving the spectral features of the substituted norbornadiene compounds used in them. For this reason, spectroscopy plays a central role in testing the performance of MOST systems. Spectroscopy can be performed both experimentally (via chromatography methods) [19] or computationally, via methods described in this thesis which are time and resource intensive. Their analysis and improvement can lead to better understanding of the factors that render some norbornadiene variants better suited than others as photoswitches for MOST systems.

In particular, obtaining computational absorption spectra for these compounds is a task which can be accomplished by means of TDDFT calculations (further details are provided in Chapter 2 on Methodology). TDDFT is a method regularly employed in computational quantum chemistry to determine the excitation energy spectra of molecules and consequently the absorption spectra.

However, the determination of a realistic absorption spectrum of a chemical compound is not a trivial task, for which a single determination of TDDFT roots (solutions) would be adequate. As an example, the experimentally obtained absorption spectrum of the norbornadiene variant **N3** (Figure 2.3) is compared to the computationally obtained spectrum (Figure 2.7).

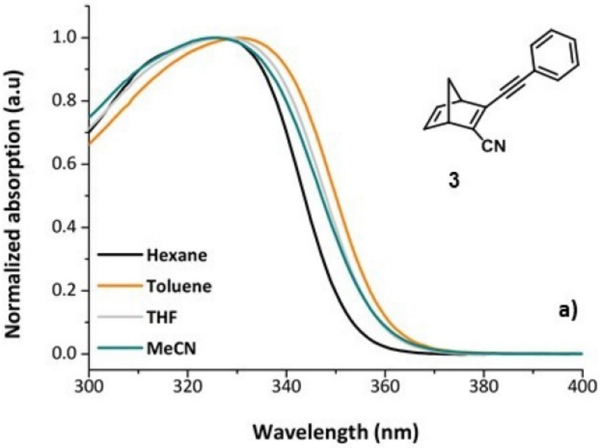

Figure 2.3: Experimentally derived spectra of **3** in various solvents. In toluene, the onset of absorption is present around 370 nm (3.35 eV). [3]

Via the computational process, the TDDFT roots of a single molecular geometry can be determined, which can provide information about the excited energy levels for this particular geometry (Figure 2.4).

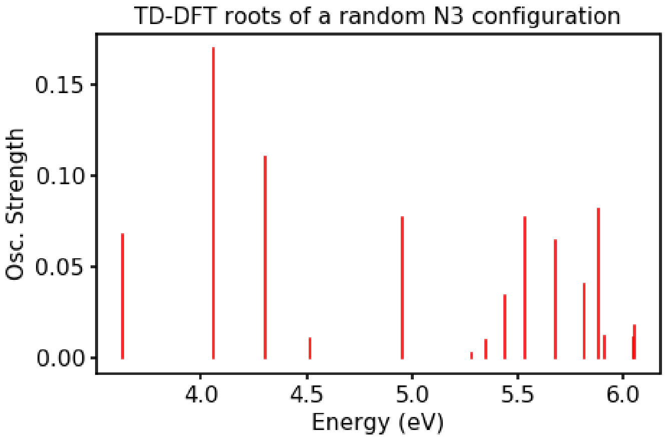

Figure 2.4: TDDFT roots of a random configuration of **3** in MeCN solvent.

In order to obtain a realistic computational spectrum, molecular dynamics (MD) simulations are performed, producing statistically decorrelated configurations of each norbornadiene variant in each solvent. After obtaining several different configurations, a large number of excitation energies computed by TDDFT is required to obtain theoretical absorption spectra. The first root is shown to be of particular importance, since it is correlated with the onset of absorption for each molecule towards the red area of the spectrum. Several such energies are computed for each configuration (Figure 2.5).

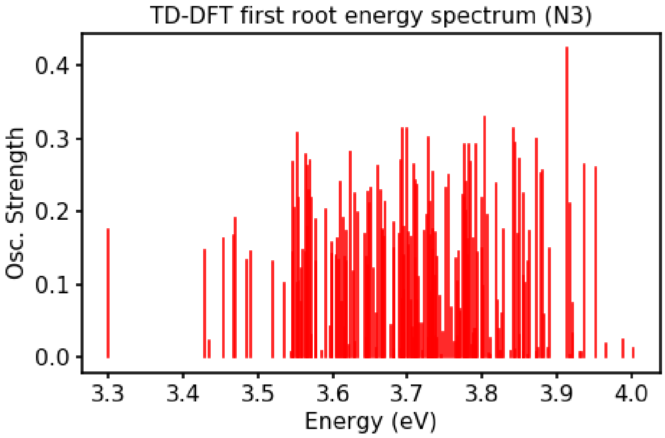

Figure 2.5: Computationally obtained roots of the HOMO-LUMO transitions of **3** in toluene, calculated by TDDFT (this sample contains 200 values in total).

The absorption spectrum is obtained by performing Gaussian broadening over the entire array of computed energy values. This approximation assumes that the real spectrum can be represented closely by a Gaussian interpolation over the absorption lines obtained by TDDFT, with the σ parameter of the Gaussian broadening adjusted empirically to 0.15 eV (Figure 2.6).
Finally, a close approximation of the absorption spectrum requires also higher order TDDFT roots, which can then give a complete theoretical spectrum (Figure 2.7)

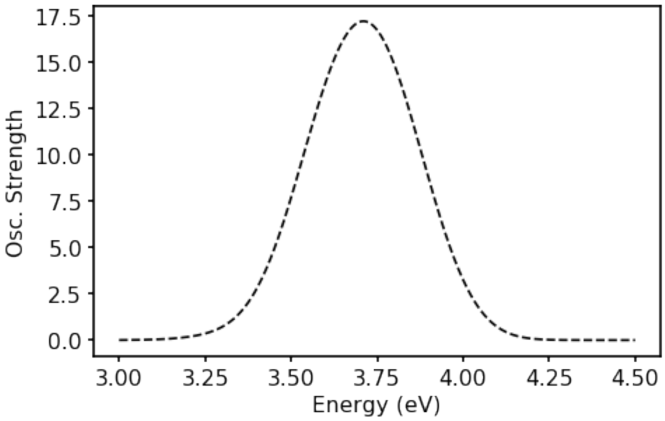

Figure 2.6: The contribution of the first order HOMO-LUMO transitions to the absorption spectrum of **3** in toluene. This transition determines the onset of absorption, which is present around 370 nm.

with some resemblance to the experimentally obtained absorption spectrum (Figure 2.3). When compared to the experimental spectrum, the computational spectra usually exhibit additional features, which can be attributed to undersampling [3]. For example, this theoretical spectrum is derived from 200 random molecular conformations of **3** in toluene.

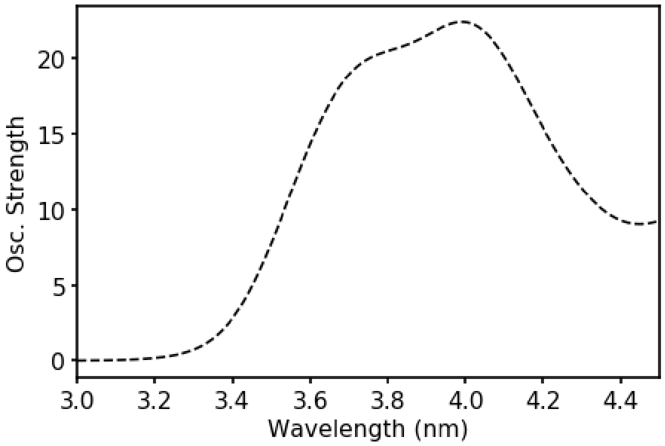

Figure 2.7: Computationally obtained spectrum of **3** in toluene.

This quick presentation of the computational procedure hopefully shows that this task is, as mentioned, not trivial and requires both time and resources to reach the final result.

2.2.1 ML for molecular quantum mechanics

It has been shown that the absorption spectra of norbornadienes can be obtained as a result of quantum mechanical computations. The main question explored in this thesis is if a ML model can be trained to predict these results with a small error compared to the ones produced by the methods described above. The motivation behind the use of ML models here is the decrease of the computational effort required. The accuracy trade-off of using ML methods is also discussed.

This trade-off is to be expected when one employs estimation methods. Approximation of solutions to quantum mechanical problems is a century-old concept [20] and ML has already offered hope for new solutions to this problem. ML methods have already exhibited some success in the prediction of atomization energies of organic molecules [21]. For this purpose, multiple ML algorithms have been tried, among them feedforward neural networks [22, 23] and convolutional neural networks [21, 24]. While the problem of predicting atomization energies from molecular geometries has been explored by past research, the prediction of light absorption spectra has not been explored to such an extent yet.

2.2.2 Past computational research on norbornadiene derivatives

The computational procedure described in this chapter and the idea for the ML model is a result of past research and experience on the norbornadiene - quadricyclane system and its variants. Current research is considering the improvement of norbornadiene-quadricyclane by exploring the photochemical properties of a number of different variants for use in MOST systems [25]. Computational research in particular focuses on this subject, with the Computational Materials Research group at the Department of Physics at Chalmers University of Technology exhibiting significant advancement in the understanding of the electronic structure and photochemical properties of different norbornadiene variants.

A first insight in the computational methods and electronic structure has been published in 2016 [7]. Electronic structure calculations offer significant information on the mechanisms influencing the properties of the studied compounds. The relationship between the light absorption and the geometric features of the molecules has been established, especially the influence of the aryl group and the π orbital coupling angles on the absorption profile of norbornadienes. Several DFT functionals have been also assessed, documenting the effect of the underestimation of the calculated energies which has been taken into account in this work.

At the same time, a first computational insight of the effect of donors and acceptors on the absorption spectrum of norbornadiene derivatives was achieved. The advantages of high solubility and large concentrations of specific norbornadiene derivatives was also taken into account together with their molar attenuation coefficient, providing a guide to engineer improved MOST systems in the future. Also, the effect of molecular mass on the light absorption capabilities of different compounds has been shown to be of great importance [12]. These observations led to the investigation of norbornadiene derivatives with small molecular mass [19]. These compounds can achieve a high energy storage density, unlike compounds with higher molecular mass, leading to a trade-off between desirable light absorption capability and energy storage density [12].

Further research has explored norbornadiene derivatives in liquid form [26] and broadening the energy gap required for the back-conversion process of quadricyclane to norbornadiene to achieve higher stability required for prolonged energy storage [11]. The addition of dithiafulvene as an improved electron donor group has also been explored computationally in an effort to decrease the S_0-S_1 band gap [27]. Lastly, an important research for this thesis project was published in 2019 exploring the effect of solvents and the effect of the linker type between aryl groups and norbornadiene, how different configurations effect the excited state energies and the absorption spectrum, the static zero-Kelvin configurations and their absorption lines and the vibrational properties of several compounds [3].

3
Methodology

The implementation of the theory presented in the previous chapter in ML-based estimation of absorption spectra of norbornadiene variants is discussed in this chapter. The data is presented alongside the methods that produced it and the ML algorithms.

There is geometrical and spectral information regarding four different norbornadiene variants in the data base, each one simulated in four different solvents. The goal of the neural network is to find correlations between the geometrical data and the spectra, in order to be able to predict them with the minimum error possible.

