

Benchmarking ab initio protocols for high throughput cluster calculations

TDK Thesis

KÍGYÓSI MÁTÉ

Supervisor:

JÁNOS DARU

Department of Organic Chemistry

Eötvös Loránd University



EÖTVÖS LORÁND
UNIVERSITY | BUDAPEST

Contents

1	Abstract	III
2	Introduction	1
2.1	Coupled Cluster Molecular Dynamics	1
3	Aims and Objectives	5
4	Theoretical Background	6
4.1	Basic approximations of Quantum Chemistry	6
4.2	Hartree-Fock method	7
4.3	Correlation methods	9
4.3.1	Configurational Interaction	9
4.3.2	Møller–Plesset Perturbational theory	10
4.3.3	Coupled-Cluster theory	11
4.3.4	Local-Correlation Methods	13
4.3.4.1	DLPNO-CCSD(T)	13
4.3.4.2	LNO-CCSD(T)	15
4.3.4.3	Density Fitting	16
4.4	Basis sets	17
4.4.1	CBS extrapolation	19
4.4.2	Density-Based Basis Set Corrections	19
5	Results and Discussion	21
5.1	Water Cluster	21
5.1.1	Computational scheme optimization for high-throughput calculations	21
5.1.2	Error analysis of quantum chemical methods	24
5.2	Anionic cluster	34

6 Conclusion	41
References	43

Chapter 1

Abstract

The combination of machine learning potential development based on local correlation reference calculations and converged path integral simulations has been proposed by Daru et al.¹. It was demonstrated that the resulting CCMD approach can yield results in excellent agreement with experimental data. However, the more than 3.5 million core hours necessary for the reference calculations imposes considerable limitations on the method’s potential for routine application. In this work, I have investigated the cost-accuracy compromises of using different basis sets and approximations within the local coupled cluster approach to reduce the computational costs associated with the reference calculations employed in the CCMD framework. My results demonstrate that a suitably chosen combination of these settings and system size can potentially lead to 1-2 orders of magnitude speed-up, while potentially also increasing the accuracy of the reference calculations. In particular, we have shown that LNO-CCSD(T) in combination with DBBSC corrections, can lead to chemically accurate results with unprecedented throughput using moderate basis sets.

Chapter 2

Introduction

2.1 Coupled Cluster Molecular Dynamics

Quantum chemistry is centered around the idea of understanding and describing chemical reactions on a molecular level. This approach facilitates accurate predictions, thus removing the need for trial-and-error methods even when studying completely novel phenomena, thus accelerating the design of drug molecules and catalysts while simultaneously reducing the associated costs. Commonly employed Density Functional Theory (DFT) and Wave Function Theory (WFT) based methods can achieve results well below chemical accuracy (i.e., within 1 kcal/mol of measured values) for gas phase systems. However, these methods are very limited when it comes to the description of condensed phase phenomena, which is most relevant to the industrial applications mentioned. A different family of methods called *ab initio* Molecular Dynamics (AIMD) is much better at handling such complex systems—including solutes and surfaces—while also providing dynamical properties. However, due to the high computational costs associated with these methods, compromises in accuracy must be made. Commonly, only lower levels of theory are applied, and Nuclear Quantum Effects (NQEs) are neglected.

Recently, Daru et al. introduced^{1,2} a novel technique that significantly improves upon the AIMD approach, reducing the computational costs and yielding more accurate results. The method, known as Coupled Cluster Molecular Dynamics (CCMD), integrates modern local Coupled Cluster (local CC) calculations with Machine Learning (ML), while also incorporating NQEs via path integral molecular dynamics. Unlike other methods, this procedure does not necessitate reference electronic structure calculations on periodic structures, as it relies on finite clusters, extracted from Periodic Boundary Condition (PBC) simulations instead (see Figure 2.1). This way, a local CC Potential Energy Surface (PES) can be constructed and applied via a machine learning potential in a subsequent PBC simulation.

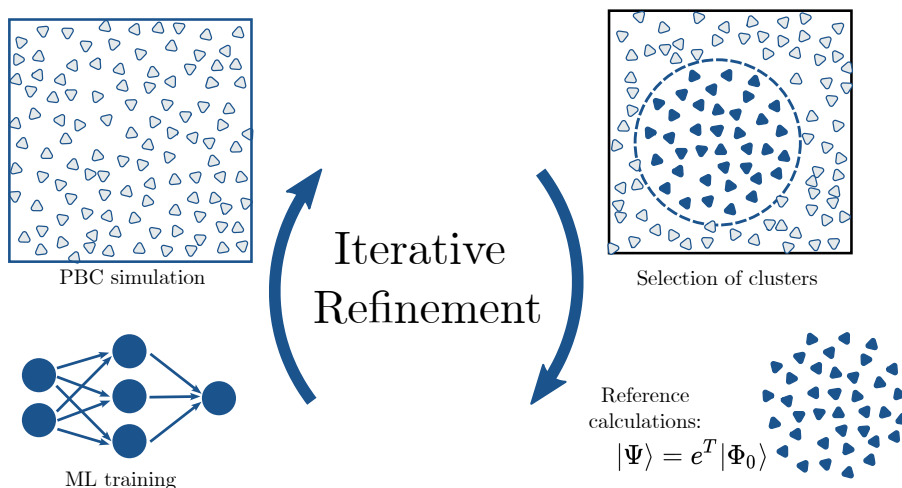


Figure 2.1: Schematic representation of the original CCMD workflow

The effectiveness of the method was demonstrated by applying it to liquid water, which traditional AIMD simulations struggle with, as the neglect of proper NQE and high-level electron correlation treatment leads to results akin to those of amorphous ice. In contrast, the CCMD approach reproduced structural (radial distribution functions) and dynamical properties (diffusion constant and orientational relaxation times) in exceptional agreement with experimental data (see Figure 2.2).

With its remarkable agreement with experimental properties, the CCMD approach holds significant promise for accurately modeling reactions in the condensed phase. However, its broader application is currently limited by the substantial computational cost of reference calculations, which required circa 3 million core hours in the original study. Thus, optimizing the method and achieving substantial speed-ups are crucial for extending its applicability to other systems.

To address this, two key areas for optimization have been identified by our group:

The first of these involves the use of DFT in place of the DLPNO-MP2 method used originally at the initial level of a two-level delta-learning protocol. In this protocol, the difference between DLPNO-MP2 and an analytical two-body potential was initially fitted. While the revised protocol does not depend on a specific DFT functional, the meta-GGA SCAN functional³ was chosen for our initial studies. SCAN adheres to all known constraints for a meta-GGA functional and effectively describes complex bonding environments and van der Waals interactions without relying on bonded information or empirical parameters.

The second focuses on lowering the costs of the local CC PES. Originally, DLPNO-CCSD(T) calculations were performed on $\sim 3,000$ clusters, each containing 64 water molecules. To achieve the desired computational savings, the more efficient LNO-CCSD(T) method is employed, and clusters of different sizes are also a subject of investigation.

In summary, the new optimized protocol integrates two machine learning (ML) potentials: a

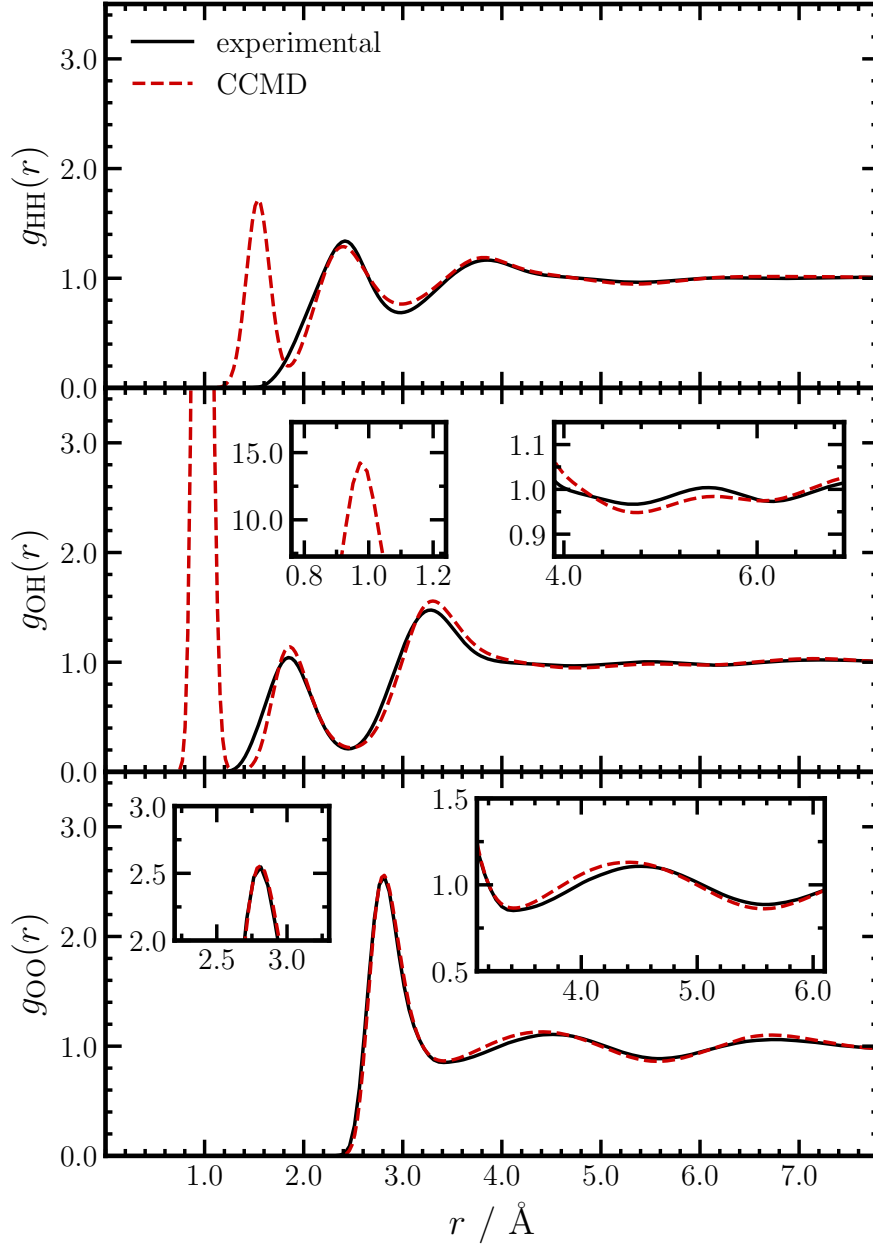


Figure 2.2: Radial Distribution Functions (RDFs) of water including nuclear quantum effects from path integral CCMD simulations at 298 K (red dashed lines) compared to experimental data from neutron diffraction ($g_{\text{HH}}(r)$ and $g_{\text{OH}}(r)$) and X-ray diffraction ($g_{\text{OO}}(r)$). The RDFs have been computed using a radial bin size of 0.01 Å as depicted and are not smoothed.

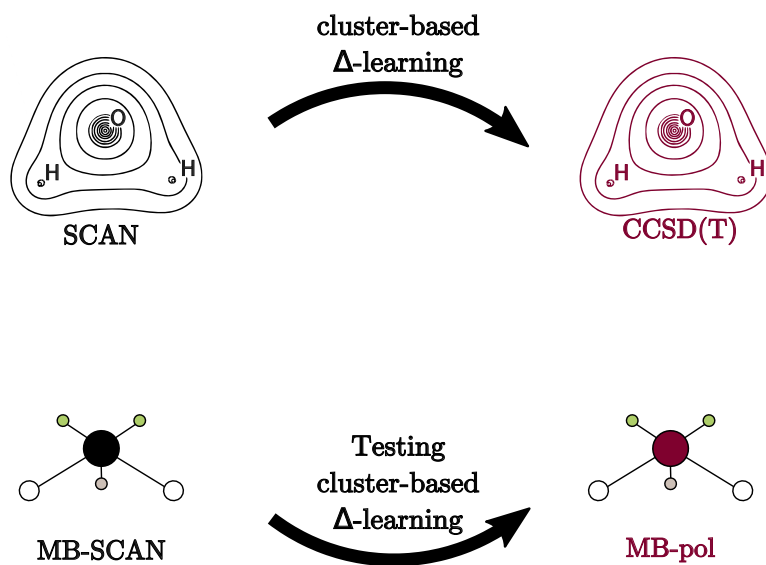


Figure 2.3: Improved, Δ -learning based scheme for CCMD

baseline model trained on periodic reference data using SCAN and a cluster-based Δ -learning model that captures electron correlation and maps to the CC PES (see Figure 2.3).

Chapter 3

Aims and Objectives

In the previous section, we introduced the exceptionally accurate but highly resource-intensive CCMD simulations and suggested possible solutions for making them more cost-effective. One of these solutions was introducing the LNO-CCSD(T) method and the possible reduction of cluster size. The main objective of this work was, therefore, to establish high-throughput methods that come at the price of a reasonable compromise in terms of accuracy—which in our case means an error in relative energies of approximately ~ 0.5 kcal/mol as this is well below the margin of error for the fitting used in the Δ -learning based scheme. Furthermore, we set out to investigate the throughput achievable for water clusters $(\text{H}_2\text{O})_n$, with $n = 5 - 64$, for which optimization of computational schemes ensuring maximal throughput must also be performed. Based on a separate study from our laboratory, 15-20 water molecules could provide a similar accuracy in the Δ -learning based scheme, as large clusters at the 100 water molecules limit.

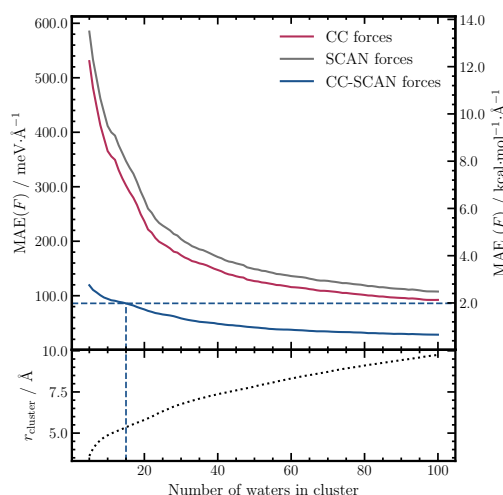


Figure 3.1: Mean absolute error comparison of forces calculated on the central oxygen of clusters wrt. the periodic reference force on the same atom. Calculated as an average of 500 clusters for each cluster size. Abbreviations: SCAN: MB-SCAN⁴, CC: MB-pol⁵. Figure courtesy of fellow group member András Szabó

Chapter 4

Theoretical Background

4.1 Basic approximations of Quantum Chemistry

Standardly employed quantum chemical methods rely on approximate solutions to the *time-independent Schrödinger-equation*:

$$\hat{H}(\mathbf{r}, \mathbf{R}) \Psi_n(\mathbf{r}, \mathbf{R}) = E_n^{\text{tot}} \Psi_n(\mathbf{r}, \mathbf{R}) \quad (4.1)$$

where $\hat{H}$ is the Hamiltonian operator, Ψ_n is the wavefunction of the n^{th} state, and E_n^{tot} is the corresponding total energy of the system, whereas $\mathbf{r}$ and $\mathbf{R}$ denote the coordinates of the electrons and the nuclei, respectively. The non-relativistic¹ molecular Hamiltonian can generally be written (in atomic units) as:

$$\begin{aligned} \hat{H} &= -\frac{1}{2} \sum_i^{\text{electrons}} \Delta_i - \sum_i^{\text{electrons}} \sum_A^{\text{Nuclei}} \frac{Z_A}{r_{iA}} + \sum_{j<i}^{\text{electrons}} \frac{1}{r_{ij}} + \sum_{B<A}^{\text{Nuclei}} \frac{Z_A Z_B}{r_{AB}} - \frac{1}{2} \sum_A^{\text{Nuclei}} \frac{1}{m_A} \Delta_A \\ &= \hat{T}_e(\mathbf{r}) + \hat{V}_{e-N}(\mathbf{r}, \mathbf{R}) + \hat{V}_{e-e}(\mathbf{r}) + \hat{V}_{N-N}(\mathbf{R}) + \hat{T}_N(\mathbf{R}) \\ &= \hat{H}_e(\mathbf{r}, \mathbf{R}) + \hat{T}_N(\mathbf{R}), \end{aligned} \quad (4.2)$$

where $\hat{H}_e$ is the *electronic Hamiltonian* — which consists of the potential energy operators $\hat{V}_{e-N}$, $\hat{V}_{e-e}$, $\hat{V}_{N-N}$ describing the electron-nucleus attraction, the electron-electron and nuclear-nuclear repulsion respectively as well as the $\hat{T}_e$ kinetic energy operator of the electrons — and $\hat{T}_{N-N}$ denotes the kinetic energy operator corresponding to the nuclei. The above equations show that nuclear and electronic motion are coupled. However, this complicated picture can be simplified by taking into account that the nuclei display a significantly greater inertia² than the electrons, and from this imparity, one can derive an adiabatic approximation in which the electrons instantaneously respond to (follow) the motion of

¹For first through third-row elements, relativistic effects are negligible.

