Data Driven Acceleration of Coupled-Cluster Calculations using Machine Learning, Multitask Learning and Physics Imposed Learning

A Dissertation Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Perera Don Varuna Sanjaya Pathirage
August 2024

Acknowledgements

I would like to express my deepest gratitude to my mentor, Dr. Konstantinos Vogiatzis, for his unwavering guidance and support throughout the years. His mentorship has been instrumental in shaping my career as a scientist.

I am also profoundly grateful to my committee members, Dr. Sharani Roy, Dr. Michael Best, Dr. Cristian Batista, and Dr. Vasileios Maroulas for their invaluable advice and encouragement.

A heartfelt thanks goes to all past and present group members, especially Dr. Jacob Townsend, Dr. Grier Jones, and Justin Phillips, whose support has been crucial to my accomplishments.

I extend my gratitude to the Sri Lankan community in Knoxville for their friendship and support. Also, I would like to thank my friends Samuditha, Nishadi, Navindra, Manoj, Ravindika, Vimukthi and Thushari for their unwavering encouragement and companionship throughout my journey.

To my wife Jayani, parents, brother Supun and his wife Narendra, your steadfast love and encouragement have been my pillars of strength and success. Thank you for standing by me through thick and thin.

Finally, I would like to express my heartfelt appreciation to the people of Sri Lanka for providing me with 17 years of free education. This invaluable gift has given me access to facilities and infrastructure that I would not have had otherwise.

Abstract

The data-driven coupled-cluster singles and doubles (DDCCSD) method aims at predicting the coupled-cluster t_2 amplitudes using MP2-level electronic structure data with machine learning. In this work we address limitations of the DDCCSD method to expand its applicability and increase its accuracy. First, we implemented localized molecular orbitals (LMO) to the DDCCSD scheme, and we observed a ten-fold increase in accuracy when LMOs are used compared to the canonical molecular orbital implementation. Next, we introduced five data selection techniques to select data for testing and training. Here we were able to achieve accuracies less than the full model that utilizes all amplitudes for training using only 10-15% of total datapoints. Then, we developed three versions of DDCC variants based on neural network architectures that introduce a variety of advantages including multitask learning, and more portable models. With these models, we were able to predict both t_1 and t_2 amplitudes simultaneously. This coupled with physics-imposed learning allowed us to achieve greater accuracies than the conventional DDCCSD methods and due to the architecture of the models we were able to expand the accuracy towards larger chemical systems. Finally, we showed that DDCC can be further expanded by introducing a data-driven formalism to the perturbative triples (DDCCSD(T)) and the exact triples (DDCCSDT).

Table of Contents

List of Tables

List of Figures

Chapter 1 Machine Learning in Quantum Chemistry

Atoms, the fundamental units of chemical structures, consist of subatomic particles such as electrons, protons, and neutrons. Due to the size of these subatomic particles, these atomic systems cannot be described precisely using classical mechanics. Quantum chemistry is the application of quantum mechanics to understand the behavior of chemical systems. Walter Heitler and Fritz London became the first scientists who applied quantum mechanics to study chemical systems when they solved the Schrödinger equation for the hydrogen atom in late 1920's.[1] This can be considered as the origin of the field of quantum chemistry. With the invention of computers, quantum chemistry gained much more power and attention since computers enhanced our ability to do complex calculations and therefore, we were able to expand the applicability of quantum chemistry.

Although we are able to solve the Schrödinger equation for simpler systems, we are still unable to find exact solutions to the many-body electron problem. Early approaches focused on solving this issue using semi-empirical methods. With the contribution from Douglas Hartree and V. A. Fock, a new theory was developed in early 1930's that can be used to solve the time-independent Schrödinger equation with Born-Oppenheimer approximation. Today, this method is termed as Hartree-Fock (HF) theory. This theory aims at capturing the electron-electron interaction in a mean field manner.

Hartree-Fock theory is fundamentally flawed since it is a mean field approximation and consequently, it cannot capture all electron-electron interactions. These missing interactions are collectively called correlation energy and they are typically grouped into two categories. The first type is the dynamic correlation that arises due to the sudden movement of electrons. The second type of correlation energy is static correlation that arises due to the degeneracies or near-degeneracies of the electronic states. Methods such as configuration interaction (CI), perturbation theory and coupled-cluster (CC) theory aim to capture the correlation energy and

were developed based on HF theory. These methods are collectively known as post-HF methods. Although these methods give more accurate description to chemical systems, the computational expense restricts the applicability of these methods to larger chemical systems. A variety of methods were developed over the years aiming the acceleration of the post-HF methods such as resolution of the identity,[2] local correlation methods,[3-9] accelerated convergence techniques,[10, 11] explicitly-correlated methods,[12, 13] tensor decomposition,[14, 15] dual-basis methods,[16] parallel computing with GPU acceleration,[17, 18] and machine learning.[19-22]

Due to the developments in the computational infrastructure, machine learning is gaining more popularity in the field of computational chemistry. According to a recent classification by Aldossary et al,[23] there are four types of machine learning applications in quantum chemistry based on the amount of physics included in the models. The first type of machine learning models learns purely on computational data for training (e.g. from DFT or semi-empirical methods). These models rely on training with large training datasets and typically aim on the fast and reliable statistical analysis of large libraries that can lead to the discovery of new molecules or materials.[24-28] The physics of the problem under consideration are not explicitly introduced, but the model eventually will learn them once it is trained on adequate and diverse data. The second type is machine learning potentials and force fields.[22, 29-31] These models learn to map interatomic interactions and to generate potential energy surfaces, and the goal is to reach the accuracy of *ab initio* methods with low-level computational data. The next type of machine learning models applies machine learning to enhance quantum chemical calculations.[32-39] These models use machine learning to improve or accelerate quantum chemical calculations either by predicting different quantum chemical intermediates or internal parameters and then performing quantum chemical calculations to predict properties. The last

type of models are machine learning *ab initio* methods.[40, 41] These methods use machine learning to directly build the wave function.

The focus of this dissertation is on the extension of the capabilities and accuracy of the data-driven coupled-cluster singles and doubles (DDCCSD) methods developed by Townsend and Vogiatzis[37] which aims at predicting the two-electron excitation amplitudes (t_2) for coupled-cluster singles and doubles (CCSD) using MP2-level electronic structure data and machine learning. These predicted t_2 amplitudes can be used in two ways. First use is to calculate the correlation energy directly using these amplitudes. By doing this we can obtain a better approximation to CCSD energy than MP2 energy at the expense of MP2. The second application of the machine learned amplitudes is as initial guesses to the iterative CCSD solver to determine the exact amplitudes. By doing this we can reduce the number of iterations to obtain the exact CCSD amplitudes. The first approach where the predicted amplitudes are being used to calculate the energy directly can be categorized under the third type of machine learning methods discussed by Aldossary et al.[23] since there we obtain a wave function with the DDCCSD t_2 amplitudes. The second approach can be categorized under the four types of the above classification since the predicted amplitudes are being used to accelerate the calculation of exact t_2 amplitudes and thus, recover the exact coupled-cluster wave function.

Although the earlier implementation of DDCCSD served as an excellent proof-of-concept especially for small molecules, there were limitations hindering the applicability of the model to large chemical systems, as discussed in Chapter 2. Such limitations involve the utilization of imbalanced data, small basis sets, and application of instance-based learning. In this project, we have addressed these limitations and we have elevated the applicability of DDCC. For example, electron excitations are local in nature. In Chapter 3 we discuss the introduction of

Boys localized molecular orbitals to DDCC. We have shown that with the implementation of locality to our DDCCSD models we were able to achieve up to a ten-fold increase in accuracy.

Main factor that hinders the applicability to large molecular systems is the number of amplitudes and their distribution. The number of amplitudes increases exponentially with the number of atoms and the complexity of the basis set. This will cause the computational expense in training the models to increase and reduce the portability of the models. As a solution we introduced five data selection techniques. We were able to achieve more accuracy than all-amplitude model with 10 – 15% of the total number of amplitudes using these schemes. These schemes will be discussed in detail in Chapter 3.[39]

Although we predict the t_2 amplitudes, in the previous models, we do not predict the t_1 amplitudes in our previous study. In the current work we discuss the development of **D**ata-**D**riven **C**oupled-**C**luster Neural **Net**works (DDCCNet) a multitask learning approach to DDCCSD where we predict t_1 and t_2 amplitudes simultaneously. Apart from predicting the t_1 amplitudes, DDCCNet also improves the portability of the DDCCSD models. Due to the flexibility of the DDCCSD models in terms of the architecture and loss functions, we were able to improve the accuracy of the models further by incorporating physics-informed learning (Chapter 4).

Finally, we present the implementation of perturbative triples correction to the DDCCSD models in Chapter 5. This is important since this allows us to get a reasonable prediction for the CCSD(T) energy which is the gold standard for quantum chemistry.

Chapter 2 Data-driven acceleration of coupled-cluster methods

2.1 Disclosure

Sections 2.2, 2.3, and 2.4 contain an adapted reproduction of the published book chapter with title "*Data-driven acceleration of coupled-cluster and perturbation theory methods*" by Grier M. Jones, P. D. Varuna S. Pathirage, and Konstantinos D. Vogiatzis that is part of the book *Quantum Chemistry in the Age of Machine Learning*.[42] P. D. Varuna S. Pathirage discussed the data-driven coupled cluster methods and Grier M. Jones discussed data-driven multiconfigurational methods under the guidance of Professor Konstantinos D. Vogiatzis.

2.2 Abstract

Molecular electronic structure theory refers to the study of the electron motion in molecules and it has become an important computational tool in chemical research. An exact solution to the full electronic Hamiltonian in the Born-Oppenheimer approximation can be obtained only for hydrogenic atoms or molecular species with one electron. A variety of powerful approximate theories (configuration interaction theory, perturbation theory, and coupled-cluster theory) have been developed since the dawn of quantum mechanics. These wave function-based methods are often described as post-Hartree-Fock methods since a Hartree-Fock wave function is typically used as their starting point. However, significant computational resources are required for reaching the expected high accuracy due to exponential scaling with respect to the system size. Machine learning has been utilized as a powerful alternative for the acceleration of the convergence of the underlying equations of post-Hartree-Fock methods. Here, the recent advances in the acceleration of the coupled-cluster iterative solver through machine learning are presented. Wave function properties obtained from computationally less demanding calculations are used as input for machine learning. The data-driven coupled-cluster scheme provides the exact coupled-cluster energy because the actual iterative coupled-cluster

equations are solved. Alternatively, accurate energetics can also be obtained by bypassing the explicit solution of the coupled-cluster projected equations. Our results show that remarkable speedups are achieved, especially when the physics of the electron correlation are encoded and used for training ML models.

2.3 Introduction

Quantum chemistry applies quantum mechanics for the theoretical description of chemical systems. The primary aim is the solution of the Schrödinger equation for atoms and molecules, which gives access to their energies, geometries, and physicochemical properties. The Born-Oppenheimer approximation decouples the electronic component from the nuclear component of the wave functions and consequently, allows the separate treatment of the electrons. Molecular electronic structure theory describes electrons in atoms and molecules, but the correlated motion of electrons introduces an intractable many-body problem that requires approximate numerical solutions. The typical starting point of the electronic structure theory is Hartree-Fock (HF) theory. In HF theory, the electronic wave function is expressed as a Slater determinant that includes the Fermi correlation, that is, correlation that originates from the spin and permutation symmetry conditions of the wave function. Coulombic correlation is the second type of electron correlation that stems from Coulomb repulsion between electrons and is accounted in HF theory, in an average, mean-field manner. Although HF theory accounts for almost 99% of the total electronic energy, the missing 1% becomes very important for the accurate computation of chemical reactivity and chemical properties. Post-HF theories provide a variety of systematic methodologies toward the exact Born-Oppeinheimer, nonrelativistic electronic energies. These are the configuration interaction (CI) theory, many-body perturbation theory, and coupled-cluster (CC) theory. For a more detailed description of these methods, we refer the reader to a recent overview.[43]

Machine learning (ML), a branch of artificial intelligence, involves the development of computational models based on input sample data (also known as "training" data) to make predictions or decisions without being explicitly programmed to do so. The strength of ML lies in its ability to provide fitting functions that can accurately infer complicated, highly nonlinear dependencies and provide accurate predictions. ML has become a revolutionary tool in many scientific disciplines due to the development of efficient algorithms, the availability of easy-to-apply software, and the improved hardware infrastructure. Domain experts (e.g., computational chemists, organic chemists, material scientists) can routinely apply ML for many challenging applications. A desirable goal of theoretical and computational chemistry is to generate ML potentials applicable to molecules of various sizes and obtain accurate geometries and gradients at force-field cost. A successful approach that enhances transferability between different molecular systems for ML is the incorporation of a physics-based representation of the chemical compounds and materials. ML models that target a global, intrinsic molecular property such as wave functions and electron densities as the learnable output have been recently explored, where the latter has led to the development of new density functionals. Increased transferability is achieved when an ML model learns the wave function since its physical properties are well-defined and analogous between similar molecules. Early efforts involved the direct solution of the electronic or vibrational Schrödinger equation with (mainly) neural networks.[41, 44-46] More recently, reinforcement learning has been used for the electronic many-body problem of molecular systems.[47, 48] An alternative route targets the training of supervised ML models on quantum chemical data (input) for the prediction of energy terms or the parameterization of complex wave functions (output). For example, combination of neural networks with a wave function representation based on atomic basis sets provided milliHartree accuracy when solving the Schrödinger equation.[35]

CC theory has also been the target of hybrid ML/quantum chemical approaches. McGibbon et al. developed a neural network model for improving the accuracy of second-order Møller-Plesset perturbation theory (MP2) with respect to accurate CCSD(T) results[49], while the Miller group used data from HF wave functions to develop a ML model based on Gaussian process regression for CCSD and CCSD(T) correlation energies.[32, 50, 51] Similarly, Margraf and Reuter used a representation based on the MP2 wave function (the MP2 amplitudes) for learning CC energies.[52] More recent works explored HF orbitals[53], one-electron reduced-density matrices (1-RDM)[54], and CCSD(T) densities[55] as input to ML. A common characteristic of these studies is that the learnable output is the correlation energy and thus, they provide a prediction of a particular energy value with some uncertainty.

We have examined a different approach to this problem, which learns the CC wave function, rather than the CC energy.[37] This route has very attractive advantages. First, a predicted wave function can be used for restarting a CC calculation to obtain the exact CC energy. Second, apart from electronic energy, molecular properties such as dipole moments and polarizabilities can be computed from the ML-trained wave function. Third, the CC wave function shares many similarities between different molecular species, which makes the new method transferable between similar molecules or clusters. In addition, strongly correlated methods such as the complete-active-space self-consistent field (CASSCF) and its extension via second-order perturbation theory (CASPT2) have been recently coupled with ML. The development and strengths of these data-driven approaches for quantum chemistry are discussed in this Chapter.

2.4 Data-driven coupled cluster

2.4.1 Basic theory

CC theory aims to capture the missing dynamic correlation from HF theory by considering electronic excitations to virtual spin-orbitals (unoccupied orbitals) from occupied spin-orbitals, as in the case of CI theory. The difference between CC and CI is that CI theory uses a linear parameterization of the wave function as shown in equation 2.1 whereas CC theory uses separable, nonlinear parameterization as shown equation 2.2.

$$|\Psi_{CI}\rangle = c_0|\Psi_0\rangle + \sum_{I,A} c_I^A \hat{\tau}_I^A |\Psi_0\rangle + \sum_{\substack{I>J \\ A>B}} c_{IJ}^{AB} \hat{\tau}_{IJ}^{AB} |\Psi_0\rangle + \sum_{\substack{I>J>K \\ A>B>C}} c_{IJK}^{ABC} \hat{\tau}_{IJK}^{ABC} |\Psi_0\rangle + \cdots \quad 2.1$$

$$|\Psi_{CC}\rangle = \left[\prod_{I,A}(1 + t_I^A \hat{\tau}_I^A)\right]\left[\prod_{\substack{I>J \\ A>B}}(1 + t_{IJ}^{AB} \hat{\tau}_{IJ}^{AB})\right]\left[\prod_{\substack{I>J>K \\ A>B>C}}(1 + t_{IJK}^{ABC} \hat{\tau}_{IJK}^{ABC})\right] \ldots |\Psi_0\rangle \quad 2.2$$

where c_I^A, c_{IJ}^{AB}, and c_{IJK}^{ABC} are the CI expansion coefficients for single, double, and triple electron excitations from the zeroth-order HF wave function Ψ_0, respectively, and t_I^A, t_{IJ}^{AB}, and t_{IJK}^{ABC} are the excitation amplitudes of the CC expansion. The excitation operators $\hat{\tau}_I^A$, $\hat{\tau}_{IJ}^{AB}$, etc. are promoting an electron from an occupied spin-orbital *I,J,K,*…of the reference wave function Ψ_0 to an unoccupied spin-orbital *A,B,C,*…. In the second-quantization formalism, the excitation operators are expressed as,

$$\hat{\tau}_I^A = a_A^\dagger a_I \quad 2.3$$

$$\hat{\tau}_{IJ}^{AB} = a_A^\dagger a_B^\dagger a_J a_I \tag{2.4}$$

$$\hat{\tau}_{IJK}^{ABC} = a_A^\dagger a_B^\dagger a_C^\dagger a_K a_J a_I \tag{2.5}$$

where $a_P^\dagger$ and a_Q are the electron creation and annihilation operators, respectively. These operators are adding and removing an electron, respectively, from an occupation-number vector that represents the zeroth-order wave function. The molecular electronic, Born-Oppenheimer, nonrelativistic Hamiltonian operator in the second-quantization formalism is given as

$$\widehat{H} = \sum_{PQ} h_{PQ} a_P^\dagger a_Q + \frac{1}{2} \sum_{PQRS} g_{PQRS} a_P^\dagger a_Q^\dagger a_S a_R \tag{2.6}$$

where indices *P, Q, R, S* represent arbitrary spin-orbitals. The h_{PQ} and g_{PQRS} terms of equation 2.6 are the one- and two-electron Hamiltonian terms, respectively, and h_{nuc} is the nuclear repulsion energy.

The exact, nonrelativistic, Born-Oppenheimer electronic energy of an atom or molecule is computed by solving the electronic Schrödinger equation with the Hamiltonian of equation 2.6, in the full excitation manifold of the CI expansion (equation 2.1) and a complete basis set (CBS) expansion. This corresponds to the full CI (FCI) method at the CBS that is only applicable for reference calculations of atoms and small molecules. For practical applications, a truncated CI expansion is considered. Similarly, the full CC expansion of equation 2.2 is equivalent to FCI, but for practical applications, a truncated Ansatz is applied. For example, a CC expansion that includes only two-electron excitations corresponds to coupled-cluster

doubles (CCD), if both one- and two-electron excitations are considered then we refer to coupled-cluster singles and doubles (CCSD), and so on.

For the derivation of the CC energy expression, it is easier to work with an exponential parameterization of equation 2.2:

$$|\Psi_{CC}\rangle = \exp\left(\sum_{I,A} t_I^A \hat{\tau}_I^A + \sum_{I>J,A>B} t_{IJ}^{AB} \hat{\tau}_{IJ}^{AB} + \cdots\right)|\Psi_0\rangle \qquad 2.7$$

This can be simplified using the cluster operator $\hat{T}$:

$$|\Psi_{CC}\rangle = \exp(\hat{T})\,|\Psi_0\rangle = \exp(\hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \cdots)\,|\Psi_0\rangle \qquad 2.8$$

where,

$$\hat{T}_1 = \sum_{I,A} t_I^A \hat{\tau}_I^A = \sum_{I,A} t_I^A a_A^\dagger a_I \qquad 2.9$$

$$\hat{T}_2 = \sum_{\substack{I>J \\ A>B}} t_{IJ}^{AB} \hat{\tau}_{IJ}^{AB} = \sum_{\substack{I>J \\ A>B}} t_{IJ}^{AB} a_A^\dagger a_B^\dagger a_J a_I \qquad 2.10$$

The t_I^A and t_{IJ}^{AB} terms are the one- and two-electron amplitudes, respectively, that need to be determined, and $\hat{\tau}_I^A$ and $\hat{\tau}_{IJ}^{AB}$ are the one- and two-electron excitation operators, respectively. Introduction of the CC wave function (equation 2.8) in the electronic Schrödinger equation leads to

$$\hat{H}exp(\hat{T})|\Psi_0\rangle = Eexp(\hat{T})|\Psi_0\rangle \tag{2.11}$$

For a truncated CC expansion, the application of the variational principle will result to complex nonlinear equations and thus, a projection technique is used for the computation of the CC energy E_{CC}. Left multiplication of equation 2.11 by exp $\left(-\hat{T}\right)$ provides the following expression:

$$exp(-\hat{T})\hat{H}exp(\hat{T})|\Psi_0\rangle = Eexp(-\hat{T})exp(\hat{T})|\Psi_0\rangle \tag{2.12}$$

We can now setup a set of equations for the computation of the CC energy and amplitudes by left projecting each determinant into equation 2.12:

$$\langle\Psi_0|exp(-\hat{T})\hat{H}exp(\hat{T})|\Psi_0\rangle = E_{CC} \tag{2.13}$$

$$\langle\mu|exp(-\hat{T})\hat{H}exp(\hat{T})|\Psi_0\rangle = 0 \tag{2.14}$$

where μ is the determinantal excitation manifold. These are the *projected coupled cluster equations* and an iterative, self-consistent solver is used for the computation of the CC energy amplitudes. Next, the Baker-Campbell-Hausdorff (BCH) expansion is introduced for the evaluation of the similarity-transformed Hamiltonian.

$$\begin{aligned} &exp(-\hat{T})\hat{H}exp(\hat{T})|\Psi_0\rangle \\ &\quad = \hat{H} + [\hat{H},\hat{T}] + \frac{1}{2!}\left[[\hat{H},\hat{T}],\hat{T}\right] + \frac{1}{3!}\left[\left[[\hat{H},\hat{T}],\hat{T}\right],\hat{T}\right] \\ &\quad + \frac{1}{4!}\left[\left[\left[[\hat{H},\hat{T}],\hat{T}\right],\hat{T}\right],\hat{T}\right] \end{aligned} \tag{2.15}$$

where $\left[\widehat{H},\widehat{T}\right]$is the commutator relation between operators $\widehat{H}$, and $\widehat{T}$. Application of the BCH expansion in the projected energy expression 2.13 leads to

$$E_{CC} = \langle\Psi_0|\widehat{H}|\Psi_0\rangle + \langle\Psi_0|\left[\widehat{H},\widehat{T}\right]|\Psi_0\rangle + \frac{1}{2!}\langle\Psi_0|\left[\left[\widehat{H},\widehat{T}\right],\widehat{T}\right]|\Psi_0\rangle + \frac{1}{3!}\langle\Psi_0|\left[\left[\left[\widehat{H},\widehat{T}\right],\widehat{T}\right],\widehat{T}\right]|\Psi_0\rangle + \frac{1}{4!}\langle\Psi_0|\left[\left[\left[\left[\widehat{H},\widehat{T}\right],\widehat{T}\right],\widehat{T}\right],\widehat{T}\right]|\Psi_0\rangle \quad 2.16$$

The above expression can be further simplified by considering that electron excitation operators acting on the HF wave function are $\langle\Psi_0|\hat{T} = 0$ and that $\langle\Psi_0|\widehat{H}\hat{T}_1|\Psi_0\rangle = 0$ due to the Brillouin theorem. In addition, the Hamiltonian operator has a maximum of two-particle terms (equation 2.6), so it can be coupled only with cluster operators that excite a maximum of two electrons. Thus, evaluation of all higher excitations will provide zero values $\langle\Psi_0|\widehat{H}\hat{T}_i|\Psi_0\rangle = 0, i > 2$ and the projected energy expression is further simplified to

$$E_{CC} = E_{HF} + \langle\Psi_0|\left[\widehat{H},\hat{T}_2\right]|\Psi_0\rangle + \frac{1}{2}\langle\Psi_0|\left[\left[\widehat{H},\hat{T}_1\right],\hat{T}_1\right]|\Psi_0\rangle \quad 2.17$$

We can now introduce the Hamiltonian (equations 2.6) and cluster operators (equations 2.9, 2.10) in the energy expression of equation 2.17. An energy expression can be obtained by using the standard algebra of creation and annihilation operators together with the Wick's theorem:

$$E_{CC} = E_{HF} + \sum_{\substack{i<j \\ a<b}} \langle ij\|ab\rangle t_{ij}^{ab} + \sum_{\substack{i<j \\ a<b}} \langle ij\|ab\rangle t_i^a t_j^b \quad 2.18$$

where the indices *i, j* and *a, b* correspond to the spin-free occupied and unoccupied orbitals, respectively, and $\langle ij\|ab\rangle$ is the standard two-electron (four-index) integrals. A set of projected equations 2.13, 2.14 are formed by considering the full determinantal excitation manifold μ.