3.1 Computational methods for theoretical spectroscopy

In this section, the standard computational methods for theoretical spectroscopy are presented and discussed. These methods include: Density Functional Theory (DFT), Time-Dependent Density Functional Theory (TD-DFT) and Molecular Dynamics (MD), which is an auxiliary method in theoretical and computational spectroscopy, required in generating a realistic spectrum.

3.1.1 Density Functional Theory (DFT) and Time-Dependent Density Functional Theory (TD-DFT)

Density functional theory is an approach to determining the correlations of many-body systems. As such, it is particularly useful in the study of molecules and other quantum mechanical systems consisting of many particles. DFT can be used to determine molecular properties and behaviors such as the electronic structure of the molecule, perform geometry optimization or calculate the normal mode vibrational frequencies of a molecule.

The main ansatz of DFT is the assumption that the aforementioned correlations (and any other property) can be represented by functionals of the ground state of the system $n_o(\mathbf{r})$, which is a function of particle position $\mathbf{r}$. DFT calculations become an optimization problem of minimizing the energy functional $E(n)$ of a system of N particles, with the constraint:

$$\int d^3rn(\mathbf{r}) = N \tag{3.1}$$

In this context, an additional proposition, the Kohn-Sahm ansatz, can replace the problem of interacting particles with a one-particle system described by an exchange-correlation functional. The computation then employs a suitable basis set which describes the pseudopotentials of the particles. In computational quantum chemistry and physics problems, the Kohn-Sahm approach successfully incorporates the kinetic energy of electrons and the interaction terms as functionals of density (Hohenberg-Kohn theorems). It has enjoyed success in this field of research and has led to a large number of related publications.

Time-Dependent Density Functional Theory (TD-DFT) is an extension of DFT by adding a time parameter to the Hohenberg-Kohn theorems. This theory is applicable in electronic excitations. Calculations of electronic excitations need to be performed in order to derive information relevant to the absorption and emission spectra of chemical compounds. Especially when studying molecules in dielectric environments (e.g. in solvents), the time evolution of the system leads to interactions between electrons which alter the bound states of the system and the excited state energies. Experimental research provides proofs of these interactions, and TD-DFT offers insights to the mechanisms and calculations, a problem which has also been studied in norbornadienes [3].[28][29]

3.1.1.1 Functionals and basis sets

There are many options among different functionals and basis sets in DFT/TD-DFT. The correct choice depends on the desired accuracy, the problem and the system studied.

Functionals can be categorized according to their degree of physical approximation, with the Kohn-Sham exchange-correlation functional $E_{XC}(\{\psi_i\})$, being on the highest order of physical approximation. On the lowest order, the Local Density Approximation (LDA) functional approximates a local electron density (e.g. in a molecular bond) as having the same density with a spatially uniform electron cloud. Above the Local Density Approximation on the scale of accuracy and cost in computational resources comes the Generalized Gradient Approximation (GGA), which does not treat the electron cloud density as approximately uniform. In GGA functionals, both the local electron density $n(\vec{r})$ and the density gradient $\nabla n(\vec{r})$ are taken into account. Two common examples of GGA functionals are the PW91 and the PBE functionals. Higher in order of accuracy, meta-GGA functionals also include the divergence of the gradient of the electron density $\nabla^2 n(\vec{r})$. TPSS is an example of a functional belonging in this category. Finally, hyper-GGA functionals are more commonly used and make use of the Kohn-Sham orbitals, which are a part of the Kohn-Sham theory and contain information of the interaction between particles.

The functional used in the research relevant to this work is the B3LYP exchange-correlation functional, which is a hybrid functional (contains the exact exchange energy $E^{exchange}$ derived from the Kohn-Sham orbitals). B3LYP combines other functionals as well, according to the formula:

$$V_{\mathrm{XC}}^{\mathrm{B3LYP}} = V_{\mathrm{XC}}^{\mathrm{LDA}} + \alpha_1(E^{\mathrm{exchange}} V_{\mathrm{X}}^{\mathrm{LDA}}) + \alpha_2(V_{\mathrm{X}}^{\mathrm{GGA}} V_{\mathrm{X}}^{\mathrm{LDA}}) + \alpha_3(V_{\mathrm{C}}^{\mathrm{GGA}} V_{\mathrm{C}}^{\mathrm{LDA}}) \tag{3.2}$$

$V_{\mathrm{X}}^{\mathrm{GGA}}$ is the Becke 88 exchange functional, $V_{\mathrm{C}}^{\mathrm{GGA}}$ is the Lee-Yang-Parr correlation functional and the α parameters are empirically chosen for performance optimization.

B3LYP has been very successful and is widely used in research involving DFT calculations. It must be noted that all functionals produce systemic errors when compared to the experiment. B3LYP alongside other functionals has been benchmarked in calculations on norbornadiene also [7].

Hybrid functionals, such as B3LYP, are not used in delocalized systems due to difficulties in numerical calculations, a problem which is solved by using localized basis sets.

The choice of a basis set is another important topic in computational chemistry. The electronic wave functions are represented as an expansion of *basis set functions* ϕ_i, with larger basis sets containing more ϕ functions, resulting also in higher accuracy but requiring more computational resources. Basis set functions contain electronic structure information which can be specific to different kind of problems, with localized basis set being more efficient in molecular calculations rather than bulk materials.[28][29]

3.1.2 Molecular Dynamics (MD)

The derivation of realistic theoretical absorption spectra requires TD-DFT calculations for a wide array of different molecular configurations in order to simulate the material under investigation at a temperature other than $0\,{}_{o}\mathrm{K}$. The production of decorrelated molecular configurations requires simulation of trajectories of moving atoms, a task performed by Molecular Dynamics (MD) simulations.[28]

In MD simulations, a system of N particles (atoms) is represented by $3N$ positions and velocities. The particles are subjected to forces and obey Newton's laws of motion. Then, a system of $6N$ first order differential equations are solved, offering information on the evolution of the system over time and additional information, such as the temperature of the system, can be calculated

$$\frac{1}{2}m\bar{u}^2 = \frac{3k_bT}{2} \tag{3.3}$$

where $\frac{1}{2}m\bar{u}^2$ is the average kinetic energy of each degree of freedom, k_b is the Boltzmann constant and T is the temperature of the system.

3.1.3 DFT for normal mode analysis

An important application of DFT is normal mode analysis. The specifics of how the results of this analysis are important in the work of this thesis are described later in the Methodology chapter. It is noted that norbornadiene variants are molecules studied in a non-zero temperature environment, thus their vibrational properties affect the geometries and the related excited state energies of the molecular configurations.

In general, DFT for vibrating molecules that interact with other atoms (solvents in the context of norbornadiene variants) involves the location of a point $\vec{r_0}$ where the energy is minimum, the atomic energy can be written as a Taylor expansion with second order terms $\partial^2 E$. The Hessian matrix of this second-order partial derivatives can be written according to the new coordinate system $\vec{x}$ with $\vec{x_0}$ as origin:

$$H_{ij} = \left[\frac{\partial^2 E}{x_i x_j}\right] \tag{3.4}$$

By calculating the force related to the i-th coordinate, F_i, a classical equation of motion can be written for this coordinate. It is related to an acceleration by the equality $F_i = m_i(d^2x_i/dt^2)$ and to the energy of the system of atoms $F_i = \partial E/\partial x_i$, and in matrix form the equation of motion of the system becomes

$$\frac{d^2\boldsymbol{x}}{dt^2} = \boldsymbol{A}\boldsymbol{x} \tag{3.5}$$

$\boldsymbol{A}$ is the mass-weighted Hessian matrix $Aij = H_{ij}/m_i$.

The eigenvectors of the mass-weighted Hessian matrix lead to a special set of solutions of these equations for which the displacements point along the eigenvectors and the amplitude shows a harmonic oscillation, defined by the related eigenvalue of the Hessian. These special solutions are the normal modes [28].

3.2 ML examination for norbornadiene variants absorption spectra

3.2.1 Vibrational modes, side-group rotations and excitation energies: the choice of molecular descriptors

The correlation between vibrational modes, side-group rotations and electronic excitation energies is a central part of this thesis. In principle, a ML mode can pick up these correlations and “learn” how to predict them on its own in various scenarios. In order to investigate this idea, an investigation beyond simple molecular geometry correlations is performed. In this investigation, the fact that specific bond angles play an important role in the electronic properties of the molecules is taken into account.

Therefore, there is a need to represent the molecular configurations in a meaningful way for a ML model. For this purpose, one must choose the correct molecular representation, also known as molecular descriptor, which acts as a mathematical representation of molecular features for the model.

Two common molecular descriptors are the Cartesian matrix and the Z-matrix of internal coordinates. The Cartesian matrix consists of four columns and as many rows as the number of atoms of the represented molecule. The first column contains the symbols of each atom and the other three columns contain its three spacial coordinates. In total, it contains $3n$ coordinates for a molecule with n atoms. The Z-matrix has the same first column as the Cartesian matrix. The second column in a Z-matrix contains the distance between the atom of each row to the atom of the previous row ($n-1$ distances in total). The third column contains $n-2$ torsion angles between triplets of atoms and the fourth column contains $n-3$ dihedral angles between planes of atoms [30]. This representation offers some advantages over

the Cartesian matrix: it is a reduced coordinate system invariant to rotations and translations, it contains $3n-6$ instead of $3n$ coordinates, and it contains information on angles between atoms, which are related to the problem investigated in this thesis. The dimensionality reduction that it offers is expected to lead to better training and better predictions in machine learning models, compared to the Cartesian matrix.
Both of these descriptors incorporate geometrical information of the molecular configurations in a straightforward way. The Cartesian matrix and the Z-matrix can be thought of as simple molecular descriptors and the neural network as a complex system with a behavior which can converge in such a way that complex correlations can arise out of data, even though the descriptor is "simple". By adding more specific geometrical information, the data is engineered in such a way that it encompasses higher complexity.
Before the performance of frequency analysis, DFT geometry optimization is performed via NWChem, with the option of tight convergence. Then, frequency analysis is performed but with the option of more accurate convergence of the positions, requiring the maximum not to exceed 10^{-8} eV/Å. Finally, TDDFT calculations are carried out for the singlet and triplet excited state energies and oscillator strengths. The correlation between vibrational modes and excited state energies can be studied as a model of more complex molecular descriptors with a linear correlation with the energies.
The study of these correlations is first performed by DFT frequency analysis and then the neural networks are assessed versus a linear model, regarding their efficiency to pick up these correlations.

3.2.2 Linear regression with SOAP and MBTR molecular descriptors

SOAP and MBTR are two molecular descriptors involved in the ML analysis of the correlation between normal modes and excited energy levels in the context of this thesis work. In general, they are more complex descriptors than the Cartesian matrix or the Z-matrix and their generation from molecular configurations is a more involved process. Their implementation is done via the DScribe Python package of molecular descriptors [31].
The SOAP descriptor is formed by implementing information about the position $\mathbf{r}$ of atoms, the Gaussian smoothed atomic density $\rho^Z(\mathbf{r})$, spherical harmonics $Y_{lm}(\theta, \phi)$ and the radial basis function $g_n(r)$ in the coefficients $c^Z_{nlm}(\mathbf{r})$:

$$c^Z_{nlm}(\mathbf{r}) = \iiint_{R^3} dV g_n(r) Y_{lm}(\theta, \phi) \rho^Z(\mathbf{r}) \tag{3.6}$$

Then, the final output is the partial power spectrum vector $\mathbf{p}(\mathbf{r})$ with elements that are defined as:

$$p(\mathbf{r})^{Z_1 Z_2}_{nn'l} = \pi \sqrt{\frac{8}{2l+1}} \sum_m c^{Z_1}_{nlm}(\mathbf{r})^* c^{Z_2}_{n'lm}(\mathbf{r}) \tag{3.7}$$

The MBTR is a descriptor that is suitable both for molecules and periodic representations and encodes structural motifs. Structural patterns are transformed into scalars by taking into account some or all of the following properties: atomic species, distances and angles. These functions can be then weighted or left as they are and transformed into a distribution with kernel density estimation.[31, 32]

3.2.3 Data preparation for the neural network-based investigation

The training and assessment of the neural networks is based on results of a series of computational simulations performed in order to extract molecular geometry configurations (Figure 3.1) and averaged absorption spectra of the norbornadiene compounds over these geometries [3].

