Ab initio analysis of meta- and para-aminobenzoic acid molecules

Master's thesis
Kalle Rantamaula
Department of Physics
University of Oulu
Spring 2022

Contents

Abbreviations and Symbols

ABA Aminobenzoic Acid

m-ABA meta-Aminobenzoic Acid or 3-Aminobenzoic Acid

1 m-ABA Rotamer 1 meta-Aminobenzoic Acid or Rotamer 1 3-Aminobenzoic Acid

2 m-ABA Rotamer 2 meta-Aminobenzoic Acid or Rotamer 2 3-Aminobenzoic Acid

p-ABA para-Aminobenzoic Acid or 4-Aminobenzoic Acid

HF Hartree-Fock

MP2 Møller and Plesset perturbation theory truncated at 2nd order

CCSD Coupled-Cluster Singles and Doubles

H Hamiltonian

PES Potential Energy Surface

MO Molecular orbital

AO Atomic orbital

LCAO Linear Combination of Atomic Orbitals

STO Slater Type Orbital

GTO Gaussian Type Orbital

STO-nG Slater Type Orbital approximated by n primitive Gaussians

cc-pCVnZ correlation-consistent polarized Core and Valence (Double/Triple/ect.) Zeta

cc-pwCVnZ correlation-consistent polarized weighted Core and Valence (Double/Triple/ect.) Zeta

cc-pVnZ correlation-consistent polarized Valence (Double/Triple/ect.) Zeta

SCF Self-Consistent Field

RHF Restricted Hartree-Fock

ROHF Restricted Open-shell Hartree-Fock

UHF Unrestricted Hartree-Fock

MCSCF Multiconfiguration Self-Consistent Field

CSF Configuration State Function or Configuration

CAS Complete Active Space

RAS Restricted Active Space

MP Møller and Plesset perturbation theory

MPn Møller and Plesset perturbation theory truncated at nth order

CI Configuration interaction

CIS Configuration Interaction Singles

CID Configuration Interaction Doubles

CISD Configuration Interaction Singles and Doubles

CC Coupled-Cluster

CCS Coupled-Cluster Singles

CCSDT Configuration Interaction Singles, Doubles and Triples

CCSD(T) Coupled-Cluster Singles, Doubles and an estimation of Triples based on perturbation theory

CISDT Configuration Interaction Singles, Doubles and Triples

CISDTQ Configuration Interaction Singles, Doubles, Triples and Quadruples

CCSDTQ Coupled-Cluster Singles, Doubles, Triples and Quadruples

Abstract

Ab initio methods such as the Hartree-Fock method, the Møller and Plesset perturbation method and the Coupled-Cluster method can be used to solve the electronic Schrödinger equation and calculate the total and orbital energies of the system. The total energy can then be used for optimizing the geometry of the system. The binding energies of the molecule can then be approximated from the orbital energies at the optimized geometry with Koopmans' theorem, which states that the one-electron orbital energy can be taken to be roughly the negative of the ionization energy of the electron from that orbital. The approximations of binding energies can then be displayed as a density of states spectra and compared to experimental binding energy spectra.

By optimizing the geometries of the two rotamers of meta-aminobenzoic acid and the para-aminobenzoic acid molecules with ab initio methods and comparing the results with the experimental binding energy spectra of meta-aminobenzoic acid it is easy to identify most of the peaks from the experimental spectra and analyze the differences between experimental and calculated binding energies.

By comparing the approximated binding energies of the two rotamers of meta-aminobenzoic acid molecules to each other and the experimental meta-aminobenzoic acid molecule binding energy spectra it can be seen that it is difficult to differentiate between rotamers based on the experimental spectra, as most of the differences are small and the larger ones are in binding energy regions where they would be difficult to notice.

The differences between experimental and calculated approximations of binding energies for meta-aminobenzoic acid molecules and the calculated approximations of para-aminobenzoic acid's binding energies can be used to form a decent approximation of what para-aminobenzoic acid's true binding energy spectra would look like.

1 Introduction

In this thesis I optimize the geometry and study the binding energies of aminobenzoic acid (ABA) molecules. First, I go through the theory needed for geometry optimization and analysis of the results. Then I present the experimental binding energy spectra of meta-aminobenzoic acid (m-ABA). I also briefly introduce the experimental setup and explain how the data was processed to form the spectra.

The geometries of rotamer 1 meta-aminobenzoic acid (1 m-ABA), rotamer 2 meta-aminobenzoic acid (2 m-ABA) and para-aminobenzoic acid (p-ABA) are optimized using ab initio methods. The geometry is first optimized using Hartree-Fock (HF) method going through different basis sets, then the geometry is optimized further first with the Møller and Plesset perturbation method truncated at 2nd order (MP2) and then the Coupled-Cluster Singles and Doubles (CCSD) method going through the same basis sets as with HF method and always using the result gained with the previous basis set or method as the starting point for the next one. These calculations also provide the orbital energies for the optimized geometries and from the orbital energies the corresponding binding energies can be approximated using Koopmans' theorem.

The approximated binding energies for the final optimization level of 1 m-ABA, 2 m-ABA are compared to each other and the measured m-ABA binding energy spectra. Based on the results of the comparison and the approximated binding energies of p-ABA an approximation of the true p-ABA binding energy spectra is then formed.

2 Theory

2.1 Quantum Mechanics

In quantum mechanics any chemical system can be described by a wave function Ψ. A wave function returns the system's properties when acted upon by appropriate operators. The Hamiltonian operator H returns the system's total energy E as an eigenvalue, as seen from the Schrödinger equation:

$$H\Psi = E\Psi. \tag{1}$$

For a molecule there are many valid eigenfunctions Ψ each with their own eigenvalue E. This means that there is a complete set of orthonormal wave functions Ψ_i with eigenvalues E_i. This leads to the equation:

$$\int \Psi_j H \Psi_i d\mathbf{r} = E_i \delta_{ij}, \tag{2}$$

where δ_{ij} is the Kronecker delta: [1]

$$\delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j. \end{cases} \tag{3}$$

The Hamiltonian H can be written as

$$\begin{aligned} H = &-\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2 - \sum_k \frac{\hbar^2}{2m_k}\nabla_k^2 \\ &- \sum_i \sum_k \frac{e^2 Z_k}{r_{ik}} + \sum_{i<j} \frac{e^2}{r_{ij}} + \sum_{k<l} \frac{e^2 Z_k Z_l}{r_{kl}}, \end{aligned} \tag{4}$$

where i and j run over electrons, k and l run over nuclei, m_e is the mass of an electron, m_k is the mass of nucleus k, Z is the atomic number of the indicated nucleus, r is the distance between the indicated particles and e is the charge of an electron. Eq.(4) can be simplified into the form

$$\begin{aligned} H = &-\sum_i \frac{1}{2}\nabla_i^2 - \sum_k \frac{1}{2m_k}\nabla_k^2 \\ &- \sum_i \sum_k \frac{Z_k}{r_{ik}} + \sum_{i<j} \frac{1}{r_{ij}} + \sum_{k<l} \frac{Z_k Z_l}{r_{kl}} \end{aligned} \tag{5}$$

by using atomic units. [1]

The first two terms of the Hamiltonian represent kinetic energy and the last three potential energy. Thus, Eq.(4) can also be written in the form

$$H = T_e + T_n + V_{ne} + V_{ee} + V_{nn}, \tag{6}$$

where the terms T_e and T_n are kinetic energies and the terms V_{ne}, V_{ee} and V_{nn} are potential energies, with e indicating electrons and n indicating nuclei. [1,2]

2.1.1   Variational principle

If an arbitrary function Φ is of the correct form to be operated by the Hamiltonian, it must be a linear combination of Ψ_i:

$$\Phi = \sum_i c_i \Psi_i, \tag{7}$$

where the individual Ψ_i or the individual coefficients c_i are not necessarily known. Due to the normalization of Φ we can write

$$\int \Phi^2 d\mathbf{r} = 1 = \sum_i c_i^2 \tag{8}$$

and using Eq. (3) we can also write

$$\int \Phi H \Phi d\mathbf{r} = \sum_i c_i^2 E_i. \tag{9}$$

Combining equations (8) and (9) leads to

$$\int \Phi H \Phi d\mathbf{r} - E_0 \int \Phi^2 d\mathbf{r} = \sum_i c_i^2 (E_i - E_0), \tag{10}$$

where E_0 is the lowest energy value in the set of all E_i. Thus, $E_i - E_0$ is always greater than or equal to zero and if the c_i is assumed to be a real number then c_i^2 must also always be greater than or equal to zero. This leads to the equation:

$$\frac{\int \Phi H \Phi d\mathbf{r}}{\int \Phi^2 d\mathbf{r}} = \mathbf{E} \geq E_0, \tag{11}$$

where $\mathbf{E}$ is the Rayleigh ratio. The Rayleigh ratio can be used to optimize trial wave functions, with a lower energy indicating a better wave function and the $\mathbf{E} = E_0$ case applying only when the trial wave function is the ground-state wave function of the system. As the Rayleigh ratio is never lower than the ground-state energy, variational methods always give an upper bound to the

system's true energy. However, there is no guarantee that a wave function optimized using the Rayleigh ratio will give good values for properties other than energy. [1,3]

Eq.(11) can also be written in the form

$$
\begin{aligned}
E &= \frac{\int \left(\sum_i c_i \Psi_i \right) H \left(\sum_j c_j \Psi_j \right) d\mathbf{r}}{\int \left(\sum_i c_i \Psi_i \right) \left(\sum_j c_j \Psi_j \right) d\mathbf{r}} \\
&= \frac{\int \sum_{ij} c_i c_j \Psi_i H \Psi_j d\mathbf{r}}{\int \sum_{ij} c_i c_j \Psi_i \Psi_j d\mathbf{r}} \\
&= \frac{\sum_{ij} c_i c_j H_{ij}}{\sum_{ij} c_i c_j S_{ij}},
\end{aligned}
\tag{12}
$$

where the S_{ij} is the overlap integral. This version of the variation theory is known as Rayleigh-Ritz method. Optimizing the energy in Eq.(12) leads to the equation:

$$
\sum_{i=1}^{N} c_i (H_{ki} - ES_{ki}) = 0 \quad \forall k,
\tag{13}
$$

where N is the number of functions Ψ that make up the wave function Φ. Eq.(12) can be used to optimize the coefficients c_i of the functions Ψ_i, by solving the N values of energy E_j from

$$
\begin{vmatrix}
H_{11} - ES_{11} & H_{12} - ES_{12} & \cdots & H_{1N} - ES_{1N} \\
H_{21} - ES_{21} & H_{22} - ES_{22} & \cdots & H_{2N} - ES_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
H_{N1} - ES_{N1} & H_{N2} - ES_{N2} & \cdots & H_{NN} - ES_{NN}
\end{vmatrix} = 0
\tag{14}
$$

and applying them to Eq.(13). [1,3]

Eq.(11) also shows that the trial wave function Φ does not need to be constructed as a linear combination of Ψ_i. Instead, the trial wave function can be constructed in any manner desired and then its quality can be evaluated by using Eq.(11). [1]

2.1.2 Born-Oppenheimer approximation

In order to simplify the required calculations, the Born-Oppenheimer approximation is used. In the Born-Oppenheimer approximation the movements of the electrons and the nuclei of the system are decoupled and then the electronic energies are computed for fixed nuclear positions. This is based on the

fact that the nuclei are moving much slower than the electrons and thus electronic movement can be taken to be instantaneous compared to the motion of the nuclei. Thus, the Schrödinger equation can be separated into two parts: the electronic Schrödinger equation and the nuclear Schrödinger equation. The electronic Schrödinger equation can be written as

$$(H + V)\Psi = E\Psi, \tag{15}$$

where H contains the terms T_e, V_{ne} and V_{ee} of the Hamiltonian in Eq.(6) and V is the nuclear-nuclear repulsion energy V_{nn}. As the electronic wave function Ψ is considered to be in a fixed nuclear position, the term V_{nn} is a constant for that position and the electron-nuclear correlation in the term V_{ne} has been negated. The nuclear Schrödinger equation contains the remaining T_n term of Eq.(6). The different solutions of the electronic Schrödinger equation Eq.(15) for different fixed nuclear geometries form the potential energy surface (PES) of the molecule. [1, 2]

Without the use of Born-Oppenheimer approximation it is impossible to analytically solve the Schrödinger equation for even the simplest of molecules. The Born-Oppenheimer approximation does not cause much error in ground state calculations, but it is less accurate for excited states. [3]

With the use of Born-Oppenheimer approximation electron-nuclear correlation has been removed, but the electron-electron correlation still remains. The simplest way to handle it is to ignore it by only considering systems with a single electron. The eigenfunctions of Eq. (15) are then one-electron wave functions and they can be considered molecular orbitals (MO) of the molecule. Thus, the eigenvalue associated with a MO is the energy of the electron in that orbital. [1]

Ignoring the electron correlation simplifies the calculations, but it does so at the cost of accuracy. Because of this multiple ways of taking the electron correlation at least partially into account have been developed. The Chapters 2.3.2 -2.3.5 introduce some of them.

2.1.3 Molecular orbital theory

Atomic orbitals (AO) are one-electron wave functions of atoms. They can be classified into 1s, 2s, 2p, 3s, 3p, 3d, 4s, 4p, 4d, 4f, etc. orbitals based on their principal quantum number, which is represented by the number part and angular momentum quantum number, which is represented by the letter part. Electrons also have a spin of 1/2 which is taken into account separately as can be seen from the methods introduced in Chapter 2.3 [3]

If the system only has one nucleus and one electron it can be solved exactly by using hydrogenic AOs. Based on this, atomic orbitals can be used

to model more complex MOs. So, the trial wave function can be constructed as a linear combination of atomic wave functions. This method is called linear combination of atomic orbitals (LCAO). [1]

As the orbital approximation is quite primitive, there are no good analytical forms for AOs of many electron atoms. However, the AOs can be approximated with Slater-type orbitals (STO) [4]. In 1950 S.F. Boys proposed the idea of using the computationally more efficient Gaussian-type orbitals (GTO) [5] instead of STOs. But because GTOs are worse than STOs at representing the orbitals at the nuclei, more of them are required to match the accuracy of using STOs. The relative shapes of STOs and GTOs can be seen from the Figure 1.

Figure 1: The difference between STO and GTO shapes

This difference in shapes is caused by the fact that radial decay of STOs is proportional to e^{-r}, while the radial decay of GTOs is proportional to e^{-r^2}, where r is the distance from the nuclei. [1,3]

To counter the extra computational effort caused by using more GTOs, they are often grouped together as contracted Gaussian functions. A contracted Gaussian function G_o is a linear combination of Gaussian functions g_i of the same atom:

$$G_o = \sum_i d_{oi} g_i, \qquad (16)$$

where the coefficients d_{oi} and the parameters of g_i are held fixed during calculations, thus reducing the number of unknown variables. The Gaussian functions g_i are called primitive Gaussian functions. [1,3]

There are also other ways to approximate the orbitals such as using plane waves, but they are not focused on in this thesis.

2.1.4 Basis sets

The group of functions used to form a wave function is called a basis set. A molecule could be represented perfectly by using an infinite number of basis functions with each MO expressed as a linear combination of basis functions to form the wave function, but it is impossible to practically implement an infinite basis set. The error in MO representation caused by using a finite basis set is called the basis-set truncation error. To minimize the basis-set truncation error, much work has been done to look for mathematical functions that allow wave functions to approach infinite basis as efficiently as possible. [1, 3]

There are three important factors in creating a good basis set. Keeping the total amount of basis functions relatively low is useful as it reduces the computational load. The functional forms of the basis set should be chosen so that their mathematical evaluation is easy. Sometimes this can lead to larger basis sets being faster to evaluate than smaller ones. The forms of the basis functions should be useful in a chemical sense. This means having large amplitudes in regions where electron density is high and small amplitudes where it is low. [1]

Basis sets can be divided into single-ζ, multiple-ζ and split-valence basis sets. Single-ζ or minimal basis sets have one basis function for each type of orbital. The STO-nG basis sets developed by Hehre, Stewart and Pople in 1969 [6] are single-ζ basis sets. The name STO-nG means STO approximated by n primitive GTOs and as can be seen from the name they are based on using a contracted Gaussian to approximate a STO. Of the STO-nG basis sets the STO-3G is the most used as it provides a good balance of speed and accuracy. [1, 2]

Multiple-ζ basis sets have multiple basis functions for each orbital, and they are further classified as double-ζ basis, triple-ζ basis and so on based on the number of basis functions they have for each orbital. The point of this is to increase the size of the basis set and thus approach infinite basis. The cc-pCVnZ basis sets developed by Woon and Dunning in 1995 [7] are multiple-ζ basis sets. The name cc-pCVnZ means correlation-consistent polarized Core and Valence n Zeta, where the n is D(Double)/T(Triple)/etc. based on the number of functions the set has per orbital. The correlation-consistent part of the name indicated that the basis sets are optimized for methods that take electron correlation into account. [1, 2]

The cc-pwCVnZ basis sets developed by Peterson and Dunning in 2002

are an improved version of the cc-pCVnZ basis sets. The added w in the name cc-pwCVnZ means weighted and it indicates that the basis set weights the core-valence correlation over the core-core correlation. [8]

Split-valence or valence-multiple-ζ basis sets have one basis function for each of the core orbitals, but the valence orbitals have multiple functions. This is done because core orbitals are only weakly affected by chemical bonding and thus are similar for the same atom in different molecules. Valence orbitals instead can vary widely as a function of chemical bonding and thus benefit more from the extra flexibility given by multiple functions per orbital. They are also further classified as valence-double-ζ basis, valence-triple-ζ basis and so on. The k-nlmG basis sets developed by Pople and co-workers are split-valence basis sets. In k-nlmG the k indicates the number of primitive GTOs used for core orbitals and the nlm part describes how the valence orbitals are handled. The number of parameters in nlm part tells how many functions the valence orbitals are split into, and the values of the parameters tell how many primitive GTOs are used for each of those parts. So, 3-21G is a valence-double-ζ basis, while 6-311G is a valence-triple-ζ basis. The k-nlmG basis sets are known as Pople basis sets. The cc-pVnZ basis sets developed by Woon and Dunning [9, 10] are also split-valence basis sets. They are split-valence versions of the cc-pCVnZ basis sets as can be seen from the name. [1, 2]

While implementing infinite basis is impossible, it is possible to extrapolate to the infinite basis limit with decent accuracy. But even for basis sets that were designed for it like the cc-pVnZ and the cc-pCVnZ, the computational cost of the extrapolation procedure becomes increasingly high as the number of functions per orbital grows. This is especially true for systems that have more than five atoms. [1]

2.2   Geometry optimization

In computational chemistry geometry optimization means looking for a molecule's lowest energy configuration starting from a given position. To optimize the geometry of a molecule the minimum required starting information is the atoms of which it is made of and their relative positions. The positions can be given in Cartesian coordinates or bond lengths and angles between the bonds. [1]

All the possible structures of a molecule can be taken into account by describing the molecule's PES. The PES is a hypersurface defined by the potential energy of a collection of atoms over all possible atomic arrangements. It is often nearly impossible to construct a complete PES for molecules that have more than 3 atoms. Because of this often only the chemically interesting parts of the PES such as the areas around minima are fully formed. The PES

has $3N - 6$ dimensions, where N is the number of atoms and $N \geq 3$. The local minima of the PES represent optimal geometries of the molecule, and the saddle points indicate paths between minima. [1, 2]

At its basis, geometry optimization is a math problem of finding the minimum of an arbitrary function of many variables. The simplest way to optimize geometry is to go through and optimize all the parameters (bond lengths and angles between bonds) one at a time. But because the minimization of total energy in polyatomic molecules often forces the individual parameters to deviate from their minimum energy configurations, this optimization process must be repeated multiple times until the overall energy finally converges. This can be a very lengthy process even for a relatively small molecule. Thus, it is better to use methods like locating the greatest downward slope in energy with respect to all coordinates instead of optimizing one parameter at a time. The geometry optimization of a complex molecule can require hundreds or thousands of energy evaluations for intermediate structures. Thus, small time savings on the calculation can multiply so that they have a significant effect on the overall calculation time. [1]

A problem with geometry optimization is that most optimization methods are good at finding the nearest minimum instead of the global minimum. Some methods also end in the nearest stationary point with no guarantee of it being a minimum. Also, in reality a molecule's geometry is fluctuating around the minimum even at 0 K due to zero-point vibrational effects. Small molecules are usually fairly close to the minimum energy structure, but larger molecules with a looser structure can be quite far from the optimized structure. At temperatures above 0 K these fluctuations can be even greater. [1]

Symmetry can be used to help in geometry optimization by removing some of the $3N - 6$ degrees of molecular freedom. This also has some drawbacks, but the positives mostly overshadow them. Using symmetry to limit the degrees of freedom might lead to the end result being a transition state instead of a minimum. This happens when the symmetry constraints prevent reaching the true minimum. This problem can be solved by relaxing some or all of the symmetry constraints in a later part of the geometry optimization process. [1]

2.3   Ab initio methods

Ab initio calculations and semiempirical methods are the two main ways of solving the Schrödinger equation Eq.(1). In ab initio calculations a trial wave function is formed, and the Schrödinger equation is solved by using only the trial wave function, fundamental constants and atomic numbers of the nuclei.