For example, for CCSD, we need to consider projection to the single and double excitation determinants $(\mu = \Psi_i, i = S, D)$. Therefore, for the single determinants, the projection equations are,

$$0 = \langle \Psi_S|\hat{H}|\Psi_0\rangle + \langle \Psi_S|\left[\hat{H},\hat{T}_1\right]|\Psi_0\rangle + \langle \Psi_S|\left[\hat{H},\hat{T}_2\right]|\Psi_0\rangle + \frac{1}{2}\langle \Psi_S|\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_1\right]|\Psi_0\rangle + \langle \Psi_S|\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_2\right]|\Psi_0\rangle + \frac{1}{3!}\langle \Psi_S|\left[\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_1\right],\hat{T}_1\right]|\Psi_0\rangle \quad 2.19$$

This expansion is cubic in the $\hat{T}_1$ operator, but only linear to the $\hat{T}_2$ operator. For the double excitation manifold, the projection equations are,

$$\begin{aligned} 0 = {} & \langle \Psi_D|\hat{H}|\Psi_0\rangle + \langle \Psi_D|\left[\hat{H},\hat{T}_1\right]|\Psi_0\rangle + \langle \Psi_D|\left[\hat{H},\hat{T}_2\right]|\Psi_0\rangle + \frac{1}{2}\langle \Psi_D|\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_1\right]|\Psi_0\rangle \\ & + \langle \Psi_D|\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_2\right]|\Psi_0\rangle + \frac{1}{2}\langle \Psi_D|\left[\left[\hat{H},\hat{T}_2\right],\hat{T}_2\right]|\Psi_0\rangle \\ & + \frac{1}{3!}\langle \Psi_D|\left[\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_1\right],\hat{T}_1\right]|\Psi_0\rangle + \frac{1}{2}\langle \Psi_D|\left[\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_1\right],\hat{T}_2\right]|\Psi_0\rangle \\ & + \frac{1}{4!}\langle \Psi_D|\left[\left[\left[\left[\hat{H},\hat{T}_1\right],\hat{T}_1\right],\hat{T}_1\right],\hat{T}_1\right]|\Psi_0\rangle \end{aligned} \quad 2.20$$

Here, the single-excitation amplitudes appear to fourth-order and the double excitation amplitudes are up to quadratic. The evaluation of commutators shown in equations 2.19 and 2.20 follows the same principles as for the energy expression.

Once the t_i^a and t_{ij}^{ab} amplitudes have been evaluated, they are introduced in the energy expression of equation 2.18 for the computation of the CCSD energy. The amplitudes are then updated and the full set of projected equations are solved again until convergence. Further inspection of the energy equation reveals that only the singles and doubles amplitudes contribute directly to the CC energy. However, all higher excitations will affect indirectly the energy expression since the one- and two-electron amplitudes are solved together with all higher-order amplitudes, so they depend on each other.

2.4.2 Data-driven coupled-cluster singles and doubles

Equation 2.18 implies that the energy function depends on the CC wave function parameters t_i^a (or t_1) and t_{ij}^{ab} (or t_2). In other words, we have to find the optimum amplitudes in order to calculate the CC energy. The solution of the projected CC equations to determine t_1 and t_2 amplitudes has a computational cost of approximately N^6, where N is the number of basis functions. Hence, this step is considered the bottleneck for CC calculations. Recently, we have developed a data-driven coupled-cluster (DDCC) model that aims to accelerate the computation of the t_2 amplitudes. *The DDCC model uses MP2-level electronic structure data to predict t_2 values for CCSD energy calculations.* Thus, DDCC "learns" the contribution of each two-electron excitation to the total correlation energy. These predicted amplitudes can either be used as an initial guess for the iterative process of solving projected CC equations ("exact DDCC") or can be directly introduced in equation 2.18 to estimate the CCSD energy ("approximate DDCC"). In the first case, the aim is to substitute the MP2 amplitudes that are used as initial guess in conventional CCSD implementations with the "learned" amplitudes. Use of the predicted t_2 amplitudes from the DDCC model reduces the number of iterations to obtain the converged t_2 amplitudes and compute the exact CCSD energy. In the latter case, substitution of the predicted amplitudes directly to the energy expression of equation 2.18 provides an estimated CCSD energy at MP2 speed, which scales as N^4. To conclude, the aim of DDCC is to learn the parameters of the CCSD wave function by utilizing a physics-based representation of the two-electron excitations from a computationally cheaper method (MP2). In our current implementation, we solve explicitly the projected equations for the t_1 amplitudes since their number is relatively small. This process does not introduce a significant computational overhead.

To understand the features used in the DDCC model, it is important to understand how MP2 amplitudes $t_{ij(MP2)}^{ab}$ are calculated.

$$t_{ij(MP2)}^{ab} = \frac{\langle ab||ij\rangle}{\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b} \quad 2.21$$

In equation 2.21 ε_k is the energy of orbital k and $\langle ab||ij\rangle$ are the two-electron integrals. It becomesclear that the MP2 amplitudes depend on the properties of the four orbitals *i, j, a,* and *b.* Thus,our focus is to encode in the DDCC representation the physics of the two-electron excitation. For the development of a transferable model to molecular systems of varying size, orientation, and atom composition, we chose to restrict the features to electronic structure properties. Therefore, the chosen features do not explicitly incorporate the identity or position of any of the atoms.

Here, the features used in the DDCC model are analyzed. First, the most obvious feature is the actual MP2 amplitude, since its relation to the CCSD t_2 amplitudes is established. Next, the numerator and the denominator terms of the MP2 amplitude are considered separately. Both are introduced as input in the DDCC model. The numerator consists of the two-electron integral and the denominator term consists of the energy difference between the occupied and virtual orbitals $\varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b$. The next set of features consists of individual energies for each orbital involved in the two-electron excitation. The orbital energies are further separated by their oneelectron Hamiltonian, the Coulombic, and exchange components. The Coulombic (J_i^a, J_j^b) and exchange terms (K_i^a, K_j^b) within the two-electron excitation manifold are also included in the feature set. The next feature is Boolean and represents whether or not the two electrons are promoted to the same virtual orbital. This feature is important since the t_2 amplitudes for an excitation of two electrons with antiparallel spin to the same virtual orbital

have a significantly greater value. In addition to these features, the log_{10}value of MP2 amplitudes, the log_{10} value of the Coulomb and Exchange integrals for individual excitations, and whether the t_2 amplitudes are greater than zero (Boolean) are also included in the representation vector.

Altogether the feature set for a single t_2 amplitude consists of 30 electronic structure properties. All features and t_2 amplitudes are extracted from the Psi4NumPy[56] software. An important advantage of our physics-based representation is that all features are calculated at the HF and MP2 levels.

As a proof-of-concept, the DDCC scheme utilizes a *k*-nearest neighbors (*k*-NN) regression algorithm to learn the t_2 amplitudes. Since *k*-NN is an instance-based learning algorithm, it provides a distinct advantage over regression algorithms that are trained by minimizing a loss function. *k*-NN is sensitive to the local environment of the feature space and treats all amplitudes equally. The Euclidean distance between the test set and training set in feature space is taken, and the *k*-NNs are determined and averaged. Algorithms trained via the minimization of loss functions tend to accurately fit the larger amplitudes, while ignoring the smaller amplitudes, which are not advantageous for the DDCC scheme.

The DDCC method uses a modified Euclidean distance function with optimized feature weights since certain features may not contribute equally for the prediction of amplitudes. The modified Euclidean distance is defined as

$$d(\boldsymbol{p}, \boldsymbol{q}) = \sqrt{[x_1(q_1 - p_1)]^2 + [x_2(q_2 - p_2)]^2 + \cdots + [x_n(q_n - p_n)]^2} \qquad 2.22$$

where **p** and **q** denote the test and training set features, respectively, and subscripts denote the number of features from 1 to n. In order to accentuate the contribution of certain features, feature weights, x_n, are determined by a parameter grid search on scaled data, where x_n = 1, 1.25, 1.5, 2, 5, 10, 100, 1000 (search space of 8 × 30 features). The most important features have been determined to be the MP2 amplitudes, elements belonging to the MP2 amplitudes such as the two-electron integrals and the differences in orbital energies, and the binary feature denoting whether the two electrons have been promoted to the same virtual orbital or not. The features are scaled using the MinMaxScaler implemented in Sci-Kit Learn.[57]

Although *k*-NNs can be used as a good proof-of-concept, the computational cost associated with increasing the number of training data makes the method cost prohibitive. Due to the computational costs of *k*-NNs, we are currently exploring alternative machine learning methodologies, such as ensemble learning algorithms, that can be applied to larger databases and larger molecular systems.

2.4.3 DDCCSD and transferability

One of the highlights of the data-driven coupled-cluster (DDCC) method is the transferability between systems with similar atomic composition or between clusters of variable size. In order to demonstrate this important feature, we have trained models with water clusters of variable size. In particular, we trained three different models for predicting the CCSD energy of four different configurations of water hexamers (cage, book, prism, and ring). The first model used data from only water monomers, the second was augmented with data from water dimers, and the third was trained on data from water monomers, dimers, and trimers. Figure 2.1 shows the deviation between the predicted DDCC energy before any iterations and the exact CCSD energy. For comparison, we have included the CCSD energy obtained from the initial MP2

amplitudes before initializing the CC solver (0th iteration of the iterative process). The model that was trained only on data from water monomers showed significantly larger deviations than the conventional 0th-iteration CCSD energy (Figure 2.1, second bar per molecular cluster). The reason is that no information on the electron excitations between two water molecules was included in our training set and the model does not include any electron correlation information of the hydrogen bond network of the water cluster. A significantly lower deviation was obtained from the addition of data from water dimers (Figure 2.1, third bar per molecular cluster). The error drops almost to zero when data from water trimers included in the training of the ML model (Figure 2.1, fourth bar per molecular cluster). In other words, the model successfully learns the physics of hydrogen bonds between water molecules and accurately estimates the CCSD wave function. As it was discussed before, the predicted DDCCSD wave function can be used as starting guess for solving the CCSD equations in order to provide the exact CCSD energy in significantly less computational time since the amplitudes are already very close to their final value. Thus, DDCC holds promise for accelerating quantum chemical calculations of large clusters by utilizing data obtained from smaller molecular species.

In a separate study, we investigated the local nature of the electron correlation and how it affects data-driven schemes similar to DDCC.[38] It is well known that electron correlation is short-ranged in nature, and the local character of electron excitations can be explored with localized molecular orbitals or domains, a process pioneered by Saebø and Pulay. For that purpose, we developed a scheme that estimates the total pair-correlation energy ϵ_{ij} corresponding to an occupied electron pair ij. The sum of all ϵ_{ij} pair terms provide the total CCSD correlation energy:

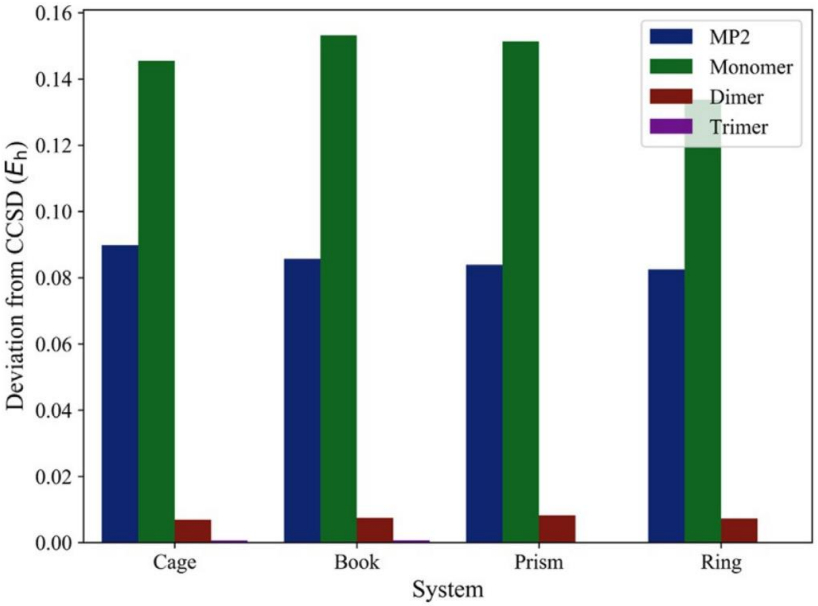

Figure 2.1: DDCC absolute deviation (Eh) from the converged CCSD plotted for the four systems containing data with the monomer, dimer, and trimer systems, as well as from CCSD using from MP2 amplitudes. *Reproduced from J. Townsend, K.D. Vogiatzis, Data-driven acceleration of the coupled-cluster singles and doubles iterative solver, J. Phys. Chem. Lett. 10 (2019) 4129–4135, Copyright 2019 American Chemical Society.*

$$E_{\text{corr}}^{\text{CCSD}} = \sum_{ij} \epsilon_{ij} \qquad 2.23$$

We have used canonical, Boys and Pipek-Mezey (PM) localized orbitals and trained a model with data from small alkanes (C_nH_{2n+2}, n = 2-7). The error of the predicted total correlation energies is shown in Figure 2.2. It is evident that the short-range nature of the electron correlationis successfully encoded in the models that use localized orbitals. The errors for the two models (Boys or PM) for longer alkane chains are constantly below 1 kcal/mol (chemical accuracy, shown with a horizontal dashed line in Figure 2.2). This is further evidence that the model learns the local nature of the two-electron excitations and the CC wave function. This scheme offers transferability and efficiency as it was found that only 140 randomly selected molecules are adequate to achieve chemical accuracy for more than 133,000 organic molecules (GDB-9 database).

2.5 Limitations of the DDCCSD schemes

Although Townsend and Vogiatzis provide a promising proof of concept for DDCCSD in their work discussed in Section 2.4, there are some limitations that prevent DDCCSD method's applicability towards large chemical system with complex basis sets.

In the previously discussed work to predict the t_2 amplitudes using MP2-level electronic structure data, we use canonical orbitals. Although the accuracy of these models is higher for comparatively smaller molecules (molecules with three or less than three non-hydrogen atoms), results from the study done to predict the pair-energies discussed in the previous Section, has shown that a significant improvement in energy prediction can be achieved using localized molecular orbitals.

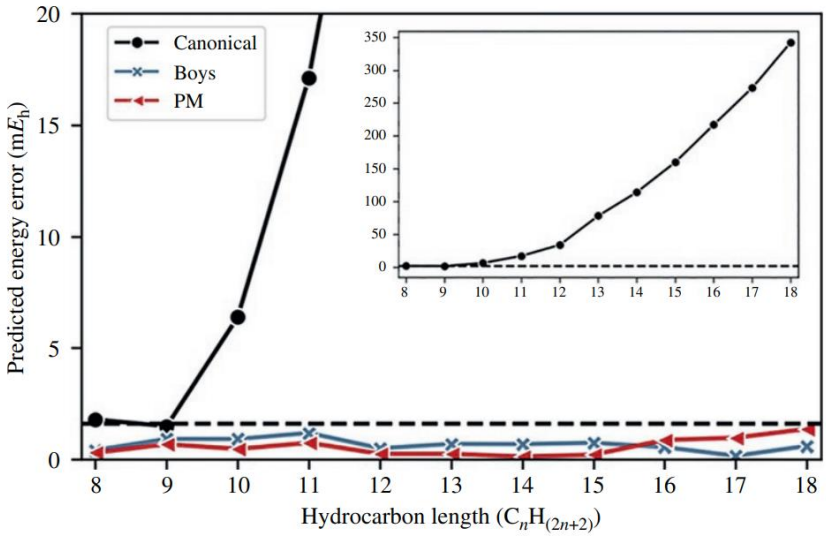

Figure 2.2: Predicted CCSD energy error for hydrocarbons octane through octadecane (n = 8-18) based on the model trained on single configurations of propane through heptane (n = 3-7). In the inset, the CCSD energy errors computed with canonical orbitals are shown in greater detail. *Reproduced from J. Townsend, K.D. Vogiatzis, Transferable MP2-based machine learning for accurate coupled-cluster energies, J. Chem. Theory Comput. 16 (2020) 7453–7461, Copyright 2020 American Chemical Society.*

Therefore, implementation of localized orbitals based DDCCSD scheme to predict t_2 amplitudes was needed. These LMO based DDCCSD schemes are discussed extensively in the next Chapter (Chapter Chapter 3).

The second limitation of the initial DDCCSD models that hinders the applicability is larger chemical systems is the distribution and the number of t_2 amplitudes. For a molecule with a given basis set, there are $n_{occ}^2 \times n_{occ}^2 \times n_{vir}^2 \times n_{vir}^2$ number of t_2 amplitudes where, n_{occ} is the number of occupied orbitals and n_{vir} is the number of virtual orbitals. Therefore, the number of amplitudes exponentially increases with the number of occupied orbitals depending on the size of the chemical system and the number of virtual orbitals which depends on the size of the basis set. In Chapter Chapter 3, five types of amplitude selection schemes that were developed to address this issue are discussed.

Next major limitation of the DDCCSD scheme discussed in previous Section is the machine learning algorithm used. Although k-nearest neighbors offers higher accuracies, it limits the applicability to larger systems due to the prediction time complexity. As a solution we have developed machine learning models using random forest and neural networks, that are extensively discussed in Chapters Chapter 3 and Chapter 4.

Furthermore, in Chapter Chapter 4 we discuss the development of physics informed, multitasking models that predicts t_1 amplitudes along with t_2 amplitude. Another improvement to initial DDCCSD schemes is the prediction of perturbative triples energy and t_2 amplitudes for CCSD full triples will be discussed in Chapter Chapter 5.

Chapter 3 Exploration of the Two-Electron Excitation Space with Data-Driven Coupled Cluster

3.1 Disclosure

This Section contains an adapted reproduction of the published manuscript in Journal of Physical Chemistry A, taken with permission. P. D. Varuna S. Pathirage developed localized molecular orbital DDCCSD models.[39] P. D. Varuna S. Pathirage and Justin T. Phillips developed five amplitude selection schemes. P. D. Varuna S. Pathirage developed score-to-bins, cluster-bins, large amplitude and electron correlation schemes and Justin T. Phillips developed probabilistic selection scheme under the guidance of P. D. Varuna S. Pathirage. P. D. Varuna S. Pathirage and Justin T. Phillips worked on manuscript preparation under the guidance of Professor Konstantinos D. Vogiatzis.

3.2 Abstract

Computational cost limits the applicability of post-Hartree–Fock methods such as coupled-cluster on larger molecular systems. The data-driven coupled-cluster (DDCC) method applies machine learning to predict the coupled-cluster two-electron amplitudes (t_2) using data from second-order perturbation theory (MP2). One major limitation of the DDCC models is the size of training sets that increases exponentially with the system size. Effective sampling of the amplitude space can resolve this issue. Five different amplitude selection techniques that reduce the amount of data used for training were evaluated, an approach that also prevents model overfitting and increases the portability of data-driven coupled-cluster singles and doubles to more complex molecules or larger basis sets. In combination with a localized orbital formalism to predict the CCSD t_2 amplitudes, we have achieved a 10-fold error reduction for energy calculations.

3.3 Introduction

Electron correlation is the heart of molecular electronic structure theory.[58] Hartree−Fock (HF) theory includes Fermi correlation in an exact sense and electron repulsion in an average, mean-field manner since it fails to describe the instantaneous Coulombic repulsion between electron pairs. Post-HF methods such as configuration interaction, many-body perturbation theory, and coupled-cluster (CC) theory introduce the missing electron correlation by considering a systematic expansion of the *N*-electron basis formed from all possible electronic configurations within a given orbital basis.[43] Among post-HF methods, coupled-cluster singles and doubles with perturbative triples [CCSD(T)][59] provide a balance between accuracy and efficiency for molecular systems without degeneracies or near-degeneracies. A variety of different approaches have been introduced that aim to accelerate the convergence of the CCSD(T) and increase its applicability, including explicitly correlated methods,[12, 13] linear scaling methods,[60] pair-natural orbital expansions,[61-65] fragmentation schemes,[66, 67] and high-performance computing.[68]

Machine learning (ML) has been recently proposed as an alternative to traditional methodologies for further acceleration of CCSD(T). We can separate these methods into two groups based on the feature space (representation) used as input to ML models.[42] In the first group, the feature space depends on atomic positions or atomic functions that are able to encode the local environment of atoms in molecules and materials.[35, 69] The "Accurate NeurAl networK engINe for Molecular Energies" (ANI) family of methods[22, 31, 70, 71] is a representative example where a vector that contains specific radial and angular chemical information on an individual atom environment is introduced as input. These models are trained with density functional theory or, partially, with CC data and offer significant speedup for the reliable calculation of energies and forces.

The second group of methods utilizes a descriptor space that is based on quantum chemical information obtained from a low-level method such as HF or second-order perturbation theory (MP2) and aims to predict results from a higher-level method [e.g., CCSD or CCSD(T)]. This type of molecular representation is indirectly correlated to the atomic positions, and it can offer direct transferability within molecular environments since they encode the underlying electronic structure properties.[46, 52, 54, 55, 72-76] Along these lines, we have developed a data-driven methodology for the prediction of CCSD two electron excitation (t_2) amplitudes using MP2-level electronic structure data.[37, 42] The data-driven coupled-cluster singles and-doubles (DDCCSD) scheme has some attractive features, such as transferability between systems of different sizes. However, as the number of basis functions and the system size increase, the number of t_2 amplitudes increases, which eventually introduces a computational bottleneck.

In this work, we extended DDCCSD by introducing two new key features that aim to expand the applicability of the method to larger molecular systems. First, we have applied a localized orbital formalism into the data-driven coupled cluster (DDCC) methodology, and second, we have developed five different approaches for the data-point selection from the amplitude space [score to bins (SB), clustering to bins (CB), large amplitudes (LA), electronic correlation (EC), and probabilistic selection (PS)]. The amplitude selection schemes reduce the computational burden of the training and testing process and allow us to apply DDCC on larger molecules and larger basis sets, which was previously unfeasible. Another important aspect of the systematic selection of training data points is the increased accuracy of the ML models by preventing overfitting.