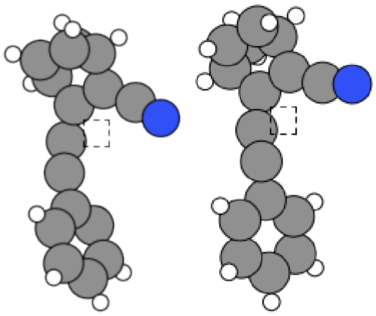

Figure 3.1: Two different configurations out of the 200 of the same compound (**3**) in MeCN solvent. The configurations are the result of Molecular Dynamics simulations. (Image produced by the Atomic Simulation Environment software [4].)

The first step for the production of these calculations was a series of molecular dynamics (MD) simulations with the GROMACS software [33]. The molecules were simulated inside a box of dimensions 5x5x5 nm. Solvent molecules were also inserted in the simulation. Then, the simulation was performed for a temperature of 300 ^oK and pressure 1000 ps. The MD simulations produced in total 800 different configurations for each molecule, with 200 configurations per solvent (Figure 3.2). The four compounds included are: **N3** (C16H11N), **N4** (C18H16N2), **N5** (C19H16) and **N6** (C14H11N). The solvents are: MeCN, toluene, tetrahydrofurane and hexane.

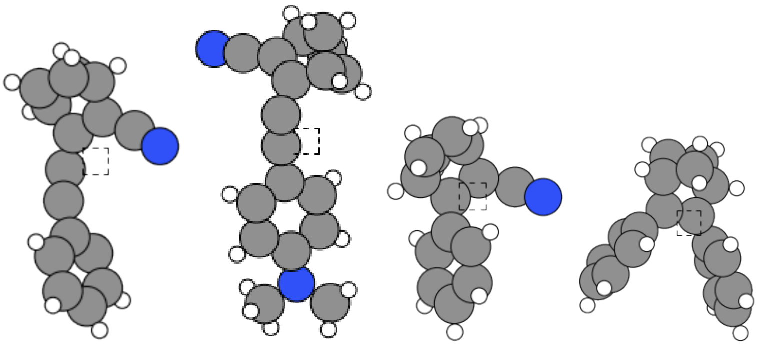

Figure 3.2: Examples of configurations of the four different norbornadiene variants included in the data base (from left to right: N3, N4, N5 and N6). The random configurations were produced by MD simulations. (Image produced by the Atomic Simulation Environment software [4].)

Quantum calculations were performed on the results of the MD simulations in order to determine their excitation energies. Time-Dependent Density Functional Theory (TDDFT) was performed for the molecular configurations with the NWCHEM software [34], in order to determine the excitation energies and the dipole strengths of the transitions. Then, after the excitation energies were calculated, they underwent Gaussian broadening and the absorption spectra were plotted.
In these calculations, the coordinate system which describes the geometries of the atomic configurations is the Cartesian matrix. The Cartesian matrix contains the positions of all atoms in the Cartesian coordinate system, according to the origin of the coordinate system as it was set in the MD simulation.
For the purpose of the work presented in this thesis, every matrix was converted to the Z-matrix reduced coordinate system. A Python script accomplishes this purpose, with the following rules: the first column contains information on the atomic species, then the distance of each atom is calculated and the second column is filled with the distances of consecutive atoms. The third column of the matrix contains the torsion and the fourth column contains the dihedral angles.

3.2.4 Neural network - based method

The central problem discussed in this thesis is the efficacy of Artificial Neural Networks (ANNs) as a regressor capable of predicting features of absorption spectra of chemical compounds that belong in the same family of molecules, in particular norbornadiene variants. Here, a brief description of ANNs and their application is presented.

3.2.5 Feedforward neural networks

Feedforward neural networks, also known as multi-layer perceptrons, are a class of algorithms that perform logistic regression and are stacked in layers, forming a network, with the final layer serving the purpose of producing the output values. The building block of this network is the McCulloch-Pitts neuron, which was invented as a model of the function of an actual biological neuron. The actual biological brain is a network of neurons, similarly an appropriately structured network of McCulloch-Pitts neurons should be able to be "trained" and "learn", so that it can infer deductions in complex classification and regression problems.
For this purpose, each neuron is connected to some inputs, which can be the input values or the output of other neurons and produces output according to the following rule:

$$n_i(t+1) = \theta\left(\sum_j w_{ij} n_j(t) - \mu_i\right), \tag{3.8}$$

where the neuron updates its output in discrete i time steps and adds input values from j inputs or neurons weighted by the *synaptic weight* w_{ij}. The calculation of the output value may also include a *threshold value* μ_i and an activation function θ [5]. In the context of this work, the Rectified Linear Unit activation function is chosen (ReLU):

$$\theta(x) = max(0, x) \tag{3.9}$$

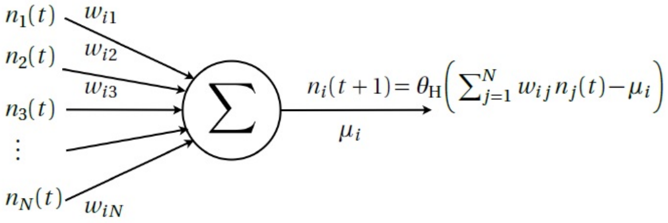

Figure 3.3: The McCuloch-Pitts neuron (image taken from Artificial Neural Networks lecture notes by B. Mehlig, University of Gothenburg [5]). in this example, the Heaviside function θ_H is used as an activation function.

3.2.6 Neural network methodology

The first category of models to study is the neural networks, in particular multi-layer perceptrons. The descriptors mentioned above, Cartesian matrices and Z-matrices, are fed as input in the neural networks, which are capable of working out the complexity of the system and correlate the geometries to the excited state energies of the molecules.
First, the matrices are turned into vectors which are fed into the input layer of the neural networks. This process is known as "flattening" of the input data and it is a necessary step in the training of the multi - layer perceptrons. For example, a Cartesian matrix of a molecule with 28 atoms (such as **3**) has a shape of 28 rows and 3 columns, for each atom and each coordinate respectively.

4

Normal mode analysis and side-group rotations

This chapter discusses the correlation between vibrational normal modes of the molecules and their spectral features. In specific, the two molecules compared here are the molecules of the norbornadiene variants N3 (C16H11N) and N5 (C14H11N), also known in literature as NBD3 and NBD2 respectively [11]. In both cases, a similar structure can be noticed: the norbornadiene part of the molecule is connected to an aromatic ring. The special feature in N3 is the presence of an ethyl linker between the two parts (Figure 4.1), whereas in N5 this feature is missing. The mobility of these two types of link is presented in existing literature [11], where it is shown that the presence of the ethyl linker offers rotational mobility between the two groups.

Consequently, the N3 molecule can obtain configurations far from the equilibrium point easier at lower temperatures than N5. Rotations away from the equilibrium position can change the interaction between the π molecular orbitals and influence the energy landscape of the S0-S1 coupling (S0 is the ground state and S1 is the first electronically excited state of the molecule) [11]. This is the reason why the absorption spectrum of the two compounds differs, with the onset of absorption for N3 exhibiting a red shift, when compared to N5. The correlation between side-group rotations and the S1 energy is studied here, which plays an important role in the onset of absorption.

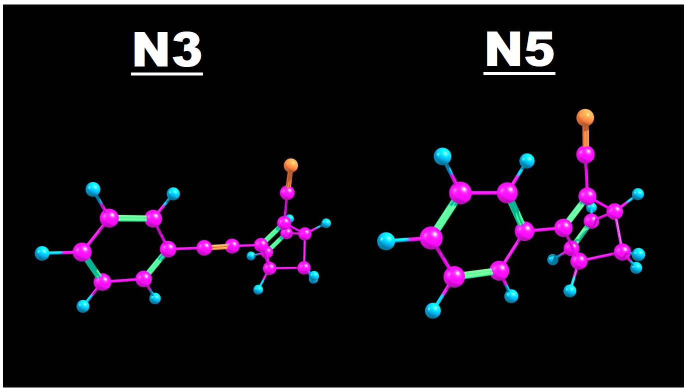

Figure 4.1: The N3 molecule features an ethyl linker between the norbornadiene part and the aromatic ring, in contrast to the N5 molecule (image produced by Chemcraft - graphical software for visualization of quantum chemistry computations. https://www.chemcraftprog.com).

4.1 Normal modes and side-group rotation

The first step towards the study of the normal molecular modes and the side-group rotations is the determination of the equilibrium geometries of both molecules (depicted in Figure 4.1).
Geometry optimization was achieved by means of Density Functional Theory, as it is implemented by the NWChem computational chemistry software. The functional of choice is the exchange-correlation B3LYP. After the completion of this task, N3 and N5 molecules are relaxed S0 ground energy level.
After the geometry optimization, frequency analysis was performed by DFT with the NWChem software. DFT frequency analysis calculated the frequencies of the normal vibrational modes of the molecules. Each molecule exhibits $3n - 3$ normal modes of vibration, where n is the number of atoms for each molecule. In the case of N3 (C16H11N) n=26, which means that there are 75 normal modes. For N5 (C14H11N) n=28, which means that there are 81 normal modes. Singlet and triplet excited state energies and oscillator strengths were then calculated with TDDFT, with XC-B3LYP exchange-correlation functional again.

4.1.1 The normal modes and the S1 excitation

The following two plots show the relation between S1 excitation energy spectrum (golden line) with some of the normal mode coordinates (blue line) for (**3**) Figure 4.2 (on the left). It can be compared to the S1 energy variance in relation to the oscillation energy (on the right).