The accuracy of an ab initio calculation depends mainly on the quality of the trial wave function and the available computational power. Semiempirical methods are a computationally less expensive alternative. They use a simplified Hamiltonian and parameters based on experimental data to lessen the computational load. [3]

2.3.1 Hartree-Fock theory

By ignoring the electron-electron correlation the Hamiltonian becomes separable, which means it can be separated into one-electron Hamiltonian operators giving it the form

$$H = \sum_{i=1}^{N} h_i, \tag{17}$$

where N is the number of electrons and the one-electron Hamiltonian h is defined as

$$h_i = -\frac{1}{2}\nabla_i^2 - \sum_{k=1}^{M} \frac{Z_k}{r_{ik}} + \sum_{j\neq i} \int \frac{\rho_j}{r_{ij}} d\mathbf{r}, \tag{18}$$

where M is the total number of nuclei, Z_k is the atomic number of the nucleus k, ρ_j is the charge density of the electron j and r_{ik} and r_{ij} are the distances between the indicated particles. The final term of Eq.(18) represents the interaction potential of an electron with all the other electrons of the system. The many electron eigenfunctions of a separable Hamiltonian can then be constructed as a product of one-electron eigenfunctions:

$$\Psi_{HP} = \psi_1\psi_2\cdots\psi_N. \tag{19}$$

This is called a Hartree-product wave function. [1,3]

In 1928 Hartree developed the self-consistent field (SCF) method [11]. In the SCF method the wave functions for all the occupied MOs are guessed and then they are used to construct the necessary one-electron Hamiltonian operators h, which are then used to solve the one-electron Schrödinger equations:

$$h_i\psi_i = \varepsilon_i\psi_i. \tag{20}$$

The resulting eigenvalues ε_i represent the energies of different MOs. With the help of the variational principle Eq.(13) these energies can then be used to optimize the coefficients of the basis functions that make up the wave functions ψ. This leads to a new set of ψ, which can then be used to generate a new set of h and this process is repeated until the set of ψ converges. The resulting set of ψ is called converged SCF orbitals. [1]

The one-electron orbital energy ε_i can be taken to be roughly equal to the negative of the ionization energy of the electron from the orbital represented by ψ_i:

$$I_i \approx -\varepsilon_i. \tag{21}$$

This approximation is known as the Koopmans' theorem. [3]

Wave functions that are just products of one-electron wave functions cannot satisfy the Pauli principle, which states that when the coordinates of two electrons are exchanged the electronic wave function changes sign. To satisfy the Pauli principle the ground-state Hartree-product wave function of a system with two electrons of the same spin α can be written as

$$\Psi_{SD} = \frac{1}{\sqrt{2}} \begin{vmatrix} \psi_a(1)\alpha(1) & \psi_b(1)\alpha(1) \\ \psi_a(2)\alpha(2) & \psi_b(2)\alpha(2) \end{vmatrix}. \tag{22}$$

This is called a Slater determinant and due to the properties of determinants it changes signs when any two rows are exchanged. As the rows correspond to different electrons, this electron exchange allows Slater determinants to satisfy the Pauli principle. A more general form of the Slater determinant is

$$\Psi_{SD} = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(1) & \chi_2(1) & \cdots & \chi_N(1) \\ \chi_1(2) & \chi_2(2) & \cdots & \chi_N(2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(N) & \chi_2(N) & \cdots & \chi_N(N) \end{vmatrix}, \tag{23}$$

where N is the number of electrons and χ is a spin-orbital. Spin orbital is a product of a spatial orbital and an electron spin eigenfunction. [1,3]

Fock used Slater determinants to expand Hartree's SCF method to include the effects of electron exchange, creating the HF SCF method. Later in 1951 Roothaan [12] and Hall [13] provided the necessary mathematical tools for doing the HF calculations using basis set representations for the MOs. In a closed-shell system HF equations can be presented using

$$f_i = -\frac{1}{2}\nabla_i^2 - \sum_k^{nuclei} \frac{Z_i}{r_{ik}} + \sum_{j \neq i} (2J_{ij} - K_{ij}), \tag{24}$$

where f_i is the one-electron Fock operator for the electron i, J is the Coulomb operator

$$J_{ab} = \iint |\psi_a(1)|^2 \frac{1}{r_{12}} |\psi_b(2)|^2 d\mathbf{r}(1) d\mathbf{r}(2) \tag{25}$$

and K is the exchange operator

$$K_{ab} = \iint \psi_a(1)\psi_b(1) \frac{1}{r_{12}} \psi_a(2)\psi_b(2) d\mathbf{r}(1) d\mathbf{r}(2). \tag{26}$$

The resulting MO energies can then be calculated from a variation of the Eq.(14), where the Hamiltonian operator H has been replaced by the Fock operator F giving it the form

$$
\begin{vmatrix}
F_{11} - ES_{11} & F_{12} - ES_{12} & \cdots & F_{1N} - ES_{1N} \\
F_{21} - ES_{21} & F_{22} - ES_{22} & \cdots & F_{2N} - ES_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
F_{N1} - ES_{N1} & F_{N2} - ES_{N2} & \cdots & F_{NN} - ES_{NN}
\end{vmatrix} = 0, \qquad (27)
$$

where N is now the number of basis functions, the overlap matrix S is now:

$$
S_{ij} = \int \phi_i \phi_j d\mathbf{r}, \qquad (28)
$$

where ϕ_i and ϕ_j are basis functions and the elements of F are

$$
\begin{aligned}
F_{\mu\nu} = & \left\langle \mu \left| -\frac{1}{2}\nabla^2 \right| \nu \right\rangle - \sum_{k}^{nuclei} Z_k \left\langle \mu \left| \frac{1}{r_k} \right| \nu \right\rangle \\
& + \sum_{\lambda\sigma} P_{\lambda\sigma} \left[(\mu\nu|\lambda\sigma) - \frac{1}{2}(\mu\nu|\lambda\sigma) \right],
\end{aligned} \qquad (29)
$$

where k represents MOs and $P_{\lambda\sigma}$ is the designated element of the density matrix $\mathbf{P}$. $P_{\lambda\sigma}$ can be written as

$$
P_{\lambda\sigma} = 2 \sum_{i}^{occ.} a_{\lambda i} a_{\sigma i}, \qquad (30)
$$

where the coefficients a represent the effect of the basis functions λ and σ on the molecular orbital i. The notation $\langle \mu|g|\nu \rangle$ means one-electron integral of the form

$$
\langle \mu|g|\nu \rangle = \int \phi_\mu (g\phi_\nu) d\mathbf{r}, \qquad (31)
$$

where g is some operator and ϕ_μ and ϕ_ν are basis functions. The notation $(\mu\nu|\lambda\sigma)$ means the integration:

$$
(\mu\nu|\lambda\sigma) = \iint \phi_\mu(1)\phi_\nu(1)\frac{1}{r_{12}}\phi_\lambda(2)\phi_\sigma(2)d\mathbf{r}(1)d\mathbf{r}(2), \qquad (32)
$$

where ϕ_μ and ϕ_ν represent the probability density of electron 1 and ϕ_λ and ϕ_σ the probability density of electron 2. This form of HF theory is known as restricted Hartree-Fock (RHF). [1,3]

For an open-shell system HF theory has two forms: the restricted open-shell HF (ROHF) theory and unrestricted HF (UHF) theory. In ROHF all

the paired electrons are restricted to doubly occupied orbitals so they can be handled in the same way as in RHF. For the unpaired electrons the factor of two is removed from their density matrix elements Eq.(30) as they do not have a pair. This leads to the form

$$P_{\lambda\sigma} = \sum_i^{occ.} a_{\lambda i} a_{\sigma i}.$$

(33)

In UHF the paired electrons are not restricted to the same spatial orbital. This leads to different forms for Eq.(29)

$$F_{\mu\nu}^{\xi} = \left\langle \mu \left| -\frac{1}{2}\nabla^2 \right| \nu \right\rangle - \sum_k^{nuclei} Z_k \left\langle \mu \left| \frac{1}{r_k} \right| \nu \right\rangle$$
$$+ \sum_{\lambda\sigma} \left[\left(P_{\lambda\sigma}^{\alpha} + P_{\lambda\sigma}^{\beta} \right) (\mu\nu|\lambda\sigma) - P_{\lambda\sigma}^{\xi}(\mu\nu|\lambda\sigma) \right]$$

(34)

and Eq.(30)

$$P_{\lambda\sigma}^{\xi} = \sum_i^{\xi-occ.} a_{\lambda i}^{\xi} a_{\sigma i}^{\xi},$$

(35)

where α and β are the spins of the electrons and ξ is either α or β. [1]

UHF can also be used for closed shell systems and the UHF energy is always equal to or lower than the corresponding ROHF or RHF energy. However, the UHF wave functions sometimes contain contributions from higher spin states, which skews the results obtained with them. The case of equal energies mostly happens with closed shells states that are near their equilibrium geometry. [1,2]

There is an important difference between the Hamiltonian operator and the Fock operator. The Hamiltonian operator is used to find the electronic energy of a many-electron system. The Fock operator instead is the set of all the interdependent one-electron operators that are used to find the one-electron MOs which are used to construct the HF wave function as a Slater determinant. [1]

HF provides a well-defined energy, which can be converged in an infinite basis set. This is called the HF limit. The difference between the converged HF energy and reality is called the electron correlation energy:

$$E_{corr} = E - E_{HF}.$$

(36)

HF theory makes the approximation that each electron moves in the static electron field created by all the other electrons and optimizes orbitals for all

electrons in a self-consisted fashion. This ignoring of the electron correlation simplifies the calculations, but it is also the cause of the main error in HF theory. In HF limit this error is the electron correlation energy E_{corr}. [1,3]

There is no guarantee that any SCF process will converge to a stable solution. A common problem is a SCF oscillation where SCF computation gets stuck in a loop and never converges. Sometimes the energies associated with each step vary wildly and randomly, but this usually evens out as the process continues. These problems are most commonly caused by problems with the initial trial wave function. One solution for a badly behaving SCF process is to try using different convergence schemes and hope that one of them works. A better solution is to use a smaller basis set as it is usually easier to converge a small basis set HF calculation and then use the results as an initial guess for a larger basis set and repeat this process until the desired basis set is reached. Due to the exponential scaling of HF calculations the early steps of this process are comparatively fast and can even end up saving time, if they save steps from calculations with larger basis sets by providing a better initial guess. [1]

The number of required calculations in HF scales as N^4 where N is the number of basis functions, but linear scaling HF implementations have also been developed. These implementations are not fast, but they are more efficient than the normal N^4 scaling version for large enough systems. [1]

HF is the base of most ab initio and semi-empirical methods. HF itself is an ab initio method and further ab initio methods can be formed by improving its accuracy. Semi-empirical methods can be formed from HF by adding more approximations into it. [2,3]

2.3.2 Multiconfiguration self-consistent field theory

To improve upon HF theory, a wave function can be constructed as a linear combination of determinants.

$$\Psi = c_0 \Psi_{HF} + c_1 \Psi_1 + c_2 \Psi_2 + \cdots, \qquad (37)$$

where Ψ_{HF} is the HF determinant. This way the electron correlation energy ignored by HF can be taken at least partially into account. It is not necessary to have Ψ_{HF} as a part of the linear combination, but it provides a good reference point for creating other determinants and it usually dominates the linear combination as the largest individual term. In a system with degenerate frontier orbitals additional determinants can be obtained by changing which frontier orbital is populated. Frontier orbitals are the highest energy occupied orbitals and the lowest energy unoccupied orbitals. Another way

to obtain additional determinants is to consider other possible excitations of the electrons. [1, 3]

As the normal SCF process does not optimize the shapes of virtual orbitals (orbitals which are not occupied in the ground state), a multiconfiguration self-consistent field (MCSCF) process must be used. MCSCF optimizes both the coefficients of the determinants in Eq.(37) and the MOs that those determinants are made from. A configuration or configuration state function (CSF) refers to the molecular spin state and the occupation numbers of the orbitals. For closed shell singlets CSFs can be presented as single determinants, but in many open shell systems they can only be presented as a combination of determinants. [1, 2]

The selection of orbitals for MCSCF requires a careful consideration of the molecule's chemistry. This selection is called active space and it is usually marked as (m,n) where m is the number of electrons and n is the number of orbitals. In complete active space (CAS) calculations all possible arrangements of the electrons are taken into account, leading to N possible CSFs singlets:

$$N = \frac{n!\,(n+1)!}{\left(\frac{m}{2}\right)!\left(\frac{m}{2}+1\right)!\left(n-\frac{m}{2}\right)!\left(n-\frac{m}{2}+1\right)!}, \tag{38}$$

where m is the number of electrons and n is the number of orbitals. CAS calculations become very demanding quickly as the number of electrons and orbitals rises. Symmetry can be used to limit the number of required CSFs for a CAS calculation. [1]

CAS calculations can be simplified by reducing the size of the CAS while including a restricted active space (RAS). While all possible electron configurations in CAS are allowed only a limited number of excitations, usually only one or two electrons, are allowed to or from the RAS orbitals, thus, limiting the number of possible RAS configurations considerably. The orbitals remaining outside of CAS and RAS are inactive orbitals. They are limited to the occupation numbers of two or zero and their occupation numbers do not change between the different CSFs. The CAS and RAS calculations can be further simplified by freezing the core orbitals to the shapes calculated at the HF level. Thus, MCSCF orbitals can be divided into four categories: Frozen orbitals, Inactive orbitals, RAS orbitals and CAS orbitals. [1]

Methods that use multi determinant wave functions as their reference point are known as multi-reference methods. Most of them are based on using a MCSCF wave function instead of an HF wave function as the starting point. [2]

2.3.3 Møller-Plesset perturbation theory

Using Rayleigh-Schrödinger perturbation theory, it is possible to use the exact eigenfunctions and eigenvalues of a simplified operator to estimate the eigenfunctions and eigenvalues of a more complete operator. An operator $\mathbf{A}$ can be written as

$$\mathbf{A} = \mathbf{A}^{(0)} + \lambda\mathbf{V}. \tag{39}$$

Where $\mathbf{A}^{(0)}$ is an operator with known eigenfunctions, $\mathbf{V}$ is a perturbing operator and λ is a dimensionless parameter the value of which varies between 0 and 1. The eigenfunctions and eigenvalues can then be expanded as Taylor series in λ. These series are usually written as

$$\Psi_0 = \Psi_0^{(0)} + \lambda\Psi_0^{(1)} + \lambda^2\Psi_0^{(2)} + \lambda^3\Psi_0^{(3)} + \cdots \tag{40}$$

and

$$a_0 = a_0^{(0)} + \lambda a_0^{(1)} + \lambda^2 a_0^{(2)} + \lambda^3 a_0^{(3)} + \cdots, \tag{41}$$

where the terms with superscripts (n) are called nth-order corrections to the zeroth order term. By using Eq.(40) and Eq.(41) the equation

$$(\mathbf{A}^{(0)} + \lambda\mathbf{V})|\Psi_0\rangle = a|\Psi_0\rangle \tag{42}$$

can be written as

$$\begin{aligned}
(\mathbf{A}^{(0)} + \lambda\mathbf{V})|\Psi_0^{(0)} + \lambda\Psi_0^{(1)} + \lambda^2\Psi_0^{(2)} + \cdots\rangle = \\
(a_0^{(0)} + \lambda a_0^{(1)} + \lambda^2 a_0^{(2)} + \cdots)|\Psi_0^{(0)} + \lambda\Psi_0^{(1)} + \lambda^2\Psi_0^{(2)} + \cdots\rangle.
\end{aligned} \tag{43}$$

As Eq.(43) is true for all values of λ, the terms with equal powers of λ can be collected to form the equations

$$\begin{aligned}
\lambda^0: \quad & \mathbf{A}^{(0)}|\Psi_0^{(0)}\rangle = a^{(0)}|\Psi_0^{(0)}\rangle \\
\lambda^1: \quad & \mathbf{A}^{(0)}|\Psi_0^{(1)}\rangle + \mathbf{V}|\Psi_0^{(0)}\rangle = a^{(0)}|\Psi_0^{(1)}\rangle + a^{(1)}|\Psi_0^{(0)}\rangle \\
\lambda^2: \quad & \mathbf{A}^{(0)}|\Psi_0^{(2)}\rangle + \mathbf{V}|\Psi_0^{(1)}\rangle = a^{(0)}|\Psi_0^{(2)}\rangle + a^{(1)}|\Psi_0^{(1)}\rangle + a^{(2)}|\Psi_0^{(0)}\rangle \\
\vdots \\
\lambda^n: \quad & \mathbf{A}^{(0)}|\Psi_0^{(n)}\rangle + \mathbf{V}|\Psi_0^{(n-1)}\rangle = \sum_{i=0}^{n} a^{(i)}|\Psi_0^{(n-i)}\rangle,
\end{aligned} \tag{44}$$

which can be used to solve the various nth-order corrections.

$$a_0^{(1)} = \langle\Psi_0^{(0)}|\mathbf{V}|\Psi_0^{(0)}\rangle \tag{45}$$

is the first-order correction and

$$a_0^{(2)} = \sum_{j>0} \frac{|\langle \Psi_j^{(0)}|\mathbf{V}|\Psi_0^{(0)}\rangle|^2}{a_0^{(0)} - a_j^{(0)}} \tag{46}$$

the second-order correction. [1, 2]

Based on the Rayleigh-Schrödinger perturbation theory Møller and Plesset developed the Møller and Plesset perturbation theory (MP) in 1934 [14]. It is often referred as the MPn where n is the order at which the perturbation theory is truncated at. In MP the operator $\mathbf{A}$ is the Hamiltonian $\mathbf{H}$, Ψ_0 is the HF wave function and the $\mathbf{H}^{(0)}$ is a sum of one-electron Fock operators:

$$\mathbf{H}^{(0)} = \sum_{i=1}^{n} f_i, \tag{47}$$

where n is the number of basis functions and f is a Fock operator as defined by Eq.(24). The perturbation $\mathbf{V}$ can be written as

$$\mathbf{V} = \mathbf{H} - \mathbf{H}^{(0)} = \sum_{i}^{occ.} \sum_{j>i}^{occ.} \frac{1}{r_{ij}} - \sum_{i}^{occ.} \sum_{j}^{occ.} \left(J_{ij} - \frac{1}{2}K_{ij} \right), \tag{48}$$

where *occ.* refers to occupied orbitals. [1, 3]

MP1 is equivalent to HF in determining the energy, as can be seen from:

$$
\begin{aligned}
a^{(0)} + a^{(1)} &= \langle \Psi_0^{(0)}|\mathbf{H}^{(0)}|\Psi_0^{(0)}\rangle + \langle \Psi_0^{(0)}|\mathbf{V}|\Psi_0^{(0)}\rangle \\
&= \langle \Psi_0^{(0)}|\mathbf{H}^{(0)} + \mathbf{V}|\Psi_0^{(0)}\rangle \\
&= \langle \Psi_0^{(0)}|\mathbf{H}|\Psi_0^{(0)}\rangle \\
&= E_{HF}.
\end{aligned}
\tag{49}
$$

The second-order energy correction in MP is

$$a_0^{(2)} = \sum_{i}^{occ.} \sum_{j>i}^{occ.} \sum_{a}^{vir.} \sum_{b>a}^{vir.} \frac{[(ij|ab) - (ia|jb)]^2}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \tag{50}$$

and the MP2 energy is the sum of $a^{(0)}$, $a^{(1)}$ and $a^{(2)}$. The *vir.* in Eq.(50) refers to virtual orbitals. [1]