This article is organized as follows: a brief theoretical background of DDCCSD together with a new approach to the calculation of the feature weights is presented in Section 3.4. The computational details related to the data generation used in this work are given in Section 3.5.

Section 3.6 introduces the definitions, rationale, and technical aspects of the five amplitude selection schemes that are examined in this study. The evaluation of the performance of different selection schemes is discussed in Section 3.7, and finally, concluding remarks are presented in Section 3.8.

3.4 Theoretical Background

3.4.1 Data-Driven Coupled-Cluster Singles and Doubles.

In CC theory, the correlation energy ($E_{\text{corr}}^{\text{CCSD}}$) is computed by using the following equation.

$$E_{\text{corr}}^{\text{CCSD}} = \sum_{\substack{a<b \\ i<j}} \langle ij \| ab \rangle t_{ij}^{ab} + \sum_{\substack{a<b \\ i<j}} \langle ij \| ab \rangle t_i^a t_j^b \qquad 3.1$$

The indices *i* and *j* correspond to occupied orbitals, *a* and *b* to virtual orbitals, t_i^a and t_{ij}^{ab} are the one- and two-electron amplitudes, respectively, and $\langle ij \| ab \rangle$ is the two-electron integral between orbitals *i*, *j*, *a*, and *b*. The variational computation of the t_i^a and t_{ij}^{ab} amplitudes is a nontrivial process, and thus, the amplitudes are computed by solving the projected CC equations iteratively.

In equation 2.13 and 2.14, Ψ_0 is the HF reference wave function, $\hat{T}$ is the cluster operator, and μ the determinantal excitation manifold (in CCSD, μ are either the singly or doubly excited determinants). Typically, the MP2 amplitudes $t_{ij(MP2)}^{ab}$ are used as an initial guess for CC t_{ij}^{ab} amplitudes (equation 2.21).

In the DDCCSD approach, the initial CC t_{ij}^{ab} amplitudes are predicted by ML using 30 input features from MP2. Since MP2 amplitudes are the initial guess for the iterative solution of the

projected CC equations, they are selected as one of the features used in the DDCCSD method. Other features of the DDCCSD are the numerator and denominator terms of the MP2 amplitude equation (equation 2.21). The denominator is further broken into individual terms, and the occupied and virtual orbital energies are also introduced in the input vector. The individual orbital energy is broken into the one-electron, Coulomb, and exchange contributions. Another feature of the DDCCSD method is whether the excited electrons are promoted to the same virtual orbital (binary). For the full feature list, see Supporting Information, Section S-I (Appendix I).

The predicted amplitudes $t_{ij(DDCC)}^{ab}$ are then introduced into the singles ($\mu = \Psi_i^a$) projected equations 2.13 and 2.14 and a single step is performed that updates the t_i^a amplitudes ($t_{i(1)}^a$or $t_{1(1)}$). The DDCCSD energy is then computed as

$$E_{\text{corr}}^{\text{DDCCSD}} = \sum_{\substack{a<b \\ i<j}} \langle ij \| ab \rangle t_{ij(DDCC)}^{ab} + \sum_{\substack{a<b \\ i<j}} \langle ij \| ab \rangle t_{i(1)}^{a} t_{j(1)}^{b} \qquad 3.2$$

Note that in the next paragraphs, we will refer to a generic t_{ij}^{ab} amplitude as t_2. Alternatively, we can introduce the predicted t_2 amplitudes to the solver of CCSD and iteratively obtain the exact CCSD energies in less computational effort, as we have shown previously.[37] In this work, we solely consider the first approach, where the predicted amplitudes from ML are used for the computation of an approximate CCSD energy.

3.4.2 Feature Weights.

In our previous work on the DDCCSD method,[37] a weight w_i was assigned to each feature, which was optimized via a grid search for a data set of water molecules in different

conformations close to the equilibrium geometry. In order to extend the transferability of DDCC, we have applied an alternative approach for weight assignment to features. We have calculated the Pearson correlation coefficient (r) between each of the features with the CCSD t_2 amplitude value for the training data

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3.3}$$

Here, x_i is the value of feature x, y_i is the value of the ith t_2 amplitude, $\bar{x}$ is the mean of feature x, and $\bar{y}$ is the mean of the value of CCSD t_2 amplitudes. This correlation coefficient value is used as the weight for each feature (see Supporting Information, Section S-II in Appendix I) and is used for the comparison of canonical molecular orbitals (CMOs) and localized molecular orbitals (LMOs) and for the comparison of different amplitude selection schemes.

3.5 Computational Details

All quantum chemical calculations, feature extraction, and output data collection (t_2 amplitudes) were performed using the Psi4NumPy software[56] as described previously.[37] For the purposes of this study, we used five molecules for the calibration and performance evaluation of the new models; those are diethyl ether, divinyl ether, vinyl alcohol, methyl vinyl ether, and allyl alcohol molecules (Figure 3.1). Initial ground-state structures were optimized with the B3LYP[77, 78] density functional and the 6-31G[79, 80] basis set, with the Psi4[81] quantum chemical program package.

In a previous study on machine-learned CCSD pair energies, we demonstrated that models trained with LMOs are more transferable than the CMOs and that the Foster-Boys (Boys) localization scheme provided slightly better accuracy than the Pipek-Mezey localization

scheme.[38] In this work, the Boys localization scheme is introduced in DDCC and its performance is compared with canonical orbitals. For the comparison of the performance of the CMOs and LMOs, we used six training sets. The first five training sets consisted of 30 conformers of an individual molecule, while the sixth training set consisted of a mixture of six conformers from each molecule. Similarly, five test sets, each consisting of 50 conformers of a molecule, were used for testing. Conformers for training were generated by rotating the C-O bond of each of the five molecules shown in Figure 3.1. Conformers for testing were generated by rotating the same bond. For this analysis, the STO-3G[82, 83] basis set was used to calculate the CCSD t_2 amplitudes.

For the evaluation of the amplitude selection schemes, we used the mixed training set since it better represents the transferability of the DDCCSD models. For testing, we used five conformers from each molecule. Initially, we used the STO-3G basis set with both LMO and CMO formalisms and then expanded our study to the cc-pVDZ and aug-cc-pVDZ basis sets with LMOs.

Three basis sets were considered in this study. Initially, we used the STO-3G basis set for the comparison of the CMO and LMO models and for determining the best set of hyperparameters for each scheme for further studies. A set of 1,180,056 data points (amplitudes) was available in total for training from the conformers when the STO-3G basis set was used. In order to study the performance of the amplitude selection schemes with increasing number of data points, we further used cc-pVDZ[84] and aug-cc-pVDZ[85] basis sets. A total of 57,357,372 data points were available for training with the cc-pVDZ basis set, and 197,707,512 data points were available with the aug-cc-pVDZ basis set (see Supporting Information, Section S-III in Appendix I, for the CCSD t_2 amplitude distribution for all basis sets considered in this study). Figure 3.2 shows a log plot of the distribution of CCSD t_2 amplitudes for the aug-cc-pVDZ basis set.

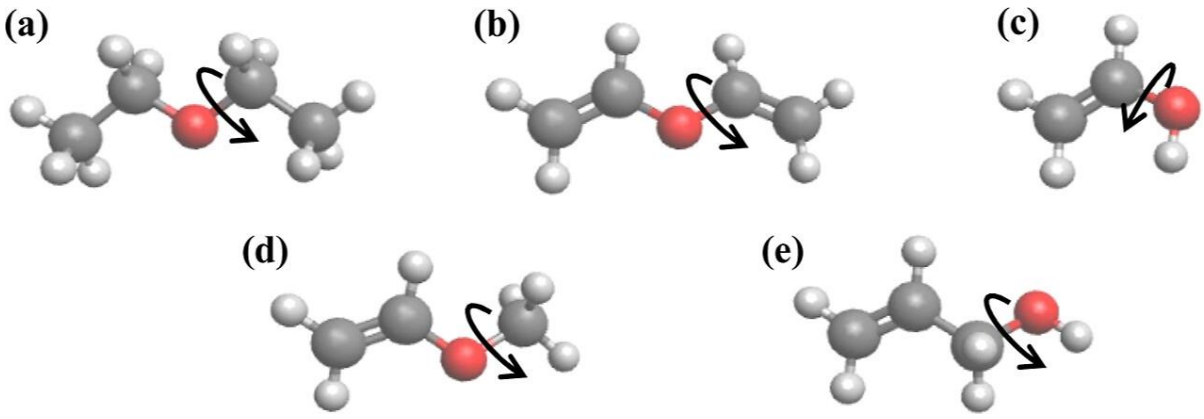

Figure 3.1: Optimized structures of five organic molecules considered in this study: (a) diethyl ether, (b) divinyl ether, (c) vinyl alcohol, (d) methyl vinyl ether, and (e) allyl alcohol. Molecular conformers for training and testing of the DDCC models we generated by rotating the dihedral angle indicated by curve.

All DDCCSD models were trained with the random forest ML algorithm[86] using the scikit-learn package.[57] During training, random forest models create a set of decision trees; the number of decision trees is treated as a hyperparameter that must be defined. The prediction for a new data point is the mean of the output from each decision tree. The first step of our study was a hyperparameter search with respect to the number of estimators (decision trees), the choice of the loss function, as well as the minimum number of data points. For the hyperparameter optimization, models were trained with 30 different water conformers and tested with 100 different water conformers with the STO-3G basis set. From this analysis, we found that the optimum set of hyperparameters was 350 estimators with squared error loss function and a minimum sample split of 2 (see Supporting Information, Section S-IV in Appendix I). These hyperparameters were used throughout this study.

For the assessment of the DDCC models trained with truncated amplitude spaces, the mean absolute error (MAE) was calculated for each training and test set combination where *n* is the number of test conformers,

$$MAE = \frac{\sum_{n=1}^{n} |E_{CCSD} - E_{DDCCSD}|}{n} \qquad 3.4$$

All the features were scaled using MinMaxScaler[57] as implemented in the scikit-learn package.[57]

3.6 Amplitude selection schemes

The main objective of this study is the development and assessment of data selection techniques for the reduction of the amplitude space needed for the DDCC model training. Two aspects were considered for decreasing the number of training data points. The first aspect is

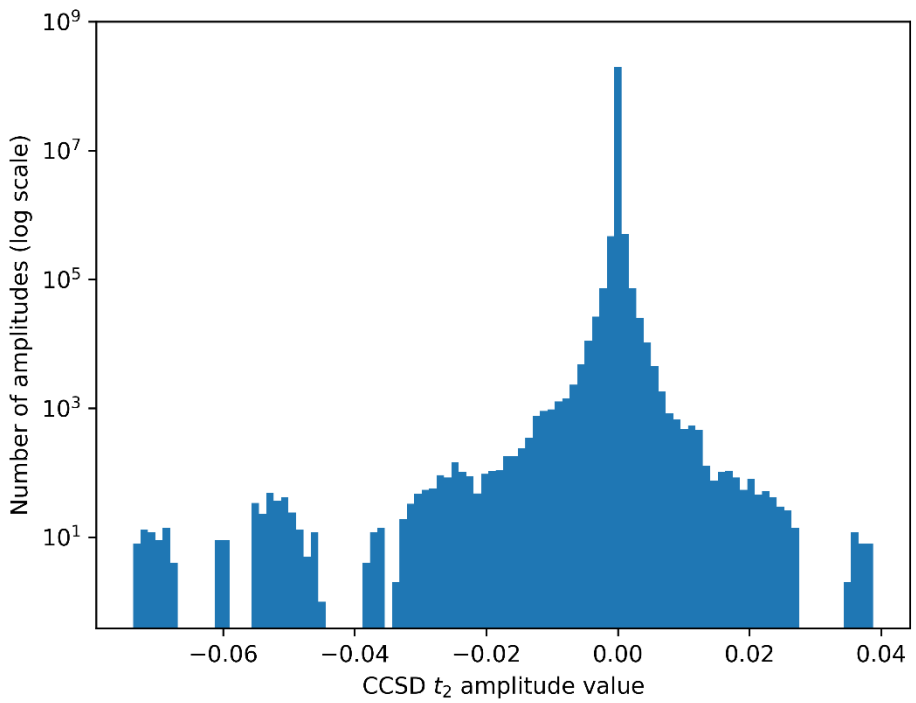

Figure 3.2: Distribution of CCSD t_2 amplitudes (on a log scale) computed with the aug-cc-pVDZ basis set.

computational cost and time reduction. For example, the random forest ML algorithm has a training time complexity of $O(n \times \log(n) \times d \times k)$ where n is the number of training data points, d is the number of features and k is the number of decision trees. Thus, it becomes evident that the number of input data directly affects the training time. The second consideration is the prevention of model overfitting since a significant number of amplitudes for a given molecule have values close to zero. The number of amplitudes with values closer to zero is also size-dependent. As a solution, we propose the selection of an amplitude subspace that is used for DDCC training without affecting the model accuracy. In this study, we discuss five selection approaches. The first method is SB, where a value (score) is assigned to each amplitude based on their scaled feature scores. The second method involves ML clustering for sorting data into bins (CB). The work of Maitra et al.[87-90] formed the basis for the LA scheme. In their study, they used a set of amplitudes that are greater than a cutoff value to train models that would predict amplitudes that are less than the cutoff value. Here, we are selecting only the amplitudes that are larger than a particular cutoff value for training and testing, and we will refer to it as a LA scheme. The fourth method considers the MP2 correlation energy of two-electron excitations $ij \rightarrow ab$ for selecting amplitudes for training and testing (EC scheme), and the fifth method involves a probabilistic selection (PS scheme) of the initial amplitude space based on feature values. These five schemes are presented in detail in the following paragraphs.

3.6.1 Score to Bins.

In the SB approach, a score S_i is assigned to each amplitude based on the values of the scaled features

$$S_i = \sqrt{\sum_{n=1}^{30} (w_n f_n^i)^2} \qquad 3.5$$

Here, w_n is the weight of feature n, and f_n^i is the value of the nth feature of the ith amplitude. The score S_i represents the distance from the origin of the feature space to each data point. Then, the amplitude space is clustered into $m \times m$ bins, where m is the number of sections that the data set is partitioned along feature score and CCSD t_2 axes. The initial choice of m was 100. Therefore, the amplitude space was divided into 10,000 bins. Each amplitude is assigned to a bin according to its feature score S_i and the value of the CCSD t_2 amplitude. Next, a common cutoff will be set to all the bins, such that the defined percentage of amplitudes will be selected when it is applied. If a bin has less or equal number of amplitudes than the cutoff, all the amplitudes are used for training. If a bin has amplitudes more than the cutoff, then the number of amplitudes that are equal to the cutoff will be randomly selected for training. Overall, ten models were trained with 5, 10, 20, 30, 40, 50, 60, 70, 80, and 90% of the total number of amplitudes. Figure S4 in the Supporting Information Section SV in Appendix I shows a heatmap of the amplitude grouping to bins.

3.6.2 Clustering to Bins.

The CB approach is similar to SB, but instead of assigning a score, *k*-means clustering[91] is used for assigning amplitudes to bins. *k*-means clustering is an unsupervised learning algorithm whose goal is to assign the data points to k number of clusters by minimizing the difference between the data point and the centroid (mean) of the cluster that has been assigned to. The algorithm starts by assigning centroids to the clusters. The number of clusters k is treated as a model hyperparameter that should be defined (vide infra). After the initial assessment of CB where a value of 100 clusters was selected for k, a hyperparameter optimization was performed to determine the best number of k clusters. For that purpose, the *k*-means cluster algorithm was executed 10 times for each clustering with different centroid seeds, and the best clustering was

selected each time. To assess the performance, ten models were trained that contained 5, 10, 20, 30, 40, 50, 60, 70, 80, and 90% of the total amplitudes using the CB scheme.

We used the 30 scaled features multiplied by the weight for each feature and the CCSD t_2 amplitude value for clustering into k bins. Then, amplitudes were selected for training from each bin using a similar approach to the SB method by setting a common cutoff value. Both SB and CB approaches attempt to select an amplitude subspace that uniformly represents the full amplitude space, which can reduce the amplitudes with values closer to zero but can also avoid model overfitting.

After the optimum percentage of amplitudes was determined using the above-mentioned procedure, the effect of the number of bins was assessed by creating models with different numbers of bins and calculating their accuracy. For the SB approach, six models were trained by selecting 20% of the total amplitudes from 50 × 50, 100 × 100, 150 × 150, 200 × 200, 250 × 250, and 300 × 300 bins. For the CB approach, six models were trained with 5% amplitudes chosen from 50, 100, 150, 200, 250, and 300 bins (see Supporting Information, Section SVIII in Appendix I).

3.6.3 Large Amplitude

The LA approach is different from the SB and CB approaches since it uses only data points with MP2 amplitude magnitude greater than a defined cutoff value. All amplitudes below this value are set to zero, while the rest are predicted by DDCC. To further validate this notion, we calculated the correlation coefficient between the MP2 amplitudes and the CCSD t_2 amplitudes (STO-3G basis set, LMOs). A correlation coefficient of 0.9330 was obtained when all amplitudes were taken into account (Figure 3.3), but when amplitudes with magnitudes greater than 0.01 were considered, the correlation coefficient increased to 0.9943. For the LA approach,

11 models were trained with data points corresponding to MP2 amplitudes within the 10^{-2}-10^{-7} range.

3.6.4 EC Scheme

The EC scheme shares similarities with the LA scheme since both approaches use a cutoff value for selecting training data points. There are two main differences between LA and EC. First, EC uses the MP2 correlation energy $e_{ij(\mathrm{MP2})}^{ab}$ as a cutoff criterion, instead of the amplitude value as in LA. For a two-electron excitation $ij \rightarrow ab$, the MP2 correlation energy $e_{ij(\mathrm{MP2})}^{ab}$ is given as,

$$e_{ij(\mathrm{MP2})}^{ab} = \langle ij \| ab \rangle t_{ij\,(\mathrm{MP2})}^{ab} \tag{3.6}$$

Optimum performance was found when a cutoff of 1×10^{-7} E_h (14.3% of the total amplitudes) was used for the STO-3G basis set, $1 \times 10^{-7.5}$ Eh (5.76% of the total amplitudes) for models trained with data from the cc-pVDZ basis set, and 1×10^{-8} E_h (4.47% of the total amplitudes) for models with data from augcc-pVDZ. All amplitudes with MP2 correlation energy greater than the cutoff value were selected for training. The second difference is related to the selection of the amplitudes that are "predicted" by the DDCC scheme. Specifically, after careful inspection of the opposite-spin (OS) MP2 correlation energy captured by an amplitude subspace (see Supporting Information, Section S-VI in Appendix I), we set a cutoff threshold based on the OS MP2 correlation energy percentage (p_cutoff). All $ij \rightarrow ab$ correlation energies $e_{ij(\mathrm{MP2})}^{ab}$ of test molecules are arranged in ascending order, and CCSD t_2 predictions are made for the amplitudes with the largest magnitude for $e_{ij(\mathrm{MP2})}^{ab}$ that capture a p_cutoff percent of the

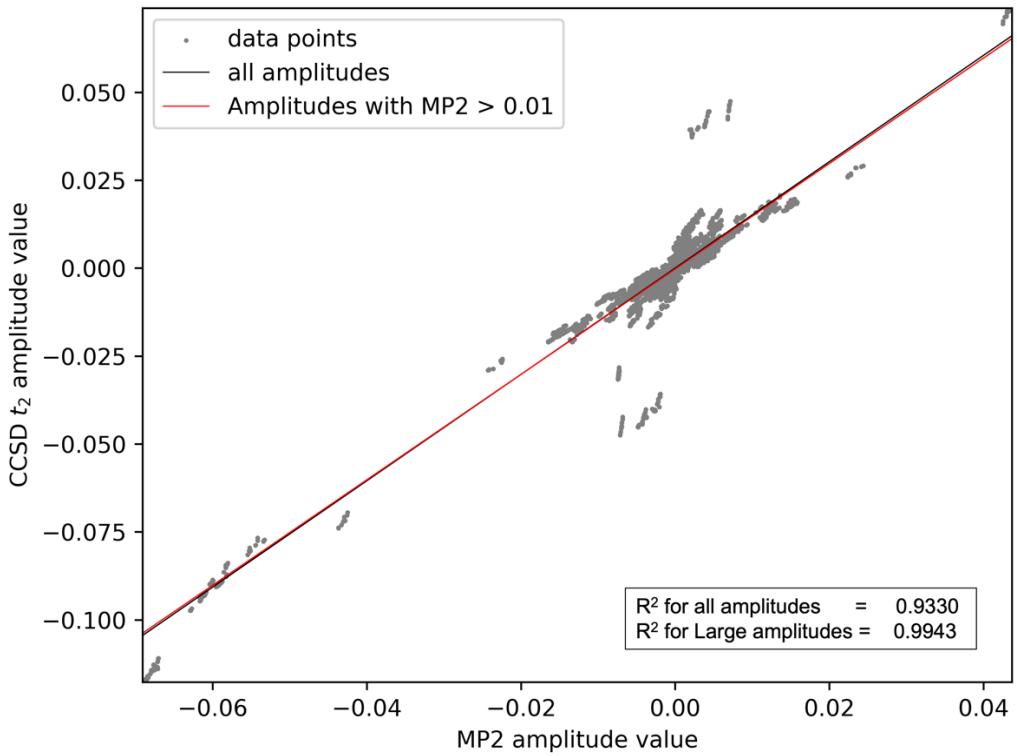

Figure 3.3: CCSD t_2 amplitude value vs MP2 amplitude value for all the amplitudes calculated with the STO-3G basis set and LMO formalism.

total MP2 OS correlation energy. In Section 3.7, we discuss the effect of different p_{cutoff} values on the model accuracy.

3.6.5 Probabilistic Selection.

The PS approach uses a weighting mechanism where each amplitude from the initial amplitude space is assigned a W_n weight defined as

$$W_n = t_{ij\,(MP2)}^{ab} \times (e_i + e_j) \tag{3.7}$$

where e_k is the energy of occupied orbital k. Once each amplitude has been assigned a weight, the probability p_n is computed as

$$p_n = \frac{W_n}{\sum_n W_n} \tag{3.8}$$

In equation 3.8, p_n is the probability assigned to the nth amplitude. Once the weights are converted to probabilities, a certain percentage of the initial amplitudes are selected in accordance with the assigned probabilities. For the PS approach, the MP2 amplitude was used as an initial weight, as it is the feature most highly correlated to the CCSD amplitude, and the percentage of amplitudes selected from the initial amplitude space was treated as a hyperparameter to be optimized to reduce the error of the model.

3.7 Results and Discussion

3.7.1 Comparison of CMO and LMO Models

Table 3.1 shows the MAE for models trained with CMOs and LMOs. These DDCC models are grouped as "individual models" (i.e., models that are trained with data from a specific molecule and tested on conformers of the same molecule) and as "mixed models", where heterogeneous data from all five molecules are used for training. These MAEs show that there is a clear gain in both accuracy and transferability from the use of localized orbitals. All individual and mixed models showed an increase in accuracy by an order of magnitude when LMOs were introduced in DDCC. For example, the average MAE of the individual and mixed models with CMOs is 8.82 and 10.21 mE_h, respectively, while the corresponding deviations from the LMO models are only 0.69 and 1.04 mE_h, respectively. LMOs also demonstrate increased transferability since the difference between mixed and individual models is 0.35 mE_h, while the same difference with CMOs is 1.39 mE_h. We also note that the deviations from LMO individual and mixed models are comparable, apart from divinyl ether (difference of 0.90 mE_h).

