Quantum Monte Carlo and Molecular Fragmentation Methods for the Treatment of Electron Correlation in Molecules and Solids

by

Amanda Dumi

Master of Chemistry, Duquesne University, 2017

Submitted to the Graduate Faculty of

the Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

DEPARTMENT OF CHEMISTRY

This dissertation was presented

by

Amanda Dumi

It was defended on

January 9, 2023

and approved by

Kenneth D. Jordan, Department of Chemistry

Dr. Daniel Lambrecht, Department of Chemistry & Physics, Florida Gulf Coast University

Dr. Jennifer Laaser, Department of Chemistry

Dr. Geoffrey Hutchison, Department of Chemistry

Quantum Monte Carlo and Molecular Fragmentation Methods for the Treatment of Electron Correlation in Molecules and Solids

Amanda Dumi, PhD

University of Pittsburgh, 2023

The quantum mechanical treatment of molecular systems in computational chemistry offers insight into the nature of chemical bonding, reaction mechanisms, and many experimental observables. An improved treatment of the molecular electronic structure is tied to an increase of computational resources. this work explores two approaches that attempt to describe the electronic structure with two families of approximations which reduce the compuational cost without significantly sacrificing accuracy.

One approximation explored is the application of Quantum Monte Carlo methods, which stochastically solve the Schrödinger equation. Here, the diffusion Monte Carlo formulation is used to provide insight into chemical systems that are not well described by the commonly utilized Density Functional Theory. Two systems are explored: a hydrogen atom chemisorbed to the surface of graphene and a model non-valence correlation-bound anion. The diffusion Monte Carlo approach is systematically improvable for most approximations, except for the fixed node error which can often be addressed through a careful choice of trial wave functions. Trial wave functions composed of multi-Slater determinants as generated by a selected configuration interaction procedure are investigated as they produce a compact determinant expansion through selecting the most important determinants for a specific system in an iterative fashion.

The second approach is the development of a fragment selection scheme through unsupervised machine learning approaches. Fragmentation approaches are motivated by the short range nature of correlation effects. The full system is approximated by subsystems which are each treated at a certain level of theory and an estimate of the interactions between them. For this approach to return valuable results, the chemical domains need to capture the most important physics of the desired problem. In this work an unsupervised machine learning based method is developed which will allow the systematic identification of important chemical domains with minimal quantum mechanical data which can improve transferability and automation of fragmentation approaches.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

> "The general theory of quantum mechanics is now almost complete, ... The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."-Dirac[1]

Electronic structure methods aim to accurately describe the electrons within a molecule or solid. This becomes computationally expensive when attempting to describe correlated electrons, generating competing goals of obtaining accurate results and feasible computational cost. One can avoid this complexity by treating the system in a mean-field approach such as Hartree-Fock, however electron correlation will not be accounted for. Neglecting electron correlation is not always appropriate for property prediction. As such, incorporation of these effects accurately and affordably is necessary. Full Configuration Interaction, which is formally exact in a complete basis, has prohibitive $N!$ scaling, where N is representative of system size. Whereas the coupled-cluster singles doubles with perturbative triples method is often regarded as the "gold standard" level of theory, but incurs an N^7 scaling. The higher scaling restricts the application of these methods to small systems while many of the chemical processes of interest involve much larger systems.

On the other hand, one of the most frequently applied methods is Density Functional Theory (DFT), a mean field approach that has a favorable scaling like HF, but approximates the missing correlation energy through an exchange-correlation functional. Although DFT has demonstrated the ability to successfully describe many molecular and periodic systems, there are many occasions in which this approximation falls short, e.g. systems which exhibit strong correlation, systems in which the physics used to construct the functional may differ from that being treated, and in situations targeting more than just total energies as exchange-correlation functionals are often benchmarked against energies only.

In this work, two approaches are explored which aim to treat only a subspace of the full electronic structure calculation which can accurately incorporate the correlation energy while reducing the computational demand. The first approach is the application of a stochastic sampling to solve the Schrödinger equation through quantum Monte Carlo approaches. The second is molecular

partitioning for fragmentation approaches determined by unsupervised machine learning methods. These two projects are introduced briefly in the Section 1.1.1 and Section 1.1.2. In the chapters to follow, relevant theoretical background is provided in Chapter 2, the QMC related works are discussed in Chapter 4 and Chapter 5, and the fragmentation work is presented in Chapter 6.

1.1 Project Summaries

1.1.1 Quantum Monte Carlo Applications

One way to allow for the treatment of the electronic structure for larger systems is to use a stochastic approach to solve the Schrödinger equation. When sampled with a Monte Carlo-type approach, these methods are collectively are known as quantum Monte Carlo (QMC) methods, though there are many formulations. We demonstrate the ability of QMC to accurately describe two challenging chemical problems where DFT provides inconsistent or incorrect results: hydrogen chemisorbed to a graphene surface and a model $(H_2O)_4$ system.

The first application is describing a hydrogen atom chemisorbed to graphene. The unique electronic, optical, and transport properties of graphene make it an important system for a wide range of applications, many of which involve or are impacted by the adsorption of atoms or molecules. To bring these applications to fruition, a deeper understanding of the interaction of atoms and molecules with graphene is required, and has been the subject of several experimental and theoretical studies.[2–14] Most computational studies of adsorption of atoms and molecules on graphene have employed DFT, primarily due to its favorable scaling with system size. However, a reliable theoretical description of interactions at the graphene surface has proven to be challenging for DFT.[2,7,8,15] The diffusion Monte Carlo (DMC)[16] method, which is a real-space stochastic approach to solving the many-body Schrödinger equation is particularly attractive given its low scaling with the number of electrons and high parallelizability. DMC also has the advantages of being systematically improvable and its energy being much less sensitive to the basis set employed than methods that work in the space of Slater determinants. DMC has been used to describe the adsorption of various species on graphene including O_2[10], a water molecule[9,17], and a platinum atom.[11] In the

present work, we use the DMC method to calculate the binding energy of H to graphene in the chemisorbed state.

The second challenging chemical system explored using QMC methods is a model $(H_2O)_4$ system, which is an example of a non-valence correlation-bound (NVCB) anion. In recent years, there has been growing interest in NVCB anions in which long-range correlation effects are crucial for the binding of the excess electron.[18–31] By definition, NVCB anions are unbound in the HF approximation. As NVCBs are relevant to image potential states of larger species such as C_{60} and graphitic systems, a method which enables an accurate ab initio treatment will allow for furthered understandings of these systems. Usual methods for treating these systems are restricted by a rather large scaling with system size, so it is of interest to know if DMC can provide an accurate description. This particular model system is useful as there is no net dipole and thus electrostatics will not play a prominent role in binding the excess electron.[18,25] Additionally, the coordinates of the water molecules can be tuned to regimes with dominant NVCB character or to regimes in which the NVCB character is lessened or non-existent. Although many of the approximation in DMC are systematically improvable, the fixed node error discussed in Section 2.3 needs care in ensuring the effects of the approximation are minimal. The nodes of the wave function are fixed to that of the initial trial wave function. This model system will allow us to also understand the quality of the trial wave function needed to allow for an accurate DMC treatment. Various sources of trial wave functions are explored, from single determinant methods like HF and DFT, to multi-Slater determinant expansions. The multi-Slater determinant trial wave functions are generated from Configuration Interaction truncation schemes. The truncations used are a manually defined orbital space/excitation degree scheme and through an automated approach of selected CI methods such as perturbatively selected configuration interaction scheme (CIPSI).[32–36] From a method development standpoint, there is interest in understanding how multideterminant trial wave functions can impact QMC treatments.[37–39]

1.1.2 Real Space Partition Methods

To compensate for high computational scaling, fragmentation approaches estimate the energy and other properties of large molecular systems by partitioning the system into small subsystems,

where the final estimate of the energy becomes the accumulation of the parts. This can be very successful as electron correlation is often a short range phenomena. The accuracy of the fragmentation approach hinges on the electronic structure treatment of each fragment, the approach used to describe the interaction between the fragments, and the way in which the molecule is partitioned. In the fragmentation schemes, the best case scenario for scaling becomes $\mathcal{O}(N^p) \rightarrow N_{\text{frag}}\mathcal{O}(f^p)$, where N_{frag} is the number of fragments, f is representative of the fragment size, and p is the exponential value dependent on the level of electronic structure theory utilized.[40,41] This partitioning of a single, very costly calculation into N_{frag} smaller calculations achieves two important objectives: 1) computational scaling with system size is reduced with reasonable fragment definition and 2) trivial parallelization is possible by treating subsystems separately, with the potential to efficiently utilize high performance computing resources.

As pointed out by Herbert, the choice of fragments for a system is not well-defined, but affects the quality of results obtained.[40] In some systems, a natural approach towards partitioning arises when there is a stark difference in the types of bonding present in the system. However, the choice of fragments is not always as clear-cut. In such cases, fragmentation requires the comparison of total energies, dipoles, or polarizabilities. Often the fragment definition is based on predefined functional groups or chosen manually.[42] However, functional group definition may become ambiguous. For example, there is no set number of monomers to include from a polymer backbone to acquire and accurate description capture the chemical behavior. Additionally, a fragment definition based only on functional groups may not consider the interacting chemical environment. If instead a more general approach could be developed that can consider the molecular system in an automatic and unbiased fashion, work would be made possible related to various fields such as quantum embedding approaches, Quantum Mechanics-Molecular Mechanics approaches, quantum computing efforts, and the description of large systems with ab initio methods. A desirable approach to choosing fragments would have low computational cost and prioritize keeping associated molecular components intact to treat fully with quantum mechanics while the estimation of their interactions should occur only at the most weakly bound points. Here we present a proof of concept for an unsupervised machine learning methods to identify the strongly interacting substructures of the system. Clustering methods are a form of unsupervised machine learning used to identify substructures in data sets. To this end, a set of clustering algorithms (spectral, agglom-

erative, *k*-means, and affinity propagation) were studied in combination with various molecular representations, including those incorporating bonding information derived from quantum mechanics. The performance for the clustering/descriptor combinations was assessed for test systems spanning a range from easily distinguishable fragments such as non-covalently bound water cluster to oligomers in which lowest-loss fragmentation is ambiguous.

2.0 Theoretical Background

Electronic structure methods are theoretical approaches which attempt to describe the behavior of electrons in a chemical system to gain insight into the properties of that system. The system can be described by a wave function Ψ, which contains all information about a state. We specifically denote the wave function of all particles as $\Psi(\vec{R}_A \ldots \vec{R}_M; \vec{r}_i \ldots \vec{r}_N)$ where r and R are position vectors for N electrons and M ions, respectively. When Ψ is acted on by an operator, observables of the system can be obtained. Specifically, let us consider the non-relativistic Schrödinger equation:

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

where $\hat{H}$ is the Hamiltonian, an operator which is applied to the wave function, Ψ, to return the eigenvalues representing the energies of the state. If Ψ is the true ground state wave function, the ground state energy of the system, E will be returned as the lowest eigenvalue. The full molecular Hamiltonian and has the following form:

$$\hat{H} = \sum_{A}^{M} -\frac{1}{2m_A}\nabla_A^2 + \sum_{i=1}^{N} -\frac{\hbar^2}{2m_e}\nabla_i^2 + \frac{e^2}{4\pi\varepsilon_0}\left(\sum_{i<j}\frac{1}{|\vec{r}_i - \vec{r}_j|} - \sum_{i=1}^{N}\sum_{A=1}^{M}\frac{Z_A}{\vec{r}_i - \vec{R}_A} + \sum_{A<B}\frac{Z_A Z_B}{\left|\vec{R}_A - \vec{R}_B\right|}\right) \tag{2}$$

where the first summation is over M ions, m_A is the mass of the ion, the second summation is over N electrons, $\hbar$ is the reduced Plank's constant, m_e is the mass of an electron, e is the elementary charge, ε_0 is the vacuum permittivity, Z is the atomic number of the nuclei,and R_A is the position of the ions. The first is the kinetic energy operator of the ions, the second term is the electronic kinetic energy operator, while the last three terms are potential energy terms resulting from Coulomb interactions between electron-electron, electron-ion, and ion-ion. To simplify this, Hartree atomic units will be used throughout this work which reduces the electron mass (m_e), the expression $4\pi\varepsilon_0$, reduced Planck's constant, and the elementary charge to 1.

In practice, finding and exact solution to Equation 2 is unfeasible except for systems containing a few particles. Describing larger systems requires approximations to be introduced. One of the most commonly utilized approximations, the Born-Oppenheimer approximation enforces that the electrons are moving within a fixed field of ions. This is a reasonable approximation as the larger

mass of ions compared to the mass of the electrons will cause the relative speed of the ions to be much slower. This allows the wave function to depend only on the coordinates of the electron ($\Psi(\vec{r}_i \ldots \vec{r}_N)$) and a simplification of the Hamiltonian to:

$$\hat{H} = \sum_i \left[-\frac{1}{2}\nabla_i^2 - \sum_{iA} \frac{Z_A}{\left|\vec{r}_i - \vec{R}_A\right|} \right] + \sum_{i<j} \frac{1}{\left|\vec{r}_i - \vec{r}_j\right|} \tag{3}$$

Separating this expression into the one-electron and two-electron terms results in:

$$\hat{H} = \sum_i \hat{h}(i) + \sum_{i<j} \hat{v}(i,j) \tag{4}$$

where $\hat{h}(i)$ is the one-electron operator which contains the electron kinetic energy and the electron-ion potential, and $\hat{v}(i,j)$ is the two-electron operator which describes the electron-electron repulsion energy.

Additional approximations are often introduced for the wave function as well as the full wave function is often too complex to fully describe. One of the first approximation is known as the Hartree product. The Hartree product wave function treats the full wave function as a product of the wave functions for each particle. This simple product is physically correct for Bosonic systems, but will fail to describe a Fermionic system as the Hartree product will not enforce that electrons are indistinguishable, i.e. that our wave function exhibits antisymmetry. To enforce antisymmetry of the wave function, Slater determinants are used. The Slater determinant has the form:

$$\Psi = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \cdots & \chi_N(\mathbf{x}_N) \end{vmatrix}$$

where χ_i is a single particle orbital and $\mathbf{x_i}$ is a collective variable to describe position and spin of an electron. The single particle orbitals in computational chemistry can have many forms. Plane waves are often used for materials calculation and molecular systems are often represented by molecular orbitals (MOs) that are a linear combination of atomic orbitals (LCAO) centered on atoms:

$$|\chi_i\rangle = \sum_v C_{vi}\phi_v \tag{5}$$

where χ are MOs and C are the linear coefficients for the atomic orbitals, ϕ. Within a single particle basis, the one- and two-electron terms can be expressed as integrals over the orbitals. The one-electron term becomes

$$\langle i|\,\hat{h}\,|j\rangle = \int d\mathbf{x}_1 \chi_i^*(\mathbf{x}_1)\hat{h}(1)\chi_j(\mathbf{x}_1) \tag{6}$$

The two-electron term can be expressed as

$$\langle ij|\,\hat{v}\,|kl\rangle = \int d\mathbf{x}_1 d\mathbf{x}_2 \chi_i^*(\mathbf{x}_1)\chi_j^*(\mathbf{x}_2)\hat{v}(1,2)\chi_k(\mathbf{x}_1)\chi_l(\mathbf{x}_2) \tag{7}$$

which is known as the physicists' notation, or one can express this in chemists' notation as:

$$[ij|kl] = \int d\mathbf{x}_1 d\mathbf{x}_2 \chi_i^*(\mathbf{x}_1)\chi_j(\mathbf{x}_1)\hat{v}\chi_k^*(\mathbf{x}_2)\chi_l(\mathbf{x}_2). \tag{8}$$

For a wave function to be exact and lead to an exact energy, an infinite basis is required. This is not feasible in computational chemistry do to the associated cost and instead the wave function is approximated by a finite basis. The variational principle can be used to inform the quality of the result obtained by the approximate wave function. The variational principle states that an approximate wave function which satisfies the boundary conditions of the problem will have an energy that is greater than the true ground state energy:

$$E_v \;= \frac{\int dR \Psi^*(R)\hat{H}\Psi(R)}{\int dR \Psi^*(R)\;\Psi(R)} \;\geq\; E_0. \tag{9}$$

Here, $R = (\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N)$ representing the positions of all electrons, E_v is the energy of the approximate wave function, and E_0 is the true ground state energy. Thus, if a method obeys the variational principle then the improvement of an approximate wave function will lead to a minimization of the variational energy.

Currently, the problem is presented as placing identical particles in orbitals. This idea lends itself well to another expression of these equations known as second quantization. Second quantization expressions will be exceptionally helpful when discussing wave function methods such as Configuration Interaction. The basics of the formalism are given here, but details of can be found in Reference [43]. Second quantization can be understood as a Slater determinant representing an occupation vector

$$|n_\alpha\rangle \;=\; |n_1, n_2, n_3, \ldots, n_\alpha, \ldots\rangle \tag{10}$$

where the values of the occupation will be 1 or 0 to indicate whether an orbital is occupied or unoccupied. This maintains that the sum over all occupied orbitals in the vector will result in the total number of electrons $\sum_\alpha n_\alpha = N$. The occupation of specific orbitals within the vector can be indicated through the use of creation operator , $a_i^\dagger$, and annihilation operators a_i. The creation operator will place a particle within the i^{th} orbital, while the annihilation operator will remove a particle from the i^{th} orbital. The molecular Hamiltonian (non-relativistic and within the Born-Oppenheimer approximation) in second quantization form is:

$$\begin{aligned} \hat{H} &= \sum_i \hat{h}(i) + \sum_{i<j} \hat{v}(i,j) \\ \hat{H} &= \sum_{pq} h_{pq} a_p^\dagger a_q + \sum_{pq,rs} v_{pqrs} a_p^\dagger a_q^\dagger a_r a_s + V_{NN} \end{aligned} \tag{11}$$

where h_{pq} are the one electron Hamiltonian terms for orbital pair pq and v_{pqrs} are the two-electron terms for $pqrs$ orbitals, and V_{NN} is the potential resulting from the fixed ions.

2.1 Mean Field Methods

2.1.1 Hartree-Fock Theory

The Hartree-Fock (HF) approach is one of the fundamental approximations for quantum chemistry applications. This method provides a description of an electron in the mean field of the other particles. The formulation of the method can be understood as finding the best set of spin orbitals for a single Slater determinant that result in the best estimation of our ground state wave function. Following from the variational principle, the best orbitals will be those that result in the lowest ground state energy. The energy expression for a single Slater determinant in the HF formalism is given by:

$$E_{HF} = \sum_i \langle i|\, h\, |i\rangle + \frac{1}{2} \sum_{ij} [ii|jj] - [ij|ji] \tag{12}$$

where i and j are the occupied orbitals. As shown in Equation 8, the full expression of the two-electron terms are:

$$[ii|jj] = \langle ii|\,\hat{v}\,|jj\rangle = \int\int d\mathbf{x}_1 d\mathbf{x}_2 \chi_i^*(\mathbf{x}_1)\chi_i(\mathbf{x}_1)\frac{1}{r_{12}}\chi_j^*(\mathbf{x}_2)\chi_j(\mathbf{x}_2) \tag{13}$$

$$[ij|ji] = \langle ii|\,\hat{v}\,|jj\rangle = \int\int d\mathbf{x}_1 d\mathbf{x}_2 \chi_i^*(\mathbf{x}_1)\chi_j(\mathbf{x}_1)\frac{1}{r_{12}}\chi_j^*(\mathbf{x}_2)\chi_i(\mathbf{x}_2). \tag{14}$$

These terms can be evaluated through the inclusion in the Fock operator that will have eigenvalues that are the optimal MOs to minimize E_{HF}.

$$f(\mathbf{x}_1) \;=\; h(\mathbf{x}_1) + \sum_{j\neq i}(\mathscr{J}_j(\mathbf{x}_1) - \mathscr{K}_j(\mathbf{x}_1)) \tag{15}$$

where $\mathscr{J}$ (Equation 16) is Coulomb interaction of electron in spatial orbital χ_i with the average charge distribution of all other electrons. The second term, $\mathscr{K}$, is the exchange operator which does not have a classical analog, but follows from the antisymmetry requirement of the wave function (Equation 17).

$$\mathscr{J}(\mathbf{x}_1) = \int d\mathbf{x}_2 |\chi_i(\mathbf{x}_2)|^2 r_{12}^{-1} \tag{16}$$

$$\mathscr{K}(\mathbf{x}_1)\chi_i(\mathbf{x}_1) = \left[\int d\mathbf{x}_2 \chi_j^*(\mathbf{x}_2) r_{12}^{-1}\chi_i(\mathbf{x}_2)\right]\chi_j(\mathbf{x}_1) \tag{17}$$

The restriction on the summation of $i \neq j$ would require a Fock matrix defined for each orbital, but since cases where $i = j$, $[\mathscr{J} - \mathscr{K}]\,\chi_i = 0$ the restriction on the sum can be removed and a single Fock matrix can be used for all orbitals. The HF equations can then be solved by employing the method of Lagrange multipliers with additional constraints to ensure that the MOs remain orthogonal.

The HF equations defined within a single particle basis result in the Roothaan equations. Using an LCAO basis from Equation 5, the matrix form of the equations become:

$$\sum_\nu F_{\mu\nu}C_{\nu i} = \varepsilon_i \sum_\nu S_{\mu\nu}C_{\nu i} \tag{18}$$

where $\mu\nu$ are indices of the atomic orbitals and i is the index of the molecular orbital, S is the overlap matrix and C contains the coefficients for the atomic orbitals. The HF approach is an example of a self-consistent method where iterations of steps of forming the Fock matrix, diagonalizing the Fock matrix, and checking for convergence. The convergence measures are typically for the energy and/or the density.

Due to the fact that HF is a mean-field approach, electrons are not correlated to one another. Correlation energy can be important to describe many chemical phenomena. In fact, the difference in energy of the true system within the nonrelativistic treatment from the energy described with the HF method is the correlation energy, $E_{corr} = E_{True} - E_{HF}$. Methods discussed in the next sections this are attempting to recover E_{corr} through approximate approaches to treating interacting electrons.

2.1.2 Density Functional Theory

Density Functional Theory (DFT) is another mean field approach but attempts to estimate the missing correlation energy. DFT is one of the most common methods used in solving the electronic structure in chemistry applications as it offers an approximation of the effects of electron correlation, while still having low computational scaling ($\sim N^3$). The discussion of this method is motivated by DFT's ability to reliable results in many cases, but also to illuminate the possible sources of error as these is a motivating facet of this work.

Fundamentally, the energy of a system can be understood as the interactions of the electrons and an external potential induced by the interaction of electrons with nuclei and additional sources:

$$E = \langle \Psi | \hat{H} | \Psi \rangle = \langle \Psi | \hat{T} + \hat{V}_{ee} | \Psi \rangle + \int dR v_{ext}(R) n(R) \tag{19}$$

where $\hat{T}$ and $\hat{V}_{ee}$ are the kinetic energy and electron-electron potential operators, respectively. Since the Ψ is dependent on v_{ext}, all terms in this equation are also dependent on v_{ext}. In other words, the energy can be understood as a functional of the external potential, which can be represented as $E[v_{ext}]$. Hohenberg and Kohn postulated that the many-particle wave function can be represented by a reduced variable of the electron density[44,45] It follows that if the exact ground state electron density is found, the exact ground state energy is recovered. The formulation of a universal functional was defined which will be valid for any V_{ext}:

$$F_{HK}[n] = \langle T \rangle + \langle V_{ee} \rangle \tag{20}$$

and an energy functional defined as:

$$E_{HK}[n, v_{ext}] = F[n] + \int dr v_{ext}(R) n(R). \tag{21}$$

For quantum chemistry applications, a further approximation of DFT is often used to simplify finding an accurate energy functional. The Hohenburg-Kohn formulation chooses the density as the quantity to form the functional around, however, this only allows for an explicit expression ford the external potential and the average Coulomb Hartree term, but the rest of the functional is unknown. Introduced by Kohn and Sham, the theorem states that a non-interacting system can be used to represent the electron density of the interacting system, where the auxiliary system is chosen to reproduce some but not all properties of a system.[46] This changes our expression of the energy to

$$E_{KS}[n] \;=\; T_{ip}[n] + \int d\mathrm{R} v_{ext}(R) n(R) + E_H[n] + E_{xc}[n] \tag{22}$$

where T_{ip} is now the independent particle kinetic energy, E_H is the Hartree potential, and E_{xc} is the exchange correlation energy can be understood as:

$$E_{xc}[n] = \langle T \rangle - T_i p[n] + \langle V_{ee} \rangle - E_H \tag{23}$$

where $\langle T \rangle$ is the true kinetic energy, T_{ip} is the kinetic energy of the system in absence of any electron-electron interactions, $\langle V_{ee} \rangle$ is the true electron interaction potential, and E_H is the Hartree potential. From Equation 22, all terms can be solved exactly except for E_{xc}, thus much of the following discussion describes how E_{xc} values are approximated.[47]

2.1.2.1 Exchange-Correlation Functionals

The E_{xc} value is determined through the use of an exchange-correlation functional. Since the exact form of the functional is not known, there have been many approximations. Families of functionals exist that are characterized by the information incorporated. The families of functionals are colloquially referred to as rungs of Jacob's ladder[48] and are discussed briefly here.[49] As more ingredients are incorporated into the functional, the accuracy of the functional generally increases. However, it must be emphasized that DFT is not systematically improvable. In the following paragraphs, a short description of the rungs are given.

The first rung is the local density approximation (LDA) functional, which is based on the uniform electron gas (UEG). Since there is no analytical form for the LDA correlation functional, parameterized data from QMC results for the UEG produced by Ceperly and Alder[50] are used.

The UEG is a good approximation for homogeneous systems and as a result the LDA functional can give reasonable descriptions in cases which resemble the UEG, but often falls short for many molecules, surfaces, and defect solids. which are inhomogenous. The second rung in the ladder are generalized gradient approximation (GGA) functionals which improve upon LDA functionals by introducing inhomogeneity correction factors through the density gradient. GGA functionals correct for the overestimation of the interaction energy of the LDA functional. The next rung of Jacob's ladder contains meta-generalized gradient approximation (mGGA) functionals, which are an extension of GGA functionals that include further improvements to the describe the density. Two different improvements can be included, the Laplacian of the density or the kinetic energy density. The flexibility in the functional form provided by the inclusion of the kinetic energy density make it a more common functional component. mGGA can improve upon GGA results, but are often more sensitive to the integration grid.

The functionals listed above all exhibit self interaction error, which results from the approximate exchange formulation. Hybrid functionals introduce exact exchange from HF in an attempt to correct for self interaction error.[51,52]. Though, the inclusion of exact exchange increases the scaling and memory requirements of the calculation. Hybrid functionals will be explored in Chapter 4 as tool to describe hydrogen chemisorbed to the surface of graphene. Further improvements to the functional form to correct for self interaction error results from range separated functionals.[53] Beyond these main rungs of Jacob's ladder, there are other improvements to functional accuracy which can be based on dispersion corrections or inclusion of data from wave function methods.[54–57]

The choices of functional are vast, but the lack of a systematically improvable framework can bring result in variability in the DFT results. Additionally, there are many systems which are a formidable challenge for DFT. Systems which exhibit strong correlation are poorly described by DFT. This is due to strong correlation requiring multireference wave functions to be captured accurately. The single reference nature of DFT leads to shortcomings in the description of these systems. The metal-insulator transition of Mott[58,59] is a well-studied example.

There have been attempts to incorporate strong correlation into various functionals.[60–62] Another approach is the use of the Hubbard model as a way to capture the effects of strong correlation.[63,64] Where the model can be used as a simplification of a solid state system where long range

interaction are ignored, and site-wise interactions can be introduced by U and J parameters. The U parameter captures strong on-site Coulomb interaction of localized electrons with an additional Hubbard-like term and J an exchange coupling term. These methods have been successful in many cases.[65] Though, the selection of the U and J parameters are often empirically determined diminishing the ab initio nature of DFT. More recently, the parameters have been determined through machine learning approaches of QMC-based calculations.[56]

There have also been effort to go beyond standard DFT to describe strongly correlated systems.[66] One such approach for materials is the use of model Hamiltonians.[67] These approaches involve downfolded Hamiltonians, tight-binding models, dynamical mean-field theory (DMFT). DMFT methods map a many-body lattice problem to a local model called an impurity model.[68,69] This approach has been used to successfully describe the Mott insulator transition mentioned above.[70]

2.2 Wave Function Methods

2.2.1 Configuration Interaction

In the Configuration Interaction (CI) approach, the wave function becomes a linear expansion of a set of Slater determinants.

$$|\Psi_{CI}\rangle \;=\; \sum_k c_k^{CI}\,|D_k^{CI}\rangle \tag{24}$$

$$E_{CI}c_k^{CI} \;=\; \sum_l \langle D_k^{CI}|\,\hat{H}\,|D_l^{CI}\rangle \tag{25}$$

The level of excitation contained in the CI can be truncated to only include certain levels of excitations.

$$|\Psi_{CI}\rangle \;=\; (1 \;+\; \sum_{ia} c_i^a a_a^\dagger a_i + \sum_{ijab} c_{ij}^{ab} a_a^\dagger a_b^\dagger a_i a_j \;+\; \ldots)\,|\Phi_{HF}\rangle \tag{26}$$

where a and $a^\dagger$ are discussed in Chapter 2 and i and j are occupied orbitals and a and b are virtual orbitals. If all possible excitations are chosen, this is known as Full Configuration Interaction (FCI). FCI provides the exact solution within a single particle basis. The approach is size extensive (energy grows linearly with the number of electrons) and and size consistent (energy of the system

is strictly separable.). However, the drawback of this method is the scaling of $N!$ as it considers all N electrons in $2n$ orbitals. This scaling can be slightly reduced by enforcing an S_z value (z component of the angular momenta), but the scaling still remains exponential.

2.2.1.1 Truncation of the FCI Space

In many instances, many of the Slater determinants in the full Hilbert space are not necessary for an accurate descriptions of the system.[71,72] For instance, in many systems the FCI expansion, the HF determinant will have the largest coefficient. If the coefficient on the HF determinant is large enough, a small determinant expansion can often provide accurate results. Once one is considering a truncated CI, the methods are no longer size-consistent or size-extensive. Determining a subspace of the full Hilbert space can be done in two ways:

- choosing only certain orbitals and/or allowing only certain excitations
- selected CI approaches: a family of algorithms which look to incorporate determinants in a system agnostic way

A reduction of the number of the determinants can be accomplished through a restriction on the level of excitation included. If only single excitations are considered, this is referred to as Configuration Interaction Singles (CIS), where Configuration Interaction Singles and Doubles (CISD) incorporates singles and doubles, and this pattern continues for including triple CISDT and quadruple CISDTQ excitations. Restrictions on the orbitals also allows for a significant reduction in the CI cost. This approach revolves around determining an active space which defines the orbitals which excitations are allowed to and from. One form of this type of restriction is freezing the core orbitals. While on the other end of things, the high-lying virtual orbitals can be frozen since the excitations into this space offer only small contribution to the wave function. However, this does lead to an ambiguity in the choice of active space for a given system. If one is defining a subset of orbitals and allowing all excitations within that subset, this approach is known as Complete Active Space Configuration Interaction (CASCI). The idea of reducing excitation level and orbital spaces are also a possibility.

In selected CI methods, the determinants are not chosen *a priori* but instead are selected during a process based on how the determinant contributes to the FCI wave function. There are various

formulations of selected CI approaches, some of which were first presented in 1960,[73] but selected CI methods are experiencing a renaissance in recent years. Some of the formulations are heat bath-CI[74,75], full configure interaction quantum Monte Carlo (FCIQMC)[76], adaptive CI[77], and perturbatively selected configuration interaction scheme (CIPSI)[32,33]. The work presented here employs the CIPSI approacj.

In the CIPSI method, an initial zeroth-order wave function is defined as an initial determinant expansion:

$$|\Psi^{(0)}\rangle = \sum_I c_I |D_I\rangle \tag{27}$$

where I can begin at 1 if using a single determinant or many if a small multi-determinant wave function was generated from a truncated CI scheme, such as CIS. The variational energy, $E^{(0)}$, can be evaluated as:

$$E^{(0)} = \frac{\langle\Psi^{(0)}|\hat{H}|\Psi^{(0)}\rangle}{\langle\Psi^{(0)}|\Psi^{(0)}\rangle}. \tag{28}$$

From the $|\Psi^{(0)}\rangle$, determinants external to this reference space are generated, $|\alpha\rangle$. The external determinants are the set of all single and double excitations from the reference space. The set of $|\alpha\rangle$ determinants are evaluated using a perturbative criteria, the Epstein-Nesbet perturbative energy (Equation 29) to give an estimate of the correlation energy recovered by the inclusion of $|\alpha\rangle$ into the variational wave function:

$$e_\alpha^{(2)} = \frac{\langle\Psi^{(0)}|\hat{H}|\alpha\rangle^2}{E^{(0)} - \langle\alpha|\hat{H}|\alpha\rangle}. \tag{29}$$

This can provide a second order energy correction to the variational energy

$$E^{(2)} = \sum_\alpha e_\alpha^{(2)} \tag{30}$$

where $E^{(2)}$ gives an approximation of the correlation energy that $\Psi^{(0)}$ is neglecting. A CIPSI iteration consists of the following steps:

1. Define a variational wave function which is an expansion over a set of *internal determinants*, $\mathscr{I}$

$$|\Psi^{(0)}\rangle = \sum_{I\in\mathscr{I}_n} c_I |D_I\rangle$$

2. Assess the variational energy according to Equation 27
3. Generate $\mathscr{A}_n$, the set of all *external determinants* $\alpha \notin \mathscr{I}_n$ and $\langle\Psi^{(0)}|\hat{H}|\alpha\rangle \neq 0$

4. Compute the perturbative contribution according to Equation 30
5. The α determinants with the largest $e^{(2)}$ contribution are incorporated into $\mathscr{I}_{n+1}$
6. Iterate until the number of determinants included in the space exceeds a certain threshold or the perturbative contribution to the energy from the external set of determinants contributes less than a certain threshold.

The CIPSI method has its own formulation of calculating excited states, forces, and dipoles.[36,78] However, it is often used a trial wave function for QMC techniques.[37,79]

2.2.2 Coupled-Cluster

The CC approach uses an exponential ansatz of the form

$$\Psi_{CC} \;=\; e^{\hat{T}} \left|\Phi_0\right\rangle \tag{31}$$

where Φ_0 is a reference wave function. The $\hat{T}$ is a cluster operator which contains degrees of excitations in the form $\hat{T} \;=\; \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + \ldots$ and $\hat{T}_1 \;=\; \sum_{ia}\; t_{ia} a_i^\dagger a_a$. The amplitudes of the cluster coefficients and the ground state energy, E_0, are found by solving the following set of nonlinear equation:

$$\left\langle \Phi_0 \right| e^{-\hat{T}} \hat{H} e^{\hat{T}} \left|\Phi_0\right\rangle \;=\; E_0, \left\langle \Phi^* \right| e^{-\hat{T}} \hat{H} e^{\hat{T}} \left|\Phi_0\right\rangle \;=\; 0 \tag{32}$$

where Φ^* is a specific excitation. This exponential ansatz guarantees the CC wave function is size consistent and size extensive.

If one considered all excitation operators, the FCI wave function would be recovered as would FCI's prohibitive scaling. In practice, only certain excitation operators are used. In Coupled Cluster Singles and Doubles (CCSD) only single and double excitations are considered. This approach scales as n^6 where where n is the number of spatial orbitals. However, failings of CCSD in certain cases motivates the inclusion of higher order excitations. If triple excitations are included, the CC scales as n^8, which becomes prohibitive for many systems. Instead, the Coupled Cluster Singles, Doubles, and Perturbative Triples (CCSD(T)) can be used that approximates the effect of the triple excitations by a perturbative correction scales and has a computational scaling of n^7. This method is often referred to as the "gold" standard in computational quantum chemistry.

Additionally, as one makes trade-offs of treating larger systems sizes with the CC methods containing a lower number of excitations to achieve a lower scaling, some chemical phenomena will be poorly described. e method which attempts to overcome having to make this trade off is Domain-based Local Pair Natural Orbital Coupled-Cluster (DLPNO)[80], which is a linear scaling approximation to the CC methods.

2.2.3 Equation-of-Motion

CC methods can also be used to obtain excited states when used with linear response theory or Equation-of-Motion (EOM), which can help explore photochemistry and spectroscopy questions.

The typical methods for excited state properties are ΔE methods are solving separately for the energy of the ground state and an excited state. Equation-of-Motion (EOM) methods instead allow us to solve for the excited state from a ground state reference.

$$\Psi_k \;=\; \mathscr{R}\,|\Phi_0\rangle \tag{33}$$

In particular, this discussion focuses on EOM-CC methods. The CC equations can be written to express the impact of the cluster expansion on the reference state in the form of a normal Hamiltonian[81]

$$\hat{H}_N = \hat{H} - \langle\Phi_0|\,\hat{H}\,|\Phi_0\rangle \tag{34}$$

where the CC equation becomes:

$$\hat{H}_N e^{\hat{T}}\,|\Phi_0\rangle \;=\; \Delta E_0 e^{\hat{T}}\,|\Phi_0\rangle \tag{35}$$

where the correlation energy is $\Delta E_0 \;=\; E_0 - \langle\Phi_0|\,e^T\hat{H}e^{-T}\,|\Phi_0\rangle$. These same expressions can be used for the excited state, Ψ_k. The solution to the Schrödinger equation becomes

$$\hat{H}_N\,|\Psi_k\rangle \;=\; \Delta E_k\,|\Psi_k\rangle \tag{36}$$

$$\hat{H}_N\mathscr{R}\,|\Psi_0\rangle \;= \Delta E_k\mathscr{R}\,|\Psi_0\rangle \tag{37}$$

where $\Delta E_k = E_k - \langle\Phi_0|\,\hat{H}\,|\Phi_0\rangle$ and E_k is the energy of the targeted state.

Subtracting the EOM Ansatz inserted into the Schrödinger equation ($\hat{H}\mathscr{R}|\Psi_0\rangle = E_{ex}\mathscr{R}|\Psi_0\rangle$) from the application of the $\mathscr{R}$ operator onto the ground state ($\mathscr{R}\hat{H}|\Psi_0\rangle = E_0\mathscr{R}|\Psi\rangle$) results in:

$$\left[\hat{H}_N, \mathscr{R}\right]|\Psi_0\rangle = \Delta\omega_k\mathscr{R}|\Psi_0\rangle \tag{38}$$

where $\omega_k = E_k - E_0$ and represents the energy of the transition studied. The CC exponential ansatz of Equation 31 can be substituted and the exponential cluster operator introduced through left multiplication resulting in

$$e^{-\hat{T}}[H, \mathscr{R}]e^{\hat{T}}|\Phi_0\rangle = \Delta\omega_k e^{-\hat{T}}\mathscr{R}_k e^{\hat{T}}|\Phi_0\rangle\,. \tag{39}$$

If the normal-ordered Hamiltonian is slightly reordered to:

$$\bar{H}_N = \langle\Phi_0|e^{-\hat{T}}\hat{H}e^{\hat{T}}|\Phi_0\rangle \tag{40}$$

allows us to express the EOM equations as an eigenvalue problem of $\bar{H}_N$:

$$\bar{H}_N\mathscr{R}|\Psi_0\rangle = \omega_k\mathscr{R}|\Psi_0\rangle \tag{41}$$

However, $\hat{H}_N$ is not Hermitian and thus, the right and left eigenvectors do not need to agree, but do exhibit biorthogonality. The right-hand solution is shown in Equation 41 The left-hand solution is a de-excitation operator, $\hat{L}$:

$$\langle\Psi_0|\hat{L}\bar{H}_N = \langle\Psi_0|\hat{L}\omega_k \tag{42}$$

The EOM solution will often target a specific property such as an excitation energy, ionization potential, electron affinity, and others. The solution results from a diagonalization of the $\hat{H}_N$ operator within the appropriate configurational subspace for the property. For example, the EE-EOM-CCSD subspace would be for $|\Phi_i^a\rangle$ and $|\Phi_{ij}^{ab}\rangle$

The excitation operator has a similar form of the cluster operator in CC methods:

$$\mathscr{R} = r_0 + \sum_i\sum_a r_i^a\, a_a^\dagger a_i + \sum_{i,j}\sum_{a,b} r_{ij}^{ab}\, a_a^\dagger a_b^\dagger a_i a_j + ... \tag{43}$$

The r coefficients can be solved through a coupled cluster approach. Due to the similarity in form, this operator can use truncation schemes akin to those used in CC. The EOM method shares similar scaling to CC when the same level of excitations are included in the operator. This means that although the method provides a clear path to excited states, the application will be restricted by scaling. DLPNO implementations of EOM exist implementations of to overcome the scaling restrictions.

2.3 Stochastic methods

The Hilbert space of quantum systems containing over a few atoms and a few orbitals is so large that many-body integration becomes impossible. This is part of the problem referred to as the *curse of dimensionality* A possible work around is to sample the phase space through a stochastic Monte Carlo (MC) method.

2.3.1 Classical Monte Carlo

Stepping away from the electronic structure perspective, Monte Carlo integration methods sample a multidimensional integral at a randomly selected set of points instead of assessing the multidimensional integral on a grid as is done in standard integration techniques. To generalize this, the naive MC approach will take a function, sample random points within a volume element that one is integrating over, and the values of the function at random points returned. Due to the law of large numbers, the expected value of the integral can be understood as the mean result of the samples. In these methods, the statistical error and statistical variance must be known to assess the reliability. This method is based on a given probability distribution, Π, a random walk can be taken through the possible states contributing to this distribution. This problem can be cast as a Markov chain where the samples become a series of events which are described as a transition matrix $T(s \to s')$, where a transition to s' only depend on the current state (s) and no others. This characteristic describes the concept of ergodicity if the starting state information is unknown after a certain number of steps. This requires that 1) one can move from one state to any other state in a finite number of steps with non-zero probability, 2) the transition matrix is non-periodic, and 3) average return time to any state is finite.

Detailed balance can be enforced to ensure only the intended distribution is sampled. In other words, the flux from state “A" to state “B" is equal to the flux of going from “B" to “A". In practice, the Metropolis-Hastings algorithm is used to enforce detailed balance. The Metropolis-Hastings algorithm is given in Algorithm **??** and a simple visualization shown in Figure 1.

Algorithm 1: Metropolis-Hasting's algorithm

1. Define the transition matrix and probability distribution
2. Set an initial state
3. Iterate M steps
   - Sample s' from $T(s^{(n)} \rightarrow s')$, where $s^{(n)}$ is the state at current iteration
   - Calculate the acceptance ratio $A = \frac{\Pi(s')T(s' \rightarrow s^{(n)})}{\Pi(s^{(n)})T(s^{(n)} \rightarrow s')}$
   - Accept or reject by drawing a random number from a $u_n \sim U(0,1)$.

$$s^{(n+1)} = \begin{cases} s' & A > u_n \\ s^{(n)} & A < u_n \end{cases}$$

2.3.2 Considerations for QMC

In this work and in many QMC methods, a correlated Slater-Jastrow wave function is used.[82–85].

$$\Psi_{SJ}(R) \; = \; det(\varphi_k(x_i))e^{\sum_{i<j} u(x_i,x_j)} \tag{44}$$

where φ_k are the orbitals, and the exponential term is the Jastrow correlation factor. This can be reformulated over all electrons and expressed as correlations through 1-, 2-, and 3-body terms as

$$U(R) \; = \; \sum_{i}^{I} u(\mathbf{r}_{iI}) + \sum_{i<j}^{N} u(\mathbf{r}_{ij}) + \sum_{i<j}^{N} \sum_{I}^{M} u(\mathbf{r}_{iI}, \mathbf{r}_{jI}, \mathbf{r}_{ij}) \tag{45}$$

The Jastrow factor can have various forms.[86,87] The forms used within this work are described in subsequent the chapters.

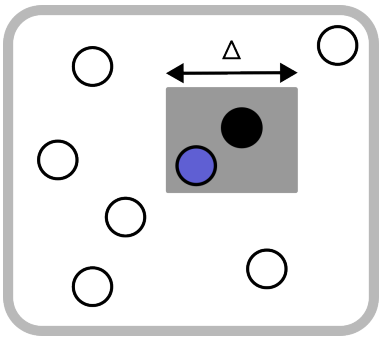


Figure 1: Schematic of a random walk, where a move is proposed from the black filled dot to the purple dot within the gray square. The acceptance of the move will be determined by the Metropolis-Hastings algorithm. Figure adapted from [45].