The MP2 method scales roughly as N^5. All orders of the MPn theory are size-consistent, meaning that doubled MPn energy of a molecule is the same as the energy of two of the same molecules with a long distance between them. [1]

MPn theory does have some problems as perturbation theory works best when the perturbation is small, but in MPn the perturbation is the full electron-electron repulsion energy, which is a large part of the total energy. Because of this MPn does not give good correlation energy values. Also, as the MPn method is not variational it can sometimes give correlation energies that are too large instead of too small. [1]

Higher orders of the perturbation theory can be used to improve the accuracy and convergence behavior of the MPn theory, but there is no guarantee of convergence even at infinite order when using a finite basis set. MP3 level is not much more expensive than MP2, but it does not offer much improvement over MP2 either. MP4 level is very costly as it scales as N^7, but it is also very accurate, usually accounting for $>95\%$ of the correlation energy with a good basis set. [1]

The results of the different MPn levels tend to oscillate as a function of n. MP2 offers improvement over HF, but the energy correction it gives over HF tends to be a little too big. MP3 corrects the results back towards HF by a variable amount and MP4 corrects them away from the HF energy, but only by little in favorable situations. For a medium sized basis set the results of MP2 level are often closer to the true value than the results of MP3 level. [1,2]

2.3.4 Configuration interaction theory

Configuration interaction (CI) is similar to MCSCF, but in it only the coefficients of the determinants in Eq.(37) are optimized. The MOs that those determinants are made of are kept fixed after the initial HF calculation. Because of this CI is not as accurate as MCSCF, but it is also computationally less demanding than MCSCF. [2,3]

Full configuration interaction (CI) is a CAS calculation that takes into account all electrons and includes all orbitals in the complete active space. The number of CSFs required for a full CI is extremely large even for a relatively small basis set. Thus, full CI calculations with large basis sets are done only for very small molecules. [1]

Full CI is the best possible calculation that can be done for a given basis set. The difference between Full CI energy and HF energy is the basis-set correlation energy and as the basis gets bigger the basis-set correlation energy approaches the electron correlation energy. With an infinite basis set full CI gives the exact solution of the non-relativistic, Born-Oppenheimer, time independent Schrödinger equation. [1,3]

As full CI is usually quite impractical, a more practical option is to limit the number of allowed excitations in the CI. For this purpose, it is useful to

write Eq.(37) in the form

$$\Psi = a_0\Psi_{HF} + \sum_i^{occ.}\sum_r^{vir.} a_i^r\Psi_i^r + \sum_{i<j}^{occ.}\sum_{r<s}^{vir.} a_{ij}^{rs}\Psi_{ij}^{rs}r + \cdots,\qquad(51)$$

where i and j are occupied MOs in the Ψ_{HF}, r and s are virtual MOs in the Ψ_{HF} and additional CSFs are generated by exciting an electron from the indicated occupied orbital to the indicated virtual orbital. So, the first term is the HF wave function, the second term includes all possible single electron excitations, the third term includes all possible two-electron excitations and so on. The N different energies E of the CI wave functions can then be solved from

$$\begin{vmatrix} H_{11} - E & H_{12} - E & \cdots & H_{1N} - E \\ H_{21} - E & H_{22} - E & \cdots & H_{2N} - E \\ \vdots & \vdots & \ddots & \vdots \\ H_{N1} - E & H_{N2} - E & \cdots & H_{NN} - E \end{vmatrix} = 0,\qquad(52)$$

where

$$H_{mn} = \langle\Psi_m|H|\Psi_n\rangle,\qquad(53)$$

H is the electronic Hamiltonian and Ψ_i is Ψ_{HF} when $i = 1$ and then Ψ_i goes through first all the one-electron excitations and then all the two-electron excitations and so on as $i \to N$. [1]

As the electronic Hamiltonian contains only one-electron and two-electron operators, two determinants that differ by more than two MOs will always result in an overlap integral between two orthogonal MOs, which is zero. Thus, the CI matrix elements H_{ij} can have a value other than zero only when the two determinants differ by 0, 1 or 2 MOs. The resulting MOs can then be written as one- and two-electron integrals over MOs. These are known as Slater-Condon rules. [2]

Based on the Slater-Condon rules all matrix elements between the ground-state wave functions Ψ_{HF} and one-electron excitation wave functions Ψ_i^r are zero:

$$\begin{aligned} H_{1n} &= \langle\Psi_{HF}|H|\Psi_i^r\rangle \\ &= \langle\phi_r|F|\phi_i\rangle \\ &= \varepsilon\langle\phi_r|\phi_i\rangle \\ &= \varepsilon_i\delta_{ir} \\ &= 0, \end{aligned}\qquad(54)$$

where F is the Fock operator, i and r represent the occupied and the virtual MOs of the one-electron excitation wave function and ε_i is the eigenvalue of

the MO. The third line is based on the fact that the MOs are eigenfunctions of the Fock operator and the last line on the fact that for a one-electron excitation wave function r cannot be equal to i. This result is known as the Brillouin's theorem. [1]

The case where only single excitations are included in the CI is called CI singles (CIS). It is not useful for ground-state calculations as based on the Brillouin's theorem the ground-state HF root is unaffected by the single excitations. Double excitations do affect the ground-state energy, so CI doubles (CID) is a more useful method. Triple excitations have no interaction with the ground state, but they mix with the double excitations and can affect the lowest energy eigenvalue that way. There are however a lot of triple excitations and because of that their inclusion is difficult to handle. Thus, triples and higher-level excitations are rarely used. Single excitations also mix with the double excitations and as there are comparatively few singles their inclusion is not too difficult. This is called CI singles and doubles (CISD). [1]

The CISD scales as N^6 with respect to the size of the used basis set. As with other methods, symmetry and the freezing of some orbitals can be used to reduce the required computational effort. [1]

The CISD method is variational and thus the CISD energy is an upper bound of the exact energy. But CISD is not size consistent, which means that the doubled CISD energy of a molecule can be smaller than the energy of two of the same molecules with a long distance between them. [1, 2]

2.3.5 Coupled-cluster theory

The coupled cluster (CC) theory developed by Čížek in 1966 [15] is a technique for estimating electron correlations. It is based on describing the full CI wave function in the form

$$\Psi = e^{\mathbf{T}} \Psi_{HF}, \tag{55}$$

where the exponential operator $e^{\mathbf{T}}$ is expanded as a series

$$e^{\mathbf{T}} = 1 + \mathbf{T} + \frac{1}{2!}\mathbf{T}^2 + \frac{1}{3!}\mathbf{T}^3 + \cdots \tag{56}$$

and the cluster operator $\mathbf{T}$ is

$$\mathbf{T} = \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 + \cdots + \mathbf{T}_n, \tag{57}$$

where n is the number of electrons and the $\mathbf{T}_i$ operator describes all possible determinants with i excitations. The CC energy is given by the equation:

$$E_{CC} = \langle \Psi_{HF} | \mathbf{H} | e^{\mathbf{T}} \Psi_{HF} \rangle. \tag{58}$$

25

The advantage CC offers over CI is that it takes into account products of excitation operators and is thus size consistent. However, it is not variational. [1,3]

CC doubles (CCD) is equivalent to CID with the products of the excitation operators added to it. Adding singles into CCD to form CCSD is worth the increased cost to gain better accuracy. CCSD scales as N^6. Adding the triples term $\mathbf{T}_3$ into CCSD forms the CC singles doubles and triples (CCSDT) method, but it is impractical for all but the smallest systems as it scales as N^8. There are various methods for estimating the effect of the triples term through perturbation theory. The most successful of these methods is known as CCSD(T). The CCSD(T) tends to slightly overestimate the triples correction, but it does so by an amount equal to the ignored quadruples. Thanks to this favorable cancellation CCSD(T) is extremely efficient in most cases and it is considered one of the best single reference calculation methods. [1]

2.4   Comparison of methods

HF, MCSCF, MPn, CI and CC are all ab initio methods. MCSCF is a multi-reference method, while HF, MPn, CI and CC are single-reference methods, though MPn, CI and CC do have multi-reference variants based on the MCSCF wave function. [1]

The HF theory tends to overemphasize the occupation of bonding orbitals leading to too short bonds between atoms, while methods that take the electron correlation into account tend to have longer bond lengths in order to lower the energy. Electron correlation methods also have a greater dependency on the basis set quality than HF methods. This has led to development of basis sets that are optimized for correlated methods. [1]

Generally, MP2 is an excellent level for geometry optimization of minima that include correlation energy. Bond angles are already quite accurate at HF level so MP2 offers little improvement for them. The geometric improvements offered by methods beyond MP2 level tend to be so small for equilibrium structures that they are not worth the added cost. MP2 is also very efficient for energy differences between minima, though this is because most minima are already well described by HF wave functions. [1]

A rough accuracy order of the different single-reference ab initio methods is

$$
\begin{aligned}
HF \ll MP2 \sim MP3 \sim CCD < CISD \\
< CCSD < MP4 < CCSD(T).
\end{aligned}
\tag{59}
$$

The accuracy of multi-reference methods is more difficult to evaluate as it depends heavily on the size of the used reference wave function. The scaling

behavior of the different single-reference ab initio methods as a function of the number basis functions N is given by the following table.

Scaling	Methods
N^4	HF
N^5	MP2
N^6	MP3, CISD, CCSD
N^7	MP4, CCSD(T)
N^8	MP5, CISDT, CCSDT
N^9	MP6
N^{10}	MP7, CISDTQ, CCSDTQ

Table 1: Scaling behavior of different ab initio methods

As seen from the Table 1 and Eq.(59), there is usually little point in doing MP3 or CISD calculations when the superior CCSD calculations have roughly the same cost. Though similar scaling behavior is not the same as similar overall time, but for small and medium sized systems they tend to be similar. [1, 2]

For ground state calculations MPn and CC are usually the preferred methods. But for excited states CI and MCSCF based methods are used more, as MPn and CC have some troubles with them. There are, however, some variants of the CC method that are good at handling excited states. [2]

3 Experiment

The experimental spectra used in this thesis is from an experiment done at the FinEstBeAMS beamline [16, 17] at the MAX IV synchrotron in Sweden in March 2020. In the experiment the valence, oxygen 1s, nitrogen 1s and carbon 1s spectra m-ABA were measured using vaporized powder samples. The powder was placed in a stainless steel crucible, which was placed inside a resistively heated oven, which was then used for vaporizing the samples inside the spectrometry equipment. To ensure proper pressure for the measurements the m-ABA sample was heated to ~ 120 °C (393.15 K). The molecules were then ionized with linearly polarized synchrotron radiation and the kinetic energies of the resulting electrons were analyzed with a Scienta R4000 hemispherical analyzer. [18]

The valence orbital spectra were measured using a photon energy of 100 eV and using water evaporated from the sample for calibration. The oxygen 1s spectra were measured using a photon energy of 600 eV, a resolution of 490 meV and using O_2 gas for calibration. The nitrogen 1s spectra were measured using a photon energy of 450 eV, a resolution of 350 meV and using N_2 gas for calibration. The carbon 1s spectra were measured using a photon energy of 350 eV, a resolution of 140 meV and using CO_2 gas for calibration. [18]

The measured kinetic energies of electrons were translated into binding energies using Einstein's Photoelectric Law:

$$
\begin{aligned}
E_j &= h\nu - I_j \\
I_j &= -(E_j - h\nu),
\end{aligned}
\tag{60}
$$

where E_j is the kinetic energy of the emitted electron, $h\nu$ is the energy of the ionizing photons and I_j is the binding energy of the emitted electron. The resulting binding energy spectra are shown by the following Figures 2-5.
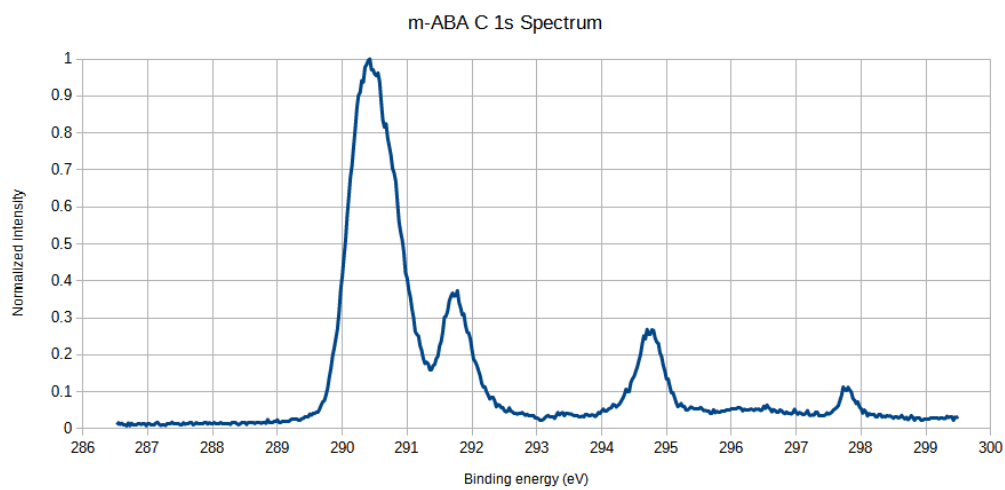
Figure 2: C 1s Spectrum of m-ABA
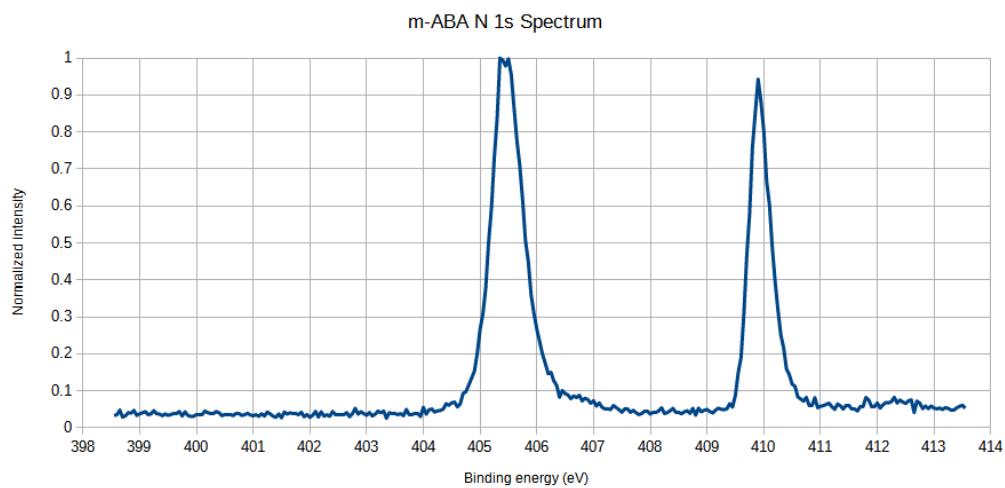
Figure 3: N 1s Spectrum of m-ABA
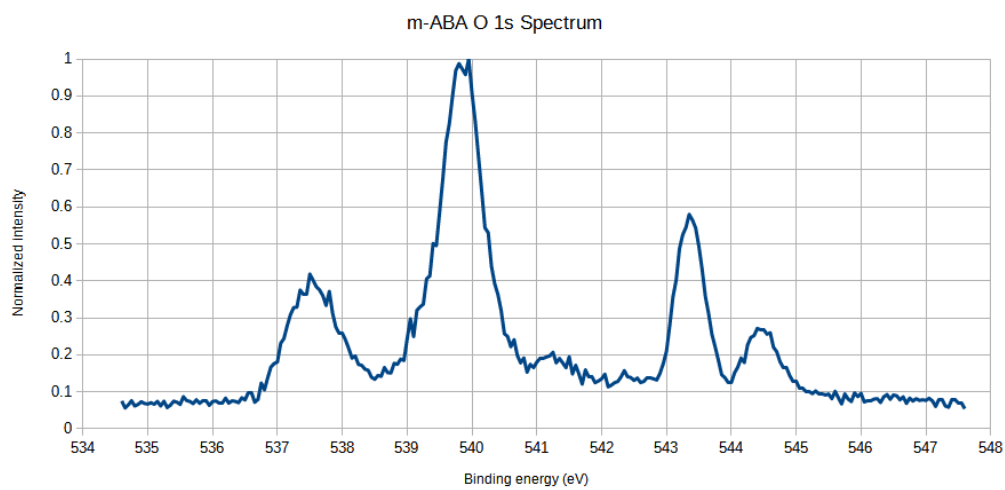
Figure 4: O 1s Spectrum of m-ABA
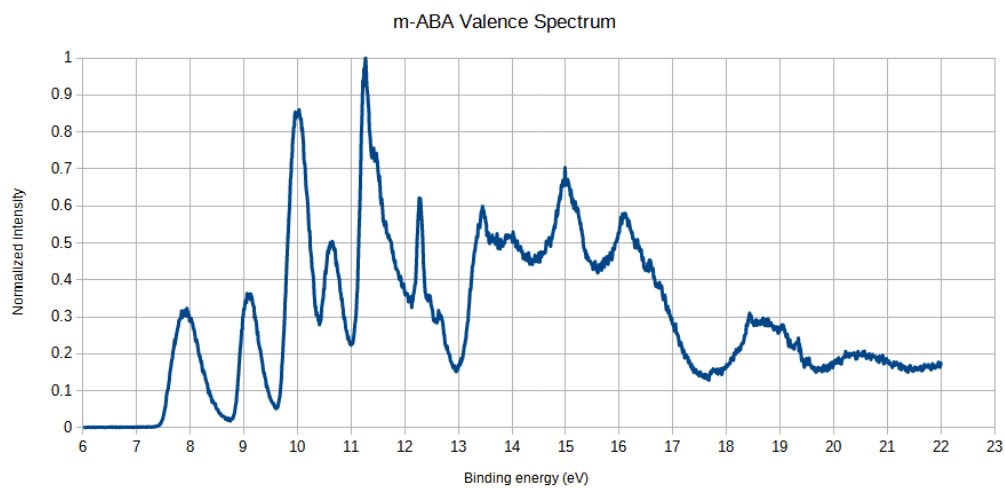
Figure 5: Valence Spectrum of m-ABA

The binding energy spectra were then calibrated by shifting them so that the peak caused by the calibration material matches its binding energy's reference value. The binding energy reference values used for calibration are as follows. Water's $1b_1$ orbital has a binding energy of 12.62 ± 0.005 eV [19]. The accuracy of water's $1b_1$ orbitals binding energy was given as the accuracy of its associated frequency in the form ± 40 cm^{-1}, which was translated into energy with

$$E = h\nu c, \tag{61}$$

where h is the Planck's constant, ν the frequency and c the speed of light. For the O_2 gas the O 1s ionization's $^4\Sigma-$state orbital has a binding energy of 543.39 ± 0.05 eV and the $^2\Sigma-$state orbital has a binding energy of 544.43 ± 0.05 eV [20]. As there is some water present in the samples, its O 1s orbital also causes a peak in the oxygen 1s spectra and the O 1s orbital of water has a binding energy of 539.9 ± 0.02 eV [21]. The N 1s orbital of N_2 has a binding energy of 409.9 ± 0.1 eV [22]. The C 1s orbital of CO_2 has a binding energy of 297.699 ± 0.03 eV [23].