3.7.2 Evaluation of the Amplitude Selection Schemes

The bar plot of Figure 3.4 shows the distribution of all amplitudes (black bars) as well as 20% of the amplitudes selected from the SB, CB, LA, PS, and EC schemes (yellow bars). These amplitudes were calculated with the STO-3G basis set and using LMOs. In addition to the above-mentioned schemes, a distribution for 20% of randomly selected data points is also shown in Figure 3.4 (e) for comparison with the data driven selection schemes. The SB, CB, LA, PS, and EC schemes select almost all the amplitudes with considerably larger magnitudes as shown in Figure 3.4 (a), (b), (c), (d), and (f) respectively, whereas the random model omits a significant number of amplitudes with larger magnitudes.

Table 3.1: Average MAE (in mE_h) of DDCC Models Trained with Canonical or Localized Orbitals and with Data from the Same Molecule (Individual Model) or with Data from All Five Molecules (Mixed Model)

Molecule	Canonical Orbitals		Localized Orbitals	
	Individual Model	Mixed Model	Individual Model	Mixed Model
Diethyl ether	11.90	12.41	0.67	0.77
Divinyl ether	11.86	11.80	1.08	1.98
Vinyl alcohol	4.22	7.17	0.34	0.46
Methyl vinyl ether	8.36	10.34	0.76	1.29
Propenol	7.77	9.34	0.60	0.69
Average	8.82	10.21	0.69	1.04

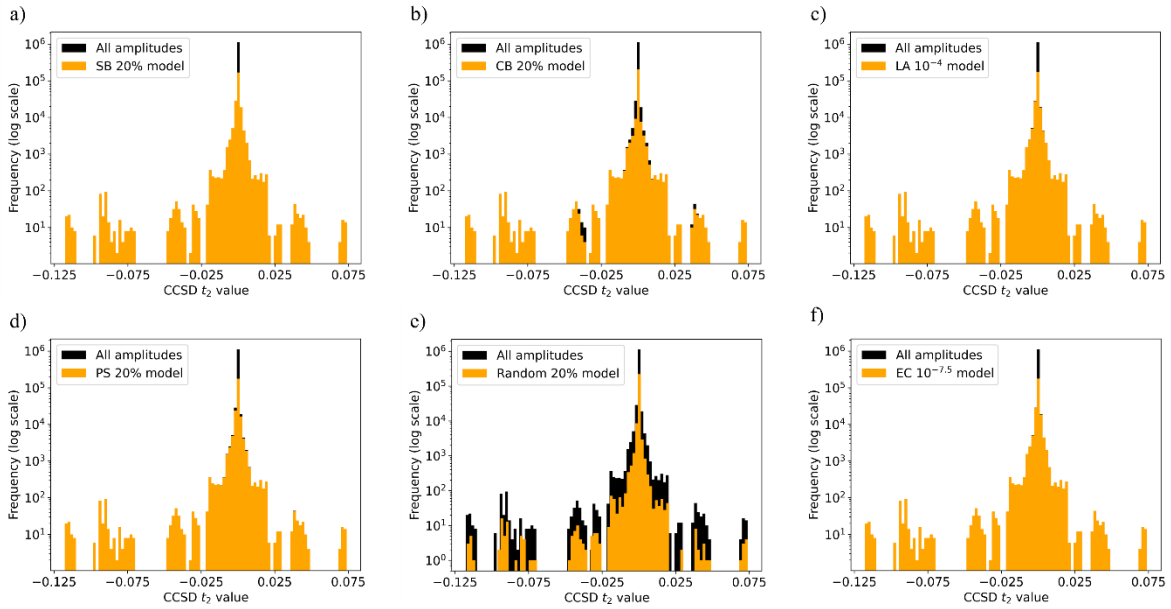

Figure 3.4: Distribution of selected CCSD t_2 amplitudes (yellow) compared with the distribution of the total amplitude space for (a) SB, (b) CB, (c) LA, (d) PS, (e) random, and (f) EC schemes.

A comparison of MAE for the calculation of energy for test conformers with different amplitude selection schemes with LMOs is shown in Figure 3.5. It is evident that the CB and PS models provide an accuracy below 1 kcal/mol with only 5% of the total amplitudes. Models with the LA approach provide an accuracy below 1 kcal/mol error with 9.81% of the total amplitudes. Both SB and randomly selected amplitude models have an accuracy below 1 kcal/mol with 20% of the total amplitudes. Overall, with LMOs, all schemes achieve less than 1 kcal/mol accuracy with a significantly lower number of amplitudes. Interestingly, the CB, LA, and random-amplitude models converge to the accuracy of the all-amplitude model as the number of training data points systematically approaches the full amplitude space (100%). SB models approach a plateau at around 60% of the total amplitude. We would also like to note that the PS model with LMOs has the best performance across all different amplitude percentages, from 5 to 90% (vide infra). On the contrary, all CMO models with amplitude selection failed to achieve below 1 kcal/mol accuracy (see Supporting Information, Section S-VII, in Appendix I). An important observation is the behavior of the LA models, which achieve significantly less error (6.95 mE_h) than the all-amplitude model (10.19 mE_h) at 18% of the total amplitudes, but the error gradually increases and approaches asymptotically the all-amplitude model when the number of amplitudes increases.

Next, we performed an analysis to investigate the effect of the number of bins on the accuracy of SB and CB models with LMO formalism. Results of this analysis are listed in the Supporting Information, Section S-VIII in Appendix I. The best number of bins for SB models was 300 × 300 bins and that for CB models was 50 bins. These results will be used for the calculations with the cc-pVDZ basis set.

The behavior of the LA models shows an interesting pattern. Errors from LA models are large at low data percentages (less than 10% of the total data points used for training). This large

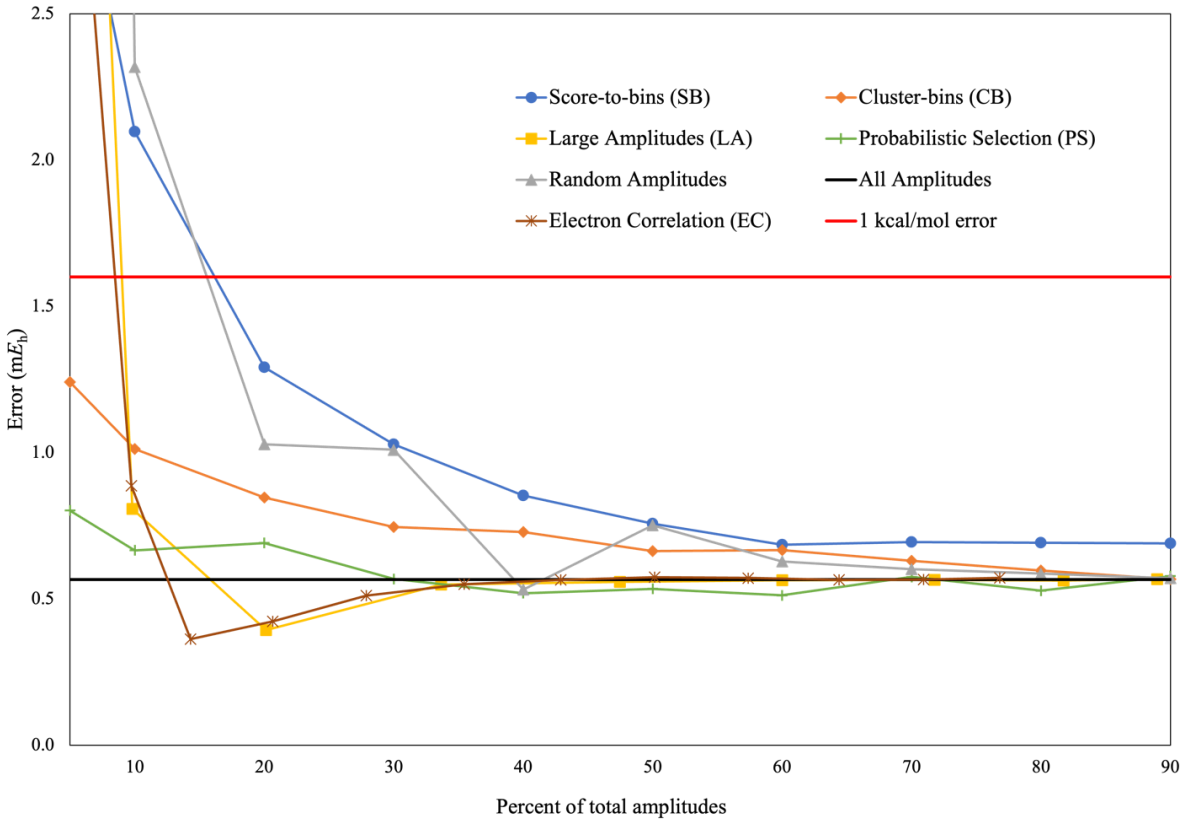

Figure 3.5: MAE vs percentage of amplitudes selected for SB (blue), CB (orange), LA (yellow), EC (brown), and PS (green) schemes. MAE for the random selection approach was also depicted in gray. CCSD amplitudes for training and testing were calculated using the STO-3G basis set and LMO formalism.

error can be attributed to neglecting contributions from a large number of amplitudes for the correlation energy calculation. The accuracy is increased when more amplitudes are added (15%). Upon incremental addition above 20%, the errors start to increase and asymptotically reach the error of the allamplitude model. Thus, with 15% of the total number of amplitudes, we avoid overfitting of the DDCCSD/LA model. This is visually shown in Figure 3.3, where the R^2 value between all MP2 and CCSD amplitudes is 0.9330, which increases to 0.9943 when only the large amplitudes are considered.

The EC scheme shows behavior similar to that of the LA scheme. At low percentages, the error is significantly high due to a lack of an adequate number of amplitudes to account for the accurate estimation of the total correlation energy.

The EC scheme approaches the lowest error of 0.363 mE_h with only 14.3% (when the cutoff is 1×10^{-7} mE_h) of the total amplitudes. Upon further addition of amplitudes, the MAE increases asymptotically due to overfitting and eventually approaches the all-amplitude model error when 42.9% (cutoff is 1×10^{-9} mE_h) of the total amplitudes are used.

Models trained with PS schemes at low percentages (between 5 and 20%) predict the CCSD energy with greater accuracy than the other schemes considered in this study. Further addition of amplitudes above 30% does not deviate from the accuracy of the all-amplitude model, and the highest accuracy for the PS scheme is at 60% (0.523 mE_h). The PS scheme has the lowest range of variation of MAE, thus, it is the most consistent scheme out of all the amplitude selection schemes we have used for STO-3G.

3.7.3 Basis Set Effects

Next, we explored basis set effects by keeping the same hyperparameters (e.g., number of bins and amplitude percentage) as they were obtained from the smaller STO-3G basis set. For example, the percentages for the CB and SB models were 5% and 20%, respectively, since those were the lowest percentage values that provided MAEs below 1 kcal/mol. The number of bins used in the CB and SB models was 50 and 300 × 300, respectively. For the LA model, a cutoff value of 1×10^{-4} was applied since the model with 1×10^{-4} cutoff provides the lowest error for the STO-3G basis set.

MAEs for models trained with SB, CB, LA, EC, and PS schemes calculated with the cc-pVDZ basis set are shown in Figure 3.6, together with the MAE between MP2 and CCSD correlation energies (45.04 mE_h, see bar with label "MP2"). The model trained with 5% of the total amplitude space using the CB approach has an MAE of 11.68 mE_h, while the SB model with 20% of amplitudes has an MAE of 10.06 mE_h. Additional models that utilize larger data percentages of the total amplitudes were trained (CB with 20%, 30% with SB), but the accuracy remained close to or above 10 mE_h (10.55 and 9.941 mE_h for CB and SB, respectively). Since the addition of larger amounts of training data (larger percentages) increases the computational effort without decreasing the model deviations, we are not considering the CB and SB approaches in the following Sections.

As mentioned above, the first model for the LA approach with the cc-pVDZ basis set was trained using the cutoff value for the best LA model for the STO-3G basis set (1×10^{-4}). CCSD energies were predicted with an accuracy of 2.447 mE_h using only 4.65% of the amplitudes with this model. Next, models were trained and tested with the LA approach by using different cutoff values to identify the best value for the ccpVDZ basis set (see Table 3.2,

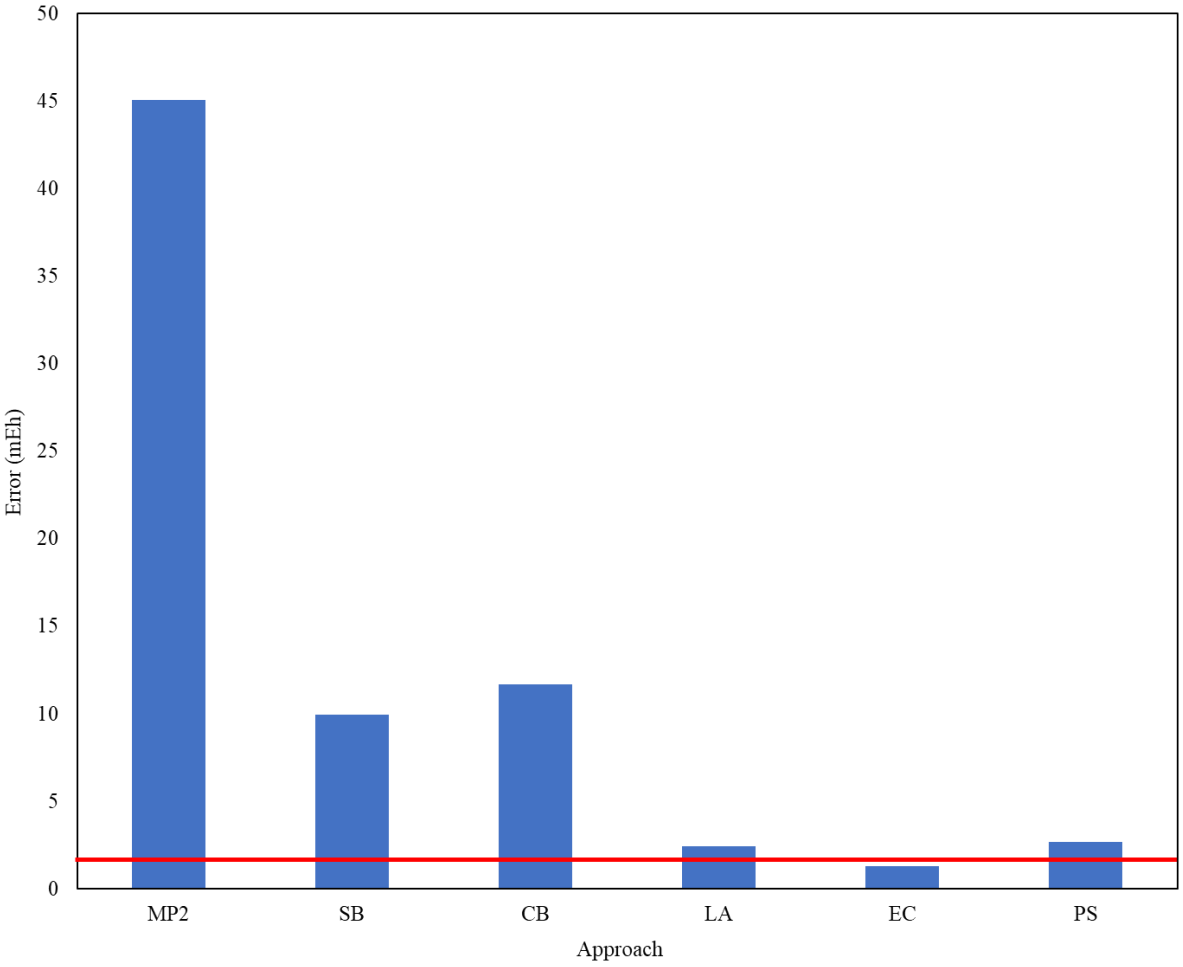

Figure 3.6: Bar plot with the MAEs for models trained with SB, CB, LA, EC, and PS schemes calculated with the cc-pVDZ basis set. The difference between MP2 and CCSD correlation energies is shown as MP2 (left bar). The 1 kcal/mol threshold is indicated by a horizontal line.

left column). As the number of amplitudes is decreased by increasing the cutoff, the MAE increases. A smaller cutoff increases the number of amplitudes for training and testing but the MAE also increases (~8.8 mE_h). Therefore, from this analysis, we found that the accuracy of the models converges at the 1×10^{-4} cutoff. Similar behavior for MAE was observed for STO-3G models (Figure 3.5).

The EC model that was tested with a p_{cutoff} of 99.15% of the total MP2 correlation was able to provide the lowest errors (see the middle column of Table 3.2). This model was able to predict the CCSD energy with an error of 1.288 mE_h.

For the PS approach, the model trained with 60% of the total amplitudes resulted in a large deviation (117.160 mE_h). To further evaluate the performance at lower percentages of amplitudes, we trained models with the PS approach from 1 to 10% (see Table 3.2, right column), and the lowest MAE (2.674 mE_h) was achieved with only 6% of the total number of amplitudes. Therefore, the PS offers a useful roadmap for avoiding model overfitting.

We selected the best scheme from the cc-pVDZ basis set calculations, which is the EC scheme, for the aug-cc-pVDZ basis set calculations. The results for the aug-cc-pVDZ basis set with different cutoff values are summarized in Table S7 of the Supporting Information, Section S-IX in Appendix I. The average difference between MP2 and CCSD correlation energy for the aug-ccpVDZ basis set is 42.532 mE_h. The EC model was trained using 4.47% of the total number of amplitudes (amplitudes with $|e_{ij(MP2)}^{ab}|$ greater than $1 \times 10^{-8} E_h$). The best cutoff for testing the EC model was 98.8%, and the MAE for this cutoff is 1.946 mE_h.

3.8 Conclusions

The DDCC model offers an alternative approach to the acceleration of CC theory. As the system

Table 3.2: MAE (in mE_h) calculated for LA, EC and PS schemes with different cutoff values (cc-pVDZ basis set). Cutoff values for the LA scheme are set for the magnitude of MP2 amplitudes, whereas the cutoff for the EC scheme are set for the correlation energy of each datapoint. For the PS scheme, cutoffs are the percentages of datapoints selected from the total amplitude space. The percentage of amplitudes selected for training EC scheme models were fixed at 5.76% of the total number of amplitudes.

LA Scheme			EC Scheme		PS scheme	
Cutoff	MAE	Percent amplitudes	p_{cutoff}	MAE	Cutoff (percent)	MAE
1×10^{-2}	493.563	0.003	99.1%	1.406	2%	5.253
1×10^{-3}	112.333	0.330	99.15%	1.288	4%	3.027
1×10^{-4}	2.447	4.654	99.2%	1.370	6%	2.674
1×10^{-5}	8.837	21.801	99.25%	1.565	8%	6.850
1×10^{-6}	8.754	49.024	99.3%	1.953	10%	10.501

and basis set sizes increase, the number of training data that are needed increases and eventually becomes the bottleneck for the application of DDCC to larger molecular systems. The main objective of this study was the t_2 amplitude space truncation, an important task that can reduce the required quantum chemical data and training time, while also increasing model accuracy. For this purpose, we introduced five amplitude selection schemes that can be grouped into two categories. The first category includes "data-driven" schemes such as the SB, the CB, and the PS. The second category contains two schemes that are based on chemical arguments and in particular on the two-electron excitation parameter t_2 (LA) or the two-electron excitation energy $e_{ij(MP2)}^{ab}$ (EC). Results obtained from the STO-3G basis set allowed us to explore the behavior and accuracy of these schemes as well as to refine model hyperparameters. We found that models that utilized 15% of the total amplitude set together with the LA, EC, and PS schemes were able to provide similar or better accuracy when compared with the model trained with all available amplitudes. These three schemes were also able to provide reasonable accuracy when used together with the data obtained from the cc-pVDZ basis set, but further assessment with the larger aug-cc-pVDZ basis set revealed that only the EC scheme was able to provide satisfactory results. Finally, our initial hypothesis of reduction of overfitting by amplitude space truncation was proven, since errors start to increase after reaching a minimum for all basis sets when LA and PS approaches were used. We are currently exploring physics-based approaches that can be coupled together with the amplitude selection schemes discussed here that can further increase the applicability of DDCC schemes. Another future direction is the exploration of electron correlation effects far from the equilibrium geometries, to develop models with improved transferability for strongly correlated systems.

Chapter 4 DDCCNet: Physics-informed multitask leaning approach to data-driven coupled-cluster

4.1 Disclosure

This section contains on going work on implementation of neural networks in DDCC. P. D. Varuna S. Pathirage developed three versions of data-driven coupled-cluster neural networks (DDCCNet) under the guidance of Professor Konstantinos D. Vogiatzis.

4.2 Abstract

Data-driven couped-cluster singles and doubles (DDCCSD) method aims at accelerating CCSD calculations using MP2 level electronic structure data with machine learning. In the current work we have introduced multitasking and physics-informed learning to DDCCSD via neural networks. Data-driven coupled-cluster neural networks (DDCCNet) models predict both t_1 and t_2 amplitudes simultaneously using multitask learning. They also improve the portability of the models compared to DDCCSD random forest or k-nearest neighbors models. In a comparison with DDCCSD random forest model, DDCCNet models showed significant increase in accuracy and portability, which demonstrates the effectiveness of multitasking learning and physics-informed workflows.

4.3 Introduction

A variety of methods were developed to capture the correlation energy that are missing from the Hartree-Fock (HF) theory such as configuration interaction (CI) theory, coupled-cluster (CC) theory, perturbation theory and density functional theory (DFT). Although these methods give more accurate description to chemical systems, due to the computational expense, applicability to larger chemical systems are restricted.

Many methods have being developed to accelerate these calculations such as resolution of the identity,[2] local correlation methods,[3-9] accelerated convergence techniques,[10, 11] explicitly-correlated methods,[12, 13] tensor decomposition,[14, 15] dual-basis methods,[16] parallel computing with GPU acceleration[17, 18] and machine learning.[19-22]

Machine learning (ML) has gained increasing popularity in the field of quantum chemistry due the recent development of computational infrastructure. Recently, Aldossary et al[23] categorized machine learning applications in quantum chemistry into four sections based on the amount of physics that are included in the ML workflows. The first category trains models with large amounts of data to predict properties of chemical systems. The second type of applications are machine learning potentials and force fields. These models map the interatomic interactions and generate potential energy surfaces to approach *ab initio* level accuracy. The third category uses machine learning to predict quantum chemical intermediates and parameters to accelerate the calculations. The fourth category of machine learning applications in quantum chemistry are *ab initio* methods that build the wave function using machine learning.

The data-driven coupled cluster singles and doubles (DDCCSD) methods[37] developed by Townsend and Vogiatzis, aims at predicting the CC singles and doubles (CCSD) two-electron excitation (t_2) amplitudes with MP2 level electronic structure data and machine learning. These predicted t_2 amplitudes have two major applications. The first application is using them directly to calculate the CCSD correlation energy. The second application is using them as initial guess for the iterative coupled cluster solver instead of MP2 amplitudes. In the first case we can obtain a better approximation to the CCSD energy than MP2 energy at the cost of MP2. The second approach aims at reducing the number of iterations, hence accelerating the CCSD convergence to obtain exact CCSD correlation energy. The first approach of DDCCSD can be categorized as under the fourth category in the above-mentioned categorization scheme since

we are building the wave function for the system using machine learning whereas the second approach can be categorized under the third category since we aim at accelerating the CCSD calculations.