2.3.3 Variational Monte Carlo

The variational Monte Carlo (VMC) approach is based on the variational theorem. The method is used to optimize the Jastrow parameters of the trial wave function. In VMC a random walk is used to sample the distribution:

$$\Pi(R) \;=\; \frac{|\Psi_T(R)|^2}{\int |\Psi_T(R)|^2 dr} \tag{46}$$

where Ψ_T is the trial wave function. Equation 9 can then be rewritten as an average value of the "local energy" over the distribution $\Pi_T(R)$.

$$E_V \;=\; \int dR \Pi_T(R) E_L(R) \;=\; \langle E_L(R) \rangle_\Pi \tag{47}$$

where the local energy is defined as:

$$E_L \;=\; Re \frac{\hat{H}\Psi_T(R)}{\Psi_T(R)} \tag{48}$$

and an average variational energy over a number of steps in the random walk can be defined as:

$$\bar{E}_V = \frac{1}{M} \sum_n E_L(R^{(n)}). \tag{49}$$

As discussed, the error bar information of the random walk is a necessary assessment of the reliability of the results. For VMC, the variance of the energy can be assessed as

$$v_T \;=\; \frac{\int dR\,|\Psi_T|^2\,(E_L(R)-E_V)^2}{\int dR\,|\Psi_T|^2} \tag{50}$$

and can be expressed as an average variance value over M steps:

$$\bar{v}_T \;=\; \frac{1}{M}\sum_{n=1}^{M}(E_L(R^{(n)}) \;-\; \bar{E}_V)^2 \tag{51}$$

This method is restricted by the quality of the Ψ_T. In other words, even with sufficient sampling, VMC will produce an exact answer only if $\Psi_T \;=\; \Psi_{exact}$.

2.3.4 Projector Monte Carlo

Projector Monte Carlo is an umbrella term for a variety of methods that all filter out the ground state by repeatedly applying a many-body projector with a random walk.[45,88]

This formulation can be understood if one begins with the time-dependent Schrödinger equation:

$$i\frac{d\Psi(R,t)}{dt} \;=\; \hat{H}\Psi(R,t) \tag{52}$$

As discussed, Ψ can be expressed a linear combinations of eigenfunctions:

$$\Psi(R) = \sum_j c_j \varphi_j(R) \tag{53}$$

$$\hat{H}\varphi_i = E_i\varphi_i \tag{54}$$

To arrive at a form to be used with projector methods, a transformation to imaginary time through the substitution $t \rightarrow i\tau$ is done

$$\frac{d\Psi(r,\tau)}{d\tau} \;=\; \hat{H}\Psi(r,\tau). \tag{55}$$

A formal solution to the imaginary time Schrödinger equation is

$$|\Psi(\tau_1+\delta\tau)\rangle = e^{-H\delta\tau}\,|\Psi(\tau_1)\rangle \tag{56}$$

this can then be expanded into energy eigenstates, as is done in Equation 53

$$|\Psi(\delta\tau)\rangle = \sum_{i=0}^{\infty} c_i e^{-\varepsilon_i\delta\tau}\,|\varphi_i\rangle\,. \tag{57}$$

Finally, a shift by E_0 is introduced to ensure a stationary solution as $\tau \to \infty$

$$|\Psi(\delta\tau)\rangle = \sum_{i=0}^{\infty} c_i e^{-\varepsilon_i - E_0 \delta\tau} |\varphi_i\rangle \tag{58}$$

Thus as imaginary time progresses through the repeated application of the projector, any state that is not orthogonal to the ground state and is higher in energy will be damped out leaving the ground state projected out.

Here, the discussion will focus on diffusion Monte Carlo (DMC) as that is the main projector Monte Carlo approach used in this work. The DMC projector is

$$\hat{G}_D \;=\; e^{-\Delta\tau(\hat{H}-E_T)} \tag{59}$$

where E_T is the trial energy and is introduced as a way to control the normalization. $\Delta\tau$ is the time step in imaginary time. However, the exponential operator is often requires through the diagonalization of $\hat{H}$ within a continuous real space basis, making $\hat{G}_D$ intractable for many interesting systems. Instead the application of the operator can be done in the short time expansion, which is made possible through Trotter's formula[89,90]:

$$E^{-\tau(\hat{T}+\hat{V})} = \lim_{n\to\infty} (e^{-\Delta\tau\hat{T}} e^{-\Delta\tau\hat{V}})^n \tag{60}$$

Furthermore, instead of sampling the all real space configurations, the projection is carried out through a random walk through the space of electron coordinates. The common approach to sampling is a branching scheme where one is duplicating and terminating random walks. One can consider the effect of the random walk through the impact of the Hamiltonian components individually. The kinetic energy term in the absence of the potential term leads to a diffusive step and the random walk progresses from R to R' according to the diffusion probability:

$$P_d(R' \leftarrow R) = (2\pi\Delta\tau)^{\frac{3N}{2}} exp(|R-R'|^2/\Delta\tau) \tag{61}$$

If the potential term is considered in the absence of the kinetic term, branching steps are defined according to $exp(V(R') - E_T)$. However, since $\hat{T}$ and $\hat{V}$ do not commute, an error is incurred. The error can be controlled through alternating between branching and diffusive steps as well as by utilizing the second order Trotter-Suzuki formula.[91]

Beyond the formulation of the method, there are a few practical considerations that must be made to the algorithm to produce reliable results with efficient sampling.

Importance Sampling: Given a long enough propagation through imaginary time the ground state will be projected out for any initial guess wave function which adheres to the boundary conditions of the problem. However this is often unstable and inefficient. Instead, one can utilize importance sampling, which introduces a guiding wave function, Ψ_G, that approximates the true ground state wave function. The projector becomes

$$G(R \leftarrow R') = \Psi_G(R) \langle R | e^{-\Delta\tau(\hat{H}-E_T)} | R' \rangle \Psi_G(R)^{-1} \tag{62}$$

which results in higher sampling in areas where Ψ_G is large.

Fixed Node Approximation: The Fermion wave function is not a probability distribution as it will have positive and negative regions. The most common way to address the sign problem is to forbid moves that would change the sign of the trial wave function. The fixed-node energy is an upper bound to the exact ground state energy. The incorporation of this approximation is referred to as FN-DMC The DMC investigations presented in Chapters 4 and 5 will utilize FN-DMC approach.

2.3.4.1 DMC for Periodic Systems

todo: add dmc periodic discussion for twist averaging, non local pseudopotential, a

In treating extended systems, there is often an interest in describing the behavior as one approaches the thermodynamic limit. Modeling an infinite system by a smaller simulation cell will incur some finite size errors as a periodic image may impact the simulation cell nonphysically. Finite size effects for QMC calculations result from changes in the correlation functions as a function of the electron density and differences resulting from representing the integration of k-points on a finite mesh in reciprocal space as a summation.[92] These sources of error impact both the kinetic and potential energy. The finite size errors from interaction between simulation cells can happen when the correlation length exceeds the simulation cell size. These errors can typically be addressed by increasing the simulation cell size, extrapolation and/or correction techniques, and twist boundary conditions.[93] Twist boundary conditions are employed in this work. Twist boundary conditions express the boundaries of a periodic wave function in terms of an angle θ_α. A

value of θ_α=0 is equivalent to periodic boundary conditions, but other angles allow for a rotation of the phase of the particles in reciprocal space as a particle reaches a boundary. This approach accomplishes a similar goal to that of Brillouin zone integration done in many DFT calculations. Twist boundary conditions only incur a small computational cost since accumulating the average this way also reduces the statistical noise of the calculation.

Although twist boundary conditions ease the impact of using a finite simulation cell, additional corrections are needed. For instance, the long range Coulomb potential cannot be modeled with the typical expression of $1/r$, but instead relies on approximations such as the Ewald summation or model periodic Coulomb potential.[94, 95] Impacts of finite size effects on the kinetic energy can be corrected using the Chiesa correction.[96] The correlation length becomes very large or diverges at phase transitions, though methods have been developed to perform finite size extrapolations.[97,98]

3.0 Fragmentation Methods

To overcome the high scaling of computational cost with system size, fragment approaches approximate the energy and other properties of the system by partitioning into subsystems, where the final result becomes the accumulation of the parts. One potential advantage of fragmentation is to reduce the computational scaling from $\mathcal{O}(N^p)$ to $\mathcal{O}(N_{\text{frag}}^q f^p)$, where p is the scaling exponent for the chosen electronic structure method, where N_{frag} is the number of fragments, f is representative of the fragment size, and q is an exponential specific to the fragmentation approach. Judicious choices in the fragment approach can lead to a reduction in overall computational scaling ($q < p$), and in best cases linear scaling in the number of fragments can be obtained, $\mathcal{O}(N_{\text{frag}} f^p) \approx \mathcal{O}(N_{\text{frag}})$, where constant fragment size f is assumed. Another appealing feature of fragment approaches is that parallelization is facile since much of the workload for the individual subsystems can be treated separately, which provides the potential to efficiently utilize high performance computing resources to reduce the overall wall time of calculations. Beyond these computational advantages, fragment approaches can quantify magnitudes and roles of inter-fragment interactions that can deepen chemical understanding similar to analysis techniques such as energy decomposition analysis or symmetry adapted perturbation theory.[99–103] A perhaps surprising application of fragment approaches was proposed for quantum computing, where it may offer a route to overcome the challenges of system-environment interactions which hamper the integrity of results for quantum simulations of large systems. Instead of treating an entire (large) system on a quantum computer, fragment approaches were suggested so that the system could be partitioned into subsystems which lead to circuits that are less susceptible to errors, while incorporating the most important chemical interactions in the quantum domain.[104] In summary, fragment approaches offer useful features ranging from computational efficiency over chemical insights to potential applications in unprecedented fields such as quantum computing.

The accuracy and efficiency of fragment approaches hinges on a number of factors, including the choice of electronic structure treatment for each fragment, approaches to incorporate inter-fragment interactions, and the specific choices for partitioning of the system. Significant attention has been devoted to the treatment of inter- and intra-fragment interactions. These approaches will

be discussed briefly, but are beyond the scope of the current work and interested readers are directed towards a number of helpful reviews.[40,105] In contrast, this work investigates how a number of strategies for the selection of fragments impact the accuracy and efficiency of the results.

3.1 Inter-fragment Treatments

Although improving the treatment of inter- and intrafragment interactions is not a focus of the current body of work, a brief overview of the main families of methods are given for context. These methods can be categorized as, many-body expansion (MBE) approaches, and multilayer composite approaches. ("ONIOM" scheme), approximate functional group additivity approaches.

The many body expansion approaches can have many flavors, but they essentially look to take the energy of each fragment and estimate the energy of including many body terms into the energy. The estimation of many-body effects to the energy is done by studying groups of fragments according to Equation 63

$$E = \sum_I E_I + \sum_I \sum_{J>I} \Delta E_{IJ} + \sum_I \sum_{J>I} \sum_{K>J} \Delta E_{IJK} + \ldots \tag{63}$$

where the ΔE terms represent additions of the many body terms to the initial fragment energy, i.e., the energy of the isolated fragments subtracted from the energy of combined fragment. For a concrete example, the two body form of this term is

$$\Delta E_{IJ} = E_{IJ} - E_I - E_J \tag{64}$$

The popular quantum chemistry method of fragment molecular orbitals fall under the umbrella of a MBE. The many-body terms will be truncated at a certain level of interaction.

A different family of methods are known as multilayer approaches which break the system into areas of interest to estimate the full system as a combination of subsystems treated at varying levels of theory. This in theory would allow one to employ a high level of theory to a realm which one knows interesting physics is occurring, but then switch to a medium or low level theory for those areas participating less in the area of interest. Methods that fall under this family are

ONIOM[106–108] and molecule-in-molecule (MIM) approaches[109–111]. The third type of fragmentation method attempts to determine the size of the fragments through an iterative scheme until convergence of the property with changing fragment size.

4.0 The Binding of Atomic Hydrogen on Graphene from Density Functional Theory and Diffusion Monte Carlo Calculations

The test and figures in this chapter have been adapted from A. Dumi, S. Upadhyay, L. Bernasconi, H. Shin, A. Benali, and K.D. Jordan The binding of atomic hydrogen on graphene from density functional theory and diffusion Monte Carlo calculations, *J. Chem. Phys.*, **2022**, 156, 144702, DOI:10.1063/5.0085982.with the permission of AIP Publishing.with permission from The author's contribution to the work included performing trial wave function generation calculations, generation of images, setting up the QMC workflow with S. Upadhyay, along with performing analysis and assisting with figures, and editing/revising the manuscript.

4.1 Summary

In this work density functional theory (DFT) and diffusion Monte Carlo (DMC) methods are used to calculate the binding energy of a H atom chemisorbed on the graphene surface. The DMC value of the binding energy is about 16% smaller in magnitude than the Perdew-Burke-Ernzerhof (PBE) result. The inclusion of exact exchange through the use of the Heyd–Scuseria–Ernzerhof (HSE) functional brings the DFT value of the binding energy closer in line with the DMC result. It is also found that there are significant differences in the charge distributions determined using PBE and DMC approaches.

4.2 Introduction

The unique electronic, optical, and transport properties of graphene make it an important system for a wide range of applications, many of which involve or are impacted by the adsorption of atoms or molecules. To bring these applications to fruition, a deeper understanding of the interaction of atoms and molecules with graphene is required, and, not surprisingly, this has been the

subject of several experimental and theoretical studies.[2–14]

The adsorption of H atoms on graphene has been the subject of multiple studies.[3–6,8,112] It is known that there is both a weakly absorbed state in which barriers for diffusion are small and a much more strongly bound chemisorbed state[113,114], which is the focus of this work. Chemisorbed H atoms open up the band gap and allow for tuning of electronic properties.[115] It has been demonstrated that even a single chemisorbed hydrogen atom causes an extended magnetic moment in the graphene sheet.[116,117] On the other hand, there is evidence that given the ready diffusion of H in the physisorbed state, the H atoms tend to pair up on the surface leading to non-magnetic species.[117] Finally, interest in the hydrogen/graphene system has also been motivated by the potential use of graphene and graphitic surfaces for hydrogen storage.[14] In spite of the interest in H chemisorbed on graphene, we are unaware of experimental values of the binding energy.

Most computational studies of adsorption of atoms and molecules on graphene have employed density functional theory (DFT), primarily due to its favorable scaling with system size, allowing for the treatment of larger periodic structures. However, a reliable theoretical description of interactions at the graphene surface has proven to be challenging for DFT.[2,7,8,15] In recent years considerable progress has been made in extending correlated wave function methods to periodic systems.[79,118–122] Among these methods, the diffusion Monte Carlo (DMC)[16] method, which is a real-space stochastic approach to solving the many-body Schrödinger equation is particularly attractive given its low scaling with the number of electrons and high parallelizability. DMC also has the advantages of being systematically improvable and its energy being much less sensitive to the basis set employed than methods that work in the space of Slater determinants. In DMC calculations, the atomic basis set is important only to the extent that it impacts the nodal surface. DMC has been used to describe the adsorption of various species on graphene including O_2[10], a water molecule[9,17], and a platinum atom.[11] In a study of a physisorbed H atom on graphene, Ma et al. found that different DFT functionals gave binding energies ranging from 5 to 97 meV, while DMC calculations gave a value of only 5 ± 5 meV.[8] Various DFT calculations utilizing the Perdew-Burke-Ernzerhof (PBE)[123] and Perdew-Wang (PW91)[124] functionals predict the chemisorbed H atom species to be bound by 480 to 1,440 meV.[125–133] However, this large spread is primarily a result of some calculations employing small supercells resulting in an unphysical description of the low-coverage situation, too small a *k*-point grid, or small atom-localized basis sets that do not

adequately describe the binding and introduce large basis set superposition error (BSSE). In the present work, we use the DMC method to calculate the binding energy of H to graphene in the chemisorbed state.

4.3 Methods

All calculations reported in this study used a 5x5x1 supercell of graphene, as it was large enough to make inconsequential the interaction between periodic images of the adsorbed hydrogen atom and to assure that there are essentially unperturbed C atoms between the buckled regions in adjacent images in the x and y directions. The geometries of graphene, both pristine and with a chemisorbed H atom, were provided by Kim et al.,[134] and were obtained using the PBE+D3 DFT method.[123,135] For all systems, a vacuum spacing of 16 Å was used.

4.3.1 Density Functional Theory Calculations

The single particle orbitals used in the trial wave functions for variational Monte Carlo (VMC) and DMC calculations were calculated using the PBE functional with the correlation consistent electron core potential (ccECP)[136,137] pseudopotentials and a plane wave basis with an energy cutoff of 3,400 eV. Monkhorst-Pack k-point grid meshes[138] were employed with a 13.6 meV Marzari-Vanderbilt-DeVita-Payne cold smearing of the occupations.[139] The PBE results were converged at a 6x6x1 k-point grid to 1 meV for graphene and graphene with an adsorbed hydrogen atom. The hydrogen atom trial was generated using a 1x1x1 k-point grid. Convergence studies can be found in Table S1 and S2 of the Supplementary Material.

In addition to the PBE calculations used to generate the trial wave functions for DMC, DFT calculations were carried out with the PBE0[140] and Heyd–Scuseria–Ernzerhof (HSE) functionals[141] to determine if inclusion of exact exchange proves important for the adsorption energy. Due to the inclusion of exact exchange, these calculations would be computationally prohibitive in a plane wave basis, particularly with the high energetic cutoff required by the ccECP pseudopotential. For this reason, they were carried out all-electron with the POB-TZVP Gaussian type orbital

(GTO) basis set.[142] Due to the use of GTOs, these calculations suffer from basis set superposition error (BSSE), which we corrected using Grimme's geometry-dependent counterpoise correction scheme.[143,144] This correction resulted in a 113 meV reduction in the magnitude of the binding energy when using the PBE0 functional. For the PBE0 and HSE calculations, a 12x12x1 k-point grid was used to assure binding energies converged to within 2meV. Convergence data are supplied in Table S3 of the Supplementary Material.

The plane wave DFT calculations were carried out with the QUANTUM ESPRESSO version 6.3 code.[145–147] The Gaussian basis DFT calculations were carried out with CRYSTAL17,[148,149] save for the HSE calculation of the isolated hydrogen atom which was carried out using NWChem version 6.8[150] using the same basis as the calculations in CRYSTAL17.

4.3.2 Quantum Monte Carlo Calculations

DMC is a projector quantum Monte Carlo (QMC) method, solving the Schrödinger equation in imaginary time $\tau = it$; any initial state $|\psi\rangle$, that is not orthogonal to the true ground state $|\phi_0\rangle$, will evolve to the ground state in the long time limit. When dealing with Fermionic particles, the DMC method requires the use of the fixed-node approximation[151] to maintain the antisymmetric property of the wave function. For efficient sampling and to reduce statistical fluctuations, we use a Slater-Jastrow trial wave function fixing the nodes through a Slater determinant comprised of single-particle orbitals, which, in this work, are expanded in a B-spline basis. The Jastrow factor is a function that reduces the variance by explicitly describing dynamic correlation. The Jastrow factor contains terms for one-body (electron-ion), two-body (electron-electron) and three-body (electron-electron-ion) interactions. The one- and two-body terms were described with spline functions[152], while the three-body terms were represented by polynomials.[153] 10 parameters were used for the one-body terms per atom type, and 10 parameters were employed per spin-channel for the two-body terms. The cutoffs on the one- and two-body terms were fixed to the Wigner-Seitz radius of the simulation cell. The three-body terms were comprised of 26 parameters per term with a cutoff of 10 Bohr. The parameters in the Jastrow factor were separately optimized for each geometry with the linear method[154] using VMC. To reduce the cost of the DMC calculations as well as to reduce the fluctuations near the ionic core regions, ccECP pseudopotentials were used to

replace the core electrons.[136,137] The ccECP pseudopotentials were designed to be used with high-accuracy many-body methods. The non-local effects due to the pseudopotentials were addressed

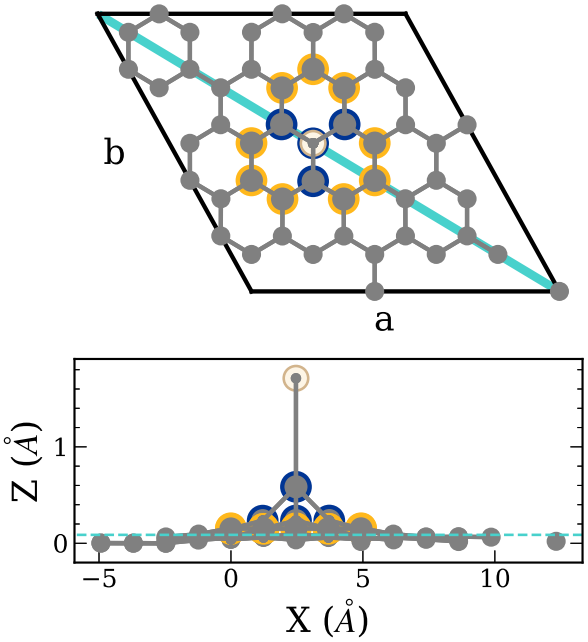


Figure 2: Perpendicular view of the simulation cell (top) and a parallel view obtained by projection onto the *xz*-plane (bottom). The carbon atoms are colored gray and the hydrogen atom is denoted as white. For the perpendicular view, the cyan line represents the slice of the cell used to visualized electron density differences. For the parallel view, the dotted cyan line represents the mean carbon *z* position. Blue outlined atoms are greater than one standard deviation away from the mean carbon *z* position, whereas yellow atoms are between 0.5-1.0 σ.

using the determinant-localization approximation along with the t-moves method (DLTM).[155,156] Finite size effects were addressed using twist averaging.[93] The twist angles were chosen to be the symmetry unique points of the 6x6x1 *k*-point grid shifted by half a grid step away from the gamma point in each direction.

The DMC calculations were performed using the branching scheme proposed by Zen et al. (ZSGMA)[157] with a population control target of 8,192 walkers and a time step of 0.005 a.u., which represented a balance between computational cost and finite timestep error in previous work.[10]

We define the binding energy as

$$E_b = E_{dgr+H} - (E_{gr} + E_H) \tag{65}$$

where E_{dgr+H} is the energy of the distorted graphene sheet with a chemisorbed atomic hydrogen, E_H is the energy of a hydrogen atom, and E_{gr} is the energy of a pristine graphene sheet. In the chemisorbed state, the hydrogen atom bonds directly over a carbon atom, causing this carbon to be pulled out of the sheet towards the hydrogen.[158,159] The adjacent carbons are also pulled in the direction of the hydrogen leading to a distorted graphene sheet.

The QMC calculations were carried out using the QMCPACK code, with the workflow between QUANTUM ESPRESSO and QMCPACK managed by Nexus.[160–162] Figures 2 and 4 were rendered with matplotlib[163] and the density plots were generated using VESTA.[164]

4.4 Results & Discussion

4.4.1 Binding Energy

Table 1 contains a summary of the binding energies of a hydrogen atom chemisorbed on graphene from this work and selected values from previous publications using the PW91 and PBE functionals. These literature values range from -790 to -980 meV. This wide spread of binding energies is caused by (1) the use in some studies of small supercells for which there are sizable interactions between the CH groups in adjacent cells, and (2) the use in some studies of small atom-centered basis sets without corrections for BSSE. Our calculations with the PBE functional in conjunction with a plane wave basis set give a binding energy of -821 meV. This should be contrasted with our -691 $\pm$ 19 meV DMC result. There are several possible sources for the difference between the PBE and DMC values of the binding energy. These include errors in the DFT calculations due to self interaction and planar graphene having more multiconfigurational character than H/graphene, with this being better described with DMC than with PBE. We note that the inclusion of the D3 dispersion correction with the PBE functional only changes the magnitude of the binding energy by 0.03 eV.

Table 1: Binding energy (meV) of a hydrogen atom chemisorbed on graphene calculated with various DFT functionals and with DMC.

Method	Binding energy
This Work	
PBE[1]	-821
PBE[2]	-871
PBE0[2]	-851 (-800)
HSE[2]	-794 (-743)
DMC	-691 $\pm$ 19
Previous Work	
PW91	-810 to -830[126], -870[127]
PBE	-790[128], -840[129], -980[130]

[1] Calculation was done in the plane wave basis

[2] Calculation was done in the Gaussian basis set with corrections for BSSE. Values in parentheses include a correction for the basis set incompleteness as described in the text.

The PBE binding energy is 51 meV lower in magnitude in the plane wave than in the GTO basis set when the same *k*-point grid is used, and this value is used as a correction for the basis set incompleteness error for the results with other functionals in Table 1. The calculations in the GTO basis set give a slightly smaller in magnitude binding energy with PBE0 than with PBE. However, with HSE, we obtain a binding energy 77 meV smaller in magnitude than the PBE result. Applying the correction for the basis set incompleteness error, we obtain -800 meV for the PBE0 binding energy and -743 meV for the HSE binding energy, with the latter being in reasonable agreement with the DMC result of -691 meV. Although the 130 meV difference between the plane-wave PBE and DMC values of the binding energy may appear to be small, this energy difference, of that

magnitude is consistent with an order of magnitude change in the hydrogen evolution current at room temperature on graphene electrodes.[134]

In order to better understand the origin of the difference in the PBE and HSE H-atom adsorption energies, we also carried out non-self-consistent calculations, using PBE densities to evaluate the HSE energies. These calculations gave a binding energy only 21 meV smaller in magnitude than obtained from the self-consistent HSE calculations. This demonstrates that the functional is more important than the density in establishing the binding energy. Detailed information can be found in Table S4 of the Supplementary Material.

Detailed results of the DMC calculations can be found in Appendix C.

4.4.2 Binding Density

It is instructive to examine the change in the electron density associated with the binding of the H atom to the distorted graphene as determined from the PBE and DMC calculations. The density change is given by

$$\rho_b = \rho_{dgr+H} - (\rho_{dgr} + \rho_H), \tag{66}$$

where ρ_H is the charge density of the hydrogen atom, and ρ_{dgr+H} and ρ_{dgr} are the charge densities of the distorted graphene sheet with and without hydrogen, respectively. For the QMC density, the density was accumulated during the VMC and DMC calculations, the mixed estimator bias was found to be insignificant, and was thus not corrected.

The ρ_b density differences for both DMC and PBE are shown in Figure 3. The dark blue and gold regions represent a loss and gain of electron density, respectively. As expected, there is a shift in electron density from the carbon atom participating in the carbon-hydrogen bond as well as to the three adjacent carbon atoms. These qualitative changes in the density are consistent with previous theoretical and experimental studies.[158,159] The rehybridization from sp^2 to sp^3 of the carbon participating in the CH bond and the weakening of the π bonds due to the distortion of the graphene lead to the electron density shift. The change in the charge distribution is similar for PBE and DMC, with the most noticeable difference being a greater increase of density at remote C atoms in the DMC than in the PBE calculations.

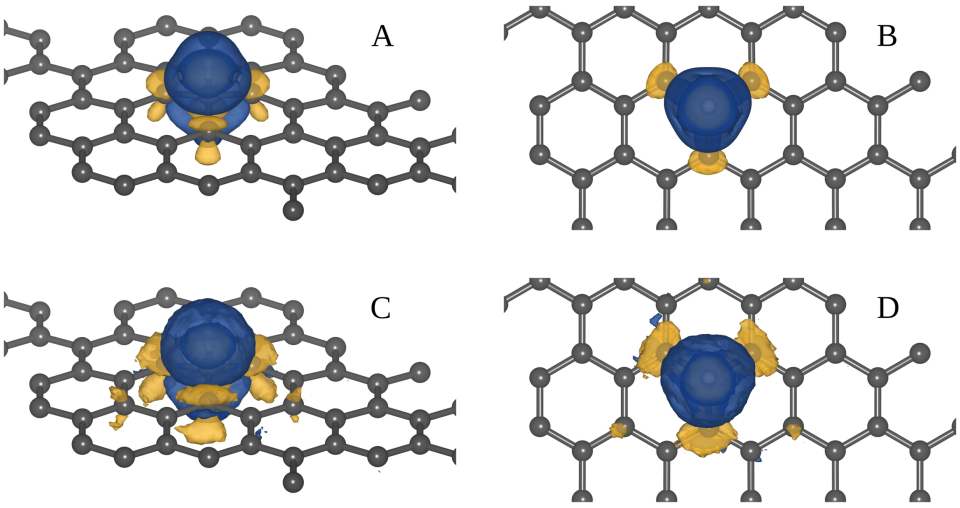


Figure 3: Change of the electron density due to the adsorption of the H atom to the distorted graphene sheet (Eq. 66). ρ_b from PBE calculations is shown from an oblique angle (A) and aligned along the c axis (B). ρ_b from DMC calculations (C) and (D) is shown from the same perspectives. Gold and blue represent a gain and loss of electron density, respectively. Note that there is a region of increased charge density at the C-H bond that is enveloped by a region of loss in the charge density. The binding density was visualized using an isovalue of 2.8×10^{-5} for DMC and 3.9×10^{-5} for PBE, in both cases capturing 95% of the differential charge density.

4.4.3 Charge Density Differences Between DMC and PBE

In this section, the difference between the DMC and PBE charge densities for distorted graphene with the adsorbed hydrogen atom as well as for pristine planar graphene without the adsorbed hydrogen atom are considered. The charge density difference for each system is calculated according to

$$\Delta\rho_{system} = \rho^{DMC}_{system} - \rho^{PBE}_{system}, \tag{67}$$

where ρ^{DMC}_{system} is the DMC charge density of a given system (either distorted graphene with the adsorbed hydrogen or pristine graphene) and ρ^{PBE}_{system} is the corresponding PBE charge density. $\Delta\rho_{gr}$ and $\Delta\rho_{dgr+H}$ are reported in Figure 4 along the 110 slice through the unit cell, which captures the carbon-hydrogen bond. From the top-down perspective in Figure 2, the 110 lattice plane bisects the cell diagonally through the longer of the two diagonals and is indicated by the solid cyan line. In Figure 4, blue represents areas where the PBE density is larger, while gold areas represent areas where the DMC density is larger. The DMC density, in comparison with the PBE density, has greater weight in the bonding region between atoms. We note that the HSE density displays similar differences as the PBE density. Figure S2 of the Supplementary Material includes a visualization of the DMC-HSE density difference. This is the case for both the planar graphene without hydrogen and the system with hydrogen chemisorbed to graphene. Even though there are significant differences between the PBE and DMC densities for both systems, the difference is similar in the two systems, consistent with it not introducing a large error in the PBE value of the binding energy.

4.5 Conclusions

Calculations of the binding energy of a hydrogen atom on a graphene sheet were carried out using various DFT methods and with DMC. The DMC calculations provide a benchmark value of the binding energy. Our best estimate of the binding energy from DMC calculations is -691 $\pm$ 19 meV. The PBE result obtained with a plane-wave basis set gives a binding energy about 20% larger in magnitude than the DMC result. The global hybrid functional, PBE0, gives a binding energy close to that of PBE. In comparison, HSE, a range-separated hybrid functional, gives a smaller

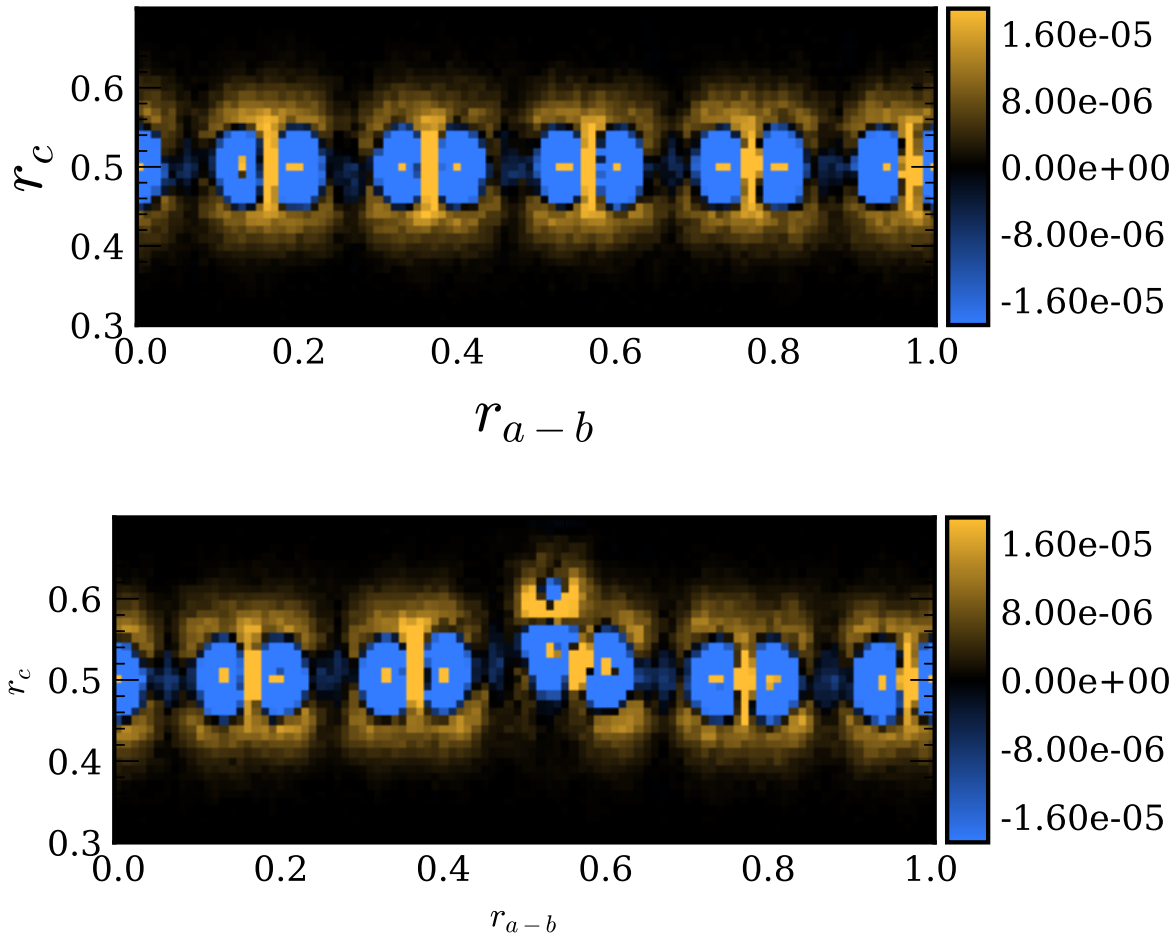


Figure 4: Visualization of the difference of PBE and DMC densities sliced along the 110 lattice plane of the unit cell for the graphene sheet, $\Delta\rho_{gr}$, (top) and H adsorbed onto graphene, $\Delta\rho_{dgr+H}$, (bottom). The abscissa represents traversing the 110 plane in fractional coordinates, while the ordinate represents traversing the c axis in fractional coordinates. Blue regions represent places where the PBE density is larger, while the gold color represents regions where the DMC density is larger.

binding energy of -743 meV, after a correction applied for the basis set incompleteness error, and is much closer to the value from DMC calculations. Interestingly, there are significant differences in the DMC and PBE charge densities of both graphene and H/graphene.

4.6 Acknowledgements

We thank Dr. Dan Sorescu for helpful discussion and for sharing the coordinates of his calculations. A.B. and H.S were supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division, as part of the Computational Materials Sciences Program and Center for Predictive Simulation of Functional Materials. An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. The DMC and plane wave DFT calculations used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under contract DE-AC02-06CH11357. The DFT calculation using Gaussian orbitals were carried out on computing resources in the University of Pittsburgh's Center for Research Computing. K.D.J. acknowledges NSF (CBET-2028826) for partial support of this work. S.U. was supported in part by the Pittsburgh Quantum Institute (PQI) Graduate Quantum Leader Award.

4.7 Data Availability Statement

The data that support the findings of this study are openly available on the Materials Database Facility at `https://acdc.alcf.anl.gov/mdf/detail/dumi_dmc_hgraphene_v1.3`, with the following DOI: 10.18126/s1wc-tya.

5.0 The Role of High-Order Electron Correlation Effects in a Model System for Non-valence Correlation-bound Anions

The test and figures in this chapter have been adapted from The Role of High-Order Electron Correlation Effects in a Model System for Non-valence Correlation-bound Anions, *J. Chem. Phys.*, **2020**, 153, 224118, DOI: 10.1063/5.0030942 with the permission of AIP. The author's contribution to the work included performing CIPSI calculations and the subsequent DMC calculations, along with performing analysis and assisting with figures, and editing/revising the manuscript.

5.1 Summary

Non-valence correlation bound anions (NVCB) are molecules in which the primary binding interaction comes from electron correlation, thus methods which capture electron correlation are necessary for an accurate description. The diffusion Monte Carlo (DMC), auxiliary field quantum Monte Carlo (AFQMC), and equation-of-motion coupled cluster (EOM-CC) methods are used to calculate the electron binding energy (EBE) of the non-valence anion state of a model $(H_2O)_4$ cluster. The viability of the CIPSI procedure to select determinants to describe the binding of the electron for the NVCB system is explored. Two geometries are considered, one at which the anion is unbound and the other at which it is bound in the Hartree-Fock (HF) approximation. It is demonstrated that DMC calculations can recover from the use of a HF trial wave function that has collapsed onto a discretized continuum solution, although larger electron binding energies are obtained when using a trial wave function for the anion that provides a more realistic description of the charge distribution, and, hence, of the nodal surface. For the geometry at which the cluster has a non-valence correlation-bound anion, both the inclusion of triples in the EOM-CC method and the inclusion of supplemental diffuse d functions in the basis set are important. DMC calculations with suitable trial wave functions give EBE values in good agreement with our best estimate EOM-CC result. AFQMC using a trial wave function for the anion with a realistic electron density gives a value of the EBE nearly identical to the EOM-CC result when using the same basis set. For the

geometry at which the anion is bound in the HF approximation, the inclusion of triple excitations in the EOM-CC calculations is much less important. The best estimate EOM-CC EBE value is in good agreement with the results of DMC calculations with appropriate trial wave functions.

5.2 Introduction

In recent years, there has been growing interest in a class of anions known as non-valence correlation-bound (NVCB) anions in which long-range correlation effects are crucial for the binding of the excess electron.[18–31] By definition, NVCB anions are unbound in the Hartree-Fock (HF) approximation. Due to their highly spatially extended charge distributions, large, flexible basis sets are required for the theoretical characterization of NVCB anions. However, with such basis sets, the wave function from Hartree-Fock (HF) calculations on the excess electron system collapses onto the neutral plus an electron in an orbital that can be viewed as a discretized representation of a continuum solution.[18] Methods that start from the HF wave function including second-order Møller-Plesset perturbation theory (MP2)[165] or coupled-cluster singles and doubles with perturbative triples (CCSD(T))[166] do not recover from this collapse onto the continuum, while methods such as orbital-optimized MP2 (OOMP2)[167] or Bruckner coupled-cluster[168] can overcome this problem.[18] The majority of calculations of NVCB anions have employed the equation-of-motion coupled-cluster singles and doubles (EOM-CCSD) method.[169] Among the NVCB anions studied computationally to date are C_{60}, C_6F_6, TCNE, $(NaCl)_2$, Xe_n clusters, large polyaromatic hydrocarbons, and certain $(H_2O)_n$ clusters.[18–25,30,31]

The EOM-CCSD method displays an $\mathscr{O}(N^6)$ scaling with system size, and higher order EOM-CC methods are even more computationally demanding. As a result, most of the calculations of NVCB anions carried out to date have not been fully converged with respect to basis set or the level of excitations treated in the EOM procedure. We note, however, that by using domain-based local pair natural orbitals (DLPNO), electron affinity EOM-CCSD calculations have recently been carried out on systems described by up to 4,500 basis functions.[80]

In the present work, we apply two quantum Monte Carlo (QMC) methods to the problem of calculating the electron binding energy (EBE) of the non-valence anion of a model $(H_2O)_4$ cluster.

The first approach considered is fixed-node diffusion Monte Carlo (DMC),[94,170–172] using various single Slater determinant (SD) and multideterminant (MD) trial wave functions. DMC is a real-space method, with the major sources of error resulting from the use of finite time steps and the fixed-node approximation. The finite time step error can be largely eliminated by running calculations at different time steps and then extrapolating to the zero time step limit. The fixed-node error results from imposition of a nodal surface via a trial wave function, which is necessary to ensure Fermionic behavior, and can be addressed by a variety of means including expanding the number of Slater determinants in the trial wave function or by applying the backflow transformation.[173] It is important to note that, by virtue of working in real space, fixed-node DMC energies are much less sensitive to the choice of the atomic basis set than methods such as EOM-CCSD that operate in a space of Slater determinants.

The second QMC approach considered is the auxiliary field QMC (AFQMC) method.[174–180] AFQMC calculations sample an over-complete space of nonorthogonal Slater determinants. The finite time step error can be mitigated as in DMC. The error that arises from constraining the phase of the wave function to zero can be systematically reduced by improving the trial wave function. Phaseless AFQMC is additionally subject to the limitations of the atomic basis set employed. DMC scales as $\sim \mathscr{O}(N^3)$ with system size, while AFQMC displays an $\sim \mathscr{O}(N^4)$ scaling in most implementations. One of the goals of these calculations is to determine whether DMC calculations can recover from the use of a trial wave function that has collapsed onto a discretized continuum orbital in the case of the excess electron. Additionally, we explore whether correlation effects that are missing in EOM-CCSD are important for electron binding.

In our calculations, we employ a model $(H_2O)_4$ cluster that has been investigated in earlier studies by our group.[18,25] In this model, depicted in Figure 5, the monomers are arranged so that the net dipole moment is zero. If the distance R is varied, with all other geometrical parameters held fixed, the system can be tuned from a regime (large R) that the excess electron weakly binds in the HF approximation to one (small R) at which it is not bound in the HF approximation. i.e., at which it is NVCB in nature.

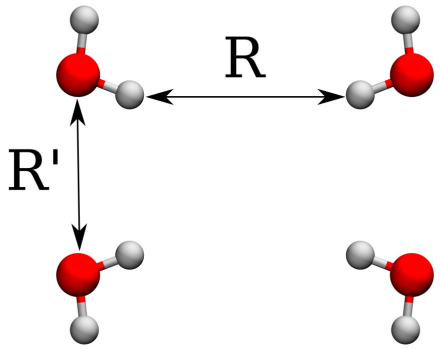


Figure 5: The model $(H_2O)_4$ system considered in this study. R' held fixed at 3.461 05 Å, and R is either 4 Å or 7 Å. Image generated using VMD.[181]

5.3 Computational Details

5.3.1 EOM Coupled Cluster

The EOM methods considered in this study are EOM-MP2,[182] EOM-CCSD,[169] EOM-CCSD(T)(a)*,[183] and EOM-CCSDT,[184,185] listed in order of increasing sophistication in terms of treatment of correlation effects. In the EOM-MP2 and EOM-CCSD methods, the neutral molecule is treated at the MP2 and CCSD levels, respectively, and the amplitudes from these calculations are used to perform unitary transformation of the Hamiltonian. This "dressed" Hamiltonian is then used to carry out a 1-particle plus 2-particle-1-hole CI calculation on the anion. In the EOM-CCSDT method, the neutral species is first treated at the CCSDT level, and the transformed Hamiltonian is used to do CI calculation on the anion that includes up to 3-particle-2-hole configurations. The EOM-CCSD(T)(a)* method includes in an approximate manner both triple excitations in the ground state coupled cluster calculations and 3-particle-2-hole excitations in the treatment of the anion.[183]

The main basis set used for the EOM calculations reported in this study is aug-cc-pVTZ+7s7p, formed by supplementing the aug-cc-pVTZ Gaussian-type orbital (GTO) basis set[186,187] with a 7s7p set of diffuse functions centered at the middle of the cluster and similar to the set from Ref.[18]. The exponents of the supplemental functions start at 0.023622, with each successive exponent

Table 2: Dependence of the total energies and the EBE of the model $(H_2O)_4$ cluster at R = 4 Å on the supplemental diffuse basis functions. Results obtained using the EOM-CCSD method.

basis set	neutral (Ha)	anion (Ha)	EBE (meV)
aug-cc-pVTZ	-305.327947	-305.331344	92.4
aug-cc-pVTZ+1s	-305.327953	-305.332359	119.9
aug-cc-pVTZ+2s	-305.327957	-305.334226	170.6
aug-cc-pVTZ+3s	-305.327958	-305.334460	176.9
aug-cc-pVTZ+7s	-305.327958	-305.334462	177.0
aug-cc-pVTZ+7s1p	-305.327979	-305.334604	180.3
aug-cc-pVTZ+7s7p	-305.327987	-305.334622	180.6
aug-cc-pVTZ+3s1p	-305.327979	-305.334602	180.2

being smaller by a factor of 3.2. However, as seen from Table 2, the supplemental 7s7p set of diffuse functions can be truncated to 3s1p without significantly impacting the EBE as calculated at the EOM-CCSD level. Moreover, as shown in Table 3, expanding the main basis set (i.e., the non-supplemented portion) from aug-cc-pVTZ to aug-cc-pVQZ[186,187] makes only a small impact on the EBE (4% at R = 4 Å) . In contrast, reducing the main basis set to aug-cc-pVDZ[186,187] leads to a 14% reduction in the EBE. (These results were obtained using the EOM-MP2 method, but as seen from comparison of the results in Tables 2 and 3, using the aug-cc-pVTZ+3s1p basis set in both cases, the EBEs from the calculations with the EOM-CCSD and EOM-MP2 methods agree to within 0.5 meV.) The smaller aug-cc-pVDZ+3s1p basis will be used in the EOM-CCSDT calculations, which would have been computationally prohibitive with aug-cc-pVTZ+7s7p or aug-cc-pVTZ+3s1p basis sets. Finally, EOM-CCSD(T)(a)* calculations were carried out with aug-cc-pVTZ+3s1p3d basis sets, where the exponents of the d functions match those of the s and p functions, to assess the importance of supplemental d functions on the EBEs. The EOM calculations utilized the frozen core approximation and were carried out using the Coupled-Cluster techniques for Computational Chemistry (CFOUR) program package.[188,189]

Table 3: Sensitivity of the EBE of the $(H_2O)_4$ model to the "core" basis set. Results obtained using the EOM-MP2 method.