The peaks of the 1s spectra were then identified with the help of the calculated orbital energies of 1 m-ABA and 2 m-ABA and the Koopmans' theorem Eq.(21). The valence peaks are the results of heavy mixing of AOs so they cannot be identified to originate from any single atom. The calibrated m-ABA spectra with the 1s peaks named are shown by the following Figures 6-9.
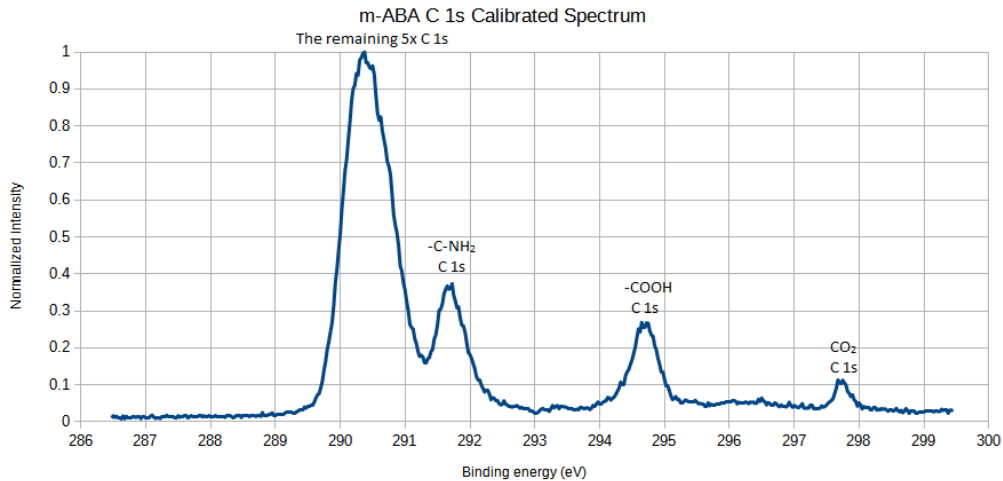
Figure 6: Calibrated C 1s Spectrum of m-ABA with the peaks named
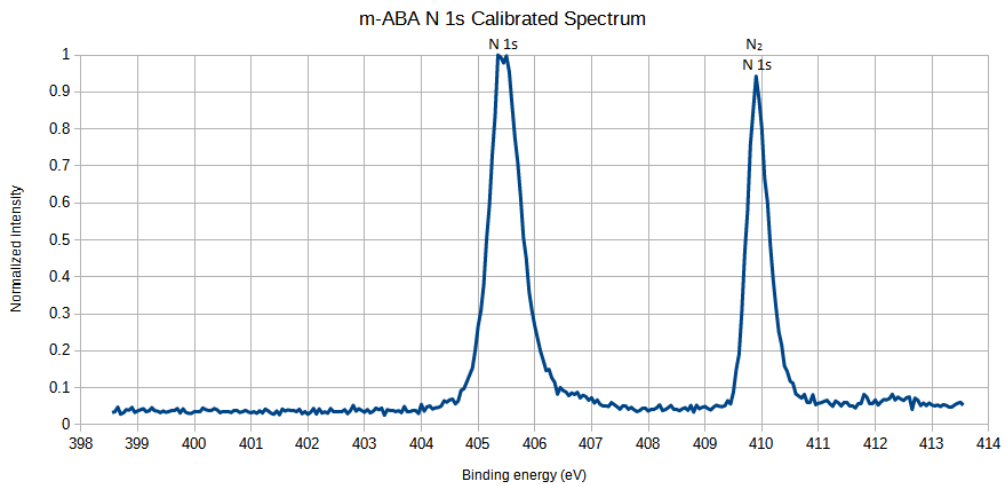
Figure 7: Calibrated N 1s Spectrum of m-ABA with the peaks named
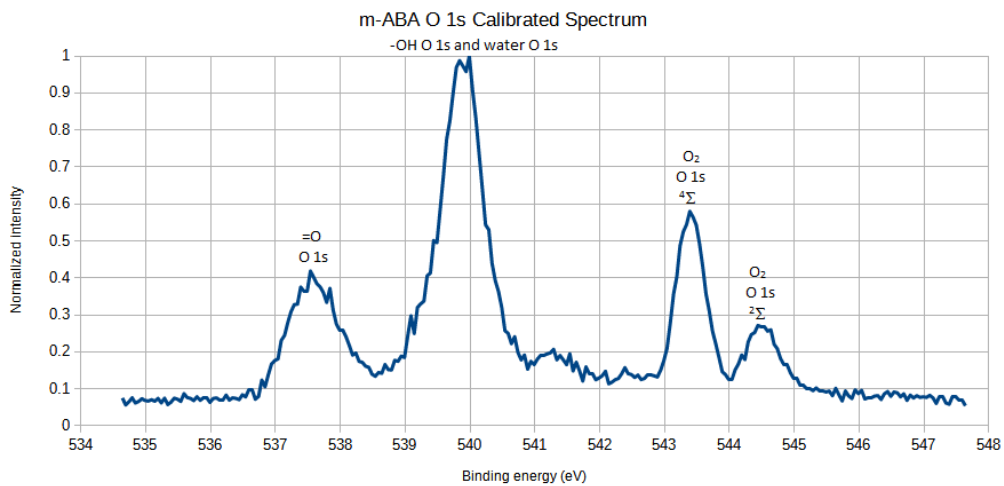
Figure 8: Calibrated O 1s Spectrum of m-ABA with the peaks named
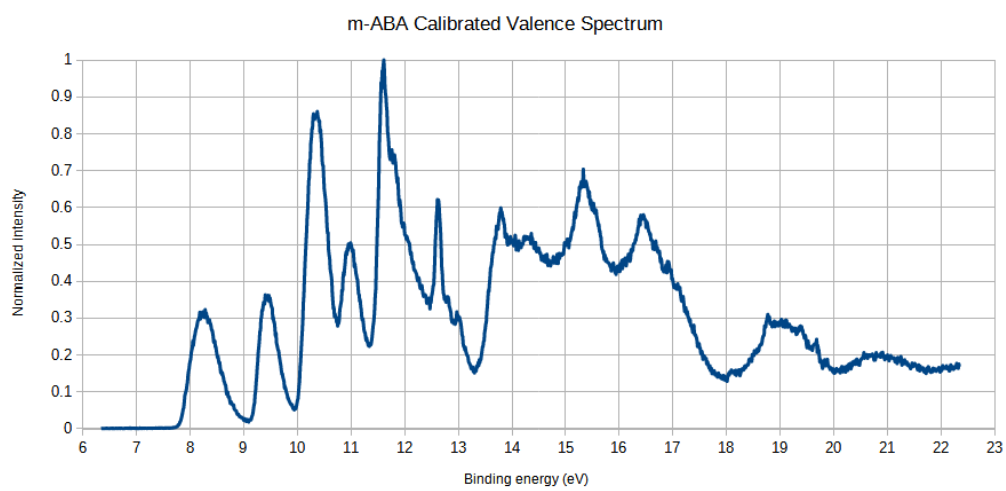
Figure 9: Calibrated Valence Spectrum of m-ABA

4 Calculations

The calculations were done on the Puhti supercomputer at CSC [24] using the MOLPRO program [25–35]. The starting guesses for geometries were done by constructing rough guesses of the 1 m-ABA, 2 m-ABA and p-ABA molecules geometries with the Macmolplt program [36] and then using Macmolplt to print out the coordinates of the atoms. The starting geometries can be read from the Tables 2, 3 and 4 in Appendix A and Figure 10 shows a graphical representation of them.

(a) 1 m-ABA   (b) 2 m-ABA   (c) p-ABA
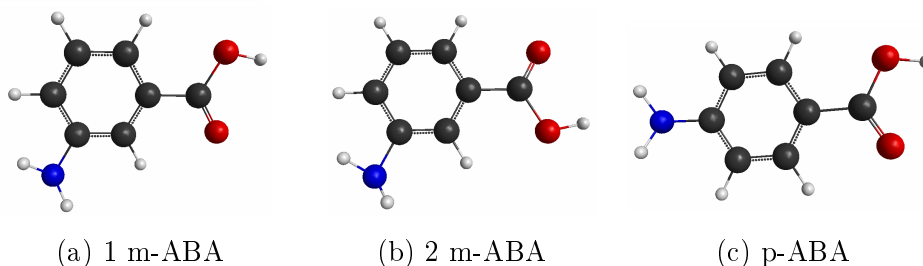
Figure 10: Starting geometries

Originally the plan was to optimize the geometries of 1 m-ABA, 2 m-ABA and p-ABA molecules first with the HF method, then the MP2 method and finally with the CCSD(T) method. Following basis sets were to be used for each method: cc-pVDZ, cc-pVTZ, cc-pVQZ, cc-pwCVDZ, cc-pwCVTZ and cc-pwCVQZ. Always using the results of previous calculations as the starting geometry for the next basis set or method. Due to memory limitations the CCSD(T) calculations could not be done, so the CCSD method had to be used instead of CCSD(T). The CCSD calculations using cc-pwCVQZ basis set also turned out to be impossible because of calculation time, memory and disk space limitations. Because of this the CCSD calculations using cc-pwCVnZ basis sets were not done for the molecules besides 1 m-ABA for which the cc-pwCVDZ and cc-pwCVTZ calculations were already done at this point and I decided to use the CCSD calculations using cc-pVQZ basis set as the final level of optimization. The total energies of 1 m-ABA, 2 m-ABA and p-ABA can be seen from the Table 5 in Appendix B and Figure 11 shows a graphical comparison of the total energies.
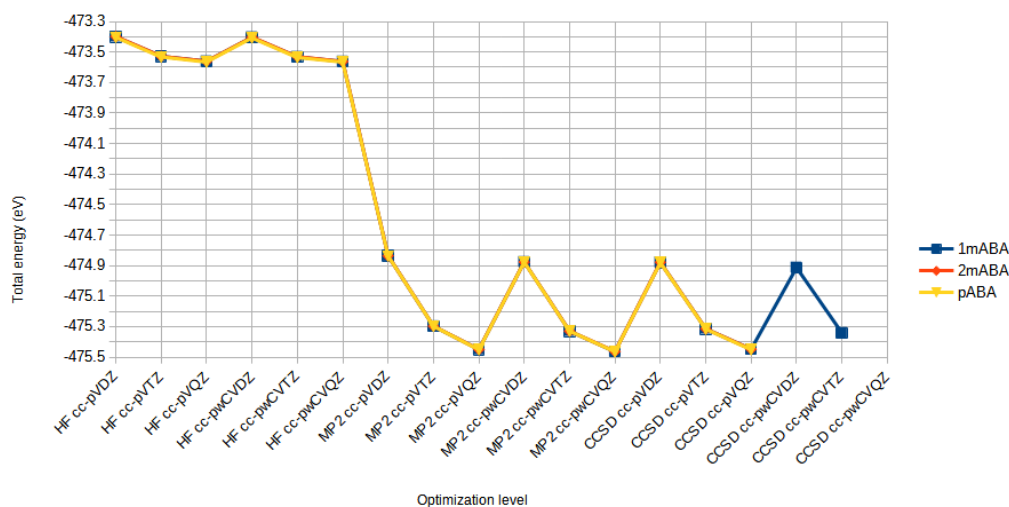
Figure 11: Total energy at different optimization levels for 1 m-ABA, 2 m-ABA and p-ABA

The Gabedit program [37] was used display the final geometries, the shapes of the orbitals and to form density of states spectra using calculated orbital energies convoluted with a Gaussian line shape for 1 m-ABA, 2 m-ABA and p-ABA.

The geometries at the end of HF calculations can be read from the the Tables 6, 7 and 8 in Appendix C and the geometries at the end of MP2 calculations can be read from the the Tables 9, 10 and 11 in Appendix D. The final geometries can be read from the Tables 12, 13 and 14 in Appendix E and Figure 12 shows a graphical representation of them.

(a) 1 m-ABA   (b) 2 m-ABA   (c) p-ABA
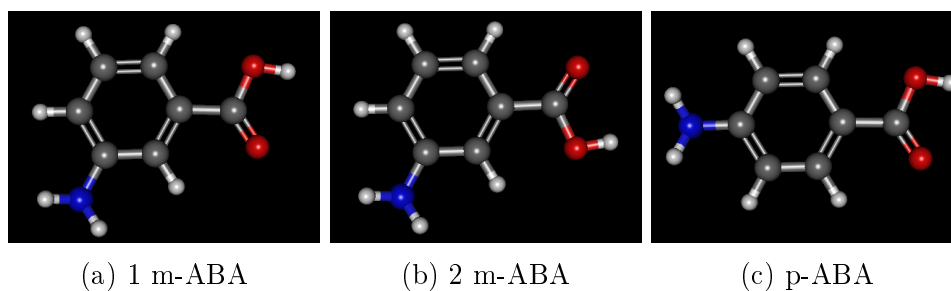
Figure 12: Final geometries

The calculated CCSD cc-pVQZ spectra can be seen from the following Figures 13-24.

35
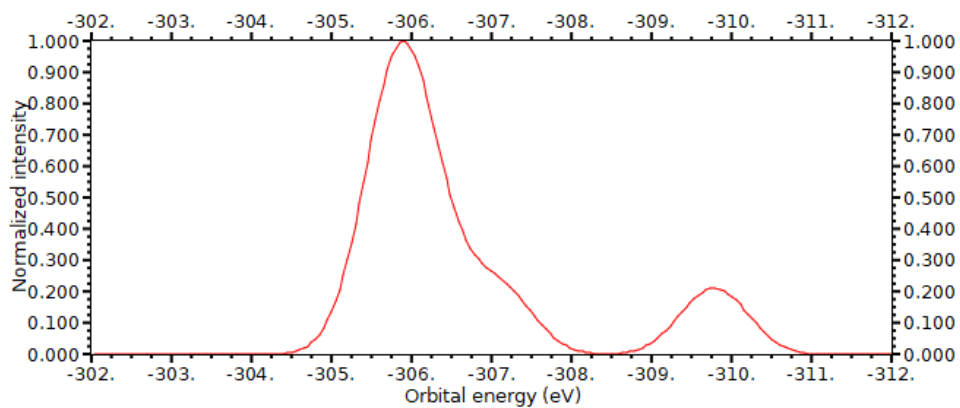
Figure 13: Calculated C 1s Spectrum of 1 m-ABA
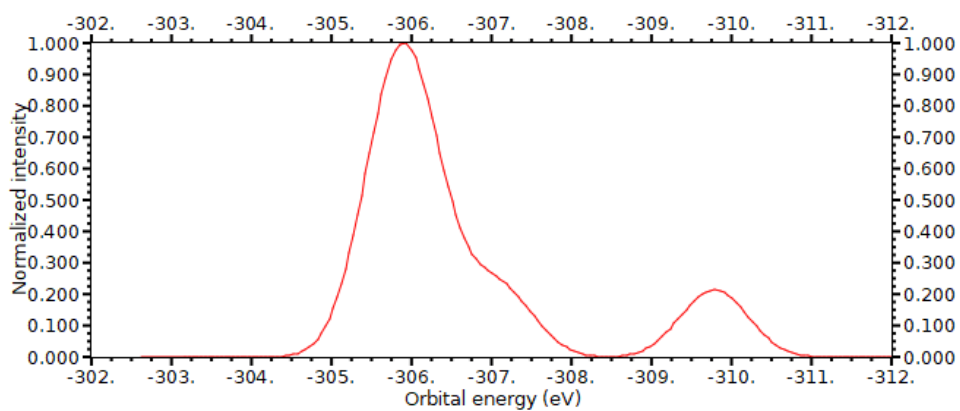
Figure 14: Calculated C 1s Spectrum of 2 m-ABA
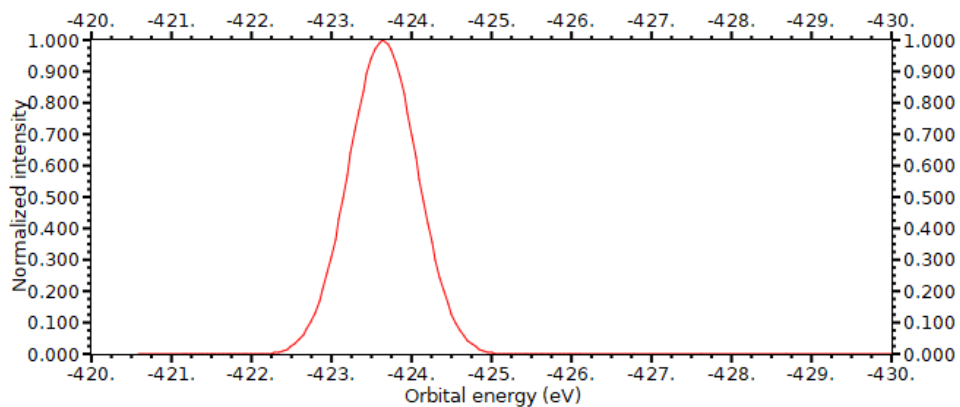
Figure 15: Calculated N 1s Spectrum of 1 m-ABA
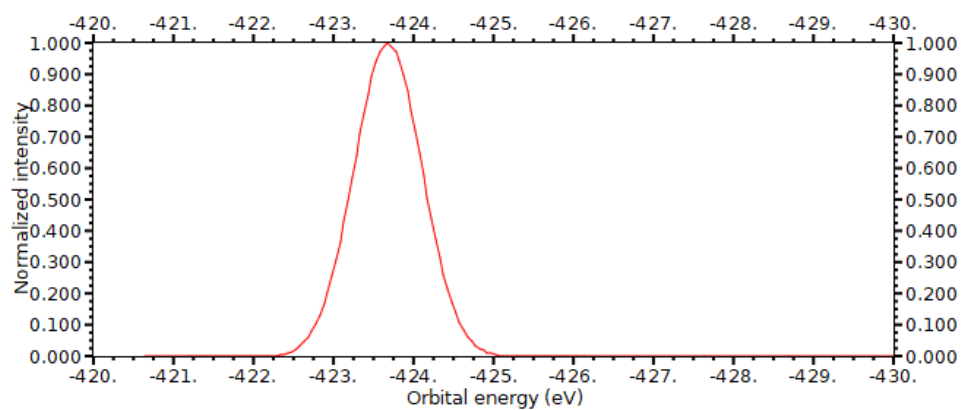
Figure 16: Calculated N 1s Spectrum of 2 m-ABA
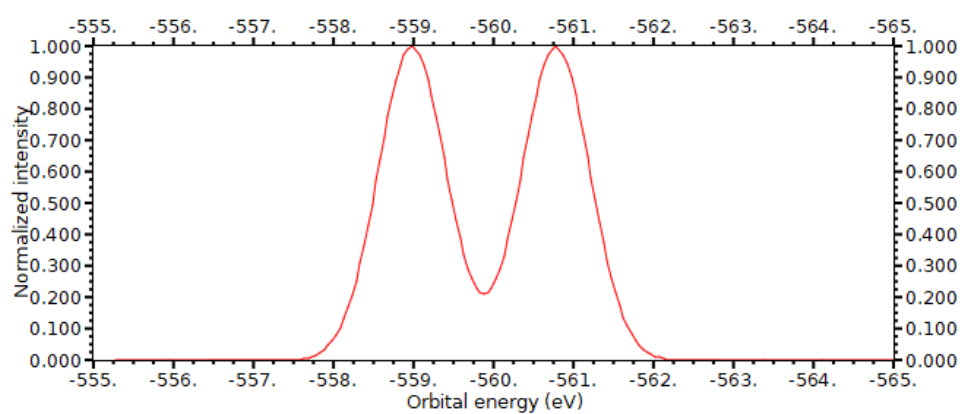
Figure 17: Calculated O 1s Spectrum of 1 m-ABA
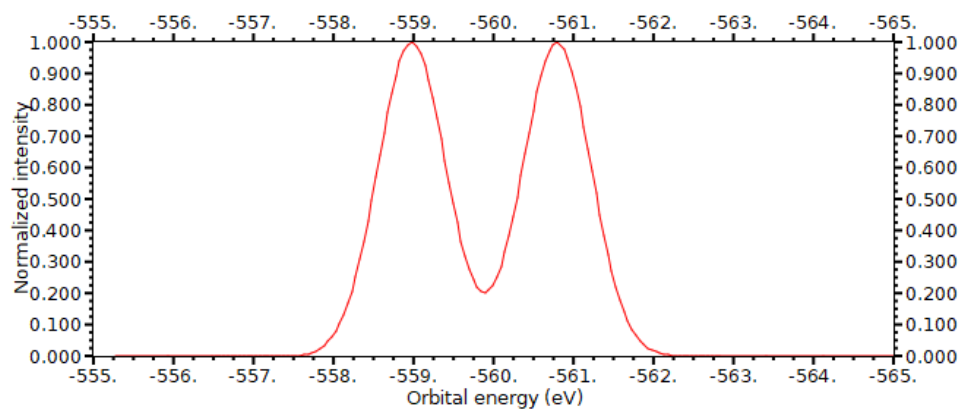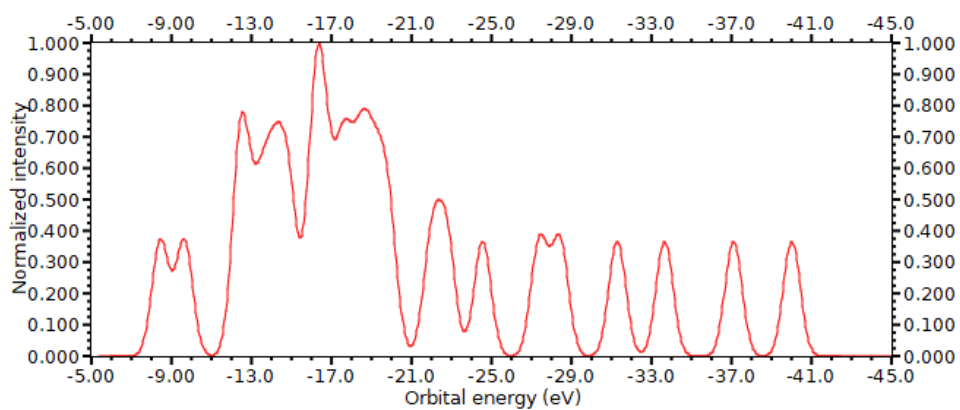