²Already for an H atom, there is a three orders of magnitude difference between the mass of the nucleus and the electron.

the nuclei. This allows for the separation of their motion by parameterizing the dependence of the electronic energy (the eigenvalues of $\hat{H}_e$) on the nuclear coordinates. It is in this framework — called the *Born-Oppenheimer* (*BO*) approximation⁶ — that the concept of *Potential Energy Surface* (*PES*) arises, as taking the potential energy acting on the nuclei at fixed nuclear arrangements and plotting it against the internuclear distances, we are left with a function of $3N-6$ variables, describing the energetically favorable structure(s) of the molecule.

Another quintessential approximation to quantum chemical methods is the *independent electron approximation*, as it enables the description of many electron systems in a computationally feasible approach. In this model, we take advantage of the ϕ *spin-orbitals*, functions that take the spatial and spin coordinates of only a single electron as their variable and thus can be written as:

$$\phi_i = \varphi_i * \theta_i \quad (4.3)$$

with φ_i being a spatial orbital and θ_i being a spin function, both only depending on the respective coordinates of the i th electron. Using these spin-orbitals, we can decompose our n electron $\Psi(1, 2, \dots, n)$ wavefunction into the product of n $\phi_i(i)$ one-electron spin-orbitals as:

$$\Psi(1, 2, \dots, n) = \phi_1(1)\phi_2(2) \dots \phi_n(n). \quad (4.4)$$

However, to satisfy the conditions set by the Pauli principle — that that the wavefunction must be anti-symmetrized with respect to the interchange of two fermions — it is best to express our wavefunction in the form of a determinant:

$$\Psi(1, 2, \dots, n) = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & \cdots & \phi_1(n) \\ \phi_2(1) & \phi_2(2) & \cdots & \phi_2(n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(1) & \phi_n(2) & \cdots & \phi_n(n) \end{vmatrix} \quad (4.5)$$

This ansatz is the so-called called *Slater-determinant*, with n being the number of electrons, $\frac{1}{\sqrt{n!}}$ a normalization factor and $\phi_i(j)$ being the i -th spin-orbital at the coordinates of the j -th electron.

4.2 Hartree-Fock method

In this work, we employed wavefunction-based methods, which involve as their first step obtaining an approximate wavefunction, using the *Hartree-Fock(-Roothaan)*⁷⁻⁹ (*HF*) method. In this approach, the wavefunction is constructed in the form of a single Slater-determinant. The primary objective of the HF method is to determine the set of spin-orbitals that minimize the electronic energy. For this we can use a variational principle, under the constraint that the spin-orbitals remain orthonormal:

$$\int \phi_i^*(x)\phi_j(x)dx = \delta_{ij} \quad (4.6)$$

Using this, we can write the electronic energy as:

$$E_{\text{HF}} = \sum_{i=1}^N \int \phi_i^*(x) \hat{h} \phi_i(x) dx + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N [\mathbf{J}_{ij} - \mathbf{K}_{ij}] + \mathbf{H}_{nuc} \quad (4.7)$$

where $\hat{h}$ is the one-electron Fock operator:

$$\hat{h} = -\frac{1}{2} \nabla^2 - \sum_A \frac{Z_A}{|\mathbf{r} - \mathbf{R}_A|}, \quad (4.8)$$

and $\mathbf{J}_{ij}$, $\mathbf{K}_{ij}$ are the Coulomb and exchange integral respectively:

$$\mathbf{J}_{ij} = \int \int \phi_i^*(x) \phi_i(x) \frac{1}{|\mathbf{r} - \mathbf{r}'|} \phi_j^*(x') \phi_j(x') dx dx' \quad (4.9)$$

$$\mathbf{K}_{ij} = \int \int \phi_i^*(x) \phi_j(x) \frac{1}{|\mathbf{r} - \mathbf{r}'|} \phi_j^*(x') \phi_i(x') dx dx' \quad (4.10)$$

and finally $\mathbf{H}$ is the nuclear repulsion term:

$$\mathbf{H}_{nuc} = \sum_{A < B} \frac{Z_A Z_B}{r_{AB}}. \quad (4.11)$$

Meaning that the energy is a functional of the orbitals, and to minimize the total electronic energy, we are optimizing the spin-orbitals iteratively. This is achieved through the self-consistent field (*SCF*) procedure, where the Fock operator, defined as

$$\hat{f}(x) = \hat{h}(x) + \sum_{j=1}^N [\mathbf{J}_j(x) - \mathbf{K}_j(x)], \quad (4.12)$$

is used to construct a set of equations for the spin-orbitals, known as the Hartree-Fock equations:

$$\hat{f}(x) \phi_i(x) = \epsilon_i \phi_i(x), \quad (4.13)$$

where ϵ_i are the orbital energies. The Coulomb operator $\mathbf{J}_j$ and exchange operator $\mathbf{K}_j$ are given by:

$$\mathbf{J}_j(x) \phi_i(x) = \int \phi_j^*(x') \phi_j(x') \frac{1}{|\mathbf{r} - \mathbf{r}'|} dx' \phi_i(x), \quad (4.14)$$

$$\mathbf{K}_j(x) \phi_i(x) = \int \phi_j^*(x') \phi_i(x') \frac{1}{|\mathbf{r} - \mathbf{r}'|} dx' \phi_j(x). \quad (4.15)$$

The SCF procedure begins with an initial guess for the spin-orbitals, followed by the construction of the Fock operator and finally the Hartree-Fock equations are solved yielding a new set of spin-orbitals. This process is repeated until the change in energy or spin-orbitals between successive iterations falls below a predefined threshold.

The resulting Hartree-Fock wavefunction is a mean-field approximation, meaning that for each electron only an average potential created by the other electrons is considered.

4.3 Correlation methods

As discussed in the previous section, in the Hartree-Fock method, we only include a term describing the average interaction between electrons and do not consider the effects arising from the instantaneous positions of electrons and the consequential correlation of their motion. This approximation leads to significant errors in terms of the most chemically interesting properties of a molecule (energies, geometry, vibrational frequencies, etc.). The resulting discrepancy in the HF energy compared to the exact non-relativistic energy of the system is called the correlation energy, and the methods used to improve upon the results of an HF calculation by incorporating the electron correlation in some approximate form are dubbed correlation methods.

4.3.1 Configurational Interaction

Following convention, the *Configurational Interaction*¹⁰ or *CI* is the first correlation method to be discussed. The main shortcomings of the Hartree-Fock method in the description of electron correlation can be attributed to the use of a single determinant built from the occupied (lowest-energy) orbitals. The *CI* method improves upon this by expanding the wavefunction with the linear combination of multiple determinants. These determinants can be generated by replacing the i th occupied orbital in the HF wavefunction with the a^{th} virtual orbital as follows:

$$\Phi_{HF} = \Phi_0 = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & \cdots & \phi_1(n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_i(1) & \phi_i(2) & \cdots & \phi_i(n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(1) & \phi_n(2) & \cdots & \phi_n(n) \end{vmatrix} \rightarrow \Phi_i^a = \frac{1}{\sqrt{n!}} \begin{vmatrix} \phi_1(1) & \phi_1(2) & \cdots & \phi_1(n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_a(1) & \phi_a(2) & \cdots & \phi_a(n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(1) & \phi_n(2) & \cdots & \phi_n(n) \end{vmatrix}.$$

When we only replace a single orbital, it is called a singly-excited determinant (denoted as Φ_i^a above); when we replace two, a doubly-excited one, and so on. Using these excited determinants, the *CI* wavefunction can be expressed as:

$$\Psi_{CI} = c_0 \Phi_0 + \sum_{i,a} c_i^a \Phi_i^a + \sum_{i<j,a<b} c_{ij}^{ab} \Phi_{ij}^{ab} + \sum_{i<j<k,a<b<c} c_{ijk}^{abc} \Phi_{ijk}^{abc} \dots \quad (4.16)$$

where each term Φ represents a different level of excitation, with Φ_0 being the ground state reference, i.e. the HF determinant. The coefficients $c_0, c_{ij}^{ab}, c_{ijk}^{abc}, \dots$ can be determined by inserting the *CI* wavefunction into the (electronic) Schrödinger equation:

$$\hat{H}\Psi_{CI} = E\Psi_{CI} \quad (4.17)$$

and employing the method of linear variation, it transforms into a matrix eigenvalue equation for the coefficients

$$\mathbf{H}\mathbf{c} = E\mathbf{c}, \quad (4.18)$$

where $\mathbf{H}$ is the matrix representation of the Hamiltonian, with elements

$$\mathbf{H}_{IJ} = \langle \Phi_I | \hat{H} | \Phi_J \rangle \quad (4.19)$$

for all pairs of Φ_I and Φ_J .

Using the complete expansion, including all possible excitations, would lead to an exact solution on the given one-electron basis. Such a method, however, is computationally too costly for all but the simplest of systems.

4.3.2 Møller–Plesset Perturbational theory

Perturbation theory is generally used in cases considering systems which resemble readily soluble problems. As a general scheme, we decompose the problem into a model system for which the solution is known and for which there is a small perturbation. The solution is then expanded in terms of the perturbation as a series expansion around this known solution. In the series expansion, higher-order terms correspond to smaller and smaller corrections corresponding to the exact value.

In computational chemistry, perturbation theory is widely used, from descriptions of vibrations (VPT) to electronic structure theory. The latter typically uses solutions from the Hartree-Fock theory as the model system to describe the electron correlation. Møller and Plesset¹¹ developed this approach, commonly abbreviated as MPn, where n is the highest-order correction applied.

In *Møller-Plesset perturbation theory*, we start with a partitioning of the exact Hamiltonian H as:

$$H = H^{(0)} + \lambda V,$$

where $H^{(0)}$ is the unperturbed Hamiltonian, which is the Fock operator from HF theory, V is the perturbation, which represents the difference between the exact electron-electron interaction and the mean-field approximation. λ is a scaling parameter providing the model solution at $\lambda = 0$ and the exact solution at $\lambda = 1$.

We expand the wavefunction and the energy in terms of λ as:

$$\Psi = \Psi^{(0)} + \lambda \Psi^{(1)} + \lambda^2 \Psi^{(2)} + \dots$$

and

$$E = E^{(0)} + \lambda E^{(1)} + \lambda^2 E^{(2)} + \dots$$

Substituting these terms into the Schrödinger-equation, we obtain

$$\left(H^{(0)} + \lambda V \right) \left(\Psi^{(0)} + \lambda \Psi^{(1)} + \lambda^2 \Psi^{(2)} + \dots \right) = \left(E^{(0)} + \lambda E^{(1)} + \lambda^2 E^{(2)} + \dots \right) \left(\Psi^{(0)} + \lambda \Psi^{(1)} + \lambda^2 \Psi^{(2)} + \dots \right).$$

By gathering the terms corresponding to different orders of λ , we are left with

$$H^{(0)} \Psi^{(0)} = E^{(0)} \Psi^{(0)},$$

for the zeroth order,

$$H^{(0)}\Psi^{(1)} + V\Psi^{(0)} = E^{(0)}\Psi^{(1)} + E^{(1)}\Psi^{(0)}$$

for the first order and

$$H^{(0)}\Psi^{(2)} + V\Psi^{(1)} = E^{(0)}\Psi^{(2)} + E^{(1)}\Psi^{(1)} + E^{(2)}\Psi^{(0)}$$

for the second order. The first-order parts offer no improvement over the Hartree-Fock calculations' results; solutions descriptive of the electron correlation are only obtained from the second order on.

The corresponding energy term can be calculated as:

$$E^{(2)} = \sum_{i < j, a < b} \frac{|\langle \Phi_{ij}^{ab} | V | \Phi_{\text{HF}} \rangle|^2}{\epsilon_i + \epsilon_j - \epsilon_a - \epsilon_b}.$$

Here: - i, j refer to occupied orbitals, and a, b refer to virtual orbitals. - Φ_{HF} is the Hartree-Fock reference determinant. - Φ_{ij}^{ab} represents a doubly excited determinant where electrons from orbitals i and j are excited to orbitals a and b . - ϵ_i are the orbital energies from the HF solution.

The MP2 total energy thus can be given as:

$$E_{\text{MP2}} = E_{\text{HF}} + E^{(2)},$$

where E_{HF} is the Hartree-Fock energy and $E^{(2)}$ is the above-mentioned second-order energy correction term, which includes an approximation of the electron-correlation. Higher order corrections may be applied but are computationally costly, to the point that they are neglected in favor of other quantum chemical methods.