	Neutral (Ha)	Anion (Ha)	EBE (meV)
	R = 4.0 Å		
aug-cc-pVDZ+3s1p	-305.0371957	-305.0428558	154.0
aug-cc-pVTZ+3s1p	-305.3092869	-305.3159306	180.8
aug-cc-pVQZ+3s1p	-305.4008845	-305.4078074	188.4
	R = 7.0 Å		
aug-cc-pVDZ+3s1p	-305.0383747	-305.0432259	132.0
aug-cc-pVTZ+3s1p	-305.3104923	-305.3157472	143.0
aug-cc-pVQZ+3s1p	-305.4021640	-305.4075716	147.1

5.3.2 DMC

The DMC calculations were carried out using trial wave functions represented as products of one or more Slater determinants with a Jastrow factor with one-, two-, and three-body terms.[160,190,191] The parameters in the Jastrow factors were optimized using variational Monte Carlo (VMC), and the resulting trial wave functions were then employed in subsequent DMC calculations. Three types of SD trial wave functions were employed. These used HF orbitals, Becke-Lee-Yang-Parr (B3LYP) DFT orbitals,[192–195] and natural orbitals (NOs) from small restricted single plus double excitation configuration interaction (SDCI) calculations designed to bind the excess electron when it is not bound in the HF approximation. In addition, DMC calculation were carried out using MD trial wave functions, with the determinants being determined either from the restricted SDCI procedure or from configuration interaction using a perturbative selection made iteratively (CIPSI) calculations.[32] Details on these calculations are provided below.

To reduce the computational cost of the DMC calculations, the ccECP pseudopotentials[136,137] were employed together with GTO basis sets that we designate as cc-pVDZ / ccECP, aug-cc-pVDZ / ccECP, aug-cc-pVDZ / ccECP+3s1p, and aug-cc-pVDZ / ccECP+7s7p. The "core" cc-pVDZ /

ccECP[136,137] basis set was designed for use with the ccECP pseudopotentials; the "aug" indicates that the diffuse aug functions from the aug-cc-pVDZ basis sets of Dunning and co-workers are included; and the 7s7p set of diffuse functions are those described above in the Section 5.3.1.[187] The T-moves scheme was used to control the localization error for nonlocal pseudopotentials.[196]

The double-zeta rather than the larger triple-zeta basis set was used as the core basis set due to the relative insensitivity of DMC calculations to the choice of the atomic basis set. For most of the DMC calculations a fixed population of 16,000 walkers and time steps of 0.001, 0.003, and 0.005 a.u. were employed, with the reported results obtained by linear extrapolation to zero time step. However, this population is much larger and the time steps much smaller than what is actually required to achieve well converged energies with minimized finite time step and fixed population errors. Indeed, DMC calculations using Hartree-Fock trial wave functions, larger time steps (specifically 0.05, 0.1, and 0.2 a.u.) and a smaller population of only 1,000 walkers produce an electron binding energy within error bars of that obtained using the smaller time steps and larger populations. Additionally, a DMC calculation with a B3LYP trial wave function with a time step of 0.05 is in agreement with the values obtained with the smaller time steps and larger populations suggesting that these parameters do not depend strongly on the choice of starting orbitals. In light of this, the 0.05 a.u. time step and smaller walker population were employed in the DMC calculations using CIPSI trial wave functions to mitigate the additional cost associated with the MD space. The VMC and DMC calculations were carried out using the QMCPACK code.[160,161] The orbitals for the SD-based trial wave functions and the restricted SDCI MD wave function were both generated using the General Atomic and Molecular Electronic Structure System (GAMESS) program package,[197–199] whereas the CIPSI wave functions were generated using the Quantum Package 2.0 code.[33]

5.3.3 Restricted CI and CIPSI-generated Trial Wave Functions for DMC Calculations

The restricted SDCI procedure employed the HF wave function for the neutral molecule and a specially tailored SDCI wave function for the anion, which included all symmetry-allowed single and double excitations, with the latter restricted so that one of the electrons excited is from the orbital occupied by the excess electron in the HF wave function. This approach, when used with

a flexible basis, gives a bound anion. NOs were generated from the SDCI wave function of the anion and were used in a SD trial wave function for subsequent DMC calculations. In addition, the SDCI wave function itself (expanded in terms of HF orbitals) was used in MD DMC calculations on the anion for R = 4 Å. In this case, a threshold of 0.001 on the magnitude of coefficients in the CI expansion was used in choosing the retained determinants. This resulted in a wave function with 1,392 Slater determinants.

By design, the restricted SDCI wave function does not allow for change of the correlation energy of the valence electrons due to the presence of the excess electron. This possibility is allowed for in the CIPSI MD trial wave functions. The CIPSI calculations were carried out using B3LYP orbitals rather than Hartree-Fock orbitals because the former avoids the problem of collapse onto a discretized continuum solution at R = 4 Å.[192–194] Since the CIPSI calculations have not approached the full configuration interaction limit as indicated by the second-order perturbative correction to the energy, a judicious choice of starting orbitals is required to construct a physically meaningful trial wave function. In order to generate compact wave functions for both the anion and the neutral, NOs were iteratively refined through successive CIPSI calculations, each beginning from a single determinant reference of natural orbitals from the previous iteration. For each NO-generating CIPSI calculation, approximately 100,000 determinants were retained and used to generate NOs for the next iteration, for a total of six NO generation cycles. With the determinant of resulting NOs as a reference, a final CIPSI calculation was carried out, stopping when at least 150,000 determinants were included in the variational space for the anion and at least 100,000 determinants for the neutral. The resulting determinant spaces were used as the DMC trial wave functions.

Both the restricted SDCI and the CIPSI calculations used to generate the trial wave functions for subsequent DMC calculations were carried out using the ccECP pseudopotentials. The aug-cc-pVDZ/ccECP+7s7p and aug-cc-pVDZ/ccECP+3s1p basis sets were used for the CIPSI and CIPSI calculations, respectively.

5.3.4 AFQMC

AFQMC[174–180] utilizes the Hubbard-Stratonovich transformation[200] to represent the imaginary-time propagator as a multi-dimensional integral over auxiliary-fields. Ground-state properties are

sampled from a random walk in the space of non-orthogonal Slater determinants subject to the phaseless constraint[177] introducing a bias which can be systematically reduced based on the quality of the nodal surface of the trial wave function employed. While sophisticated trial wave functions generated from regularized orbital-optimized MP2 (κ-OOMP2)[201] or complete active space self-consistent field (CASSCF)[202–204] are required to obtain quantitative predictions for some biradicaloids and transition metals, high accuracy has been obtained, even for systems exhibiting non-trivial electron correlation such as dipole-bound anions,[205] with single-determinant trial wave functions consisting of HF or Kohn-Sham orbitals.[205,206]

In this work, we perform calculations with a graphics processing unit (GPU) implementation of AFQMC,[207] utilizing single-precision floating-point arithmetic and two-electron integrals decomposed via a modified Cholesky decomposition (10^{-5} cutoff).[208] These calculations made use of the aug-cc-pVTZ+7s7p basis set, a small imaginary-time step of 0.005 a.u, and correlated all electrons. For the neutral species and electrostatically bound anion (R = 7 Å), the Hartree-Fock wave function was used as the trial wave function. For the NVCB anionic species (R = 4 Å), a SD trial wave function comprised of natural orbitals from the restricted SDCI calculation as detailed in Section 5.3.3 (but now carried out without pseudopotentials) was used.

5.3.5 Radial Orbital Densities

To compare the description of the charge distribution of the excess electron as calculated using different theoretical methods, we generate radial electron density plots. This choice is motivated by the fact that the excess electron occupies an orbital belonging to the totally symmetric representation. The radial electron densities are generated by numerically integrating over the angular components of the singly occupied molecular or natural orbital. First, Molden files are created from the output data from the various generating programs using cclib when supported.[209] With the Molden files as input, PySCF is used to generate the electron density on a uniform radial grid and 5810 point Lebedev-Laikov angular grid as tabulated in quadpy.[210–213] Finally, a numerical integration is performed over the angular components. An example of this workflow is presented in detail in the Supplementary Information.

5.4 Results and Discussion

The EBEs obtained from the EOM and AFQMC calculations are summarized in Table 4, and the results from the various DMC calculations are summarized in Table 5. We consider first the results obtained for R = 4 Å, for which HF calculations do not bind the excess electron.

5.4.1 Results for R = 4 Å: the Correlation Bound Region

From Table 4, it is seen that the EOM-CCSD/aug-cc-pVTZ+7s7p calculations give a value of the EBE of 181 meV for the $(H_2O)_4$ cluster model at R = 4 Å. This increases to 196 meV with the EOM-CCSD(T)(a)* method. The AFQMC calculations using the same basis set and for the anion a single determinant of NOs from the restricted SDCI calculation for the trial wave function produce an EBE value of 194 $\pm$ 10 meV, comparable to the EOM-CCSD(T)(a)* result. The EOM-CCSD(T)(a)* and EOM-CCSD(T) EBE values calculated with this basis set are nearly identical, demonstrating that the approximate treatment of triples in the former procedure introduces a negligible error in the EBE. The contribution of supplemental diffuse functions was checked using the EOM-CCSD(T)(a)* method and the aug-cc-pVTZ+3s1p2d basis set. These calculations reveal that the inclusion of the supplemental diffuse d functions leads to a $\sim$10 meV increase in the EBE. With the inclusion of this correction, we obtain an estimated EOM-CCSDT EBE of 212 meV. It is expected that the inclusion of the supplemental d functions in the basis set used for the AFQMC calculations would lead to a similar increase in the EBE obtained using that method.

The restricted SDCI procedure, by itself, is not expected to give an accurate value of the EBE and is designed to generate appropriate trial wave functions for DMC or AFQMC calculations on the anion. In fact, the EBE resulting from the HF treatment of the neutral and the restricted SDCI treatment of the anion using the aug-cc-pVTZ+7s7p basis set is 345 meV, appreciably larger than the EOM and AFQMC values. This over-binding is due in part to the fact that the restricted SDCI wave function, like the HF wave function, overestimates the magnitude of the dipole moment of the water molecules, resulting in a too favorable electrostatic interaction. We also constructed a single determinant trial wave function for the anion using the natural orbitals of the restricted SDCI expansion. We note also that the single determinant of NOs generated from the restricted

SDCI wave function and using the aug-cc-pVTZ+7s7p basis set places the anion 160 meV above the neutral when the latter is treated in the HF approximation. This is not surprising since this calculation neglects correlation effects other than those incorporated in the determination of the orbitals. What is important is that the approaches based on the restricted SDCI procedure provide a realistic description of the orbital occupied by the excess electron and avoid the collapse onto the discretized continuum as was observed with the HOMO in the HF calculations.

In light of the close agreement between the EOM-CCSD(T)(a)* and AFQMC values of the EBE of the $(H_2O)_4$ model at R = 4 Å, when using a comparable basis sets in the two approaches it is relevant to determine whether DMC calculations with sufficiently flexible trial wave functions give an EBE close to the AFQMC and EOM values consistent with these results. DMC calculations using HF trial wave functions together with the aug-cc-pVDZ/ccECP+7s7p basis set give an EBE of 183 $\pm$ 10 meV, appreciably smaller than the EOM-CCSD(T)(a)* and AFQMC values. Interestingly, essentially the same EBE is obtained from the DMC calculations using a Slater determinant of HF orbitals expanded in the aug-cc-pVDZ/ccECP basis set without the 7s7p supplemental set of diffuse functions. However, if the aug diffuse functions are also removed, the DMC calculations fail to bind the excess electron. We believe that this is a consequence of the fact that with the cc-pVDZ basis set there is a near zero probability of sampling regions of space at large distances from the molecule, which are important for describing the charge distribution of the excess electron.

A significantly larger value of the EBE is obtained from SD DMC calculations using B3LYP orbitals in place of HF orbitals. The resulting EBE of 212 $\pm$ 11 meV, within statistical error, agrees with the EOM-CCSD(T)(a)* and AFQMC values. A similar value of the EBE is obtained from DMC calculations using a single determinant of HF orbitals for the neutral cluster and a single determinant of natural orbitals from the restricted SDCI procedure described in Section 5.3.3 for the anion. DMC calculations using a SD of HF orbitals for trial wave function of the neutral and a trial wave function for the anion retaining 1,392 of the most important determinants from the restricted SDCI calculation gives an EBE of 202 $\pm$ 12 meV, close to the values obtained using the single determinants B3LYP orbitals or of NOs from the SDCI calculation (for the anion). The DMC value of the EBE resulting from the anionic trial wave function using a SD of NOs from the

[1] SD/X indicates that the trial wave function employed a single Slater determinant with X (either HF or B3LYP) orbitals. When different types of trial wave functions are used for the neutral (N) and anion (A) this is indicated by the double slash.

Table 4: EBEs of the $(H_2O)_4$ model calculated using HF, EOM, and AFQMC methods and employing the aug-cc-pVTZ+7s7p basis set.

Method	EBE (meV)
R = 4.0 Å	
HF	-0.4
EOM-CCSD	180.6
EOM-CCSD(T)(a)*	195.8
EOM-CCSDT	197.5[1](212.0)[2]
AFQMC SD/HF(N)//SD/NO SDCI(A)	194 ± 10
R = 7.0 Å	
HF	41.3
EOM-CCSD	140.2
EOM-CCSD(T)(a)*	141.7
EOM-CCSDT	143.3[1](154.2)[2]
AFQMC SD/HF	181 ± 5

[1] This EOM-CCSDT/aug-cc-pVTZ+7s7p value was estimated by adding the difference of EBEs from the EOM-CCSD(T)(a)* and EOM-CCSDT calculations with the aug-cc-pVDZ+3s1p basis set to the value from EOM-CCSD(T)(a)*/aug-cc-pVTZ+7s7p.

[2] The EOM-CCSDT/aug-cc-pVTZ+7s7p3d value was estimated by adding the difference between the EBEs calculated with the EOM-CCSD(T)(a)* with the aug-cc-pVTZ+3s1p and aug-cc-pVTZ+3s1p3d basis sets to the EOM-CCSDT/aug-cc-pVTZ+7s7p estimated value in footnote [1] to assess the effect of incorporating diffuse d functions into the basis.

Table 5: EBEs of the $(H_2O)_4$ model calculated using the DMC method and various trial wave functions[1]

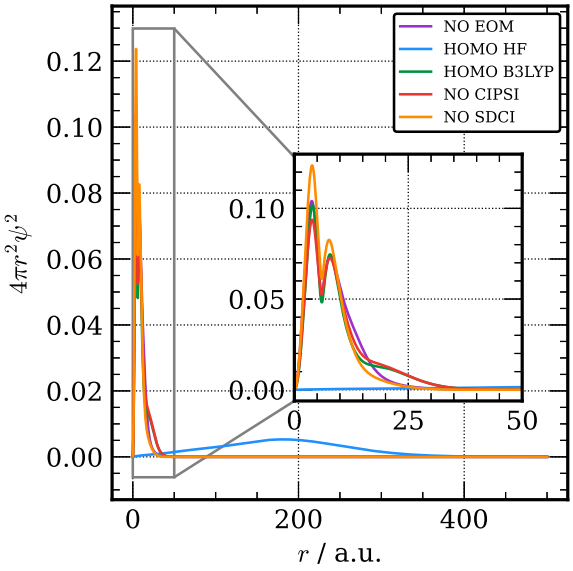
wave function	basis set	EBE (meV)
	R = 4.0 Å	
SD/HF	aug-cc-pVDZ+7s7p	183 ± 10
SD/HF	aug-cc-pVDZ	176 ± 12
SD/HF	cc-pVDZ	-528 ± 25
SD/B3LYP	aug-cc-pVDZ+7s7p	212 ± 11
SD/HF(N)//SD/NO SDCI(A)	aug-cc-pVDZ+7s7p	205 ± 10
SD/HF(N)//MD/NO SDCI(A)	aug-cc-pVDZ+7s7p	202 ± 12
MD/CIPSI NO	aug-cc-pVDZ+3s1p	190 ± 9
	R = 7.0 Å	
SD/HF	aug-cc-pVDZ+7s7p	141 ± 14
SD/B3LYP	aug-cc-pVDZ+7s7p	164 ± 9
SD/HF(N)//SD/NO SDCI(A)	aug-cc-pVDZ+7s7p	160 ± 9
MD/CIPSI NO	aug-cc-pVDZ+3s1p	159 ± 8

restricted SDCI MD calculation results is 205 $\pm$ 10meV, similar to that from DMC calculations using as trial wave functions the MD restricted SDCI wave function for the anion and the HF wave function for the neutral.

Figure 6 compares the radial charge distributions of the singly occupied orbital from the HF and B3LYP calculations on the excess electron system as well as of the NOs associated with the excess electron from EOM-CCSD, restricted SDCI and CIPSI calculations. The collapse of the singly occupied orbital from the HF calculations onto a discretized continuum orbital is readily apparent. In contrast, the NOs from the EOM-CCSD and restricted SDCI calculations and the singly occupied orbital from the B3LYP calculation on the anion are more localized and are qualitatively similar to one another. These results are consistent with the nodal surface for the anion being significantly improved when using a SD trial wave function that has a physically reasonable charge distribution for the orbital occupied by the excess electron. Thus, although DMC calculations do recover from the collapse of the HF trial wave function onto a discretized continuum solution in the case of the anion, starting with such a trial function leads to a greater nodal surface error for the anion than for the neutral cluster. However, we also note that the radial distribution function of the singly occupied orbital from the B3LYP calculation on the anion has a spurious peak near 25 atomic units from the center of the cluster. This is likely a consequence of the self-interaction error in the B3LYP functional. The relevant NO extracted from the CIPSI calculations, which were carried using B3LYP orbitals, exhibits a similar shoulder.

Our final set of DMC calculations at R = 4 Å used MD trial wave functions determined from CIPSI calculations for the neutral and anionic clusters. The strategy used in performing the CIPSI calculations was presented in Section 5.3.3, where it was noted that these calculations, unlike those with the restricted SDCI wave functions, allow for the correlation between the valence electrons change due to the the presence of the excess electron. The DMC calculations using the CIPSI trial wave function resulted in an EBE of 190 $\pm$ 9 meV for R = 4 Å, slightly under-binding compared to the single determinant DMC value of the EBE obtained using B3LYP orbitals though in close agreement with the results of DMC calculations carried out with the restricted SDCI trial wave function.

Figure 6: Radially integrated charge densities of the singly occupied orbitals from HF and B3LYP calculations and the singly occupied natural orbital from EOM-CCSD, SDCI, and CIPSI calculations of the model ($(H_2O)_4$) cluster anion at R = 4 Å. All plots generated using Matplotlib.[163]



.

5.4.2 Results for R = 7 Å: the Electrostatically Bound Region

We now consider the results obtained for the $(H_2O)_4$ cluster model at R = 7 Å, for which HF calculations with the aug-cc-pVTZ+7s7p basis set bind the excess electron by 41 meV. In this case, the EOM-CCSD and EOM-CCSD(T)(a)* calculations give EBEs of 140 meV and 142 meV, respectively. Thus unlike the situation for R = 4 Å, the inclusion of triples in the EOM-CC procedure is relatively unimportant at R = 7 Å. The DMC calculations using SD HF trial wave functions give an EBE of 141 $\pm$ 14 meV, while the DMC calculations using as trial wave functions single determinants of B3LYP orbitals, single determinants generated using the restricted SDCI procedure, or MD trial wavefunctions generated using the CIPSI procedure give similar EBEs values ranging from 159 $\pm$ 8 to 164 $\pm$ 9 meV.

Since the anion is bound in the HF approximation at R = 7 Å, we also were able to calculate EBEs using separate, frozen-core coupled-cluster calculations for the neutral and anion with the following coupled-cluster methods: coupled-cluster singles, doubles, and a perturbative treatment of triples ΔCCSD(T)[166], coupled-cluster singles, doubles, and triples (ΔCCSDT)[214–217], and CCSDT with the perturbative treatment of quadruple excitations (ΔCCSDT(Q)).[218] methods. The Δ indicates that the EBE is derived from the energy difference between the separate calculations on the neutral and anion. The ΔCCSDT and ΔCCSDT(Q) calculations were carried out with only the aug-cc-pVDZ+3s1p basis set. These calculations indicate that full treatment of the triples, and even approximate treatment of the quadruple excitation contributions, has less than a 1 meV effect on the EBE of the $(H_2O)_4$ cluster model at R = 7.0 Å. On the other hand, the inclusion of diffuse d function in the supplemental set of functions leads to a 12 meV increase in the EBE. With this correction we obtain an estimated EOM-CCSDT EBE of 154 meV, which is in good agreement with the DMC results using suitable trial wave functions.

The AFQMC calculations give an EBE of 181 $\pm$ 5 meV, significantly larger than the EOM-CC results or DMC values. This most likely reflects an inadequacy of the HF wave function used for the anion in the AFQMC calculations. Support for this interpretation is provided by examination of Figure 7, which shows the radial charge distribution of the excess electron for the $(H_2O)_4$ model at R = 7 Å. From this figure it is seen that although that the HF wave function has not collapsed onto the continuum as it did in the R = 4 Å cluster, it is still much more diffuse than that from

calculations that include correlation effects. It is also seen from comparisons of Figures 6 and 7 that the charge distribution associated with the NO occupied by the excess electron in the EOM-CCSD calculations for the cluster with R = 7 Å, is more radially extended than that at R = 4 Å. Another noticeable difference between the charge density plots for R = 7 Å and 4 Å is the reduction of the long-range shoulder in the radial charge distribution of the HOMO from the B3LYP calculations on the anion and in the relevant NO from the CIPSI calculations on the anion carried out using B3LYP orbitals, suggesting that self-interaction errors are less problematical at R = 7 Å.

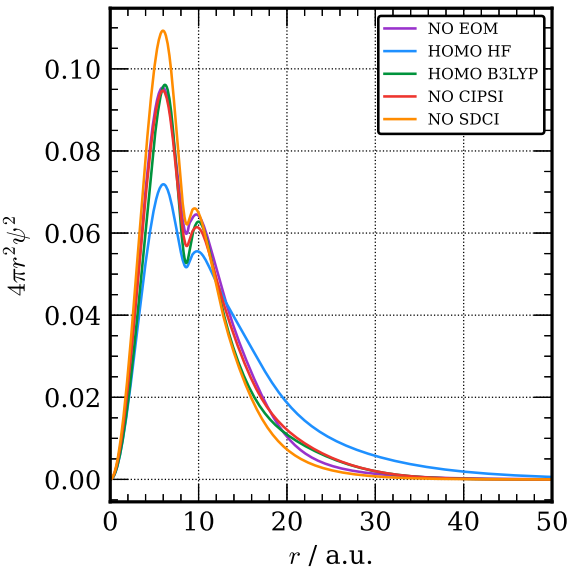
5.5 Conclusion

In this study we have applied various EOM-CC methods and two different quantum Monte Carlo methods to calculate the EBE of a model $(H_2O)_4$ cluster at two geometries, one at which the anion is bound in the HF approximation and the other at which it is not. Diffusion Monte Carlo calculations using single determinant trial functions based on Hartree-Fock orbitals are shown to bind the excess electron even when the initial wave function for the anion has collapsed onto the neutral plus discretized continuum orbital. However, such calculations significantly underestimate the EBE, whereas SD DMC calculations using trial wave functions for the anion with a more realistic charge distribution for the excess electron give larger EBE values that are in close agreement with our best estimate EOM-CCSDT values for both geometries considered.

For R = 4 Å, at which the anion is correlation bound, use of such trial wave functions accurately reflecting the physical charge density resulted in AFQMC-predicted EBE values in agreement with the EOM-CCSD(T)(a)* result (when using comparable basis sets). However, at R = 7 Å, AFQMC calculations with HF trial wave functions significantly overestimate the EBE compared to EOM-CC and DMC values, suggesting the need for an improved trial wave functions in this case. For the $(H_2O)_4$ model system, the restricted SDCI represents an economical way to create trial wave functions for QMC calculations on non-valence anions that are not bound in the Hartree-Fock approximation. However, it remains to be seen if this strategy will be as effective for systems in which the neutral species is more strongly correlated than the model $(H_2O)_4$ cluster.

Finally, we note that at R = 4 Å, for which the anion is NVCB in nature, the most frequently

Figure 7: Radially integrated charge densities of the singly occupied orbitals from HF and B3LYP calculations and the singly occupied natural orbital from EOM-CCSD, restricted SDCI, and CIPSI calculations of the model ($(H_2O)_4$) cluster anion at R = 7 Å.



used method to characterize such anions, EOM-CCSD, underestimates the EBE by about 10% compared to the result of EOM-CCSDT calculations. Both DMC and AFQMC are viable alternatives to high order EOM methods, and while more computationally demanding for the $(H_2O)_4$ cluster, they demonstrate lower scaling with system size than EOM methods, making them attractive for the characterization of non-valence anions of much larger systems.

6.0 Characterizing Chemical Fragmentation Definition Through Unsupervised Learning Methods

The author's contribution to the work included development of descriptors, writing of code, running calculations, performing analysis, and writing/editing the manuscript. This work is in preparation and will be submitted to the Journal of Chemical Physics.

6.1 Summary

In order to reduce the computational cost of large calculations, we look to fragmenting large molecules into smaller subsystems which can best represent the whole. This work proposes a scheme for automatic molecular fragmentation through unsupervised learning approaches in which each fragment is chosen to best retain the important features of the bonding environment. This chapter highlights the efforts at benchmarking the performance of our proposed method on a set of test systems. To this end, a set of clustering algorithms (spectral, agglomerative, *k*-means, and affinity propagation) were studied in combination with various molecular representations, including those incorporating bonding information derived from quantum mechanics. The performance for the clustering/descriptor combinations was assessed for test systems spanning a range from easily distinguishable fragments such as non-covalently bound water cluster to oligomers in which lowest-loss fragmentation is ambiguous. Overall, it is found that spectral clustering works well in all systems tested, showing very little sensitivity to the representation employed. Spectral, agglomerative, and *k*-means clustering produce reasonable fragments for systems with clear fragmentation patterns. Though in the oligomer system, spectral clustering achieves the best performance as assessed by offering a trade-off between lowest error and highest speed-up and is thus recommended as the most robust clustering approach for molecular systems. The approach has the potential to improve reproducibility and transferability by replacing manual fragmentation with quantitative partitioning criteria.

6.2 Introduction

Common quantum chemistry methods that provide an accurate description of molecules are often restricted to small systems in terms of number of atoms or basis set size due to high scaling of the computational cost with system size, N. For example, coupled-cluster singles doubles with perturbative triples methods is often regarded as the"gold standard" level of theory, but incurs an N^7 scaling. Full configuration interaction (FCI), which is formally exact in a complete basis, has prohibitive $N!$ scaling. The higher scaling restricts the application of these methods to small systems while many of the chemical processes of interest involve large molecules. To compensate for this high scaling, fragmentation approaches estimate the energy and other properties of large molecular systems by partitioning the system into small subsystems, where the final estimate of the energy becomes the accumulation of the parts. The accuracy of this approach hinges on the electronic structure treatment of each fragment, the approach used to describe the interaction between the fragments, and the way in which the molecule is partitioned. The possible inter- and intra-fragment treatment approaches are vast, but beyond the scope of the current article, though interested readers are directed towards a number of helpful reviews.[40,105] In the fragmentation schemes, the best case scenario for scaling becomes $\mathscr{O}(N^p) \rightarrow N_{\text{frag}}\mathscr{O}(f^p)$, where N_{frag} is the number of fragments, f is representative of the fragment size, and p is the exponential value dependent on the level of electronic structure theory utilized.[40,41] This partitioning of a single, very costly calculation into N_{frag} smaller calculations achieves two important objectives: 1) Computational scaling with system size is reduced with reasonable fragment definition and 2) trivial parallelization is possible by treating subsystems separately, with the potential to efficiently utilize high performance computing resources. In addition to enabling the treatment of larger systems, fragmentation methods can provide detailed insight into interfragment interactions when combined with analysis techniques such as energy decomposition analysis or symmetry adapted perturbation theory.[99–103] Fragmentation approaches can also assist recent efforts of enabling quantum chemistry via quantum computers. Current quantum computation is restricted as the integrity of the results can be sacrificed due to interactions of the hardware with the environment. Fragmentation methods have been suggested as ways to treat larger molecular systems on quantum computers. By breaking molecules into smaller domains before treatment on a quantum computer, the most important chemical interactions can

be efficiently described without succumbing to errors resulting from the quantum computing hardware.[219]

As pointed out by Herbert, the choice of fragments for a system is not well-defined, but affects the quality of results obtained.[40] In some systems, a natural approach towards partitioning arises when there is a stark difference in the types of bonding present in the system, such as in non-covalent molecular clusters. For covalent systems, however, the choice of fragments is not always as clear-cut. In such cases, fragmentation requires the comparison of total energies, dipoles, or polarizabilities. Often the fragment definition is based on predefined functional groups or chosen manually.[42] Some methods of energy estimation are defined based on specific fragmentation schemes such as the systematic molecular fragmentation (SMF) and systematic molecular fragmentation by annihilation (SMFA).[220–224] In these methods, fragments are built around functional groups or larger fragments made from their groupings. Ultimately, the level of fragmentation is at the discretion of the user to achieve the desired level of accuracy. However, functional group definition may become ambiguous. For example, there is no set number of monomers to include from a polymer backbone to acquire and accurate description capture the chemical behavior. Additionally, a fragment definition based only on functional groups may not consider the interacting chemical environment.

A desirable approach to choosing fragments would have low computational cost and prioritize keeping associated molecular components intact to treat fully with quantum mechanics while the estimation of their interactions should occur only at the most weakly bound points. To this end, we propose an approach utilizing clustering methods to identify the strongly interacting substructures of the system, which we term Automatic Fragmentation of Molecules using Clustering (AFMC) approach. Clustering methods are a form of unsupervised machine learning used to identify substructures in data sets, as a result these approaches are fundamental to data-mining procedures. Previous work in chemistry utilized clustering methods to identify structure-property relationships in large databases[225], to determine the number of residues to treat in quantum refinement methods[226], and to partition large proteins into peptides using an amino acids representation using graph based methods.[227] The application of clustering methods to produce logical fragments of individual molecules at an atomic level is an unexplored direction. The motivation is that these unsupervised machine learning algorithms (UML) , given a certain level of molecular information,

will be able to group the atoms interacting most strongly with each other, ensuring that segmentation occurs between atoms which are the most weakly connected. This approach is able to operate independent from functional group definition which will become useful for capturing non-covalent interactions and non-local interactions in materials or biomolecules. Additionally, this approach is expected to overcome shortcomings in other fragmentation approaches such as severing of double bonds or ring structures, since the molecular representation should be designed to avoid this.

The article is organized as follows, Section 6.3.1 will discuss the clustering approaches used, Section 6.3.2 will describe the representation of chemical data used as input for the clustering approach, with validation methods described in Section 6.3.3, results and a discussion on clustering performance follow in Section 6.4.

6.3 Methods

In UML approaches, the clustering depends fundamentally on two factors: the features used to describe each data point and the algorithm used to identify domains within the data. This section introduces UML (clustering) approaches along with molecular representations (features) upon which fragment selection is based. Additionally, metrics are discussed to assess the quality of a chosen partitioning and molecular test systems are presented. Four representations are explored in this work and are described in Section 6.3.2. Several clustering approaches selected from a range of different families of algorithm were considered. The main article focuses on agglomerative, *k*-means, and spectral clustering; additional algorithms tested can be found in the supporting information. These clustering methods are described in Section 6.3.1. The code used to explore automatic molecular fragmentation using clustering can be found at `https://github.com/amandadumi/molfrag`. Clustering approaches are utilizing the implementations contained within the scikit-learn Python package.[228]

6.3.1 Clustering approaches

This subsection describes the clustering methods explored in this work. This work explores the application of three clustering algorithms, representing different approaches for the selection of fragments: agglomerative, k-means, and spectral clustering.[229–233] At a minimal level, all chosen algorithms require only one user input: the number of subsets to identify within the data, i.e. the desired number of fragments. Starting from this user input, the clustering approaches automatically determine fragments based on selected molecular representations, as outlined below.

Agglomerative clustering[234] is performed in a bottom-up fashion where in each iteration the most similar clusters are merged. In the initial iteration, all of the atoms are regarded as individual clusters, which are then merged into larger clusters. Merging occurs according to a linkage criterion which describes the similarity (or distance) between clusters. In this work, the Ward linkage criterion is used, which chooses which clusters so that the variance of Euclidean distances within each cluster is minimized. Here, the variance of a cluster is calculated as the residual sum of squares (RSS) of all variables in a cluster C,

$$d_C = \sum_{i,j \in C} ||x_i - x_j||^2, \tag{68}$$

where i, j are observations (atoms) within a cluster with associated data points x. The Ward criterion[234] results in a more regular distribution of cluster sizes compared to other choices, which is advantageous for the speed in a fragment calculation, as even distribution of fragment sizes is associated with equal computational cost distribution. This process is repeated until the user-specified number of fragments has been obtained. A visualization of this algorithm is shown in Figure 8, where the progression of this algorithm can be understood as a dendrogram. The dots are grouped through iterations until the desired number of clusters are found. The time complexity of the agglomerative clustering method is $\mathcal{O}(n^3)$, but can be reduced to $\mathcal{O}(n^2)$ with various optimizations, where n is the number of data points.[229] A shortcoming of this method is that it is a greedy algorithm, or in other words once an atom is assigned to a fragment only considering the local environment and this assignment will not be reassessed.

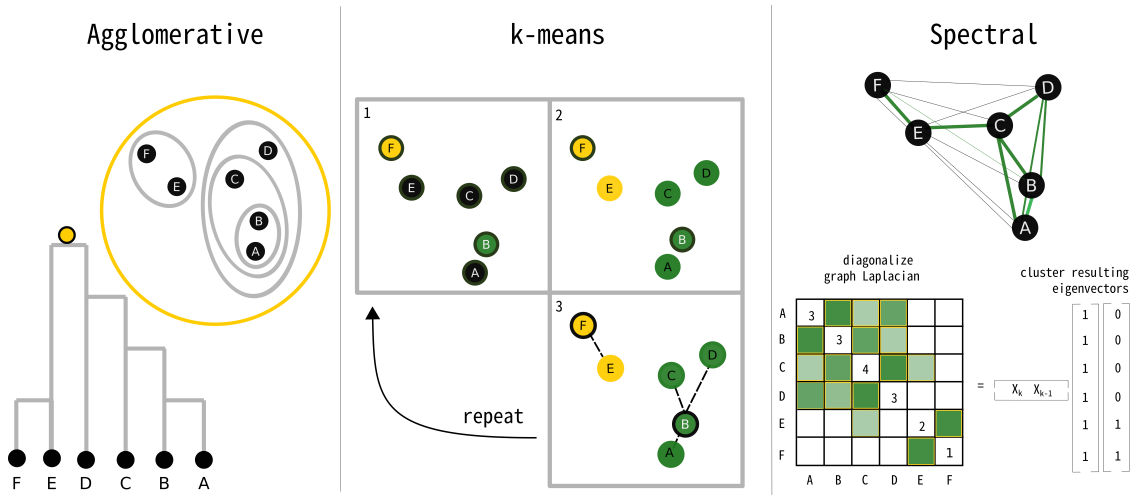


Figure 8: Visual representations of the three clustering approach explored in this work. The dots are a minimal example of a set of data. Agglomerative clustering (left), k-means clustering (center), and spectral clustering (right). Detailed descriptions of each method are found in the text.

The k-means clustering approach which iteratively minimizes the distance between the data points x_i and the cluster centers $\bar{x}_c$ according to:

$$\text{Minimize} \left\{ J \equiv \sum_{c=1}^{k} \sum_{i=1}^{n_c} \left\| x_i^{(c)} - \bar{x}_c \right\|^2 \right\} \tag{69}$$

where c denotes a particular cluster, $x_i^{(c)}$ a data point within this cluster, and $\bar{x}_c$ is the centroid associated with cluster c. k is the number of fragments requested by the user. During the first iteration, the centroids are chosen randomly among the data points. Data points are then assigned to different centroids (clusters) based on shortest distances; centroids are updated by averaging over the data points associated with a given cluster,

$$\bar{x}_c = \sum_{j \in c} x_j^{(c)}. \tag{70}$$

This procedure of assigning points based on proximity to the centroids and update of centroids is repeated iteratively until changes between centroids falls below a threshold. An illustration of this approach is given in Figure 8.

It is interesting to note some differences and similarities between agglomerative and k-means clustering. First, it is noted that at first glance the sum of squares used as objective function within k-means clustering is related to Ward's linkage criteria used within agglomerative clustering. However, an important difference is that the k-means approach is non-greedy, meaning that a reassignment of points between different clusters is possible between different iterations. This property implies that k-means has the potential to generate a global solution, whereas agglomerative clustering aims at a locally optimal solution. Another interesting difference is that k-means clustering is an iterative procedure, meaning that termination and thus runtime depend on the specific data, whereas agglomerative clustering is guaranteed to terminate after a certain number of steps when the user-requested number of clusters has been identified. The convergence of k-means clustering implementations, specifically of Lloyd's algorithm[231], varies significantly between average-case and worst-case scenarios. For typical applications, k-means is observed to converge in few iterations, leading to an average-case observed complexity that is linear, $O(n)$, where we use O to denote an approximate, observed scaling. In practice, k-means is therefore found to be more readily applicable to larger data sets compared to agglomerative clustering. However, we do note that under worst-case scenarios the number of iterations required for k-means convergence scales in a superpolynomial fashion, leading to lower bounds of $2^{\Omega(\sqrt{n})}$ for the worst-case computational scaling, where Ω denotes a lower bound.[235] Lastly, we note that the k-means procedure is typically repeated for different (random) choices of initial centroids to ensure convergence to a global minimum, since initial centroids may impact the convergence to a specific solution. For this reason, this work reports k-means results over ten runs with different initial centroids for each calculation. In contrast, agglomerative clustering always produces the same results for a given set of input data except in the case of identical cluster distances; in such a case the final clusters would depend on the sorting of the input data.[236] For molecular representations using floating point numbers to map molecular structure, it is unlikely to encounter exact degeneracies. However, some of the molecular representations rounding these representations to integers, it would be possible to encounter degeneracies and thus a dependency of agglomerative clustering results on sorting of the input data.

The third clustering algorithm explored here is spectral clustering.[237–239] A visualization of this method can be seen in Figure 8. Spectral clustering begins by defining an affinity matrix A_{ij} which

describes the similarity of each pair of data points, ij. In the present context, the similarity between two data points can be understood as the presence of a bond or the strength of the bond. From the affinity matrix, a degree matrix, D, is built, which sums the rows of A onto the diagonal of D. The off-diagonal elements of A combined with and D are combined to form the graph Laplacian of the data as $L = D\text{-}A$. The eigenvectors and eigenvalues resulting from the diagonalization of L represent the data in a lower dimension space that leads to clearer separation for linear cuts. Following the spectral decomposition, the eigenvalues are then clustered by another method such as k-means, a discretized approach, or others.[237,238,240]. In this work, the k-means approach is utilized. The time complexity of the spectral clustering method is $\mathcal{O}(n^3)$, where n is the number of data points.

Additionally, the application of affinity propagation and mean shift clustering were explored, but were unsuccessful in producing useful molecular fragments. These methods are of interest since the number of clusters are chosen automatically through an a variety of approaches to analyze the density of a given set of data. However, these methods did not produce consistent results with the molecular representations explored in this work and, in many cases, no viable subsystems resulted from the fragmentation for the representations explored in this work. Although affinity propagation had some success, it was not consistent across test sets. Results for these two additional methods are included in the SI. These finding do not rule out the possible use of these clustering methods for chemical fragmentation as a tailored the molecular representation for this method may be needed.

6.3.2 Molecular Representations

This subsection describes the molecular representation used to describe the chemical system. The success of clustering depends on the representation, i.e. the features used to describe the relationship between the data points (here atoms). This work explores the application of four different representations. Two descriptors are derived from structure information alone: Cartesian-based and a covalent radii based bond matrix-based descriptors. The structure-only derived representation provide a low-cost descriptor as no quantum mechanical information is incorporated. Alternatively, incorporating bonding information from a quantum mechanical treatment should provide a more detailed descriptor, though at a higher computational cost. Two descriptors which incorporate

quantum mechanic information are explored: the Mayer bond matrix descriptor and the rounded Mayer bond matrix descriptor.

The descriptors are presented in a way that they represent an affinity or similarity matrix between objects. The clustering methods utilize this information in different ways. Spectral clustering uses the affinity matrix to perform the subspace search, agglomerative clustering will invert the affinity matrix to indicate that those data points more strongly interacting are closer in space, and the k-mean approach will use each row of the affinity matrix as a description of the dimensions in which the vector norm is measured. The representation is the reciprocal of each element which is handled automatically within the `molfrag` code.

The Cartesian representation describes the position of each atom as x, y, and z components. Distances between any two atoms, A and B, are calculated via the conventional Euclidean distance, $R_{AB} = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2 + (z_A - z_B)^2}$. If used as a precomputed similarity matrix to describe the strength of the interactions between molecules, the representation is given as $G^{xyz}_{AB} = \frac{1}{R_{AB}}$. In k-means and agglomerative clustering, the Cartesian coordinates are fed directly to the clustering algorithm. In the spectral clustering algorithms, the Cartesian data must be represented by an affinity matrix. The affinity matrix, A, defined as describes the strength of interactions between data pairs.

When a descriptor exhibits a more block diagonal structure in the descriptor, many clustering methods including those explored here are able to distinguish easily between these sections. To explore if this idea can assist in molecular fragmentation definition, a representation is constructed around a bond matrix based on the covalent radii of the atoms (cr). The covalent radii has been used in supervised machine learning as a way to numerically capture a chemical environment, thus exploring this as a feature for unsupervised machine learning is sensible.[241,242] This is a Boolean matrix in which one indicates the presence of a bond as determined by

$$G^{cr}_{AB} = \begin{cases} 1 \text{ if } R_{AB} \leq 1.1(A_{cr} + B_{cr}) \\ 0 \text{ otherwise} \end{cases}$$

where the G^{cr}_{AB} is the descriptor entry, R_{AB} is the distance between atom A and atom B, and A_{cr} (B_{cr}) is the covalent radii of A (B).[243,244]

The remaining two representations incorporate information of the bonding environment via the Mayer bond order as a surrogate for density matrix and thereby quantum mechanical bonding information. The Mayer bond order is defined in terms of spin orbitals as:

$$G_{AB}^{Mbm} = 2 \sum_{\mu \in A} \sum_{\nu \in B} (\mathbf{PS})_{\mu\nu} (\mathbf{PS})_{\nu\mu}, \tag{71}$$

where P is the density matrix and S is overlap matrix in an atomic orbital basis μ and ν.[245] The Mayer bond order matrix (G^{Mbm}) representation is a slight modification of the form of the values into an affinity matrix, where the magnitude of the Mayer bond matrix element represents the similarity between two atoms. As previously mentioned, some clustering methods benefit from a more block diagonal structure of the descriptor, we also look to coarse grain the descriptor through rounding the values of the matrix. This results in the rounded Mayer bond order matrix (G^{rMbm}) representation which rounds G_{AB}^{Mbm} according to:

$$G_{AB}^{rMbm} = \begin{cases} \lceil B_{AB} \rceil, \text{ if } \{B_{AB}\} >= 0.5 \\ \lfloor B_{AB} \rfloor, \text{ otherwise.} \end{cases} \tag{72}$$

The intention of this representation is to dampen out insignificant pairs, allowing only the most strongly interacting pairs to be considered in the descriptor and making the cuts between clusters more obvious. In general, it is expected that incorporating the bond order into the descriptor will enable the fragmentation approach to preserve bonds between the most strongly interacting parts of the molecules.

6.3.3 Validation

In this section, the means of defining successful partitioning of a molecular system is outlined. In this work, two different approaches are used: one which ensures the expected fragmentation is produced and another that quantifies the recovered energy of the full system. The set of water clusters and methylthiophene tetramers were chosen that should produce very clear clustering for a requested number of clusters. The performance of the clustering methods on these test cases

can be assessed through external validation, which compares the resulting cluster labels to a expected/correct cluster labels.[246,247] The external validation statistic used in this work is the Adjusted Rand Index (ARI). The Rand index, 73, measures the frequency of occurrence of agreement over the total pairs.

$$\mathscr{R}\left(\mathscr{C},\mathscr{C}'\right) = \frac{N_{11}+N_{00}}{n(n-1)/2} \tag{73}$$

Here $\mathscr{C}$ is the resulting clustering and $\mathscr{C}'$ is the expected clustering consisting of n total points. N_{00} and N_{11} are the number of data point pairs in the same clusters and different clusters respectively for C and C'. However, since there is a small probability these data points could end up in the same cluster by chance, the ARI is used. The ARI corrects for chance by using a baseline of expected similarity from a random model. A value close to unity represents total agreement between the expected and the actual fragmentation. The ARI approaching zero reflects an increase in the difference between the expected and returned fragmentation. In the oligomer systems, the correct fragmentation is ambiguous and thus correct clustering is not known ahead of time. for these systems the percent error of the fragmented system from supermolecular energy is used as the metric of success.