Figure 18: Calculated O 1s Spectrum of 2 m-ABA
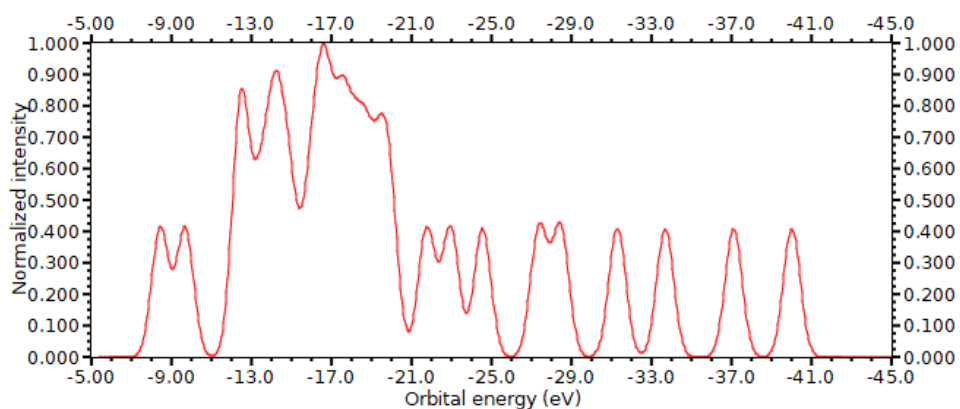
Figure 19: Calculated Valence Spectrum of 1 m-ABA
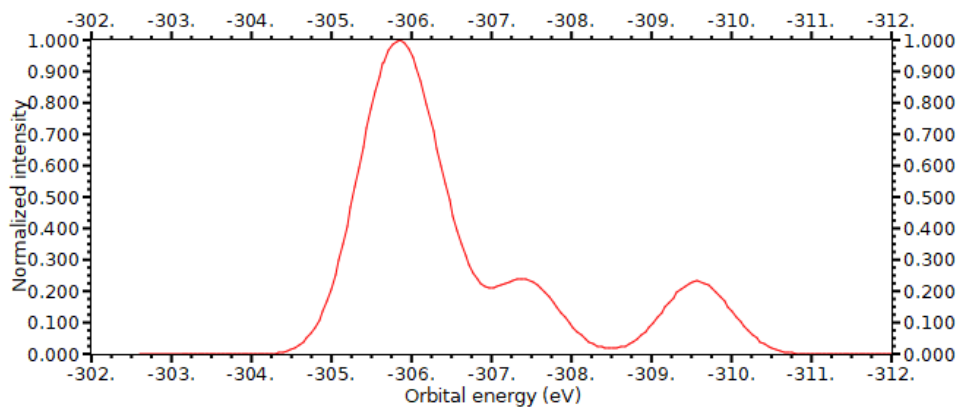

Figure 20: Calculated Valence Spectrum of 2 m-ABA
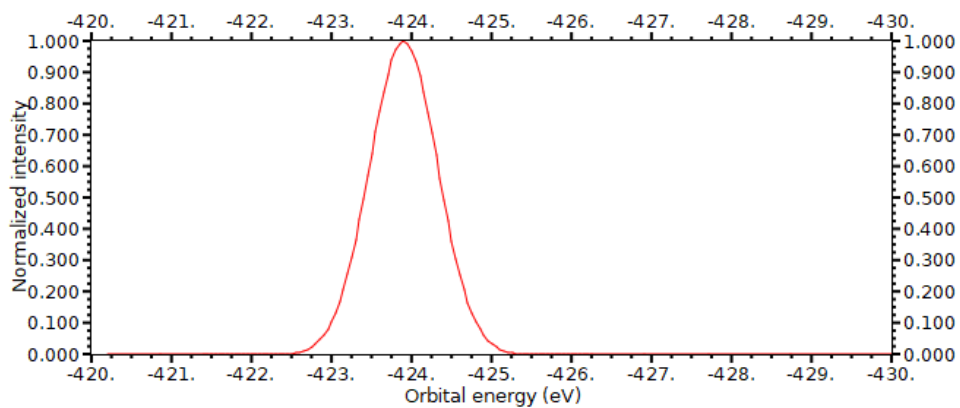

Figure 21: Calculated C 1s Spectrum of p-ABA

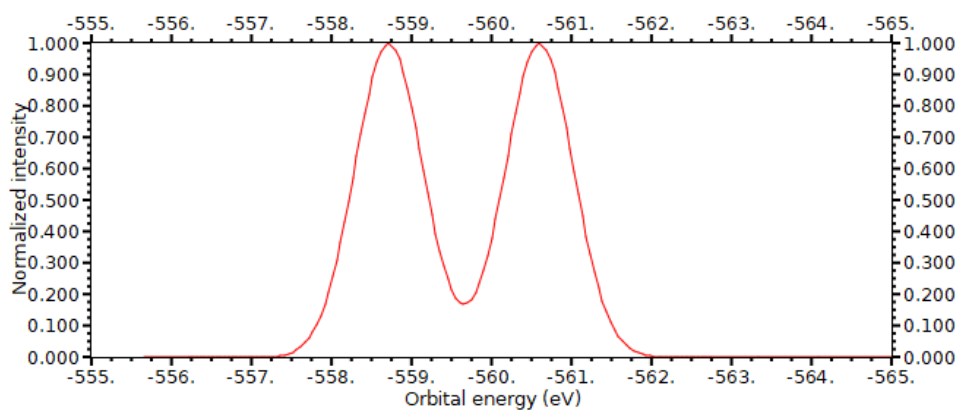Figure 22: Calculated N 1s Spectrum of p-ABA

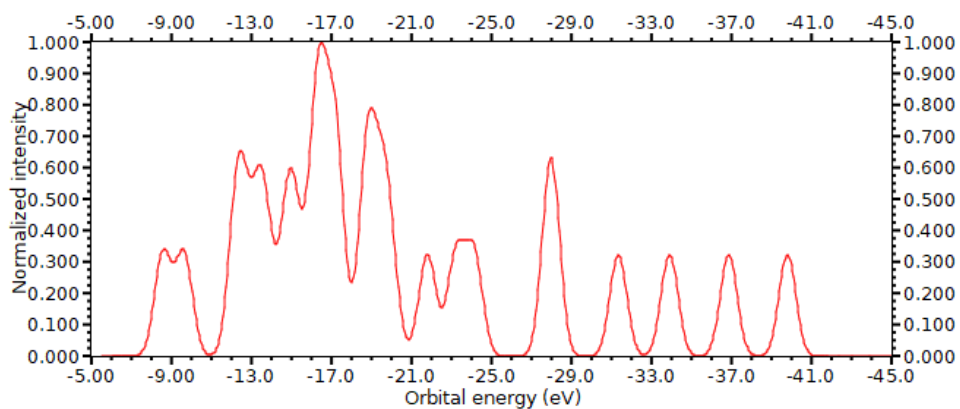Figure 23: Calculated O 1s Spectrum of p-ABA

Figure 24: Calculated Valence Spectrum of p-ABA

The x-axes of these spectra have been mirrored in order to make comparison with the binding energy spectra easier. The calculated CCSD cc-pVQZ values of the orbital energies for 1 m-ABA, 2 m-ABA and p-ABA that were used to form those spectra can be seen from the table in Appendix F. The first ten energies in the table are mostly from the different O, N and C atoms' 1s orbitals. The rest marked as valence are the results of heavy mixing of the AOs and as such cannot be said to originate from any single atom. The shapes of the orbitals can be seen from Figures 25-27 in Appendices G-I.

The 1 m-ABA calculations were done first and because of this they required multiple restarts as I adjusted the memory and time allocations for them. The HF calculations took $\sim$ 15 h 36 min and required $\sim$ 880.9 GB of disk space for 1 m-ABA, $\sim$ 14 h 56 min and required $\sim$ 813.53 GB of disk space for 2 m-ABA and $\sim$ 19 h 7 min and required $\sim$ 790.77 GB of disk space for p-ABA. The time and disk space estimations for 1 m-ABA calculations are affected by the fact that they had to be restarted between the cc-pwCVTZ and cc-pwCVQZ basis set calculations because I had underestimated the time they required.

The MP2 calculations took $\sim$ 2 d 3 h 19 min and required $\sim$ 1078.46 GB of disk space for 1 m-ABA, $\sim$ 22 h 52 min and required $\sim$ 1077.64 GB of disk space for 2 m-ABA and $\sim$ 1 d 9 h 29 min and required $\sim$ 1041.62 GB of disk space for p-ABA. The time and disk space estimations for 1 m-ABA and p-ABA calculations are affected by the fact that they had to be restarted to allocate more memory and time for them, the 1 m-ABA calculation had to be restarted four times between the different basis set calculations and the p-ABA calculations had to restarted once during the cc-pwCVQZ basis set calculations.

The CCSD cc-pVnZ calculations took $\sim$ 4 d 15 h 54 min and required $\sim$ 1134.37 GB of disk space for 1 m-ABA, $\sim$ 5 d 8 h 14 min and required $\sim$ 1137.23 GB of disk space for 2 m-ABA and $\sim$ 6 d 22 h 47 min and required $\sim$ 1127.97 GB of disk space for p-ABA. The time and disk space estimations for 1 m-ABA, 2 m-ABA and p-ABA calculations are affected by the fact that they had to be restarted to allocate more memory and time for them or as they hit the maximum time allocation of three days, all of the CCSD cc-pVnZ calculations had to be restarted at least once as they hit the maximum time limit of three days. In addition to this the 1 m-ABA calculations had to be restarted two additional times before that to allocate more memory and time for them and the p-ABA calculations had to be restarted one additional time as they hit the maximum time limit of three days again. The CCSD cc-pwCVDZ and cc-pwCVTZ calculations for 1 m-ABA took $\sim$ 1 d 15 h 13 min and required $\sim$ 1134.37 GB of disk space. Most of the time for the CCSD calculations was spent on the cc-pVQZ basis set calculations. For 1 m-ABA

they took $\sim$ 3 d 8 h 48 min, for 2 m-ABA they took $\sim$ 4 d 16 h 44 min and for p-ABA took $\sim$ 6 d 9 h 24 min.

All of the calculations used $\sim$ 1 TB of disk space, because they were optimized to do so as $\sim$ 1 TB was the maximum allowed disk space usage. Besides the worse scaling behavior of the CCSD method compared to the other methods the CCSD calculations were also slowed down by the fact that they required much more memory per calculation process, which limited the number of parallel processes that could be used. The number of parallel processes for CCSD calculation was also limited by the fact that each used parallel process also increased the amount of disk space required by the calculations and each calculation process of the CCSD method required more disk space than the other methods calculation processes.

5 Analysis

5.1 Computational results

As can be seen from the Figure 11 shown earlier, the total energies of 1 m-ABA, 2 m-ABA and p-ABA are very close to each other at every optimization level. 1 m-ABA has a slightly lower total energy than 2 m-ABA and p-ABA has a lower total energy than 1 m-ABA. At the final optimization level, the 1 m-ABA total energy is ~ 0.35 meV lower than the 2 m-ABA total energy and the p-ABA total energy is ~ 3 meV lower than the m-ABA variants total energies.

There is a clear difference of over 1 eV in energy when moving from the HF level calculations to MP2 level, but the MP2 level and CCSD level results are very close to each other. The difference between total energies for MP2 level and CCSD level using the same basis set is ~ 42.9 meV for cc-pVDZ, ~ 17.5 meV for cc-pVTZ and ~ 1.3 meV for cc-pVQZ. The used basis set has more effect in energy on the MP2 and CCSD levels than on the HF level.

The difference between cc-pVnZ and cc-pwCVnZ basis sets of the same n for the same method is quite small and it gets smaller as the basis sets get bigger: for the cc-pVDZ and cc-pwCVDZ the average difference is ~ 26.0 meV, for the cc-pVTZ and cc-pwCVTZ the average difference is ~ 20.9 meV and for the cc-pVQZ and cc-pwCVQZ the average difference is ~ 7.4 meV. Based on this using CCSD cc-pVQZ as the final optimization level is justified as it should give total energies that are quite close to the ones that using CCSD cc-pwCVQZ would have given.

From the coordinates of the atoms at different optimization levels shown by the Tables 2-4, 6-14 in Appendices A, C-E the distance an atom moved between optimization levels can be calculated, as the length of a vector between an atom's locations at different geometries. The average distance the atoms moved between the starting guess and the end of the HF level was ~ 0.1359 Å. The average distance the atoms moved between the end of the HF level and the end of the MP2 level was ~ 0.0155 Å. The average distance the atoms moved between the end of the MP2 level and the end of the CCSD level was ~ 0.0056 Å. As could be expected the largest change in the atoms locations was between the starting guess and the end of the HF level and from there on the changes got significantly smaller at each step.

The multiple restarts that had to be done during calculations make comparing the calculation times difficult, because whenever the calculations hit a memory or time limit, part of the progress was lost so after a restart part of the calculations had to be redone. This can be seen most clearly from the MP2 calculations of 1 m-ABA that had four restarts and took over twice as

much time as the 2 m-ABA MP2 calculations that had no restarts. From the calculation times it can be said that on the Puhti supercomputer with optimized memory and time allocations the HF calculations can be done in under one day, the MP2 calculations can be done in under one and a half days and the CCSD cc-pVnZ calculations can be done in under a week. It is clear that the CCSD cc-pVQZ calculations were the most time-consuming part of the calculations.

5.2 Comparison between the experiment and the calculations

The calculated orbital energy values can be transformed into binding energies with the Koopmans' theorem Eq.(21). Compared to these calculated approximations of the binding energies the measured m-ABA C 1s binding energies are $\sim$ 15 eV lower in energy, the measured m-ABA N 1s binding energy is $\sim$ 18 eV lower in energy and the measured m-ABA O 1s binding energies are $\sim$ 21 eV lower in energy. For the measured valence spectrum, the two left most peaks match the two outermost valence orbitals energies quite well with a difference in energies around 0.2 eV, but the rest of the peaks do not really match up to the remaining calculated binding energies.

The large energy differences between measured and calculated 1s binding energies are caused by the fact that Koopman's theorem does not take into account the electronic relaxation (the changes that happen to a molecules electron density distribution after ionization) in the ionized molecule. The electronic relaxation can be taken into account by calculating the binding energy as the difference between the energies of the ground state and the ionized state, but it would have to be done separately for each ionization. Due to the removal of one electron at least one of the two states is always an open-shell state, because of this these calculations have to use methods based on ROHF or UHF theory. The error caused by ignoring the electronic relaxation is mostly seen in the 1s binding energies as the effect of electronic relaxation is smaller for the outer orbitals and it is often hidden by a favorable cancellation with electron correlation effects. [1]

The differences in calculated and measured inner valence region binding energies are caused by multiple factors. One big factor is that the potential rotational and vibrational excitations during ionization have not been taken into account and they can distort the peaks of the measured spectra greatly. Electron correlation has a large effect in the valence region, but the CCSD method does take it partially into account and electronic relaxation, which has a favorable cancellation with electron correlation also has a small effect

in the valence region, so it is difficult to say what their combined effect is. Due to the electron correlation effects the ionization of one electron can cause the excitation or ionization of another electron. These are known as shake-up (excitation) and shake-off (ionization) processes and they can be seen as additional peaks in the spectra. [1,38]

As can be seen from Figures 25-27 in Appendices G-I the outer most orbital marked as orbital 36. is located mainly on the benzene ring and around the nitrogen atom for 1 m-ABA, 2 m-ABA and p-ABA. For p-ABA it also has a small presence around the doubly-bound oxygen atom and for 2 m-ABA an even smaller presence at the same location. The second outermost orbital marked as orbital 35. is located mainly on two opposite edges of the benzene ring for 1 m-ABA, 2m-ABA and p-ABA. For 1 m-ABA and 2 m-ABA it also has a small presence around the doubly-bound oxygen atom.

For 1 m-ABA and 2 m-ABA the difference in approximated binding energies is small, on average it is ~ 0.05 eV. There are a couple of binding energies where the difference reaches ~ 0.1 eV or higher and one where the difference is ~ 0.26 eV. Most of these larger differences are in the middle of valence spectra where the measured and calculated spectra do not really match so they would be difficult to spot from a spectrum and the one ~ 0.1 eV difference that is not in the valence region belongs to one of the five carbon atoms which form a combined peak, so it is also difficult to spot from a spectrum as it gets lost within the other carbon atoms' energies.

Based on the calculated approximations of binding energies and the measured binding energy spectra an approximation of what the true p-ABA binding energy spectra would look like can be made. First the difference between calculated and measured binding energies is assumed to be similar for m-ABA and p-ABA. So the O 1s binding energies are ~ 21 eV lower in energy, the N 1s binding energies are ~ 18 eV lower in energy, the C 1s binding energies are ~ 15 eV lower in energy and the two outermost valence orbitals are ~ 0.2 eV lower in energy compared to the calculated values. The O 1s spectrum would be similar to the m-ABA spectrum having two peaks with energies ~ 539.6 eV for the -OH peak and ~ 537.7 eV for the =O peak. The two peaks would be slightly lower in energy for p-ABA, but the difference in energy between them would be around the same for m-ABA and p-ABA. The N 1s peak would have an energy of ~ 405.9 eV, which would be slightly higher than the energy of the m-ABA N 1s peak. The C 1s spectrum would have peaks with energies of ~ 294.6 eV for the -COOH peak, ~ 292.4 eV for the $-C-NH_2$ peak and ~ 290.9 eV for the combined peak from the remaining five carbon atoms. Thus, the energies of the combined peak from the five carbon atoms and the $-C-NH_2$ peak would be higher for p-ABA than m-ABA, but the energy difference would be bigger between the $-C-NH_2$ peaks. The

-COOH peaks energy would be slightly lower for p-ABA than m-ABA. This would lead to the combined peak from the five carbon atoms and the -C-NH$_2$ peak being further from each other and the -C-NH$_2$ peak and -COOH peak being closer to each other for p-ABA than m-ABA. The two lowest energy valence peaks would have energies of $\sim$ 8.3 eV and $\sim$ 9.4 eV. These peaks would be closer together than for m-ABA, as the lowest energy peak's energy would be slightly higher, and the second lowest energy peak's energy would be slightly lower for p-ABA than m-ABA. It is difficult to say anything for the rest of the valence spectrum as the m-ABA valence peaks' energies do not really match up to the calculated valence binding energies and the situation is probably similar for p-ABA.

6 Summary

In this thesis I used the HF, MP2 and CCSD ab initio methods to optimize the geometries of 1 m-ABA, 2 m-ABA and p-ABA molecules and then estimated the binding energies of the molecules based on their orbital energies at the optimized geometries. I then compared the calculated 1 m-ABA and 2 m-ABA binding energies to each other and the experimental m-ABA binding energy spectra and analyzed their differences. I also used the calculated binding energies to identify most of the peaks from the experimental spectra.

The total energies of MP2 and CCSD levels were very close to each other at the cc-pVQZ level, while the CCSD calculation took considerably longer to complete. As there was not much difference in the total energies between MP2 and CCSD, the MP2 results were probably already quite accurate. The results might differ if CCSD(T) calculations could have been done, but they probably would have taken even longer to complete than the CCSD calculations. The MP2 level provided decently accurate results much faster than CCSD so it would probably have been good enough for ABA molecules' geometry optimization.

The 1 m-ABA and 2 m-ABA total energies were very close to each other and the differences in their approximated binding energies were also mostly small enough that they would be difficult to notice from a spectrum. The largest differences were in the middle of the valence binding energies and as the approximations of valence spectrum binding energies did not match well with the measured valence spectrum it would be difficult to say how they shows up in a measured valence spectrum. The largest differences in the 1s spectra were in the C 1s spectra, but as they were in the binding energies of the five carbon atoms which energies are so close to each other that they form a combined peak the differences are also very difficult to notice. Thus, it is difficult to differentiate between 1 m-ABA and 2 m-ABA based on measured spectra.