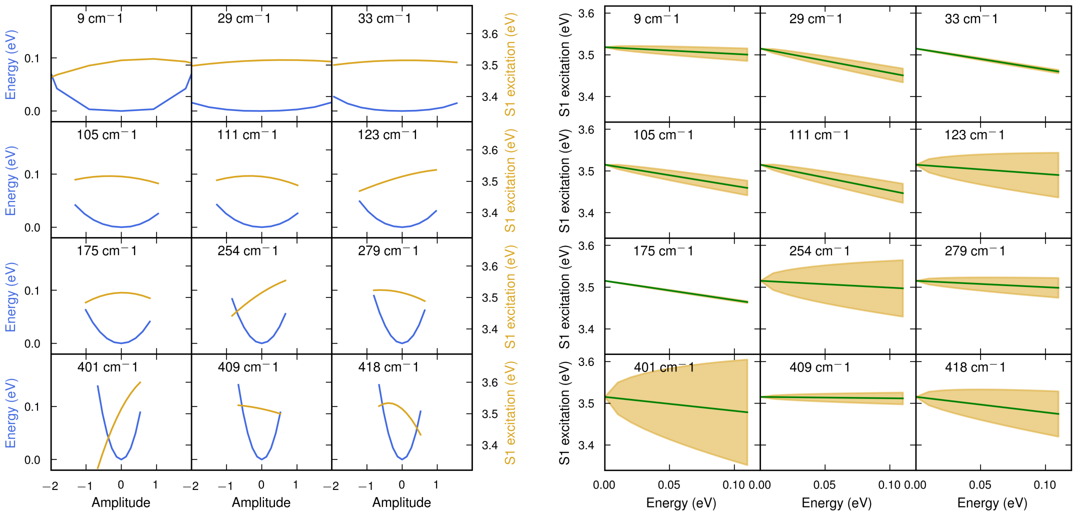

Figure 4.2: Left: N3 S1 excitation energy (in gold) and oscillator strength (in blue). Right: S1 energy variance.

An important observation is that for some normal modes, the effect of the oscillations of the molecule is not very influential on the S1 excitation energy but in other cases the difference in the oscillation direction has a dramatic impact in the S1 excitation energy. This is more obvious in **5**, with normal modes which are related to more extreme variations in the excitation energy Figure 4.3.

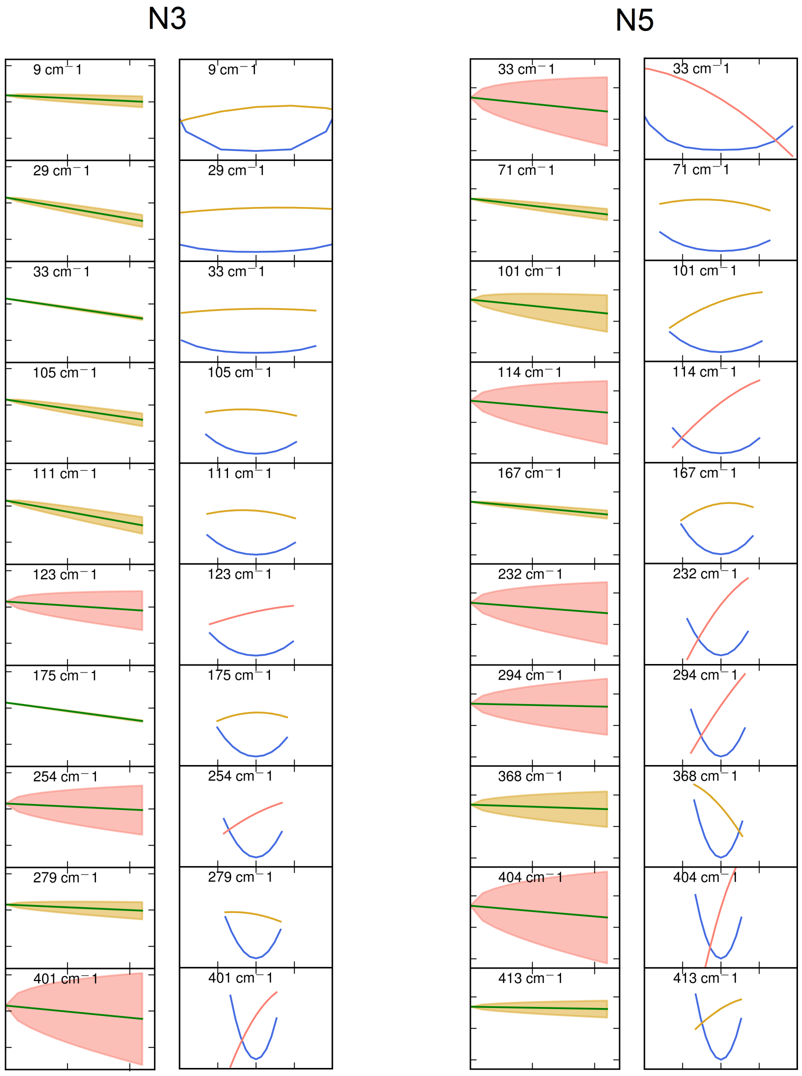

Figure 4.3: Comparison of S1 excitations and their variance along several modes for **3** and **5**. Some modes are not related to large variance of the S1 excitation energy (shown with yellow color), while others are related to greater variance (shown with light red color), thus are more important for absorption of light with larger wave length. **5** facilitates normal modes that exhibit more extreme variance of S1 excitation energy, resulting in a red shift.

Obviously, these normal modes play an important role in shaping the absorption spectrum of each molecule. A further investigation of these two categories of normal modes shows some important differences between them.

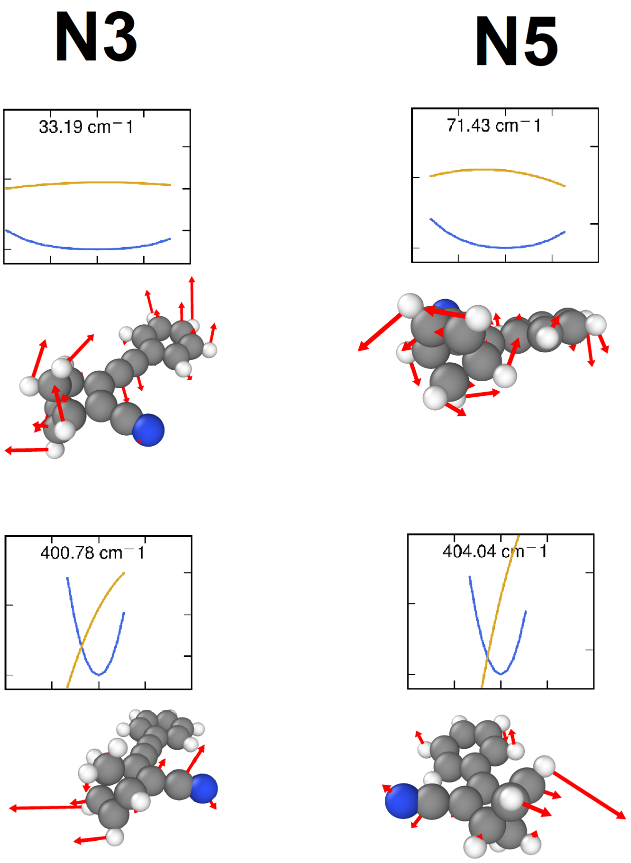

Figure 4.4: Normal modes with symmetric (first row) and asymmetric (second row) influence on the S1 excitation energy.

Examples (Figure 4.4) show that modes which affect mostly the changes in the excitation energy are the ones that break symmetries and move atoms closer and further away to one another. In **3** and **5**, the first row in Figure 4.4 shows two modes that produce a symmetric wave-like motion, where the side groups move "up" and "down". This motion does not result in a big change in the S1 excitation energy level.

The second row of Figure 4.4 shows atomic motions that deform the side group and the norbornadiene part, resulting in atoms in the norbornadiene part moving close and further away from each other. These deformations have more profound and asymmetric effect on the change of the excitation energy.

This shows how different normal modes have a different effect on the spectrum. The following plot Figure 4.5 shows the correlation between the spectral shift and broadening and the normal modes for the two molecules. The modes tend to cluster, with some outliers having a greater effect.

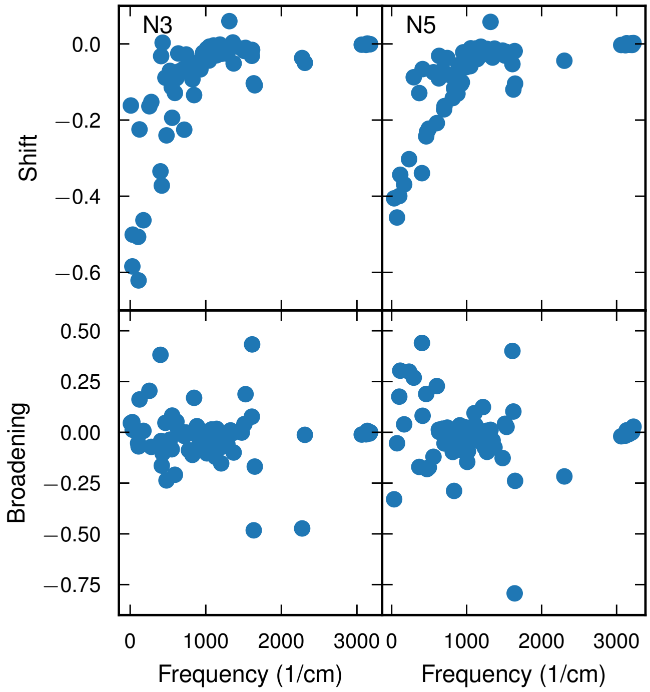

Figure 4.5: Spectral shift and broadening.

Then, we can look into specific modes and which elements of the z-matrix are mostly correlated with the most important normal modes. Defining which geometry features affect the S1 excitation energy the most can lead to development of more functional molecular descriptors. Figure 4.6 shows two "important" modes and their corresponding frequencies for **3** and **5** respectively.

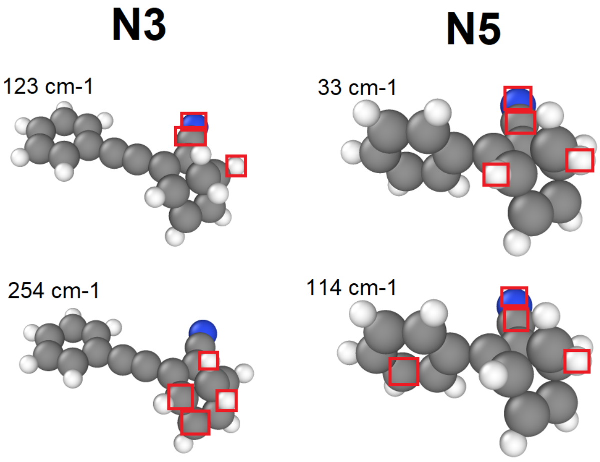

Figure 4.6: Normal modes that affect S1 excitation energy the most are related to deformations of the norbornadiene part of each compound. The highlighted atoms are the ones mostly involved in these modes.

The norbornadiene atoms are the ones that play the most important role in the onset of absorption. After all, this is the compound that research focuses on improving.

However, it has been shown that different groups and linkers between them and norbornadiene affect the absorption of light. In **3** and **5**, the difference is in the ethynyl linker. Different linkers affect the rotation of the sidegroup relative to the norbornadiene differently.