6.3.4 Systems

To measure the success of the AFMC approach, families of molecules are considered to explore the performance in cases where the partitioning introducing the lowest error is apparent for noncovalently and covalently bound molecules and a case of more ambiguous fragmentation in oligomers.

Water Clusters: The performance of the AFMC with the various descriptors on noncovalently bound molecules, a set of water clusters were explored. Water clusters, $(H_2O)_n$ for n=2 to n=21 optimized with the TIP4P water model were used from from the Wales cluster database.[248,249]

Methylthiophenes: To benchmark performance of the fragmentation methods on covalently bound systems with a clear desired fragmentation, a set of methylthiophene tetramers were explored. When partitioning these systems, 4 fragments were requested. Coordinates for the tetramers were generated with Open Babel by providing a SMILES representation with defined linkage atoms, indicated by a box in Figure 9.[250] The geometry of each molecule was determined at two lev-

els of theory to study the sensitivity of the fragmentation to small perturbations in the structure. The levels of theory used optimize the structures were Hartree-Fock/6-311G** and ωB97X-D/6-311G**.[251,252] Introduction of broken bonds within the molecule were treated by hydrogen capping. The hydrogen cap contributions were then treated by subtracting the energy of all hydrogen atoms from that of the fragment calculation.

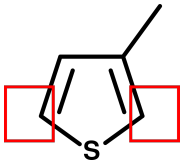


Figure 9: The methylthiophene monomer 2-D structure with the linkage atoms highlighted in red.

Silyl ketene oligomers: Silyl ketene oligomers (SKs) provide a set of systems where ideal fragmentation is ill-defined due to the backbone and side group interactions which may be essential to describe their chemistry and as a result, not clear fragmentation pattern.SKs have the general form of (RC=C=O), where R is a SiR_3 group (Figure 10(left)). This class of molecule is a candidate for chain polymerization and can avoid the undesired 2+2 cycloadditions observed in aryl and alkyl ketenes. Previous work in our group has aimed to predict stable structures and polymerization mechanisms.[253] As large polymer units are considered, the computational cost grows and fragmentation becomes an attractive and possibly necessary option. In this work, oligomers of the SK monomers act as a test system to explore the performance of the clustering approaches in terms of the clustering ability to reduce computational time while minimize the difference in error when compared to the supermolecular calculation. The systems explored consist of a dimer and trimer of the tert-Butyldiphenylsilyl monomer, with a methylonate nucleophile to begin the polymerization displayed in Figure 10 (center, right). The SK structures were generated with Avogadro2 and optimized at the Becke-3 Parameter-Lee-Yang-Parr (B3LYP) including the Becke-Johnson dispersion correction (-D3(BJ)) level of theory with the pc-1 basis set.[192–195] The work by Mardirossian et al. suggests this optimization level represents a balance of computational cost and accuracy.[49] Introduction of broken bonds within the molecule were treated by hydrogen capping. The hydrogen cap contributions were then treated by subtracting the energy of all hydrogen atoms from that of the fragment calculation.

This test case also aimed to determine the sensitivity of the clustering methods to the level of theory used for the Mayer bond order calculations. Using the optimized structure above, the Mayer bond matrix was also calculated at varying levels of theory: Hartree-Fock, B3LYP[192–195], and ωB97M-V[254] each level of theory is paired with four different basis sets: STO-3G[255,256], 3-21G[257,258], cc-pVDZ, cc-pVTZ.[259,260]

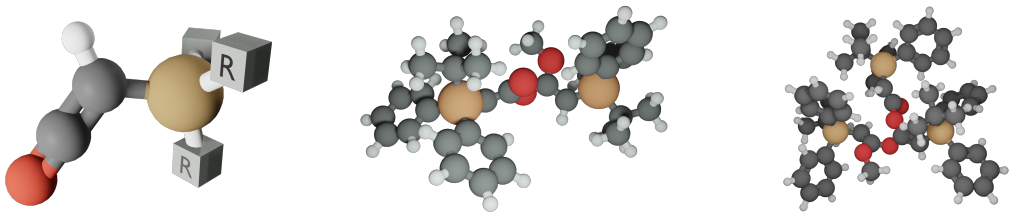


Figure 10: Structure of a generic silyl ketene unit (left), the dimer (center) and trimer (right) used for the clustering benchmark.

6.4 Results and Discussion

6.4.1 Water Clusters

The agglomerative, *k*-means, and spectral clustering return the expected clustering as a comparison of the expected and true clustering results with an average ARI of 1 for all descriptors. This indicates the non-covalently bound water structures are correctly clustered when the number of clusters is predefined. This is expected as non-covalently bound systems have apparent non-interacting or weakly interacting points are clearly represented by the representation and resolved by the clustering algorithm. These results suggests these clustering methods could reliably find reasonable fragments for non-covalently bonded systems.

In order to recover the total energy of a water cluster, often more than one water monomer should be included in a fragment, placing importance on clustering methods which allow for the selection of cluster numbers are valuable. Ideally, the clustering algorithms should resolve to keep individual water molecules intact within each cluster as opposed to separating a molecule between

two fragments.[40] Figure 11 demonstrates the fragmentation for spectral clustering with the G^{xyz} descriptor in cases where the number of clusters is lower than the number of water monomers. In this case the ideal returned fragments are defined as including one or more water monomers without any segmentation of O-H bonds. Preservation of covalent bonds is observed for spectral clustering and k-means clustering, but not for agglomerative clustering.

In addition to ensuring preservation of covalent bonds, achieving a relatively balanced number of molecules (N_{mol}) in the clusters is also a priority. The balanced of the cluster sizes was used as a metric to probe the clustering method/descriptor pairs. The $(H_2O)_{21}$ molecule was partitioned $N_{frag} = 2 \ldots 21$ using the G^{Mbm} and G^{xyz} descriptors with both spectral and k-means clustering. The results shown in Figure 12 demonstrate that spectral clustering with both descriptors return fragments of a similar size over the range tested. The k-means method with G^{xyz} descriptorreturns fragments of a similar size. Though when paired with the G^{Mbm}, requesting a smaller number of fragments returns unevenly sized fragments, as shown in the black circle, where the green fragment contains most of the water molecules and all other fragments contain one-two water molecules.

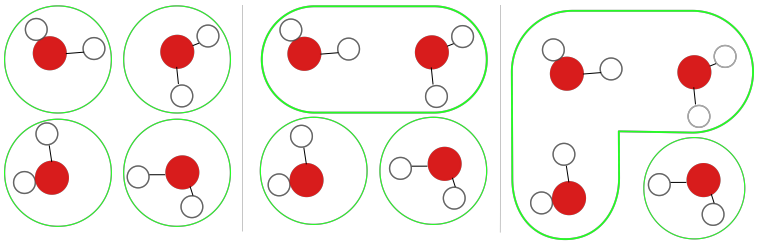


Figure 11: A demonstrative result for the preservation of covalent bonds in cases when the N_{frag} is less than the N_{mol}. The N_{frag} requested decreases from 4 (requested on the left) to 2 (requesting on the right), clusters are designated by the green outlines. This example is performed with spectral clustering on the Cartesian representation for $(H_2O)_4$. Covalent bonds are preserved for all descriptor/clustering algorithm combinations.

The timing of the clustering itself is another important metric to consider, and the AFMC has negligible cost for generating fragments. Once the features of the descriptor are calculated, the routine to set up the molecule, generate fragments, and create output files took approximately 2 seconds for all clustering methods on an 4 core laptop with a i5-5200 CPU. The AFMC approach thus offers a computationally efficient way to generate fragments which can then be used in combi-

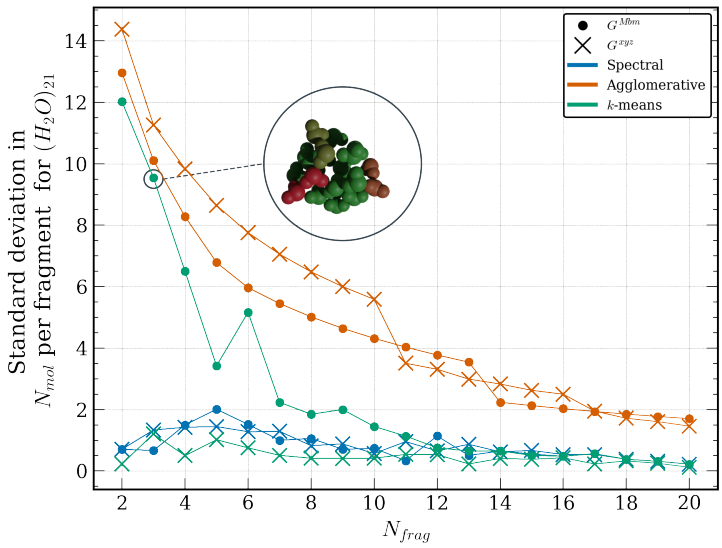


Figure 12: The standard deviation in cluster size for cases when the number of requested clusters is less than the number of monomers. For for $(H_2O)_{21}$, the N_{frag} requested increases from 2 to 20. Shown here for the G^{Mbm} and G^{xyz} descriptor.

nation with any interfragment treatment approaches to estimate the energy and properties of large molecules.

6.4.2 Methylthiophenes

The clustering method performance for the covalently bound methylthiophene tetramers into four fragments are shown in Figure 13; The effects of geometry optimization on the fragmentation results are also presented. Optimization are performed at the Hartree-Fock/6-311G** level of theory subfigure a and ωB97X-D/6-311G** level of theory was used in subfigure b. Spectral clustering performs well for all molecular representations with no dependence on the descriptor, level of theory, or basis set used. Other clustering methods have a strong dependence on the descriptor used and variation as the level of theory used in the geometry optimization changes. These structures are challenging due to the descriptors maintaining less of a block diagonal structure, i.e more non-local interactions. A less block diagonal structure in the representation means cluster

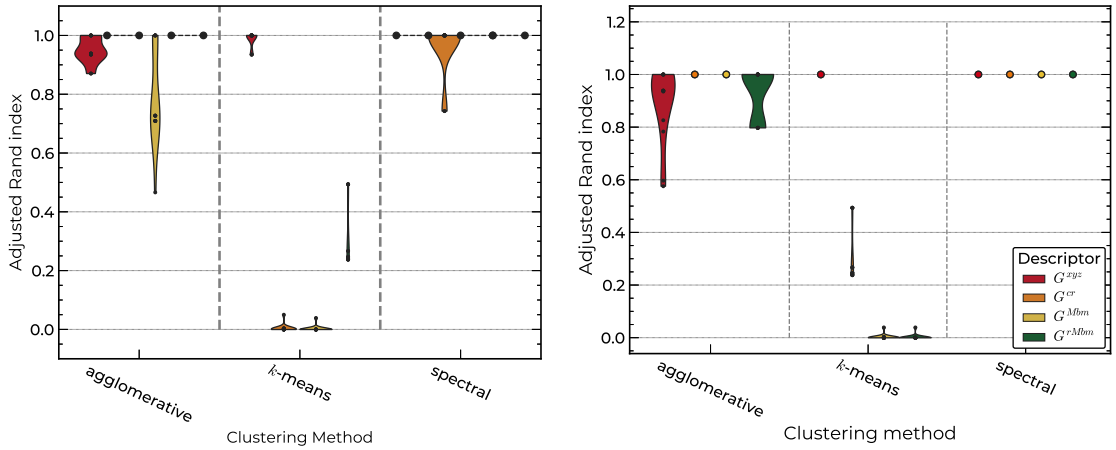


Figure 13: Performance of clustering for the methylthiophene test set, measured as the average ARI across the set of methylthiophenes. The colored dots at 1 indicate successful clustering of the full test set, as the average ARI is 1 for the descriptor/clustering algorithm combination.

boundaries are much less clear leading to problems for certain unsupervised learning algorithms. However, as spectral clustering first embeds the representation into a lower dimensional space before clustering, it is able to resolve the primary interactions. Notably, agglomerative, *k*-means and spectral clustering perform well with the Cartesian descriptor.

6.4.3 Silyl Ketene

Silyl ketenes, realistic systems with more ambiguity in the choice of partitioning, are challenging applications for these clustering approaches. The dimer results are found in Figure **??** and trimer results can be seen in Figure **??**. The findings across the dimers and trimers are comparable and suggest spectral clustering is the most robust clustering algorithm, achieving the lowest errors with the largest speedups across all descriptor/method pairs. Interestingly, a fragmentation pattern with the lowest error is found when the covalent radii bond matrix is used for the dimer or the Cartesian descriptor is used for the trimer. The favorable fragmentation using these purely structure-based descriptors suggests it may be possible to select fragments that result in high accuracies in molecular properties without relying on the incorporation quantum mechanical

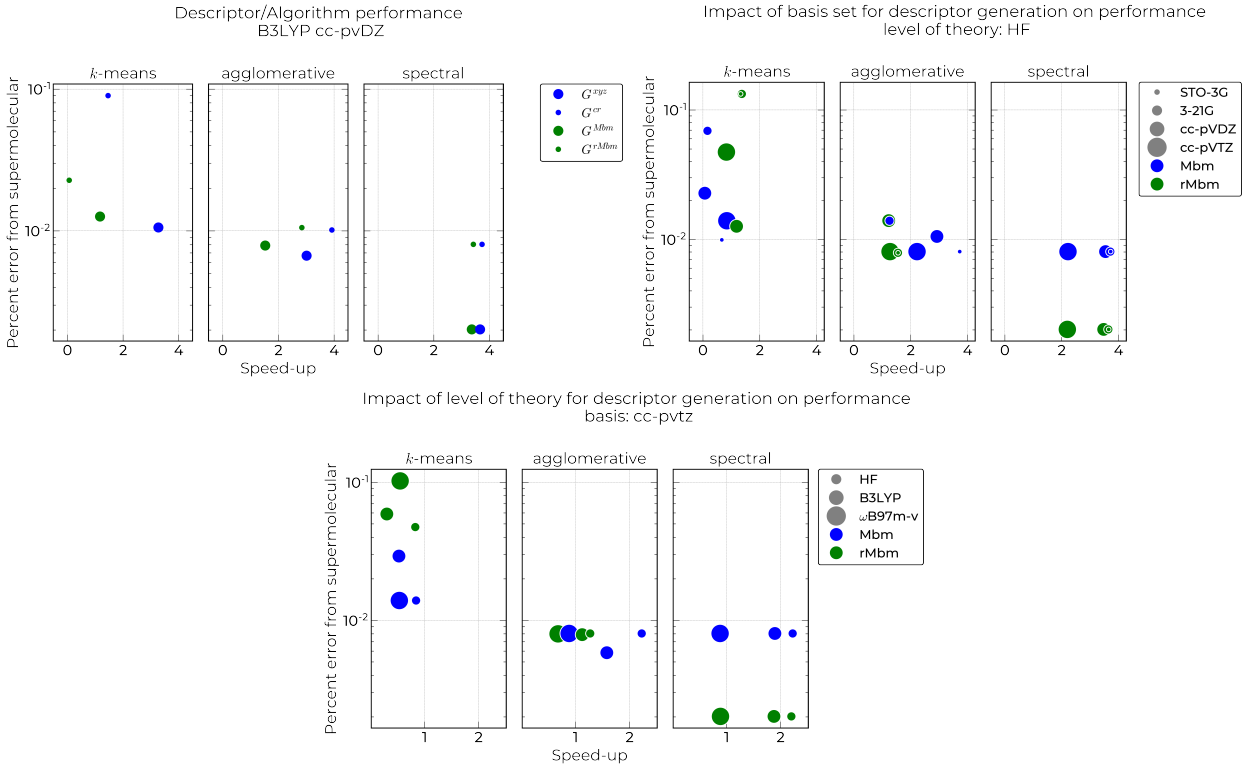


Figure 14: Assessment of fragmentation schemes on silyl ketene trimers: Presented is the percent error of the energy and the speedup over the supermolecular calculation for the SK trimer to assess the performance of the fragment approaches and descriptor quality.

Figure 15: Representative visualization of resulting fragmentation for silyl ketenes. Results from agglomerative (left) and spectral (right) clustering on the G^{Mbm} descriptor. Colors represent fragment identity.

information in the descriptor for some systems.

Agglomerative clustering produces acceptable fragments, though in general, a larger deviation from the supermolecular result is observed. On the other hand, only the Cartesian descriptor with the *k*-means clustering algorithm yields reasonable result, which is likely due to the conserved spatial information allowing for an accurate choice of centroids opposed to representations which employ bonding environment as the features. Overall, the clustering approaches investigated are relatively insensitive to the level of theory and basis set used to generate the Mayer bond matrix. which is encouraging as it may allow for future computational savings in future applications on larger more complex systems where a purely structure base descriptor may not incorporate the necessary interaction important to the structure. If cases arise in larger systems where the quantum mechanics based descriptors become necessary, the low level approximation to the bond order will suffice as long as the bond matrix is meaningful.

6.5 Conclusion

In this work we explore the Automatic Fragmentation of Molecules through Clustering (AFMC) as an automated partitioning scheme for molecular systems. AFMC approach utilizes UML methods to determine molecular domains in computationally efficient and transferable ways. Three classes of systems were studied to assess the performance of the AFMC approach: water clusters to investigate non-covalently bonded molecules, methylthiophene tetramers to probe the behavior of AFMC with ring structures, and two silyl ketene oligomers to explore a realistic chemical case with ambiguous fragmentation choices. Several clustering approaches and molecular representations were tested. We find that this approach with various clustering methods can accurately identify meaningful molecular domains for non-covalently bound molecular systems (water clusters) and in the case of covalently bonded systems with aromatic units. Overall, a spectral clustering approach was able to produce balanced and sensible molecular fragments as can be seen by the low error and high speed up for the silyl ketene structures. The clustering is performed on molecular representations derived from either the molecular structure alone or a low-level quantum mechanics prediction of the Mayer bond matrix. Both classes of representations performed well for the

systems studied, though it is not yet clear whether more intricate, correlation-dependent, bonding schemes will benefit from the quantum mechanical informed descriptors and we aim to explore this in future work. The combination of Spectral clustering with the Cartesian descriptor provides the most reliable clustering with minimal preparation cost, thus we recommend this combination for molecules alike to those studied here. The combination of these descriptors and UML techniques provide a low cost way to determine acceptable fragments for computational chemistry for further, more accurate, quantum mechanical calculations.

7.0 Conclusions

In this work, two approaches were explored to approximate the description of electron correlation in electronic structure calculations while minimizing or reducing the computational demand. The first direction explored was through the application of stochastic sampling to solve the Schrödinger equation through quantum Monte Carlo (QMC) approaches. Also explored was the use of UML methods to determine molecular partitions for fragmentation schemes.

For the quantum Monte Carlo (QMC) applications, two systems were explored that are not well described by DFT: a hydrogen chemisorbed to the surface of graphene and a non-valence correlation bound anion model system, $(H_2O)_4$. Calculations of the binding energy of a hydrogen atom on a graphene sheet were carried out using various DFT methods and with DMC. The DMC calculations provide a benchmark value of the binding energy. Our best estimate of the binding energy from DMC calculations is -691 $\pm$ 19 meV. The PBE result obtained with a plane-wave basis set gives a binding energy about 20% larger in magnitude than the DMC result. The global hybrid functional, PBE0, gives a binding energy close to that of PBE. In comparison, HSE, a range-separated hybrid functional, gives a smaller binding energy of -743 meV, after a correction applied for the basis set incompleteness error, and is much closer to the value from DMC calculations. Interestingly, there are significant differences in the DMC and PBE charge densities of both graphene and H/graphene. Moving forward, there are cases that an inaccurate density can be the primary source of DFT errors. In these cases density-corrected DFT can be used.[261–266] Often HF densities are used, but exploring cases where QMC data may help is of interest. Additionally, exploring further observables for the system may provide more insight into the difference in binding energy observed between DMC and DFT.

For the $(H_2O)_4$ system, various EOM-CC methods and two different quantum Monte Carlo methods to calculate the EBE of a model $(H_2O)_4$ cluster at two geometries, one at which the anion is bound in the HF approximation and the other at which it is not. Diffusion Monte Carlo calculations using single determinant trial functions based on Hartree-Fock orbitals are shown to bind the excess electron even when the initial wave function for the anion has collapsed onto the neutral plus discretized continuum orbital. However, such calculations significantly underestimate

the EBE, whereas SD DMC calculations using trial wave functions for the anion with a more realistic charge distribution for the excess electron give larger EBE values that are in close agreement with our best estimate EOM-CCSDT values for both geometries considered. For the $(H_2O)_4$ model system, the restricted SDCI represents an economical way to create trial wave functions for QMC calculations on non-valence anions that are not bound in the Hartree-Fock approximation. However, it remains to be seen if this strategy will be as effective for systems in which the neutral species is more strongly correlated than the model $(H_2O)_4$ cluster. Finally, we note that at R = 4 Å, for which the anion is NVCB in nature, the most frequently used method to characterize such anions, EOM-CCSD, underestimates the EBE by about 10% compared to the result of EOM-CCSDT calculations. Both DMC and AFQMC are viable alternatives to high order EOM methods, and while more computationally demanding for the $(H_2O)_4$ cluster, they demonstrate lower scaling with system size than EOM methods, making them attractive for the characterization of non-valence anions of much larger systems.

The other approach explored was the molecular partitioning for fragmentation approaches as determined by UML methods. In this work we explore the Automatic Fragmentation of Molecules through Clustering (AFMC) as an automated partitioning scheme for molecular systems. AFMC approach utilizes UML methods to determine molecular domains that is both computationally efficient and system agnostic. Several clustering approaches and molecular representations were tested. The molecular representations are derived from either the molecular structure alone or a low-level quantum mechanics prediction of the Mayer bond matrix. Three classes of systems were studied to assess the performance of the AFMC approach: water clusters to investigate non-covalently bonded molecules, methylthiophene tetramers to probe the behavior of AFMC with ring structures, and two silyl ketene oligomers to explore a realistic chemical case with ambiguous fragmentation choices. We find that this approach with various clustering methods can accurately identify meaningful molecular domains for non-covalently bound molecular systems (water clusters) and in the case of covalently bonded systems with aromatic units. Overall, a spectral clustering approach was able to produce balanced and sensible molecular fragments as can be seen by the low error and high speed up for the silyl ketene structures. Both classes of representations performed well for the systems studied, though it is not yet clear whether more intricate, correlation-dependent, bonding schemes will benefit from the quantum mechanical informed de-

scriptors and we aim to explore this in future work. The combination of spectral clustering with the Cartesian descriptor provides the most reliable clustering with minimal preparation cost, thus we recommend this combination for molecules alike to those studied here. Beyond an exploration of the quantum mechanics-derived descriptors on larger molecules, there are also interesting direction which can look to creating descriptors that are based around a target area such as an adsorbate on a surface or a protein binding pocket.

Appendix A Additional works

As some of the research directions are not tied directly to the dissertation theme or are still being developed, they are mentioned here. These works can briefly be described as contribution to improving quantum computing emulators, an extension of the AFMC approach to regional embedding fragment definition, and a density functional theory analysis of silyl ketene systems to inform experimental synthesis.

Appendix B Q-GPU: A Recipe of Optimizations for Quantum Circuit Simulation Using GPUs

B.1 Summary

In recent years, quantum computing has undergone significant developments and has established its supremacy in many application domains. Unfortunately, modern quantum computing is still positioned in the Noisy Intermediate-Scale Quantum (NISQ) era that is limited by the number of qubits, short qubit lifetime, and imperfect operations. While quantum hardware is accessible to the public through the cloud environment, a robust and efficient quantum circuit simulator is necessary to investigate the constraints and foster quantum computer development, such as quantum algorithm development and quantum device architecture exploration. In this paper, we observe that most of the publicly available quantum circuit simulators (e.g., QISKit from IBM) are not optimized and suffer from slow simulation and poor scalability. To this end, we systematically studied the deficiencies in modern quantum simulators and propose Q-GPU, a framework that leverages GPUs with comprehensive optimizations to allow efficient and scalable quantum circuit simulation (QCS). Specifically, Q-GPU features i) proactive state amplitude transfer, ii) zero state amplitudes pruning, iii) delayed qubit involvement, and iv) non-zero state compression. Experimental results across eight representative quantum circuits indicate that Q-GPU significantly improves the simulation performance over the state-of-the-art GPU-based QCS by 2.53$\times$ on average. It also outperforms the most recent OpenMP CPU implementation, the Google Qsim-Cirq simulator, and the Microsoft QDK simulator.

B.2 Introduction

Quantum computing is a promising computing paradigm that has the potential to solve problems that cannot be handled by classical computers in a feasible amount of time[267]. In the past decade, there has been steady progress towards building a large quantum computer. The number

of qubits in a real quantum machine has increased from 14 in 2011[268] to 76 in 2020[269]. IBM promises 1000 qubits quantum machine by the year 2023[270]. Despite this rapid progress, current quantum computing is still positioned in the Noisy Intermediate-Scale Quantum (NISQ) era where the public has very limited access to quantum machines. These machines are also constrained by the limited number of qubits, short lifetimes of qubits, and imperfect operations[271]. Thus, quantum circuit simulation (QCS) toolsets provide an essential platform to satisfy many needs, e.g., developing many different algorithms with a large number of qubits, validating and evaluating newly proposed quantum circuits, and design space exploration of future quantum machine architectures. Many companies, such as IBM, Google, Intel, and Microsoft have developed their quantum circuit simulators to provide precise end-end simulation.

In general, QCS is challenging as it is both compute-intensive and memory-intensive[272,273]. The reasons are: i) fully and accurately tracking the evolution of quantum system through classical simulation[274] requires storing all the quantum state amplitudes, which carries a memory cost that grows exponentially as the number of qubits in the simulated quantum circuit increases, and ii) applying a gate within a quantum circuit requires a traversal of all the stored state amplitudes, leading to exponentially scaling computational complexity. Modern GPUs have been used to fuel QCS in high-performance computing (HPC) platforms. Specifically, when applying a gate to a n-qubit quantum circuit, the 2^n state amplitudes are evenly divided into groups, and each group of amplitudes is updated independently in parallel by GPU threads. However, the promising parallelism of GPUs is diminished by the limited GPU on-board memory capacity. For example, simulating a quantum circuit with 34 qubits requires 256 GB of memory to store state amplitudes, which is beyond the memory capacity of any modern GPUs.

There exist several works optimizing QCS, including multi-GPU supported simulation[275,276], OpenMP and MPI based CPU simulation[277–279], and CPU-GPU collaborative simulation[280]. Most of these works focus on distributed simulation while failing to benefit from GPU execution due to the memory constraint. In particular, our characterization shows that the state-of-the-art GPU-based simulation[280] has low GPU utilization when the number of qubits in the quantum circuit is large. As a result, most state amplitudes are stored and updated on the CPU, failing to take advantage of the GPU parallelization. Moreover, the static and unbalanced allocation of state amplitudes introduces frequent amplitude exchange between CPU and GPU, which introduces

additional data movement and synchronization overheads.

In this paper, we aim to provide a high-performance and scalable QCS using GPUs. We propose *Q-GPU*, a framework that significantly enhances the simulation performance for practical quantum circuits. The proposed framework leverages modern GPUs as the main execution engine and is featured with several end-to-end optimizations to fully take advantage of the rich computational parallelism on GPUs, while maintaining a minimum amount of data movement between the CPU and GPU. Specifically, our approach includes four optimizations. First, instead of statically assigning state amplitudes on GPU and CPU as done in prior works[280], Q-GPU dynamically allocates groups of state amplitudes on the GPU and proactively exchanges the state amplitudes between CPU and GPU. Doing so maximizes the overlap of data transfer between CPU and GPU, thereby reducing the GPU idleness. Second, Q-GPU prunes zero state amplitudes to avoid unnecessary data movement between CPU and GPU. Third, we also propose compiler-assisted quantum gate reordering (complying with the gate dependencies) to enlarge the opportunity of pruning zero state amplitudes. Finally, we propose efficient GPU-supported lossless data compression to further reduce data transfer caused by non-zero amplitudes. This paper makes the following contributions:

- We use the popular IBM QISKit-Aer with its state-of-the-art CPU-GPU implementation[281], and conduct an in-depth characterization of the simulation performance. We observe that the performance degrades significantly as the number of qubits increases due to the unbalanced amplitudes assignment, where most of the computation is done by the CPU.
- We implement a dynamic state amplitude assignment to allow the GPU to update all state amplitudes. However, such an implementation did not provide any performance improvements and even worsened compared to the CPU execution due to the massive and expensive data movement between CPU and GPU.
- We propose Q-GPU, a framework comprising end-to-end optimizations to mitigate the data movement overheads and unleash the CPU capability in QCS. Specifically, the proposed Q-GPU is featured with the following major optimizations: i) dynamic state amplitudes allocation and proactive data exchange between CPU and GPU, ii) dynamic zero state amplitude "pruning", iii) dependency-aware quantum gate reordering to enlarge the potential of zero amplitude pruning, and iv) GPU-supported efficient lossless compression for non-zero amplitudes.

- We evaluate the proposed Q-GPU framework using eight practical quantum circuits. Experimental results indicate that in all circuits tested, Q-GPU significantly improves the QCS performance and outperforms the baseline by 2.53× on average. We also compare Q-GPU with Google Qsim-Cirq[282] and Microsoft QDK[283], and results show that Q-GPU approach outperforms Qsim-Cirq and QDK by 1.02× and 9.82×, respectively.

B.3 Background

B.3.1 Quantum Basics

Similar to the *bit* concept in classical computation, quantum computation is built upon the *quantum bit* or *qubit* for short[284]. A qubit is a two-level quantum system defined by two computational orthonormal basis states $|0\rangle$ and $|1\rangle$. A quantum state $|\psi\rangle$ can be expressed by any linear combination of the basis states.

$$|\psi\rangle = a_0|0\rangle + a_1|1\rangle, \tag{74}$$

where a_0 and a_1 are complex numbers whose squares represent the probability amplitudes of basis states $|0\rangle$ and $|1\rangle$, respectively. Note that we have $|a_0|^2 + |a_1|^2 = 1$, meaning that after measurement, the read out of state $|\psi\rangle$ is either $|0\rangle$ or $|1\rangle$, with probabilities $|a_0|^2$ and $|a_1|^2$, respectively. The states of a quantum system are generally represented by *state vectors* as

$$|0\rangle = \begin{bmatrix}1\\0\end{bmatrix}, |1\rangle = \begin{bmatrix}0\\1\end{bmatrix}. \tag{75}$$

To be more general, for an n-qubit system, there are 2^n state amplitudes. Then, the quantum state $|\psi\rangle$ can be expressed as a linear combination

$$|\psi\rangle = a_{0\ldots00}|0\ldots00\rangle + a_{0\ldots01}|0\ldots01\rangle + \cdots + a_{1\ldots11}|1\ldots11\rangle. \tag{76}$$

Similarly, the state of a n-qubit system can also be represented by a state vector with 2^n dimensions as

$$|\psi\rangle = a_{0\ldots00}\begin{bmatrix}1\\0\\\vdots\\0\end{bmatrix} + a_{0\ldots01}\begin{bmatrix}0\\1\\\vdots\\0\end{bmatrix} + \cdots + a_{1\ldots11}\begin{bmatrix}0\\0\\\vdots\\1\end{bmatrix} = \begin{bmatrix}a_{0\ldots00}\\a_{0\ldots01}\\\vdots\\a_{1\ldots11}\end{bmatrix}. \tag{77}$$

Quantum computation describes changes occurring in this state vector. A quantum computer is built upon a *quantum circuit* containing *quantum gates* (or quantum operations), and a quantum algorithm is described by a specific quantum circuit. In simple terms, quantum gates are represented by unitary operations that are applied on qubits to map one quantum state to another. A quantum gate that acts on k qubits is represented by a $2^k \times 2^k$ unitary matrix.

To illustrate how a quantum gate is applied to a state vector, let us consider a 2-qubit system with a Hadamard gate/operation operating on qubit 0. A Hadamard gate can be represented as

$$H \equiv \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{78}$$

Then the state vector of this 2-qubit system is updated through

$$\begin{bmatrix} a'_{00} \\ a'_{01} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} a_{00} \\ a_{01} \end{bmatrix}, \tag{79}$$

$$\begin{bmatrix} a'_{10} \\ a'_{11} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} a_{10} \\ a_{11} \end{bmatrix}. \tag{80}$$

For an n-qubit system, when a H gate is applied to qubit j the amplitudes are transformed as[285]:

$$\begin{bmatrix} a'_{\times\cdots\times 0_j \times\cdots\times} \\ a'_{\times\cdots\times 1_j \times\cdots\times} \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} a_{\times\cdots\times 0_j \times\cdots\times} \\ a_{\times\cdots\times 1_j \times\cdots\times} \end{bmatrix} \tag{81}$$

Therefore, the indices of every pair of amplitudes have either 0 or 1 in the jth bit, while all other bits remain the same[1]. Note that each pair of amplitudes can be updated in parallel.

B.3.2 Quantum Circuit Simulation (QCS)

The purpose of QCS is to mimic the dynamics of a quantum system[274], and to reproduce the outcomes of a quantum circuit with high accuracy. There are several approaches to simulating a quantum circuit, each offering different advantages and drawbacks. We summarize the three most widely used approaches below.

- **Schrödinger style simulation:** Schrödinger simulation describes the evolution of a quantum system by tracking its quantum state. It tracks the transformations of the state vector according

[1] "$\times$" can be 0 or 1; the "$\times$" in the same position of $a_{\times\times\times\times\times\times\times 0}$ and $a_{\times\times\times\times\times\times\times 1}$ are the same.

to Equation 81. Note that one can also track the density matrix $\rho = |\psi\rangle\langle\psi|$, which is useful when measurement is required during simulation[274,275]. In this work, we only consider quantum measurements at the end of circuits.

- **Stabilizer formalism:** Simulation based on the stabilizer formalism is efficient for a restricted class of quantum circuits[274,284,286]. Specifically, stabilizer circuits (a.k.a Clifford circuits) can be simulated in $O(poly(n))$ space and time costs. Rather than tracking the state vector, the quantum state is uniquely represented and tracked by its stabilizers, which is essentially a group of operators derived from the Clifford group. A detailed description can be found in[286].
- **Tensor network:** Tensor network simulators are useful when a single or few amplitudes of the full state vector are being updated as tensor networks[287–290]. For example, one type of tensor network that are extremely common are matrix product states (MPS). When applied to a single amplitude in Equation 76, the resulting state resembles a long string of matrix multiplications

$$|\psi\rangle = \sum_{j_0 \cdots j_{n-1} j_n} a_{j_0 \cdots j_{n-1} j_n} |j_0 \cdots j_{n-1} j_n\rangle \tag{82}$$

$$= \sum_{j_0 \cdots j_{n-1} j_n} Tr[A^{j_0} \ldots A^{j_{n-1}} A^{j_n}] |j_0 \cdots j_{n-1} j_n\rangle \tag{83}$$

The matrices A (rank-2 tensors) in Equation 83 can be thought of as a decomposition of the full coefficient tensor a. Despite the restriction of returning a limited number of amplitudes, tensor networks states are efficient as they compress the dimension of the problem from $O(2^n)$ to $O(nd^2)$ where d is the dimension of the individual tensors in Equation 83.

Among all these simulation methods, Schrödinger style simulation is widely used as the mainstream simulation method, and has been widely adopted in prior research works[272,277–280,285,291–294]. Also, industrial quantum circuit simulators such as IBM QISKit[281], Google Qsim-Cirq[277,292] and Microsoft QDK[283] use full state vector simulations. In this work, we build Q-GPU based on IBM QISKit-Aer, a high-performance C++ simulation backend of QISKit, since it contains the state-of-the-art GPU support.

Table 6: List of quantum circuit benchmarks.

Abbrv.	Application
hchain	Linear hydrogen atom chain[295]
rqc	Random quantum circuit[267]
qaoa	Quantum approximate optimization algorithm[296]
gs	Graph state[297,298]
hlf	Hidden linear function[299]
qft	Quantum Fourier transform[300]
iqp	Instantaneous quantum polynomial-time[301,302]
qf	Quadratic form[303]

B.4 Characterization of QCS

B.4.1 Quantum Circuit Benchmarks

In this paper, we characterize the performance of QCS using a rich set of quantum circuits. Table 6 lists the circuit benchmarks.

- `hchain`: This circuit which describes a system of hydrogen atoms arranged linearly is a representative quantum chemistry application[304–308]. This circuit incorporates increased circuit depth and an early entanglement in terms of total operations.
- `rqc`: The random quantum circuit from Google[267,309] is used to represent the quantum supremacy compared to classical computers.
- `qaoa`: Quantum approximate optimization is a promising quantum algorithm in the NISQ era that produces approximate solutions for combinatorial optimization problems[296].
- `gs`: This circuit is used to prepare graph states[310] that are multi-particle entangled states. Examples include many-body spin states of distributed quantum systems that are important in quantum error correction[311].
- `hlf`: This benchmark circuit solves the 2D hidden linear function problem[299].
- `qft`: The quantum Fourier transform circuit[300] is the quantum analog of the inverse discrete Fourier transform. It is an important function in Shor's algorithm[312].
- `iqp`: The instantaneous quantum polynomial circuit provides evidence that sampling the output probability distribution of a quantum circuit is difficult when using classical approaches[301,302].

- `qf` : This circuit implements a quadratic form on binary variables encoded in qubit registers. It is used to solve the quadratic unconstrained binary optimization problems[303].

B.4.2 Baseline QCS

Step 1: State vector partitioning: QISKit-Aer first partitions the state vectors into "chunks". Chunk is the granularity used in the simulator to update the state vector. For illustrative purposes, let us assume we have a 7-qubit circuit, i.e., that there are in total 2^7 different state amplitudes from $a_{0000000}$ to $a_{1111111}$. All the states are stored in a vector (i.e., the state vector), and this state vector is partitioned into chunks. For example, assuming we divide the state vector into 8 chunks, each chunk contains 16 state amplitudes as shown in Figure 16. The three most significant bits are used to index the chunks, and the remaining bits are as offsets within a chunk.

Step 2: Static chunk allocation: After partitioning, these chunks are allocated into GPU memory based on the GPU memory availability. As illustrated in Figure 16, if a GPU can only store 3 chunks, the remaining 5 chunks will be stored in the host CPU memory. For example, when 64 GB memory is needed to simulate 32 qubits, the first 16 GB is allocated in GPU memory (in P100 GPU with 16 GB memory) and the remaining 48 GB is in the CPU memory.

Step 3: Reactive chunk exchange: During circuit simulation, a chunk exchange between the GPU and the CPU arises when the requested state amplitudes are not locally available on the GPU. In QISKit-Aer, the chunk exchange between the CPU and the GPU is triggered on-demand. That is, when both the chunks on the CPU and the GPU are involved in one state-update calculation, the corresponding CPU chunks are transferred to GPU for updating. After the operation, the updated chunks are transferred back to the CPU. Note that, the amount of data exchange in the following scenarios is dependent on the qubits in the specific gate simulation.

- **Case 1: All the indices of the qubits involved in the current gate are smaller than the chunk size:**. For example, a gate on qubit 0 requires amplitudes $a_{\times\times\times\times\times\times\times 0}$ and $a_{\times\times\times\times\times\times\times 1}$ (see Equation 81). In this case, each chunk can be updated independently without requiring extra data movement.
- **Case 2: Some indices of qubits involved in the current gate are outside the chunk boundary:** In this scenario, let us assume there is a gate that operates on q_6, thereby the required pairs

of amplitudes are $a_{\times 0 \times \times \times \times \times \times}$ and $a_{\times 1 \times \times \times \times \times \times}$. However, as depicted in Figure 16, none of the chunks contains a pair of required amplitudes, i.e., the computation for updating amplitudes involves more than one chunk. Specifically, to update the pairs of amplitudes, we need ($chunk_0$, $chunk_2$), ($chunk_1$, $chunk_3$), ..., and ($chunk_5$, $chunk_7$). However, ($chunk_1$, $chunk_3$) involves one chunk on the GPU and one chunk on the CPU. In this scenario, data exchange is required. In the baseline QISKit-Aer simulation, the requested chunks are always copied from CPU to GPU. That is, in the example above, the CPU copies $chunk_3$ to GPU. After the $chunk_3$ is updated together with $chunk_1$, it is copied back to the CPU memory.

Note that, as the GPU memory capacity is much less compared to the CPU host memory, a large number of chunks are statically allocated on CPU memory when the number of qubits is large. For instance, on the P100 GPU with 16 GB memory, we observe from experiments that when simulating a circuit that has 34 qubits, the state vector is divided into 8192 chunks, 496 chunks are allocated on GPU, while the remaining 7696 chunks are all on CPU. Therefore, one can expect that most of the time, the CPU does the state amplitude update without benefiting from the GPU acceleration.

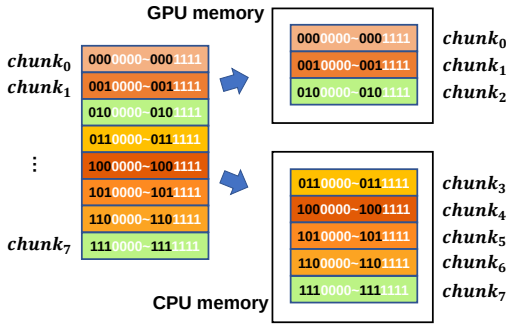


Figure 16: Example of baseline execution where the state vector is statically partitioned and allocated on CPU and GPU.

B.4.3 Characterization and Observations

In this section, we quantify the simulation performance of the baseline QISKit-Aer. We first study the scalability when the number of qubits increases. We observe that, if there are less than 30

qubits in the circuit, the baseline GPU simulates much faster than compared CPU-based simulation (e.g. 9.67× speedup for 29-qubit circuits on average), since the entire state vector fits in the P100 GPU memory and there is no need for data exchange and synchronization. However, the baseline GPU performance significantly drops when the number of qubits is larger than 30. It becomes even worse than running on the CPU alone when the number of qubits reaches 32. In particular, we observe a factor of 1.8× slowdown for `qft_33`[2] as an example.

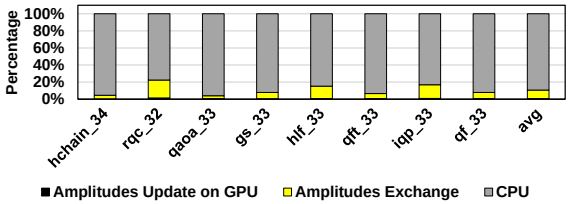


Figure 17: Baseline execution time breakdown.

To investigate the reason for this slowdown, we show the breakdown of the execution time in Figure 17. One can observe that, on average, 89.34% of the execution is spent on the CPU, indicating that the GPUs are not properly used in the baseline execution for large number qubit circuits. Moreover, the overheads involve amplitude exchange and synchronization occupies 9.91% of the average execution time, and the computation time of GPU only occupies 0.71% of total time on average. *In other words, most of the computation is performed by the CPU and the GPU is idle due to the static state chunk allocation in the baseline GPU execution.* In Figure 21, ① depicts the execution timeline of the baseline.

B.4.4 Will a Naive Optimization Work?

To improve the GPU utilization during simulation, an intuitive optimization would dynamically allocate the chunks and transfer the chunks to GPU for updates. In this section, we investigate whether the naive implementation works well or not.

We implemented the dynamic state vector chunk allocation in QISKit-Aer. Figure 18 depicts the execution time of the naive optimization normalized to the baseline execution. Surprisingly, none of the quantum circuits we studied show improvements when using dynamic allocation. To

[2]In this paper, we use *n* in the circuit name (e.g., `circ_n`) to represent a circuit with *n* qubits.

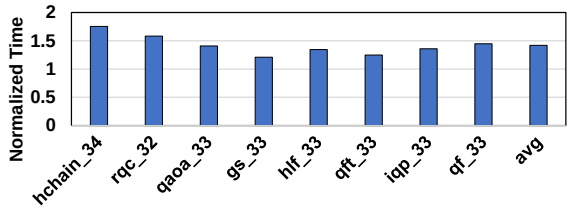


Figure 18: Normalized execution time of naive approach.