Based on the results of the ab initio calculations for m-ABA and p-ABA and the measured spectra of m-ABA a decently accurate estimations of the different p-ABA 1s binding energy spectra could be constructed and an approximation for the lowest energy part of the p-ABA valence binding energy spectrum could also be made.

Bibliography

[1] Cramer, C. J. (2004). *Essentials of computational chemistry: Theories and models* (2. ed.). Wiley.

[2] Jensen, F. (1999). *Introduction to computational chemistry*. Wiley.

[3] Atkins, P. & Friedman, R. (2011). *Molecular quantum mechanics* (5th ed.). Oxford University Press.

[4] Slater, J. C. (1930). Atomic Shielding Constants. *Physical review, 36*(1), 57-64.

[5] Boys, S. F. (1950). Electronic Wave Functions. I. A General Method of Calculation for the Stationary States of Any Molecular System. *Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences, 200*(1063), 542-554.

[6] Hehre, W. J., Stewart, R. F. & Pople, J. A. (1969). Self-Consistent Molecular-Orbital Methods. I. Use of Gaussian Expansions of Slater-Type Atomic Orbitals. *The Journal of chemical physics, 51*(6), 2657-2664.

[7] Woon, D. E. & Dunning, T. H. (1995). Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *The Journal of chemical physics, 103*(11), 4572-4585.

[8] Peterson, K. A. & Dunning, T. H. (2002). 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. *The Journal of chemical physics, 117*(23), 10548-10560.

[9] Dunning, T. H. (1989). Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *The Journal of chemical physics, 90*(2), 1007-1023.

[10] Woon, D. E. & Dunning, T. H. (1993). Gaussian basis sets for use in correlated molecular calculations. III: The atoms aluminum through argon. *The Journal of chemical physics, 98*(2), 1358-1371.

[11] Hartree, D. R. (1928). The Wave Mechanics of an Atom with a Non-Coulomb Central Field. Part I-III. *Theory and Methods. Mathematical proceedings of the Cambridge Philosophical Society, 24*(1,3), 89-110, 111-32, 426-437.

[12] Roothaan, C. C. J. (1951). New Developments in Molecular Orbital Theory. *Reviews of modern physics, 23*(2), 69-89.

[13] Hall, G. G. (1951). The Molecular Orbital Theory of Chemical Valency. VIII. A Method of Calculating Ionization Potentials. *Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences, 205*(1083), 541-552.

[14] Møller, C. & Plesset, M. S. (1934). Note on an Approximation Treatment for Many-Electron Systems. *Physical review, 46*(7), 618-622.

[15] Čížek, J. (1966). On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *The Journal of chemical physics, 45*(11), 4256-4266.

[16] Pärna, R., Sankari, R., Kukk, E., Nõmmiste, E., Valden, M., Lastusaari, M., . . . Huttula, M. (2017). FinEstBeaMS – A wide-range Finnish-Estonian Beamline for Materials Science at the 1.5 GeV storage ring at the MAX IV Laboratory. *Nuclear instruments & methods in physics research. Section A, Accelerators, spectrometers, detectors and associated equipment, 859*, 83-89.

[17] Kooser, K., Kivimäki, A., Turunen, P., Pärna, R., Reisberg, L., Kirm, M., . . . Kukk, E. (2020). Gas-phase endstation of electron, ion and coincidence spectroscopies for diluted samples at the FinEstBeAMS beamline of the MAX IV 1.5 GeV storage ring. *Journal of synchrotron radiation, 27*(4), 1080-1091.

[18] Abid, A. R. (2021). *X-ray absorption and fragmentation as initial steps of radiation damage in free organic molecules and nanoparticles.* University of Oulu.

[19] Potts, A. W. & Price, W. C. (1972). Photoelectron Spectra and Valence Shell Orbital Structures of Groups V and VI Hydrides. *Proceedings of the Royal Society of London. Series A, Mathematical and physical sciences, 326*(1565), 181-197.

[20] Sorensen, S. L., Børve, K. J., Feifel, R., de Fanis, A. & Ueda, K. (2008). The O 1s photoelectron spectrum of molecular oxygen revisited. *Journal of Physics B, 41*(9),

[21] Sankari, R., Ehara, M., Nakatsuji, H., Senba, Y., Hosokawa, K., Yoshida, H., . . . Ueda, K. (2003). Vibrationally resolved O 1s photoelectron spectrum of water. *Chemical physics letters, 380*(5)

[22] Jolly, W., Bomben, K. & Eyermann, C. (1984). Core-electron binding energies for gaseous atoms and molecules. *Atomic data and nuclear data tables, 31*(3), 433-493.

[23] Myrseth, V., Bozek, J., Kukk, E., Sæthre, L. & Thomas, T. (2002). Adiabatic and vertical carbon 1s ionization energies in representative small molecules. *Journal of electron spectroscopy and related phenomena, 122*(1), 57-63.

[24] The author wishes to acknowledge CSC – IT Center for Science, Finland, for computational resources.

[25] Werner, H., Knowles, P. J., Knizia, G., Manby, F. R. & Schütz, M. (2012). Molpro: A general-purpose quantum chemistry program package. *Wiley interdisciplinary reviews. Computational molecular science, 2*(2), 242-253.

[26] Werner, H., Knowles, P. J., Manby, F. R., Black, J. A., Doll, K., Heßelmann, A., . . . Sibaev, M. (2020). The Molpro quantum chemistry package. *The Journal of chemical physics, 152*(14), 144107.

[27] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, S. J. Bennie, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, S. J. R. Lee, Y. Liu, A. W. Lloyd, Q. Ma, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, T. F. Miller III, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, M. Wang, and M. Welborn, MOLPRO, version 2021.2, a package of ab initio programs, see https://www.molpro.net.

[28] Lindh, R., Ryu, U. & Liu, B. (1991). The reduced multiplication scheme of the Rys quadrature and new recurrence relations for auxiliary function based two-electron integral evaluation. *Journal of Chemical Physics, 95*(8), 5889-5897.

[29] El Azhary, A., Rauhut, G., Pulay, P. & Werner, H. (1998). Analytical energy gradients for local second-order Møller–Plesset perturbation theory. *The Journal of chemical physics, 108*(13), 5185-5193.

[30] Eckert, F., Pulay, P. & Werner, H. (1997). Ab initio geometry optimization for large molecules. *Journal of computational chemistry, 18*(12), 1473-1483.

[31] Hampel, C., Peterson, K. A. & Werner, H. (1992). A comparison of the efficiency and accuracy of the quadratic configuration interaction (QCISD), coupled cluster (CCSD), and Brueckner coupled cluster (BCCD) methods. *Chemical physics letters, 190*(1), 1-12.

[32] Deegan, M. J. & Knowles, P. J. (1994). Perturbative corrections to account for triple excitations in closed and open shell coupled cluster theories. *Chemical physics letters, 227*(3), 321-326.

[33] Kats, D. & Manby, F. R. (2013). Communication: The distinguishable cluster approximation. *The Journal of chemical physics, 139*(2), 021102.

[34] Kats, D. (2014). Communication: The distinguishable cluster approximation. II. The role of orbital relaxation. *The Journal of chemical physics, 141*(6), 061101.

[35] Kats, D., Kreplin, D., Werner, H. & Manby, F. R. (2015). Accurate thermochemistry from explicitly correlated distinguishable cluster approximation. *The Journal of chemical physics, 142*(6), 064111.

[36] Bode, B. M. & Gordon, M. S. (1998). Macmolplt: A graphical user interface for GAMESS. *Journal of molecular graphics & modelling, 16*(3), 133-138.

[37] Allouche, A. (2011). Gabedit-A graphical user interface for computational chemistry softwares. *Journal of computational chemistry, 32*(1), 174-182.

[38] Carlson, T. A. (1975). *Photoelectron and Auger spectroscopy.*

A   Starting geometries

Atom	X	Y	Z
O	2.18721294	1.88313794	-0.00000000
O	3.31737399	-0.26936299	-0.00000000
N	-0.35456800	-3.94935989	0.00000300
C	-0.35456800	0.31064001	-0.00000000
C	0.85353202	-0.38686001	-0.00000000
C	0.85353202	-1.78185999	-0.00000000
C	-0.35456800	-2.47936010	-0.00000000
C	-1.56266797	-1.78185999	-0.00000000
C	-1.56266797	-0.38686001	-0.00000000
C	2.18721199	0.38313800	-0.00000000
H	-0.35456800	1.41033995	-0.00000000
H	1.80593204	-2.33175993	-0.00000000
H	-2.51506901	-2.33175993	-0.00000000
H	-2.51506901	0.16304000	-0.00000000
H	3.22430301	2.24980497	-0.00000000
H	0.35736799	-4.30602694	0.71471697
H	-1.31822097	-3.88110995	0.46000600

Table 2: 1 m-ABA starting geometry in angstroms

Atom	X	Y	Z
O	3.48593211	-0.36631599	0.04050000
O	2.18748999	1.68766296	-0.03523600
N	-0.35456800	-3.94935989	0.00000300
C	-0.35456800	0.31064001	-0.00000000
C	0.85353202	-0.38686001	-0.00000000
C	0.85353202	-1.78185999	-0.00000000
C	-0.35456800	-2.47936010	-0.00000000
C	-1.56266797	-1.78185999	-0.00000000
C	-1.56266797	-0.38686001	-0.00000000
C	2.18721199	0.38313800	0.00000000
H	-0.35456800	1.41033995	-0.00000000
H	1.80593204	-2.33175993	-0.00000000
H	-2.51506901	-2.33175993	-0.00000000
H	-2.51506901	0.16304000	-0.00000000
H	4.32207489	0.34840900	0.03423300
H	0.35736799	-4.30602694	0.71471697
H	-1.31822097	-3.88110995	0.46000600

Table 3: 2 m-ABA starting geometry in angstroms

Atom	X	Y	Z
O	2.18721294	1.88313794	-0.00000000
O	3.31737399	-0.26936299	-0.00000000
N	-2.90254998	-2.47092605	0.05666900
C	-0.35456800	0.31064001	-0.00000000
C	0.85353202	-0.38686001	-0.00000000
C	0.85353202	-1.78185999	-0.00000000
C	-0.35456800	-2.47936010	-0.00000000
C	-1.56266797	-1.78185999	-0.00000000
C	-1.56266797	-0.38686001	-0.00000000
C	2.18721199	0.38313800	-0.00000000
H	-0.35456800	1.41033995	-0.00000000
H	1.80593204	-2.33175993	-0.00000000
H	-2.51506901	0.16304000	-0.00000000
H	3.22430301	2.24980497	-0.00000000
H	-3.71849298	-1.90613794	0.45683801
H	-2.87091589	-3.43257594	0.52476001
H	-0.35456800	-3.57810593	0.00000000

Table 4: p-ABA starting geometry in angstroms

B Total energies

Method	Basis set	1 m-ABA (eV)	2 m-ABA (eV)	p-ABA (eV)
HF	cc-pVDZ	-473.39986021	-473.39947992	-473.40410956
HF	cc-pVTZ	-473.52891265	-473.52849295	-473.53334372
HF	cc-pVQZ	-473.56155390	-473.56112588	-473.56599046
HF	cc-pwCVDZ	-473.40347435	-473.40309548	-473.40771542
HF	cc-pwCVTZ	-473.53204586	-473.53162495	-473.53646208
HF	cc-pwCVQZ	-473.56268760	-473.56225943	-473.56712369
MP2	cc-pVDZ	-474.83670701	-474.83638181	-474.83721146
MP2	cc-pVTZ	-475.29783203	-475.29748535	-475.29870378
MP2	cc-pVQZ	-475.45079757	-475.45044244	-475.45176324
MP2	cc-pwCVDZ	-474.87683895	-474.87651589	-474.87734126
MP2	cc-pwCVTZ	-475.33007519	-475.32971838	-475.33095181
MP2	cc-pwCVQZ	-475.46451137	-475.46415496	-475.46548274
CCSD	cc-pVDZ	-474.87913521	-474.87882925	-474.88112422
CCSD	cc-pVTZ	-475.31484589	-475.31450993	-475.31727986
CCSD	cc-pVQZ	-475.44898234	-475.44863457	-475.45152500
CCSD	cc-pwCVDZ	-474.91344328		
CCSD	cc-pwCVTZ	-475.34205511		
CCSD	cc-pwCVQZ			

Table 5: Total energies of 1 m-ABA, 2 m-ABA and p-ABA for different methods and basis sets

C Geometries at HF cc-pwCVQZ optimization level

Atom	X	Y	Z
O	2.1099841752	1.6499185941	-0.0479225914
O	3.224454906	-0.2362328243	0.0663273727
N	-0.3742033285	-3.8118420334	0.0788467751
C	-0.3233865752	0.3719043492	-0.0208676525
C	0.8573233564	-0.3468576926	0.0206594493
C	0.8441331073	-1.7289822175	0.0707026636
C	-0.3549309362	-2.4204277464	0.0804648945
C	-1.5397093506	-1.6954208049	0.0371953698
C	-1.5193264246	-0.3188436884	-0.0112271382
C	2.1864054088	0.328392554	0.0167495306
H	-0.3095560272	1.440250471	-0.059019882
H	1.7764140372	-2.2565672435	0.0985161937
H	-2.4789920327	-2.2160099304	0.0368708882
H	-2.4462140338	0.2200851259	-0.0430149094
H	2.988256243	1.9954554305	-0.0458551572
H	0.4475547549	-4.2409317309	0.4338640068
H	-1.1937082098	-4.2199694722	0.4624361563

Table 6: 1 m-ABA geometry at HF cc-pwCVQZ optimization level in angstroms

Atom	X	Y	Z
O	3.2318535454	-0.3762913833	0.0475809418
O	2.2318269358	1.5738060607	-0.0492512527
N	-0.3610715935	-3.8291904431	0.0872603989
C	-0.3104186625	0.3530549023	-0.0292138844
C	0.8714523592	-0.3648097581	0.0136993534
C	0.8587515778	-1.7464596744	0.0686424122
C	-0.3437888903	-2.4366743327	0.0830497229
C	-1.526445805	-1.7127004136	0.0379394949
C	-1.5053740145	-0.3342964693	-0.0165136803
C	2.1537958523	0.3953751514	-0.0005461796
H	-0.2819624767	1.4215355991	-0.0711632853
H	1.7825754253	-2.2874323295	0.0961500913
H	-2.4659989955	-2.2328071761	0.0412561153
H	-2.432116648	0.2047255055	-0.04975237
H	3.9925979489	0.1821116967	0.0323131506
H	0.4527942795	-4.255036147	0.4638067623
H	-1.1873647182	-4.2348135986	0.4589651788

Table 7: 2 m-ABA geometry at HF cc-pwCVQZ optimization level in angstroms

Atom	X	Y	Z
O	2.0776365425	1.6516472365	-0.0303200743
O	3.2211976097	-0.2216321112	-0.025394928
N	-2.7596188882	-2.4181936345	0.0481474019
C	-0.3408758453	0.3291092315	0.0238295954
C	0.8628189626	-0.3592810574	0.0118212205
C	0.8359870038	-1.7471565455	0.0265458454
C	-0.3534310266	-2.4304590803	0.0533910151
C	-1.5621158837	-1.7365211293	0.0666365792
C	-1.5386445631	-0.345044711	0.0511626163
C	2.1700018963	0.3256442702	-0.015992565
H	-0.34058682	1.3997076649	0.0126948961
H	1.7635285169	-2.2829516062	0.0166177959
H	-2.4617373577	0.202404998	0.0549277514
H	2.9531141156	2.0030575382	-0.0487053485
H	-3.5552628394	-1.9173215986	0.3629991165
H	-2.7366133052	-3.3549116805	0.3724057703
H	-0.3569359082	-3.5036657951	0.0575003319

Table 8: p-ABA geometry at HF cc-pwCVQZ optimization level in angstroms

D Geometries at MP2 cc-pwCVQZ optimization level

Atom	X	Y	Z
O	2.091344032	1.6717125804	-0.0488813964
O	3.250276929	-0.2465004253	0.0653054996
N	-0.3736880018	-3.8184809169	0.0715288107
C	-0.3277553371	0.3831171089	-0.0231132094
C	0.8630351679	-0.3411273665	0.0202513133
C	0.8518599232	-1.7321262977	0.0730616351
C	-0.3553884831	-2.4277988562	0.0842868041
C	-1.5487421878	-1.6992586878	0.0398449434
C	-1.5302160845	-0.3119294855	-0.0117171588
C	2.1847055667	0.3244073324	0.0161780643
H	-0.3058157666	1.4603726275	-0.063230075
H	1.7972900529	-2.2572661714	0.1024400422
H	-2.4934366991	-2.2276077813	0.0412604414
H	-2.4653321162	0.2287962374	-0.0434170306
H	3.0049160335	1.9887263625	-0.0451477344
H	0.4577341335	-4.250099138	0.438846841
H	-1.2062880919	-4.2310159823	0.4572281791

Table 9: 1 m-ABA geometry at MP2 cc-pwCVQZ optimization level in angstroms

Atom	X	Y	Z
O	3.2395919366	-0.4040153724	0.050863366
O	2.2381746856	1.6015522021	-0.0533388277
N	-0.3635893223	-3.8363864412	0.0793727688
C	-0.3118850662	0.3617932903	-0.0314543743
C	0.8788384108	-0.3614875051	0.0132286118
C	0.8675307101	-1.7530175527	0.0709818323
C	-0.3436534503	-2.4447198714	0.0865921631
C	-1.535606145	-1.7156885991	0.0401539434
C	-1.515574575	-0.3279538028	-0.0169615123
C	2.1492555256	0.396814879	-0.0013323698
H	-0.2718327192	1.4394292241	-0.0753236277
H	1.8021479806	-2.2947844176	0.1006571202
H	-2.4804982119	-2.2436438264	0.0450230858
H	-2.4496610531	0.2143300794	-0.0500751928
H	3.9955637325	0.1986896416	0.033360652
H	0.4623689207	-4.2666912007	0.4603926056
H	-1.2000652394	-4.244123537	0.4620827256

Table 10: 2 m-ABA geometry at MP2 cc-pwCVQZ optimization level in angstroms

Atom	X	Y	Z
O	2.0587235295	1.6733017867	-0.029262511
O	3.2468508566	-0.2310360156	-0.0226252463
N	-2.7752800733	-2.4264103795	0.0279057769
C	-0.3411908942	0.3425120645	0.0216518713
C	0.8643714965	-0.3579573398	0.0104288055
C	0.8466346157	-1.7516752	0.0245761073
C	-0.3553237948	-2.4348468464	0.0502951262
C	-1.5692236762	-1.7407464635	0.0646244017
C	-1.5435006145	-0.3435899197	0.0476880166
C	2.1714849577	0.3230396444	-0.014954716
H	-0.3341636486	1.4217269894	0.0117421475
H	1.7864406168	-2.284476971	0.016723436
H	-2.4763580895	0.2048120565	0.051266947
H	2.968524588	1.9998885695	-0.0456822632
H	-3.5662374015	-1.9168112074	0.3835117308
H	-2.7415910109	-3.3663814422	0.3844534174
H	-0.3616992473	-3.5169173362	0.0559239727

Table 11: p-ABA geometry at MP2 cc-pwCVQZ optimization level in angstroms

E Final geometries (CCSD cc-pVQZ optimization level)

Atom	X	Y	Z
O	2.0993693127	1.6685009706	-0.0447092559
O	3.2441657880	-0.2415616497	0.0625475164
N	-0.3751646268	-3.8231058888	0.0659578338
C	-0.3265566193	0.3797528389	-0.0217160938
C	0.8607072887	-0.3436425729	0.0190507232
C	0.8496391641	-1.7326952769	0.0680523486
C	-0.3561482730	-2.4262499255	0.0784241858
C	-1.5480307609	-1.6995604305	0.0355970275
C	-1.5291774686	-0.3143015553	-0.0120649972
C	2.1874167642	0.3278856567	0.0158868062
H	-0.3066335164	1.4563369988	-0.0588007870
H	1.7921671497	-2.2602580309	0.0958470245
H	-2.4921867261	-2.2273818732	0.0354839269
H	-2.4630186239	0.2276994636	-0.0427751935
H	3.0046387837	1.9929094121	-0.0411966299
H	0.4542169613	-4.2452879067	0.4453217167
H	-1.2009055270	-4.2251190903	0.4738198173