4.2 Side-group rotation and the S1 energy spectrum

The side group rotations of **3** and **5** (Figure 4.7) have been shown to effect the S1 excitation energy [3]. This phenomenon has been studied because the two compounds are very similar, but **3** features a triple bond as a linker between the norbornadiene and the aryl group (ethynyl linker), whereas **5** does not have this feature. This difference makes it easier to understand what kind of functionality the aryl side group offers to each molecule. It has been shown that rotations of the aryl side group away from the optimal geometry of each compound reduce the alignment between the highest occupied and the lowest unoccupied molecular orbitals. Side group rotations are more accessible on N3 and the distance between the two parts of the molecule is greater [3]. Despite having very similar geometries, this feature can make two molecules have different first order excitation energies. Normal mode analysis performed by DFT expands on this observation and shows how rotations and deformations influence the first excitation energy, which is a key feature in predicting the spectrum of a compound. Rotations of the side group have been linked with the onset of absorption of each compound, which is desired to be shifted to the red part of the spectrum. Consequently, it is expected that the Z-matrix, which contains information on intramolecular angles, will perform better than the Cartesian matrix and prove to be a suitable molecular descriptor. In the work for this thesis, the effect of the side-group rotations is presented in the following plots in Figure 4.8:

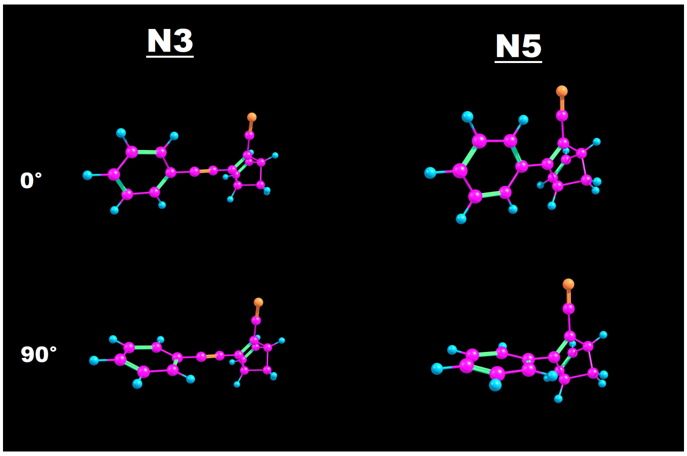

Figure 4.7: Visualization of side-group rotation by 90^o (image produced by Chemcraft - graphical software for visualization of quantum chemistry computations. https://www.chemcraftprog.com).

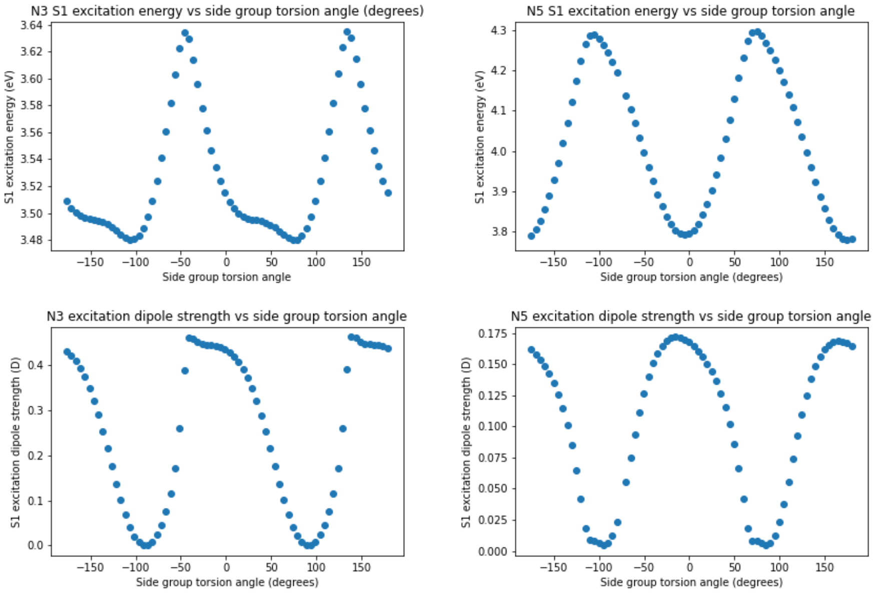

Figure 4.8: The effect of the side-group rotation on the S1 excitation energy and transition dipole strength.

4.3 Normal modes classification

The frequency distribution of the four different norbornadiene variants is presented here.

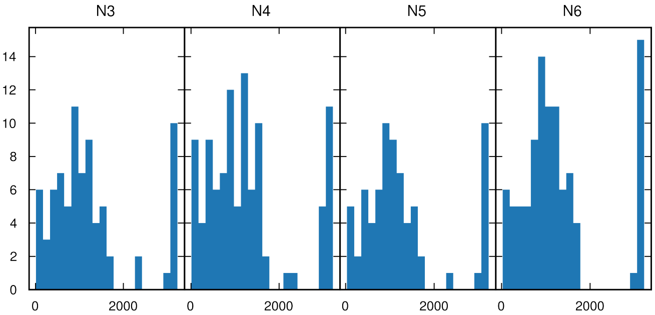

Figure 4.9: Distributions of normal modes along the frequency spectrum (in cm^{-1}).

Qualitatively, the spectrum of the normal modes shows an interesting "partitioning" of the frequency distribution, with modes that tend to cluster in separate groups according to the characteristics of the vibrations (Figure 4.10).

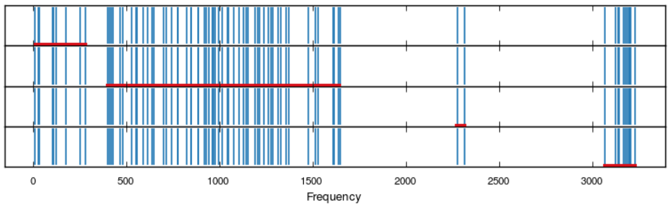

Figure 4.10: Four distinct clusters of normal modes tend to form in the N3 normal mode frequency spectrum.

The lower modes are the ones that affect mostly the excited energy levels of each molecule. These modes tend to cluster in the first group of lower frequencies and are related to the side-group rotation about the axis of the ethynyl linker (Figure 4.11):

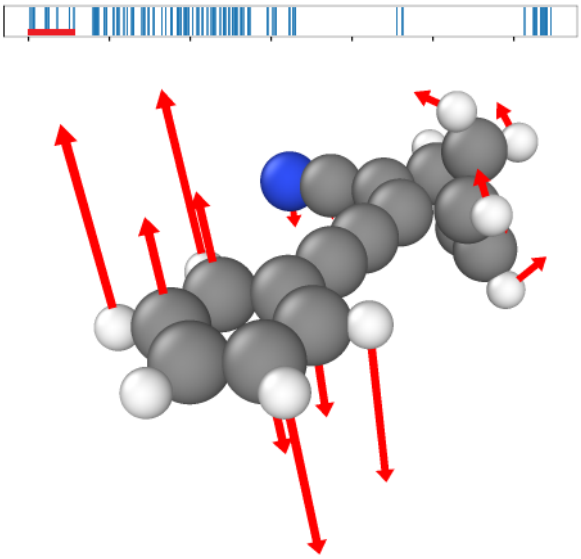

Figure 4.11: Atomic displacements when following the softest mode ($9\,\mathrm{cm}^{-1}$) in the frequency spectrum for **3**. This mode primarily involves a rotational motion of the side group about the axis of the ethynyl linker.

A look into the second cluster of modes that tend to group together shows that they are mostly related to group deformations (Figure 4.12):

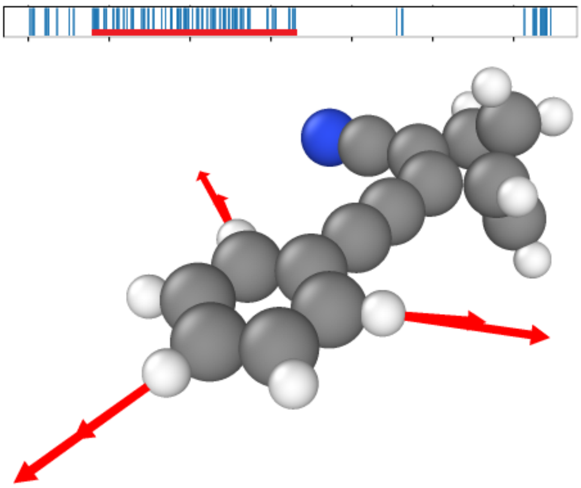

Figure 4.12: Atomic displacements exhibited in the normal mode with frequency 1015 cm^{-1} (for **3**). This mode primarily involves a deformation (stretching) of the side ring group.

A small cluster which involves only 2 modes is related to harmonic oscillation of atoms involved in the ethynyl linker (Figure 4.13):

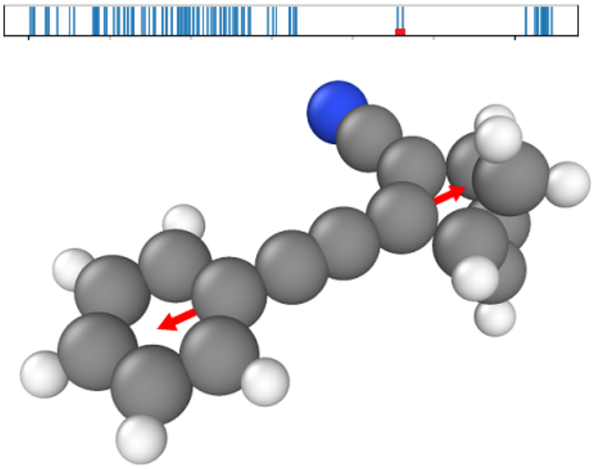

Figure 4.13: Atomic displacements exhibited in the normal mode with frequency 2275 cm^{-1} (for **3**). This mode primarily involves oscillations of the atoms along the ethynyl linker.

This small cluster of modes related to atomic displacements of ethynyl linkers is not present in **6**, which has no single linker. In **5**, where the linker is not ethynyl, only one mode is present in this category (Figure 4.9).
Finally, a look into the 4th cluster of modes of **3** shows that they are related to displacements of hydrogen atoms (Figure 4.14). These are the stiffest modes explored by DFT, and they do not affect the excited energy spectrum of the molecule as much as the softest normal modes.

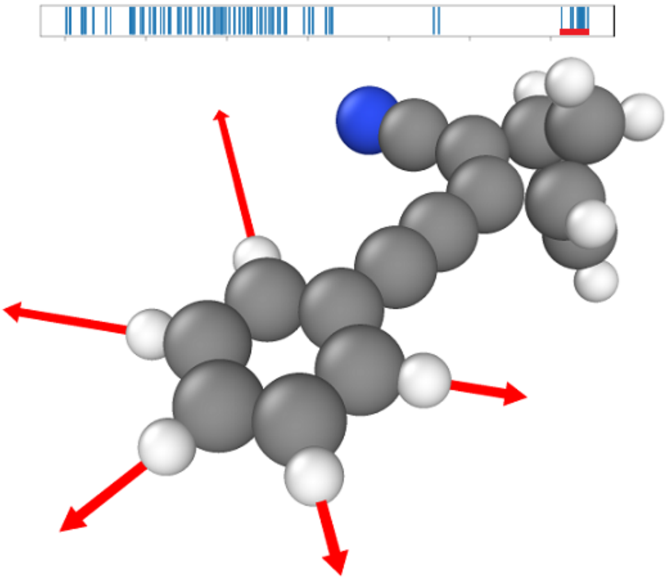

Figure 4.14: Atomic displacements exhibited in the normal mode with frequency 3199 cm^{-1} (for **3**). This mode primarily involves displacements of hydrogen atoms.

4.4 Comments

The plots presented in this chapter show generally greater S1 energy gaps for N5 in comparison to N3, both for modes with higher and lower frequencies. For N5, S1 excitation energy can rise or decline steeply along the normal mode coordinates. This can be attributed to the overlapping of orbitals, which can reduce the HOMO-LUMO gap [19].
In N5, the S1 excitation energy gap can become large for most vibration modes. This is attributed to the stronger interaction between the norbornadiene group and the aryl side group in N5, which results in more resistance to rotation.
N3 exhibits more broadening and shift in lower frequencies. Since we examine the first singlet excitation, we can derive some conclusions about the HOMO-LUMO gap. In the case of N3 it seems to be smaller. There is an explanation which can be found in literature [3]: rotations in N3 are seamless when compared to N5, due to the presence of the ethynyl linker. Apart from the S1 transition energy, this is expected to also have an effect on the dipole strength, in particular the HOMO-

LUMO transition in N5 is expected to have a smaller dipole strength.