4.3.3 Coupled-Cluster theory

In Coupled Cluster^{12,13} we also use excited determinants to include an approximation of the electron-correlation. We define the following (cluster) operator for a system of n electrons:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots + \hat{T}_n \quad (4.20)$$

where the term $\hat{T}_i$ acting on the reference HF determinant Φ_0 generates all possible i th excitations as:

$$\begin{aligned} \hat{T}_1 \Phi_0 &= \sum_{i,a} t_i^a \Phi_i^a \\ \hat{T}_2 \Phi_0 &= \sum_{i < j} \sum_{a < b} t_{ij}^{ab} \Phi_{ij}^{ab} \\ &\vdots \end{aligned} \quad (4.21)$$

Here, the coefficients t are referred to as amplitudes. To obtain the CC wavefunction Ψ_{CC} , this operator is applied in an exponential ansatz:

$$\Psi_{CC} = e^{\hat{T}} \Phi_0 \quad (4.22)$$

Alternatively, we can rewrite the exponential operator as a Taylor expansion:

$$\Psi_{CC} = \left(1 + \hat{T} + \frac{1}{2}\hat{T}^2 + \frac{1}{3!}\hat{T}^3 + \dots \right) \Phi_0 \quad (4.23)$$

Finally, upon further expansion, we are left with³

$$\Psi_{CC} = \left[1 + \hat{T}_1 + \left(\hat{T}_2 + \frac{1}{2}\hat{T}_1^2 \right) + \left(\hat{T}_3 + \hat{T}_2\hat{T}_1 + \frac{1}{3!}\hat{T}_1^3 \right) + \dots \right] \Phi_0. \quad (4.24)$$

Substituting this into the Schrödinger equation, we obtain

$$\hat{H}e^{\hat{T}}\Phi_0 = E_{CC}e^{\hat{T}}\Phi_0. \quad (4.25)$$

From this, we can derive an equation for the CC energy by calculating the expectation value of the Hamiltonian:

$$\langle \Phi_0 | \hat{H}e^{\hat{T}} | \Phi_0 \rangle = E_{CC} \langle \Phi_0 | e^{\hat{T}} | \Phi_0 \rangle \quad (4.26)$$

which simplifies to:

$$E_{CC} = \langle \Phi_0 | \hat{H}e^{\hat{T}} | \Phi_0 \rangle \quad (4.27)$$

due to the normalization of the reference wavefunction Φ_0 . Although the exact energy expression only includes amplitude terms up to doubles excitations, we cannot neglect amplitudes of higher excitations as to determine their optimal values; we need to project the Schrödinger equation onto the space of determinants of different levels of excitations as:

$$\langle \Phi_i^a | \hat{H}e^{\hat{T}} | \Phi_0 \rangle = E_{CC} \langle \Phi_i^a | e^{\hat{T}} | \Phi_0 \rangle \quad (4.28)$$

$$\langle \Phi_{ij}^{ab} | \hat{H}e^{\hat{T}} | \Phi_0 \rangle = E_{CC} \langle \Phi_{ij}^{ab} | e^{\hat{T}} | \Phi_0 \rangle \quad (4.29)$$

$$\langle \Phi_{ijk}^{abc} | \hat{H}e^{\hat{T}} | \Phi_0 \rangle = E_{CC} \langle \Phi_{ijk}^{abc} | e^{\hat{T}} | \Phi_0 \rangle \quad (4.30)$$

$$\vdots \quad (4.31)$$

which creates a system of nonlinear equations. Thus, the t_i^a and t_{ij}^{ab} terms also depend on higher amplitudes. This problem can only be solved iteratively.

By including all terms of $\hat{T}$, all possible excited determinants would be generated, resulting in a wavefunction equivalent to the one obtained by full CI. This, however, is once again manageable only for the smallest of systems. To ensure the feasibility of computations for larger systems, a truncation must be employed. Most commonly, this truncation is chosen as:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 \quad (4.32)$$

which is labeled as *Coupled Cluster Singles and Doubles* or $CCSD$ ¹⁴ for short. Although the largest contributions usually come from the corresponding amplitudes of t_i^a and t_{ij}^{ab} , for the reasons discussed

³Here, the terms are collected in parenthesis based on the level of excited states they generate.

above, it is best to include the contributions of some higher-level terms as well. However, an alternative method has been developed to reduce the costs of the otherwise computationally demanding approach of CC calculations where higher terms are included explicitly, incorporating an estimate of these higher excitations. For $CCSD$ we include a correction term $(T)^{15}$ which is an approximation of the contribution of triple excitations. This contribution of the $\hat{T}_3$ term is treated as a perturbation, with the triples amplitudes being determined by using the $CCSD$ amplitudes as follows:

$$t_{ijk}^{abc} = \frac{\langle \psi_i \psi_j \psi_k | | \psi_a \psi_b \psi_c \rangle}{\Delta_{ijk}^{abc}}. \quad (4.33)$$

where ψ_i, ψ_j, ψ_k are occupied and ψ_a, ψ_b, ψ_c are virtual orbitals, furthermore the energy denominator Δ_{ijk}^{abc} can be given as:

$$\Delta_{ijk}^{abc} = \epsilon_i + \epsilon_j + \epsilon_k - \epsilon_a - \epsilon_b - \epsilon_c, \quad (4.34)$$

with ϵ_o , ($o = i, j, k$) and ϵ_v , ($v = a, b, c$) being the Hartree-Fock orbital energies of the occupied and virtual orbitals respectively.

The energy correction term resulting from (T) is then given by the following expression:

$$E^{(T)} = \sum_{i,j,k}^{\text{occ}} \sum_{a,b,c}^{\text{virt}} \frac{|\langle \psi_i \psi_j \psi_k | | \psi_a \psi_b \psi_c \rangle t_{ijk}^{abc}|}{\Delta_{ijk}^{abc}}. \quad (4.35)$$

and by expanding the triples amplitudes in the above equation, we obtain:

$$E^{(T)} = \sum_{i,j,k}^{\text{occ}} \sum_{a,b,c}^{\text{virt}} \frac{|\langle \psi_i \psi_j \psi_k | | \psi_a \psi_b \psi_c \rangle|^2}{\left(\Delta_{ijk}^{abc}\right)^2}. \quad (4.36)$$

The $CCSD(T)$ is a very effective method, providing reliable, accurate results, and is therefore commonly chosen as a reference method for benchmarking computations.

4.3.4 Local-Correlation Methods

As accurate as the $CCSD(T)$ method is, it is limited in its applicability for larger systems (20-25+ atoms) due to its scaling of $\mathcal{O}(N^7)$ with the system size N . To improve upon this *local-correlation*, methods have been developed. The main idea behind these methods is that the electron correlation decays steeply with the distance. This phenomenon can be exploited by decomposing the system in question into subsystems and considering the electron correlation between spatially close parts, using *Localized Molecular Orbitals* (*LMOs*). These can be obtained by transforming a set of occupied canonical molecular orbitals.

4.3.4.1 DLPNO-CCSD(T)

The first set of widely employed local-correlation methods to be discussed are the ones based on the *Domain-Based Local Pair Natural Orbital* (*DLPNO-*) approach^{16,17}, of which I will detail the parts

relevant to a DLPNO-CCSD(T) calculation. As the first step of this method, the occupied orbitals are localized by a unitary transformation, followed up by defining the external space spanned by the unoccupied orbitals

$$|\tilde{\mu}'\rangle = (1 - \hat{P})|\mu\rangle, \quad (4.37)$$

where μ are atomic orbitals, and $\hat{P}$ is the projection operator of the localized occupied space, and the resulting $\tilde{\mu}$'s are called *Projected Atomic Orbitals (PAOs)*. Then, orbital domains from occupied orbitals $|i\rangle$ and two sets of pair domains for each electron pair ij are constructed. Then, a prescreening of the electron pairs ij is performed by evaluating differential overlap integrals (DOI):

$$DOI_{ij} = \sqrt{\int |f_i(\mathbf{r})|^2 |g_j(\mathbf{r})|^2 d\mathbf{r}}, \quad (4.38)$$

where $f_i(\mathbf{r})$ and $g_j(\mathbf{r})$ are functions describing sparsity as well as computing an upper bound to the dipole-dipole interaction energy. If the values thus derived exceed some threshold pre-set for each method respectively, the given ij pair is then subjected to LMP2 treatment. If neither of the values surpasses their respective limit, the pair is excluded from the LMP2 calculations, and the corresponding energy contributions are derived from the following equation for the dipole-dipole interaction energy $\varepsilon_{ij}^{\text{DIP}}$:

$$\varepsilon_{ij}^{\text{DIP}} = -\frac{4}{R_{ij}^6} \sum_{\tilde{\mu}_i, \tilde{\nu}_j} \frac{\left\{ \langle i | \mathbf{r} | \tilde{\mu}_i \rangle \langle j | \mathbf{r} | \tilde{\nu}_j \rangle - 3 \left(\langle i | \mathbf{r} | \tilde{\mu}_i \rangle \cdot \hat{\mathbf{R}}_{ij} \right) \left(\langle j | \mathbf{r} | \tilde{\nu}_j \rangle \cdot \hat{\mathbf{R}}_{ij} \right) \right\}^2}{\varepsilon_{\tilde{\mu}_i} + \varepsilon_{\tilde{\nu}_j} - F_{ii} - F_{jj}},$$

where:

- R_{ij} is the distance between orbitals i and j .
- $\tilde{\mu}_i$ and $\tilde{\nu}_j$ denote the dipole moments, each associated with orbitals i and j .
- $\mathbf{r}$ is the position vector of the electrons.
- $\hat{\mathbf{R}}_{ij}$ is the vector pointing from orbital i to j .
- F_{ii} and F_{jj} denote the Fock matrix elements.

For the pairs included in the LMP2 treatment, amplitudes and pair correlation energies are obtained. Using these amplitudes, pair-specific densities are constructed:

$$\mathbf{D}^{ij} = \tilde{\mathbf{t}}^{ij} \mathbf{t}^{ij+} + \tilde{\mathbf{t}}^{ij+} \mathbf{t}^{ij} \quad (4.39)$$

Diagonalization of this leads to the *Pair Natural Orbitals (PNOs)* and their occupation numbers (n^{ij})

$$\mathbf{D}^{ij} \mathbf{d}^i j = n^{ij} \mathbf{d}^i j \quad (4.40)$$

these occupation numbers are then used to filter the PNOs included in the higher-order correction calculations.

Thus, in summary, the preparatory phase of a DLPNO-CCSD calculation consists of the localization of the occupied orbitals, the construction of orbital domains, and the construction of two sets of pair domains for each electron pair ij .

4.3.4.2 LNO-CCSD(T)

The most relevant local-correlation method to this work is the *Local Natural Orbital* (LNO-) CCSD(T) method^{18–21}. To discuss this, let us first look at a general case. It can be stated that the correlation energy in given method M1 can be expressed exactly as a sum of contributions from occupied orbitals:

$$E_{\text{corr}}^{\text{M1}} = \sum_{ijab} I_{ij,ab} C_{ij,ab}^{\text{M1}} = \sum_i^\delta E_i^{\text{M1}} \quad (4.41)$$

where $C_{ij,ab}^{\text{M1}}$ represents the wavefunction parameters and $I_{ij,ab}$ the electron repulsion integrals (ERIs) in canonical *MOs*. In the more specific case of $\text{M1} = \text{CCSD}(T)$ the energy is obtained as:

$$E^{\text{CCSD}(T)} = \sum_i \delta E_i^{\text{CCSD}} + \sum_i \delta E_i^{(T)} = \sum_i \delta E_i^{\text{CCSD}(T)}, \quad (4.42)$$

, where the CCSD contributions δE_i^{CCSD} are derived as:

$$\delta E_i^{\text{CCSD}} = 2 \sum_a f_{ai} t_i^a + \sum_{abj} L_{ij}^{ab} (t_{ij}^{ab} + t_i^a t_j^b) \quad (4.43)$$

with f_{ai} being an element of the Fock-matrix, terms t_i^a and t_{ij}^{ab} cluster amplitudes for single and double excitations, and

$$L_{ij}^{ab} = 2(ai|bj) - (aj|bi)$$

with the two-electron ERIs $(pq|rs)$.

Local approximations can be applied by recognizing that the δE_i^{corr} contributions remain unchanged under independent unitary transformations of both occupied and virtual orbitals. This invariance allows the substitution of the i canonical occupied orbitals with localized *MOs*, i' . Each such LMO i' defines an extended domain $(\text{ED}, \mathcal{E}_i)$ containing itself and the LMOs it forms a strong pair with. Strong and distant pairs are distinguished by evaluating MP2 pair correlation energy $\delta E_{i'j'}^{\text{MP2}}$ in the $i'j'$ pair-specific domains $(\mathcal{P}_{i'j'})$. The distant pairs are excluded from higher-level computations. The virtual space of each ED is then limited to the local virtual orbitals (PAOs) near the LMOs it contains. Then, for each ED, the $\delta E_{i'}^{\text{MP2}}(\mathcal{E}_{i'})$ MP2 correlation energy contributions are computed, which also allows for the construction of local natural orbitals (LNOs) through the use of the MP2 amplitudes²².

A local interacting subspace (LIS, $P_{i'}$) is then created for each central LMO i' by compressing the full LMO-PAO basis of the ED into a reduced set of LNOs. The LNO-CCSD(T) energy is then given as:

$$E^{\text{LNO-CCSD(T)}} = \sum_{i'} \left[\delta E_{i'}^{\text{CCSD(T)}}(P_{i'}) + \Delta E_{i'}^{\text{MP2}} \right], \quad (4.44)$$

with the correction term

$$\Delta E_{i'}^{\text{MP2}} = \delta E_{i'}^{\text{MP2}}(\mathcal{E}_{i'}) - \delta E_{i'}^{\text{MP2}}(P_{i'}) + \frac{1}{2} \sum_{j'}^{\text{distant}} \delta E_{i'j'}^{\text{MP2}}(P_{i'j'}), \quad (4.45)$$

accounting for pair and LNO approximations at the MP2 level.

4.3.4.3 Density Fitting

For further reducing the computational costs of the local correlation calculations, the method of *density fitting* (DF)^{23,24} is widely employed. More specifically, this method is used to remedy the bottleneck caused by the $\mathcal{O}(N^4)$ scaling in the number of significant four-center two-electron integrals:

$$(pq|rs) = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \phi_p^*(\mathbf{r}_1) \phi_q(\mathbf{r}_1) \frac{1}{r_{12}} \phi_r^*(\mathbf{r}_2) \phi_s(\mathbf{r}_2), \quad (4.46)$$

where ϕ_i s represent some general MOs, while $\mathbf{r}_1$ and $\mathbf{r}_2$ denote the coordinates of electrons 1 and 2 respectively. The products $|pq\rangle$ and $|rs\rangle$ may also be taken as densities ρ_{pq} and ρ_{rs} respectively. To resolve the issue of high-order scaling, in the *density fitting* method, we introduce an auxiliary basis set, which can be used to approximate the densities as:

$$\rho_{pq}(\mathbf{r}) \approx \tilde{\rho}_{pq}(\mathbf{r}) = \sum_P d_P^{pq} \chi_P(\mathbf{r}) \quad (4.47)$$

where the d_K^{pq} fitting coefficients are obtained by minimizing the deviation between the fitted and actual densities:

$$\Delta_{pq} = (pq - \tilde{p}q | pq - \tilde{p}q) \quad (4.48)$$

which in turn minimizes the error in the electric field and leads to the following set of linear equations:

$$d_P^{pq} = \sum_P (pq|P) [\mathbf{J}^{-1}]_{PQ} \quad (4.49)$$

where

$$(pq|P) = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \phi_p(\mathbf{r}_1) \phi_q(\mathbf{r}_1) \frac{1}{r_{12}} \chi_P(\mathbf{r}_2) \quad (4.50)$$

and

$$J_{PQ} = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \chi_P(r_1) \frac{1}{r_{12}} \chi_Q(\mathbf{r}_2) \quad (4.51)$$

Thus we can replace the four-index, two-electron integral $(pq|rs)$ with

$$\begin{aligned}
 (pq|rs) &\approx (\tilde{p}q|\tilde{r}s) = \int d\mathbf{r}_1 \int d\mathbf{r}_2 \tilde{\rho}_{pq}(\mathbf{r}_1) \frac{1}{r_{12}} \rho_{rs}(\mathbf{r}_2) \\
 &= \int d\mathbf{r}_1 \int d\mathbf{r}_2 \sum_Q d_Q^{pq} \chi_Q(\mathbf{r}_1) \frac{1}{r_{12}} \phi_r(\mathbf{r}_2) \phi_s(\mathbf{r}_2) \\
 &= \sum_Q d_Q^{pq} (Q|rs) \\
 &= \sum_{PQ} (pq|P) [\mathbf{J}^{-1}]_{PQ} (Q|rs).
 \end{aligned} \tag{4.52}$$

By reducing the four-index integral to the product of three-index ones this way, a speed-up of about a factor of 10 can be achieved, and it also leads to a significant down-scale in the memory requirements of the computations. Therefore, density fitting is extensively employed during HF calculations, as well as various parts of the (local-)correlation methods.