Although the initial DDCCSD models were able to predict the t_2 amplitudes and gain acceleration for solving CCSD iteratively, application to larger chemical systems and complexed basis sets was hindered by the amount of training data and the distribution of the amplitudes. Next, we developed five schemes to select amplitudes for training and testing.[39] Out of the five schemes electron correlation (EC) scheme had the highest accuracy and large amplitude (LA) and probabilistic selection (PS) schemes were the next most accurate models. We were able to achieve below 1 kcal/mol accuracy for predicting energy for 25 conformers from five organic molecules that have three to five non-hydrogen atoms by coupling EC scheme with localized molecular orbital (LMO) implementation. In the initial DDCCSD models k-nearest neighbors algorithm was used. Due to testing time for k-nearest neighbors increasing exponentially with the number of data points, it is infeasible to use for larger systems. Therefore, for the study on amplitude selection schemes we used random forest algorithm. Although we were able to achieve greater accuracies at a lesser computational expense, we still have limitations with the random forest algorithm. The first limitation is the portability of the models. In order to achieve greater accuracies with random forest, we have to use models that require extensive amounts of memory. The next limitation we had with the random forest models is the difficulty to implement multitask learning to predict one-electron and two-electron amplitudes simultaneously. During the process of determining the solving the CCSD amplitudes, t_1 and t_2 amplitudes are being predicted simultaneously. Implementation of such a model with random forest is difficult.

A better solution for this is the use of neural networks. "Accurate NeurAl networK engINe for Molecular Energies" (ANI) family of methods[22, 31, 70, 71] provides a great example for portable neural network machine learning models. Due to the flexibility in designing the models, multitask learning can be easily implemented with neural networks.[92-94] Another advantage is the ability to implement physics-informed leaning with neural networks.[95, 96] Therefore, in the current study we developed three versions of neural networks that use multitask learning to predict t_1 and t_2 amplitudes simultaneously. We collectively called the **D**ata-**D**riven **C**oupled-**C**luster neural **Net**works (DDCCNet). DDCCNet version 3 also implements physics informed learning by predicting the intermediates of the CC iterative solver. We use a set of loss functions to implement the physics-informed learning and multitask learning that will be discussed in detail under the theoretical background section.

From a comparison done using methanol molecule conformers we have seen that there is a significant improvement in energy prediction using DDCCNet models compared to DDCCSD random forest model which shows the dominance of the neural network models over the random forest models.

The organization of this article is as follows: introduction to CC theory and DDCCSD, architecture of the DDCCNet models and loss functions are presented in theoretical background section. The Computational Details section contains information about data generation and train and test conformers. A comparison of DDCCSD random forest model and DDCCNet models are discussed under results and discussion section followed by concluding remarks.

4.4 Theoretical Background

4.4.1 Residual coupled-cluster equations

In this section we will discuss the residual t_1 and t_2 amplitude equations.

A variational solution to the coupled-cluster equations for the determination of the t_1 and t_2 amplitudes is non-trivial. Therefore, they are determined by solving a set of non-liner equations derived from projected CC equations (equations 2.13 and 2.14).

For the ease of computation and storage, residual CC t_1 and t_2 amplitudes are calculated by determining a set of intermediates as shown in equations 4.1 and 4.2.[97]

$$\Delta t_i^a D_i^a = f_{ia} + \sum_e t_i^e F_{ae} + \sum_{me} t_{im}^{ae} F_{me} - \sum_{nf} t_n^f \langle na \| if \rangle - \frac{1}{2} \sum_{mef} t_{im}^{ef} \langle ma \| ef \rangle - \frac{1}{2} \sum_{men} t_{mn}^{ae} \langle nm \| ei \rangle \quad 4.1$$

$$\begin{aligned} \Delta t_{ij}^{ab} D_{ij}^{ab} = {} & \langle ij \| ab \rangle + P_-(ab) \sum_e t_{ij}^{ae} \left(F_{be} - \frac{1}{2} \sum_m t_m^b F_{me} \right) \\ & - P_-(ij) \sum_m t_{im}^{ab} \left(F_{mj} - \frac{1}{2} \sum_e t_j^e F_{me} \right) + \sum_{mn} \tau_{mn}^{ab} W_{mnij} \\ & + \sum_{ef} \tau_{ij}^{ef} \langle ab \| ef \rangle - P_-(ab) \sum_m t_m^a Z_{mbij} \\ & + P_-(ij) P_-(ab) \sum_{me} \left(t_{im}^{ae} W_{mbej} + t_i^e t_m^a \langle mb \| ej \rangle \right) \\ & + P_-(ij) \sum_e t_i^e \langle ab \| ej \rangle - P_-(ab) \sum_m t_m^a \langle mb \| ij \rangle \end{aligned} \quad 4.2$$

In the above equations, Δt_i^a is the residual term for t1 amplitudes and Δt_{ij}^{ab} is the residual term for t_2 amplitudes. The terms D_i^a and D_{ij}^{ab} are the difference of energies for occupied and virtual orbitals given as:

$$D_i^a = \varepsilon_i - \varepsilon_a \tag{4.3}$$

$$D_{ij}^{ab} = \varepsilon_i + \varepsilon_j - \varepsilon_a - \varepsilon_b, \tag{4.4}$$

where ε_p is the energy of the p^{th} spin orbital.

For the calculation of residual t_1 and t_2 amplitudes, we need to derive a set of intermediates. The expressions for those intermediates are given below.

$$F_{ae} = (1 - \delta_{ae}) f_{ae} - \frac{1}{2}\sum_m f_{me} t_m^a + \sum_{mf} t_m^f \langle ma \| fe \rangle - \frac{1}{2}\sum_{mnf} \tilde{\tau}_{mn}^{af} \langle mn \| ef \rangle \tag{4.5}$$

$$F_{mi} = (1 - \delta_{mi}) f_{mi} + \frac{1}{2}\sum_e f_{me} t_i^e + \sum_{en} t_n^e \langle mn \| ie \rangle + \frac{1}{2}\sum_{nef} \tilde{\tau}_{in}^{ef} \langle mn \| ef \rangle \tag{4.6}$$

$$F_{me} = f_{me} + \sum_{nf} t_n^f \langle mn \| ef \rangle \tag{4.7}$$

$$W_{mnij} = \langle mn \| ij \rangle + P_-(ij) \sum_e t_j^e \langle mn \| ie \rangle + \frac{1}{4}\sum_{ef} \tau_{ij}^{ef} \langle mn \| ef \rangle \tag{4.8}$$

$$\begin{aligned} W_{mbej} = \langle mb \| ej \rangle + \sum_f t_j^f \langle mb \| ef \rangle \\ - \sum_n t_n^b \langle mn \| ej \rangle - \sum_{nf} (\frac{1}{2} t_{jn}^{fb} + t_j^f t_n^b) \langle mn \| ef \rangle \end{aligned} \tag{4.9}$$

$$Z_{mbij} = \sum_{ef} \tau_{ij}^{ef} \langle mb \| ef \rangle \tag{4.10}$$

$$\tau_{ij}^{ab} = t_{ij}^{ab} + t_i^a t_j^b - t_i^b t_j^a \tag{4.11}$$

$$\tilde{\tau}_{ij}^{ab} = t_{ij}^{ab} + \frac{1}{2}(t_i^a t_j^b - t_i^b t_j^a) \tag{4.12}$$

The permutation operator $P_{-}(ij)$ used above can be defined as,

$$P_{\pm}(pq) = 1 \pm \mathcal{P}(pq) \tag{4.13}$$

$\mathcal{P}(pq)$ permutates p and q.

Additionally, there is another intermediate calculated during the process of determining Δt_{ij}^{ab} that arises due to the spin-factorization of W_{mbej}.

$$W_{mbej} = -\langle mb||je\rangle + \sum_f t_j^f \langle mb||fe\rangle - \sum_n t_n^b \langle mn||je\rangle - \sum_{nf} (\frac{1}{2} t_{jn}^{fb} + t_j^f t_n^b)\langle mn||fe\rangle \tag{4.14}$$

We use these intermediates as targets for each of the blocks when implementing physics informed learning which will be discussed in the next Section.

4.4.2 Data-driven coupled-cluster neural networks (DDCCNet)

In the current study, we have developed three different types of neural networks to predict CCSD t_1 and t_2 amplitudes based on the MP2 level-electronic structure data. Main goal of these neural networks is to implement multitask learning in predicting the t_1 and t_2 amplitudes. When solving the CC iterative equations, the t_1 and t_2 amplitudes are calculated simultaneously. This is the reason behind using multitask leaning. We collectively term these neural networks DDCCNet.

The first version consists of two separate linear networks to predict the t_1 and t_2 amplitudes. These networks are being updated using a combined loss function to implement the interdependency of these models. Second version in the second version of DDCCNet, the feature vector is partitioned into four sections and four different linear networks are being used to predict t_1 and t_2 amplitudes. In the third version of DDCCNet we are using separate linear networks to predict the intermediates of t_1 and t_2 amplitudes. Later we combine them to predict the t_1 and t_2 amplitudes. The architectures, hyperparameters and the loss functions that were used are described in the following sections.

DDCCNet version1 (DDCCNet_v1) - Multitask Evaluation

Consists of two separate neural networks with seven linear layers to predict the t_1 and t_2 amplitudes separately (Figure 4.1). Each hidden layer consists of 196 nodes.

A set of 30 features were used for t_2 prediction and 14 features were used for t_1 amplitude prediction.

DDCCNet version2 (DDCCNet_v2)

For the DDCCNet version 2 we change the T1 and T2 block with two neural networks that has the architecture shown in Figure 4.2. Here for the linear layer block, we use seven linear layers with 196 nodes in each hidden layer.

For these models the feature space is partitioned into four sections.

- **Orbital-based features:** Orbital energies, One electron, Coulomb and Exchange contribution to orbital energies

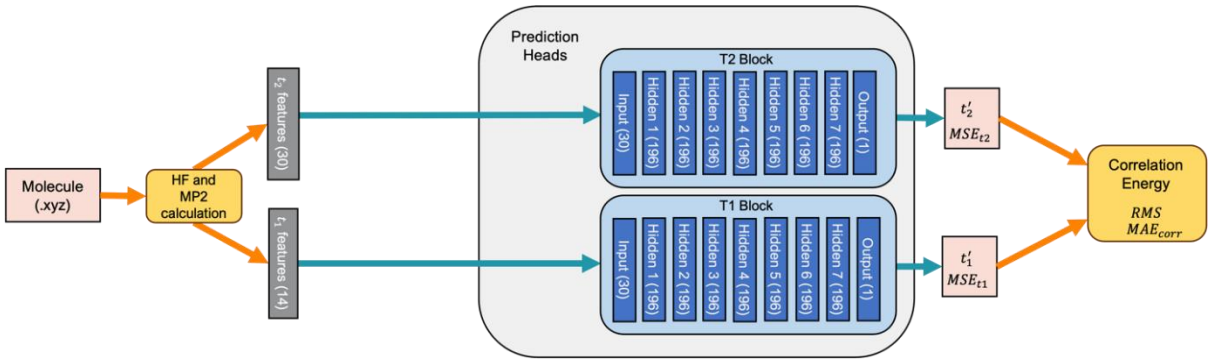

Figure 4.1: Architecture of DDCCNet version 1.

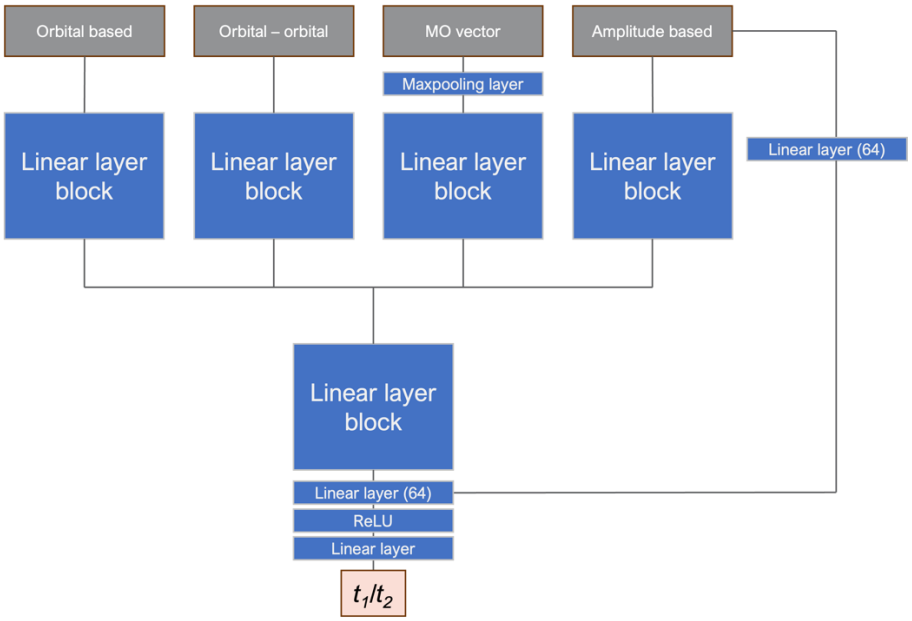

Figure 4.2: Architecture of t_1 and t_2 block of the DDCCNet version 2.

- **Orbital-orbital related features:** Coulomb and exchange integral between occupied and virtual orbitals and log values of their magnitudes
- **MO vector:** Atomic orbital coefficients for the molecular orbitals. Here a maxpooling layer is used to keep the input for the model constant.
- **Amplitude based:** MP2 amplitude, Difference of orbital energies, two electron-integral, sign of the MP2 amplitude, whether the electrons are excited to the same virtual orbital and log value of the magnitude of the MP2 amplitude

Features of each separate section goes through separate linear block. Then the output from each linear block is combined (summed) and passed through a combined block with architecture similar to the linear block. Here we include the amplitude-based features as a skip connection since those are the features with a higher correlation with the amplitude values.

DDCCNet version3 (DDCCNet_v3)

The major difference of this version of DDCCNet is the implementation of physics-informed learning.

For the DDCCNet version 2 we change the T1 and T2 block with two neural networks that has the architecture shown in Figure 4.3. Here for the linear layer block, we use seven linear layers with 196 nodes in each hidden layer.

In these models we use separate blocks to determine the contribution to the amplitudes by each intermediate (equations 4.5, 4.6, 4.7, 4.8, 4.9, 4.10, 4.11, and 4.14) calculated in the coupled cluster iterations. Then those values are being combined and passed through amplitude prediction block to obtain predictions for t_1 or t_2 amplitudes.

In the T2 prediction head for DDCCNet version 3, we have five intermediate blocks for the intermediates calculated for the calculation of the t_2 amplitudes. In the T1 prediction head we have included three intermediate blocks.

For DDCCNet version 3 we use features that are highly correlated with the amplitudes instead of the 30-dimensional features.

For the T2 prediction head and for the prediction for t_{ij}^{ab} amplitude we use,

- t_{ij}^{ab} MP2 amplitude
- $\langle ij||ab\rangle$: two electron-integral
- Reciprocal of the orbital energy differences
- $t_i^a t_j^b$: multiplication of the t1 amplitudes. These t_1 amplitudes are calculated with MP2 t_2 amplitudes
- $t_{ij}^{ab} + t_i^a t_j^b$

For the T1 prediction head we use three features to predict t_1 amplitude.

- t_i^a value calculated with MP2 t_2 amplitudes.
- Reciprocal of the orbital energy differences of occupied and virtual orbital.
- f_i^a: Fock matrix element

Loss Functions

For DDCCNet versions 1 and 2 use three loss functions combined to update the models. The first type loss function accesses the t_1 and t_2 values of the individual prediction blocks (Equations 4.15, 4.16, 4.17, and 4.18). Loss functions in equation 4.19 and 4.20 are the ones

that combine both t_1 and t_2 amplitudes. These two loss functions serve as the convergence criteria for the CCSD iterative solver.

$$MSE_{t1} = \frac{\sum(t_1 - t_1')^2}{n_{t1}} \tag{4.15}$$

$$MSE_{t2} = \frac{\sum(t_2 - t_2')^2}{n_{t2}} \tag{4.16}$$

$$MAE_{t1} = \frac{\sum|t_1 - t_1'|}{n_{t1}} \tag{4.17}$$

$$MAE_{t2} = \frac{\sum|t_2 - t_2'|}{n_{t2}} \tag{4.18}$$

$$RSS = \sqrt{\sum(t_1 - t_1')^2 + \sum(t_2 - t_2')^2} \tag{4.19}$$

$$MAE_{corr} = |E_{corr} - E'_{corr}| \tag{4.20}$$

Here t_1' and t_2' are the predicted t_1 and t_2 amplitudes and t_1 and t_2 are the exact CCSD t_1 and t_2 amplitudes. Total number of t_1 and t_2 amplitudes in a batch are denoted by n_{t1} and n_{t2} respectively.

For DDCCNet version 1 we use three loss functions each for the T1 and T2 blocks:

The loss function for T1 block is,

$$Total\ Loss_{t1}^{v1} = MSE_{t1} + RSS + MAE_{corr} \tag{4.21}$$

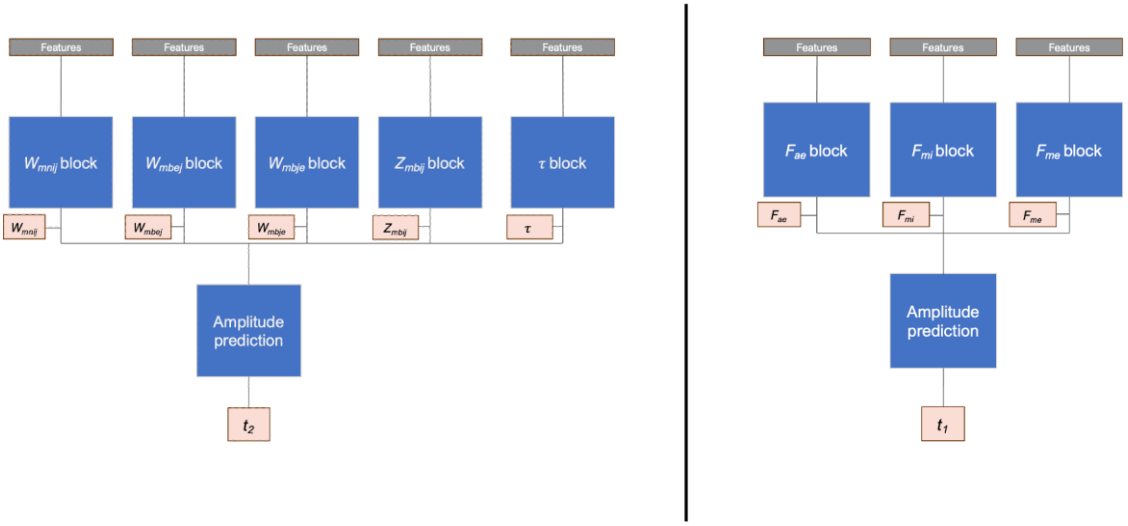

Figure 4.3: Architecture of DDCCNet_v3, where a) shows the architecture of the t_2 prediction head and b) shows the architecture of the t_1 prediction head.

Similarly, the loss function for T2 block of DDCCNet version 1 is,

$$Total\ Loss_{t2}^{v1} = MSE_{t2} + RSS + MAE_{corr} \tag{4.22}$$

However, upon hyperparameter search we determined that best loss functions for DDCCNet version 2 are MAE_{t1} and MAE_{t2} instead of MSE_{t1} and MSE_{t2} respectively. Therefore, the total loss functions for t_1 and t_2 are,

$$Total\ Loss_{t1}^{v2} = MAE_{t1} + RSS + MAE_{corr} \tag{4.23}$$

$$Total\ Loss_{t2}^{v2} = MAE_{t2} + RSS + MAE_{corr} \tag{4.24}$$

Since we predict the outcome of each individual output from intermediate blocks for DDCCNet version 3 with the aim of making the model physics-informed, we add additional loss functions to the T1 and T2 blocks.

$$MSE_I = \frac{\sum(I - I')^2}{n_I} \tag{4.25}$$

Here I denotes the value of the intermediate and I' denotes the predicted value of the intermediate. The number of intermediates per batch is denoted by n_I

The total loss function for T1 and T2 blocks can be denoted as follows.

$$Total\ Loss_{t1}^{v3} = MSE_{t1} + RSS + MAE_{corr} + \sum MSE_{I}^{t1} \quad 4.26$$

$$Total\ Loss_{t2}^{v3} = MSE_{t2} + RSS + MAE_{corr} + \sum MSE_{I}^{t2} \quad 4.27$$

We are using the ReLU activation function and 10^{-6} as the learning rate for all the DDCCNet versions.

Large amplitude scheme

Due to the large number of amplitudes generated for training and testing and the distribution of the amplitudes, we faced two major difficulties with the initial DDCCSD models. The first problem we faced was the difficulty to handle large amounts of data and when training and storing machine learning models and the second problem is the overfitting of machine learning models due to large number of amplitudes clustered closer to zero. As a solution to these problems we have developed five data selection techniques in our previous work.[39] The most accurate technique among them was the electron correlation (EC) scheme. In this scheme we use a cutoff to the magnitude of the MP2 electron correlation captured by an amplitude. The second most accurate model was the large amplitude (LA) scheme. In LA scheme instead of applying the cutoff to the electron correlation, the cutoff is applied to the magnitude of the MP2 amplitude. Due to the ease of implementation of the LA scheme compared to EC scheme, we have used LA scheme for the current study.

We also use the symmetry of the amplitudes,

$$t_{ij}^{ab} = t_{ji}^{ba} \quad 4.28$$

By using this relationship, we can reduce the number of amplitudes for training and testing from n_{t2} to $\frac{n_{t2}+\sqrt{n_{t2}}}{2}$ per molecule before applying LA scheme.

4.5 Computational Details

The Psi4Numpy[56] software package was used to perform quantum chemical calculations and feature and target extraction. The first part of this study compares the performance of DDCCSD random forest models and three versions of DDCCNet. The training set for this study consists of 50 conformers of methanol and the test set consists of 25 conformers of methanol. The models were trained and tested using localized molecular orbitals (LMOs) and cc-pVDZ basis set was used.

The hyperparameters for the random forest model used in the comparison were determined in our previous study.[39] We use 350 estimators (trees), with mean square error loss function and minimum sample split of two. We use Sci-kit learn package[57] to train the random forest models and Pytorch[98] to train DDCCNet models.

For testing the accuracy of the DDCCNet architectures, here we train models to predict the energies of CO_2 dimers. In this case we use 50 conformers of CO_2 dimers to train the models and test with 25 conformers of CO_2 dimers. These dimers were created with LAMMPS[99] software at 300 K temperature using GROMOS_54A7[100, 101] forcefield. We use cc-pVDZ basis set and for this study and due to the large number of t_2 amplitudes, we use LA scheme with a cutoff value of 1×10^{-4} discussed in the previous section to select amplitudes for training and

testing. In order to determine the transferability of the DDCCNet models we tested the models trained with the above mentioned 50 CO_2 dimers with five CO_2 trimer conformers generated using the LAMMPS software and GROMOS_54A7 forcefield at 300K.