An interesting feature of these plots is the asymmetry in the spectral peaks, which can be attributed to the asymmetry of the molecular orbitals that overlap as the side groups rotate. This feature can be picked by a machine learning model. In the next section, it is shown that a linear model based on ridge regression and molecular descriptors of higher complexity than the Z-matrix can be trained more successfully to predict the S1 energy gap than the neural network - based models explored in the next chapter.

5

Artificial Neural network-based analysis

A review of the results of the neural network-based approach is included in this chapter. Assessments of the following parameters is presented: the choice of the z-matrix as the descriptor over the xyz representation and the Coulomb matrix, the optimization algorithm, the neural network hyper-parameters, its performance in different norbornadiene variants and finally a comparison between the multi-layer perceptron-based approach and other machine learning algorithms in the scope of this thesis.

Each model is evaluated by the mean squared error acting as the loss function. The optimization algorithm used is the stochastic gradient descent, as implemented in the keras.optimizers python library with a TensorFlow 1.12.0 backend. A plot of the predicted versus true values is presented with the R^2 coefficient and the slope of the linear regression of the predicted vs true values. Ideally, if every predicted value was correct, the R^2 coefficient would be equal to unity and the slope of the regression line would be equal to unity as well, so a major criterion of the performance of the machine learning models presented here is how close these values are to the ideal.

An important issue regarding the evaluation of each neural network is the training time. The mean squared errors exhibit some variance, which means that every training session is unique and if the model actually approaches the desirable global minimum, the approximation is not always the same for each training session. For each model, an average of the mean squared error of ten unique training sessions is evaluated in order to acquire a clear picture of the model's performance. Each training session is performed with a different random split between train, test (validation) and hold-out data sets. This was necessary because the data set for each norbornadiene variant is small, with a size of only 200 data points for every compound in each solvent. A 5% split is done, producing a hold-out and a test (validation) data set of 10 data points each, so after 10 different training sessions, the plot contains 10 predictions for each one of the 10 hold-out data sets, in total 100 predictions in every plot. The test data set is used by the optimization algorithm. Generally, the training of each neural network can be time-consuming even for such small data sets, especially if the neural network contains many hidden layers (more than two).

The output value of the neural network is the first root energy (in eV) produced by TDDFT calculations. An assumption is that if a neural network can be optimized to predict this root as accurately as possible, it can perform similarly for the other roots as well (in total there are 15 roots included in the database).

5.1 Assessment of the z-matrix molecular descriptor

An important issue investigated in this thesis is the superiority of the z-matrix molecular descriptor for the study of the correlation of molecular geometries to absorption spectra over the Cartesian matrix. The neural network used here (Table 5.1) consists of one input layer with as many neurons as the z-matrix input for the molecule **3**. Then, two hidden layers which operate with the Rectified Linear Unit (ReLU) activation function and finally an output layer which consists of one neuron with no activation function, because the requested output is only the first root.

Table 5.1: The summary of the neural network model (output of Keras model.summary() command).

Layer (type)	Output shape	Number of weights
Flatten	84	0
Dense	84	7140
Dense	84	7140
Dense	1	85
Trainable parameters: 14365		

The output value is the energy of the first root (in eV). Plots of the mean squared error evaluated for the test data set over the training epochs is presented (Figure 5.1), as well as prediction plots for the training and hold-out data sets. A minimum can be seen close to 4000 epochs. In most cases the minimum appeared there, and after the 4000 epochs mark the error increased. This increase is caused by overfitting of the regression model to the training data set.

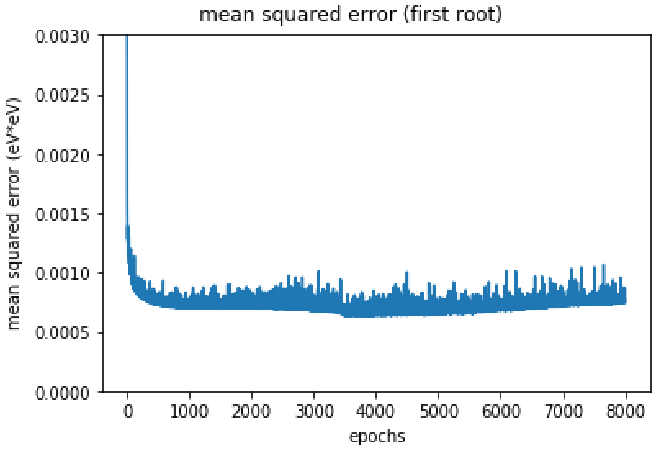

Figure 5.1: The mean squared error of the model on the test data set over the training epochs plot, with the z-matrix descriptor as input. This is a sample training session output.

Finally, the plots of predicted over true values are presented (Figure 5.2), both for the training and the holdout data set after 4000 training epochs. It is important to note that when the neural network is trained for more epochs, the mean squared error in the predictions of the training data set decreased, while the mean squared error in the predictions of the holdout data set increased, which is a sign of overfitting.

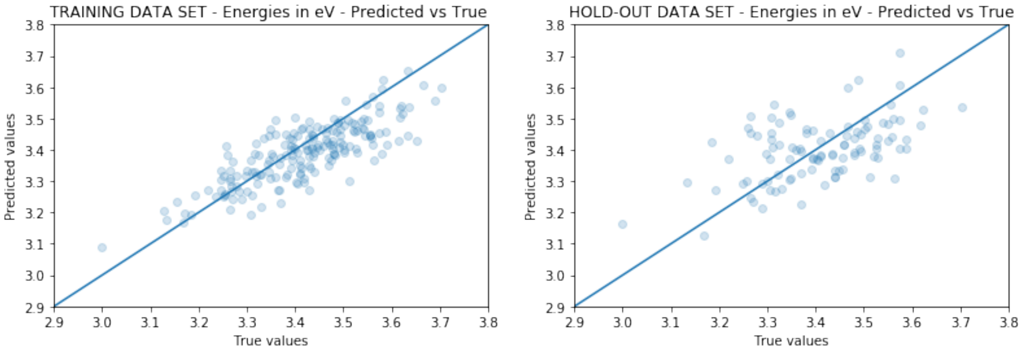

Figure 5.2: The fit of the predictions for the training data set is better than the predictions in the hold-out data set.

The average mean squared error after 10 training sessions, the slope of the line fit in the predicted over true values of the hold-out data set and the R^2 coefficient of the fit is also calculated, in order to assess the performance of the neural network.

Table 5.2: Neural network model predictive performance with z-matrix descriptor as input.

Quantity	Value
mean squared error (MSE) (eV^2)	0.000650
Slope	0.411291
Coefficient of determination R^2	0.267261

5.1.1 Comparison with the Cartesian coordinates descriptor

The same procedure was followed with the Cartesian coordinates as input. The same plots as in the z-matrix assessment are presented here (Figure 5.3, Figure 5.4).

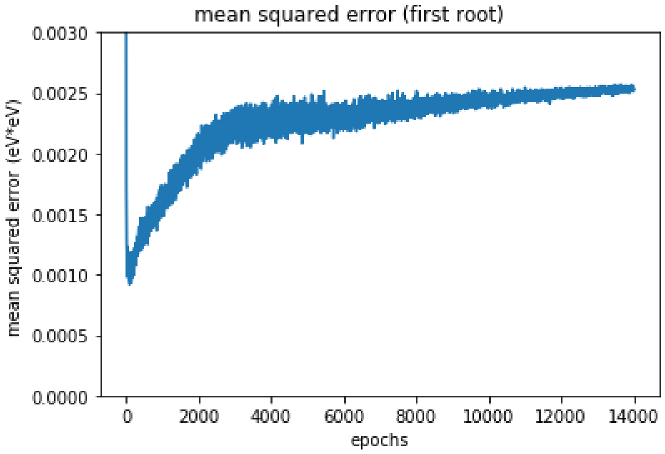

Figure 5.3: The mean squared error of the model on the test data set over the training epochs plot, with the Cartesian-matrix descriptor as input. The best possible results seem to occur after 400 epochs, a sign of underfitting with no effective training afterwards. This is a sample training session output.

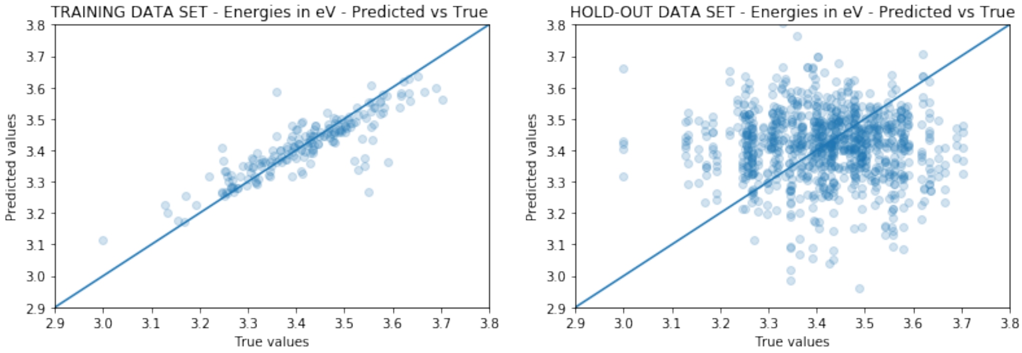

Figure 5.4: The fit of the predictions for the training data set is successful, but the predictions fail in the hold-out data set.

Table 5.3: Neural network model predictive performance with Cartesian descriptor as input.

Quantity	Value
MSE (eV^2)	0.001245
Slope	-0.021606
Coefficient of determination R^2	0.000684

The mean squared error is large compared to the z-matrix results, the regression line slope should have been close to 0.5 and the R^2 coefficient is close to zero. The results show that z-matrix is preferable over the Cartesian coordinates molecular descriptor.

5.2 Multi-layer perceptron hyperparameter optimization

In this section, the hyperparameter optimization is discussed. The neural network can consist of at least two layers, one input and one output layer. However, in order to solve any problem the addition of one hidden layer with many hidden neurons (sufficiently wide) is preferred[35]. Generally the presence of more than two hidden layers is not necessary and it can make the training sessions more time consuming. The effects of adding hidden layers is studied here.

5.2.1 1 wide hidden layer

With z-matrix molecular descriptor as input and one hidden layer which is wide (2x the neurons of the input layer) the results are comparable to the neural network model with two hidden layers which is already studied in section 4.1.

Table 5.4: Neural network model predictive performance with one wide hidden layer.

Quantity	Value
MSE (eV^2)	0.000661
Slope	0.333309
Coefficient of determination R^2	0.227982

However, the number of epochs needed to reach a local minimun in the mean squared error is now around 6000 epochs instead of 4000 (Figure 5.5).