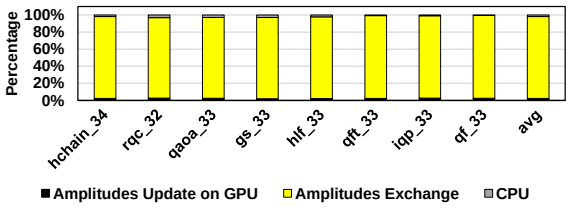


Figure 19: Execution time breakdown of naive optimization.

further investigate the reason, we break down the execution time and show the results in Figure 19. As can be seen from the figure, while CPU execution time significantly reduces and the data movement dominates, indicating that the GPU is waiting for data most of the time during execution. Therefore, naive dynamic allocation alone does not work to deliver good QCS performance. More sophisticated end-to-end optimizations are required to systematically improve the QCS performance and scalability.

B.5 Q-GPU

In this paper, we propose *Q-GPU*, a framework that features several end-to-end optimizations. Figure 20 depicts the high-level overview of Q-GPU. (❶) Q-GPU performs proactive state amplitude transfer to fully utilize the bi-directional data transfer bandwidth between CPU and GPU (Section B.5.1). (❷) Before copying state amplitudes to GPU, Q-GPU performs dynamic redundancy elimination that prunes zero state amplitudes to avoid unnecessary data movements (Section

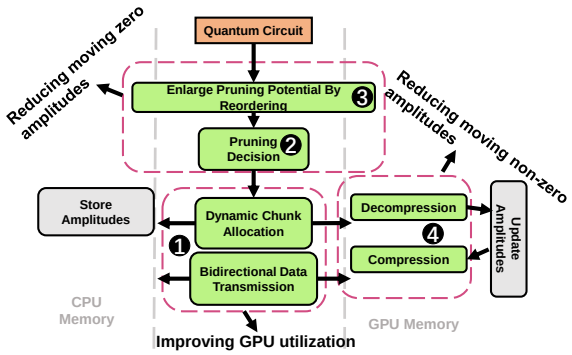


Figure 20: High level overview of Q-GPU.

B.5.2). (❸) Q-GPU features a compiler-assisted, dependency-aware quantum gate reordering to enlarge the potential of pruning (i.e., the number of zero amplitudes). (❹) Q-GPU implements a GPU-supported, lossless amplitude compression to further reduce the data transfer caused by non-zero state amplitudes with minimal runtime overheads (Section B.5.4).

B.5.1 Proactive State Amplitudes Transfer

In the naive execution, one reason behind the poor GPU utilization is the sequential state amplitude transfer between CPU and GPU. Specifically, when the GPU finishes updating all local chunks, those chunks are first copied back to CPU memory before the CPU can transfer the next batch of un-updated chunks to the GPU. This restriction is reasonable in the scenarios when particular chunks are involved in consecutive updates since the chunks being copied from the GPU's memory cannot be overwritten during the copying. In other words, data movements are synchronized to avoid data conflicts. However, if the subsequent chunks from the CPU are not copied to the same memory locations on the GPU where current chunks are stored, such data conflict does not exist. As a result, one can transfer the chunks simultaneously from the CPU to the GPU and from the GPU to the CPU.

In our work, Q-GPU leverages CUDA streams to enable concurrent and bi-directional chunk copy to fully utilize the available bandwidth between the CPU and GPU. To avoid potential data conflict, Q-GPU implements two CUDA streams and partitions the GPU memory into two halves.

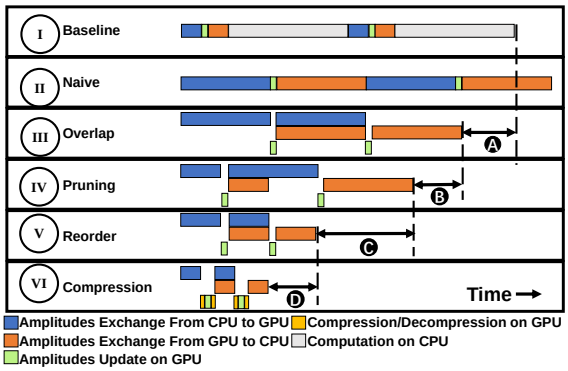


Figure 21: Time-line graph showing the benefits of each optimization in Q-GPU.

One stream is responsible for the first half partition that acts as a buffer holding the chunks the GPU is currently updating. The other stream is responsible for the second half partition that acts as a buffer for "prefetching" the next chunks for the GPU to update. The two memory partitions work as "circular buffers" to feed the GPU with the required chunks. These two streams can potentially overlap and execute concurrently.

Figure 21 illustrates the timeline of the baseline and each of our optimizations. The proposed proactive state amplitude transfer (Ⓘⓘⓘ) achieves Ⓐ cycles savings compared with the baseline (Ⓘ). We also show that the naive approach (Ⓘⓘ) performs worse than the baseline.

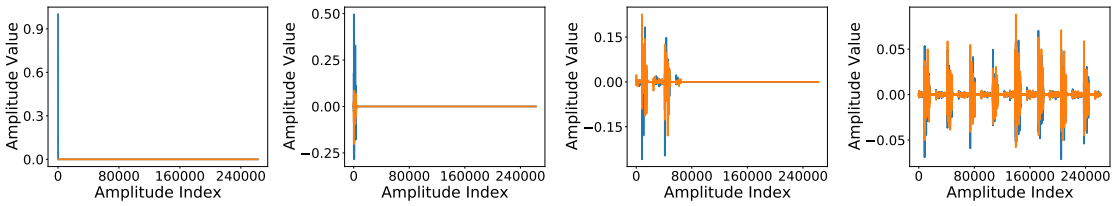
B.5.2 Pruning Zero State Amplitudes

While overlapping improves the bandwidth utilization, the total amount of amplitudes that are transferred remains unchanged. To reduce the data movement, we observe that there exist a considerable amount of zero state amplitudes that do not need to be updated during simulation. Thus, those zero state amplitudes can be pruned before transferring the chunks.

Source of zero amplitudes: Let us assume there are n qubits, the initial states are usually set as $|0\rangle^{\otimes n}$ in the general QCS, indicating that all qubits have zero probability of being measured as $|1\rangle$. Hence, all state amplitudes are zeros, except for $a_{0_1 0_2 \ldots 0_n}$ which is 1. As the state of a particular qubit is unchanged until an operation is being applied on it, its state remains $|0\rangle$ until that operation happens. For instance, if a particular qubit q_k is $|0\rangle$, all the state amplitudes $a_{\times \cdots \times 1_k \times \cdots \times}$

are zeros since q_k has zero probability to be measured as $|1\rangle$. In general, if m of n-qubits are not involved, amplitudes $a_{\times 0_{k_1} \times 0_{k_2} \cdots \times 0_{k_m} \times \times}$ are possible to be non-zero values, whereas the remaining amplitudes are guaranteed to be zero values, i.e. $2^n - 2^{n-m}$ amplitudes are zero values. Therefore, even if only one qubit is not involved, then half of the state amplitudes are zeros.

Pruning potential: To investigate the potential of pruning, Table 7 lists the number of total operations and the number of operations before all qubits are involved. For circuits like `iqp`, we can expect a significant reduction of data movement after pruning since many qubits are not involved until the end of execution. However, for `qft` and `qf`, all qubits are involved at the beginning of execution, diminishing the potential of pruning benefits. We also use `hchain_18` as an example and plot the distribution of state amplitudes after each operation (i.e., quantum gate) being applied in a quantum circuit. Figure 22 shows the state amplitude distribution after 0, 30, 60 and 90 operations. One can observe that a large portion of state amplitudes are zeros at the beginning of the simulation. During simulation, the amplitudes are gradually updated to non-zero values since more qubits are involved.



(a) (b) (c) (d)

Figure 22: State amplitudes distribution of `hchain_18`, after 0, 30, 60 and 90 operations from left to right. Blue and orange lines denote real and imaginary parts of an amplitude respectively.

In general, let us assume we have an operation involving m states, if all of the states are zero, these m states remain zeros after applying any operation. As a result, we do not need to transfer the zero state amplitudes to the GPU as their values will not change. Therefore, one can reduce the data movement between CPU and GPU by pruning the zero state amplitudes. One intuitive approach is to check each state value by traversing all states. However, a more efficient approach can be adopted, as we illustrate below.

Pruning Mechanism: In the proposed Q-GPU, we use bits in a binary string as flags to indicate

Table 7: The number of total operations and the number of operations before all qubits are involved for all circuits with 34 qubits.

Circuit	Total Operations	Operations Before Completely Involved
hchain	1786	272
rqc	124	54
qaoa	754	19
gs	37	16
hlf	48	16
qft	184	13
iqp	146	132
qf	222	16

whether a qubit has been involved after a set of gate operations (denoted as *involvement* in Algorithm 2). Initially, all the bits in *involvement* are set to 0. When q_k is involved, the kth bit in *involvement* is set to 1. Recall that the state vector is partitioned into chunks, the index of a chunk, i.e., *iChunk*, determines whether a chunk will be transferred or not. To compare *iChunk* with flag bits in *involvement*, we define *iChunk*$'$ as the left-shifted *iChunk* to align with *involvements*. When *iChunk*$'$ is larger than *involvement*, it indicates that at least one bit of *iChunk*$'$ is 1 and the corresponding flag bit in *involvement* is 0. In this situation, the corresponding qubit (i.e., indexed by this flag bit) has not been involved by any operation. As such, we skip the remaining chunks and stop the iteration (line 5). On the other hand, if *iChunk*$'$ is smaller than or equal to *involvement*, the redundancy within a chunk is determined by *iChunk*$'$ & *involvement* (line 8). For a qubit whose corresponding bit in *iChunk*$'$ is 1, if it has already been involved by previous operations, its corresponding bit in *involvement* is also 1. Therefore, for all the qubits that is 1 in *iChunk*$'$, if all of them have already been involved by previous operations, *iChunk*$'$ & *involvement* results in *iChunk*$'$ itself. Otherwise, all the state amplitudes within this chunk are zeros, and we can prune this chunk. Moreover, the *chunkSize* here is dynamically determined rather than a statically fixed value, which enhances the benefit of the above-discussed strategy. Specifically, we select *chunkSize* by finding the least non-zero bit of *involvement*. This is useful, especially at the beginning of the simulation where many state amplitudes are zeros. For instance, assuming we have an 8-qubit circuit and the *involvement* flag is 00000011 at the early execution stage, the *chunkSize* is dynamically set to 2, which has fewer zeros within a chunk compared to a larger chunk. The *involvement* flag bits are

Algorithm 2: Pruning zero state amplitudes.

```
   Variable list:  N          Total chunks number in CPU,
                   involvement Flag indicating which qubits are involved
1  /* Determine chunkSize by locating the least non-zero bit of involvement */
2  chunkSize, N = getChunkSize(involvement)
3  for iChunk ← 0 to N − 1 do
4      iChunk' = iChunk << chunkSize
5      if iChunk' > involvement then
6          break
7      if iChunk'&involvement ≠ iChunk' then
8          continue
9      /* Amplitudes update */
10     ...
11 updateInvolvement(involvement)
```

updated according to the qubits involved in each operation (line 14). In Figure 21, the proposed pruning mechanism (Ⓘⓥ) further saves (❽) cycle over Ⓘⓘⓘ.

B.5.3 Reordering to Delay Qubit Involvement

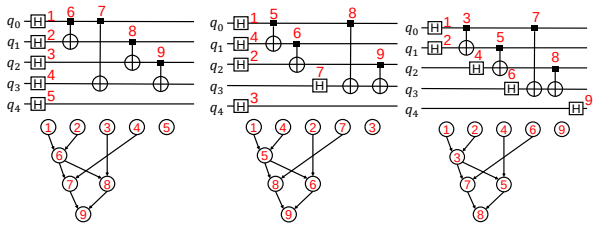


Figure 23: A walk-through example to illustrate the reordering benefits using gs_5. The red number denotes the operation orders before and after reordering.

In order to enlarge the potential of pruning, such that more state amplitudes are zeros during simulation, we propose compiler-assisted, dependency-aware quantum operation reordering to delay the involvement of qubits. Specifically, when applying a gate, we choose the one that incurs the minimum number of additional qubits to be involved with those qubits that have been already involved by previous operations. For example, Figure 23a shows the gs_5 circuit in the original execution order. The first five gates are H gates, where each gate applies to an individual qubit. As a result, once these gates have been applied, all the five qubits are involved. The next operation

is a CNOT gate applied to qubits q_0 and q_1 ($CNOT_6$). All the state amplitudes are likely non-zero because the qubits are involved by the H gates. Therefore, applying this CNOT gate requires updating all the non-zero amplitudes in the state vector, leading to moving and traversing the entire state vector on the GPU. However, the $CNOT_6$ can be executed before some of the H gates without violating the circuit semantics. This gate reordering allows more zero state amplitudes (fewer data movements) when simulating the $CNOT_6$ gate. It is also important to emphasize that any reordering must ensure that the gate dependencies are presented. For instance, $CNOT_6$ and $CNOT_7$ cannot be reordered due to the dependency on q_0.

To this end, we propose a compiler-assisted optimization to reorder the gate sequence with the goal of delaying the qubit involvement. Specifically, gates that are applied on different qubits in a quantum circuit can be executed independently in any order and the execution sequence of these independent gates does not affect the final simulation result[272,296,313]. This provides us the opportunity to reorder the independent gates, we use a directed acyclic graph (DAG) to represent the gate dependency in a circuit. Based on the DAG, we reorder the independent gates such that the simulation sequence involves the minimum number of new qubits when simulating each gate. Specifically, we investigate two heuristic strategies: 1) greedy reordering, and 2) forward-looking reordering.

Greedy reordering: greedy reordering traverses the DAG in topological order and greedily selects the gate (i.e., node in the DAG) that introduces the minimum number of new qubits to the list of updated qubits. The details of this method are illustrated in Algorithm 3. First, gates without predecessors in the DAG can be executed at the first steps and are put into *exeList*. Second, we traverse the gates in *exeList* and find the one that introduces the minimum number of newly involved qubits (lines 13 to 19). Then, we remove this gate from *exeList* and append it to the list of re-ordered gates. Third, we traverse the descendants of this gate and if a descendant does not have any predecessors other than this current gate, it will be added to *exeList* (lines 22 to 27). The second and the third steps are repeated until *exeList* is empty. In the rest of this section, we use Figure 23a as the example to illustrate how we perform reordering. At first, the *exeList* is [g_1, g_2, g_3, g_4, g_5]. Since each of these five gates involves one new qubit, we randomly select one gate among them to start simulation. In this example, g_1 is selected as the starting gate. After traversing all its descendants, no new gates can be added into *exeList*. Next, the *exeList* becomes [g_2, g_3, g_4, g_5]. In the next

Algorithm 3: Quantum operation reorder.

```
   Input :  DAG       A DAG representing circuit dependencies.
   Output:  gatesList  List of gates after reordering,
1  gatesList = [ ]
2  exeList = [ ]
3  /* First we build DAG and push gates without predecessors to an execution list                    */
4  for g in DAG do
5      if g.numPredecessors() == 0 then
6          exeList.append(g)
7  /* Then we traverse DAG in topological order and greedily decides the execution order of the gates  */
8  while exeList ≠ ∅ do
9      nextGate = NULL
10     minCost = 0
11     for g in exeList do
12         cost = g.getCost()
13         if cost < minCost then
14             minCost = cost
15             nextGate = g
16     exeList.erase(nextGate)
17     gatesList.append(nextGate)
18     for g in nextGate.descendants() do
19         g.numPredecessors() = g.numPredecessors() − 1
20         if g.numPredecessors() == 0 then
21             exeList.append(g)
```

three steps, we randomly select g_3, g_5 and g_2 since no new gates can be executed and all gates in *exeList* have equal priority. Then the *exeList* becomes [g_4, g_6]. At this time, *involvedQubits* is [q_0, q_1, q_2, q_4]. Therefore, g_4 involves one new qubit (q_3), whereas g_6 will not introduce any new qubits since it acts on q_0 and q_1 that are already in the involved list. Therefore, we will greedily select g_4 to execute since it involves the least new qubits. One can follow these reordering steps to reach the new ordering shown in Figure 23b. As a result, the number of involved qubits at each step is $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 4 \rightarrow 4 \rightarrow 5 \rightarrow 5 \rightarrow 5$. Since the baseline is $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5 \rightarrow 5 \rightarrow 5 \rightarrow 5 \rightarrow 5$, the final involvement is delayed by two steps. However, a better solution for reordering is to select g_2 and g_6 in the second and the third step, since applying these two gates only adds one qubit to *exeList*, while applying g_3 and g_5 adds two. Thus greedy reordering may misses the optimal choice.

Forward-looking reordering: To address the deficiency in greedy-reordering, we propose Forward-looking reordering that looks ahead of all the equal-priority gate candidates before making a decision. We implemented a *cost* counter to determine the priority of the gates in *exeList*. In *greedy reordering*, the *cost* is simply computed by counting new involved qubits (line 3-8 in Algorithm 4). The *cost* in forward-looking reordering is computed using Algorithm 4. Note that, *exeList*

Algorithm 4: Cost calculation in forward-looking reordering.

```
   Input :  g          Gates from exeList,
            exeList    List of gates that are executable,
            involvedQubits Set of qubits which have already been acted on.
   Output:  cost       Potential involved qubits after executing g.
1  costCurrent = 0, costLookAhead = 0
2  /* First we compute additional qubits that will be acted on by executing current gate   */
3  for q in g.qubits() do
4      if q not in involvedQubits then
5          costCurrent = costCurrent + 1
6          involvedQubits.insert(q)
7  exeList.erase(g)
8  for g' in g.descendants() do
9      if g'.numPredecessors() == 1 then
10         exeList.push(g')
11 /* Then we traverse current exeList and compute the cost of selecting a gate that involve least
      additional qubits                                                                    */
12 for g'' in exeList do
13     curCostLookAhead = 0
14     for q' in g''.qubits() do
15         if q' not in involvedQubits then
16             curCostLookAhead = curCostLookAhead + 1
17     if curCostLookAhead < costLookAhead then
18         costLookAhead = curCostLookAhead
19 cost = costCurrent + costLookAhead
20 return cost
```

and *involvedQubits* are just copies of the original ones, thus their original values are not changed. In forward-looking reordering, the *cost* of selecting a gate in *exeList* consists of two components: *costCurrent* and *costLookAhead* (line 1). The *costCurrent* is the same with the *cost* used in greedy reordering. Let us still use the example in Figure 23a to illustrate Algorithm 4. Initially, the *exeList* is also [g_1, g_2, g_3, g_4, g_5]. We take g_1 as an example to explain the computation of *costLookAhead*. First, we assume g_1 has already been executed. Then, the *costCurrent* is 1 and *involvedQubits* becomes [q_0] (lines 3-8). Since no descendants of g_1 can be executed, the *exeList* becomes [g_2, g_3, g_4, g_5] (lines 9-14). Then, we traverse the *exeList*. For each gate in *exeList*, we compute the cost of selecting this gate by counting the new involved qubits (lines 18-21) and selecting the least cost as *costLookAhead*. Now, executing any gate in *exeList* will involve one new qubit, thus *costLookAhead* is computed as 1 (lines 16-26). Similarly, one can find that all gates at the first step have equal priority. For the purpose of illustration, we assume g_1 is randomly selected. Then the *exeList* becomes [g_2, g_3, g_4, g_5]. Although all gates still have equal *costCurrent*, we can find that g_2 has the least *costLookAhead*. The reason is that, when we assume executing g_2 and look ahead from g_2, we find that executing g_6 introduces no new qubits. In contrast, look ahead after execut-

ing other gates will introduce new qubits. Finally, we get the result of forward-looking reorder as shown in Figure 23c. Clearly, the *involvement* at each step become $1 \rightarrow 2 \rightarrow 2 \rightarrow 3 \rightarrow 3 \rightarrow 4 \rightarrow 4 \rightarrow 4 \rightarrow 5$. Compared with greedy reordering, we further delay the final involvement by two steps.

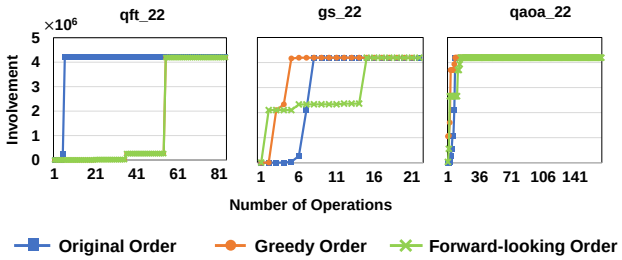


Figure 24: Qubit *Involvement* during simulation in three representative circuits.

Reorder effectiveness: To assess the performance of the reordering algorithms discussed above, we implement them to reorder the original operation sequences for all benchmark circuits that have 22 qubits and plot the *involvement* (Algorithm 2 in Section B.5.2) after each gate has been applied. For the purpose of illustration, we depict the results of three representative benchmark circuits in Figure 24. For each order, i.e original order, greedy-reorder, and forward-looking reorder, the "speed" of reaching the maximum *involvement* indicates the pruning potential. We observe that, forward-looking reordering results in the largest pruning potential, while greedy reordering only works for `qft_22` and even results less pruning potential than baseline for `gs_22`. Particularly. for `gs_22` and `qft_22`, forward-looking reordering effectively delays the involvement of qubits. Thus, we can expect the pruning potentials of these circuits to be enlarged by forward looking reordering. However, for `qaoa_22`, none of the reordering algorithms work due to the prevalent dependencies among the gates. Refering back to Figure 21, when reordering (Ⓥ) is employed, we can prune more chunks, which saves additional ❸ cycles compared to Ⓘⓥ.

B.5.4 Non-zero State Compression

Compressibility: While pruning removes the zero state amplitudes, those non-zero amplitudes still cause data movement overheads especially for circuits that do not have large pruning potentials (e.g., `qaoa` in Figure 24). Targeting reducing the data movement caused by non-zero state amplitudes, we investigate the potential compressibility and propose a GPU-supported efficient

lossless data compression in Q-GPU. Specifically, we observe that many non-zero entries within a state vector, after each operation, have similar amplitude values. In other words, there is a significant "spatial" similarly among consecutive state amplitudes in the state vector. To demonstrate the compressibility, we use `qaoa_20` and `iqp_20` as examples and show the residuals by subtracting the consecutive state amplitudes. As one can observe from Figure 25, for `qaoa_20`, most of the residuals are zero or very close to zero, indicating a potential for residual-based compression. However, `iqp` will be less compressible due to more diverse distribution.

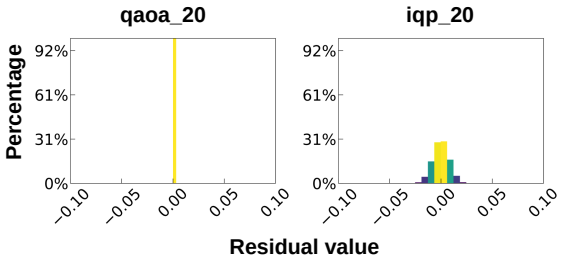


Figure 25: Residual distributions for `qaoa_20` and `iqp_20`.

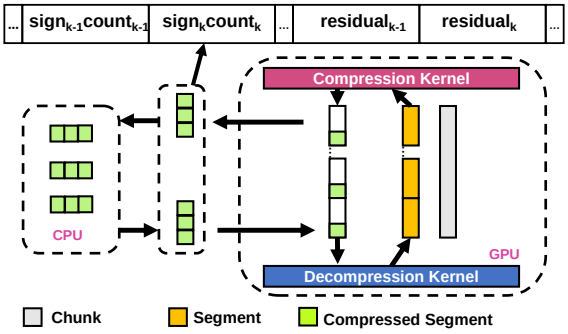


Figure 26: Overview of compression in Q-GPU.

Compression Strategy: We use the GFC algorithm[314] in Q-GPU. We implement the GFC as GPU kernels to perform the compression in parallel, thereby reducing the compression and decompression overheads. Specifically, the amplitudes on the GPU are partitioned into micro-chunks with a size of 32 amplitudes. Each GPU warp iteratively compresses/decompresses in parallel. Figure 26 (on the top) shows the compressed format. For the 32 values of a micro-chunk, we first store a 4-bit prefix for each of them, where one bit is used to record the sign of the residual and

another three bits are a count of leading zero bytes of the residual. Figure 26 also illustrates the GPU support of compression and decompression in Q-GPU. The compression is performed on the GPU after updating the chunk before copying it to the CPU. All of the chunks are equally divided into "segments". We empirically choose the segment size to match the GPU parallelism such that the GPU is properly utilized during compression. The compressed segments are transferred to the CPU instead of the original state chunks. The CPU keeps the compressed segments and copies the compressed segments to the GPUs upon request. Once the chunks are copied to the GPU, the amplitudes are decompressed, updated, and then compressed. As can be seen from Figure 21, compression (Ⓥ) saves ❶ cycles over Ⓤ and introduces negligible overhead. Later, in section B.6, we quantify the overheads incurred by the compression and decompression procedures.

B.6 Experimental Evaluation

In this section, we evaluate Q-GPU using the eight circuits in Table 6. We implement Q-GPU by substantially extending IBM QISKit-Aer. The evaluation is conducted on the same CPU-GPU platform used for characterization. For all experiments, the default optimizations in QISKit-Aer are turned on in both baseline and Q-GPU evaluation. To show the effectiveness of each optimization, we test *six different versions* of executions for all quantum circuit benchmarks:

- *Baseline:* This version is the implementation with state-of-the-art GPU support[281] in QISKit-Aer that supports GPU acceleration. As illustrated in Section B.4.2, state amplitudes are statically allocated on the GPU and CPU in this version.
- *Naive:* This version is the intuitive implementation discussed in Section B.4.4, which dynamically allocates state amplitudes to GPU. The performance of this version is dominated by expensive data movements.
- *Overlap:* This version implements the first optimization – proactive state amplitude transfer – in Q-GPU. This version is built upon the *Naive* version and its details are discussed in Section B.5.1.
- *Pruning:* This version adds the proposed pruning mechanism (Section B.5.2) to *Overlap*. By skipping the data movement of zero state amplitudes, the amount of data movement is reduced.

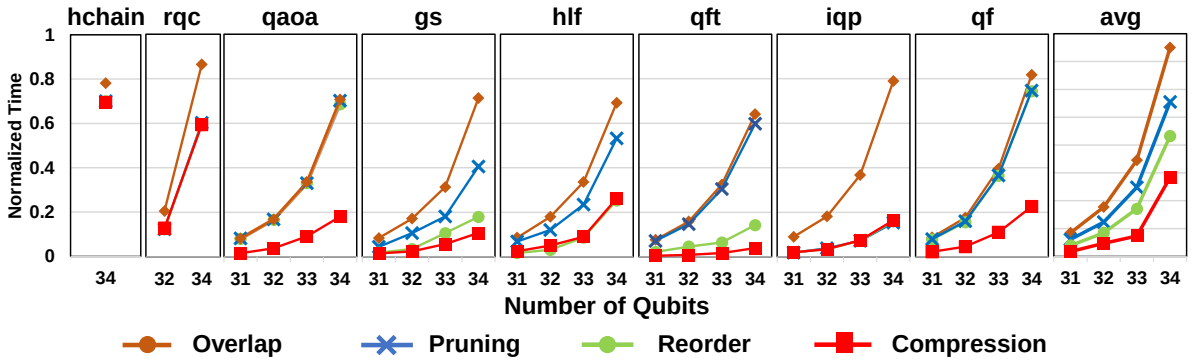


Figure 27: Normalized simulation time for circuits with different number of qubits (the lower the better).

- *Reorder:* In this version, we implement *forward-looking reorder* algorithm (Section B.5.3) to enlarge the potential for pruning. This reordering is performed by a simple compiler pass integrated in the Q-GPU.
- *Compression/Q-GPU:* In this version, all optimizations are employed with compression. We also call it *Q-GPU*. Compression (Section B.5.4) is added on top of *Reorder*. This version achieved the best performance.

B.6.1 Overall Performance

Figure 27 shows the overall performance and scalability among the six versions for all eight quantum circuits. The y-axis in the figure denotes the normalized execution time to the *Baseline* version. From the figure, one can make the following observations. First, by adding the proposed optimization in Q-GPU, our approach significantly reduces the execution time of QCS across all the circuits. Specifically, *Overlap*, *Pruning*, *Reorder*, and *Compression/Q-GPU* see a 24.96%, 44.54%, 56.78%, and 71.66% execution time reduction over the baseline execution for the largest number of qubits that can run on our platform. Second, the scalability of QCS performances is significantly improved by "breaking" the memory capacity in Q-GPU. The average achieved performance outperforms baseline by 2.53× for 34 qubits. Although we only simulate up to 34 qubits due to the CPU memory limitation (384 GB) in our system (Section B.4.3), one can infer from

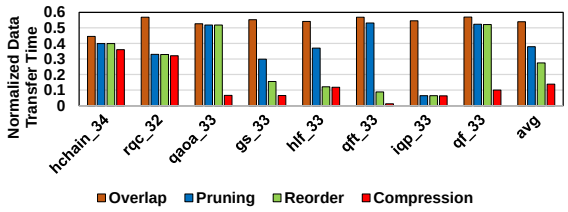


Figure 28: Normalized data transfer time (lower the better).

the trend that our optimizations are scalable to larger sized circuits. Third, Q-GPU has different accelerations for different circuits. Specifically, for `gs`, `qft`, `qaoa` and `iqp`, higher execution time reduction is observed, whereas for `hchain` and `rqc`, less speedup is observed. This is because, for `hchain` and `rqc`, reordering cannot enlarge the pruning potential because of dependent gates. Their amplitude residuals also have disperse distribution (similar to `iqp` in Figure 25). Thus, either *Reorder* or *Compression* improves little for these two benchmarks. Finally, for different circuits, a certain version may not have the same acceleration effects. For example, *Overlap* version generates a similar execution time reduction in all circuits tested. However, for *Pruning*, *Reorder* and *Compression*, the runtime reduction is different between different circuits. For example, *Pruning* and *Reorder* improve little for `qaoa` and `qf` because these two circuits do not have much potential of pruning the zero amplitudes. That is, their qubits get involved quickly with dependent operations. However, `qaoa` achieves significant benefits by compression as the great potential of compressibility. (discussed in Section B.5.4).

To further understand the execution reduction, Figure 28 plots, for each version, the exposed data movement time. In this figure, the y-axis represents the data movement time normalized to the *Naive* version. Clearly, one can observe a step-wise data movement reduction in the versions with our optimizations. First, *Overlap* uniformly reduces the data transfer time by an average of 46.14%. Note that, the savings generated in *Overlap* are independent of circuit types, that is the reason behind execution time reduction in Figure 27. For *Pruning* and *Reorder*, the reduction of data movement time varies in different circuits. This is because the number of zero state amplitudes and the potential of pruning heavily rely on the circuit type. For example, `qaoa`, `qft`, and `qf` get all qubits involved at early stage of simulation. Hence, pruning is less effective for these circuits

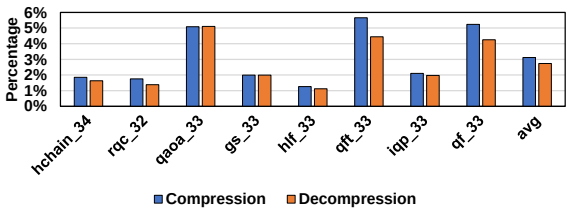


Figure 29: Compression and decompression overheads.

compared to others. Also, as discussed in Section B.5.3, *Reorder* has little effects on `hchain`, `rqc`, `qaoa`, and `qf` due to dependent operations in these circuits. Therefore, *Reorder* delivers similar data transfer time reduction with *Pruning* for these circuits. However, for those circuits with less dependent operations, *Reorder* significantly reduces their data movement time by enlarging the pruning potential. For circuits like `qaoa`, `gs`, `qft` and `qf`, *Compression* effectively reduces the data movement by leveraging the spatial similarity discussed in Section B.5.4. In a nutshell, for all circuit benchmarks tested, the reductions of data transfer time are the main reason behind the execution time reduction in Figure 27.

We also quantify the computation time of compression and decompression in Figure 29. Overall, the compression and decompression overhead is 3.12% and 2.74% of the GPU execution time. Potentially one may further optimize the compression and decompression by overlapping them on GPU, but we found the overhead is negligible compared to the significant reduction in execution time that we achieved. We also want to emphasize that the execution times reported in Figure 27 have all the sources of overhead included.

B.6.2 Comparison with OpenMP

Many publicly available quantum simulators and existing works employ OpenMP to parallelize the QCS on CPUs[272,279,292]. We compare Q-GPU with these OpenMP implementations. Specifically, we chose the OpenMP implementation in the most recent QISKit simulator and plot the results in Figure 30. We also compared our approach with other simulators in the next section, where the OpenMP is used by default in the simulators. On average, across eight circuits, Q-GPU outperforms the OpenMP QISKit by 1.79×. Particularly, Q-GPU achieved 12.79× speedup in `qft`.

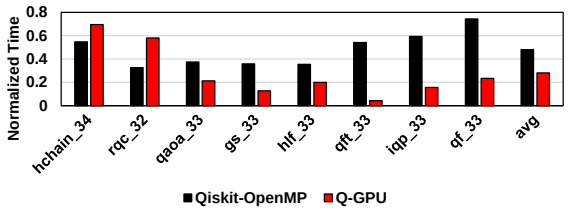
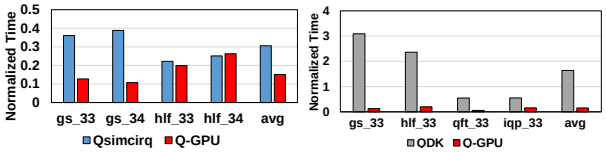


Figure 30: Comparison with OpenMP.

For `gs`, `iqp` and `qf`, Q-GPU achieves more than 2× speedup. However, for `hchain` and `rqc`, Q-GPU performs worse than OpenMP. This is because the pruning potential and the compressibility are low in both circuits where Q-GPU is less effective.

B.6.3 Comparison with Other Simulators



(a) Comparison with Qsim-Cirq (b) Comparison with QDK

Figure 31: Comparisons of Q-GPU to the simulator from Microsoft QDK v0.15 and Google Qsim-Cirq v0.8.0.

.

We compare Q-GPU with other simulators, including Google Qsim-Cirq v0.8.0 plus Cirq v0.9.2[282] and Microsoft QDK v0.15[283]. In our experiments, we run these simulators on the same CPU (Section B.4.2). Note that, both Qsim-Cirq and QDK are OpenMP enabled and we observe that they used all available threads during execution on the CPU. We report the results in Figure 31.

It is important to note that, to enable the simulation of the same circuits on Qsim-Cirq, we need to first transform our circuit benchmarks into OpenQASM codes[315]. Then, we need to import the OpenQASM codes to Qsim-Cirq for execution. Unfortunately, not all the transformed circuits can

be simulated on Qsim-Cirq due to the lack of support for particular gates (i.e., the "cp" gate cannot be recognized by Qsim-Cirq). As a result, we can only run `gs` and `hlf` successfully. This motivates our future research on uniform support of Quantum programming models. Figure 31a shows the normalized speedup of the proposed Q-GPU compared to Qsim-Cirq. Q-GPU outperforms the Google Qsim-Cirq by 1.02× on average.

To run the same quantum circuit on Microsoft QDK v0.15, we have to further convert the OpenQASM codes to "qsharp", i.e., the quantum language used in Microsoft. The conversion only succeeded for `qft`, `iqp`, `hlf`, and `gs`. The normalized simulation time is plotted in Figure 31b. On average, Q-GPU performs 9.82× better than Microsoft QDK.

B.7 Related Works

To the best of our knowledge, Q-GPU is the first work that systematically optimizes quantum circuit simulation on a GPUs. We summarize the related prior efforts below.

Prior works have focused on QCS optimizations on different platforms, from readily available devices to cloud environments[272,277–279,288,291,292,316]. Thomas et al.[277] simulated 45-qubits circuit using 8,192 nodes. They optimized single node performance by using automatic code generation and optimization of compute kernels. Edwin et al.[279] claimed to simulate more than 49 qubits by partitioning quantum circuits to "subcircuits" and delay their entanglements. In[278], the authors proposed lossy data compression to reduce the memory requirement of simulating large-scale quantum circuits. Aneeqa et al.[272] focused on fully exploiting single CPU performance for simulating a large number of qubits. The developed algorithm aims to reorder circuits such that more gates can be simulated in parallel. Compared with all these efforts, Q-GPU takes advantage of GPUs while managing the data movement between CPU and GPU. First, we identify the source of zero state amplitudes in QCS, and propose a pruning mechanism to safely reduce unnecessary computation on these states, which saves not only computation but also data movement. Unlike prior works using reordering to aggregate gates,[272,296,317] we propose reordering algorithms to enlarge the pruning potential. Moreover, Q-GPU is the first framework that leverages the GPU to implement a lossless compression that does not affect accuracy of QCS. Finally, it is important to

emphasize that Q-GPU is complementary to existing cloud-based quantum simulation frameworks, and can be integrated within these frameworks for further QCS improvements.

There are also several works that utilize GPUs to accelerate QCS[275,276,280,294,318–320]. Most of these works have limited capability in simulating large quantum circuits due to the limited memory capacity of GPUs. Ang et al.[275] proposed a multi-GPU centric QCS framework that tracks the density matrix. However, their framework cannot simulate a large number of qubits since it is limited by the aggregated memory capacity of multi-GPUs. For a single-node, they can only simulate up to 14 qubits on an NVIDIA V100 GPU. Jun et al.[280] proposed a CPU-GPU co-simulation method that enables simulation using a GPU even when the required memory exceeds the GPU memory capacity. Their method is also integrated into the IBM QISKit and is used as the baseline in this paper. In summary, compared to prior work, Q-GPU breaks the GPU memory capacity limitation, i.e., it is able to simulate 34 qubits which require 256 GB memory on a 16 GB memory GPU, and fully takes advantage of GPU parallelization. The fundamental design innovation behind this is to dynamically and proactively transfer the state amplitudes through end-to-end optimizations to minimize the data movement overheads caused by state amplitudes transfer.

B.8 Concluding Remarks

In this paper, we propose Q-GPU, a framework tailored with GPU optimizations to effectively improve the quantum circuit simulation performance for quantum circuits with a large number of qubits. The Q-GPU is able to deliver scalable simulation performance based on the four internal end-to-end optimizations, including i) proactive state amplitudes transfer, ii) zero state amplitudes pruning, iii) delayed qubit involvement, and iv) lossless non-zero state compression. Experimental results across eight representative quantum circuits indicate that Q-GPU achieves 2.53× average execution time reduction on a single GPU. It also outperforms the most recent OpenMP CPU implementation and other publicly available quantum simulators.

B.9 AFMC applications Towards Regional Embedding.

Collaborators: Shiv Upadhyay and Daniel S. Lambrecht.

In the previous chapter, we looked at partitioning an entire molecule into important domains in systems where there is no focal interaction. However, there are many cases in which a focal area of the calculation is known such as an adsorbate to a surface, a polymer chain growth site, or an enzyme in a binding pocket. One way in which these calculation are made more affordable is through localization schemes. A recent scheme was proposed for a regional embedding approach in which localization of the occupied and virtual molecular orbitals (MOs) occurs separately for a fragment area of interest. This fragment region was not defined in the original work requiring a set of convergence test for each system studied, taking away from the speed-ups granted by the method. In this chapter, the utility of the AFMC approach to determining this fragment area is presented. Specifically, the spectral clustering approaches as proposed in our previous work with a newly-developed descriptor that is tailored to target an area of interest within a chemical system. This approach allows for a reliable definition of the fragment area over which localization of the molecular orbitals occurs, which removes the necessity of convergence testing for fragment size.

Though these preliminary results are brief, they do suggest a path forward in using fragmentation approaches with regional embedding schemes, or other localization schemes. To move this work forward, the characterization of the current descriptors must be applied to a wider variety of systems to understand how system agnostic the descriptors can be. Additionally, other forms of descriptors should be investigated.

B.10 Characterization of silyl ketenes

Collaborators Krista Schoonover, Ian Baster, Sarah Mitchell, Emily Pentzer and Daniel S. Lambrecht.

A joint experimental and computational study to explore the preference for nucleophilic addition or deprotonation of combinations of different anions and different silyl ketene units. Specifically, the silyl ketene with TBDPS functional groups paired with various nucleophiles were ex-

plored with DFT. The addition reaction was thermodynamically favored, while deprotonation was primarily kinetically favored. The author's contribution to this work were in the calculation of the molecular electrostatic potential and small contributions in discussion to the machine learning study which was led and performed by Daniel. S. Lambrecht. This manuscript is currently in preparation.

Appendix C Supplemental Material for Chapter 4

The Supplementary Material document includes the total energies and error bars for the quantum Monte Carlo calculations, the total energies for the DFT calculations, and details of the convergence of the DFT total energies with respect to the *k*-point grid and kinetic energy cutoff of the plane wave basis, and a comparison of the density difference of DMC-PBE and DMC-HSE.

C.1 Convergence of PBE total and binding energies with respect to relevant parameters

In this section, the sensitivity of PBE total and binding energies with respect to *k*-point grid and cutoff of the kinetic energy of the wave function for the plane wave basis.

Table 8: PBE total energies of each system and binding energies (eV) with respect to increasing the Monkhorst-Pack *k*-point grid. The hydrogen total energy value used to calculate the binding energy was for a 1x1x1 *k*-grid with a 250 Ry wave function cutoff.

k-point grid	hydrogen on graphene	graphene	H	binding energy
2x2x1	-7759.736	-7745.293	-13.604	-0.839
3x3x1	-7759.751	-7745.339	-13.604	-0.808
4x4x1	-7759.749	-7745.323	-13.604	-0.823
5x5x1	-7759.749	-7745.322	-13.604	-0.824
6x6x1	-7759.749	-7745.325	-13.604	-0.820

Table 9: PBE total and binding energies (eV) with respect to varying the kinetic energy cutoff values of the wave function on a 4x4x1 k-point grid. The hydrogen total energy value used to calculate the binding energy was for a 1x1x1 *k*-grid with a 250 Ry wave function cutoff.

kinetic energy cutoff	hydrogen on graphene	graphene	H	binding energy
150	-7758.435	-7744.019	-13.603	-0.813
250	-7759.749	-7745.323	-13.604	-0.823
300	-7759.790	-7745.362	-13.604	-0.824
350	-7759.799	-7745.372	-13.604	-0.824

C.2 Hybrid functional total and binding energies (eV) with respect to *k*-point grid

Table 10: PBE0 and HSE total energies and binding energies (eV) with respect to increasing the Monkhorst-Pack *k*-point grid.

functional	*k*-point grid	hydrogen on graphene	graphene	H	binding energy
PBE0	6x6x1	-51820.644	-51806.227	-13.537	-0.880
PBE0	10x10x1	-51820.644	-51806.255	-13.537	-0.851
PBE0	12x12x1	-51820.644	-51806.253	-13.537	-0.854
HSE	6x6x1	-51820.264	-51805.892	-13.565	-0.807
HSE	10x10x1	-51820.264	-51805.907	-13.565	-0.792
HSE	12x12x1	-51820.264	-51805.905	-13.565	-0.794

C.3 Non-self-consistent field calculations details

Table 11: Non-self-consistent field calculations: system and binding energies used to calculate the differences in Table II of the manuscript, reported without basis set superposition error correction.

density functional	energy functional	graphene	H	Hgraphene	binding energies
pbe	pbe	-1903.884	-0.496	-1904.415	-.962
hse	pbe	-1903.883	-0.496	-1904.414	-.947
pbe	hse	-1903.989	-0.498	-1904.520	-.879
hse	hse	-1903.989	-0.498	-1904.521	-.905

C.3.1 DMC Calculation Details

C.3.1.1 Total Energies for DMC calculations

Table 12: Total and binding energies (eV) of a hydrogen atom on graphene calculated with QMC.

Trial wave function	hydrogen on graphene	graphene	hydrogen	binding energy
PBE	-7748.200 $\pm$ 0.012	-7733.893 $\pm$ 0.014	-13.616 $\pm$ 0.002	-0.691 $\pm$ 0.019

C.3.1.2 Total Energies per twist angle

Table 13: Total energies (eV) at each twist angle used for graphene DMC treatment.

Twist number	Total Energy
1	-7733.872 $\pm$ 0.046
2	-7734.059 $\pm$ 0.057
3	-7734.074 $\pm$ 0.058
4	-7733.851 $\pm$ 0.041
5	-7733.492 $\pm$ 0.052
6	-7733.467 $\pm$ 0.048
7	-7733.080 $\pm$ 0.057
8	-7734.137 $\pm$ 0.047
9	-7734.080 $\pm$ 0.041
10	-7733.853 $\pm$ 0.047
11	-7734.444 $\pm$ 0.050
12	-7733.983 $\pm$ 0.048

Table 14: Total energies (eV) at each twist angle used for hydrogen atom chemisorbed on graphene DMC treatment.