4.6 Results and Discussion

4.6.1 Comparison of DDCCSD random forest and DDCCNet models

The first part of our study is to compare the results of DDCCSD random forest model with different DDCCNet versions. Here we compare the mean absolute errors of the CC correlation energy as given by the following equation.

$$MAE = \frac{\sum_{n=1}^{n} |E_{CCSD}^{corr} - E_{DDCCSD}^{corr}|}{n} \qquad 4.29$$

For this study we used 5000 epochs for training each DDCCNet model.

Figure 4.4 (a) shows the mean absolute error for the CC correlation energy of conformers of methanol molecules calculated with DDCCSD random forest model and three versions of DDCCNet. There is a clear gain in accuracy when using DDCCNet models. The train and test mean absolute error for the prediction of energy using DDCCSD random forest model were 7.561 mE_h and 7.629 mE_h respectively. This is more than 25 times greater than the next model DDCCNet_v1. The train and test mean absolute error for DDCCNet_v1 were 0.273 mE_h and 0.251 mE_h respectively. When the DDCCNet net models were compared (Figure 4.4 (b)), DDCCNet_v3 model is the most accurate model. It has a slightly more accurate mean absolute error for training and testing compared to the other two versions. Mean absolute error of DDCCNet_v3 for testing is 0.198 mE_h which is slightly more accurate than DDCCNet_v2 (0.229 mE_h) and DDCCNet_v1(0.251 mE_h). According to the spread of data shown in Figure

4.4 (b) DDCCNet_v2 has a lower spread for train and test errors than for DDCCNet_v1 and DDCCNet_v3. Therefore, DDCCNet_v2 can be attributed as the most consistent method out of the three DDCCNet models.

Next, we wanted to compare the performance of the DDCCNet_v3 with the number of epochs (Figure 4.4 (c)). Here we see, when the number of epochs increases it can be seen that the test mean absolute error drops to 0.149 mE_h and the spread of mean absolute error decreases. Therefore, we can see that the accuracy and the precision of the models can be improved by increasing the number of epochs.

As the next step DDCCNet was used to predict energies of CO_2 molecules to access the applicability and the transferability of DDCCNet.

4.6.2 CO_2 dimers

In order to study the accuracy of the DDCCNet models for the chemical systems with intermolecular interactions, we next applied the models to CO_2 dimers. The process of obtaining the structures and training the models were discussed under computational details.

For this study we used DDCCNet versions 2 and 3.

Table 4.1 shows the comparison of test and train errors of DDCCNet version. DDCCNet version 2 predicts the energy with a greater accuracy than version3. Version 2 achieves 1 kcal/mol chemical accuracy while version 3 has errors slightly above 1 kcal/mol. The conformers with greater energies have been identified as either having C-C distance more than 4.0 Å or elongated C-O bond distances.

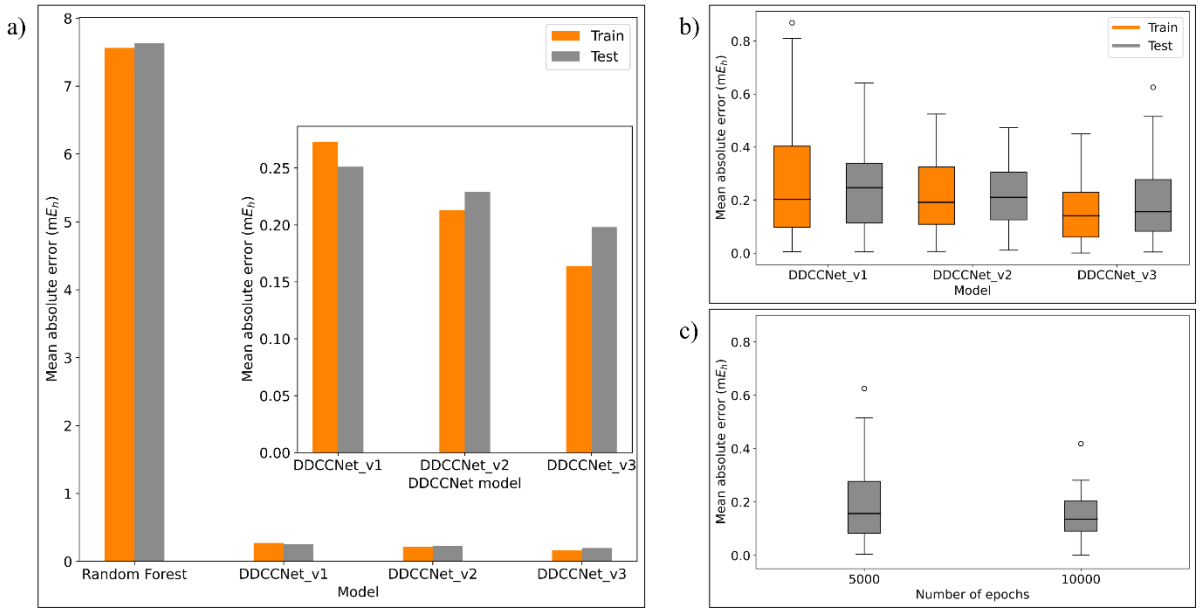

Figure 4.4: Results for comparing the performance of different machine learning models in predicting mean absolute error (in mEh). a) Bar plot that compares the train and test error for DDCCSD random forest and DDCCNet models. The inset shows a comparison of DDCCNet versions 1, 2, and 3. b) A box plot that compares train and test error of DDCCNet versions. c) A box plot that compares the mean absolute error of DDCCNet_v3 when number of epochs are increased.

Table 4.1: Mean absolute errors and R^2 values calculated for train and test conformers of CO_2 dimers. A set of 50 CO_2 dimer conformers were used for training and 25 conformers used for testing.

	Mean absolute error (mE_h)		**R^2**	
Model	**Train**	**Test**	**Train**	**Test**
DDCCNet_v2	1.270	1.470	0.783	0.761
DDCCNet_v3	1.370	1.839	0.800	0.715

Next, we tested the transferability of the models by using the same models trained with 50 dimer conformers to test accuracy of five CO_2 trimer conformers. The results are shown in Figure 4.5. The accuracy for version 3 model is significantly increased when the triples were considered. The mean absolute error for correlation energy for the trimer conformers calculated using DDCCNet version 2 was 4.630 mE_h while that for version 3 was 0.902 mE_h. From these results, it can be inferred that the transferability of the DDCCSD version 3 models is greater than version 2 models.

4.7 Conclusions

From the results we can see a clear improvement in accuracy with the use of multitask physics-informed DDCCNet models. This can be attributed to the flexibility offered by neural networks in terms of the architecture, hyperparameters and the loss functions. There is a 25-fold increase in the accuracy when using DDCCNet models compared to DDCCSD random forest models. We also have shown that DDCCNet models can predict the correlation energies for systems with intermolecular interactions with the chemical accuracy of 1 kcal/mol using CO_2 dimers. Another significant observation is the transferability of the DDCCNet version 3 when predicting the correlation energies for CO_2 trimers achieving accuracy beyond 1 mE_h.

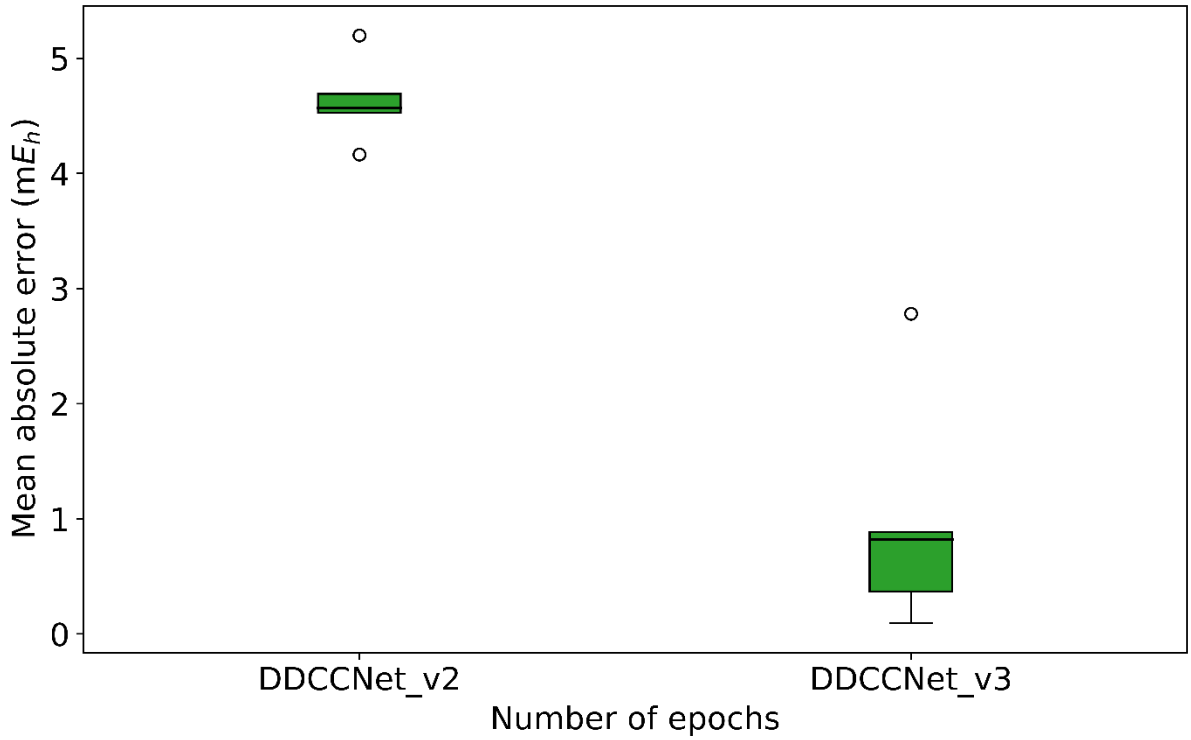

Figure 4.5: Boxplot that indicates the accuracy for testing with five conformers of CO2 trimers. The models were trained with 50 conformers of CO2 dimers.

Chapter 5 Perturbative triples approximation and prediction of two electron excitation amplitudes for CCSDT

5.1 Disclosure

This section contains on going work in developing DDCCSD perturbative triples and DDCCSD exact triples methods. P. D. Varuna S. Pathirage developed the code to calculate the perturbative triples correlation energy from the predicted amplitudes under the guidance of Professor Konstantinos D. Vogiatzis. Features and exact t_2 amplitudes from CCSDT/aug-cc-pVDZ level of theory for the development of DDCCSDT model were provided to us by Dr. Run Li and Professor Eugene DePrince III from Florida State University. The machine learning models were developed by P. D. Varuna S. Pathirage under the guidance of Professor Konstantinos D. Vogiatzis.

5.2 Abstract

DDCCSD method aims at predicting the t_1 and t_2 amplitudes at the cost of MP2 using machine learning. Here we present two approaches to approximate CCSD(T) and CCSDT energies using the data-driven approach. The DDCCSD(T) approach uses predicted amplitudes to calculate the perturbative energy. We have achieved less than 1 kcal/mol accuracy for CCSD(T) energy predictions. The DDCCSDT approach aims at predicting t_2 amplitudes for CCSDT by using CCSDT t_2 amplitudes as the target. Therefore, we focus at approximating CCSDT energy at the cost MP2.

5.3 Introduction

Coupled-cluster singles and doubles with perturbative triples [CCSD(T)][59] is considered the gold standard of quantum chemistry due to the accuracy and the efficiency of the method. Perturbative triples term is a non-iterative approximation to the CC singles, doubles and triples (CCSDT). The triples perturbation energy is calculated using CCSD t_1 and t_2 amplitudes.

Perturbative triples implementation in DDCCSD will be discussed in detail in the theoretical background section.

In coupled cluster theory, the correlation energy is calculated as in equation 3.1 for any level of truncation because the Hamiltonian operator is a two-particle operator (see Chapter Chapter 2). Therefore, we can calculate the energies for CCSD and higher levels of truncations if their respective t_2 amplitudes are known. Here we also present a method to predict the t_2 amplitudes for CCSDT hence the CCSDT correlation energy can be determined.

5.4 Theoretical Background

5.4.1 Data-driven coupled-cluster singles and doubles with perturbative triples (DDCCSD(T))

The CCSD(T) correlation energy can be calculated with equation 5.1.

$$E_{(T)} = \frac{1}{36} \sum_{ijkabc} t_{ijk}^{abc}(c) D_{ijk}^{abc} \left[t_{ijk}^{abc}(c) + t_{ijk}^{abc}(d) \right] \qquad 5.1$$

Here the term D_{ijk}^{abc} is the difference between occupied (*i, j, k*) and virtual (*a, b, c*) orbitals that can be given by

$$D_{ijk}^{abc} = \varepsilon_i + \varepsilon_j + \varepsilon_k + \varepsilon_a + \varepsilon_b + \varepsilon_c, \qquad 5.2$$

where ε_p is the energy of spin orbital energy *p*.

Terms $t_{ijk}^{abc}(c)$ and $t_{ijk}^{abc}(d)$ denote connected and disconnected triples amplitudes for the perturbative triples term (T) respectively, which can be calculated using equations 5.3 and 5.4.

$$D_{ijk}^{abc} t_{ijk}^{abc}(c) = P(i/jk)P(a/bc)\left[\sum_{e} t_{jk}^{ae}\langle ei\|bc\rangle - \sum_{m} t_{im}^{bc}\langle ma\|jk\rangle\right] \quad 5.3$$

$$D_{ijk}^{abc} t_{ijk}^{abc}(\mathrm{d}) = P(i/jk)P(a/bc)t_{i}^{a}\langle jk\|bc\rangle \quad 5.4$$

The permutation operator $P(p/qr)$ acting on function $f(pqr)$ can be denoted as,

$$P(p/qr)f(pqr) = f(pqr) - f(qpr) - f(rqp) \quad 5.5$$

In the current work we calculate the $E_{(T)}$ energy using the predicted amplitudes. We use the same 30 features as discussed in Chapter Chapter 3 for the prediction of t_2 amplitudes. Once the t_2 amplitudes are predicted one CCSD iteration is performed to obtain the t_1 amplitudes. These t_1 and t_2 amplitudes are then being used to calculate DDCCSD(T) correlation energy.

5.5 Computational Details

All quantum chemical calculations and feature extraction were performed using Psi4Numpy[56] package and machine learning was performed using Scikit-learn[57] library for python. For training the models to predict DDCCSD(T) energy, 20 conformers of water molecules were used and tested with 30 conformers of water molecules. The bond angle for the water molecules were fixed at 109.5° and the bonds for both train and test molecules were stretched from 0.8 – 1.8 Å. Three basis sets were used for this study: STO-3G,[82, 83] 6-31G[79, 80] and cc-pVDZ.[84] The k-nearest neighbor machine learning algorithm was used.

5.6 Results and Discussion

5.6.1 Predicting DDCCSD(T) energy

Figure 5.1 shows the dissociation curves for water calculated with CCSD, CCSD(T), DDCCSD and DDCCSD(T) methods and the comparison between $E_{(T)}$ and data-driven perturbative triples error ($\delta E_{DD(T)}$) for STO-3G, 6-31G and cc-pVDZ basis sets.

From the dissociation curves, we can observe that the DDCCSD and DDCCSD(T) energies agree well with the CCSD and CCSD(T) energies around the equilibrium bond distance, while the energies deviate from the exact methods as the bond distance increases. Another important observation is the deviations of DDCCSD and DDCCSD(T) methods show same behavior especially when the energies deviate from their exact counterparts. The reason behind this the dependence of the (T) energy on the t_1 and t_2 amplitudes.

In table 5.1 we list the errors from DDCCSD, DD(T), and percentage error for DD(T). Although the error for (T) increased from STO-3G to 6-31G to cc-pVDZ, percentage error for DD(T) decreased from STO-3G to cc-pVDZ. This can be explained by the comparison of the $E_{(T)}$ and the $\delta E_{DD(T)}$. Since the amount of the (T) energy captured by the basis sets increase from STO-3G to 6-31G to cc-pVDZ. This causes the percentage error to decrease although the error is increased.

5.7 Conclusions

DDCCSD(T) and DDCCSDT methods provide evidence to the fact that DDCC methods can expand further to estimate the energies with more accurate methods using machine learning. Results from the DDCCSD(T) show that the percent accuracy of the models increase with the

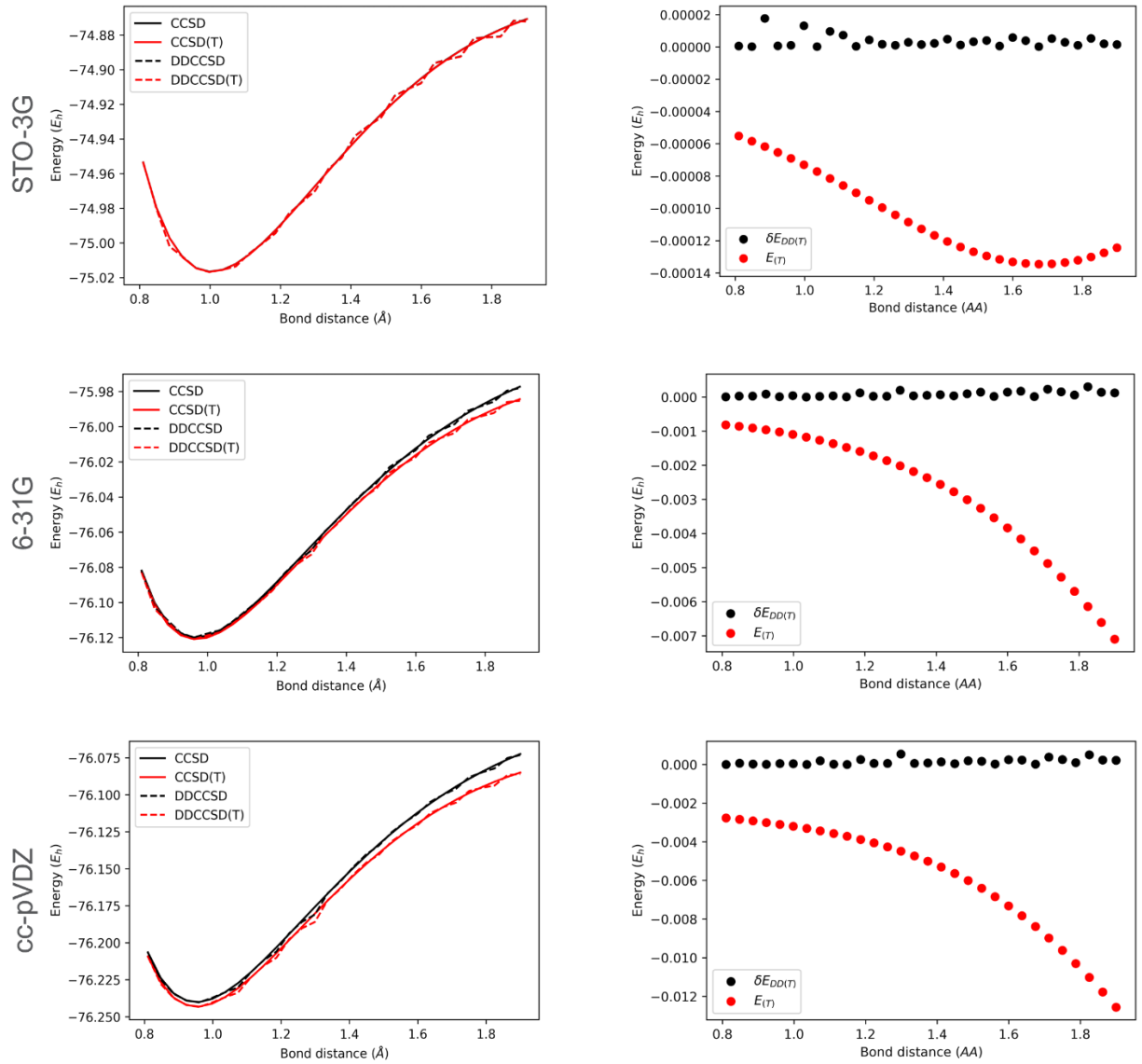

Figure 5.1: The dissociation curves and comparison of perturbative triples energy $E_{(T)}$ and error for perturbative energy $\delta E_{DD(T)}$.

Table 5.1: The DDCCSD mean absolute errors, DD(T) mean absolute errors and percentage (T) error calculated for 30 conformers of water molecule for STO-3G, 6-31G, and cc-pVDZ basis sets.

Basis set	**DDCCSD error (mE_h)**	**(T) Error (mE_h)**	**Percent (T) Error**
STO-3G	1.552	0.004	3.9%
6-31G	0.826	0.081	2.9%
cc-pVDZ	0.928	0.140	2.3%

increasing complexity of the basis set. Another observation from this study is that the accuracy of the DDCCSD(T) depends on the accuracy of the t_2 amplitudes predicted. Therefore, although k-nearest neighbors algorithm provides a proof of concept, DDCCSD(T) has potential to improve with the more accurate models such as DDCCNet.

Another advantage of DDCC methods is that MP2-level data can be used to predict t_2 amplitudes for any level of truncation (doubles, triples, quadruples, etc.), hence can be used to approximate energy for any level of truncation for CC due to the fact that CC correlation energy can be calculated using only the t_1 and t_2 amplitudes.

Chapter 6

Conclusions

In our current work we show how we address the limitations of DDCCSD method. First, by implementing LMOs to DDCCSD we were able to increase the accuracy of the models by ten-fold. This can be attributed to the local nature of electronic excitations. Next, we were able to achieve greater accuracies at lesser training cost using five amplitude selection techniques. There we observed EC, LA and PS models performed significantly better than the SB and CB models. They also performed better than the all-amplitude model at percentages as low as 10-40%. This can be attributed towards the ability of the techniques to prevent overfitting of the models.

Then by introducing neural networks, we were able to implement multitask learning to DDCCSD. These DDCCNet models were able to predict both t_1 and t_2 amplitudes simultaneously. A comparison with the DDCCSD random forest models showed that DDCCNet models perform significantly better. In DDCCNet version 3 we were able to implement physics-informed learning to the DDCCNet.

Finally, we expand the scope of DDCC models by introducing DDCCSD(T) and DDCCSDT approximation methods. There we showed that we can achieve the chemical accuracy for (T) calculations at the cost of MP2. Another key takeaway is that we can approximate CC energy at any level of truncation starting from the singles and doubles using DDCC method due since only the t_1 and t_2 amplitudes are required to calculate the CC correlation energy.

Bibliography

1. Heitler, W.; London, F., Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik. *Zeitschrift für Physik* **1927,** *44* (6), 455-472.

2. Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R., RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. *Chem. Phys. Lett.* **1998,** *294* (1-3), 143-152.

3. Werner, H.-J.; Schütz, M., An efficient local coupled cluster method for accurate thermochemistry of large systems. *J. Chem. Phys.* **2011,** *135* (14).

4. Schütz, M.; Hetzer, G.; Werner, H.-J., Low-order scaling local electron correlation methods. I. Linear scaling local MP2. *The Journal of chemical physics* **1999,** *111* (13), 5691-5705.

5. Schütz, M., Low-order scaling local electron correlation methods. III. Linear scaling local perturbative triples correction (T). *The Journal of Chemical Physics* **2000,** *113* (22), 9986-10001.