4.4 Basis sets

In the previous sections we relied on the concept of spin-orbitals ϕ . During practical calculations, to best represent these orbitals of unknown form, we expand them in the basis of known functions:

$$\phi_i = \sum_k^{\infty} c_{ik} \chi_k \approx \sum_k^K c_{ik} \chi_k, \tag{4.53}$$

Where the $\{\chi_k\}$ set of functions is the *basis set*, which — as indicated — unless a complete set, can only approximate the exact form of the spin-orbitals. This approximation imposes an inherent limitation on the accuracy of our calculations, regardless of the quantum chemical method employed. The range of this inaccuracy is dependent on the type and size of the basis set. The more accurately the (combination of) basis functions can represent the orbitals, the better our results will be. One suitable choice in the search for the best possible representation of the orbitals is the set of functions derived from the analytical solution for the non-relativistic Schrödinger equation for hydrogen-like atoms, as these are exact and known. These contain radial $R(r)$ and angular $Y(\theta, \phi)$ parts. However, it became evident that the use of these was not practical and thus was replaced by so-called *Slater-type orbitals* (STOs). The angular part of these functions also consists of spherical harmonics, but unlike the hydrogen-like orbitals, include some $e^{-\alpha r}$ like term to describe the angular dependence. After some time, the STOs, too, proved overly tedious to work with, and in search of alternatives *GTOs*, *Gaussian-type orbitals* were introduced. To ease computations, the angular part of these functions relies on Cartesian coordinates of the form $x^i y^j z^k$, while the radial part contains a $e^{-\alpha r^2}$ like term. While the use of GTOs is — as mentioned — computationally less cumbersome, this comes at the cost

of significant deviations in describing the actual behavior of atomic orbitals. To illustrate one of the key differences, let us consider the derivatives of the exponential term in the radial part of STOs and GTOs, respectively:

$$\left[e^{-\alpha r}\right]'_{STO} = -\alpha \cdot e^{-\alpha r}; \text{ let } r = 0 : -\alpha \cdot e^{-\alpha 0} = -\alpha \quad (4.54)$$

$$\left[e^{-\alpha r^2}\right]'_{GTO} = -2\alpha r \cdot e^{-\alpha r^2}; \text{ let } r = 0 : -2\alpha \cdot 0 \cdot e^{-\alpha 0^2} = 0. \quad (4.55)$$

The vanishing derivative of GTOs at the point of the nucleus — as opposed to the cusp observed in STOs — leads to an incorrect description of electron density close to the nucleus. Another critical difference between the two types of basis functions is the steeper decay of GTOs at large values of r (due to their dependence on r^2), leading to inaccuracies in long-range interactions. To overcome these hindrances, a pre-set combination of Gaussian functions (dubbed *contracted* GTOs, *CGTOs* or *contracted basis*), better imitating the behavior of STOs, may be used:

$$\chi_k^{CGTO} = \sum_n^N d_{kn} g_n, \quad (4.56)$$

with g_n being so-called *primitive Gaussians*, and d_{kn} being the corresponding coefficients, which are usually not optimized further during a calculation, unlike those corresponding to the AOs χ_k . The essence of using contracted bases lies in representing an STO as a linear combination of multiple GTOs. The simplest case with practical relevance is when each STO corresponding to an atomic orbital is approximated by 3 primitive Gaussian functions, a scheme denoted as STO-3G. By increasing the number of primitive functions, one can construct the basis set STO- n G, where n denotes the number of primitive functions.

A more complex but also more accurate approach involves using so-called split-valence shell basis sets. The concept of these is that the core orbitals are described by a single ' ζ ' contracted basis, while the valence orbitals are split into multiple ζ bases, i.e., several differently composed contracted bases. For instance, the 3-21G designation means that the core orbitals are built from 3 primitive functions (as indicated by the leading '3'), and the valence orbitals are described by double- ζ bases, one of which consists of 2 primitive functions and the other of 1 primitive function.

Polarization and diffuse functions are also crucial to accurately describe certain types of physical and chemical phenomena. Polarization functions introduce orbitals corresponding to higher angular momenta than those of the valence orbitals, to facilitate the description of the spatial distribution of the electron density, allowing for the distortion and polarization of the wavefunction. Diffuse functions, characterized by a small exponent α in the radial term $e^{-\alpha r^2}$, leading to a slow decay, which is important to describe long-range interactions such as van der Waals forces and hydrogen bonds, as well as to capture the diffuse nature of anionic systems.

We must also discuss the *correlation-consistent* (*cc*) basis sets developed by Dunning and colleagues as we relied on this family of basis sets during our calculations. These bases inherently include polarization functions, enhancing the reliability of the results obtained with them. They are generally denoted as cc-pVXZ, which stands for 'correlation-consistent polarized valence X- ζ ' basis set, where X can be $D, T, Q, 5, 6, \dots$ (double, triple, quadruple, $\dots$). If the given basis set also incorporates diffuse functions, the designation is extended with the prefix 'aug-', leading to notations like aug-cc-pVXZ. For density fitting calculations expanded, auxiliary versions of these basis sets also exist.

4.4.1 CBS extrapolation

A great advantage to this group of basis sets is that they are designed so that by increasing the cardinal number X they produce results which systematically converge to the best theoretical result achievable with a given quantum chemical method, i.e. one using an infinite number of basis functions. This is called the *Complete Basis Set (CBS)* limit. One can therefore perform extrapolations to the CBS limit to further increase the accuracy, based on results achieved with multiple correlation consistent basis. Multiple extrapolation schemes exist, here I will detail only the ones employed during our work.

To approximate the convergence of the Hartree-Fock energy to the CBS limit, we used the two-point extrapolation scheme introduced by *Karton* and *Martin*²⁵:

$$E_{\text{HF}}^{(X)} = E_{\text{HF}}^{(\infty)} + Ae^{-\alpha\sqrt{X}} \quad (4.57)$$

which, when rearranged takes the form:

$$E_{\text{HF}}^{(\infty)} = \frac{E_{\text{HF}}^{(X-1)} \cdot e^{-\alpha\sqrt{X}} - E_{\text{HF}}^{(X)} \cdot e^{-\alpha\sqrt{X-1}}}{e^{-\alpha\sqrt{X}} - e^{-\alpha\sqrt{X-1}}}. \quad (4.58)$$

Values for α were taken from *Neese* and *Valeev*²⁶ for $X = T, Q$ and from *Karton* and *Martin*²⁵ for $X = 5$.

Finally, for the correlation energy the cubic formula of *Helgaker et al.*²⁷ was used:

$$E_{\text{corr}}^{(\infty)} = \frac{(X-1)^3 E_{\text{corr}}^{(X-1)} - X^3 E_{\text{corr}}^X}{(X-1)^3 - X^3} \quad (4.59)$$

4.4.2 Density-Based Basis Set Corrections

Wave function theory-based calculations of the correlation energy notoriously suffer from a slow convergence with the size of the one-electron basis set. This hindrance stems from the inability of conventionally employed Gaussian basis sets to properly handle the electron-electron cusp of the wavefunctions. To mitigate this a number of methods have been introduced. In this work we have employed the recently emerged *Density-Based Basis Set Correction (DBBSC)* approach,^{28,29} more specifically its developer-version implemented in the MRCC program suite.³⁰ This method combines DFT and

WFT to address short- and long-range interactions, respectively, leveraging each approach for the interactions they handle more effectively, to estimate the correlation energy at the CBS limit. For a given combination of an incomplete basis set $\mathcal{B}$ and a CC method, this approximate CBS correlation energy ($E_{\text{CC},c}^{\text{CBS}}$) can be expressed as:

$$E_{\text{CC},c}^{\text{CBS}} \approx E_{\text{CC},c}^{\mathcal{B}} + E_{\text{DBBSC}}^{\mathcal{B}}[n] \quad (4.60)$$

where $E_{\text{CC},c}^{\mathcal{B}}$ is the CC correlation energy in the finite basis set and $E_{\text{DBBSC}}^{\mathcal{B}}[n]$ is a basis-set dependent complementary density functional correction, with n being the one-electron density. The integration of DFT and WFT is accomplished by constructing an effective operator that represents the electron-electron Coulomb operator in a localized real-space form projected onto $\mathcal{B}$. This can be done in a way to include the divergence displayed by the Coulomb operator at $r_{12} = 0$. Under proper conditions, this effective operator ($W^{\mathcal{B}}$) can be constructed so that it satisfies the following condition:

$$\int \int \frac{n_2^{\mathcal{B}}(\mathbf{r}_1 \mathbf{r}_2)}{r_{12}} d\mathbf{r}_1 d\mathbf{r}_2 = \int \int W^{\mathcal{B}}(\mathbf{r}_1, \mathbf{r}_2) n_2^{\mathcal{B}}(\mathbf{r}_1, \mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 \quad (4.61)$$

The effective operator constructed this way can then be linked to a local range-separation function that dynamically adjusts to account for the spatial incompleteness of $\mathcal{B}$, for which the following formula has been proposed:

$$\mu(\mathbf{r}) = \sqrt{\frac{\pi}{2}} W^{\Psi^{\mathcal{B}}}(\mathbf{r}, \mathbf{r}). \quad (4.62)$$

Using this, the complementary energy correction functional can be approximated as:

$$E_{\text{DBBSC}}^{\mathcal{B}}[n] = \int n(\mathbf{r}) \varepsilon_{\text{MD},c}^{\text{sr-PBE}}(n(\mathbf{r}), s(\mathbf{r}), \zeta(\mathbf{r}), \mu(\mathbf{r})) d\mathbf{r}, \quad (4.63)$$

where $\varepsilon_{\text{MD},c}^{\text{sr-PBE}}(n(\mathbf{r}))$ is a special Perdew-Burke-Ernzerhof³¹ type short-range correlation functional with multideterminantal (MD) reference, $s(\mathbf{r})$ is the reduced density gradient, and $\zeta(\mathbf{r})$ is the spin polarization. Although the DBBSC approach proves to be quite effective in the description of the correlation energy in the CBS limit, the results are dependent on the HF result, which itself is affected by the incompleteness of the basis sets used in a calculation. However, it has been shown that the inclusion of CABS corrections to the HF energies leads to significant improvements in the accuracy of the calculations. Therefore, in the MRCC implementation of this approach, the final results of a CC calculations are obtained based on the following expression:

$$E_{\text{CC}}^{\text{CBS}} \approx E_{\text{HF}}^{\mathcal{B}} + E_{\text{CABS}}^{\mathcal{B}} + E_{\text{CC},c}^{\mathcal{B}} + E_{\text{DBBSC}}^{\mathcal{B}}[n]. \quad (4.64)$$

Chapter 5

Results and Discussion

5.1 Water Cluster

5.1.1 Computational scheme optimization for high-throughput calculations

We began our optimization of high-throughput quantum chemical calculation schemes by trying out several different computational approaches for water clusters $(\text{H}_2\text{O})_n$, with $n = 5-64$. We focused on varying the number of parallel jobs run simultaneously on a single node and, thus, the number of cores and Rapid Access Memory (RAM) distributed among them. To standardize this investigation, we performed LNO-CCSD(T)/cc-pVTZ Single-Point Energy (SPE) calculations with Tight-LNO-CC thresholds on the aforementioned water clusters, using the MRCC program system^{32,33}. Given its common use in contemporary quantum chemical research, this medium-sized basis set was chosen as a reasonable and representative option. The Tight threshold settings were selected because they represent an (approximate) upper limit for wall times (or a lower limit for throughput). Following some prior trials on the largest cluster ($(\text{H}_2\text{O})_{64}$), we decided to investigate the three following computational schemes: 8 parallel jobs with 16 cores and 26 GB of memory per job (*8P*); 16 parallel jobs with 8 cores and 13 GB of memory per job (*16P*); 32 parallel jobs with 4 cores and 6.5 GB of memory per job (*32P*). The figures of core and RAM are derived from distributing the resources of a single node of the CPU partition of the Komondor HPC system. For further details see the caption of Figure 5.1¹ where the corresponding results are shown, excluding the two smallest clusters ($n = 5, 10$) to aid visualization.

¹Or visit: <https://hpc.kifu.hu/en/komondor>

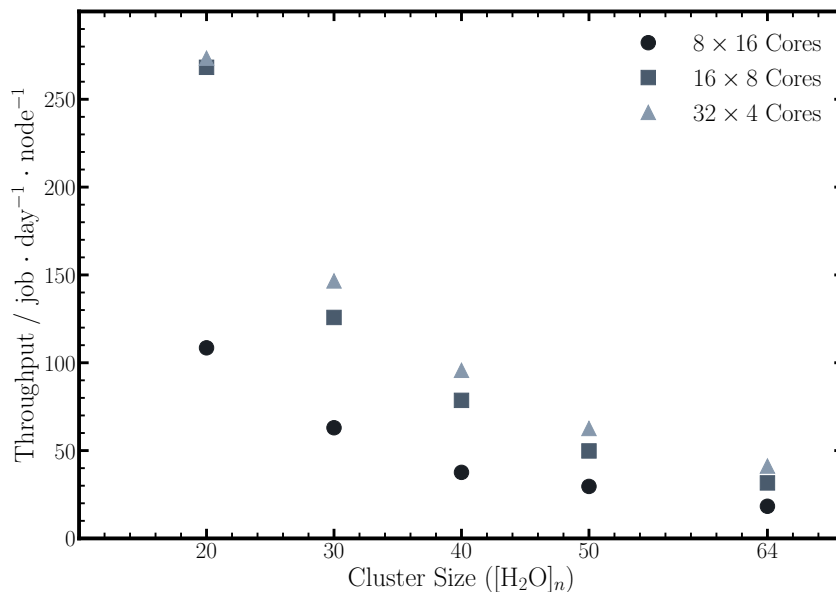


Figure 5.1: Performance of different computational schemes investigated on water clusters of varying sizes, as indicated by throughput values. Markers of different shapes and colors represent the computational schemes employed, as follows: Circle—*8P*: 8 parallel jobs with 16 cores and 26 GB of memory per job; Square—*16P*: 16 parallel jobs with 8 cores and 13 GB of memory per job; Triangle—*32P*: 32 parallel jobs with 4 cores and 6.5 GB of memory per job. All calculations were performed on a single node equipped with 2×64 -core AMD EPYC™ 7763 (Milan) CPUs and 256 GB of RAM, available on the CPU partition of the Komondor HPC facility of KIFU (Governmental Agency for IT Development, Hungary, <https://ror.org/01s0v4q65>). The throughput values were obtained from LNO-CCSD(T)/cc-pVTZ calculations using the Tight threshold settings for the LNO-CC method.