Twist number	Total Energy
1	-7748.443 ± 0.057
2	-7748.474 ± 0.045
3	-7748.283 ± 0.046
4	-7748.294 ± 0.047
5	-7748.006 ± 0.050
6	-7747.736 ± 0.056
7	-7748.748 ± 0.045
8	-7748.092 ± 0.063
9	-7747.601 ± 0.042
10	-7748.488 ± 0.047
11	-7747.941 ± 0.041
12	-7747.779 ± 0.050
13	-7747.761 ± 0.051
14	-7748.275 ± 0.045
15	-7748.655 ± 0.054
16	-7748.380 ± 0.053
17	-7748.037 ± 0.058
18	-7748.615 ± 0.047

C.3.2 DMC-DFT density difference

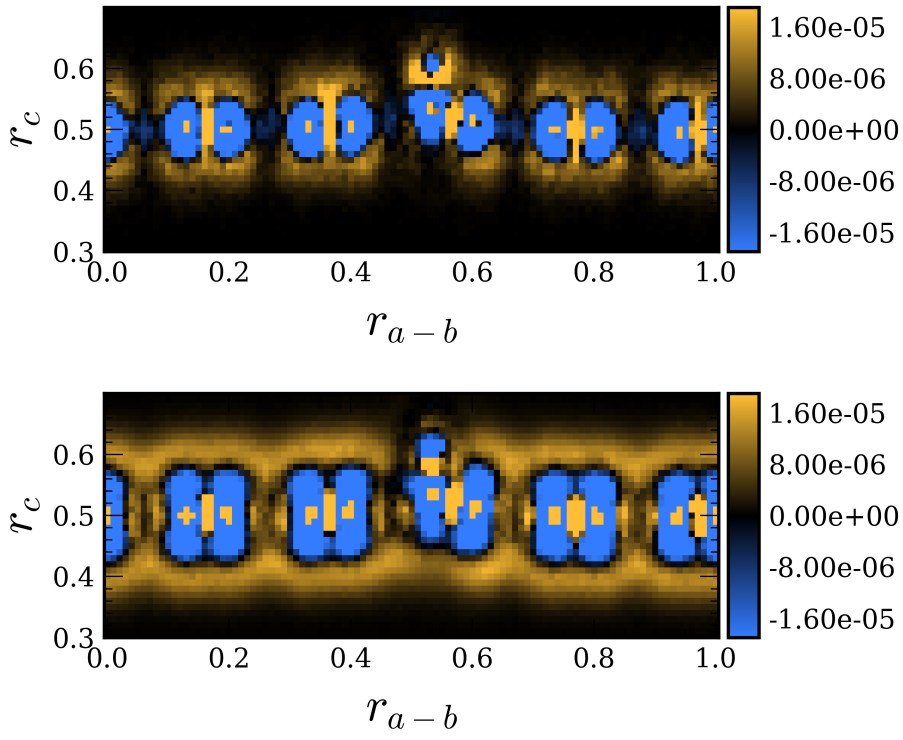


Figure 32: Similar to Figure 3 in the main text. DMC-PBE density (top) is reproduced from the main text and compared with the DMC-HSE density (bottom). We attribute the differences to primarily the difference in basis sets, plane waves vs Gaussian orbtials, and the use of pseudopotentials in the plane wave calculation.

Appendix D Supplemental Material for Chapter 5

Additional simulation details such as input files are available at: `https://github.com/shivupa/Water4_JCP_Special_Issue_Supplemental_Material`.

D.1 Geometries

The geometries are given in xyz format for the $(H_2O)_4$ structures in Angstroms.

D.1.1 Geometry at R = 4 Å

```
13
WATER 4 SYSTEM 4.0 Angstroms
X   0.000000   0.000000   0.000000
O   0.000000   1.730527   2.893437
H   0.000000   1.387569   2.000001
H   0.000000   2.681413   2.785434
O   0.000000  -1.730527   2.893437
H   0.000000  -1.387569   2.000001
H   0.000000  -2.681413   2.785434
O   0.000000   1.730527  -2.893437
H   0.000000   1.387569  -2.000001
H   0.000000   2.681413  -2.785434
O   0.000000  -1.730527  -2.893437
H   0.000000  -1.387569  -2.000001
H   0.000000  -2.681413  -2.785434
```

D.1.2 Geometry at R = 7 Å

```
13
WATER 4 SYSTEM 7.0 Angstroms
X   0.000000   0.000000   0.000000
O   0.000000   1.730527   4.393437
H   0.000000   1.387569   3.500001
H   0.000000   2.681413   4.285434
O   0.000000  -1.730527   4.393437
H   0.000000  -1.387569   3.500001
H   0.000000  -2.681413   4.285434
O   0.000000   1.730527  -4.393437
H   0.000000   1.387569  -3.500001
H   0.000000   2.681413  -4.285434
O   0.000000  -1.730527  -4.393437
H   0.000000  -1.387569  -3.500001
H   0.000000  -2.681413  -4.285434
```

D.2 Supplemental Basis Functions

The exponents for the 7s7p and 3s1p diffuse Gaussian type orbitals are given below in the GAMESS format.

D.2.1 7s7p

```
S   1
  1   0.02362232    1.0
S   1
  1   0.00738198    1.0
S   1
  1   0.00230687    1.0
S   1
  1   0.00072090    1.0
S   1
  1   0.00022528    1.0
S   1
  1   0.00007040    1.0
S   1
  1   0.00002200    1.0
P   1
  1   0.02362232    1.0
P   1
  1   0.00738198    1.0
P   1
  1   0.00230687    1.0
P   1
  1   0.00072090    1.0
P   1
```

```
  1   0.00022528    1.0
P   1
  1   0.00007040    1.0
P   1
  1   0.00002200    1.0
```

D.2.2 3s1p

```
S   1
  1   0.02362232    1.0
S   1
  1   0.00738198    1.0
S   1
  1   0.00230687    1.0
P   1
  1   0.02362232    1.0
```

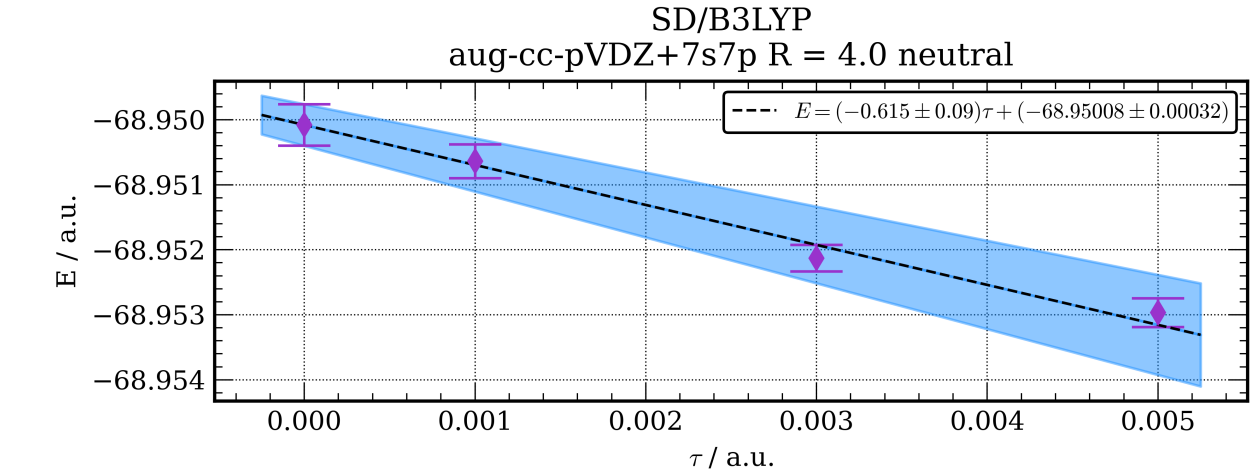
D.2.3 3s1p3d

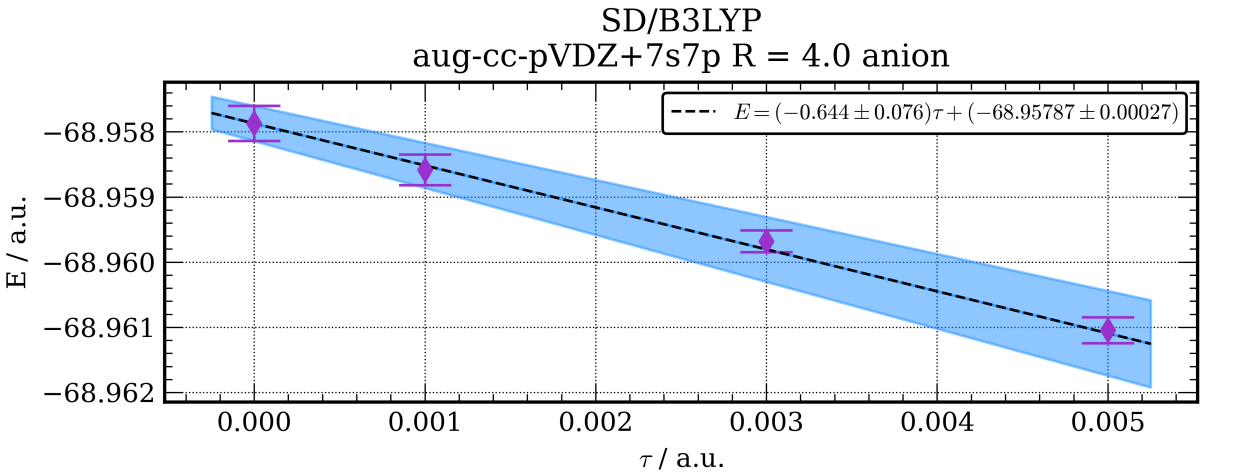
```
S   1
  1   0.02362232    1.0
S   1
  1   0.00738198    1.0
S   1
  1   0.00230687    1.0
P   1
  1   0.02362232    1.0
D   1
  1   0.02362232    1.0
D   1
  1   0.00738198    1.0
```

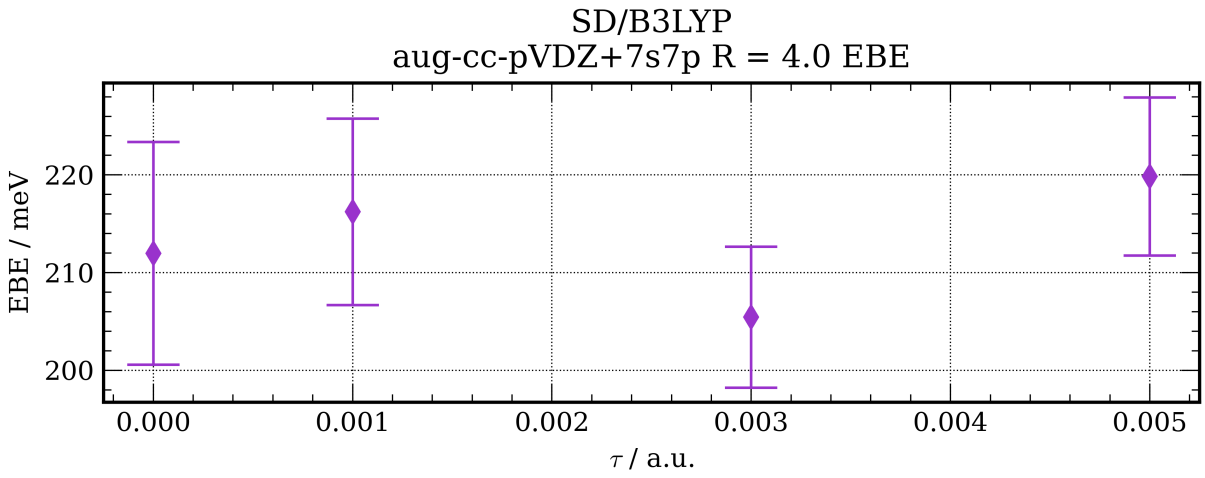
```
D   1
  1   0.00230687    1.0
```

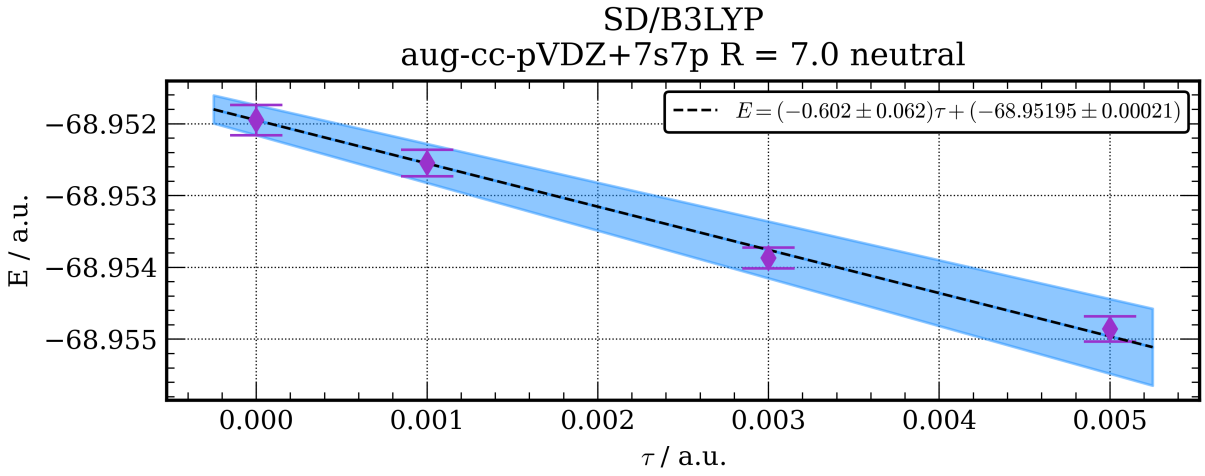
D.3 DMC Extrapolation

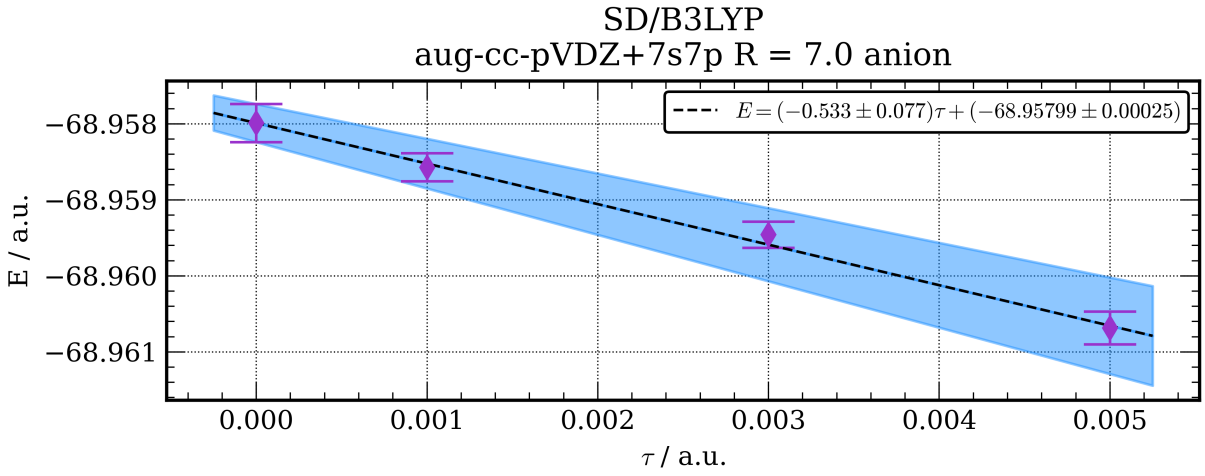
Summaries of the zero time step linear extrapolation are plotted below for the $w(H_2O)_4$ system, the R parameter is indicated in each plot title. The blue shaded region corresponds to the error in the fit of the DMC energies at the three timesteps (0.001, 0.003, 0.005).

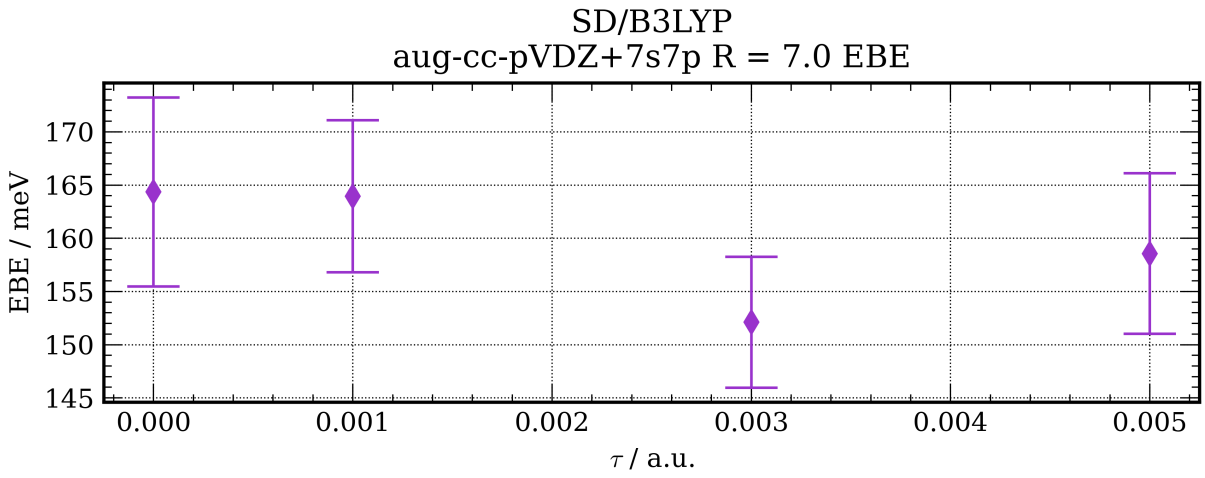


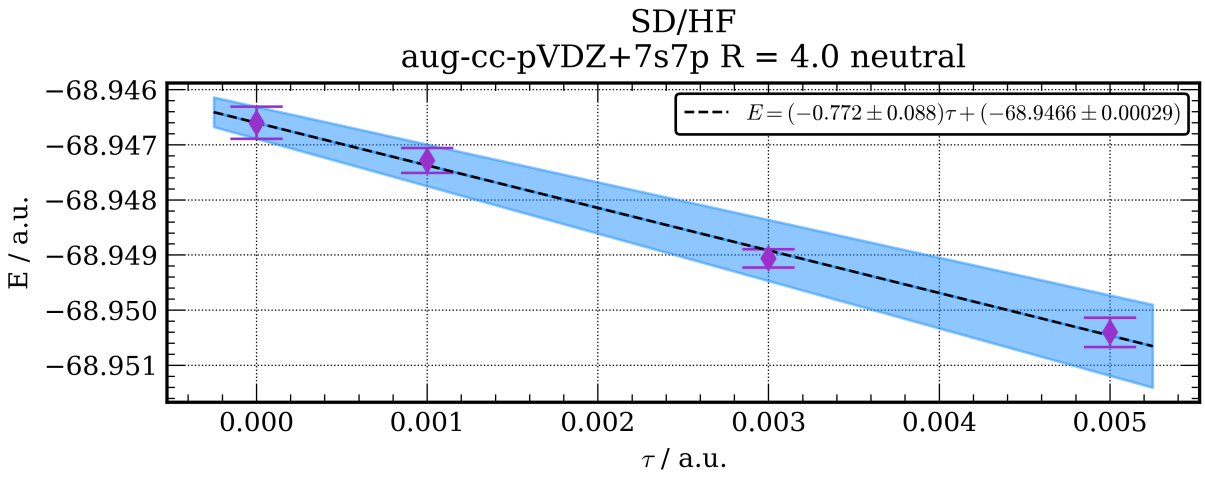


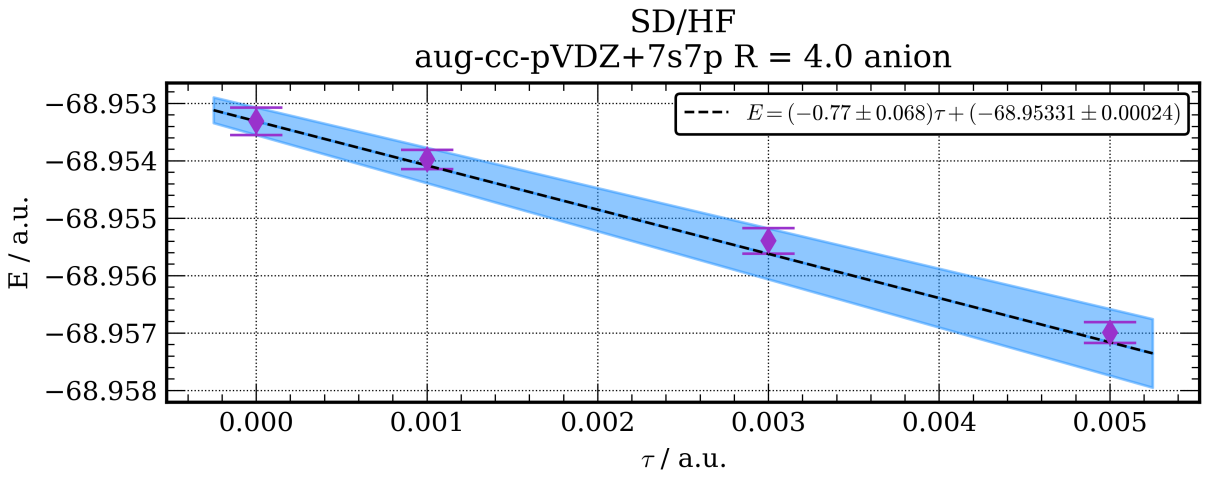

SD/B3LYP
aug-cc-pVDZ+7s7p R = 4.0 EBE
EBE / meV
220
210
200
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

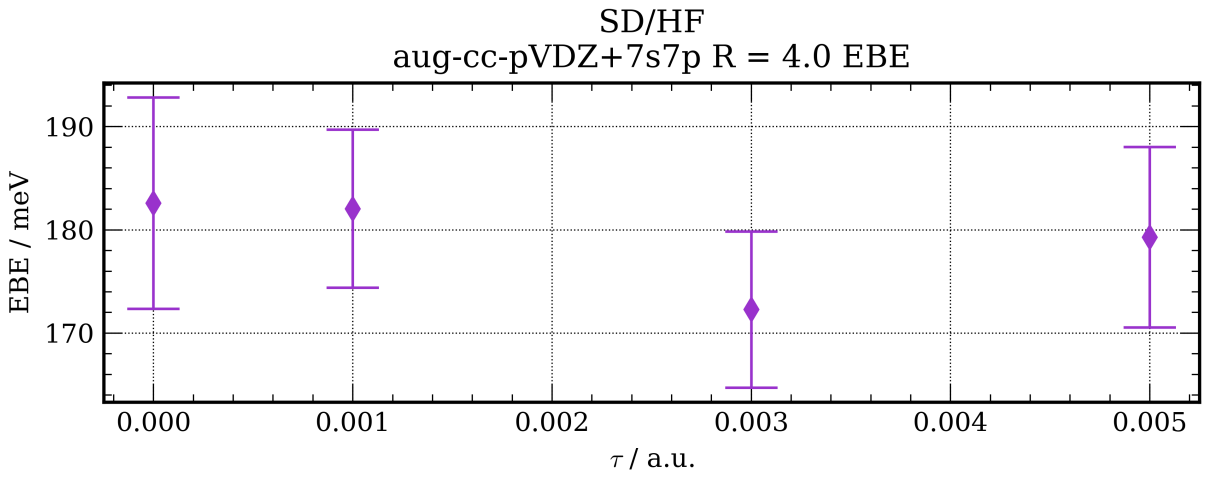

SD/B3LYP
aug-cc-pVDZ+7s7p R = 7.0 neutral
$E = (-0.602 \pm 0.062)\tau + (-68.95195 \pm 0.00021)$
E / a.u.
−68.952
−68.953
−68.954
−68.955
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

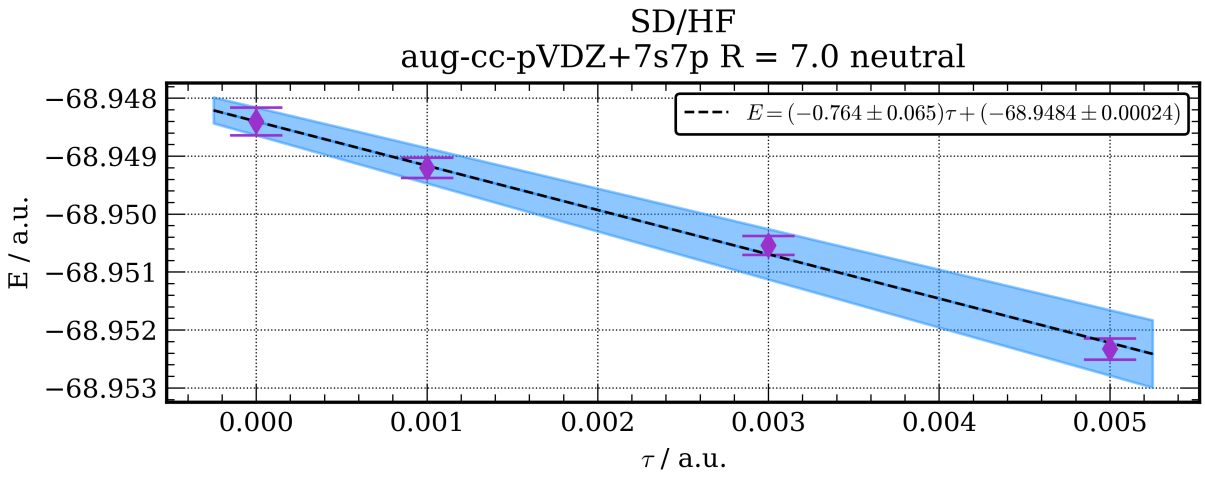

SD/B3LYP
aug-cc-pVDZ+7s7p R = 7.0 anion
$E = (-0.533 \pm 0.077)\tau + (-68.95799 \pm 0.00025)$
E / a.u.
−68.958
−68.959
−68.960
−68.961
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

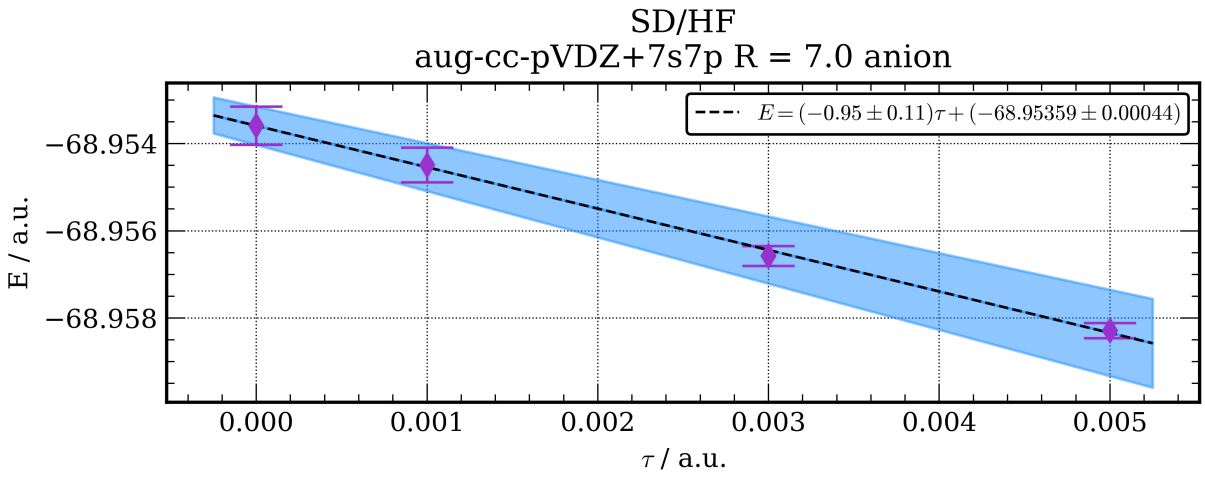

SD/B3LYP
aug-cc-pVDZ+7s7p R = 7.0 EBE
EBE / meV
170
165
160
155
150
145
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

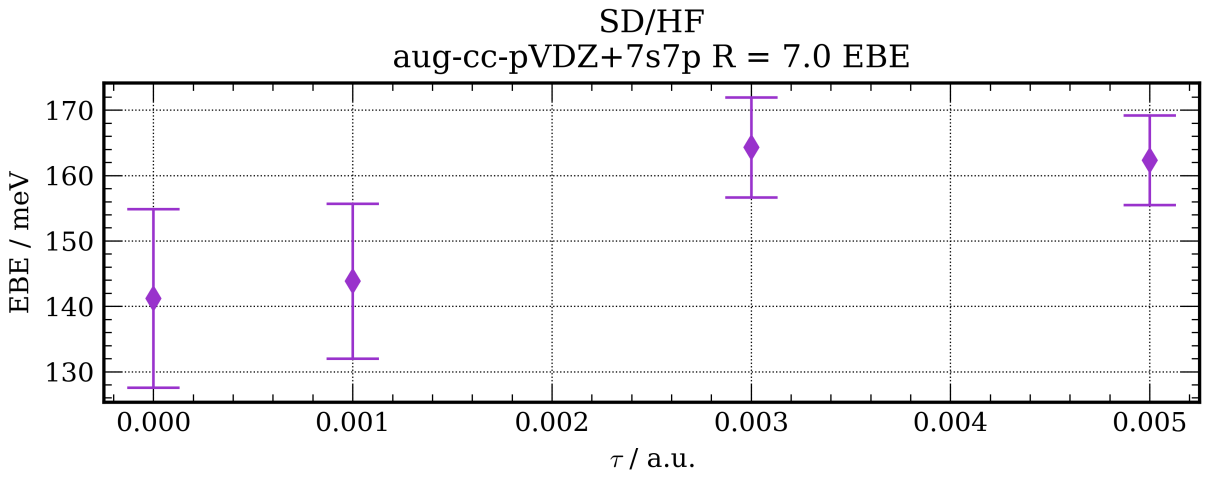

SD/HF
aug-cc-pVDZ+7s7p R = 4.0 neutral
E = (−0.772 ± 0.088)τ + (−68.9466 ± 0.00029)
E / a.u.
−68.946
−68.947
−68.948
−68.949
−68.950
−68.951
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

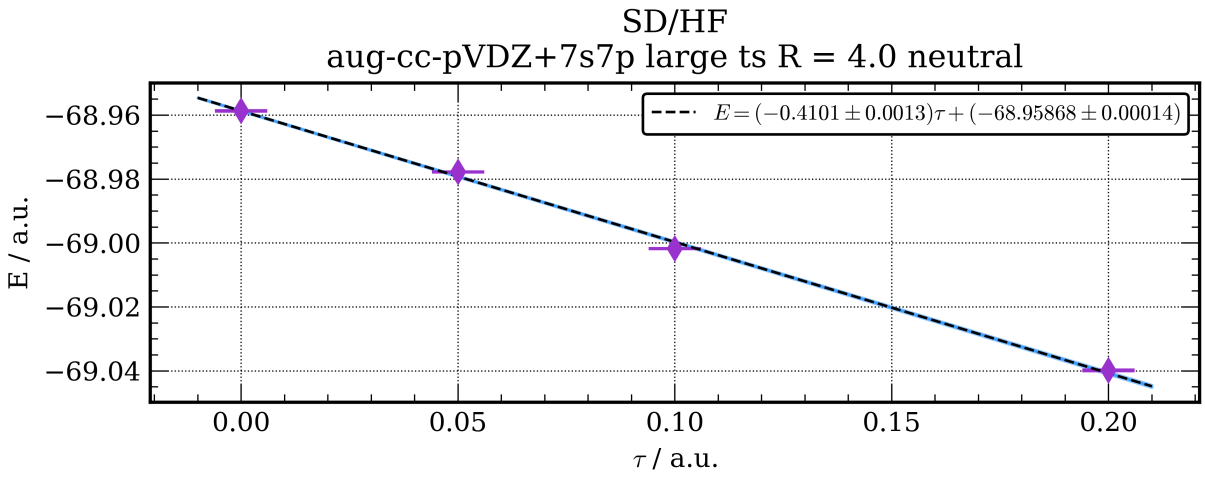

SD/HF
aug-cc-pVDZ+7s7p R = 4.0 anion
E = (−0.77 ± 0.068)τ + (−68.95331 ± 0.00024)
E / a.u.
−68.953
−68.954
−68.955
−68.956
−68.957
−68.958
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

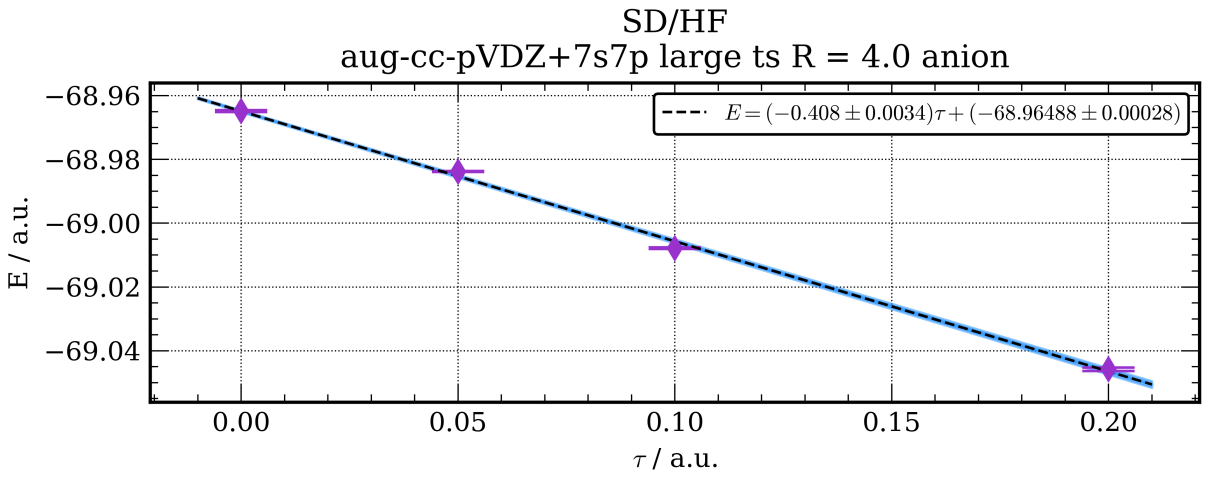

SD/HF
aug-cc-pVDZ+7s7p R = 4.0 EBE
EBE / meV
190
180
170
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

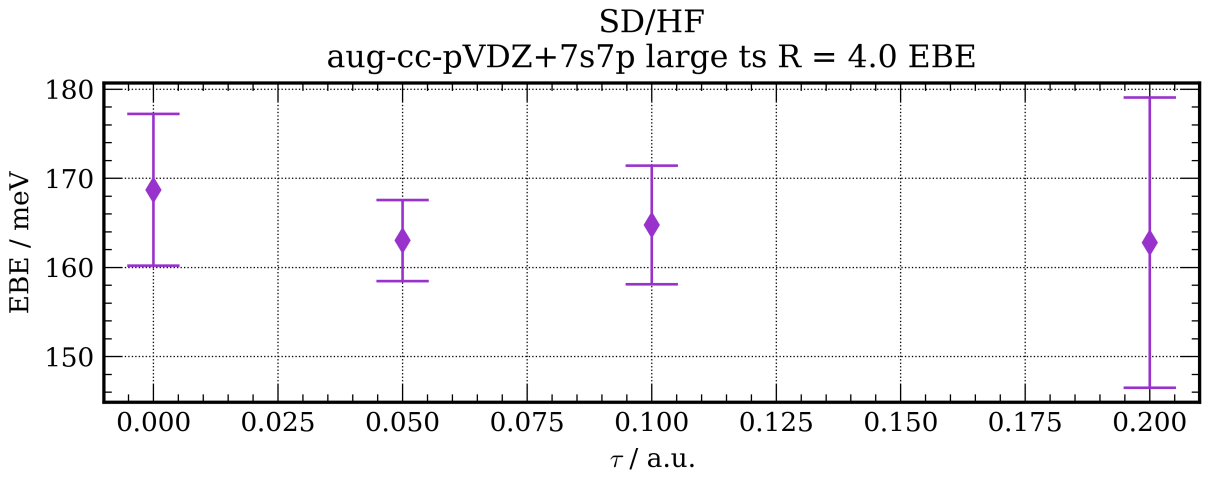

SD/HF
aug-cc-pVDZ+7s7p R = 7.0 neutral
E = (−0.764 ± 0.065)τ + (−68.9484 ± 0.00024)
E / a.u.
−68.948
−68.949
−68.950
−68.951
−68.952
−68.953
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

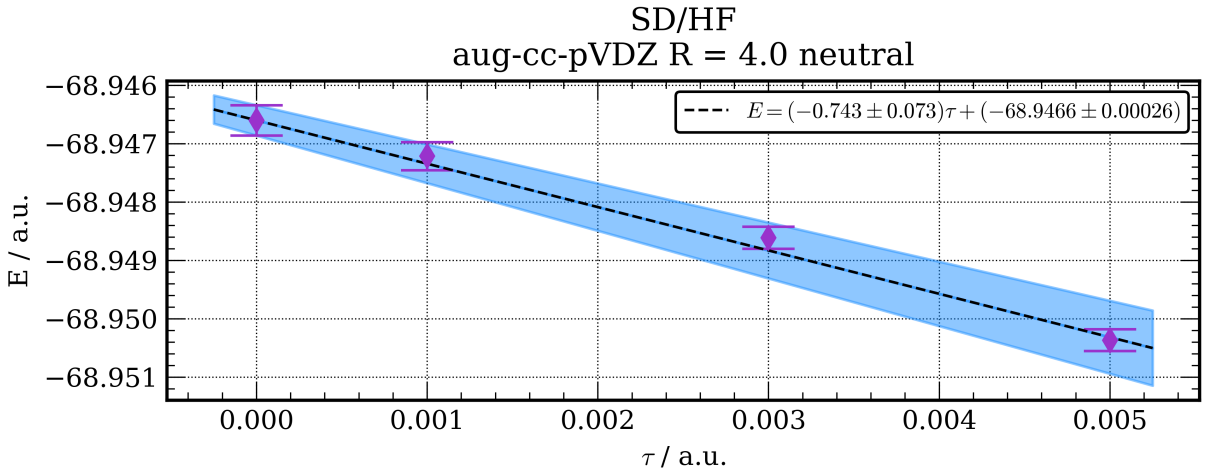

SD/HF
aug-cc-pVDZ+7s7p R = 7.0 anion
E = (−0.95 ± 0.11)τ + (−68.95359 ± 0.00044)
E / a.u.
−68.954
−68.956
−68.958
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

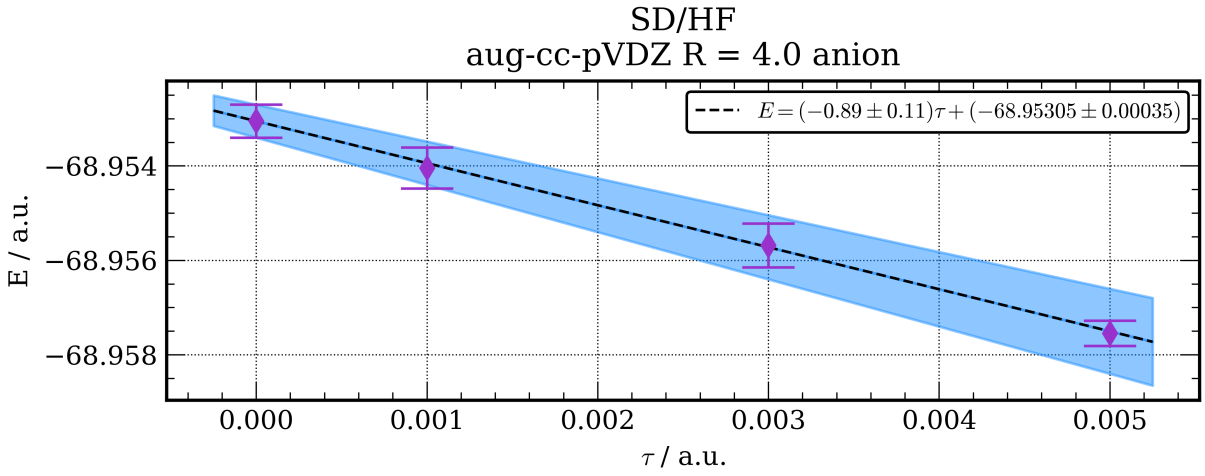

SD/HF
aug-cc-pVDZ+7s7p R = 7.0 EBE
EBE / meV
170
160
150
140
130
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

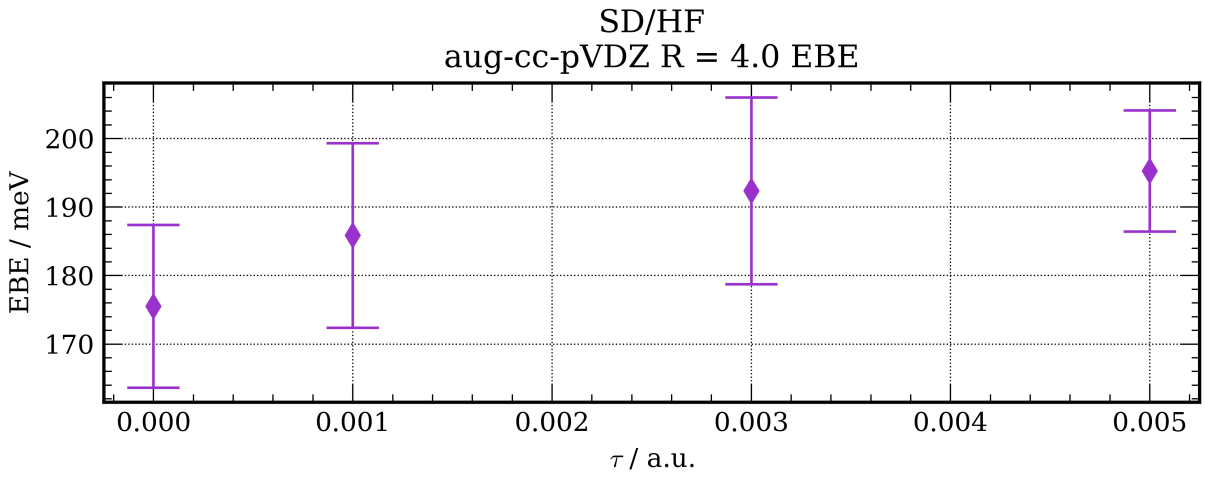

SD/HF
aug-cc-pVDZ+7s7p large ts R = 4.0 neutral
E = (−0.4101 ± 0.0013)τ + (−68.95868 ± 0.00014)
E / a.u.
−68.96
−68.98
−69.00
−69.02
−69.04
0.00
0.05
0.10
0.15
0.20
τ / a.u.

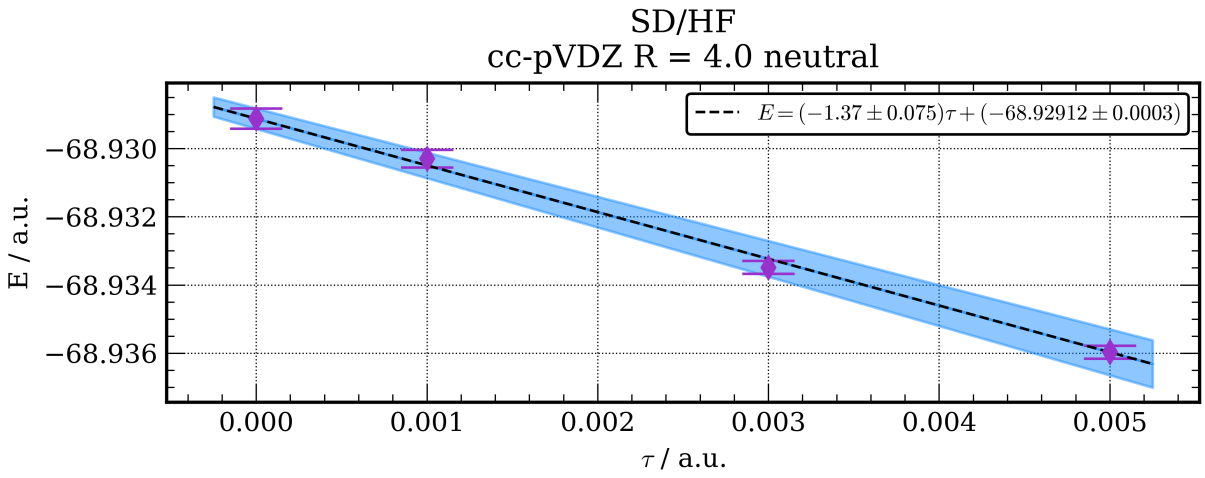

SD/HF
aug-cc-pVDZ+7s7p large ts R = 4.0 anion
E = (−0.408 ± 0.0034)τ + (−68.96488 ± 0.00028)
E / a.u.
−68.96
−68.98
−69.00
−69.02
−69.04
0.00
0.05
0.10
0.15
0.20
τ / a.u.