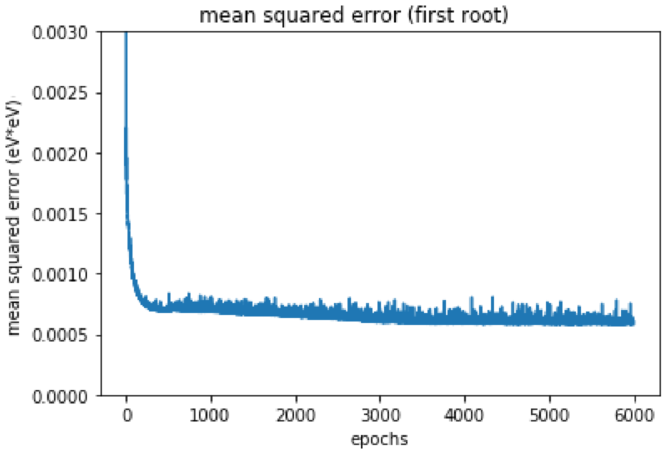

Figure 5.5: The mean squared error over the training epochs plot (model with 1 wide hidden layer, trained with the Z-matrix molecular descriptor).

5.2.2 3 and 4 hidden layers

With more than 2 hidden layers, the neural network model appears to perform slightly worse. This can be attributed to overfitting due to the large number of trainable parameters. Local minima of the mean squared error appear in the 3000-4000 epochs region. A comparison with the performance of the 2 hidden layer model is presented in the following table.

Table 5.5: Comparison of models with 2, 3 and 4 hidden layers.

	2 hidden layers	3 hidden layers	4 hidden layers
MSE (eV^2)	0.000650	0.000673	0.000722
Slope	0.411291	0.330017	0.319726
R^2	0.267261	0.234537	0.212121

5.3 Predictions in different norbornadiene variants

The number of atoms of the molecule for which the neural network model is trained can influence the prediction quality of the model. If, for example, a molecule consists of 26 atoms, the trainable parameters of the densely connected multi-layer perceptron are drastically reduced when compared to a molecule with 35 atoms, like **4**. The reason is that a smaller z-matrix requires also a neural network with narrower layers, since the dimensionality of the input vector depends on the number of atoms. The neural network for the prediction of the first energy root of N5 has the following structure (Table 5.6):

Table 5.6: The summary of the neural network model for **5** (output of Keras model.summary() command).

Layer (type)	Output shape	Number of weights
Flatten	78	0
Dense	78	6162
Dense	78	6162
Dense	1	79

Trainable parameters: 12403

This is obviously easier to train than the neural network for **3**, the compound that has been used as a benchmark in the previous sections. The neural network for **3** contains 14365 trainable parameters (Table 5.1), whereas the model for **5** contains 12405 trainable parameters. The training sessions also produce improved predictions, with better metrics, such as smaller mean squared error. Interestingly, the minimum mean squared error in the predictions of the hold-out data set appears in the region of 12000-13000 epochs instead of 4000 epochs (Figure 5.6), which was the case for **5**. More trainable parameters (e.g. with more hidden layers) can lower the optimal number of training epochs where the minimum mean squared error appears, however they are not always desirable because they result in worse overall metrics and more overfitting.

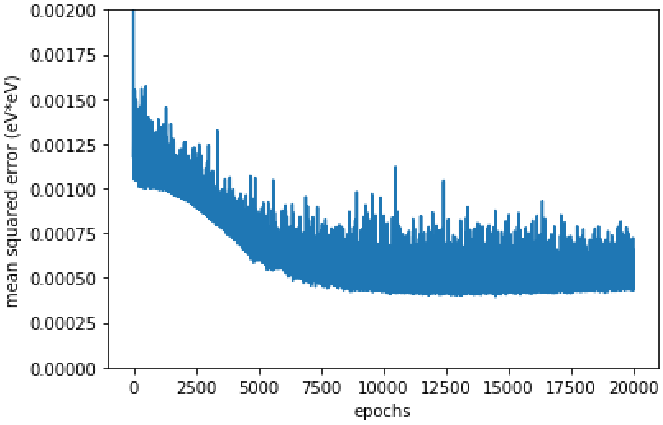

Figure 5.6: The mean squared error over the training epochs plot for **5** (two hidden layers, Z-matrix descriptor).

The **5** norbornadiene variant is molecule with the fewer atoms in the database (26 atoms). The other extreme is the **4** variant, with 35 atoms. The neural network consists of more trainable parameters (Table 5.8) and training becomes much harder(Figure 5.7), to the point that it practically fails for this compound and there are no useful predictions.

Table 5.7: Comparison of model predictions for **4** and **5**.

	N4	N5
atoms	35	26
MSE	0.001466	0.000663
Slope	-0.057614	0.558538
Coefficient of determination R^2	0.008038	0.469450

Table 5.8: The summary of the neural network model for **4** (output of Keras model.summary() command).

Layer (type)	Output shape	Number of weights
Flatten	105	0
Dense	105	11130
Dense	105	11130
Dense	1	106

Trainable parameters: 22366

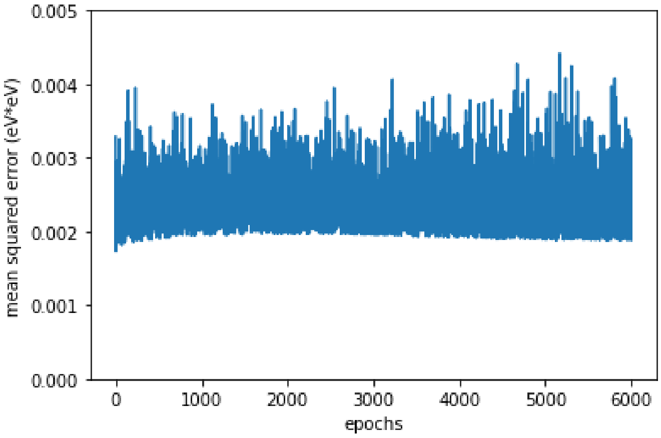

Figure 5.7: The mean squared error over the training epochs plot for **4** (2 hidden layers, Z-matrix descriptor).

5.4 Performance of other regression models

The performance of a few other machine learning algorithms and regression models has been tested. The mean squared error is smaller for the multi-layer perceptron presented in the previous section. This can be attributed to the fact that a neural network has a structure that can be optimized for every specific problem. This can be both an advantage and a disadvantage, since it can pick up more nuanced correlations and perform better, however the experience in the context of this thesis shows that it can be by far the most consuming method, both in hyperparameter optimization and training.

5.4.1 Scikit-learn machine learning library models.

The following table (Table 5.9) shows the performance of several regression machine learning models implemented in the scikit-learn machine learning library.

Table 5.9: Comparison of different algorithms.

algorithm	MSE
linear regression	0.019394
random forest	0.000931
kernel ridge	0.001095
neural network	0.000650

5.4.2 Ridge regression with Many-body Tensor Representation and Smooth Overlap of Atomic Positions molecular descriptors

In contrast to the neural network - based models where the Z-matrix, a simple molecular descriptor of a reduced coordinate system, produced better results, a linear model with molecular descriptors of higher complexity can be shown to have even better performance.

The S1 excitation energy distribution plots exhibit several linear relations between the normal mode space and the S1 energy spectrum. In such a case, ridge regression is expected to perform well.

The information of the normal mode distribution is then fed as input to the Many-body Tensor Representation (MBTR) and Smooth Overlap of Atomic Positions (SOAP) molecular descriptors, as implemented in the DScribe Python package implementation of molecular descriptors [31]. In the case of MBTR, the descriptor contains information on the inverse square and cosine of the normal mode of vibration, and in the case of SOAP geometrical information of the molecules. The data points for each molecule, N4a and N4d are 200, with the training data set containing 190 and the test data sets 10 molecular data entries. The parameters chosen for the MBTR model were: term $k = 2$ with inverse distance geometry function, grid parameters $min = 0$, $max = 1$, $n = 10$, $sigma = 0.1$, weighting with exponential function, scale 0.5 and cutoff $1e-3$, term $k = 3$ with cosine geometry function, grid parameters $min = -1$, $max = 1$, $n = 100$, $sigma = 0.1$, weighting with exponential function, scale 0.5 and cutoff $1e-3$. The SOAP model was tested with the parameters: inner averaging, rcut 8, nmax for integer values in the range between 2 and 8 and lmax in the range between 2 and each value of nmax.

The comparison between the two methods and the neural network trained with a Z-matrix descriptor gives the following results in the test data set (Table 5.10):

Table 5.10: Many-body tensor representation and smooth overlap of atomic positions model performance with the mean squared error values and the coefficient of determination (R^2).

	N3	N6
atoms	28	36
MBTR/Ridge MSE	0,00908	0,01191
MBTR/Ridge coefficient of determination	0.3080	-0.0985
SOAP/Ridge MSE	0,00706	0,00849
SOAP/Ridge coefficient of determination	0.4622	0.2170
Z-Matrix/NN MSE	0,000650	0,000933
Z-Matrix/NN coefficient of determination	0.2673	0.0806

Although the neural network trained with a Z-matrix descriptor exhibits smaller mean squared error, the R^2 value is arguably better in the case of the ridge regression model trained with the SOAP molecular descriptor. As it was shown in the previous chapter, the R^2 value can be a better criterion of a well-trained model than the mean squared error, because it shows that the model has actually "learned" the correlation between the molecular descriptor and the spectral features of the studied molecule

Especially for the N4d molecule, the ridge regression/SOAP descriptro combination shows the best R^2 value. The neural networks were failing to train properly for large molecules, as it was shown in the previous chapter. The MBTR descriptor failed as well, while the SOAP consistently produces better R^2 values, even in the case of N4d, showing a better way to train a machine learning model than neural networks.

It also has to be noted that although the generation of the SOAP descriptor requires some parametrization and arguably more complex calculations than the Z-matrix, the overall training time of the ridge regression model is way smaller than the neural network model, rendering this model more practical.

5.5 The effect of the size of the data set

Neural networks are data-hungry algorithms. With a 5% hold-out dataset split, from 200 molecules of N4a in toluene solvent, the model is left with 190 molecules/data points to train. Larger splits can have a significant impact on the accuracy of the model and its prediction capabilities.

An interesting phenomenon is that with fewer data points in the training set, the mean squared error does not increase significantly.

Table 5.11: Comparison of different training data set sizes.

dataset size	MSE
190	0.000650
180	0.000686
170	0.000732
160	0.000719
150	0.000692
100	0.000671

However, other metrics that determine the quality of the predictions worsen. The R^2 coefficient drops closer to 0 and the slope of the regression line in the predicted vs. true values plot is not close to 0.5. For example, with 100 data points in the training data set (Figure 5.8):

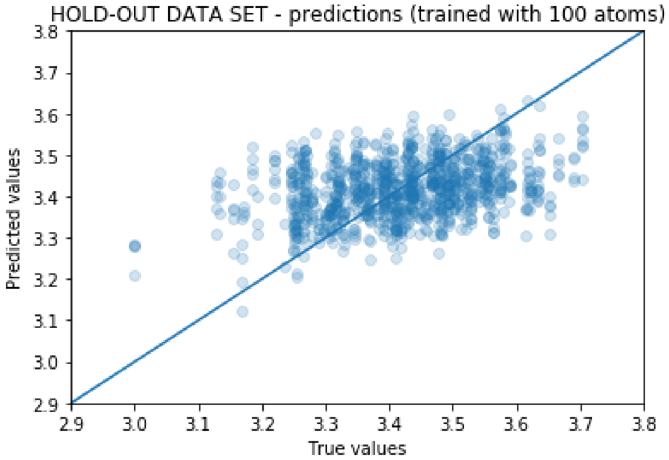

Figure 5.8: The quality of predictions of a neural network trained with just 100 molecular configurations.