As seen in the above figure, using more cores and memory on a single job improves the achievable throughput as much as increasing the number of parallel jobs, as a single calculation cannot be parallelized beyond a certain limit. Increasing the cluster size n leads to scaling in wall times of approximately n^2 in all cases. Clearly, the scheme with 8 parallel jobs performs significantly worse than the other two approaches. The methods with 16 and 32 parallel jobs running are quite close in terms of throughput, with the latter slightly outperforming the former. Based on the results presented thus far, this setup seems to be the ideal choice, however our follow-up investigations determined that due to the low core-to-job and memory-to-job ratio associated with the method, it is no longer viable for calculations with larger basis sets. Therefore, we settled on employing the scheme with 16 parallel jobs, each with 8 cores and 13 GB RAM to work with. The fact further supports this choice that advances in our research group reduced the necessary cluster size for optimal training of Machine Learning (ML) models to $n = 20$; and the *32P* and *16P* schemes for the $(\text{H}_2\text{O})_{20}$ clusters are almost equal. However, it is not clear whether this is a systematic occurrence or just an artifact of a malfunction in the node

of the $32P$ scheme.

Finally, we note that when performing DLPNO-CCSD(T) calculations with the same triple- ζ basis set (as implemented in the ORCA software package, used formerly in the CCMD scheme), only a single parallel job/node could be successfully completed, thus leading to far lower throughput values than the LNO-CC method. For a cluster of 64 molecules, a throughput of only $3.6 \text{ jobs} \cdot \text{day}^{-1} \cdot \text{node}^{-1}$ was observed, whereas for a cluster of 20 we determined the throughput value to be $55 \text{ jobs} \cdot \text{day}^{-1} \cdot \text{node}^{-1}$ when using the DLPNO-CCSD(T) method. Therefore, by simply replacing the DLPNO-CCSD(T) method used in the reference cluster calculations in the CCMD scheme with the LNO-CCSD(T) method, we already gain an 8-fold speed-up in the time necessary for the PES development. While if we also decrease the system size to just 20 molecules of water—which has been confirmed not to interfere with the accuracy of the results by separate results from our laboratory—we can achieve a 70-fold increase in throughput, when compared to the approach used in the original work (DLPNO-CCSD(T) reference calculations on $(\text{H}_2\text{O})_{64}$ clusters).

5.1.2 Error analysis of quantum chemical methods

Given that promising results were achieved for the Δ -learning model using smaller clusters, this next section of our work focuses on water clusters containing 20 species.

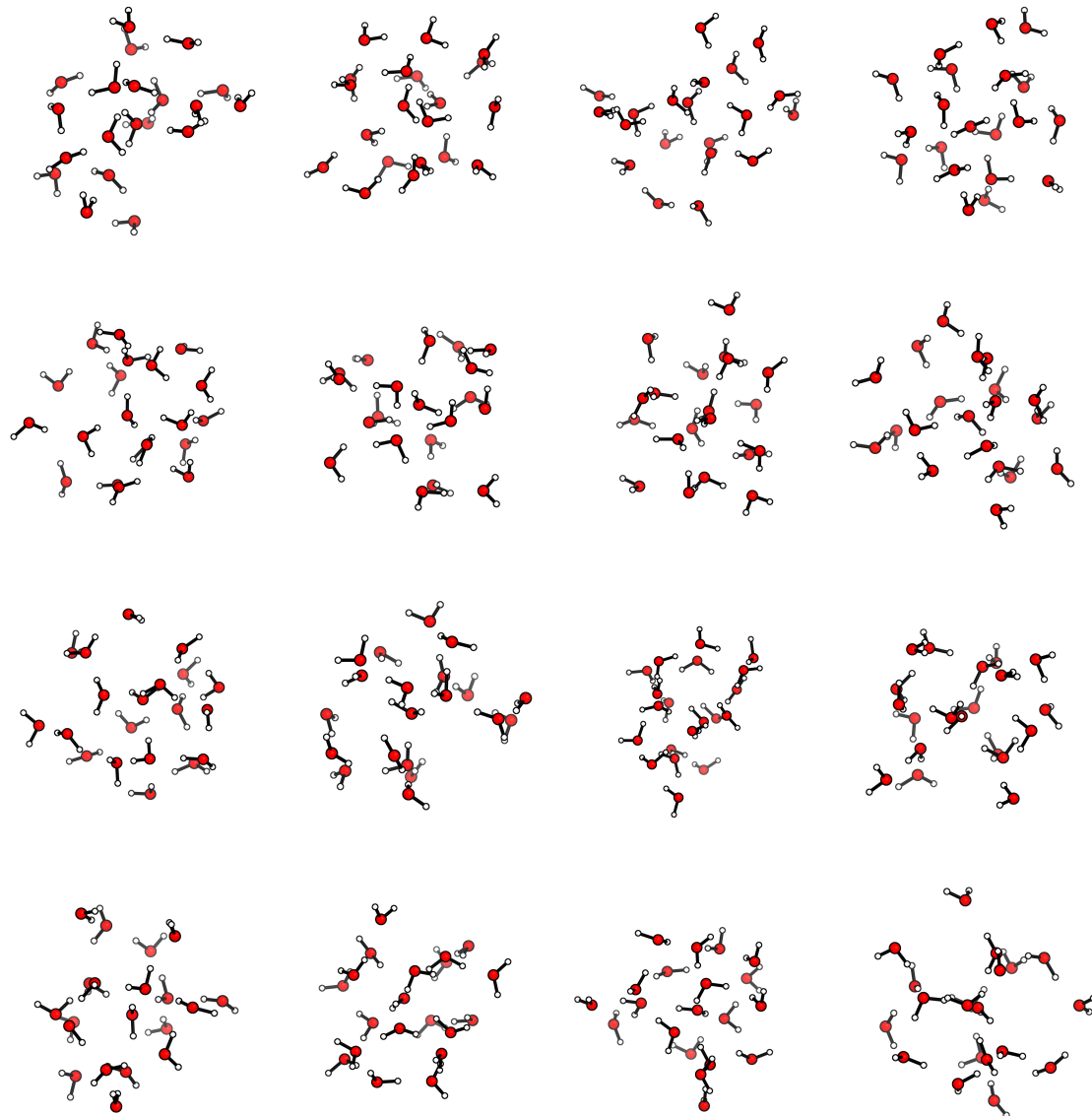


Figure 5.2: Geometries of the $(\text{H}_2\text{O})_{20}$ clusters used in our investigations.

Most often, one finds that quantum chemical methods offering highly-accurate results come at the price of significant computational costs. Here, we attempt to look for an ab initio, wavefunction-based method that leads to reliable results in the context of our work in the research group (i.e.: errors well-within the error margin of the model fitting), but still facilitates a high-throughput. In this spirit, we investigated the LNO-CCSD(T) method with Dunning-type correlation consistent basis sets, with cardinal numbers 2–4 (and their diffuse augmented counterparts), while leveraging different

possible local-correlation thresholds (namely Loose-, Normal- and Tight-LNO-CC). Additionally, we were among the first to probe the effectiveness of the DBBSC approach^{30,34}. In all cases, Single Point Energy (SPE) calculations were performed, on geometries obtained from MB-SCAN⁴ simulations.

Although the computational scheme proposed in the previous section generally performed well for calculations with the most extensive employed basis sets, a high rate of job failures prompted a deviation from this approach for the Normal—AQZ, Tight—QZ, ATZ, and AQZ calculations. Specifically: For both Normal-LNO-CC/aug-cc-pVQZ and Tight-LNO-CC/aug-cc-pVTZ calculations, two nodes were utilized, each running 8 parallel jobs. The node resources were allocated according to the *8P* scheme described earlier. Even more resources were required for optimal performance for Tight-LNO-CC/cc-pVQZ calculations and their counterparts with diffuse-augmented basis sets. These calculations were performed using 4 parallel jobs on a single node, with each job allocated 32 cores and 52 GB of RAM. Otherwise, all calculations were performed using the previously described *16P* scheme, using 16 different geometries.

To establish a reliable method, to be used as a reference in the subsequent analysis, we have also performed Tight-LNO-CCSD(T)/CBS(QZ,5Z) and Tight-LNO-CCSD(T)/CBS(ATZ,AQZ) calculations on (H₂O)₂₀ clusters and compared their performance against benchmark results available for the same geometries. These results are presented in the tables below.

	CBS(QZ,5Z)	CBS(ATZ,AQZ)
dodecahedron	-0.5	-0.1
face-cubes	0.9	0.6
face-sharing	0.6	0.7
Mean Absolute Error	0.7	0.5

Table 5.1: Comparison of the accuracy of Tight-LNO-CCSD(T)/CBS(QZ,5Z) and Tight-LNO-CCSD(T)/CBS(ATZ,AQZ) calculations for three different (H₂O)₂₀ cluster geometries. The errors of relative energies in kcal/mol units as compared to the MP2-F12/CBS(ATZ-F12,AQZ-F12) results as published by *Manna et al.*³⁵

In Table 5.1, we see that both methods we employed underestimate the energy compared to the reference results in the case of the dodecahedron structure, while they overestimate the values corresponding to face-cubes and face-sharing structures. Looking at the Mean Absolute Error (MAE) values, we see that the CBS extrapolation involving the diffuse augmented basis sets with lower cardinal numbers leads to more accurate results.

	CBS(QZ,5Z)	CBS(ATZ,AQZ)
dodecahedron	-2.0	-1.6
face-cubes	1.7	1.4
face-sharing	0.8	0.9
Mean Absolute Error	1.5	1.3

Table 5.2: Comparison of the accuracy of Tight-LNO-CCSD(T)/CBS(QZ,5Z) and Tight-LNO-CCSD(T)/CBS(ATZ,AQZ) calculations for three different (H₂O)₂₀ cluster geometries. The table shows errors of relative energies in kcal/mol units as compared to the MP2-F12/CBS(ATZ-F12,AQZ-F12)+High-level corrections from *n*-body analysis (HLC) results as published by *Manna et al.*³⁵

Higher deviations from the reference values are observed in the case of Table 5.2, but once again, the CBS(ATZ,AQZ) extrapolated results yield more accurate results.

	CBS(QZ,5Z)	CBS(ATZ,AQZ)
dodecahedron	-0.2	0.2
face-cubes	0.8	0.5
face-sharing	0.7	0.8
Mean Absolute Error	0.6	0.5

Table 5.3: Comparison of the accuracy of Tight-LNO-CCSD(T)/CBS(QZ,5Z) and Tight-LNO-CCSD(T)/CBS(ATZ,AQZ) calculations for three different H₂O)₂₀ cluster geometries. The table shows errors of relative energies in kcal/mol units as compared to the inc-CCSD(F12*)(T)|MP2-F12/cc-pVQZ-F12 results, as published by *Anacker et al.*³⁶

The errors of the employed methods are closest when compared to the reference values of Table 5.3, but still, CBS(ATZ,AQZ) results lie closer to the reference data. Based on these results, we concluded that the Tight-LNO-CCSD(T)/CBS(ATZ,AQZ) results are the more reliable choice. Therefore, these will be set as the reference point for our in-house benchmarking of protocols.

Our results for the 20-species water clusters are summarized in the Figures 5.3, 5.4 and 5.4 below, in terms of (in)accuracy—throughput using different kind of error metrics. We have considered the Mean Absolute Error of Energies (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |E_i - E_i^{\text{ref}}|, \quad (5.1)$$

Root Mean Square Error of Energies (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N |E_i - E_i^{\text{ref}}|^2}, \quad (5.2)$$

and the Maximum Absolute Error of Energies (MAXAE):

$$\text{MAXAE} = \max(|E_i - E_i^{\text{ref}}|) \quad (5.3)$$

as statistical descriptors, with averages considered over the results obtained for the 16 different geometries (see Figure 5.2) used in our calculations. Instead of deriving relative energies from the energy difference of species i and the one corresponding to the minimum, we used the average of the absolute energies as a reference point, as these values are more robust, thus providing better grounds for comparing different methods.

All figures presented hereinafter use the same general layout and system of data visualization, which can be summarized as follows:

- Values are derived from the following calculations: Plain or "raw" LNO-CCSD(T)/(aug-)cc-pVXZ (top figure); LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation (middle figure); LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections (bottom figure). Markers of different shapes represent different basis sets, as indicated in the legend. Empty markers represent basis sets augmented with diffuse functions, i.e.: AXZ.
- Colors indicate different thresholds for local-correlation approximations: Lime—Loose, Cyan—Normal, Dark Magenta—Tight for LNO-CC methods, as implemented in the MRCC program system.
- All error metrics presented were derived with Tight-LNO-CCSD(T)-CBS(ATZ,AQZ) results as the reference.

Figure 5.3 uses Mean Absolute Error (MAE) values as our metric for accuracy. Here, we observe that the different thresholds applied for the local-correlation approximations (i.e., Loose, Normal, Tight) do not result in significant deviations in terms of accuracy for a given method and basis set combination, as the corresponding MAE values line up quasi-horizontally. This phenomenon may be explained away by the fact that even the lower-level (local-)MP2 calculations can capture the electron correlation of water reasonably well, being the small system it is. Thus, accuracy is less affected by increasing the number of pairs in a domain included in an LNO-CCSD(T) calculation. Relying on this observation, hereinafter, I will mainly consider results from a given method-basis set pairing group-wise, neglecting further analysis according to the local-correlation thresholds employed. Furthermore,

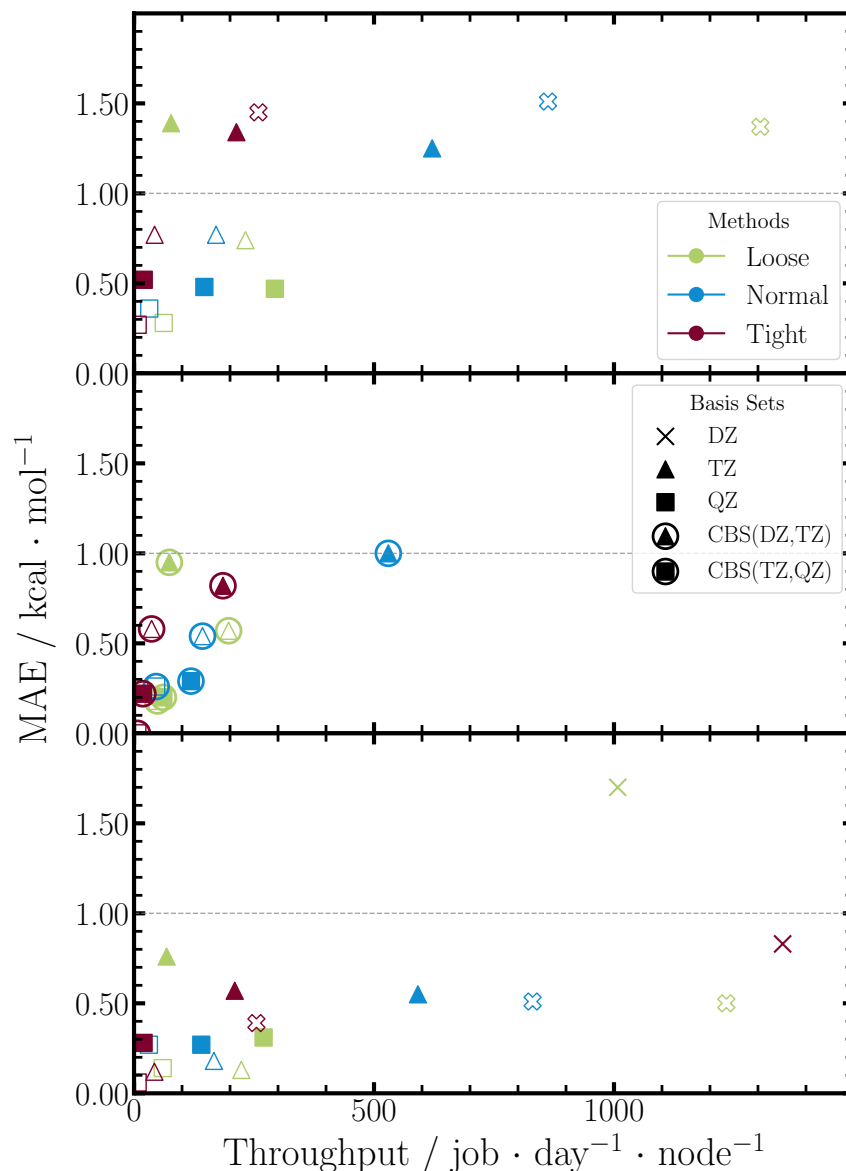


Figure 5.3: Comparison of the performance of different methods on the $(\text{H}_2\text{O})_{20}$ clusters in terms of throughput and accuracy as indicated by Mean Absolute Errors (MAE) of relative energies. Top figure: raw LNO-CCSD(T)/(aug-)cc-pVXZ (top; Middle figure: LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation based on raw results; Bottom figure: LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections. Results corresponding to the cc-pVDZ basis set are missing from the top figure, due to the cutoff at 2 kcal/mol as they display MAE values of circa ~ 3.6 kcal/mol (averaged over Loose, Normal and Tight LNO-CC results). In the bottom figure, Normal-LNO-CCSD(T)/cc-pVDZ values are victim of a cutoff of a different kind, one due to their exceptionally high throughput of 3214 jobs · day⁻¹ · node⁻¹.

based on this, one could argue that choosing a lower-level threshold may be desirable, as it leads to higher throughput. However, this comes at the cost of reducing the generality of our approach.