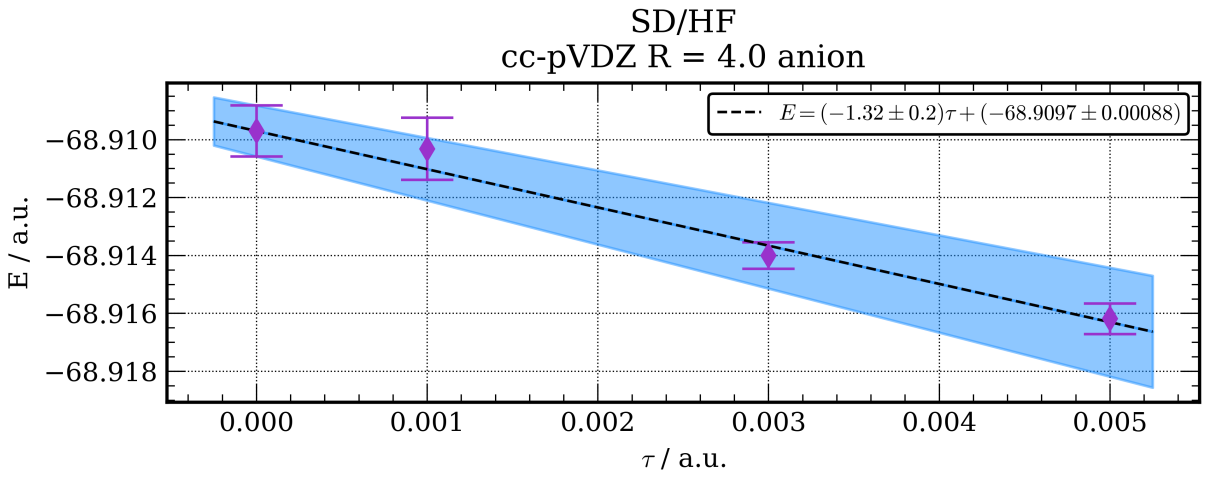

SD/HF
aug-cc-pVDZ+7s7p large ts R = 4.0 EBE
EBE / meV
180
170
160
150
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
τ / a.u.

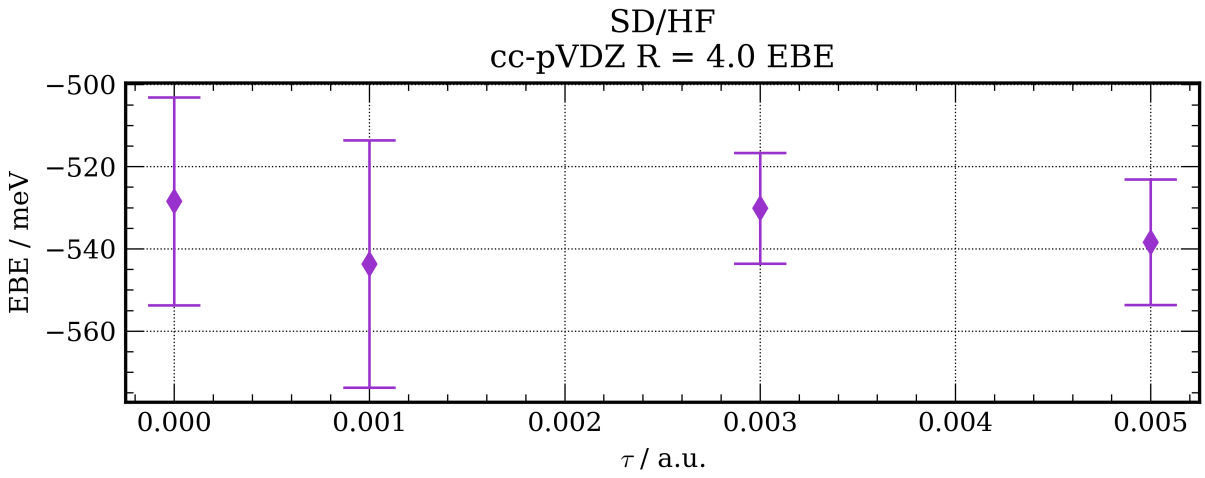

SD/HF
aug-cc-pVDZ R = 4.0 neutral
E = (−0.743 ± 0.073)τ + (−68.9466 ± 0.00026)
E / a.u.
−68.946
−68.947
−68.948
−68.949
−68.950
−68.951
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

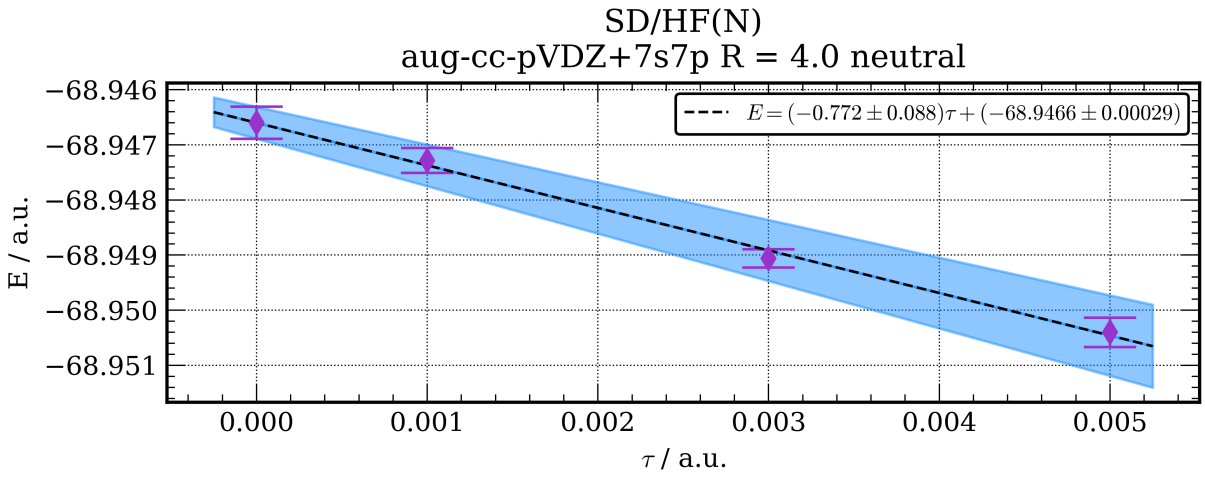

SD/HF
aug-cc-pVDZ R = 4.0 anion
E = (−0.89 ± 0.11)τ + (−68.95305 ± 0.00035)
E / a.u.
−68.954
−68.956
−68.958
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

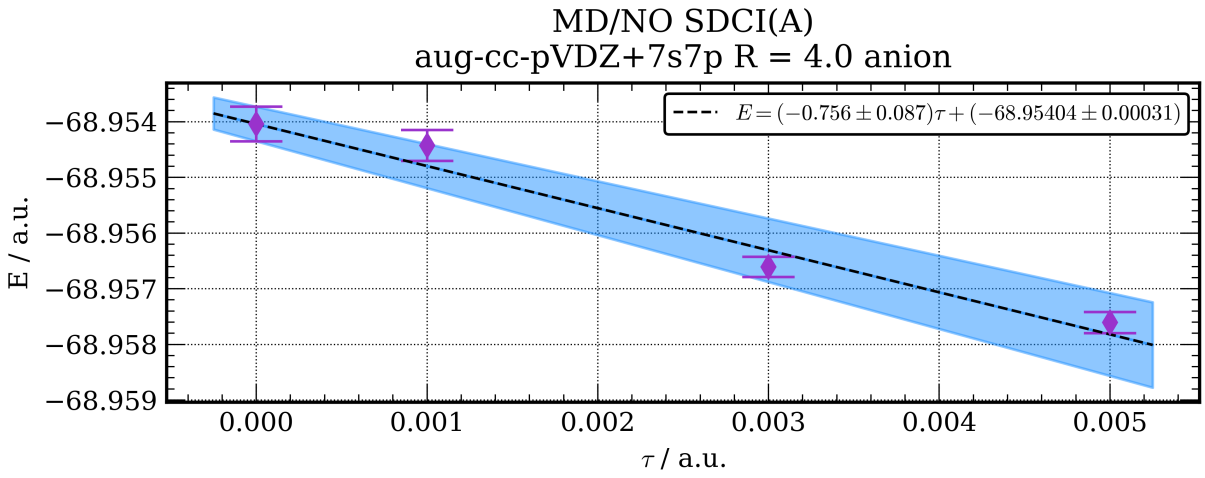

SD/HF
aug-cc-pVDZ R = 4.0 EBE
EBE / meV
200
190
180
170
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

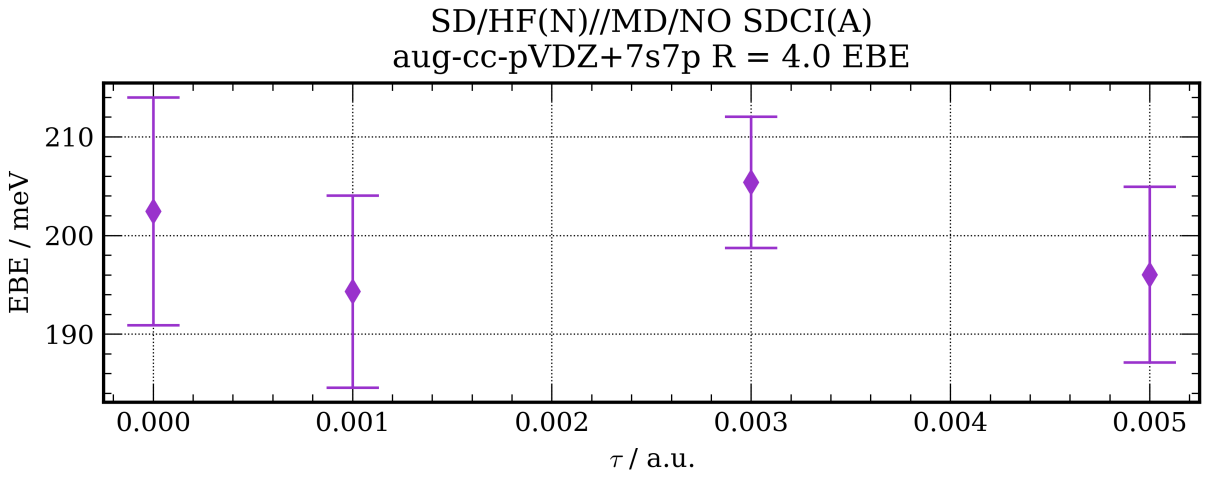

SD/HF
cc-pVDZ R = 4.0 neutral
$E = (-1.37 \pm 0.075)\tau + (-68.92912 \pm 0.0003)$
E / a.u.
−68.930
−68.932
−68.934
−68.936
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

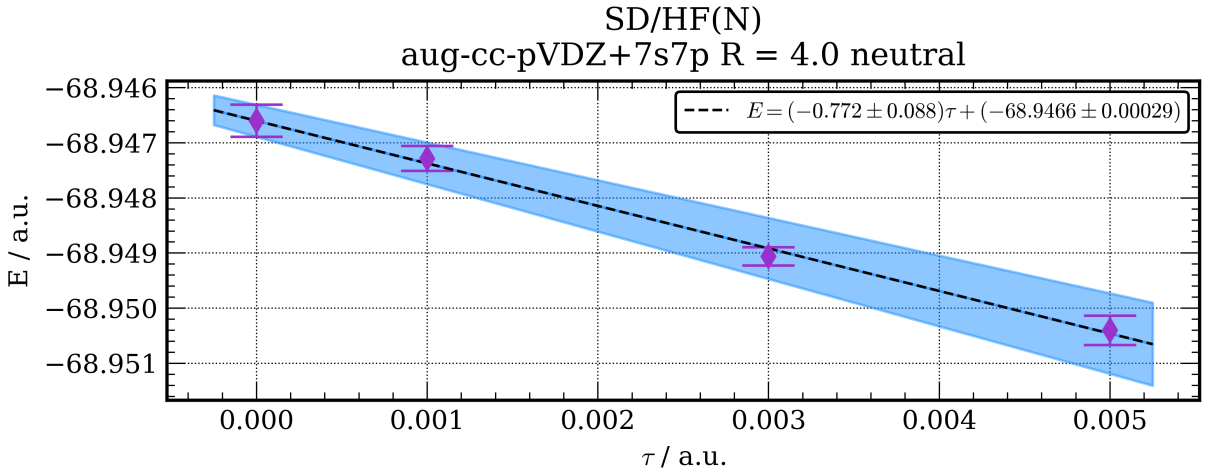

SD/HF
cc-pVDZ R = 4.0 anion
$E = (-1.32 \pm 0.2)\tau + (-68.9097 \pm 0.00088)$
E / a.u.
−68.910
−68.912
−68.914
−68.916
−68.918
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

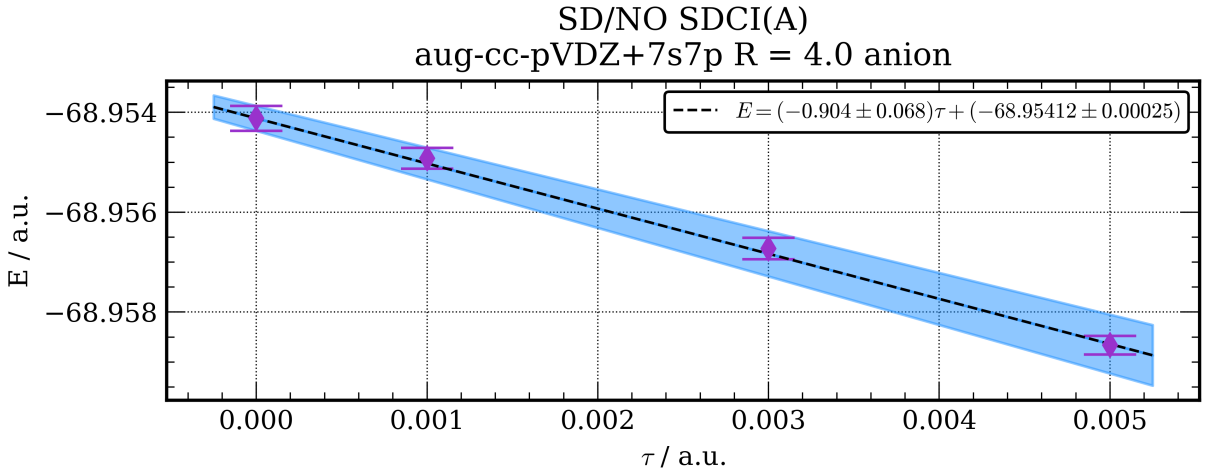

SD/HF
cc-pVDZ R = 4.0 EBE
EBE / meV
−500
−520
−540
−560
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

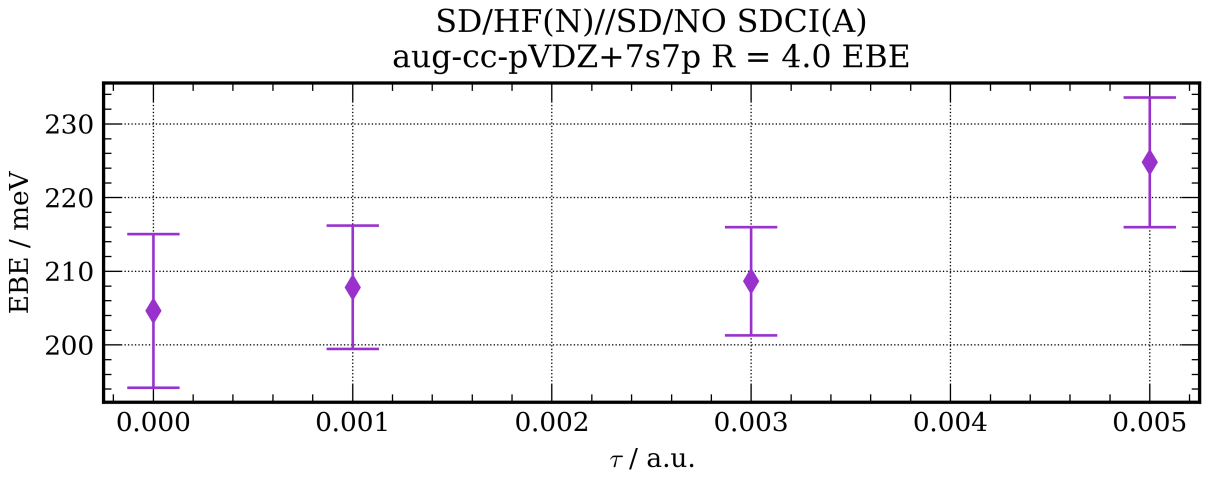

SD/HF(N)
aug-cc-pVDZ+7s7p R = 4.0 neutral
E / a.u.
−68.946
−68.947
−68.948
−68.949
−68.950
−68.951
$E = (-0.772 \pm 0.088)\tau + (-68.9466 \pm 0.00029)$
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

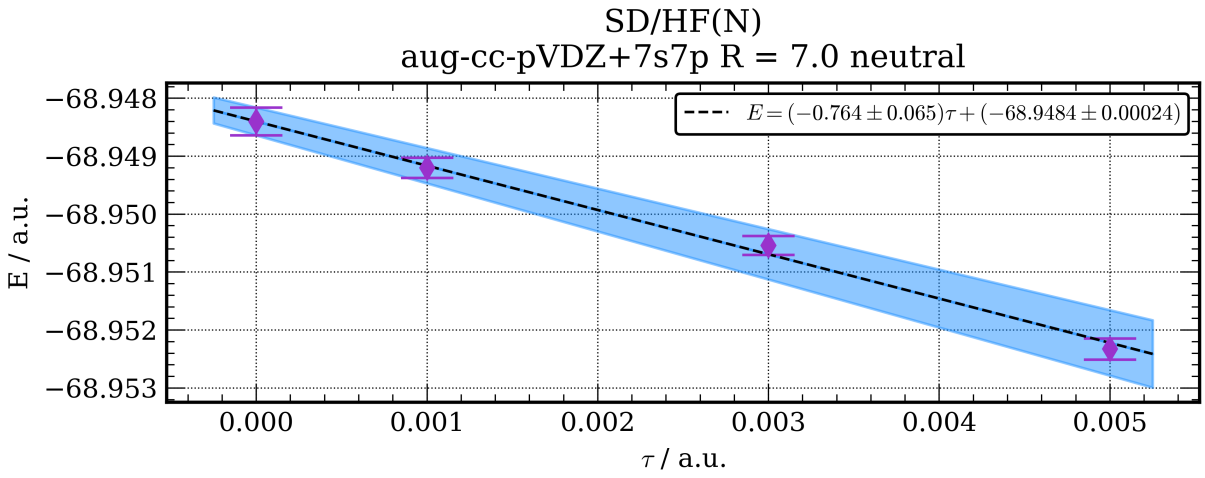

MD/NO SDCI(A)
aug-cc-pVDZ+7s7p R = 4.0 anion
E / a.u.
−68.954
−68.955
−68.956
−68.957
−68.958
−68.959
$E = (-0.756 \pm 0.087)\tau + (-68.95404 \pm 0.00031)$
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

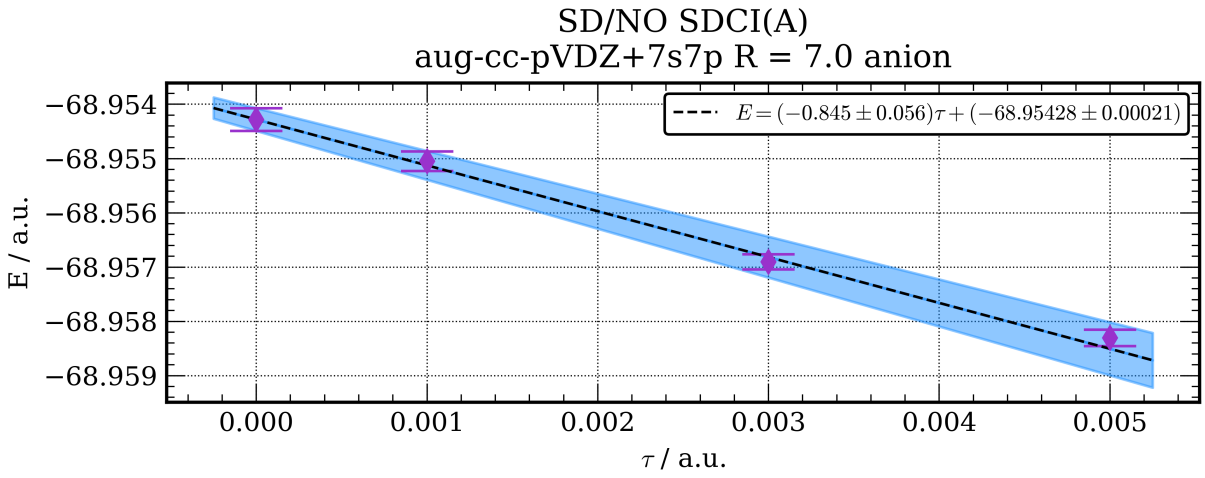

SD/HF(N)//MD/NO SDCI(A)
aug-cc-pVDZ+7s7p R = 4.0 EBE
EBE / meV
210
200
190
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

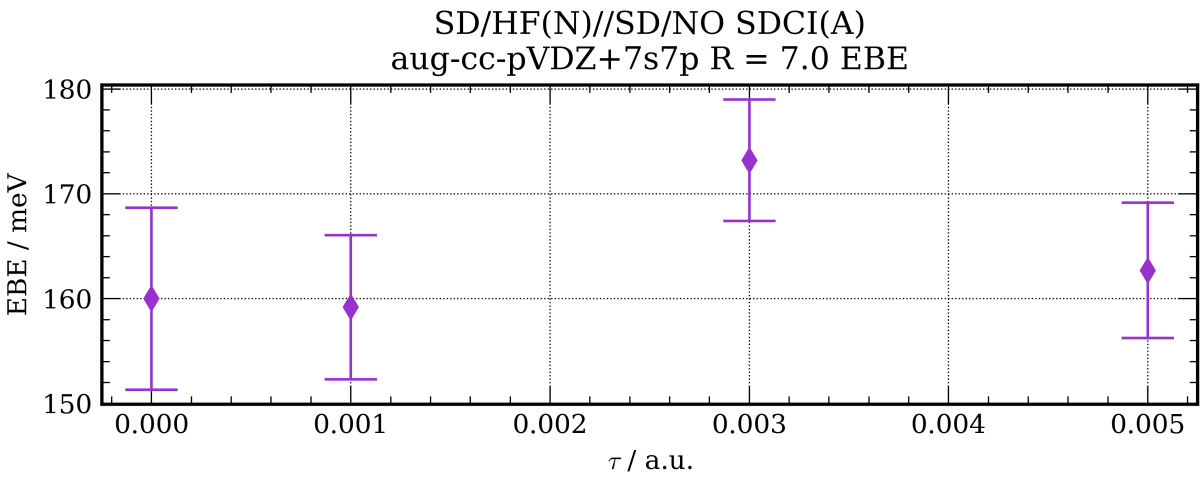

SD/HF(N)
aug-cc-pVDZ+7s7p R = 4.0 neutral
E = (−0.772 ± 0.088)τ + (−68.9466 ± 0.00029)
E / a.u.
−68.946
−68.947
−68.948
−68.949
−68.950
−68.951
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.


SD/NO SDCI(A)
aug-cc-pVDZ+7s7p R = 4.0 anion
E = (−0.904 ± 0.068)τ + (−68.95412 ± 0.00025)
E / a.u.
−68.954
−68.956
−68.958
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.


SD/HF(N)//SD/NO SDCI(A)
aug-cc-pVDZ+7s7p R = 4.0 EBE
EBE / meV
230
220
210
200
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.


SD/HF(N)
aug-cc-pVDZ+7s7p R = 7.0 neutral
E = (−0.764 ± 0.065)τ + (−68.9484 ± 0.00024)
E / a.u.
−68.948
−68.949
−68.950
−68.951
−68.952
−68.953
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.


SD/NO SDCI(A)
aug-cc-pVDZ+7s7p R = 7.0 anion
E = (−0.845 ± 0.056)τ + (−68.95428 ± 0.00021)
E / a.u.
−68.954
−68.955
−68.956
−68.957
−68.958
−68.959
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.


SD/HF(N)//SD/NO SDCI(A)
aug-cc-pVDZ+7s7p R = 7.0 EBE
EBE / meV
180
170
160
150
0.000
0.001
0.002
0.003
0.004
0.005
τ / a.u.

D.4 AFQMC energies

Table 15: AFQMC total energies at used to calculate the electron binding energies. AFQMC calculation details can be found in the manuscript.

R (Å)	neutral energy (Ha)	anion energy (Ha)
4	305.464859 $\pm$ 0.000274	-305.472004 $\pm$ 0.000266
7	-305.432678 $\pm$ 0.000121	-305.439315 $\pm$ 0.000121

D.5 Radial orbital density plots

The radial orbital density, $\psi^2(r)$, plots are created by integrating over angular portion of the norm of the single particle wave function.

$$\psi^2(r) = \int_0^{2\pi} \int_0^{\pi} \psi^2(r,\theta,\phi) r^2 \sin(\theta) d\theta d\phi \tag{84}$$

Discretizing this expression using a uniform radial grid and a Lebedev-Laikov quadrature for the angular components, yields a form that can be readily evaluated.

$$\psi^2(r_i) = 4\pi r_i^2 \sum_j^{N^{ang}} w_j^{ang} \psi^2(r_i,\theta_j,\phi_j) \tag{85}$$

The function $\psi^2(r_i)$ from Equation 85 can be plotted with the points r_i serving as the abscissa. Since the singly occupied orbitals are normalized, the proximity of the sum of the radial quadrature to unity is used as a check.

$$\sum_i^{N^{rad}} \psi^2(r_i) w_i^{rad} = \sum_i^{N^{rad}} \psi^2(r_i) \Delta r \approx 1 \tag{86}$$

D.5.1 Required software versions

Required software	version
`numpy`	1.18.4
`quadpy`	0.16.2
`pyscf`	1.7.0
`cclib`	1.6.3

1. Step 1: Generating a Molden file Molden files were generated using `cclib`, with the exception of the natural orbital from the CIPSI calculations. Since QuantumPackage is not supported by `cclib`, Molden files were created using the native utility in QuantumPackage 2.0. For the Molden files generated with cclib, the `-g/-ghost` flag indicates the presence of a ghost atom. By default the only molecular orbitals can be written to a Molden file, therefore the `-n/-naturalorbtials` flag was created to allow natural orbitals to be

written in place of molecular orbitals. This flag is not yet available in the official distribution, but a request to incorporate it in the official distribution has been opened (https://github.com/cclib/cclib/pull/948).

```
$ ccwrite molden -g "X" -n QUANTUM_CHEMISTRY_OUTPUT_FILE
```

2. Step 2: Integrating over the angular components of the singly occupied orbital

   quadpy was used to generate the Lebedev-Laikov integration weights and points. The singly occupied molecular/natural orbital was evaluated at these points using PySCF.

```
import numpy
import quadpy
import pyscf
import pyscf.tools

filename = "FILENAME.molden"
r_max = 100
num_radial_pts = 1000
mo_idx = 20
# use pyscf to load the molden
mol, mo_energy, mo_coeff, mo_occ, irrep_labels, spins = pyscf.tools.molden.load(
    filename
)
# extract the singly occupied orbital coefficients
singly_occ_orb = mo_coeff[:, mo_idx]
# generate the angular points and weights using quadpy
lebedev_laikov = quadpy.u3.schemes["lebedev_131"]()
angular_pts = lebedev_laikov.theta_phi
angular_weights = lebedev_laikov.weights
num_angular_pts = len(angular_pts[0])
# generate the radial points and weights using numpy
radial_pts = numpy.linspace(r_max, 0, num_radial_pts, endpoint=False)[::-1]
radial_weights = numpy.ones_like(radial_pts) * (radial_pts[1] - radial_pts[0])
```

```
# a helper function to convert radial and angular points to cartesian
def sph2cart(r, theta_phi):
    theta = theta_phi[0]
    phi = theta_phi[1]
    x = r * numpy.cos(theta) * numpy.sin(phi)
    y = r * numpy.sin(theta) * numpy.sin(phi)
    z = r * numpy.cos(phi)
    return numpy.vstack((x, y, z)).T
# integrate over the angular points for each radial point
values = []
for r in radial_pts:
    r_pts = r * numpy.ones(num_angular_pts)
    coords = sph2cart(r_pts, angular_pts)
    ao = mol.eval_gto("GTOval_cart", coords)
    value = angular_weights @ ao @ singly_occ_orb
    values.append(4 * numpy.pi * r * r * value ** 2)
# output the values
values = numpy.array(values)
numpy.savetxt("{}_values.txt".format(filename), values)
numpy.savetxt("{}_r.txt".format(filename), radial_pts)
# check the norm of the orbital
print(radial_weights @ values)
```

Appendix E Supplemental Material for Chapter 6

This appendix includes both a more detailed view of the averaged data presented in the manuscript and presents the data for methods that did not produce viable fragments when paired with the current molecular representations. For the test systems below, the ARI for each molecule in the test set is given. Each plot corresponds to a clustering algorithm, where each molecular representation is included. The data for the water clusters is presented in Section E.1 followed by a the additional information for the methylthiophenes in Section E.2. For the raw data, structures, and production scripts please see at GitHub hosted data at `https://github.com/amandadumi/a1_supporting_info`.

E.1 Water Clusters

For the test set of water clusters, the ARI across the test set are reported for each structure. A more detailed view of agglomerative, spectral, and k-means clustering are provides. Additionally, data is provided for the clustering methods that did not provide chemically-relevant fragments, either inconsistently or not at all.

The additional methods explored which did not produce viable fragments with the current system, either inconsistently or not at all, are also included. Affinity propagation, shown in Figure 33. Mean shift, shown in Figure 34. For affinity propagation, the Cartesian descriptor resulted in a cluster number greater than the number of water monomers for systems that included more than 6 monomers. This resulted in segmented covalent bonds. This likely results from the fact that the difference in the representation between bonded and non-bonded pairs was not sensitive enough. Affinity propagation can be sensitive to the preference of each data-point, which basically describes how likely a given point is to be a cluster center. In this work, we also looked to tuning this preference value by trying a few different options. The typical affinity matrix construction was attempted by placing the sum of distances for a given molecule along the diagonal. A second approach was explored which placed the sum of the covalent-radii determined bond matrix

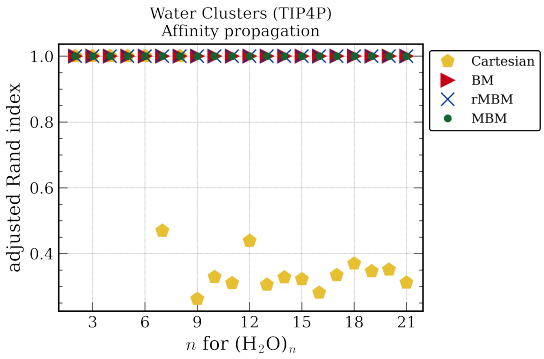


Figure 33: Water cluster test set treated with affinity propagation clustering. All molecular representations are shown and their markings indicated in the legend.

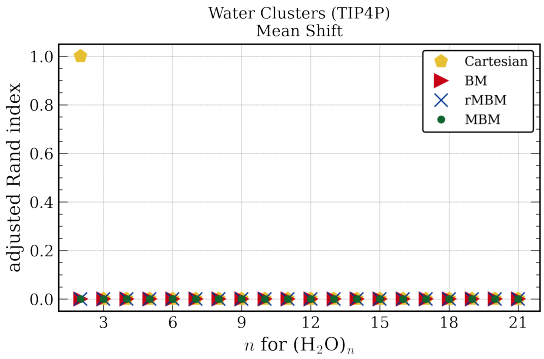


Figure 34: Water cluster test set treated with mean shift clustering. All molecular representations are shown and their markings indicated in the legend. This clustering algorithm with our current descriptor formulation did not produce meaningful fragments.

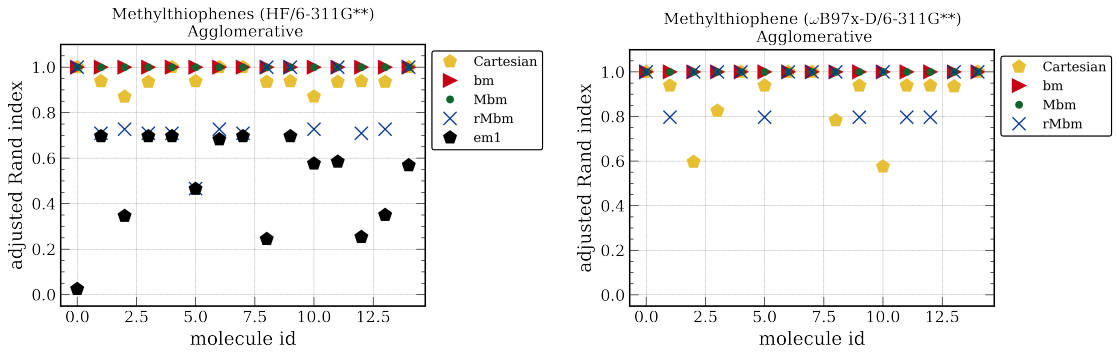


Figure 35: Methylthiophene test set treated with agglomerative clustering. The structures were optimized at either the Hartree-Fock level of theory (left) or ω-B97X-D level of theory (right).

valued for a given atom along the diagonal. We then looked to instead incorporating the Mayer bond matrix of values for a given molecule along the diagonal. The later did slightly improve the results, though not quite enough to give us reliable results. The quantum mechanics-based (QM) descriptors, rMbm and Mbm, when used with with affinity propagation performed well, though this performance was not maintained in the methylthiophene molecular test sets as seen in the next section. The mean shift clustering results are shown in Figure 34. This algorithm did not perform well with the current descriptors as can be seen by the ARI of 0 for all molecules. Again, more clusters than the number of water monomers were chosen.

E.2 Methylthiophenes

The result below are for the methylthiophene test set of molecules at both the HF and ω-B97X-D level of theory. Methods which did not produce molecular fragments in accordance with chemical intuition when used to cluster the current molecular representations are also included.

Agglomerative clustering results are shown in Figure 35, *k*-means clustering results are shown in Figure 36, spectral clustering shown in Figure 37. The effect of the level of theory used to optimize the structure had on the clustering/molecular representation can be seen by comparing

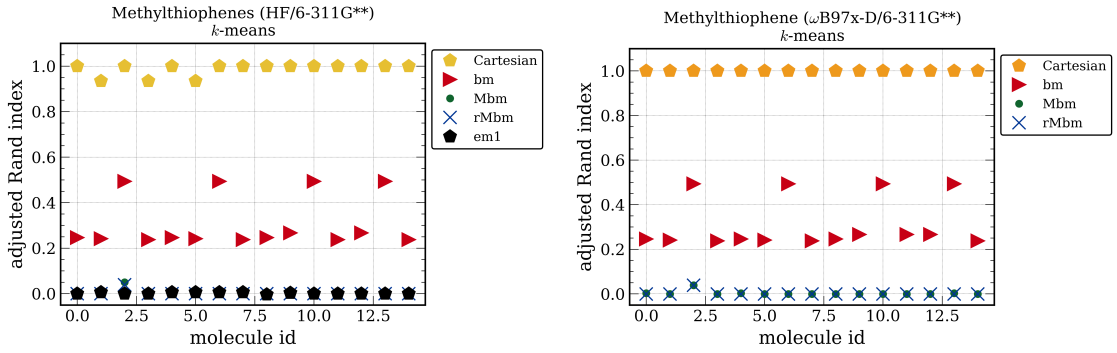


Figure 36: Methylthiophene test set treated with k-means clustering. The structures were optimized at either the Hartree-Fock level of theory (left) or ω-B97X-D level of theory (right).

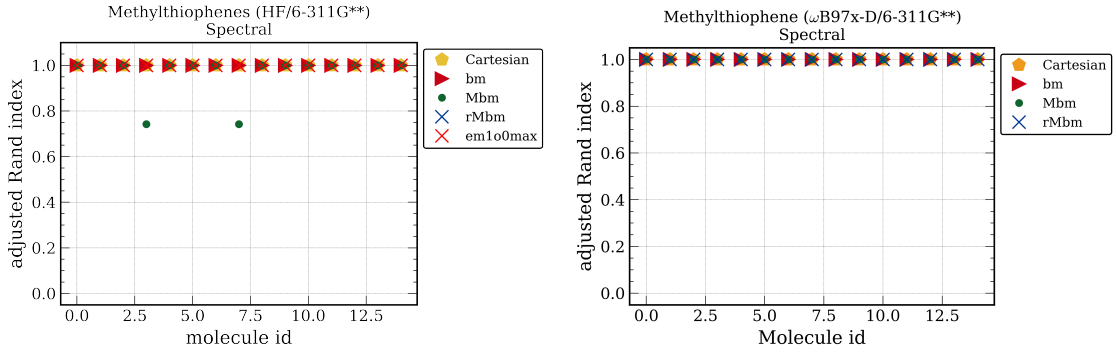


Figure 37: Methylthiophene test set treated with spectral clustering, which worked well for all descriptors. The structures were optimized at either the Hartree-Fock level of theory (left) or ω-B97X-D level of theory (right). All molecular representations are shown and their markings indicated in the legend. Two molecules resulting in different clustering upon the improved level of theory.

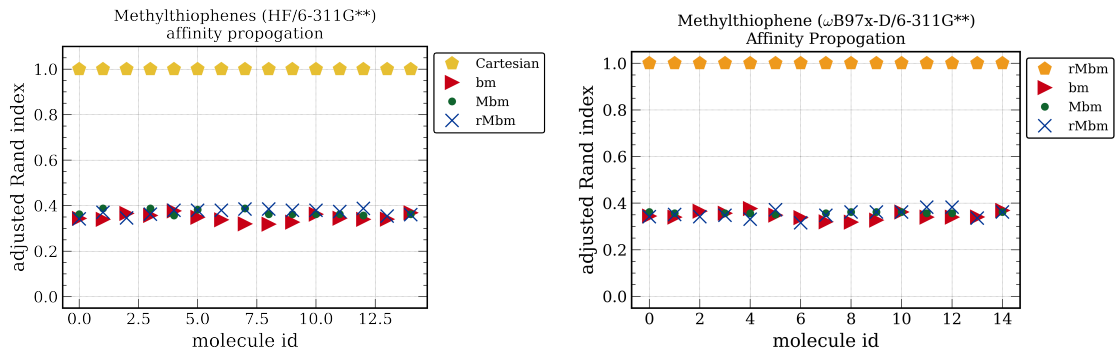


Figure 38: Methylthiophene test set treated with affinity propagation clustering. The structures optimized were at either the Hartree-Fock level of theory (left) or ω-B97X-D level of theory (right). All molecular representations are shown and their markings indicated in the legend.

the right and left image from each figure. Generally, agglomerative clustering was able to perform well for the molecular representation explored, except the rounded Mayer bond matrix which had lost too much detail in the descriptor, making any meaningful groupings difficult for this algorithm. Worth noting is that agglomerative clustering was improved when a better level of theory was used to generate the Cartesian descriptor, while more variation was introduced for the rounded Mayer bond Matrix. The *k*-means clustering with the Cartesian descriptor show a few structures form fragments with slight variations of the correct fragments, while other descriptors show very little or no variation. Generally performance is poor for all but the Cartesian descriptor. Spectral clustering was the most robust, showing a high ARI for all representations. The effect of the level of theory is small, but two molecules in the test set did depend on the higher level of theory to produce the correct clusters.

Results for affinity propagation clustering results are included in Figure 38. Oddly, for affinity propagation, the Cartesian descriptor which performed poorly in the water cluster test set, is the only descriptor that produces reliable fragments for the methylthiophene test set. Although we do see good performance of affinity propagation in both test sets, the reliance on molecular representation warrants further exploration and potentially a more widely applicable descriptor generated for this descriptor. Since a single representation was not viable for both test sets, this method was

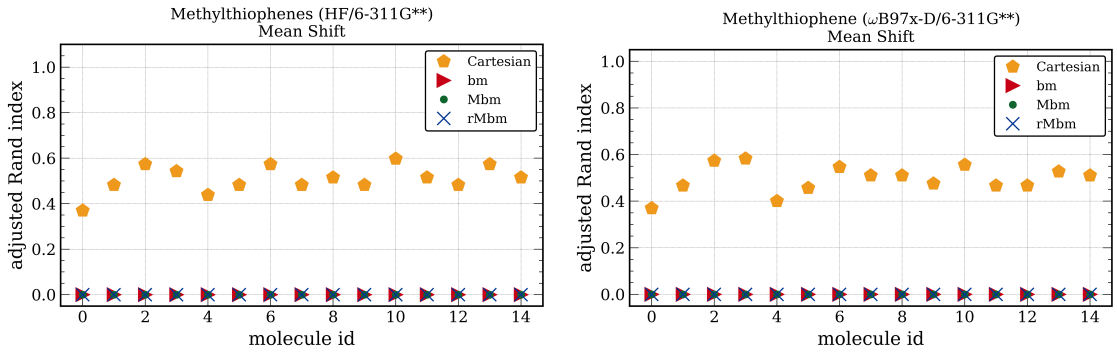


Figure 39: Methylthiophene test set treated with mean shift clustering. The structures were optimized at either the Hartree-Fock level of theory (left) or ω-B97X-D level of theory (right). The explored representations were not formulated for mean shift, and the clustering results were not viable.

not pursued further. Mean shift clustering results are shown in 39. Mean shift clustering on the explored descriptors results in a low ARI value for all representation indicating the current molecular descriptors are not constructed in a way with compliments this clustering method.

E.3 Silyl Ketenes

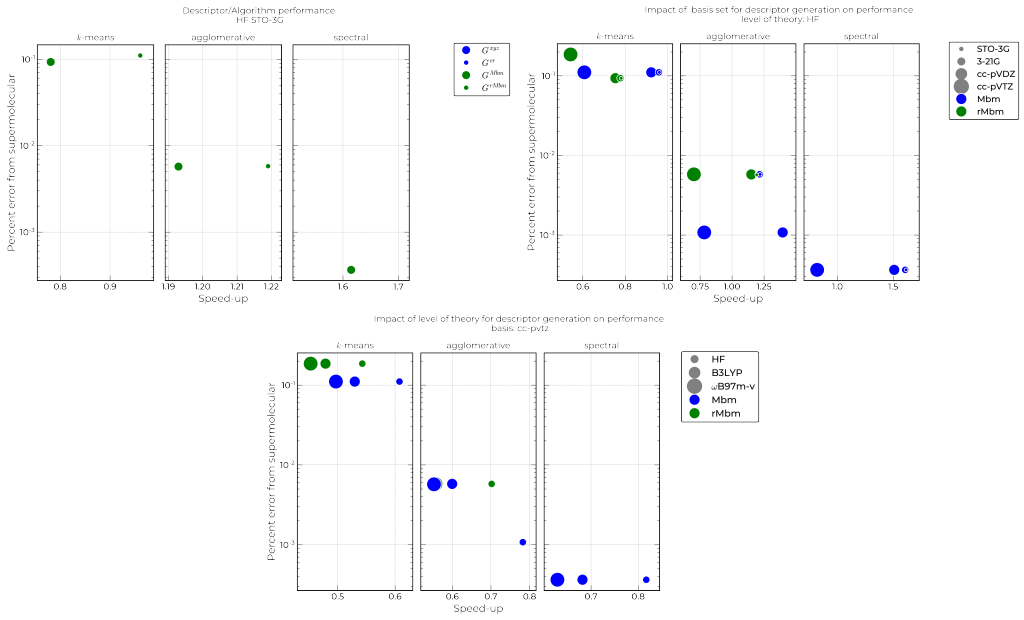


Figure 40: Percent error of the energy and the speedup over the supermolecular calculation for the SK dimer to assess the performance of the fragment approaches and descriptor quality.

Appendix F Bibliography

[1]P. A. M. Dirac and R. H. Fowler, "Quantum mechanics of many-electron systems", Proc. R. Soc. Lond. Ser. Contain. Pap. Math. Phys. Character **123**, 714–733 (1929).

[2]T. J. Hughes, R. A. Shaw, and S. P. Russo, "Computational investigations of dispersion interactions between small molecules and graphene-like flakes", J. Phys. Chem. A **124**, 9552–9561 (2020).

[3]S. Niaz, T. Manzoor, and A. H. Pandith, "Hydrogen storage: Materials, methods and perspectives", Renew. Sustain. Energy Rev **50**, 457–469 (2015).

[4]A. Fomkin, A. Pribylov, I. Men'shchikov, A. Shkolin, O. Aksyutin, A. Ishkov, K. Romanov, and E. Khozina, "Adsorption-based hydrogen storage in activated carbons and model carbon structures", Reactions **2**, 209–226 (2021).

[5]C. Ataca, E. Aktürk, S. Ciraci, and H. Ustunel, "High-capacity hydrogen storage by metallized graphene", Appl. Phys. Lett. **93**, 043123 (2008).

[6]G. K. Dimitrakakis, E. Tylianakis, and G. E. Froudakis, "Pillared graphene: A new 3-D network nanostructure for enhanced hydrogen storage", Nano Lett. **8**, 3166–3170 (2008).

[7]M. Bartolomei, M. I. Hernández, J. Campos-Martínez, R. Hernández-Lamoneda, and G. Giorgi, "Permeation of chemisorbed hydrogen through graphene: A flipping mechanism elucidated", Carbon **178**, 718–727 (2021).

[8]J. Ma, A. Michaelides, and D. Alfè, "Binding of hydrogen on benzene, coronene, and graphene from quantum Monte Carlo calculations", J. Chem. Phys. **134**, 134701 (2011).