Table 5.12: Neural network model predictive performance with a small training dataset.

Quantity	Value
MSE (eV^2)	0.000720
Slope	0.192421
Coefficient of determination R^2	0.111544

These results show that the mean squared error can be low, because the points in the plot above tend to cluster close to the line of predicted=true values, however the other metrics show that the neural network was hardly able to pick up any correlations between the atomic geometries and the first root energy. In an extreme case of underfitting, the neural network would be unable to learn these correlations and would just produce wrong predictions which would be dispersed (low R^2 coefficient) with a mean value close to the mean value of the actual energy of the first root. In this case, the regression line slope would be equal to zero. With a data set of size 190, R^2 was equal to 0.267261 and the regression line slope was 0.411291. With even larger data sets, we can assume that the R^2 coefficient would be closer to 1 and the regression line slope would closer to the ideal 0.5, indicating that the model is able to make more accurate predictions close to the true values. The size of the data set was a limiting factor, however the neural network was able to learn the correlation between molecular geometries and root energies efficiently for the two smaller molecules (**3** and **5**).

6
Conclusion

Overall, the goal of the thesis project to show that a molecular descriptor with a reduced coordinate system, the Z-matrix, can perform better than the Cartesian matrix descriptor. In this case, the Z-matrix encompasses the molecular geometry in a simple way, while the neural network is capable of resolving the complex correlation between the geometry and the spectral features of the compound. These correlations have been exhibited by analyzing the normal modes and how they are related to specific Z-matrix elements. Even though no method resulted in an acceptable R^2 coefficient value with a practical functionality (R^2 should be at least equal to 0.8), there are differences which indicate which model is the most successful in picking up these correlations.
However, this approach fails in two areas:

- The neural networks need a lot of time to train when compared to linear regression models.
- They fail in larger molecules. They increase the dimensionality of the problem and a possible solution is a larger data set.

Interestingly, a linear model approach seems to be able to work out complex correlations, but a new molecular descriptor (SOAP) has to be produced which is more complex, is not physically intuitive but it is a mathematical construct, and some parametrization is also needed. The overall process is more efficient however, especially when taking into account the almost instant training time when compared to the neural network's training time. The R^2 value for the linear model is arguably better than the neural network's output. The mean squared error may not always be the best parameter to determine the quality of the output, although it is a common choice for the loss function in the optimization algorithm during the training process of the machine learning model. The R^2 coefficient needs to be taken into account to assess the predictive capabilities of each model.
In every case, many data points must be produced first by time consuming computational methods (DFT and TDDFT). There is no evidence that these models can be trained to predict spectral features in new compounds for the moment. This has worked in previous research for the case of atomization energy, which is already a well researched topic in the field of machine learning for computational quantum chemistry. The complex correlations between molecular geometries and absorption spectra/excited energy levels have not yet produced such successful results. Machine learning application in computational quantum chemistry is an ongoing field of research and more relevant studies may be published in the future.

Bibliography

[1] Chalmers University of Technology. Emissions-free energy system saves heat from the summer sun for winter. *Chemistry and Chemical Engineering department*, 2019.

[2] Martina Viková. *Methodology of measurement of photochromic materials, kapitola 15. v knize: Somani, P.R., editor, Chromic Materials, Phenomena and their Technological Applications*, pages 509–536. 09 2010.

[3] Maria Quant, Alice Hamrin, Anders Lennartson, Paul Erhart, and Kasper Moth-Poulsen. Solvent Effects on the Absorption Profile, Kinetic Stability, and Photoisomerization Process of the Norbornadiene–Quadricyclanes System. *The Journal of Physical Chemistry C*, 123(12):7081–7087, March 2019. Publisher: American Chemical Society.

[4] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, jun 2017.

[5] B. Mehlig. Artificial Neural Networks. February 2019. arXiv: 1901.05639.

[6] Nathan S. Lewis and Daniel G. Nocera. Powering the planet: Chemical challenges in solar energy utilization. *Proceedings of the National Academy of Sciences of the United States of America*, 103(43):15729–15735, October 2006.

[7] Mikael J. Kuisma, Angelica M. Lundin, Kasper Moth-Poulsen, Per Hyldgaard, and Paul Erhart. Comparative Ab-Initio Study of Substituted Norbornadiene-Quadricyclane Compounds for Solar Thermal Storage. *The Journal of Physical Chemistry C*, 120(7):3635–3645, February 2016.

[8] Zen-ichi Yoshida. New molecular energy storage systems. *Journal of Photochemistry*, 29(1-2):27–40, May 1985.

[9] H.W. Wu, A. Emadi, G. de Graaf, J. Leijtens, and R.F. Wolffenbuttel. Design and fabrication of an albedo insensitive analog sun sensor. *Procedia Engineering*, 25:527 – 530, 2011. EurosensorsXXV.

[10] Victor Gray, Anders Lennartson, Phasin Ratanalert, Karl Börjesson, and Kasper Moth-Poulsen. Diaryl-substituted norbornadienes with red-shifted

absorption for molecular solar thermal energy storage. *Chem. Commun.*, 50(40):5330–5332, Physical Review 2014.

[11] Martyn Jevric, Anne U. Petersen, Mads Mansø, Sandeep Kumar Singh, Zhihang Wang, Ambra Dreos, Christopher Sumby, Mogens Brøndsted Nielsen, Karl Börjesson, Paul Erhart, and Kasper Moth-Poulsen. Norbornadiene-Based Photoswitches with Exceptional Combination of Solar Spectrum Match and Long-Term Energy Storage. *Chemistry - A European Journal*, 24(49):12767–12772, September 2018.

[12] Mikael J. Kuisma, Angelica M. Lundin, Kasper Moth-Poulsen, Per Hyldgaard, and Paul Erhart. Optimization of norbornadiene compounds for solar thermal storage by first-principles calculations. *Chemistry-Sustainability-Energy-Materials*, 9(14):1744–1899, July 2016.

[13] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters*, 108(5), January 2012.

[14] Bernd Möller, Eva Wiechers, Urban Persson, Lars Grundahl, Rasmus Søgaard Lund, and Brian Vad Mathiesen. Heat roadmap europe: Towards eu-wide, local heat supply strategies. *Energy*, 177:554 – 564, 2019.

[15] Jessica Orrego-Hernández, Ambra Dreos, and Kasper Moth-Poulsen. Engineering of Norbornadiene/Quadricyclane Photoswitches for Molecular Solar Thermal Energy Storage Applications. *Accounts of Chemical Research*, July 2020.

[16] Anders Lennartson, Anna Roffey, and Kasper Moth-Poulsen. Designing photoswitches for molecular solar thermal energy storage. *Tetrahedron Letters*, 56(12):1457–1465, March 2015.

[17] Xu-wu An and Yin-de Xie. Enthalpy of isomerization of quadricyclane to norbornadiene. *Thermochimica Acta*, 220:17–25, June 1993.

[18] Olaf Brummel, Daniel Besold, Tibor Döpper, Yanlin Wu, Sebastian Bochmann, Federica Lazzari, Fabian Waidhas, Udo Bauer, Philipp Bachmann, Christian Papp, Hans-Peter Steinrück, Andreas Görling, Jörg Libuda, and Julien Bachmann. Energy Storage in Strained Organic Molecules: (Spectro)Electrochemical Characterization of Norbornadiene and Quadricyclane. *ChemSusChem*, 9(12):1424–1432, 2016.

[19] Maria Quant, Anders Lennartson, Ambra Dreos, Mikael Kuisma, Paul Erhart, Karl Börjesson, and Kasper Moth-Poulsen. Low Molecular Weight Norbornadiene Derivatives for Molecular Solar-Thermal Energy Storage. *Chemistry – A European Journal*, 22(37):13265–13274, September 2016.

[20] R. H. Dalitz and Rudolf Peierls. Paul adrien maurice dirac. 8 august 1902-20 october 1984. *Biographical Memoirs of Fellows of the Royal Society*, 32:139–185, 1986.

[21] Logan Ward, Ben Blaiszik, Ian Foster, Rajeev S. Assary, Badri Narayanan, and Larry Curtiss. Machine learning prediction of accurate atomization energies of organic molecules from low-fidelity quantum chemical calculations. *MRS Communications*, 9(3):891–899, 2019.

[22] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5), Jan 2012.

[23] Grégoire Montavon, Matthias Rupp, Vivekanand Gobre, Alvaro Vazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, and O Anatole von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, sep 2013.

[24] Xin Chen, Mathias Jørgensen, Jun Li, and Bjørk Hammer. Atomic energies from a convolutional neural network. *Journal of Chemical Theory and Computation*, 14, 05 2018.

[25] Jessica Orrego-Hernández, Ambra Dreos, and Kasper Moth-Poulsen. Engineering of norbornadiene/quadricyclane photoswitches for molecular solar thermal energy storage applications. *Accounts of Chemical Research*, 53(8):1478–1487, August 2020.

[26] Ambra Dreos, Zhihang Wang, Jonas Udmark, Anna Ström, Paul Erhart, Karl Börjesson, Mogens Brøndsted Nielsen, and Kasper Moth-Poulsen. Liquid Norbornadiene Photoswitches for Solar Energy Storage. *Advanced Energy Materials*, 8(18):1703401, 2018.

[27] Mads Mansø, Martin Drøhse Kilde, Sandeep Kumar Singh, Paul Erhart, Kasper Moth-Poulsen, and Mogens Brøndsted Nielsen. Dithiafulvene derivatized donor–acceptor norbornadienes with redshifted absorption. *Physical Chemistry Chemical Physics*, 2019.

[28] David S. Sholl, Janice A. Steckel. Density Functional Theory: A Practical Introduction, 2009.

[29] Richard M. Martin. *Electronic Structure: Basic Theory and Practical Methods.* Cambridge University Press, Cambridge, 2004.

[30] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors. WileyVCH, Weinheim*, volume 11. 09 2000.

[31] Lauri Himanen, Marc O. J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, February 2020.

[32] Haoyan Huo and Matthias Rupp. Unified Representation of Molecules and Crystals for Machine Learning. *arXiv:1704.06439 [cond-mat, physics:physics]*, January 2018. arXiv: 1704.06439 version: 3.

[33] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19 – 25, 2015.

[34] M. Valiev, E.J. Bylaska, N. Govind, K. Kowalski, T.P. Straatsma, H.J.J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T.L. Windus, and W.A. de Jong. Nwchem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications*, 181(9):1477 – 1489, 2010.

[35] Stephen Marsland. Machine Learning: An Algorithmic Perspective, Second Edition, 2014.

A

Appendix 1

A.1 Density Functional Theory calculations in NWChem

The following input scripts were used in NWChem to initiate DFT calculations. DFT geometry optimization was performed via NWChem, with the option of tight convergence:

```
dft
  xc b3lyp
  mult 1
  convergence gradient 1e-6
end

driver
  maxiter 200
  tight
end

task dft optimize
```

Then, frequency analysis was performed but with the option of more accurate convergence, with a gradient $1e-8$.

```
dft
  xc b3lyp
  mult 1
  convergence gradient 1e-8
end

task dft freq
```