The tendency of increasing accuracy, with increasing cardinal number X in the (aug-)cc-pV X Z basis sets employed, shows clearly in the figure, with a ~ 2.3 , ~ 0.8 kcal/mol average decrease in MAE values for DZ $\rightarrow$ TZ, TZ $\rightarrow$ QZ respectively. Meanwhile, comparing the errors of the diffuse augmented and unaugmented version of a basis set with the same cardinal number, we observe a decrease with a factor of 2. Though the same general tendencies show for the values resulting from DBBSC calculations, these differences are far smaller, which is understandable, given the already low MAE values achievable with this method.

Looking at the values derived from extrapolation to the CBS limit, we see that the CBS(ADZ,ATZ) values are roughly on the same scale of accuracy as the non-augmented, raw QZ results. Similarly, the unaugmented CBS(TZ,QZ) values lead to errors similar to those of raw AQZ calculations. Meaning that a CBS($\{X-1\}$ Z,XZ) with non-augmented calculation results in similar accuracy as one using the raw AXZ basis set, though the latter comes at a price of lower throughput or extrapolation using diffuse-augmented basis CBS(A $\{X-1\}$ Z,AXZ), gives similar results as a unaugmented $\{X+1\}$ Z type calculation.

The bottom figure — summarizing the results achieved with the density-based corrections (DBBSC) — demonstrates it well, how powerful this method is. Generally, it can be stated that this correction reduces the mean absolute errors of a technique using a given basis set to approximately the same range as its uncorrected counterpart with the next higher cardinal number. The most prominent improvements can be observed for the smaller basis sets (DZ, TZ, ADZ), as larger ones usually already lead to quite small differences in energy. For example, it improves upon cc-pVDZ results almost with a factor of 3, lowering the Tight-LNO/DZ result to within the limit of chemical accuracy from its raw counterpart of 3.32 kcal/mol! It also brings results from both aug-cc-pVDZ and cc-pVTZ to well below that same 1 kcal/mol limit. A quite fascinating result is that comparably high accuracy may be achieved with the density-based corrections of the Tight-LNO/ADZ results and with either the DBBSC corrected QZ results or the raw AQZ, while achieving a significantly ($\sim 10\times$) higher throughput than those possible with the latter methods. In fact, while significantly reducing the errors displayed by the raw calculations, almost no noticeable increase in computational costs is observed. In another approach, we can compare the MAE values resulting from a CBS($\{X-1\}$ Z,XZ) extrapolation scheme to those resulting from a calculation with the density-based corrections of XZ and see that usually, it is the latter which leads to higher accuracy, while once again offering higher throughput.

Finally, we note the clear general trend in throughput displayed by the Loose, Normal, and Tight LNO-CC calculations, consistent with their increasing levels of approximation. However, we

must also highlight the anomalously low throughput of the Loose-LNO-CC/cc-pVDZ and Loose-LNO-CC/cc-pVTZ calculations, which are the only cases where this trend is broken. Quantifying this in the case of the former — as it is not visible in the figure — we see the following results: Loose—1007 jobs · day⁻¹ · node⁻¹, Normal—3214 jobs · day⁻¹ · node⁻¹, Tight—1351 jobs · day⁻¹ · node⁻¹.² In both the case of Loose-DZ and Loose-TZ calculations, the SCF convergence in the Hartree-Fock method was achieved exceptionally slowly, far exceeding the wall times measured for their Normal and Tight counterparts. These deviations are most likely an artifact of a hardware or software malfunction in the computational node where the calculations were performed.

The same general trends can be observed for the RMSE values as well, although these values display a slightly higher sensitivity for the Loose, Normal and Tight LNO-CC thresholds, especially in the case of triple- ζ basis functions (both diffuse augmented and unaugmented).

There is a noticeable systematic shift of errors toward higher values compared to those observed for the MAE. This shift arises from the sensitivity of the metric to outliers, as it involves squaring the errors.

We note here the surprisingly high accuracy achievable with the DBBSC inclusive ADZ calculations, which are — in the case of Normal and Loose thresholds — on par with the results of the raw cc-pVQZ or those of cc-pVTZ with the density-based corrections. Even higher accuracy may be achieved with the ADZ basis functions using the Tight settings, bringing the results close to the accuracy displayed by raw aug-cc-pVQZ(!) or DBBSC inclusive cc-pVQZ calculations. In fact the Loose-LNO-CC/cc-pVQZ results align very closely with those of Tight-LNO-CC/aug-cc-pVDZ, both in terms of accuracy (with MAE // RMSE values of: 0.31 kcal/mol // 0.51 kcal/mol and 0.39 kcal/mol // 0.54 kcal/mol respectively) and throughput (~270 jobs · day⁻¹ · node⁻¹ and 255 jobs · day⁻¹ · node⁻¹ respectively). However, given the inclusion of diffuse functions and the Loose threshold settings, the latter method likely displays a higher degree of robustness across different systems.

Although producing slightly less accurate results (MAE = 0.83 kcal/mol, RMSE = 1.26 kcal/mol) the set-up based on Tight-LNO-CC/cc-pVDZ results with the density-based corrections also seems to be a promising method, due to the exceptionally high throughput (1351 jobs · day⁻¹ · node⁻¹) associated with it, as well as the fact that when compared to the raw Tight-LNO-CC/cc-pVTZ results (the LNO-CC equivalent of the DLPNO-based approach used in the original CCMD publication¹), we still see a $\sim 2/3$ -fold reduction in the errors.

Moving on to the discussion of Maximum Absolute Errors we see that —in line with our expectations— there is a general shift to higher values when compared to the corresponding MAE and RMSE data.

²These values are derived from DBBSC inclusive LNO-CC calculations, but the same trend shows for their raw counterparts.

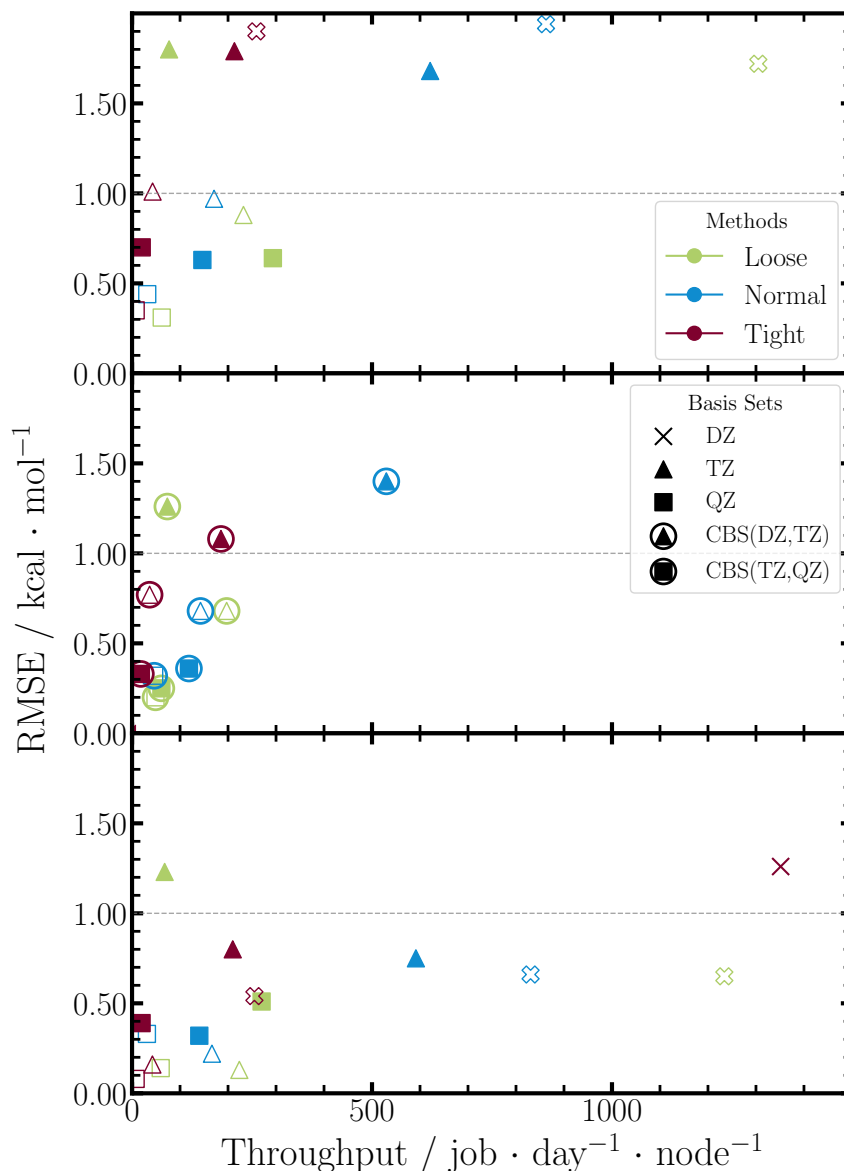


Figure 5.4: Comparison of the performance of different methods on the $(\text{H}_2\text{O})_{20}$ clusters in terms of throughput and accuracy as indicated by Root Mean Square Errors (RMSE) of relative energies. Top figure: raw LNO-CCSD(T)/(aug-)cc-pVXZ (top; Middle figure: LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation based on raw results; Bottom figure: LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections. Values corresponding to Loose, Normal, and Tight raw LNO-CC/cc-pVDZ results, as well as Loose and Normal LNO-CC/cc-pVDZ results, including density-based basis set corrections, are not shown due to their RMSE values exceeding the 2 kcal/mol limit employed.

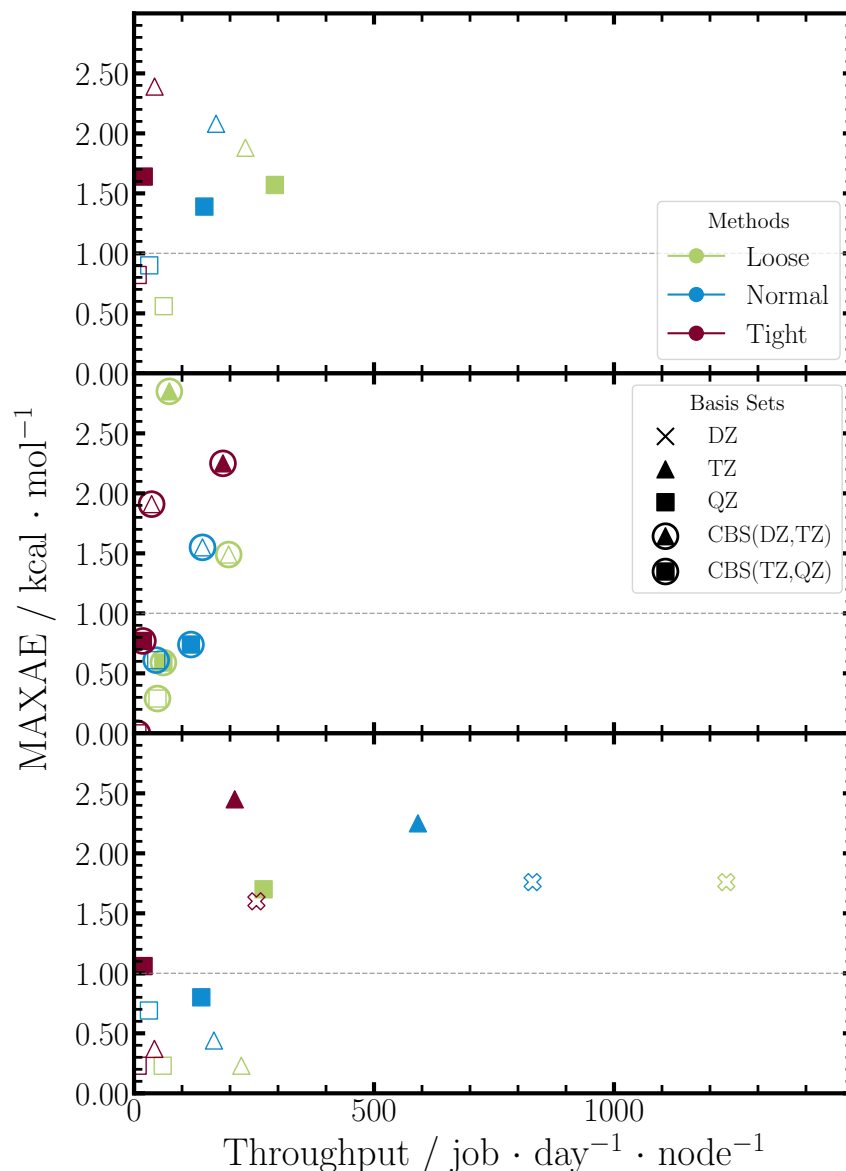


Figure 5.5: Comparison of the performance of different methods on the $(\text{H}_2\text{O})_{20}$ clusters in terms of throughput and accuracy as indicated by Maximum Absolute Errors (MAXAE) of relative energies. Top figure: raw LNO-CCSD(T)/(aug-)cc-pVXZ (top; Middle figure: LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation based on raw results; Bottom figure: LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections. Missing results (namely those corresponding to: Loose, Normal and Tight LNO-CC/cc-pVDZ — both top and bottom figures; Loose LNO-CC/cc-pVTZ — both top and bottom figure; Normal and Tight LNO-CC/cc-pVTZ — top figure; Normal LNO-CC/CBS(DZ,TZ) — middle figure) are not shown due to their MAXAE values exceeding the 3 kcal/mol cutoff limit employed.

A more interesting result is that when looking at the MAXAE data, we found that the previously observed near-horizontal alignment of Loose-, Normal- and Tight-LNO-CC results obtained within a given basis set no longer holds. Instead, the MAXAE values exhibit significantly greater sensitivity to the local-correlation thresholds. Contrary to initial expectations, another deviation from the pattern is observed—namely, the MAXAE values do not consistently decrease with increasing "tightness" of the threshold settings.