Table 12: 1 m-ABA final geometry (CCSD cc-pVQZ optimization level) in angstroms

Atom	X	Y	Z
O	3.2413029920	-0.3952579531	0.0488668223
O	2.2388593646	1.5937785519	-0.0487915942
N	-0.3634009687	-3.8404939340	0.0739024696
C	-0.3120276170	0.3598384156	-0.0295467553
C	0.8757673533	-0.3627345474	0.0128206803
C	0.8648097655	-1.7518127960	0.0667196746
C	-0.3448454886	-2.4427346431	0.0811205255
C	-1.5349314915	-1.7161702453	0.0365042283
C	-1.5148191553	-0.3297246770	-0.0169746024
C	2.1537718828	0.3971727435	-0.0003152193
H	-0.2754235880	1.4366652445	-0.0705623493
H	1.7975248562	-2.2947177961	0.0943491076
H	-2.4794061896	-2.2434277843	0.0398197806
H	-2.4480498078	0.2131556228	-0.0491148915
H	3.9988589291	0.1968897421	0.0328618895
H	0.4586730960	-4.2608104728	0.4708621522
H	-1.1955578129	-4.2395182810	0.4717010514

Table 13: 2 m-ABA final geometry (CCSD cc-pVQZ optimization level) in angstroms

Atom	X	Y	Z
O	2.0670188887	1.6698866642	-0.0276765128
O	3.2410486919	-0.2262173224	-0.0210315797
N	-2.7744082688	-2.4262286815	0.0288372664
C	-0.3420548663	0.3386319301	0.0207101726
C	0.8641182448	-0.3581685633	0.0101969537
C	0.8431750124	-1.7515129797	0.0238092025
C	-0.3554371535	-2.4368932345	0.0481694395
C	-1.5678059997	-1.7398337801	0.0611812363
C	-1.5453754558	-0.3430505339	0.0454404518
C	2.1726602139	0.3258554203	-0.0139631063
H	-0.3358420332	1.4172975013	0.0106882025
H	1.7815664826	-2.2851959612	0.0160235909
H	-2.4773597700	0.2050643449	0.0485113867
H	2.9691462477	2.0019996848	-0.0433003132
H	-3.5614857604	-1.9165002248	0.3878044182
H	-2.7383240353	-3.3626282267	0.3894809178
H	-0.3621782290	-3.5180740476	0.0533852936

Table 14: p-ABA final geometry (CCSD cc-pVQZ optimization level) in angstroms

F   Orbital energies for CCSD cc-pVQZ level

	Orbital type	1 m-ABA (eV)	2 m-ABA (eV)	p-ABA (eV)
1.	O1s(-OH)	-560.773744	-560.785636	-560.598639
2.	O1s(=O)	-558.974147	-558.976133	-558.704781
3.	N1s	-423.633528	-423.674835	-423.895546
4.	C1s(COOH)	-309.774557	-309.775836	-309.563478
5.	C1s(C-NH$_2$)	-307.076330	-307.097310	-307.412772
6.	C1s	-306.183226	-306.194709	-306.187171
7.	C1s	-305.947384	-305.933017	-306.107932
8.	C1s	-305.845451	-305.851519	-305.706618
9.	C1s	-305.753721	-305.846838	-305.646971
10.	C1s	-305.740170	-305.671815	-305.636032
11.	Valence	-39.984683	-39.984901	-39.766366
12.	Valence	-37.067704	-37.073391	-36.827237
13.	Valence	-33.620457	-33.645219	-33.862937
14.	Valence	-31.257066	-31.260522	-31.302210
15.	Valence	-28.387000	-28.418619	-28.024571
16.	Valence	-27.359062	-27.341484	-27.845412
17.	Valence	-24.519064	-24.516262	-24.090213
18.	Valence	-22.723684	-22.918627	-23.187312
19.	Valence	-21.979181	-21.717679	-21.762361
20.	Valence	-19.860285	-19.897428	-19.851414
21.	Valence	-19.314397	-19.427488	-19.419923
22.	Valence	-18.767720	-18.779285	-18.803530
23.	Valence	-18.341209	-18.291630	-18.712726
24.	Valence	-17.738531	-17.701061	-17.218930
25.	Valence	-17.320102	-17.323503	-17.143391
26.	Valence	-16.553720	-16.666593	-16.521692
27.	Valence	-16.337009	-16.458263	-16.323213
28.	Valence	-16.021574	-15.872646	-16.072405
29.	Valence	-14.902942	-14.907404	-15.107109
30.	Valence	-14.416865	-14.339666	-14.720925
31.	Valence	-13.894569	-13.987551	-13.667354
32.	Valence	-13.325280	-13.364192	-13.250394
33.	Valence	-12.523578	-12.499496	-12.477754
34.	Valence	-12.381344	-12.373943	-12.194946
35.	Valence	-9.600123	-9.645838	-9.58679
36.	Valence	-8.407502	-8.411420	-8.53374

G   1 m-ABA orbital shapes for CCSD cc-pVQZ level
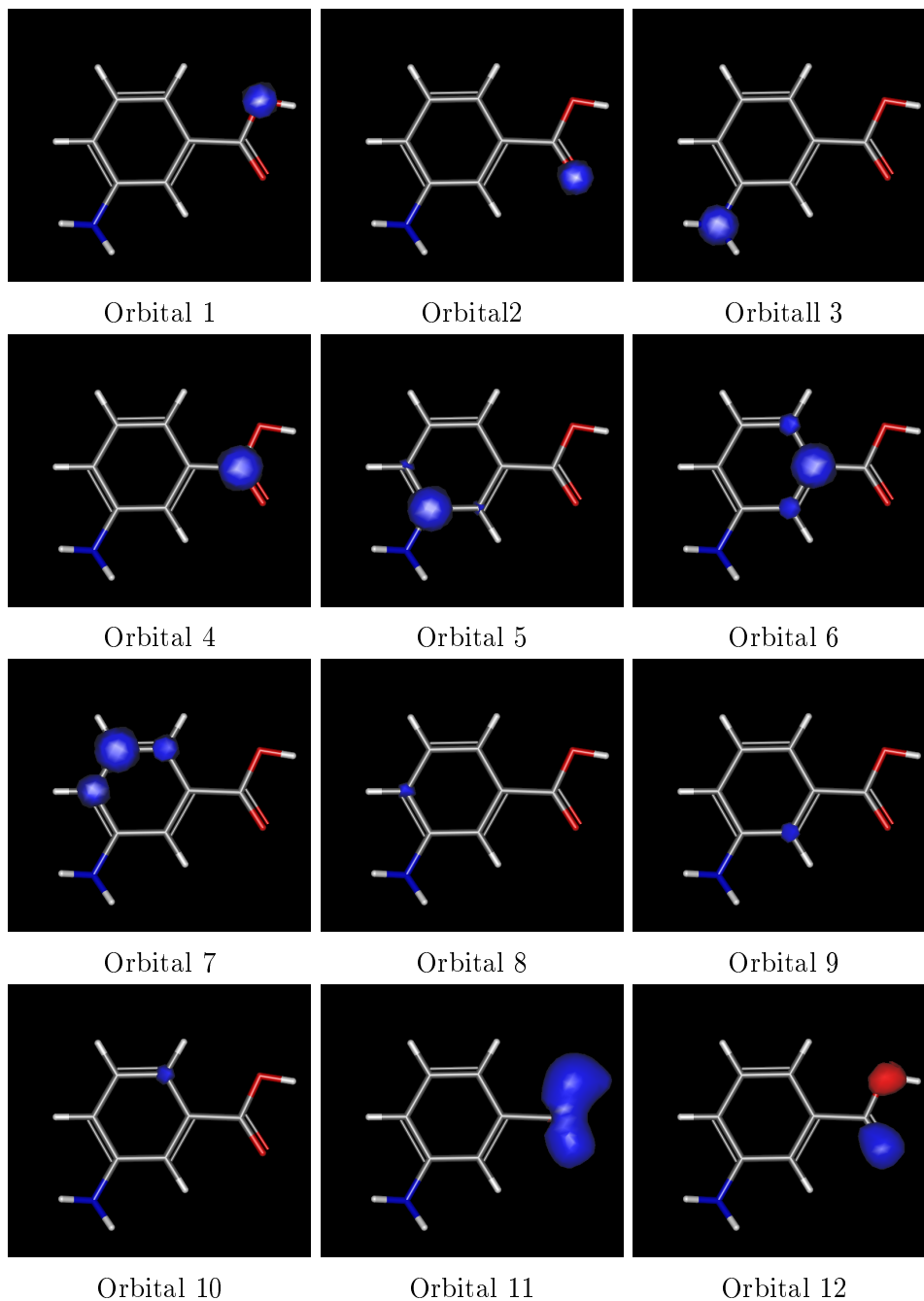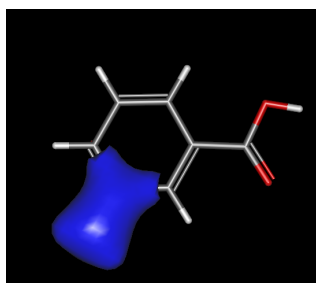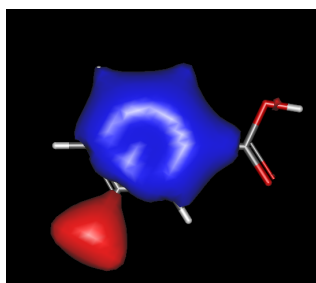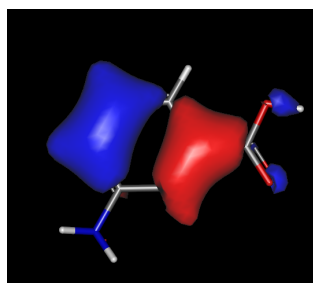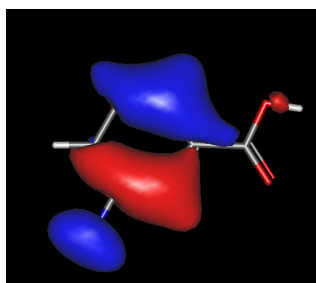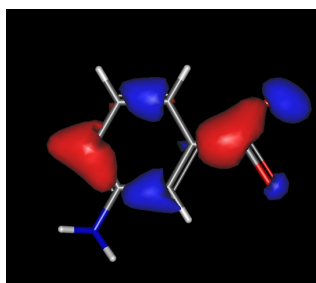
Orbital 1

Orbital2

Orbitall 3

Orbital 4

Orbital 5

Orbital 6

Orbital 7

Orbital 8

Orbital 9

Orbital 10

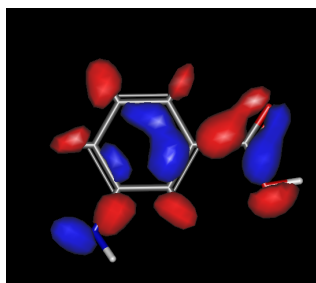Orbital 11

Orbital 12

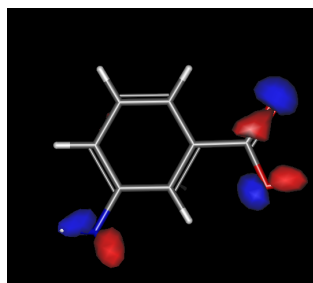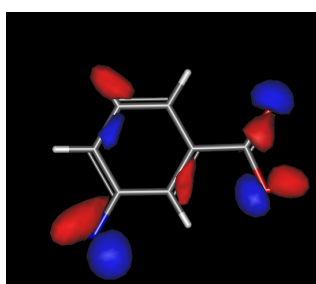Orbital 13   Orbital 14   Orbital 15

Orbital 16   Orbital 17   Orbital 18

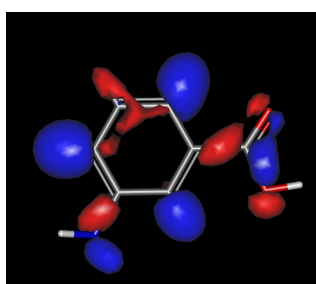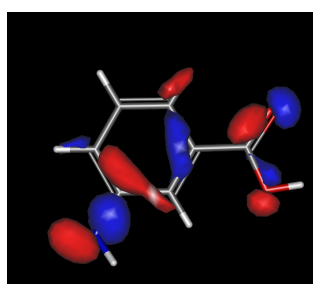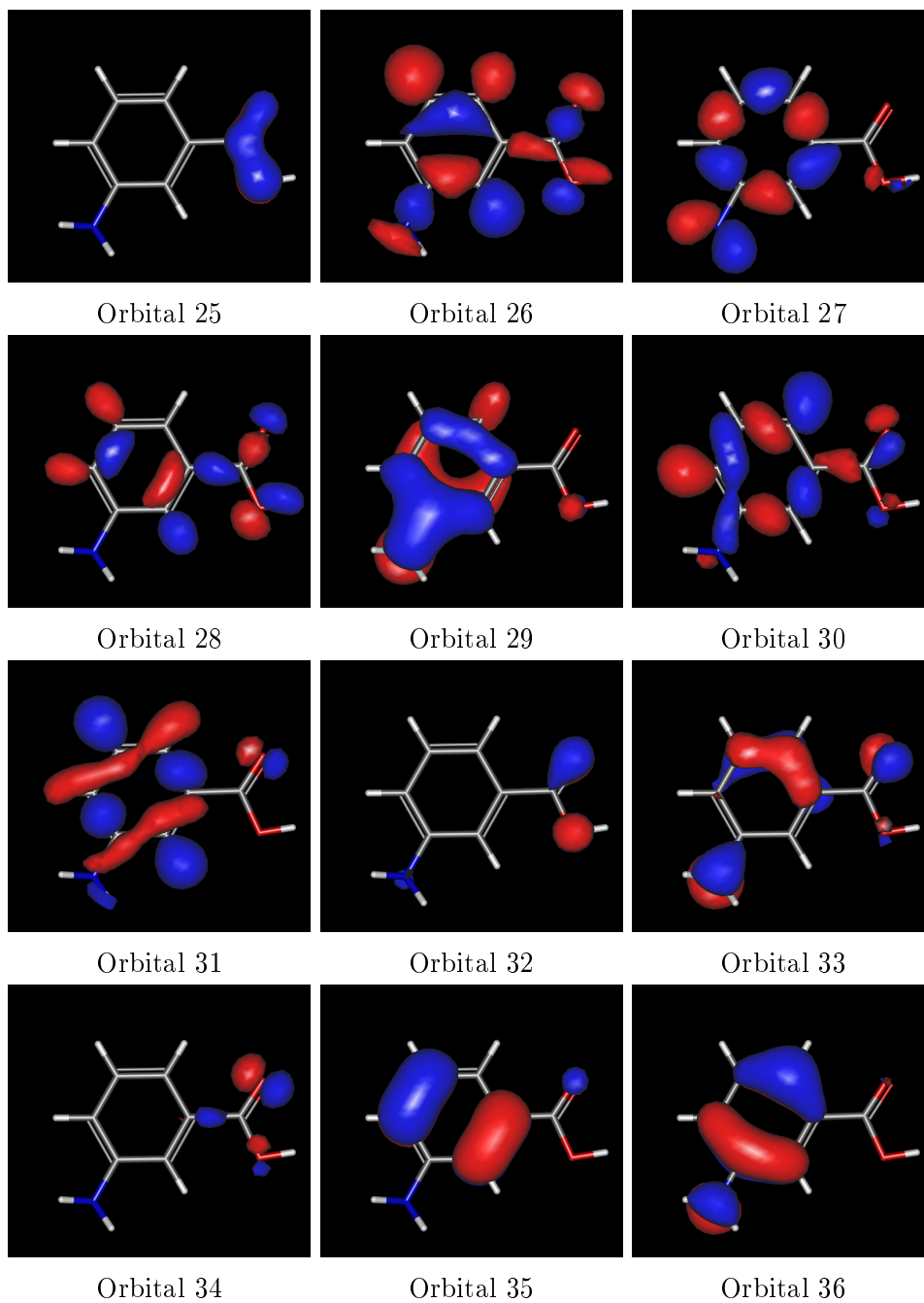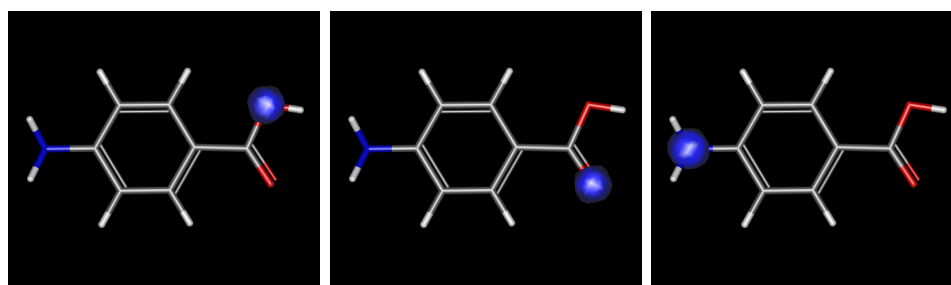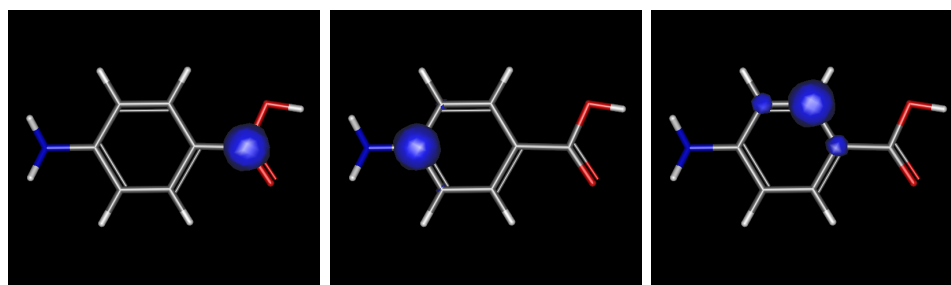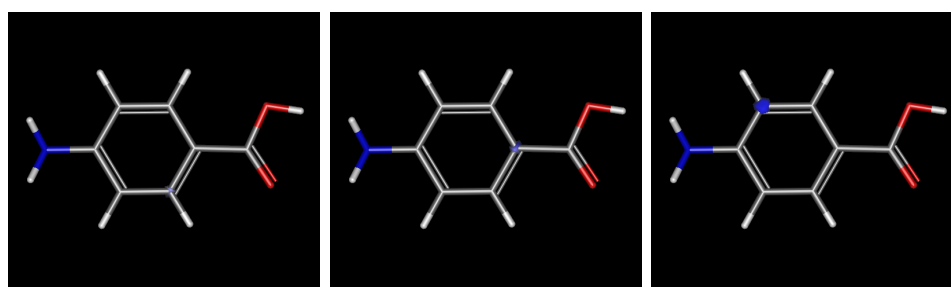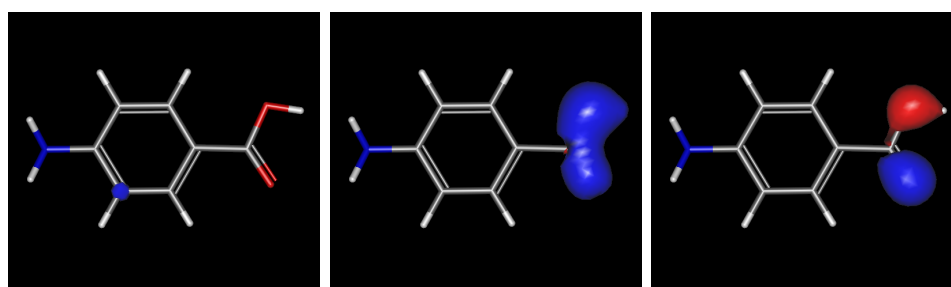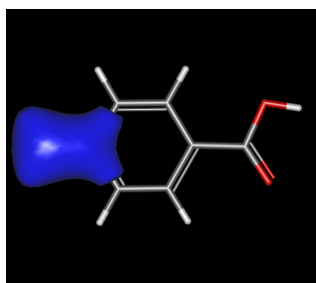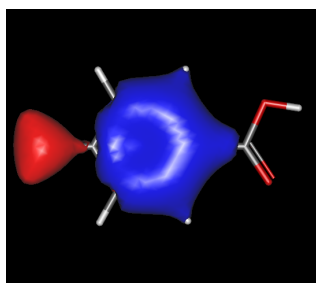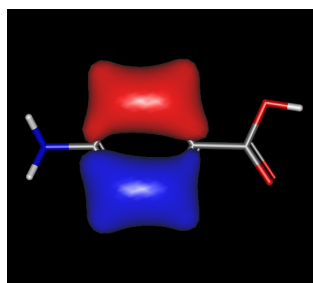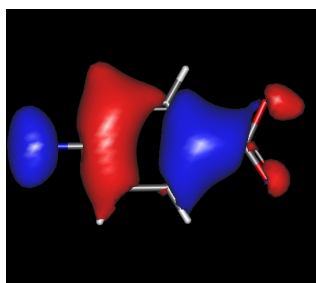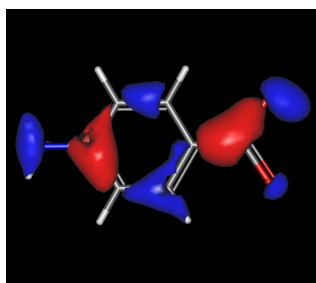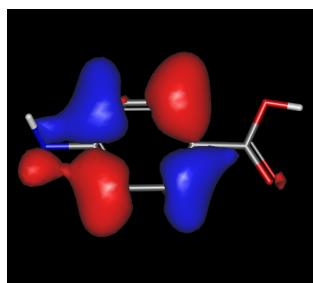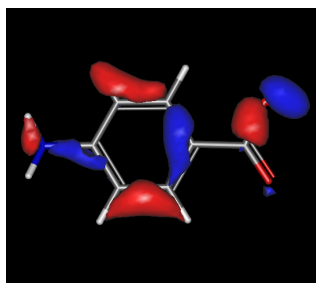Orbital 19   Orbital 20   Orbital 21