6. Schütz, M.; Werner, H.-J., Low-order scaling local electron correlation methods. IV. Linear scaling local coupled-cluster (LCCSD). *J. Chem. Phys.* **2001,** *114* (2), 661-681.

7. Schütz, M., Low-order scaling local electron correlation methods. V. Connected triples beyond (T): Linear scaling local CCSDT-1b. *J. Chem. Phys.* **2002,** *116* (20), 8772-8785.

8. Neese, F.; Hansen, A.; Liakos, D. G., Efficient and accurate approximations to the local coupled cluster singles doubles method using a truncated pair natural orbital basis. *J. Chem. Phys.* **2009,** *131* (6).

9. Riplinger, C.; Neese, F., An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013,** *138* (3), 034106.

10. Pulay, P., Convergence acceleration of iterative sequences. The case of SCF iteration. *Chem. Phys. Lett.* **1980,** *73* (2), 393-398.

11. Kudin, K. N.; Scuseria, G. E.; Cances, E., A black-box self-consistent field convergence algorithm: One step closer. *J. Chem. Phys.* **2002,** *116* (19), 8255-8261.

12. Hattig, C.; Klopper, W.; Kohn, A.; Tew, D. P., Explicitly correlated electrons in molecules. *Chem. Rev.* **2012,** *112* (1), 4-74.

13. Kong, L.; Bischoff, F. A.; Valeev, E. F., Explicitly correlated R12/F12 methods for electronic structure. *Chem. Rev.* **2012,** *112* (1), 75-107.

14. Hohenstein, E. G.; Parrish, R. M.; Martínez, T. J., Tensor hypercontraction density fitting. I. Quartic scaling second-and third-order Møller-Plesset perturbation theory. *Journal of chemical physics* **2012,** *137* (4), 044103.

15. Parrish, R. M.; Hohenstein, E. G.; Martínez, T. J.; Sherrill, C. D., Tensor hypercontraction. II. Least-squares renormalization. *J. Chem. Phys.* **2012,** *137* (22).

16. Liang; Head-Gordon, M., Approaching the basis set limit in density functional theory calculations using dual basis sets without diagonalization. *J. Phys. Chem. A* **2004,** *108* (15), 3206-3210.

17. Ufimtsev, I. S.; Martinez, T. J., Quantum chemistry on graphical processing units. 1. Strategies for two-electron integral evaluation. **2008,** *4* (2), 222-231.

18. Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X., Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015,** *113* (2), 184-215.

19. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A., Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012,** *108* (5), 058301.

20. Chmiela, S.; Tkatchenko, A.; Sauceda, H. E.; Poltavsky, I.; Schütt, K. T.; Müller, K.-R., Machine learning of accurate energy-conserving molecular force fields. *Science advances* **2017,** *3* (5), e1603015.

21. Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R., Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **2017,** *8* (1), 872.

22. Behler, J.; Parrinello, M., Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007,** *98* (14), 146401.

23. Aldossary, A.; Campos-Gonzalez-Angulo, J. A.; Pablo-García, S.; Leong, S. X.; Rajaonson, E. M.; Thiede, L.; Tom, G.; Wang, A.; Avagliano, D.; Aspuru-Guzik, A., In silico chemical experiments in the Age of AI: From quantum chemistry to machine learning and back. *Adv. Mater.* **2024**, 2402369.

24. Jones, G. M.; Smith, B. A.; Kirkland, J. K.; Vogiatzis, K. D., Data-driven ligand field exploration of Fe (iv)–oxo sites for C–H activation. *Inorganic Chemistry Frontiers* **2023,** *10* (4), 1062-1075.

25. Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A., What is high-throughput virtual screening? A perspective from organic materials discovery. *Annual Review of Materials Research* **2015,** *45*, 195-216.

26. Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S., Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry–the Harvard Clean Energy Project. *Energy Environmental Science* **2014,** *7* (2), 698-704.

27. Wang, T.; Tan, X.; Wei, Y.; Jin, H., Accurate bandgap predictions of solids assisted by machine learning. *Materials Today Communications* **2021,** *29*, 102932.

28. Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A., Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **2019,** *10* (1), 5316.

29. Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A., Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017,** *8* (1), 13890.

30. Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R., Schnet–a deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018,** *148* (24).

31. Smith, J. S.; Isayev, O.; Roitberg, A. E., ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017,** *8* (4), 3192-3203.

32. Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F., OrbNet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *J. Chem. Phys.* **2020,** *153* (12).

33. Zheng, P.; Zubatyuk, R.; Wu, W.; Isayev, O.; Dral, P. O., Artificial intelligence-enhanced quantum chemical method with broad applicability. *Nat. Commun.* **2021,** *12* (1), 7022.

34. Bosia, F.; Zheng, P.; Vaucher, A.; Weymuth, T.; Dral, P. O.; Reiher, M., Ultra-fast semi-empirical quantum chemistry for high-throughput computational campaigns with Sparrow. *J. Chem. Phys.* **2023,** *158* (5).

35. Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R.; Maurer, R. J., Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019,** *10* (1), 5024.

36. Gastegger, M.; McSloy, A.; Luya, M.; Schütt, K. T.; Maurer, R. J., A deep neural network for molecular wave functions in quasi-atomic minimal basis representation. *J. Chem. Phys.* **2020,** *153* (4).

37. Townsend, J.; Vogiatzis, K. D., Data-driven acceleration of the coupled-cluster singles and doubles iterative solver. *J. Phys. Chem. Lett.* **2019,** *10* (14), 4129-4135.

38. Townsend, J.; Vogiatzis, K. D., Transferable MP2-based machine learning for accurate coupled-cluster energies. *J. Chem. Theory Comput.* **2020,** *16* (12), 7453-7461.

39. Pathirage, P. D. V. S.; Phillips, J. T.; Vogiatzis, K. D., Exploration of the Two-Electron Excitation Space with Data-Driven Coupled Cluster. *J. Phys. Chem. A* **2024,** *128* (10), 1938–1947.

40. Hermann, J.; Schätzle, Z.; Noé, F., Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry* **2020,** *12* (10), 891-897.

41. Han, J.; Zhang, L.; Weinan, E., Solving many-electron Schrödinger equation using deep neural networks. *J. Comput. Phys.* **2019,** *399*, 108929.

42. Jones, G. M.; Pathirage, P. V. S.; Vogiatzis, K. D., Data-driven acceleration of coupled-cluster and perturbation theory methods. In *Quantum Chemistry in the Age of Machine Learning*, Dral, P. O., Ed. Elsevier: 2023; pp 509-529.

43. Townsend, J.; Kirkland, J. K.; Vogiatzis, K. D., Post-Hartree-Fock methods: configuration interaction, many-body perturbation theory, coupled-cluster theory. In *Mathematical Physics in Theoretical Chemistry*, House, J. E., Ed. Elsevier: 2019; pp 63-117.

44. Lagaris, I. E.; Likas, A.; Fotiadis, D. I., Artificial neural network methods in quantum mechanics. *Comput. Phys. Commun.* **1997,** *104* (1-3), 1-14.

45. Sugawara, M., Numerical solution of the Schrödinger equation by neural network and genetic algorithm. *Comput. Phys. Commun.* **2001,** *140* (3), 366-380.

46. Manzhos, S., Machine learning for the solution of the Schrödinger equation. *Mach. learn.: sci. technol.* **2020,** *1* (1), 013002.

47. Carleo, G.; Troyer, M., Solving the quantum many-body problem with artificial neural networks. *Science* **2017,** *355* (6325), 602-606.

48. Yang, P.-J.; Sugiyama, M.; Tsuda, K.; Yanai, T., Artificial neural networks applied as molecular wave function solvers. *J. Chem. Theory Comput.* **2020,** *16* (6), 3513-3529.

49. McGibbon, R. T.; Taube, A. G.; Donchev, A. G.; Siva, K.; Hernández, F.; Hargus, C.; Law, K.-H.; Klepeis, J. L.; Shaw, D. E., Improving the accuracy of Møller-Plesset perturbation theory with neural networks. *J. Chem. Phys.* **2017,** *147* (16).

50. Welborn, M.; Cheng, L.; Miller III, T. F., Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **2018,** *14* (9), 4772-4779.

51. Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F., A universal density matrix functional from molecular orbital-based machine learning: Transferability across organic molecules. *J. Chem. Phys.* **2019,** *150* (13).

52. Margraf, J. T.; Reuter, K., Making the coupled cluster correlation energy machine-learnable. *J. Phys. Chem. A* **2018,** *122* (30), 6343-6348.

53. Chen, Y.; Zhang, L.; Wang, H.; Weinan, W., Ground state energy functional with Hartree–Fock efficiency and chemical accuracy. *J. Phys. Chem. A* **2020,** *124* (35), 7155-7165.

54. Peyton, B. G.; Briggs, C.; D'Cunha, R.; Margraf, J. T.; Crawford, T. D., Machine-learning coupled cluster properties through a density tensor representation. *J. Phys. Chem. A* **2020,** *124* (23), 4861-4871.

55. Ikabata, Y.; Fujisawa, R.; Seino, J.; Yoshikawa, T.; Nakai, H., Machine-learned electron correlation model based on frozen core approximation. *J. Chem. Phys.* **2020,** *153* (18), 184108.

56. Smith, D. G.; Burns, L. A.; Sirianni, D. A.; Nascimento, D. R.; Kumar, A.; James, A. M.; Schriber, J. B.; Zhang, T.; Zhang, B.; Abbott, A. S., Psi4NumPy: An interactive quantum chemistry programming environment for reference implementations and rapid development. *J. Chem. Theory Comput.* **2018,** *14* (7), 3504-3511.

57. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011,** *12*, 2825-2830.

58. Tew, D. P.; Klopper, W.; Helgaker, T., Electron correlation: The many-body problem at the heart of chemistry. *J. Comput. Chem.* **2007,** *28* (8), 1307-1320.

59. Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M., A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **1989,** *157* (6), 479-483.

60. Ochsenfeld, C.; Kussmann, J.; Lambrecht, D. S., Linear-scaling methods in quantum chemistry. In *Reviews in computational chemistry*, Lipkowitz, K. B.; Cundari, T. R., Eds. 2007; Vol. 23, p 1.

61. Riplinger, C.; Neese, F., An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013,** *138* (3).

62. Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F., Sparse maps—A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016,** *144* (2), 024109.

63. Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F., Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013,** *139* (13), 134101.

64. Ma, Q.; Werner, H. J., Explicitly correlated local coupled-cluster methods using pair natural orbitals. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018,** *8* (6), e1371.

65. Vogiatzis, K. D.; Barnes, E. C.; Klopper, W., Interference-corrected explicitly-correlated second-order perturbation theory. *Chem. Phys. Lett.* **2011,** *503* (1-3), 157-161.

66. Stoll, H., The correlation energy of crystalline silicon. *Chem. Phys. Lett.* **1992,** *191* (6), 548-552.

67. Vogiatzis, K. D.; Klopper, W.; Friedrich, J., Non-covalent Interactions of CO2 with Functional Groups of Metal–Organic Frameworks from a CCSD (T) Scheme Applicable to Large Systems. *J. Chem. Theory Comput.* **2015,** *11* (4), 1574-1584.

68. Datta, D.; Gordon, M. S., A massively parallel implementation of the CCSD (T) method using the resolution-of-the-identity approximation and a hybrid distributed/shared memory parallelization model. *J. Chem. Theory Comput.* **2021,** *17* (8), 4799-4822.

69. Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M., Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **2021,** *121* (16), 9759-9815.

70. Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E., Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018,** *148* (24), 241733.

71. Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E., Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019,** *10* (1), 2903.

72. Qiao, Z.; Christensen, A. S.; Welborn, M.; Manby, F. R.; Anandkumar, A.; Miller III, T. F., Informing geometric deep learning with electronic interactions to accelerate quantum chemistry. *Proc. Natl. Acad. Sci. U.S.A.* **2022,** *119* (31), e2205221119.

73. Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R., OrbNet Denali: A machine

learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *J. Chem. Phys.* **2021,** *155* (20), 204103.

74. Dral, P. O., Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **2020,** *11* (6), 2336-2347.

75. Husch, T.; Sun, J.; Cheng, L.; Lee, S. J.; Miller, T. F., Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **2021,** *154* (6), 064108.

76. Han, R.; Rodriguez-Mayorga, M.; Luber, S., A machine learning approach for MP2 correlation energies and its application to organic compounds. *J. Chem. Theory Comput.* **2021,** *17* (2), 777-790.

77. Becke, A. D., Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A: At. Mol. Opt. Phys.* **1988,** *38* (6), 3098.

78. Lee, C.; Yang, W.; Parr, R. G., Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B: Condens. Matter* **1988,** *37* (2), 785.

79. Ditchfield, R.; Hehre, W. J.; Pople, J. A., Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules. *J. Chem. Phys.* **1971,** *54* (2), 724-728.

80. Hehre, W. J.; Ditchfield, R.; Pople, J. A., Self—consistent molecular orbital methods. XII. Further extensions of Gaussian—type basis sets for use in molecular orbital studies of organic molecules. *J. Chem. Phys.* **1972,** *56* (5), 2257-2261.

81. Smith, D. G.; Burns, L. A.; Simmonett, A. C.; Parrish, R. M.; Schieber, M. C.; Galvelis, R.; Kraus, P.; Kruse, H.; Di Remigio, R.; Alenaizan, A., PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **2020,** *152* (18), 184108.

82. Hehre, W. J.; Stewart, R. F.; Pople, J. A., Self-consistent molecular-orbital methods. I. Use of Gaussian expansions of Slater-type atomic orbitals. *J. Chem. Phys.* **1969,** *51* (6), 2657-2664.

83. Hehre, W.; Ditchfield, R.; Stewart, R.; Pople, J., self-consistent molecular orbital methods. iv. use of Gaussian expansions of Slater-type orbitals. Extension to second-row molecules. *J. Chem. Phys.* **1970,** *52* (5), 2769-2773.

84. Dunning Jr, T. H.; Hay, P. J., Gaussian basis sets for molecular calculations. In *Methods of electronic structure theory*, Schaefer, H. F., Ed. Springer: 1977; pp 1-27.

85. Kendall, R. A.; Dunning, T. H.; Harrison, R. J., Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992,** *96* (9), 6796-6806.

86. Ho, T. K. In *Random decision forests*, Proceedings of 3rd international conference on document analysis and recognition, IEEE: 1995; pp 278-282.

87. Agarawal, V.; Roy, S.; Shrawankar, K. K.; Ghogale, M.; Bharathi, S.; Yadav, A.; Maitra, R., A hybrid coupled cluster–machine learning algorithm: Development of various regression models and benchmark applications. *J. Chem. Phys.* **2022,** *156* (1), 014109.

88. Agarawal, V.; Patra, C.; Maitra, R., An approximate coupled cluster theory via nonlinear dynamics and synergetics: The adiabatic decoupling conditions. *J. Chem. Phys.* **2021,** *155* (12), 124115.

89. Patra, C.; Agarawal, V.; Halder, D.; Chakraborty, A.; Mondal, D.; Halder, S.; Maitra, R., A synergistic approach towards optimization of coupled cluster amplitudes by exploiting dynamical hierarchy. *Chem. Phys. Chem.* **2023,** *24* (4), e202200633.

90. Agarawal, V.; Roy, S.; Chakraborty, A.; Maitra, R., Accelerating coupled cluster calculations with nonlinear dynamics and supervised machine learning. *J. Chem. Phys.* **2021,** *154* (4), 044110.

91. Lloyd, S., Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **1982,** *28* (2), 129-137.

92. Queen, O.; McCarver, G. A.; Thatigotla, S.; Abolins, B. P.; Brown, C. L.; Maroulas, V.; Vogiatzis, K. D., Polymer graph neural networks for multitask property learning. *npj Comput. Mater.* **2023,** *9* (1), 90.

93. Yang, Z.; Zhong, W.; Lv, Q.; Chen, C. Y.-C., Multitask deep learning with dynamic task balancing for quantum mechanical properties prediction. *PCCP* **2022,** *24* (9), 5383-5393.

94. Liu, Z.; Lin, L.; Jia, Q.; Cheng, Z.; Jiang, Y.; Guo, Y.; Ma, J., Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *Journal of chemical information modeling* **2021,** *61* (3), 1066-1082.

95. Ji, W.; Qiu, W.; Shi, Z.; Pan, S.; Deng, S., Stiff-pinn: Physics-informed neural network for stiff chemical kinetics. *J. Phys. Chem. A* **2021,** *125* (36), 8098-8106.

96. Raissi, M.; Perdikaris, P.; Karniadakis, G. E., Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019,** *378*, 686-707.

97. Stanton, J. F.; Gauss, J.; Watts, J. D.; Bartlett, R. J., A direct product decomposition approach for symmetry exploitation in many-body methods. I. Energy calculations. *The Journal of Chemical Physics* **1991,** *94* (6), 4334-4345.

98. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A., Automatic differentiation in pytorch. **2017**.

99. Plimpton, S., Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995,** *117* (1), 1-19.

100. Daura, X.; Mark, A. E.; Van Gunsteren, W. F., Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *J. Comput. Chem.* **1998,** *19* (5), 535-547.

101. Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F., Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European biophysics journal : EBJ* **2011,** *40* (7), 843-56.

Appendix

Supplementary information for chapter 3: Exploration of the Two-Electron Excitation Space with Data-Driven Coupled Cluster

S-I. Features used for the prediction of CCSD t_2 amplitudes

List of features used for predicting the CCSD t_2 amplitudes.

1. Energy of the occupied orbital 1 (**Eocc1**)
2. Energy of the occupied orbital 2 (**Eocc2**)
3. Energy of the virtual orbital 1 (**Evir1**)
4. Energy of the virtual orbital 2 (**Evir2**)
5. One-electron contribution to energy of occupied orbital 1 (**Hocc1**)
6. One-electron contribution to energy of occupied orbital 2 (**Hocc2**)
7. One-electron contribution to energy of virtual orbital 1 (**Hvir1**)
8. One-electron contribution to energy of virtual orbital 2 (**Hvir2**)
9. Coulomb contribution to energy of occupied orbital 1 (**Jocc1**)
10. Coulomb contribution to energy of occupied orbital 2 (**Jocc2**)
11. Coulomb contribution to energy of virtual orbital 1 (**Jvir1**)
12. Coulomb contribution to energy of virtual orbital 2 (**Jvir2**)
13. Exchange contribution to energy of occupied orbital 1 (**Kocc1**)
14. Exchange contribution to energy of occupied orbital 2 (**Kocc2**)
15. Exchange contribution to energy of virtual orbital 1 (**Kvir1**)
16. Exchange contribution to energy of virtual orbital 2 (**Kvir2**)
17. Coulomb integral between occupied orbital 1 and virtual orbital 1 (**Jia1**)
18. Coulomb integral between occupied orbital 2 and virtual orbital 2 (**Jia2**)
19. Exchange integral between occupied orbital 1 and virtual orbital 1 (**Kia1**)
20. Exchange integral between occupied orbital 2 and virtual orbital 2 (**Kia2**)
21. Are the two electrons excited to the same molecular orbital. 1 for yes and 0 for no (**diag**)
22. Difference between occupied and virtual orbital energies also is the denominator of the MP2 amplitude equation (**orbdiff**)
23. Two-electron integral between the two occupied and virtual orbitals. It is the numerator of the MP2 amplitude equation (**doublecheck**)
24. MP2 amplitude value (**t2start**)
25. Log10 of the magnitude of the MP2 amplitude (**t2mag**)

26. Log10 of the magnitudes of Coulomb integral between occupied orbital 1 and virtual orbital 1 (**Jia1mag**)
27. Log10 of the magnitudes of Coulomb integral between occupied orbital 2 and virtual orbital 2 (**Jia2mag**)
28. Log10 of the magnitudes of exchange integral between occupied orbital 1 and virtual orbital 1 (**Kia1mag**)
29. Log10 of the magnitudes of exchange integral between occupied orbital 2 and virtual orbital 2 (**Kia2mag**)
30. Sign of the MP2 amplitude. 1 if positive and 0 if negative (**t2sign**)

S-II. Correlation matrix

Weights for the features of the current study were chosen to be the correlation coefficient between each feature and the CCSD t_2 amplitude value. Figure S1 and Figure S2 shows the heatmap of the correlation coefficient matrix for the 30 features and the CCSD t_2 amplitudes calculated for all amplitude model with STO-3G basis set[80, 82]. Figure S1 shows the heatmap of the correlation matrix calculated for localized orbital formalism whereas Figure S2 shows the heatmap of the correlation matrix calculated for canonical orbital formalism. Last row (or last column) was used as the weights. Weights are calculated for each model separately for the respective set of selected amplitudes.

S-III. Distribution of Amplitudes

Figure S3 shows the distribution of the CCSD t_2 amplitudes calculated with STO-3G, cc-pVDZ and aug-cc-pVDZ basis sets. Number of amplitudes are in logarithmic scale.