It generally holds true that MAXAE values lie about ~ 3 times higher than their MAE counterparts, with results corresponding to non-diffuse augmented basis sets displaying a factor closer to ~ 3.5 , while the augmented basis functions producing results usually lying about ~ 2.5 higher.

Furthermore, here we remark that although the CBS-extrapolated values tend to be more sensitive to the cardinal numbers of the basis sets used, results obtained via the DBBSC approach suggest that these calculations are less influenced by the overall basis set size and are instead more affected by the presence or absence of diffuse functions. This trend is observed across all error metrics considered but is most pronounced in the MAXAE values. Specifically, increasing the cardinal number improves accuracy on average by factors of 4 and 2 for the CBS extrapolation and DBBSC, respectively. Whereas, augmenting a basis set with diffuse functions leads to a 1.5-fold and 4-fold average increase, respectively.

We have previously noted the impressive results achievable with the Tight-LNO-CC/aug-cc-pVDZ calculations when employing the density-based corrections. It is worth mentioning that the robustness of this approach is further evidenced by the 1.60 kcal/mol max error corresponding to it, as this is still comparable to the errors of the model fitting, which is around 1.5 kcal/mol, while MAE and RMSE values fall well below this limit.

Although the same robustness is not observed for the density-corrected Tight-LNO-CC/cc-pVDZ results, even the maximal errors displayed (4.00 kcal/mol) using this approach can be considered to be within tolerance as they still fall below those observed for the raw Tight-LNO-CC/cc-pVTZ calculations (4.58 kcal/mol), which can be considered the equivalent of the method originally used in the CCMD scheme. Therefore, these two combination of methods seem to be promising candidates for chemically accurate high-throughput calculations.

5.2 Anionic cluster

We continued our investigations on a single chloride anion microsolvated by 20 water molecules. This system is more interesting from a quantum chemical perspective, as introducing a negatively charged species induces substantial perturbations in the electronic structure of the cluster and, therefore, seems like a great model system to further probe the effectiveness of the selected approaches.

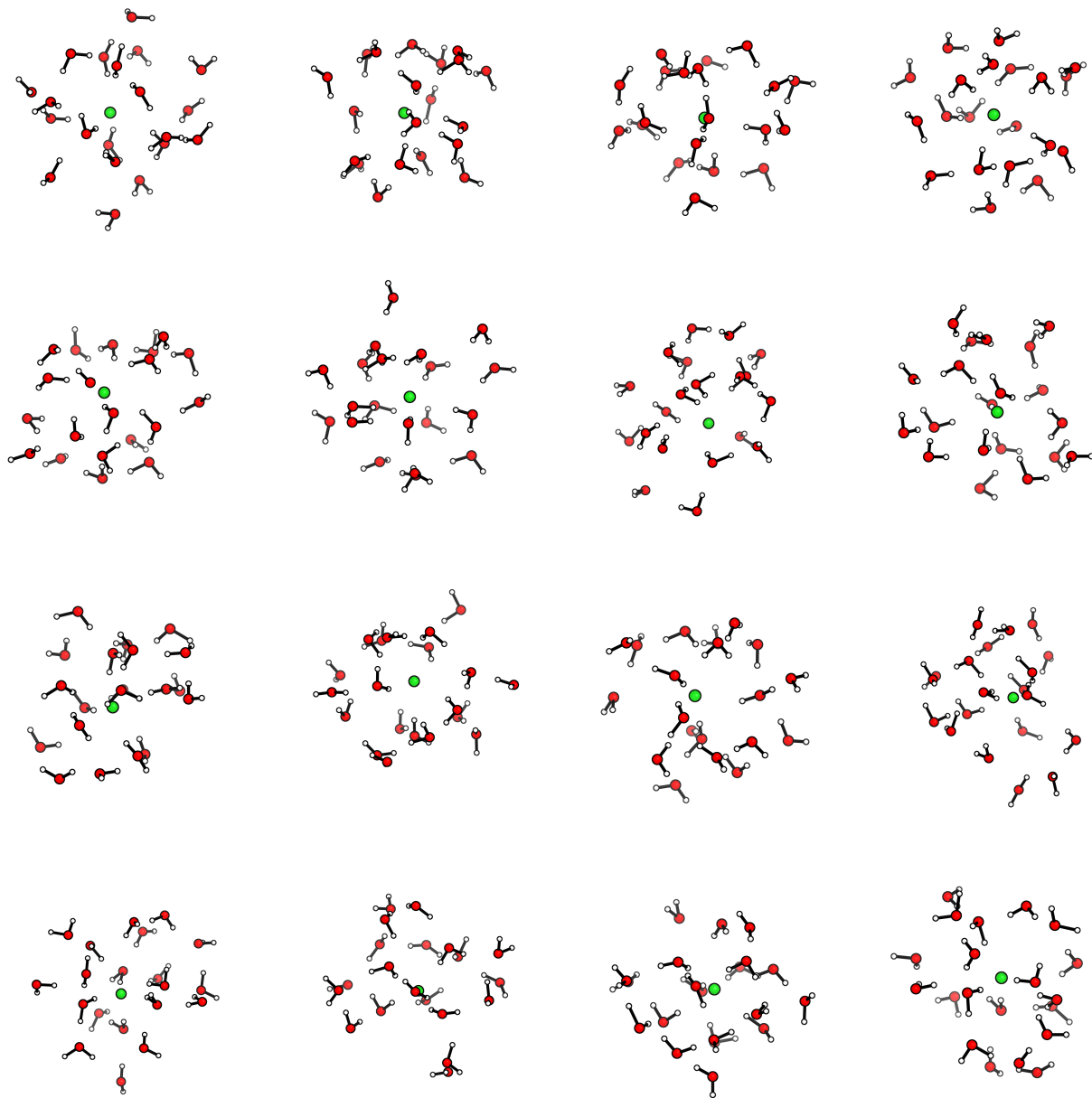


Figure 5.6: Geometries of $[\text{Cl}(\text{H}_2\text{O})_n]^-$ clusters used during our calculations

Accurate modeling of anionic systems requires careful treatment of electron correlation and basis set effects due to the diffuse nature of the wavefunction around the Cl^- . Moreover, this model system is

ideal for testing the local-correlation methods employed due to the long-range interactions characteristic of anionic systems.

We performed calculations using the same combinations of methods, basis sets, and corrections as for the water-only clusters and present the results below in Figures 5.7, 5.8 and 5.5. The exact layout and data visualization methods are employed on these figures as those detailed for the neutral clusters (see subsection 5.1.2).

Following the pattern established in the previous section, let us first discuss our results in terms of the Mean Absolute Errors determined for the anionic system. Looking at the corresponding data obtained via raw LNO-CC calculations—presented on Figure 5.7—we see that our results line up nicely with the well-known trend that by increasing the number of basis functions in a set, increasing accuracy may be achieved. I.e., both increasing the cardinal number or augmenting a set with diffuse functions yields more accurate results. Using diffuse augmented basis sets in a raw LNO-CC calculation produces results consistent with those seen for the water-only cluster, more specifically the diffuse augmented aug- XZ basis sets yields results of similar accuracy as their non-augmented $\{X+1\}Z$ counterpart. This trend shows clearly on the figure with the ATZ-QZ basis sets, but also holds for the ADZ-TZ pair as well, as the latter produce errors of 2.17, 2.28, 2.29 kcal/mol and 2.10, 2.19, 2.20 kcal/mol with Loose, Normal and Tight thresholds respectively. This indicates that while generally augmenting a basis set with a diffuse function improves upon the results, this is not due to the unique properties of the diffuse basis, but rather the increase in the number of functions included in the set. In other words our findings suggest that there is no significant dependence on incorporating diffuse functions in the basis set used during the raw calculations for the anionic clusters. This can be explained by the fact that our system models a condensed phase, and the chloride ion is in the center of the cluster, surrounded by the basis functions of the water molecules the contributions of which are enough to adequately describe the diffuse nature of the negative charge.

Considering the CBS extrapolated values, we see an increase in accuracy by about a factor 4 when switching over from $\text{CBS}(\{X-1\}Z, XZ)$ to $\text{CBS}(XZ, \{X+1\}Z)$ for basis sets both augmented and unaugmented with diffuse functions, whereas the augmentation itself (i.e., $\text{CBS}(\{X-1\}Z, XZ) \rightarrow \text{CBS}(A\{X-1\}Z, AXZ)$) leads to a 2.5-fold increase.

It seems to hold true for the $[\text{Cl}(\text{H}_2\text{O})_{20}]^-$ clusters too, that for a given basis set $(A)XZ$ the density-based corrections bring an increase in accuracy so that they are comparable to the raw $(A)\{X+1\}Z$ results, while simultaneously displaying 2-4 times higher throughput values. Such a comparison is harder to make against the values derived from CBS extrapolation, but it can be generally stated that for a basis set $(A)XZ$ comparable or higher accuracy can be achieved with the calculations incorporating density-based correction, than what is achievable with CBS extrapolation based on raw

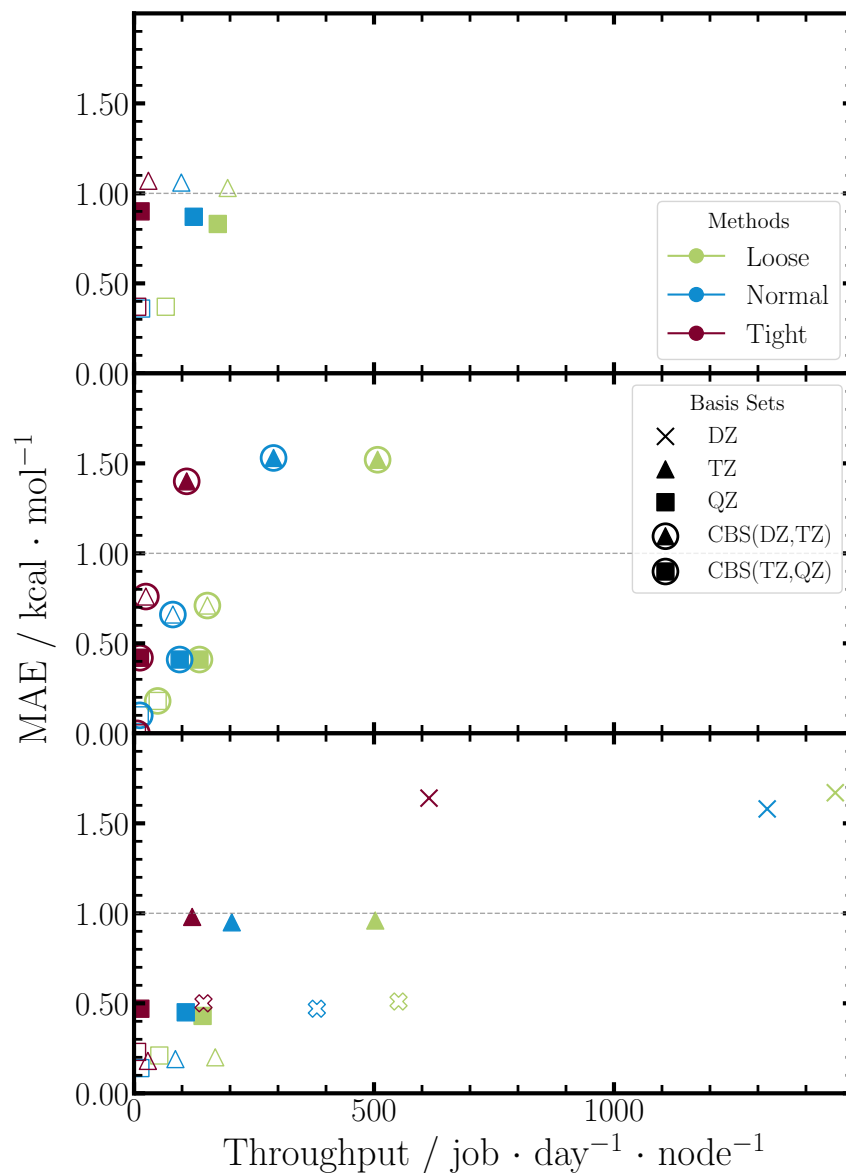


Figure 5.7: Comparison of the performance of different methods on the $[\text{Cl}(\text{H}_2\text{O})_{20}]^-$ clusters in terms of throughput and accuracy as indicated by Mean Absolute Errors (MAE) of relative energies. Top figure: raw LNO-CCSD(T)/(aug-)cc-pVXZ (top; Middle figure: LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation based on raw results; Bottom figure: LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections. Results derived from calculations using the cc-pVDZ basis set are missing from the top figure due to the cutoff at 2 kcal/mol.

(A){X-1}Z,(A)XZ results (while the respective throughput values are on par with one another).

While introducing the anionic clusters, we remarked that due to the long-range interactions characteristically present in systems with a negative electric charge, the compared performance of local-correlation methods with different threshold settings for the local approximations is of interest. However, here we see that similarly to the case of neutral clusters, the accuracy of the results does not significantly depend on the thresholds employed. This is best demonstrated in the bottom figure, presenting the results which include density-based corrections, but the statement holds generally true for the raw LNO-CC results as well.

In case of the RMSE values, just like before, we observe the same systematic shift to higher errors in RMSE values when compared to the corresponding MAE data points as seen for the water-only clusters, while all general trends discussed for the MAE values still hold true and therefore we move on to discuss the maximum errors.

Here, the otherwise strict trend of the near-horizontal alignment of results derived with the same type of calculations (excluding the local-correlation thresholds applied) is once again broken, with a surprising pattern that, in many cases, the Loose results yield the lowest errors. The aforementioned dissimilarity that CBS extrapolated values are more sensitive to the number of basis functions, while the accuracy of results derived using the density-based corrections is more heavily affected by the types of functions used (namely whether the set is augmented with diffuse functions or not) is also observed. Quantifying this, we see that increasing the cardinal number leads to an average of 3.5-fold (CBS) and 1.5-fold (DBBSC) improvement in terms of accuracy for the different methods of basis set corrections, respectively. In contrast, we found a 2-fold (CBS) and 3.5-fold (DBBSC) average increase in accuracy when comparing results derived using diffuse unaugmented and augmented versions of basis sets with the same cardinal number.