Orbital 22   Orbital 23   Orbital 24

Orbital 25   Orbital 26   Orbital 27

Orbital 28   Orbital 29   Orbital 30

Orbital 31   Orbital 32   Orbital 33

Orbital 34   Orbital 35   Orbital 36
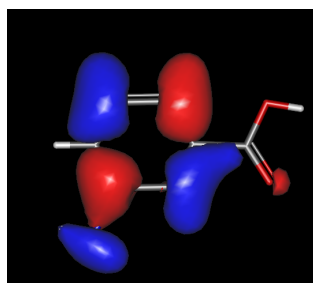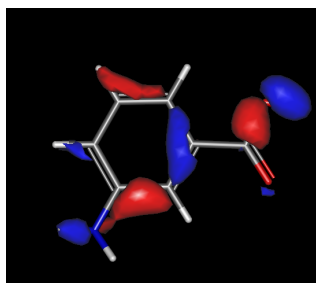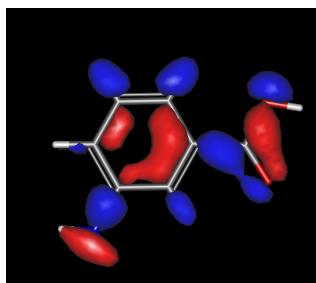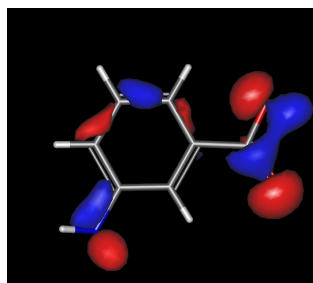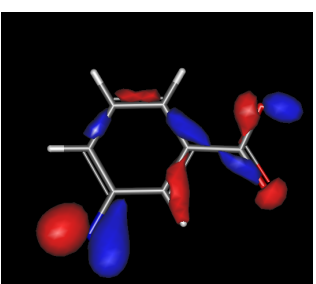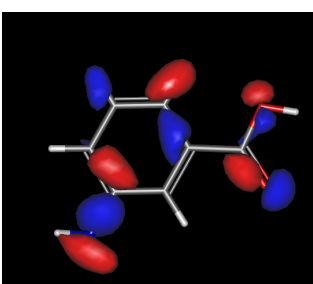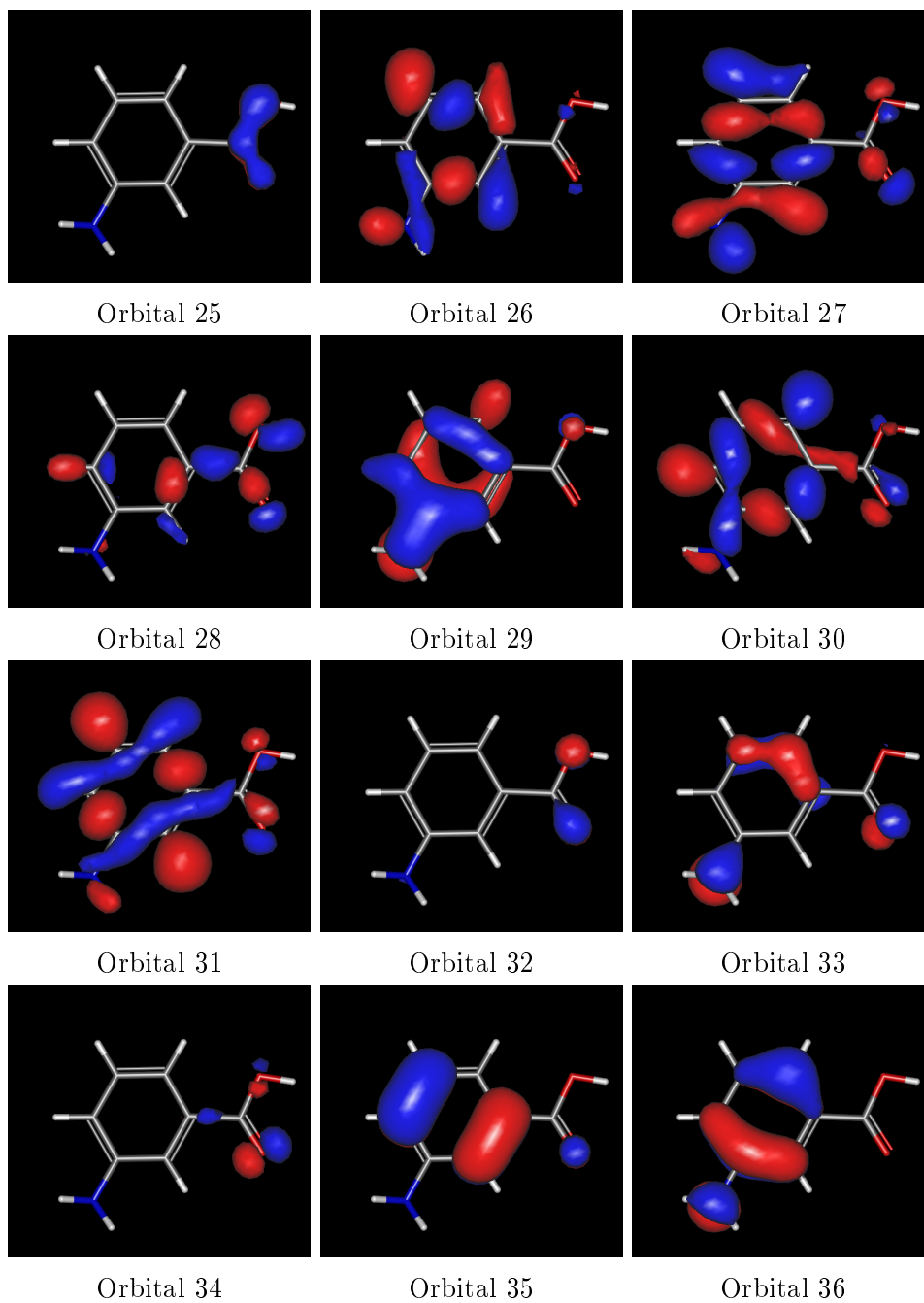
Figure 25: 1 m-ABA orbitals

H 2 m-ABA orbital shapes for CCSD cc-pVQZ level
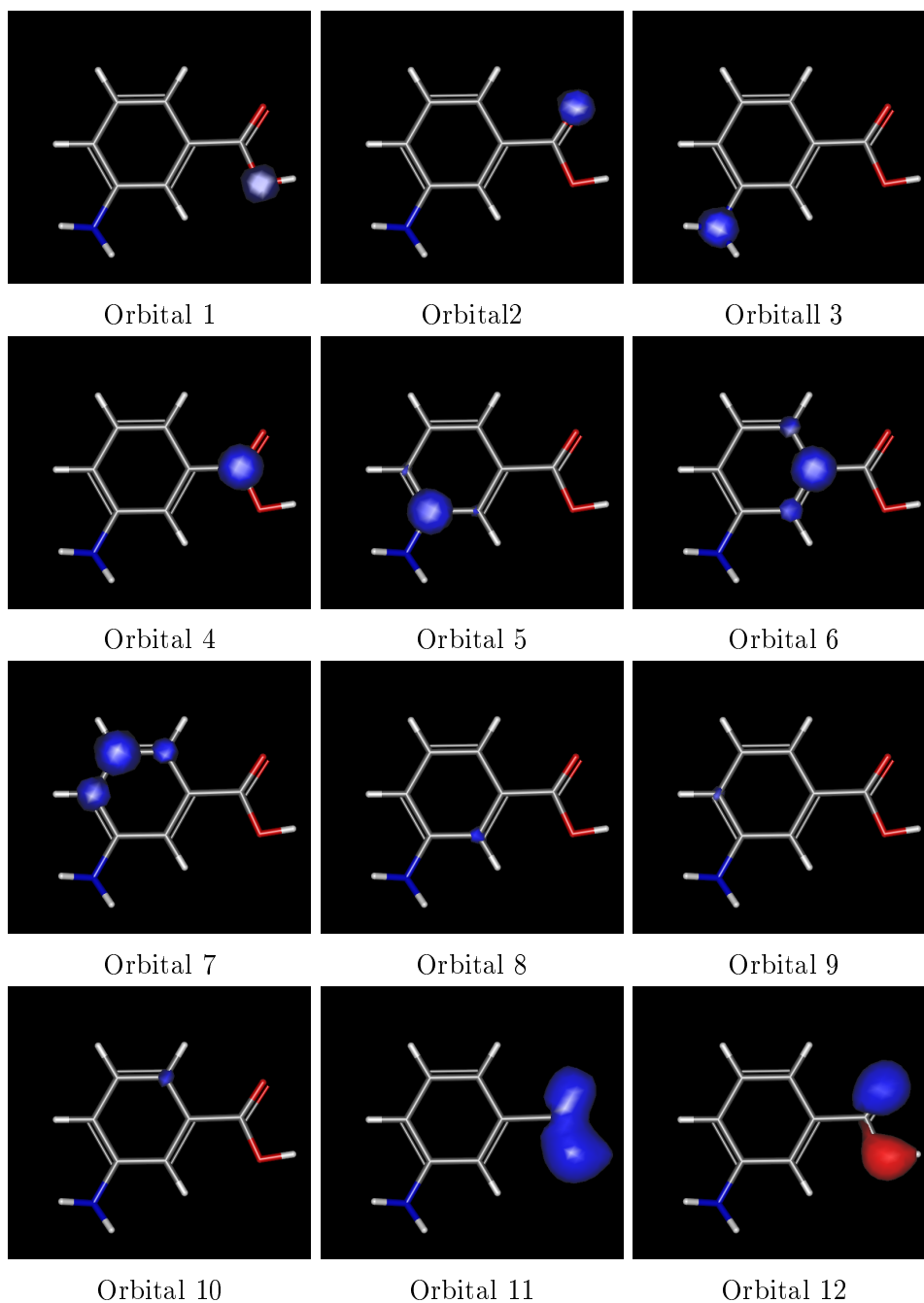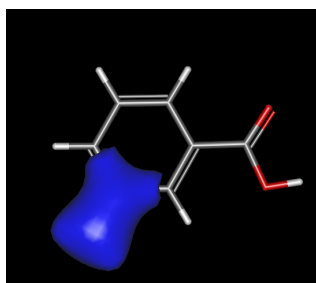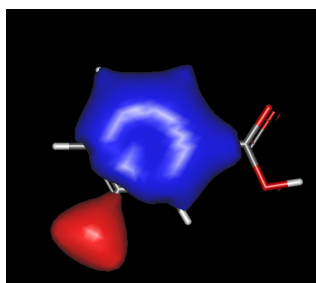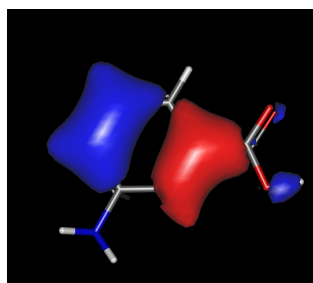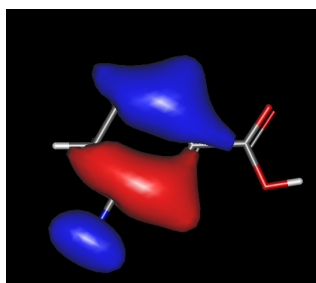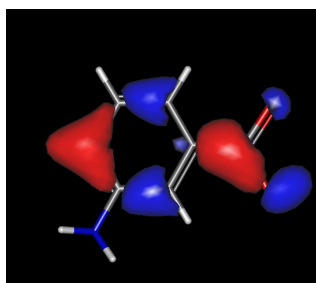
Orbital 1

Orbital2

Orbitall 3

Orbital 4

Orbital 5

Orbital 6

Orbital 7

Orbital 8

Orbital 9

Orbital 10
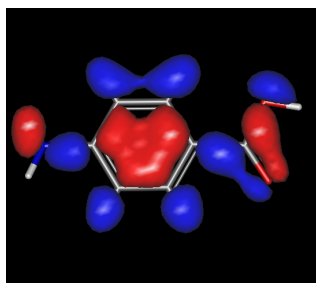

Orbital 11

Orbital 12

Orbital 13

Orbital 14
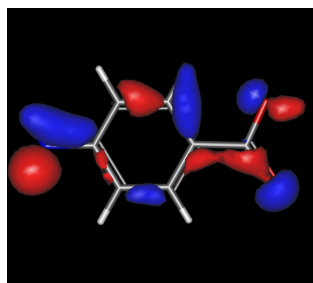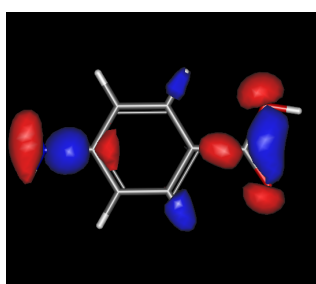

Orbital 15

Orbital 16

Orbital 17

Orbital 18

Orbital 19

Orbital 20

Orbital 21

Orbital 22

Orbital 23

Orbital 24

Orbital 25   Orbital 26   Orbital 27

Orbital 28   Orbital 29   Orbital 30
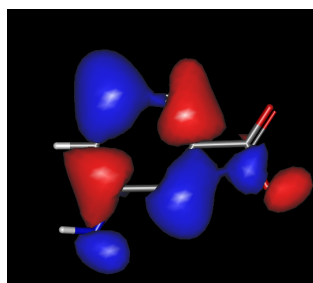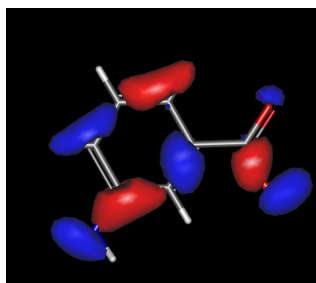
Orbital 31   Orbital 32   Orbital 33

Orbital 34   Orbital 35   Orbital 36

Figure 26: 2 m-ABA orbitals

I p-ABA orbital shapes for CCSD cc-pVQZ level

Orbital 1

Orbital2

Orbitall 3

Orbital 4

Orbital 5

Orbital 6

Orbital 7

Orbital 8

Orbital 9

Orbital 10

Orbital 11

Orbital 12
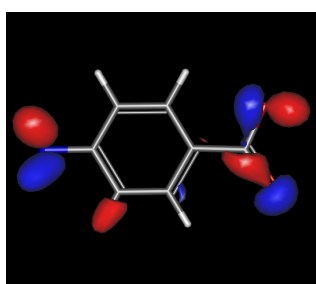
Orbital 13
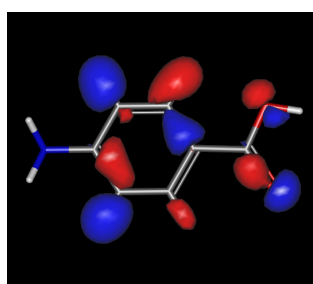

Orbital 14

Orbital 15

Orbital 16

Orbital 17

Orbital 18

Orbital 19

Orbital 20

Orbital 21

Orbital 22

Orbital 23

Orbital 24

Orbital 25   Orbital 26   Orbital 27

Orbital 28   Orbital 29   Orbital 30
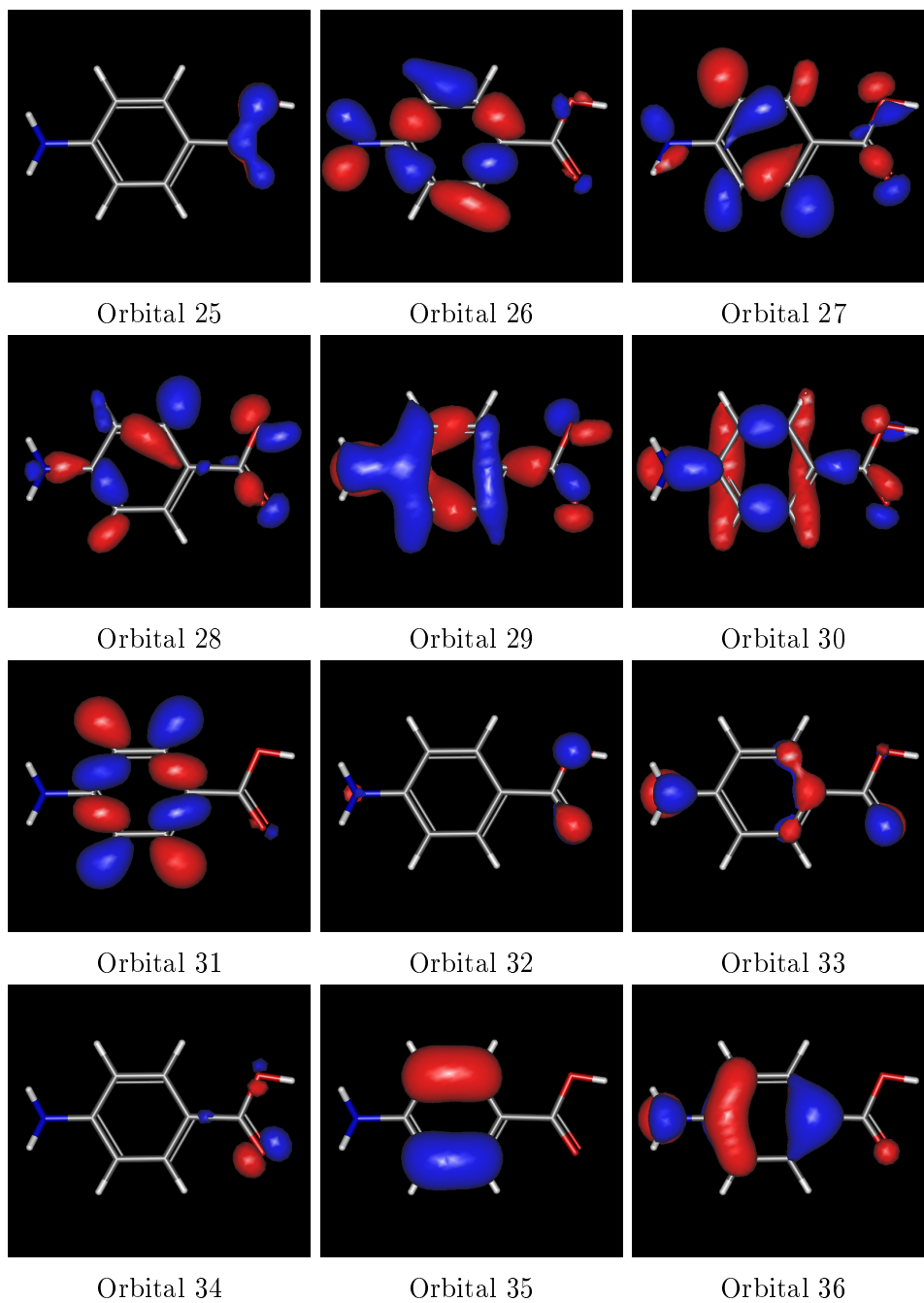
Orbital 31   Orbital 32   Orbital 33

Orbital 34   Orbital 35   Orbital 36

Figure 27: p-ABA orbitals