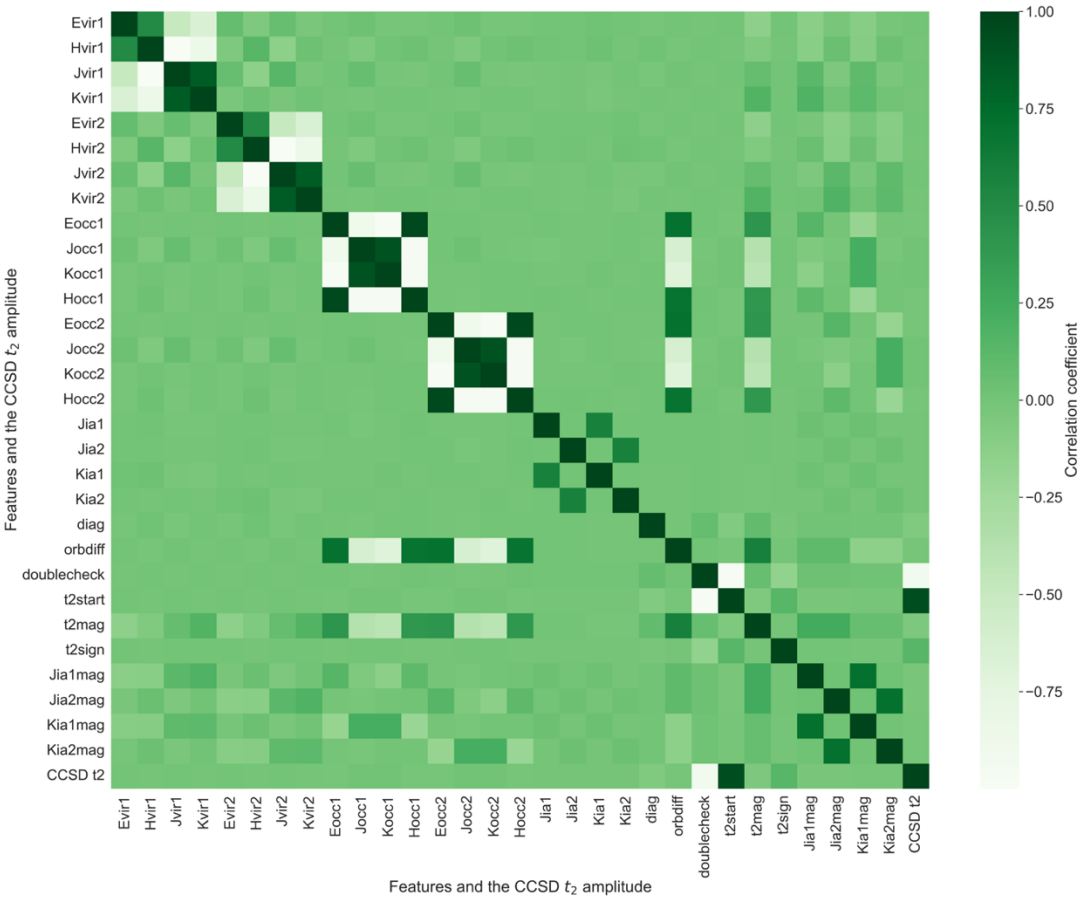

Figure S1: Heatmap for correlation coefficients calculated for the 30 features and CCSD t_2 amplitudes using STO-3G basis set and localized orbital formalism

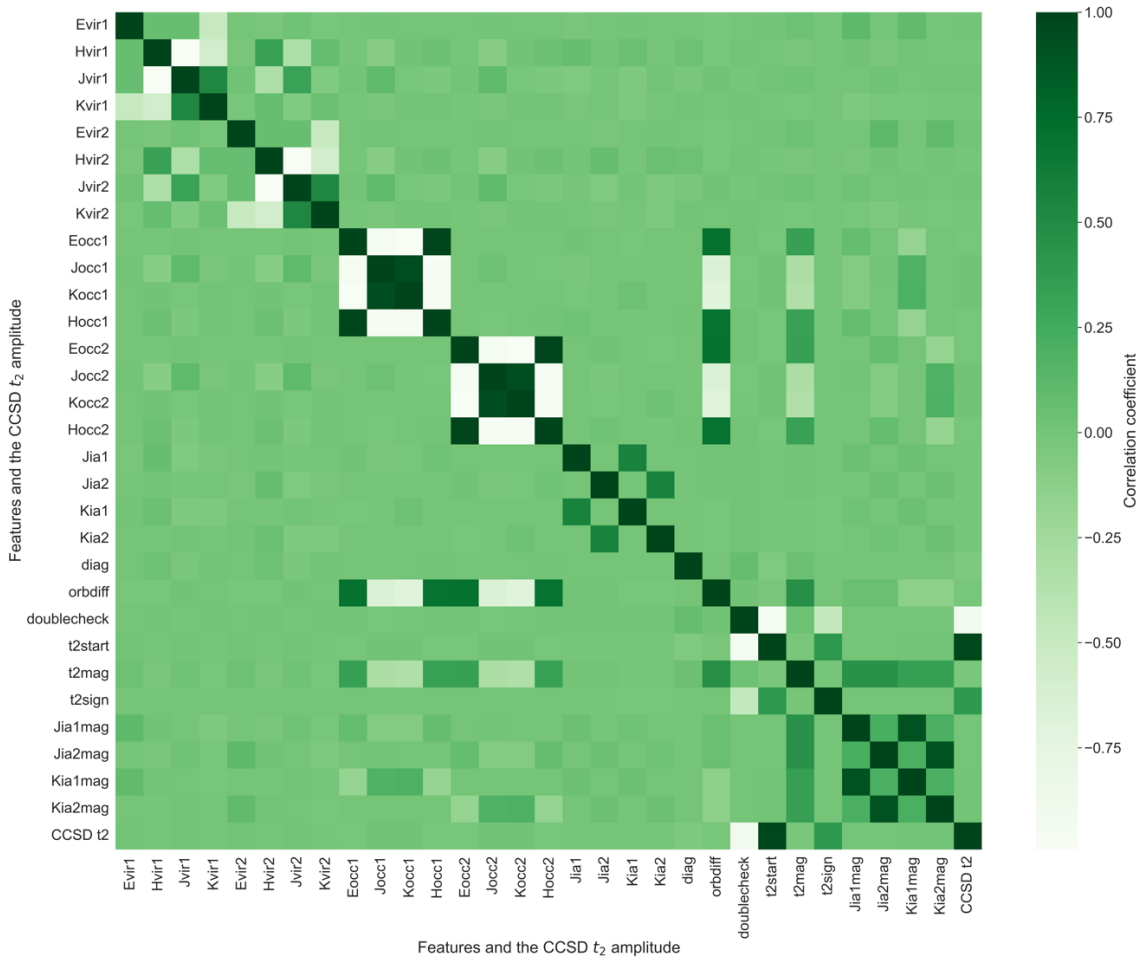

Figure S2: Heatmap for correlation coefficients calculated for the 30 features and CCSD t2 amplitudes using STO-3G basis set and canonical orbital formalism.

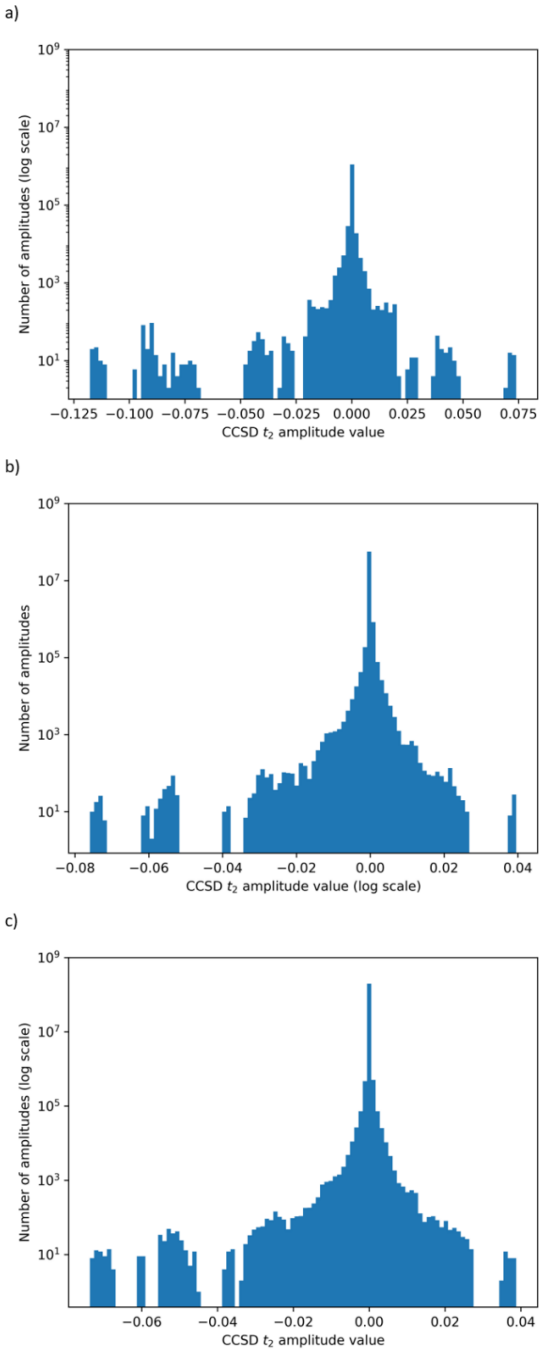

Figure S3: Distribution of CCSD t_2 amplitudes computed with a) STO-3G basis set, b) cc-pVDZ basis set, and c) aug-cc-pVDZ basis set. Number of amplitudes are in log scale.

S-IV. Hyperparameter optimization for random forest models

Three hyperparameters were optimized for training the random forest[86] ML models. These models were trained with Sci-Kit Learn package[57] in python.

I. Number of estimators (decision trees)
II. The function used to calculate the goodness of a split (criterion)
III. Minimum number of data points to split branches of a decision tree (minimum split)

For this hyperparameter optimization, models were trained with 30 water conformers and tested with 100 water conformers. Features and CCSD t_2 amplitudes were calculated with cc-pVDZ basis set.[84]

The first step of hyperparameter optimization was to identify the range of number of estimators where each of the criteria absolute error and squared error performed best with minimum split of 2. Five models with each set of hyperparameters were trained and the average of the mean absolute error (MAE) was used as the criteria for selection of the best model.

Form that data (Table S1), top three number of estimators for each of the two criterions were selected. For squared error best three number of estimators were 300 and 350 and for absolute error best number of estimators were 200 and 350.

Then for each of number of estimators, 7 models were trained with 2, 5, 10, 20, 30, 40, and 50 minimum splits.

According to the results from this hyperparameter optimization results (Table S2), it was concluded that the best hyperparameters were,

Table S1: Average mean absolute error for models trained with 10 to 500 number of estimators for absolute error and squared error criteria. Models were trained and tested five times and the averages of MAEs are used as the criterion for selecting best models. Average MAE for top two models for each criterion, absolute error and squared error are bolded.

Number of estimators	Average CCSD MAE in mE_h	
	Absolute error	Squared error
10	2.1282	2.3982
50	1.8768	1.9034
100	1.8500	1.8720
150	1.7907	2.0455
200	1.7223	1.8293
250	1.8008	1.8184
300	1.8678	1.7388
350	1.7007	1.6720
400	1.7642	1.7826
450	1.8315	1.7860
500	1.8193	1.7616

Table S2: Average mean absolute error for models trained with top three number of estimators for squared error and absolute error criterion. Models were trained and tested five times and the averages of MAEs are used as the criterion for selecting best models. Best model with least average MAE is bolded.

Number of estimators	Criterion	Minimum split	Average MAE in mE_h
200	absolute error	2	1.7223
200		5	2.0216
200		10	3.7048
200		20	10.2964
200		30	13.5084
200		40	15.2616
200		50	16.4807
350		2	1.7007
350		5	2.0058
350		10	3.7490
350		20	10.3690
350		30	13.6601
350		40	15.2419
350		50	16.6475
300	squared error	2	1.7319
300		5	2.1911
300		10	3.5749
300		20	8.9551
300		30	12.9470
300		40	13.4527
300		50	20.3416
350		2	1.6771
350		5	2.1531
350		10	3.5897
350		20	8.9634
350		30	13.0125
350		40	13.6849
350		50	20.0752

I. Number of estimators: 350

II. Criterion: squared error

III. Minimum split: 2

S-V. Separation of amplitudes to bins in score-to-bin scheme

Figure S4 shows the how the amplitudes were distributed for all-amplitude model for STO-3G with localized orbital formalism.

S-VI. Analysis of the effect of percentage of opposite-spin MP2 correlation captured and the accuracy of the EC models

The first step for this analysis is to use a common cutoff (similar to the approach used in LA schemes). For this approach, we trained and tested with cutoffs $1\times10^{-7.25}$ mE_h and $1\times10^{-7.5}$ mE_h and the MAE is calculated for each type of molecule separately (Table S3). For these calculations the same cutoff is used for training and testing and cc-pVDZ basis set is used.

From this analysis it was observed that the error is minimum when the percent of MP2 opposite-spin correlation is between 99.1 to 99.3%. There is an exception for ethenol molecule since it's percent of MP2 opposite-spin correlation captured is greater than 99.5 for both cutoff values.

As the next step, we used the model trained with $1\times10^{-7.5}$ mE_h cutoff and studied the variation of MAE with different cutoffs for percent MP2 opposite-spin correlation captured for the test conformers. Table S5 shows the results for these calculations. According to these results minimum MAE is 1.288 mE_h and the MAE is below 1 kcal/mol (1.59 mE_h) for the range 99.1

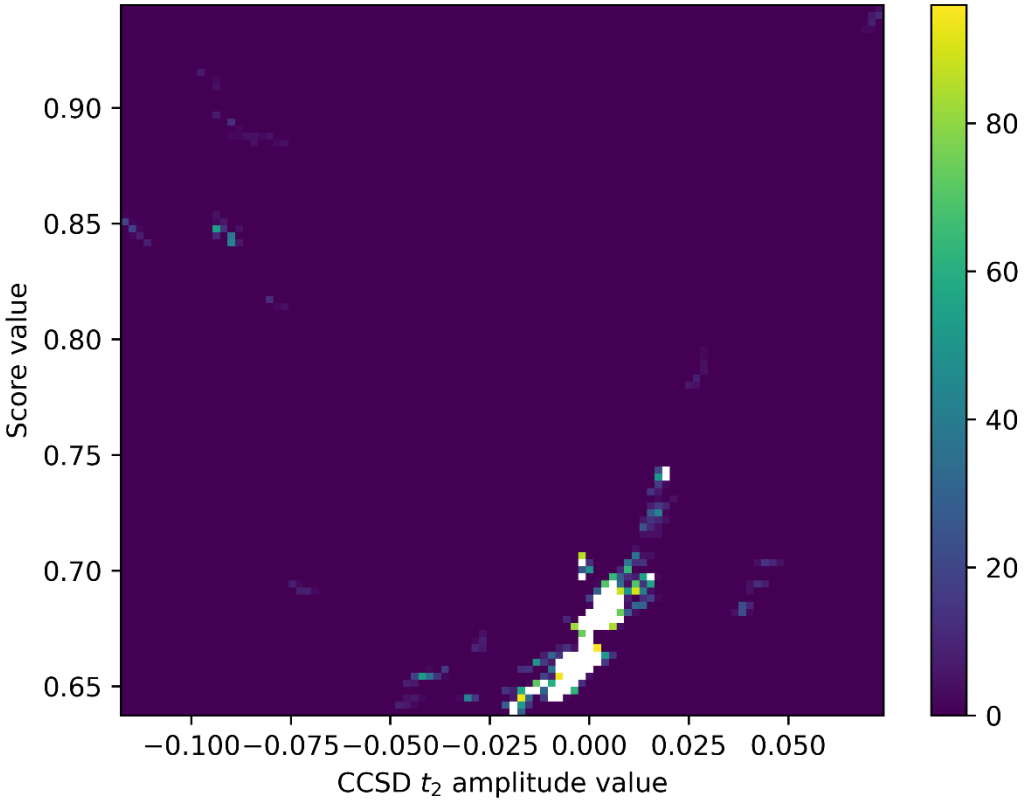

Figure S4: Distribution of the amplitudes calculated with all-amplitude model, to 100×100 bins. These amplitudes were calculated with STO-3G basis set and localized orbital formalism. For the purpose of better visualization, maximum number of amplitudes in a bin was set to 100.

Table S3: The MAE calculated for each type of test molecules separately using cc-pVDZ basis set

Cutoff	Average Error (mE_h)					
	Diethyl ether	Divinyl ether	Ethenol	Methoxyethene	Propenol	Average
$1\times10^{-7.25}$	4.019	0.641	3.779	2.899	0.751	2.418
$1\times10^{-7.5}$	0.584	2.990	5.248	5.087	3.221	3.426

Table S4: MAE calculated using cc-pVDZ basis set for each type of molecule and number of amplitudes predicted, percent of amplitudes predicted, and percent of opposite-spin (OS) correlation captured.

Molecule	Count		Percent of Amplitudes		Percent of MP2 OS correlation captured		MAE (mE_h)	
	$1\times10^{-7.25}$	$1\times10^{-7.5}$	$1\times10^{-7.25}$	$1\times10^{-7.5}$	$1\times10^{-7.25}$	$1\times10^{-7.5}$	$1\times10^{-7.25}$	$1\times10^{-7.5}$
Diethylether	119118	155906	2.7559	3.6071	98.9266	99.2814	4.019	0.584
Divinylether	99021	128336	4.1807	5.4184	99.1148	99.4237	0.641	2.990
Ethenol	50587	62719	14.0519	17.4219	99.5014	99.6999	3.779	5.248
Methoxyethene	81363	104557	6.4862	8.3352	99.1953	99.4853	2.899	5.087
Propenol	77202	98996	6.1545	7.8919	99.2480	99.5196	0.751	3.221

Table S5: Variation of the MAE with the percent of MP2 opposite-spin correlation captured by the amplitudes selected to be predicted. Predictions were made using a model trained with amplitudes whose magnitude of correlation energy is greater than $1\times10^{-7.5}$.

Cutoff	99.1	99.15	99.2	99.25	99.3
MAE (mE_h)	1.406	1.288	1.370	1.565	1.953

to 99.25%. Therefore, we used percent of MP2 opposite-spin correlation captured as the cutoff when testing for models trained with EC scheme using cc-pVDZ and aug-cc-pVDZ basis sets.

S-VII. Results for canonical orbital (CMO) formalism

Figure S5 depicts the MAEs for models trained with SB, CB, LA, and PS schemes. STO-3G basis set and canonical orbital formalism was used to calculate features and CCSD t_2 amplitudes. The EC scheme, which is a variant of the LA scheme was not used for the evaluation of performance with STO-3G basis set.

S-VIII. Effect of number of bins on the MAE of SB and CB model

To study the effect of number of bins on the accuracy of SB and CB schemes, we used 20% of amplitudes for SB scheme models and 5% of amplitudes for CB scheme models. These hyperparameters were chosen from the study done with the STO-3G basis set. A summary of MAE calculation for SB and CB schemes with varying bin sizes is given in Table S6. For the SB approach, the MAE drops drastically from 50×50 to 150×150 bins, while larger number of bins does not significantly affect the model accuracy. For the CB approach, the lowest MAE

was found on a model that utilized 50 bins, while more bins tend to negatively affect the model accuracy (increased MAE).

S-IX. Results for aug-cc-pVDZ basis set

Results for aug-cc-pVDZ basis set with LA and EC schemes are summarized in Table S7.

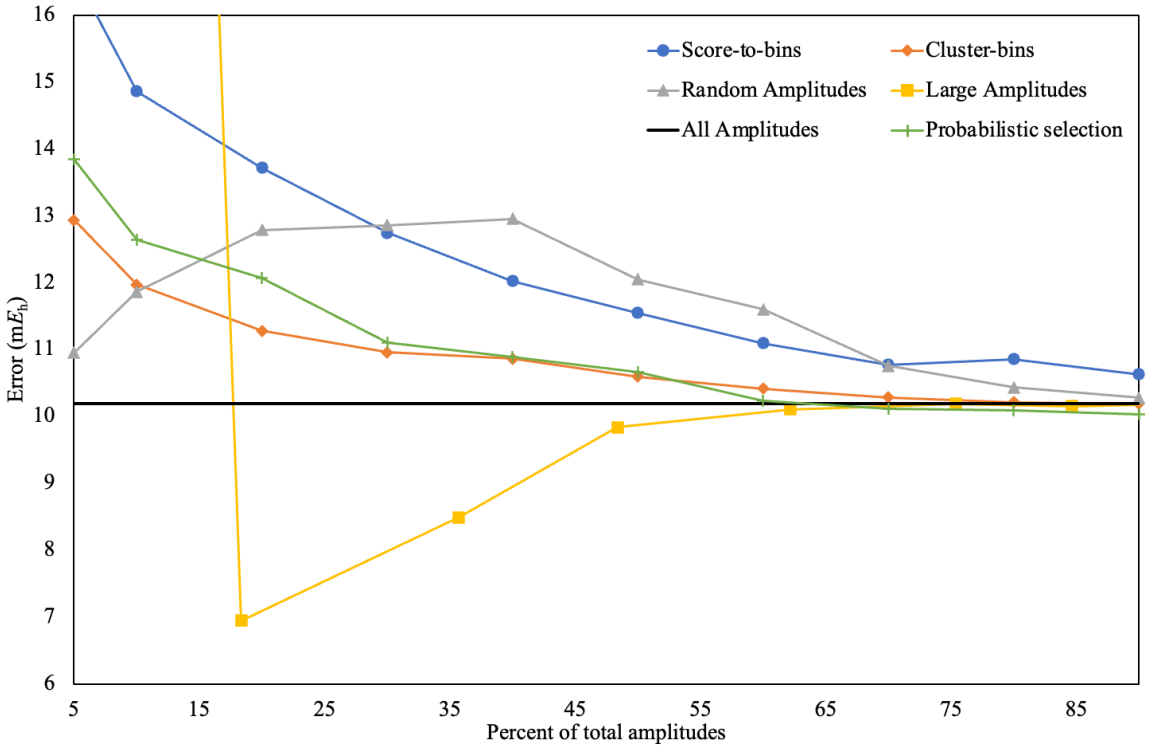

Figure S5: Mean absolute error vs. percentage of amplitudes selected for SB (blue), CB (orange), LA (yellow), and PS (green) schemes. MEA for Random selection approach was also depicted in gray. CCSD amplitudes for training and testing were calculated using STO-3G basis set and CMO formalism.

Table S6: Comparison of MAE with changing bin size for SB and CB schemes. For SB scheme 20% of the total amplitudes were selected and for CB approach 5% of the total amplitudes were selected. Calculation of CCSD amplitudes was done using STO-3G basis set.

Score to bin (SB) scheme		Cluster bins (CB) scheme	
Number of bins	MAE (mEh)	Number of bins	MAE (mEh)
10×10	1.723	10	2.803
25×25	2.224	25	1.088
50×50	1.946	50	1.038
100×100	1.292	100	1.241
150×150	0.826	150	1.440
200×200	0.856	200	1.483
250×250	0.839	250	1.576
300×300	0.791	300	1.430

Table S7: MAE calculated for LA, and EC schemes with different cutoff values. Features and CCSD amplitudes were calculated with aug-cc-pVDZ basis set. Cutoffs for LA scheme are set for the magnitude of MP2 amplitudes, whereas cutoffs for EC scheme are set for the correlation energy of each datapoint. The percentage of amplitudes selected for training EC scheme models were fixed at 4.47% of the total number of amplitudes.

LA Scheme			EC Scheme	
Cutoff	MAE (mE_h)	Percent amplitudes	Cutoff	MAE (mE_h)
1×10^{-3}	209.406	0.101	98.75%	2.011
$1\times10^{-3.25}$	38.105	0.242	98.8%	1.946
$1\times10^{-3.5}$	47.311	0.539	98.9%	2.013
$1\times10^{-3.75}$	47.272	1.203	99.0%	2.306
1×10^{-4}	4.921	2.077	99.1%	2.962
$1\times10^{-4.25}$	4.978	3.174	99.2%	3.787

S-X. Comparison of time taken to by each scheme to predict CCSD energy

A summary of the time taken to estimate CCSD energy for a diethyl ether conformer is listed in Table S8. For this calculation cc-pVDZ basis set was used. For SB, CB and PS schemes 4% of the total amplitudes were selected. For LA scheme 4.65% of the total amplitudes were selected using 1×10^{-4} as the cutoff for the absolute value of MP2 energy. For EC scheme 5.76% of the total amplitudes were selected using $1\times10^{-7.5}$ as the cutoff for the absolute value of the correlation energy of an amplitude.

Table S8: Comparison of the time taken to estimate DDCCSD energy with different schemes are shown. Time gain is the difference between time taken for exact CCSD calculation and DDCCSD calculation. Absolute difference between the exact CCSD energy and DDCCSD energy is given as the error.

Scheme	Time (s)				Error (mE_h)
	MP2	DDCCSD	CCSD	Time gain	
SB	95.916	129.098	169.431	40.334	11.822
CB	97.946	139.649	163.781	24.131	13.511
PS	99.915	130.756	168.955	38.199	12.371
LA	97.190	113.704	168.556	54.852	2.864
EC	100.957	119.591	172.599	53.009	0.918

Vita

Perera Don Varuna Sanjaya Pathirage was born and raised in Kolonnawa, Sri Lanka. He completed his primary, secondary, GCE ordinary level and GCE advanced level studies at Ananda College, Colombo, Sri Lanka. Then he entered University of Colombo, Sri Lanka in 2014 and completed the bachelor's in science degree majoring in computational chemistry in 2018. In 2019, Varuna joined the University of Tennessee to pursue his PhD in physical chemistry. Varuna published a book chapter and a manuscript while at the University of Tennessee.