Finally, let us discuss the results of our selected approaches of Tight-LNO-CC/cc-pVDZ and Tight-LNO-CC/aug-cc-pVDZ with density-based corrections for the anionic system. While neither method seems to be an exception from the general trend of shifting to higher errors mentioned in the overall review of our results for the chloride-containing clusters, this shift seems to affect the set-up involving the unaugmented basis set more heavily, as a shift from a MAE value of 0.83 kcal/mol to 1.64 kcal/mol is observed, while for the scheme employing the basis set with diffuse functions, we do not see this doubling of MAE values (0.39 kcal/mol→0.5 kcal/mol). This is in line with our observation that the density-based corrections are more sensitive to including diffuse functions. Concerning the maximal errors corresponding to these approaches, we once again find that the results derived with the cc-pVDZ basis (4.53 kcal/mol) still fall within tolerance when considering that they lie below those observed for the raw Tight-LNO-CC/cc-pVTZ calculations (4.99 kcal/mol) —which is taken as an estimate of the

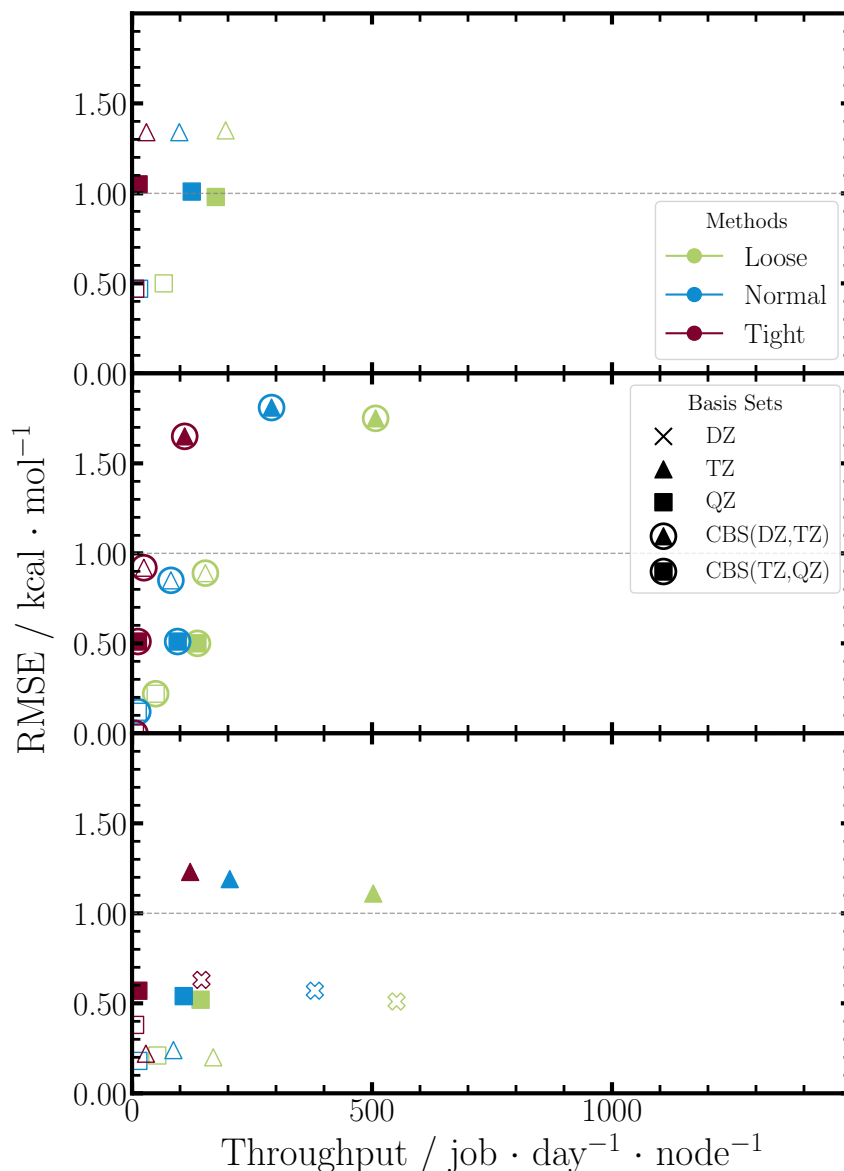


Figure 5.8: Comparison of the performance of different methods on the $[\text{Cl}(\text{H}_2\text{O})_{20}]^-$ clusters in terms of throughput and accuracy as indicated by Root Mean Square Errors (RMSE) of relative energies. Top figure: raw LNO-CCSD(T)/(aug-)cc-pVXZ (top; Middle figure: LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation based on raw results; Bottom figure: LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections. Values corresponding to the raw Loose-, Normal- and Tight-LNO-CC/(aug-)cc-pVDZ calculations exceed the 2 kcal/mol cutoff limit employed here, and the same holds for the values derived with the inclusion of DBBSC corrections for the cc-pVDZ basis set at all levels of local-approximation.

DLPNO-based approach employed in the original CCMD scheme. Whereas the protocol employing the aug-cc-pVDZ led to a MAXAE value of only 1.24 kcal/mol, which further supports the robustness of this procedure, as this is safely below the 1.5 kcal/mol error bar of the fitting model.

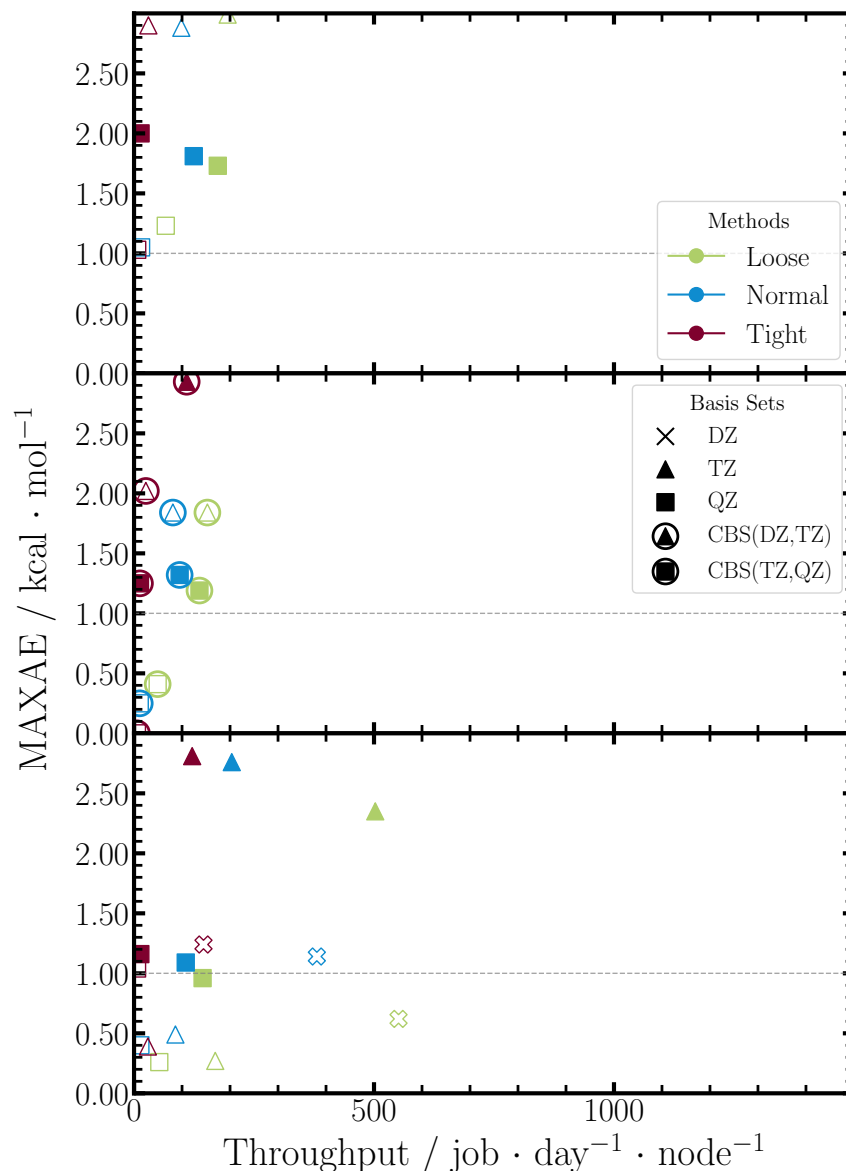


Figure 5.9: Comparison of the performance of different methods on the $[\text{Cl}(\text{H}_2\text{O})_{20}]^-$ clusters in terms of throughput and accuracy as indicated by Maximum Absolute Errors (MAXAE) of relative energies. Top figure: raw LNO-CCSD(T)/(aug-)cc-pVXZ (top; Middle figure: LNO-CCSD(T) with CBS(A/{X-1}Z,XZ) extrapolation based on raw results; Bottom figure: LNO-CCSD(T)/(aug-)cc-pVXZ with DBBSC corrections. Missing results, (namely those corresponding to: Loose, Normal and Tight LNO-CC/cc-pVDZ — both top and bottom figures; Loose, Normal and Tight LNO-CC/aug-cc-pVTZ — top figure; Loose, Normal and Tight LNO-CC/cc-pVTZ — top figure; Loose and Normal LNO-CC/CBS(DZ,TZ) — middle figure) are not shown due to their MAXAE values exceeding the 3 kcal/mol cutoff limit employed.

Chapter 6

Conclusion

We set out to explore possible ways to accelerate the reference cluster calculations to be used in the Coupled Cluster Molecular Dynamics framework by investigating a number of different combinations of local approximations, basis sets, and corrections thereof, including the novel and highly effective Density Based Basis Set Corrections. For this, we established the speed-up achievable by reducing the system size from 64 to 20 molecules of water and determined the optimal levels of parallelity.

Finally, by analyzing the cost-accuracy compromise of the different set-ups, we have identified two promising protocols: One relying on LNO-CC/aug-cc-pVDZ calculations with Tight local-correlation thresholds and employing the density-based corrections which lead to a ~ 70 -fold increase in throughput values, and another also relying on the CABS+DBBSC corrections, to the Tight-LNO-CC/cc-pVDZ results, leading to a very impressive ~ 340 -fold speed-up. The choice between these protocols can be facilitated by considering the desired accuracy, as results with the cc-pVDZ basis with CABS and DBBSC correction yield error comparable or lower as the raw triple- ζ used in the original framework, while the use of the aug-cc-pVDZ basis set decreases these errors by a factor of 3. The robustness of these protocols has been tested by also performing calculations on a model system consisting of 20 molecules of water and a chloride ion, for which the suggested set-ups yielded similar results both in terms of throughput (40- and 150-fold speed-up) and accuracy (1.4- and 4-fold increase).

In summary, we can conclude that using these protocols, the reference clusters calculations of the CCMD scheme can be performed in 1-4 node-days, as opposed to the $\sim 6 \times 10^5$ node-days of the original work.

Acknowledgements

I would like to thank my supervisor, János Daru, for his help and guidance throughout this project. Furthermore, I thank fellow group members András Szabó and Bence Balázs Mészáros for sharing their results on the convergence of Δ -Learning for water clusters. Péter R. Nagy is gratefully acknowledged for fruitful discussions and technical support.

The project was supported financially by the National Research, Development, and Innovation Office (NKFIH, Grant No. FK147031). We also thank KIFŰ (Governmental Agency for IT Development, Hungary, <https://ror.org/01s0v4q65>) for awarding us access to the Komondor HPC facility based in Hungary.

References

- [1] J. Daru, H. Forbert, J. Behler and D. Marx, *Physical Review Letters*, 2022, **129**, 226001.
- [2] N. Stolte, J. Daru, H. Forbert, J. Behler and D. Marx, *arXiv preprint arXiv:2410.17675*, 2024.
- [3] J. Sun, A. Ruzsinszky and J. P. Perdew, *Physical Review Letters*, 2015, **115**, 036402.
- [4] S. Dasgupta, E. Lambros, J. Perdew and F. Paesani, *Nature Communications*, 2021, **12**, 6359.
- [5] F. Paesani, *Accounts of Chemical Research*, 2016, **49**, 1844–1851.
- [6] M. Born and J. R. Oppenheimer, *Annalen der Physik*, 1927, **84**, 457–484.
- [7] D. R. Hartree, Mathematical Proceedings of the Cambridge Philosophical Society, 1928, pp. 89–110.
- [8] V. Fock, *Zeitschrift für Physik*, 1930, **61**, 126–148.
- [9] C. C. J. Roothaan, *Reviews of modern physics*, 1951, **23**, 69.
- [10] A. Szabo and N. S. Ostlund, *Modern quantum chemistry: introduction to advanced electronic structure theory*, Courier Corporation, 1996.
- [11] C. Møller and M. S. Plesset, *Physical review*, 1934, **46**, 618.
- [12] J. Čížek, *The Journal of Chemical Physics*, 1966, **45**, 4256–4266.
- [13] J. F. Stanton, *Chemical Physics Letters*, 1997, **281**, 130–134.
- [14] S. A. Kucharski and R. J. Bartlett, *The Journal of chemical physics*, 1998, **108**, 5243–5254.
- [15] K. Raghavachari, G. W. Trucks, J. A. Pople and M. Head-Gordon, *Chemical Physics Letters*, 1989, **157**, 479–483.
- [16] C. Riplinger and F. Neese, *The Journal of Chemical Physics*, 2013, **138**, 034106.

- [17] C. Riplinger, B. Sandhoefer, A. Hansen and F. Neese, *The Journal of Chemical Physics*, 2013, **139**, 134101.
- [18] P. R. Nagy and M. Kállay, 2017, **146**, 214106.
- [19] P. R. Nagy, G. Samu and M. Kállay, 2018, **14**, 4193.
- [20] P. R. Nagy and M. Kállay, 2019, **15**, 5275.
- [21] P. R. Nagy, *Chemical Science*, 2024, **15**, 14556–14584.
- [22] P. R. Nagy, G. Samu and M. Kállay, 2016, **12**, 4897.
- [23] J. L. Whitten, *The Journal of Chemical Physics*, 1973, **58**, 4496–4501.
- [24] M. Feyereisen, G. Fitzgerald and A. Komornicki, *Chemical physics letters*, 1993, **208**, 359–363.
- [25] A. Karton and J. M. Martin, *Theoretical Chemistry Accounts*, 2006, **115**, 330–333.
- [26] F. Neese and E. F. Valeev, *Journal of Chemical Theory and Computation*, 2011, **7**, 33–43.
- [27] T. Helgaker, W. Klopper, H. Koch and J. Noga, *The Journal of Chemical Physics*, 1997, **106**, 9639–9646.
- [28] E. Giner, B. Pradines, A. Ferté, R. Assaraf, A. Savin and J. Toulouse, *The Journal of Chemical Physics*, 2018, **149**, 194301.
- [29] P.-F. Loos, B. Pradines, A. Scemama, J. Toulouse and E. Giner, *The Journal of Physical Chemistry Letters*, 2019, **10**, 2931–2937.
- [30] D. Mester, P. R. Nagy and M. Kállay, *Journal of Chemical Theory and Computation*, 2024, **20**, 7453–7468.
- [31] J. P. Perdew, K. Burke and M. Ernzerhof, *Physical review letters*, 1996, **77**, 3865.
- [32] M. Kállay, P. R. Nagy, D. Mester, Z. Rolik, G. Samu, J. Csontos, J. Csóka, P. B. Szabó, L. Gyevi-Nagy, B. Hégyel, I. n Ladjánszki, L. Szegedy, B. Ladóczki, K. Petrov, M. Farkas, P. D. Mezei and Á. Ganyecz, 2020, **152**, 074107.
- [33] M. Kállay, P. R. Nagy, D. Mester, L. Gyevi-Nagy, J. Csóka, P. B. Szabó, Z. Rolik, G. Samu, J. Csontos, B. Hégyel, Á. Ganyecz, I. Ladjánszki, L. Szegedy, B. Ladóczki, K. Petrov, M. Farkas, P. D. Mezei and R. A. Horváth, MRCC, *a quantum chemical program suite*. See <https://www.mrcc.hu/>, Accessed August 1, 2023.

- [34] D. Traore, J. Toulouse and E. Giner, *Faraday Discussions*, 2024.
- [35] D. Manna, M. K. Kesharwani, N. Sylvetsky and J. M. Martin, *Journal of chemical theory and computation*, 2017, **13**, 3136–3152.
- [36] T. Anacker and J. Friedrich, *Journal of computational chemistry*, 2014, **35**, 634–643.