[9]J. Ma, A. Michaelides, D. Alfè, L. Schimka, G. Kresse, and E. Wang, "Adsorption and diffusion of water on graphene from first principles", Phys. Rev. B **84**, 033402 (2011).

[10]H. Shin, Y. Luo, A. Benali, and Y. Kwon, "Diffusion Monte Carlo study of O_2 adsorption on single layer graphene", Phys. Rev. B **100**, 075430 (2019).

[11]J. Ahn, I. Hong, G. Lee, H. Shin, A. Benali, and Y. Kwon, "Adsorption of a single Pt atom on graphene: spin crossing between physisorbed triplet and chemisorbed singlet states", Phys. Chem. Chem. Phys. **23**, 22147–22154 (2021).

[12]A. K. Geim, "Graphene: Status and prospects", Science **324**, 1530–1534 (2009).

[13]K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov, "Two-dimensional gas of massless Dirac fermions in graphene", Nature **438**, 197–200 (2005).

[14]O. K. Alekseeva, I. V. Pushkareva, A. S. Pushkarev, and V. N. Fateev, "Graphene and graphene-like materials for hydrogen energy", **15**, 273–300 (2020).

[15]F. Hummel, T. Tsatsoulis, and A. Grüneis, "Low rank factorization of the Coulomb integrals for periodic coupled cluster theory", J. Chem. Phys. **146**, 124105 (2017).

[16]W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, "Quantum monte carlo simulations of solids", Rev. Mod. Phys. **73**, 33–83 (2001).

[17]J. G. Brandenburg, A. Zen, M. Fitzner, B. Ramberger, G. Kresse, T. Tsatsoulis, A. Grüneis, A. Michaelides, and D. Alfè, "Physisorption of water on graphene: Subchemical accuracy from many-body electronic structure methods", J. Phys. Chem. Lett. **10**, 358–368 (2019).

[18]V. K. Voora, A. Kairalapova, T. Sommerfeld, and K. D. Jordan, "Theoretical approaches for treating non-valence correlation-bound anions", J. Chem. Phys. **147**, 214114 (2017).

[19]V. K. Voora and K. D. Jordan, "Nonvalence Correlation-Bound Anion State of C_6F_6: Doorway to Low-Energy Electron Capture", J. Phys. Chem. A **118**, 7201–7205.

[20]T. H. Choi and K. D. Jordan, "Model potential study of non-valence correlation-bound anions of (C_60)$_n$ clusters: the role of electric field-induced charge transfer", Faraday Discuss. **217**, 547–560.

[21]V. K. Voora and K. D. Jordan, "Nonvalence Correlation-Bound Anion States of Polycyclic Aromatic Hydrocarbons", J. Phys. Chem. Lett. **6**, 3994–3997.

[22]T. Sommerfeld, B. Bhattarai, V. P. Vysotskiy, and L. S. Cederbaum, "Correlation-bound anions of NaCl clusters", J. Chem. Phys. **133**, 114301.

[23]V. G. Bezchastnov, V. P. Vysotskiy, and L. S. Cederbaum, "Anions of xenon clusters bound by long-range electron correlations", Phys. Rev. Lett. **107**, 133401.

[24]A. Kairalapova, K. D. Jordan, M. F. Falcetta, D. K. Steiner, B. L. Sutter, and J. S. Gowen, "Prediction of a non-valence temporary anion state of (NaCl)$_2$", J. Phys. Chem. B **123**, 9198–9205 (2019).

[25]A. Kairalapova, K. D. Jordan, D. N. Maienshein, M. C. Fair, and M. F. Falcetta, "Prediction of a nonvalence temporary anion shape resonance for a model $(H_2O)_4$ system", J. Phys. Chem. A **123**, 2719–2726 (2019).

[26]J. P. Rogers, C. S. Anstöter, and J. R. R. Verlet, "Ultrafast dynamics of low-energy electron attachment via a non-valence correlation-bound state", Nat. Chem. **10**, 341–346.

[27]J. P. Rogers, C. S. Anstöter, and J. R. R. Verlet, "Evidence of Electron Capture of an Outgoing Photoelectron Wave by a Nonvalence State in $(C_6F_6)_n^-$", J. Phys. Chem. Lett. **9**, 2504–2509.

[28]J. N. Bull, C. S. Anstöter, and J. R. R. Verlet, "Ultrafast valence to non-valence excited state dynamics in a common anionic chromophore", Nat. Commun. **10**, 5820.

[29]J. R. R. Verlet, C. S. Anstöter, J. N. Bull, and J. P. Rogers, "Role of Nonvalence States in the Ultrafast Dynamics of Isolated Anions", J. Phys. Chem. A **124**, 3507–3519.

[30]V. K. Voora, L. S. Cederbaum, and K. D. Jordan, "Existence of a Correlation Bound s-Type Anion State of C60", J. Phys. Chem. Lett. **4**, 849–853.

[31]V. K. Voora and K. D. Jordan, "Nonvalence Correlation-Bound Anion States of Spherical Fullerenes", Nano Lett. **14**, 4602–4606.

[32]B. Huron, J. P. Malrieu, and P. Rancurel, "Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions", J. Chem. Phys. **58**, 5745–5759 (1973).

[33]Y. Garniron, T. Applencourt, K. Gasperich, A. Benali, A. Ferté, J. Paquier, B. Pradines, R. Assaraf, P. Reinhardt, J. Toulouse, P. Barbaresco, N. Renon, G. David, J. P. Malrieu, M. Véril, M. Caffarel, P. F. Loos, E. Giner, and A. Scemama, "Quantum Package 2.0: An Open-Source Determinant-Driven Suite of Programs", J. Chem. Theory Comput. **15**, 3591–3609 (2019).

[34]M. Dash, S. Moroni, A. Scemama, and C. Filippi, "Perturbatively selected configuration-interaction wave functions for efficient geometry optimization in quantum {M}onte {C}arlo", J. Chem. Theory Comput. **14**, 4176–4182 (2018).

[35]F. Kossoski, Y. Damour, and P.-F. Loos, "Hierarchy configuration interaction: Combining seniority number and excitation degree", J. Phys. Chem. Lett. **13**, 4342–4349 (2022).

[36]M. Dash, S. Moroni, C. Filippi, and A. Scemama, "Tailoring CIPSI expansions for QMC calculations of electronic excitations: The case study of thiophene", J. Chem. Theory Comput. **17**, 3426–3434 (2021).

[37]F. D. Malone, A. Benali, M. A. Morales, M. Caffarel, P. R. C. Kent, and L. Shulenburger, "Systematic comparison and cross-validation of fixed-node diffusion Monte Carlo and phaseless auxiliary-field quantum Monte Carlo in solids", Phys. Rev. B **102**, 161104 (2020).

[38]S. Pathak, B. Busemeyer, J. N. B. Rodrigues, and L. K. Wagner, "Excited states in variational Monte Carlo using a penalty method", J. Chem. Phys. **154**, 034101 (2021).

[39]S. D. Pineda Flores and E. Neuscamman, "Excited state specific multi-slater jastrow wave functions", J. Phys. Chem. A **123**, 1487–1497 (2019).

[40]J. M. Herbert, "Fantasy versus reality in fragment-based quantum chemistry", J. Chem. Phys. **151**, 1–38 (2019).

[41]J. Y. Hasegawa, "Fragment-based configuration interaction wave function to calculate environmental effect on excited states in proteins and solutions", Chem. Phys. Lett. **571**, 77–81 (2013).

[42]S. Müller, "Flexible heuristic algorithm for automatic molecule fragmentation: application to the UNIFAC group contribution model", J. Cheminformatics **11**, 57 (2019).

[43]T. Helgaker, P. Jørgensen, and J. Olsen, "Second quantization", in *Molecular electronic-structure theory* (John Wiley & Sons, Ltd, 2000), pp. 1–33.

[44]P. Hohenberg and W. Kohn, "Inhomogeneous electron gas", Phys. Rev. **136**, B864–B871 (1964).

[45]R. M. Martin, L. Reining, and D. M. Ceperley, *Interacting electrons: Theory and computational approaches* (Cambridge University Press, Cambridge, 2016).

[46]W. Kohn and L. J. Sham, "Self-consistent equations including exchange and correlation effects", Phys. Rev. **140**, A1133–A1138 (1965).

[47]R. Jones, "Density functional theory: Its origins, rise to prominence, and future", Rev. Mod. Phys. **87**, 897–923 (2015).

[48]J. P. Perdew, A. Ruzsinszky, J. Tao, V. N. Staroverov, G. E. Scuseria, and G. I. Csonka, "Prescription for the design and selection of density functional approximations: more constraint satisfaction with fewer fits", J Chem Phys **123**, 62201.

[49]N. Mardirossian and M. Head-Gordon, "Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals", Mol. Phys. **115**, 2315–2372 (2017).

[50]D. M. Ceperley and B. J. Alder, "Ground state of the electron gas by a stochastic method", Phys. Rev. Lett. **45**, 566–569 (1980).

[51]T. Schmidt and S. Küümel, "One- and many-electron self-interaction error in local and global hybrid functionals", Phys. Rev. B **93**, 165120.

[52]S. Śmiga and L. A. Constantin, "Unveiling the physics behind hybrid functionals", J. Phys. Chem. A **124**, 5606–5614.

[53]M. Brütting, H. Bahmann, and S. Kümmel, "Hybrid functionals with local range separation: Accurate atomization energies and reaction barrier heights", J. Chem. Phys. **156**, 104109 (2022).

[54]S. Grimme, S. Ehrlich, and L. Goerigk, "Effect of the damping function in dispersion corrected density functional theory", J. Comput. Chem. **32**, 1456–1465 (2011).

[55]S. Grimme, A. Hansen, J. G. Brandenburg, and C. Bannwarth, "Dispersion-corrected mean-field electronic structure methods", Chem. Rev. **116**, 5105–5154 (2016).

[56]P. Tavadze, R. Boucher, G. Avendaño-Franco, K. X. Kocan, S. Singh, V. Dovale-Farelo, W. Ibarra-Hernández, M. B. Johnson, D. S. Mebane, and A. H. Romero, "Exploring DFT+U parameter space with a bayesian calibration assisted by markov chain monte carlo sampling", npj Computational Materials **7**, 182.

[57]M. Yu, S. Yang, C. Wu, and N. Marom, "Machine learning the hubbard U parameter in DFT+U using bayesian optimization", npj Computational Materials **6**, 180.

[58]N. F. Mott, "Metal-insulator transition", Rev. Mod. Phys. **40**, 677–683.

[59]M. Imada, A. Fujimori, and Y. Tokura, "Metal-insulator transitions", Rev. Mod. Phys. **70**, 1039–1263 (1998).

[60]F. Malet and P. Gori-Giorgi, "Strong Correlation in Kohn-Sham Density Functional Theory", Phys. Rev. Lett. **109**, 246402 (2012).

[61]A. D. Becke, "Density functionals for static, dynamical, and strong correlation", J. Chem. Phys. **138**, 074109 (2013).

[62]A. D. Becke, "Communication: Two-determinant mixing with a strong-correlation density functional", J. Chem. Phys. **139**, 021104 (2013).

[63]B. Himmetoglu, A. Floris, S. de Gironcoli, and M. Cococcioni, "Hubbard-corrected DFT energy functionals: The LDA+U description of correlated systems", Int. J. Quantum Chem. **114**, 14–49 (2014).

[64]S. A. Tolba, K. M. Gameel, B. A. Ali, H. A. Almossalami, and N. K. Allam, "The DFT+U: Approaches, accuracy, and applications", in *Density functional calculations*, edited by G. Yang (IntechOpen, Rijeka, 2018).

[65]R. Tesch and P. M. Kowalski, "Hubbard U parameters for transition metals from first principles", Phys. Rev. B **105**, 195153 (2022).

[66]E. Pavarini, "Solving the strong-correlation problem in materials", Riv. Nuovo Cimento **44**, 597–640 (2021).

[67]R. Requist and E. K. U. Gross, "Model Hamiltonian for strongly correlated systems: Systematic, self-consistent, and unique construction", Phys. Rev. B **99**, 125114.

[68]J. Tilly, P. V. Sriluckshmy, A. Patel, E. Fontana, I. Rungger, E. Grant, R. Anderson, J. Tennyson, and G. H. Booth, "Reduced density matrix sampling: Self-consistent embedding and multiscale electronic structure on current generation quantum computers", Phys. Rev. Research **3**, 033230 (2021).

[69]C. J. C. Scott and G. H. Booth, "Extending density matrix embedding: A static two-particle theory", Phys. Rev. B **104**, 245114.

[70]G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. A. Marianetti, "Electronic structure calculations with dynamical mean-field theory", Rev. Mod. Phys. **78**, 865–951 (2006).

[71]L. Bytautas and K. Ruedenberg, "A priori identification of configurational deadwood", Chem. Phys., Moving Frontiers in Quantum Chemistry: **356**, 64–75 (2009).

[72]J. S. M. Anderson, F. Heidar-Zadeh, and P. W. Ayers, "Breaking the curse of dimension for the electronic Schrödinger equation with functional analysis", Comput. Theor. Chem. **1142**, 66–77 (2018).

[73]C. F. Bender and E. R. Davidson, "Studies in Configuration Interaction: The First-Row Diatomic Hydrides", Phys. Rev. **183**, 23–30 (1969).

[74]A. A. Holmes, N. M. Tubman, and C. J. Umrigar, "Heat-bath configuration interaction: An efficient selected configuration interaction algorithm inspired by heat-bath sampling", J. Chem. Theory Comput. **12**, 3674–3680 (2016).

[75]A. A. Holmes, C. J. Umrigar, and S. Sharma, "Excited states using semistochastic heat-bath configuration interaction", J. Chem. Phys. **147**, 1–11 (2017).

[76]G. H. Booth, A. J. W. Thom, and A. Alavi, "Fermion Monte Carlo without fixed nodes: A game of life, death, and annihilation in Slater determinant space", J. Chem. Phys. **131**, 054106 (2009).

[77]J. B. Schriber and F. A. Evangelista, "Communication: An adaptive configuration interaction approach for strongly correlated electrons with tunable accuracy", J. Chem. Phys. **144**, 161106 (2016).

[78]M. Dash, J. Feldt, S. Moroni, A. Scemama, and C. Filippi, "Excited States with Selected Configuration Interaction-Quantum Monte Carlo: Chemically Accurate Excitation Energies and Geometries", J. Chem. Theory Comput. **15**, 4896–4906 (2019).

[79]A. Benali, K. Gasperich, K. D. Jordan, T. Applencourt, Y. Luo, M. C. Bennett, J. T. Krogel, L. Shulenburger, P. R. C. Kent, P.-F. Loos, A. Scemama, and M. Caffarel, "Toward a systematic improvement of the fixed-node approximation in diffusion Monte Carlo for solids–A case study in diamond", J. Chem. Phys. **153**, 184111 (2020).

[80]A. K. Dutta, M. Saitow, B. Demoulin, F. Neese, and R. Izsák, "A domain-based local pair natural orbital implementation of the equation of motion coupled cluster method for electron attached states", J. Chem. Phys. **150**, 164123.

[81]M. Musiał, "Equation-of-motion coupled-cluster models", in *Quantum chemistry and dynamics of excited states* (John Wiley & Sons, Ltd, 2020), pp. 77–108.

[82]E. A. Hylleraas, "Neue Berechnung der Energie des Heliums im Grundzustande, sowie des tiefsten Terms von Ortho-Helium", Z. Physik **54**, 347–366 (1929).

[83]A. Bijl, "The lowest wave function of the symmetrical many particles system", Physica **7**, 869–886 (1940).

[84]R. Dingle, "LI. The zero-point energy of a system of particles", Lond. Edinb. Dublin Philos. Mag. J. Sci. **40**, 573–578 (1949).

[85]R. Jastrow, "Many-body problem with strong forces", Phys. Rev. **98**, 1479–1484.

[86]K. Esler, J. Kim, D. Ceperley, and L. Shulenburger, "Accelerating quantum monte carlo simulations of real materials on GPU clusters", Comput. Sci. Eng. **14**, 40–51 (2012).

[87]V. Natoli and D. M. Ceperley, "An optimized method for treating long-range potentials", Journal of Computational Physics **117**, 171–178.

[88]E. Pavarini, E. Koch, and S. Zhang, eds., *Many-Body Methods for Real Materials*, Vol. 9, Schriften Des Forschungszentrums Jülich. Modeling and Simulation (Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag, Jülich, Sept. 2019).

[89]H. F. Trotter, "On the Product of Semi-Groups of Operators", Proc. Am. Math. Soc. **10**, 545–551 (1959).

[90]E. Nelson, "Feynman integrals and the schrödinger equation", J. Math. Phys. **5**, 332–343 (1964).

[91]M. Suzuki, "Improved Trotter-like formula", Physics Letters A **180**, 232–234 (1993).

[92]M. Holzmann, R. C. Clay, M. A. Morales, N. M. Tubman, D. M. Ceperley, and C. Pierleoni, "Theory of finite size effects for electronic quantum Monte Carlo calculations of liquids and solids", Phys. Rev. B **94**, 035126 (2016).

[93]C. Lin, F. H. Zong, and D. M. Ceperley, "Twist-averaged boundary conditions in continuum quantum Monte Carlo algorithms", Phys. Rev. E **64**, 016702 (2001).

[94]W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal, "Quantum Monte Carlo simulations of solids", Rev. Mod. Phys. **73**, 1–51 (2001).

[95]P. P. Ewald, "Die Berechnung optischer und elektrostatischer Gitterpotentiale", Annalen der Physik **369**, 253–287 (1921).

[96]S. Chiesa, D. M. Ceperley, R. M. Martin, and M. Holzmann, "Finite-size error in many-body simulations with long-range interactions", Phys. Rev. Lett. **97**, 076404 (2006).

[97]J. Zinn-Justin, *Quantum field theory and critical phenomena 5e*, 5th ed., International Monographs on Physics (Oxford University Press, New York, 2021).

[98]K. Binder and D. W. Heermann, *Monte Carlo simulation in statistical physics: an introduction*, 5th ed, Graduate Texts in Physics (Springer, Heidelberg ; New York, 2010), 200 pp.

[99]B. Jeziorski, R. Moszynski, and K. Szalewicz, "Perturbation Theory Approach to Intermolecular Potential Energy Surfaces of van der Waals Complexes", Chem. Rev. **94**, 1887–1930 (1994).

[100]K. Szalewicz, "Symmetry-adapted perturbation theory of intermolecular forces", WIREs Comput. Mol. Sci. **2**, 254–272 (2012).

[101]D. S. Lambrecht, "Generalizing energy decomposition analysis to response properties to inform expedited predictive models", Comput. Theor. Chem. **1149**, 24–30 (2019).

[102]E. J. Berquist and D. S. Lambrecht, "A first principles approach for partitioning linear response properties into additive and cooperative contributions", ChemRXiv **10.26434/c**, 1–25 (2018).

[103]R. Z. Khaliullin, E. A. Cobar, R. C. Lochan, A. T. Bell, and M. Head-Gordon, "Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals", J. Phys. Chem. A **111**, 8753–8765 (2007).

[104]Y. Kawashima, E. Lloyd, M. P. Coons, Y. Nam, S. Matsuura, A. J. Garza, S. Johri, L. Huntington, V. Senicourt, A. O. Maksymov, J. H. V. Nguyen, J. Kim, N. Alidoust, A. Zaribafiyan, and T. Yamazaki, "Optimizing electronic structure simulations on a trapped-ion quantum computer using problem decomposition", Commun. Phys. **4**, 245 (2021).

[105]M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko, "Fragmentation Methods: A Route to Accurate Calculations on Large Systems", Chem. Rev. **112**, 632–672 (2012).

[106]L. W. Chung, W. M. C. Sameera, R. Ramozzi, A. J. Page, M. Hatanaka, G. P. Petrova, T. V. Harris, X. Li, Z. Ke, F. Liu, H.-B. Li, L. Ding, and K. Morokuma, "The ONIOM Method and Its Applications", Chem. Rev. **115**, 5678–5796 (2015).

[107]T. Vreven, K. S. Byun, I. Komáromi, S. Dapprich, J. A. J. Montgomery, K. Morokuma, and M. J. Frisch, "Combining Quantum Mechanics Methods with Molecular Mechanics Methods in ONIOM", J. Chem. Theory Comput. **2**, 815–826 (2006).

[108]L. D. Jacobson, R. M. Richard, K. U. Lao, and J. M. Herbert, "Chapter Two - Efficient Monomer-Based Quantum Chemistry Methods for Molecular and Ionic Clusters", in , Vol. 9, edited by R. A. Wheeler, Annual Reports in Computational Chemistry (Elsevier, 2013), pp. 25–58.

[109]N. J. Mayhall and K. Raghavachari, "Molecules-in-Molecules: An Extrapolated Fragment-Based Approach for Accurate Calculations on Large Molecules and Materials", J. Chem. Theory Comput. **7**, 1336–1343 (2011).

[110]K. V. J. Jose and K. Raghavachari, "Evaluation of Energy Gradients and Infrared Vibrational Spectra through Molecules-in-Molecules Fragment-Based Approach", J. Chem. Theory Comput. **11**, 950–961 (2015).

[111]A. Saha and K. Raghavachari, "Analysis of Different Fragmentation Strategies on a Variety of Large Peptides: Implementation of a Low Level of Theory in Fragment-Based Methods Can Be a Crucial Factor", J. Chem. Theory Comput. **11**, 2012–2023 (2015).

[112]R. M. Ferullo, N. F. Domancich, and N. J. Castellani, "On the performance of van der Waals corrected-density functional theory in describing the atomic hydrogen physisorption on graphite", **500**, 283–286 (2010).

[113]X. Sha and B. Jackson, "First-principles study of the structural and energetic properties of H atoms on a graphite (0001) surface", Surf. Sci. **496**, 318–330 (2002).

[114]L. Jeloaica and V. Sidis, "DFT investigation of the adsorption of atomic hydrogen on a cluster-model graphite surface", Chem. Phys. Lett. **300**, 157–162 (1999).

[115]S. Sahu and G. C. Rout, "Band gap opening in graphene: a short theoretical study", Int. Nano Lett. **7**, 81–89 (2017).

[116]H. González-Herrero, J. M. Gómez-Rodríguez, P. Mallet, M. Moaied, J. J. Palacios, C. Salgado, M. M. Ugeda, J.-Y. Veuillen, F. Yndurain, and I. Brihuega, "Atomic-scale control of graphene magnetism by using hydrogen atoms", Science **352**, 437–441 (2016).

[117]H. González-Herrero, E. C.-d. Río, P. Mallet, J.-Y. Veuillen, J. J. Palacios, J. M. Gómez-Rodríguez, I. Brihuega, and F. Ynduráin, "Hydrogen physisorption channel on graphene: a highway for atomic H diffusion", 2d Mater. **6**, 021004 (2019).

[118]T. N. Mihm, A. R. McIsaac, and J. J. Shepherd, "An optimized twist angle to find the twist-averaged correlation energy applied to the uniform electron gas", J. Chem. Phys. **150**, 191101 (2019).

[119]J. M. Callahan, M. F. Lange, and T. C. Berkelbach, "Dynamical correlation energy of metals in large basis sets from downfolding and composite approaches", J. Chem. Phys. **154**, 211105 (2021).

[120]G. H. Booth, A. Grüneis, G. Kresse, and A. Alavi, "Towards an exact description of electronic wavefunctions in real solids", Nature **493**, 365–370 (2013).

[121]T. Schäfer, F. Libisch, G. Kresse, and A. Grüneis, "Local embedding of coupled cluster theory into the random phase approximation using plane waves", J. Chem. Phys. **154**, 011101 (2021).

[122]T. Schäfer, B. Ramberger, and G. Kresse, "Quartic scaling MP2 for solids: A highly parallelized algorithm in the plane wave basis", J. Chem. Phys. **146**, 104101 (2017).

[123]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple", Phys. Rev. Lett. **77**, 3865–3868 (1996).

[124]J. P. Perdew, "Unified theory of exchange and correlation beyond the local density approximation", in Electron. Struct. Solids 91, Vol. 17, edited by P. Ziesche and H. Eschrig, Physical Research (1991), pp. 11–20.

[125]T. Roman, W. A. Diño, H. Nakanishi, H. Kasai, T. Sugimoto, and K. Tange, "Hydrogen pairing on graphene", Carbon **45**, 218–220 (2007).

[126]Ž. Šljivančanin, E. Rauls, L. Hornekær, W. Xu, F. Besenbacher, and B. Hammer, "Extended atomic hydrogen dimer configurations on the graphite(0001) surface", J. Chem. Phys. **131**, 084706 (2009).

[127]P. O. Lehtinen, A. S. Foster, Y. Ma, A. V. Krasheninnikov, and R. M. Nieminen, "Irradiation-induced magnetism in graphite: A density functional study", Phys. Rev. Lett. **93**, 187202 (2004).

[128]Y. Lin, F. Ding, and B. I. Yakobson, "Hydrogen storage by spillover on graphene as a phase nucleation process", Phys. Rev. B **78**, 041402 (2008).

[129]S. Casolo, O. M. Løvvik, R. Martinazzo, and G. F. Tantardini, "Understanding adsorption of hydrogen atoms on graphene", J. Chem. Phys. **130**, 054704 (2009).

[130]R. H. Miwa, T. B. Martins, and A. Fazzio, "Hydrogen adsorption on boron doped graphene: an ab initio study", Nanotechnology **19**, 155708 (2008).

[131]A. Ishii, M. Yamamoto, H. Asano, and K. Fujiwara, "DFT calculation for adatom adsorption on graphene sheet as a prototype of carbon nanotube functionalization", J. Phys. Conf. Ser. **100**, 052087 (2008).

[132]W. Li, M. Zhao, T. He, C. Song, X. Lin, X. Liu, Y. Xia, and L. Mei, "Concentration dependent magnetism induced by hydrogen adsorption on graphene and single walled carbon nanotubes", J. Magn. Magn. Mater. **322**, 838–843 (2010).

[133]D. W. Boukhvalov, M. I. Katsnelson, and A. I. Lichtenstein, "Hydrogen on graphene: Electronic structure, total energy, structural distortions and magnetism from first-principles calculations", Phys. Rev. B **77**, 035427 (2008).

[134]M. A. Kim, D. C. Sorescu, S. Amemiya, K. D. Jordan, and H. Liu, "Real-time modulation of hydrogen evolution activity of graphene electrodes using mechanical strain", ACS Appl. Mater. Interfaces **14**, 10691–10700 (2022).

[135]S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu", J. Chem. Phys. **132**, 154104 (2010).

[136]A. Annaberdiyev, G. Wang, C. A. Melton, M. C. Bennett, L. Shulenburger, and L. Mitas, "A new generation of effective core potentials from correlated calculations: 3d transition metal series", J. Chem. Phys. **149**, 134108.

[137]M. C. Bennett, C. A. Melton, A. Annaberdiyev, G. Wang, L. Shulenburger, and L. Mitas, "A new generation of effective core potentials for correlated calculations", J. Chem. Phys. **147**, 224106.

[138]H. J. Monkhorst and J. D. Pack, "Special points for Brillouin-zone integrations", Phys. Rev. B **13**, 5188–5192 (1976).

[139]N. Marzari, D. Vanderbilt, A. De Vita, and M. C. Payne, "Thermal contraction and disordering of the {A}l(110) surface", Phys. Rev. Lett. **82**, 3296–3299 (1999).

[140]C. Adamo and V. Barone, "Toward reliable density functional methods without adjustable parameters: The PBE0 model", J. Chem. Phys. **110**, 6158–6170 (1999).

[141]A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, "Influence of the exchange screening parameter on the performance of screened hybrid functionals", J. Chem. Phys. **125**, 224106 (2006).

[142]M. F. Peintinger, D. V. Oliveira, and T. Bredow, "Consistent Gaussian basis sets of triple-zeta valence with polarization quality for solid-state calculations", J. Comput. Chem. **34**, 451–459 (2012).

[143]H. Kruse and S. Grimme, "A geometrical correction for the inter- and intra-molecular basis set superposition error in Hartree-Fock and density functional theory calculations for large systems", J. Chem. Phys. **136**, 154101 (2012).

[144]J. G. Brandenburg, M. Alessio, B. Civalleri, M. F. Peintinger, T. Bredow, and S. Grimme, "Geometrical correction for the inter- and intramolecular basis set superposition error in periodic density functional theory calculations", J. Phys. Chem. A **117**, 9282–9292 (2013).

[145]P. Giannozzi, O. Baseggio, P. Bonfà, D. Brunato, R. Car, I. Carnimeo, C. Cavazzoni, S. de Gironcoli, P. Delugas, F. Ferrari Ruffino, A. Ferretti, N. Marzari, I. Timrov, A. Urru, and S. Baroni, "Quantum ESPRESSO toward the exascale", J. Chem. Phys. **152**, 154105 (2020).

[146]P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, "QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials", J Phys Condens Matter **21**, 395502 (19pp) (2009).

[147]P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. D. Jr, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. Otero-de-la-Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, "Advanced capabilities for materials modelling with QUANTUM ESPRESSO", J. Phys.: Condens. Matter **29**, 465901 (2017).

[148]R. Dovesi, A. Erba, R. Orlando, C. M. Zicovich-Wilson, B. Civalleri, L. Maschio, M. Rérat, S. Casassa, J. Baima, S. Salustro, and B. Kirtman, "Quantum-mechanical condensed matter simulations with CRYSTAL", Wiley Interdiscip. Rev. Comput. Mol. Sci **8**, e1360 (2018).

[149]R. Dovesi, F. Pascale, B. Civalleri, K. Doll, N. M. Harrison, I. Bush, P. D'Arco, Y. Noël, M. Rérat, P. Carbonnière, M. Causà, S. Salustro, V. Lacivita, B. Kirtman, A. M. Ferrari, F. S. Gentile, J. Baima, M. Ferrero, R. Demichelis, and M. De La Pierre, "The CRYSTAL code, 1976–2020 and beyond, a long story", J. Chem. Phys. **152**, 204111 (2020).

[150]E. Aprà, E. J. Bylaska, W. A. de Jong, N. Govind, K. Kowalski, T. P. Straatsma, M. Valiev, H. J. J. van Dam, Y. Alexeev, J. Anchell, V. Anisimov, F. W. Aquino, R. Atta-Fynn, J. Autschbach, N. P. Bauman, J. C. Becca, D. E. Bernholdt, K. Bhaskaran-Nair, S. Bogatko, P. Borowski, J. Boschen, J. Brabec, A. Bruner, E. Cauët, Y. Chen, G. N. Chuev, C. J. Cramer, J. Daily, M. J. O. Deegan, T. H. Dunning, M. Dupuis, K. G. Dyall, G. I. Fann, S. A. Fischer, A. Fonari, H. Fr uchtl, L. Gagliardi, J. Garza, N. Gawande, S. Ghosh, K. Glaesemann, A. W. Götz, J. Hammond, V. Helms, E. D. Hermes, K. Hirao, S. Hirata, M. Jacquelin, L. Jensen, B. G. Johnson, H. Jónsson, R. A. Kendall, M. Klemm, R. Kobayashi, V. Konkov, S. Krishnamoor-

thy, M. Krishnan, Z. Lin, R. D. Lins, R. J. Littlefield, A. J. Logsdail, K. Lopata, W. Ma, A. V. Marenich, J. Martin del Campo, D. Mejia-Rodriguez, J. E. Moore, J. M. Mullin, T. Nakajima, D. R. Nascimento, J. A. Nichols, P. J. Nichols, J. Nieplocha, A. Otero-de-la-Roza, B. Palmer, A. Panyala, T. Pirojsirikul, B. Peng, R. Peverati, J. Pittner, L. Pollack, R. M. Richard, P. Sadayappan, G. C. Schatz, W. A. Shelton, D. W. Silverstein, D. M. A. Smith, T. A. Soares, D. Song, M. Swart, H. L. Taylor, G. S. Thomas, V. Tipparaju, D. G. Truhlar, K. Tsemekhman, T. Van Voorhis, Á. Vázquez-Mayagoitia, P. Verma, O. Villa, A. Vishnu, K. D. Vogiatzis, D. Wang, J. H. Weare, M. J. Williamson, T. L. Windus, K. Woliński, A. T. Wong, Q. Wu, C. Yang, Q. Yu, M. Zacharias, Z. Zhang, Y. Zhao, and R. J. Harrison, "NWChem: Past, present, and future", J. Chem. Phys. **152**, 184102 (2020).

[151]J. B. Anderson, "Quantum chemistry by random walk: Higher accuracy", J. Chem. Phys. **73**, 3897–3899 (1980).

[152]K. Esler, J. Kim, D. Ceperley, and L. Shulenburger, "Accelerating quantum monte carlo simulations of real materials on GPU clusters", Comput. Sci. Eng. **14**, 40–51 (2012).

[153]N. D. Drummond, M. D. Towler, and R. J. Needs, "Jastrow correlation factor for atoms, molecules, and solids", Phys. Rev. B **70**, 235119 (2004).

[154]C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig, "Alleviation of the fermion-sign problem by optimization of many-body wave functions", Phys. Rev. Lett. **98**, 110201 (2007).

[155]A. Zen, J. G. Brandenburg, A. Michaelides, and D. Alfè, "A new scheme for fixed node diffusion quantum Monte Carlo with pseudopotentials: Improving reproducibility and reducing the trial-wave-function bias", J. Chem. Phys. **151**, 134105 (2019).

[156]M. Casula, S. Moroni, S. Sorella, and C. Filippi, "Size-consistent variational approaches to nonlocal pseudopotentials: Standard and lattice regularized diffusion Monte Carlo methods revisited", J. Chem. Phys. **132**, 154113 (2010).

[157]A. Zen, S. Sorella, M. J. Gillan, A. Michaelides, and D. Alfè, "Boosting the accuracy and speed of quantum Monte Carlo: Size consistency and time step", Phys. Rev. B **93**, 241118 (2016).

[158]D. Cortés-Arriagada, S. Gutiérrez-Oliva, B. Herrera, K. Soto, and A. Toro-Labbé, "The mechanism of chemisorption of hydrogen atom on graphene: Insights from the reaction force and reaction electronic flux", J. Chem. Phys. **141**, 134701 (2014).

[159]H. Jiang, M. Kammler, F. Ding, Y. Dorenkamp, F. R. Manby, A. M. Wodtke, T. F. Miller, A. Kandratsenka, and O. Bünermann, "Imaging covalent bond formation by H atom scattering from graphene", Science **364**, 379–382 (2019).

[160]J. Kim, A. D. Baczewski, T. D. Beaudet, A. Benali, M. C. Bennett, M. A. Berrill, N. S. Blunt, E. J. L. Borda, M. Casula, D. M. Ceperley, S. Chiesa, B. K. Clark, R. C. Clay, K. T. Delaney, M. Dewing, K. P. Esler, H. Hao, O. Heinonen, P. R. C. Kent, J. T. Krogel, I. Kylänpää, Y. W. Li, M. G. Lopez, Y. Luo, F. D. Malone, R. M. Martin, A. Mathuriya, J. McMinis, C. A. Melton, L. Mitas, M. A. Morales, E. Neuscamman, W. D. Parker, S. D. P. Flores, N. A. Romero, B. M. Rubenstein, J. A. R. Shea, H. Shin, L. Shulenburger, A. F. Tillack, J. P. Townsend, N. M. Tubman, B. V. D. Goetz, J. E. Vincent, D. C. Yang, Y. Yang, S. Zhang, and L. Zhao, "QMCPACK: an open source ab initio quantum Monte Carlo package for the electronic structure of atoms, molecules and solids", J. Phys.: Condens. Matter **30**, 195901.

[161]P. R. C. Kent, A. Annaberdiyev, A. Benali, M. C. Bennett, E. J. Landinez Borda, P. Doak, H. Hao, K. D. Jordan, J. T. Krogel, I. Kylänpää, J. Lee, Y. Luo, F. D. Malone, C. A. Melton, L. Mitas, M. A. Morales, E. Neuscamman, F. A. Reboredo, B. Rubenstein, K. Saritas, S. Upadhyay, G. Wang, S. Zhang, and L. Zhao, "QMCPACK: Advances in the development, efficiency, and application of auxiliary field and real-space variational and diffusion quantum Monte Carlo", J. Chem. Phys. **152**, 174105 (2020).

[162]J. T. Krogel, "Nexus: A modular workflow management system for quantum simulation codes", Comput. Phys. Commun. **198**, 154–168 (2016).

[163]J. D. Hunter, "Matplotlib: A 2D graphics environment", Comput. Sci. Eng. **9**, 90–95 (2007).

[164]K. Momma and F. Izumi, "*VESTA3* for three-dimensional visualization of crystal, volumetric and morphology data", J. Appl. Crystallogr. **44**, 1272–1276 (2011).

[165]C. Moller and M. S. Plesset, "Note on an approximation treatment for many-electron systems", Phys. Rev. **46**, 618–622.

[166]K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, "A fifth-order perturbation comparison of electron correlation theories", Chem. Phys. Lett. **157**, 479–483 (1989).

[167]R. C. Lochan and M. Head-Gordon, "Orbital-optimized opposite-spin scaled second-order correlation: An economical method to improve the description of open-shell molecules", J. Chem. Phys. **126**, 164101 (2007).

[168]J. Čížek, J. Paldus, and L. Šroubková, "Cluster expansion analysis for delocalized systems", Int. J. Quantum Chem. **3**, 149–167 (1969).

[169]D. C. Comeau and R. J. Bartlett, "The equation-of-motion coupled-cluster method. Applications to open- and closed-shell reference states", Chem. Phys. Lett. **207**, 414–423 (1993).

[170]R. C. Grimm and R. G. Storer, "Monte-Carlo solution of Schrödinger's equation", J. Comput. Phys. **7**, 134–156 (1971).

[171]J. B. Anderson, "A random-walk simulation of the Schrödinger equation: H_3^+", J. Chem. Phys. **63**, 1499–1503.

[172]J. B. Anderson, "Quantum chemistry by random walk. H ^{2}P, H3+ D_{3h} ^{1}A$'_1$, H2 $^3\Sigma^+_u$, H4 $^1\Sigma^+_g$, Be ^{1}S", J. Chem. Phys. **65**, 4121–4127 (1976).

[173]P. López Ríos, A. Ma, N. D. Drummond, M. D. Towler, and R. J. Needs, "Inhomogeneous backflow transformations in quantum Monte Carlo calculations", Phys. Rev. E **74**, 066701.

[174]R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, "Monte Carlo calculations of coupled boson-fermion systems. I", Phys. Rev. D **24**, 2278–2286 (1981).

[175]G. Sugiyama and S. Koonin, "Auxiliary field Monte-Carlo for quantum many-body ground states", Ann. Phys. **168**, 1–26 (1986).

[176]D. J. Scalapino and R. L. Sugar, "Monte Carlo calculations of coupled boson-fermion systems. II", Phys. Rev. B **24**, 4295–4308 (1981).

[177]S. Zhang and H. Krakauer, "Quantum Monte Carlo Method using Phase-Free Random Walks with Slater Determinants", Phys. Rev. Lett. **90**, 136401.

[178]S. Zhang, J. Carlson, and J. E. Gubernatis, "Constrained path Monte Carlo method for fermion ground states", Phys. Rev. B **55**, 7464–7477.

[179]M. Motta and S. Zhang, "Ab initio computations of molecular systems by the auxiliary-field quantum Monte Carlo method", Wiley Interdiscip Rev Comput Mol Sci **8**, e1364 (2018).

[180]S. Zhang, "Ab initio electronic structure calculations by auxiliary-field quantum monte carlo", in *Handbook of materials modeling : Methods: Theory and modeling* (Springer International Publishing, 2018), pp. 1–27.

[181]W. Humphrey, A. Dalke, and K. Schulten, ""VMD: Visual molecular dynamics"", J. Mol. Graphics **14**, 33–38 (1996).

[182]J. F. Stanton and J. Gauss, "Perturbative treatment of the similarity transformed Hamiltonian in equation-of-motion coupled-cluster approximations", J. Chem. Phys. **103**, 1064–1076 (1995).

[183]D. A. Matthews and J. F. Stanton, "A new approach to approximate equation-of-motion coupled cluster with triple excitations", J. Chem. Phys. **145**, 124102 (2016).

[184]S. A. Kucharski, M. Włoch, M. Musiał, and R. J. Bartlett, "Coupled-cluster theory for excited electronic states: The full equation-of-motion coupled-cluster single, double, and triple excitation method", J. Chem. Phys. **115**, 8263–8266 (2001).

[185]K. Kowalski and P. Piecuch, "The active-space equation-of-motion coupled-cluster methods for excited electronic states: Full EOMCCSDt", J. Chem. Phys. **115**, 643–651 (2001).

[186]T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen", J. Chem. Phys. **90**, 10.1063/1.456153 (1989).

[187]R. A. Kendall, T. H. Dunning, and R. J. Harrison, "Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions", J. Chem. Phys. **96**, 10.1063/1.462569 (1992).

[188]J. F. Stanton, J. Gauss, L. Cheng, M. E. Harding, D. A. Matthews, and P. G. Szalay, *CFOUR, coupled-cluster techniques for computational chemistry, a quantum-chemical program package*.

[189]D. A. Matthews, L. Cheng, M. E. Harding, F. Lipparini, S. Stopkowicz, T.-C. Jagau, P. G. Szalay, J. Gauss, and J. F. Stanton, "Coupled-cluster techniques for computational chemistry: The CFOUR program package", J. Chem. Phys. **152**, 214108 (2020).

[190]R. Jastrow, "Many-body problem with strong forces", Phys. Rev. **98**, 1479–1484.

[191]N. D. Drummond, M. D. Towler, and R. J. Needs, "Jastrow correlation factor for atoms, molecules, and solids", Phys. Rev. B **70**, 235119.

[192]A. D. Becke, "Density-functional exchange-energy approximation with correct asymptotic behavior", Phys. Rev. A **38**, 3098–3100.

[193]C. Lee, W. Yang, and R. G. Parr, "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density", Phys. Rev. B **37**, 785–789.

[194]S. H. Vosko, L. Wilk, and M. Nusair, "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis", Can. J. Phys. **58**, 1200–1211 (1980).

[195]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, and M. J. Frisch, "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields", J. Phys. Chem. **98**, 11623–11627 (1994).

[196]M. Casula, "Beyond the locality approximation in the standard diffusion Monte Carlo method", Phys. Rev. B **74**, 161102 (2006).

[197]M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, and J. A. Montgomery, "General atomic and molecular electronic structure system", J. Comput. Chem. **14**, 1347–1363 (1993).

[198]M. S. Gordon and M. W. Schmidt, "Chapter 41 - Advances in electronic structure theory: GAMESS a decade later", in *Theory and Applications of Computational Chemistry*, edited by C. E. Dykstra, G. Frenking, K. S. Kim, and G. E. Scuseria (Elsevier), pp. 1167–1189.

[199]G. M. J. Barca, C. Bertoni, L. Carrington, D. Datta, N. De Silva, J. E. Deustua, D. G. Fedorov, J. R. Gour, A. O. Gunina, E. Guidez, T. Harville, S. Irle, J. Ivanic, K. Kowalski, S. S. Leang, H. Li, W. Li, J. J. Lutz, I. Magoulas, J. Mato, V. Mironov, H. Nakata, B. Q. Pham, P. Piecuch, D. Poole, S. R. Pruitt, A. P. Rendell, L. B. Roskop, K. Ruedenberg, T. Sattasathuchana, M. W. Schmidt, J. Shen, L. Slipchenko, M. Sosonkina, V. Sundriyal, A. Tiwari, J. L. Galvez Vallejo, B. Westheimer, M. Włoch, P. Xu, F. Zahariev, and M. S. Gordon, "Recent developments in the general atomic and molecular electronic structure system", J. Chem. Phys. **152**, 154102.

[200]J. Hubbard, "Calculation of partition functions", Phys. Rev. Lett. **3**, 77 (1959).

[201]J. Lee, F. D. Malone, and M. A. Morales, "Utilizing essential symmetry breaking in auxiliary-field quantum Monte Carlo: Application to the spin gaps of the C36 fullerene and an iron porphyrin model complex", J. Chem. Theory Comput. **16**, 3019–3027 (2020).

[202]B. Rudshteyn, D. Coskun, J. L. Weber, E. J. Arthur, S. Zhang, D. R. Reichman, R. A. Friesner, and J. Shee, "Predicting ligand-dissociation energies of 3d coordination complexes with auxiliary-field quantum monte carlo", J. Chem. Theory Comput. **16**, 3041–3054 (2020).

[203]J. Shee, B. Rudshteyn, E. J. Arthur, S. Zhang, D. R. Reichman, and R. A. Friesner, "On achieving high accuracy in quantum chemical calculations of 3d transition metal-containing systems: A comparison of auxiliary-field quantum monte carlo with coupled cluster, density functional theory, and experiment for diatomic molecules", J. Chem. Theory Comput. **15**, 2346–2358 (2019).

[204]M. Kumar, J. Shee, B. Rudshteyn, D. R. Reichman, R. A. Friesner, C. E. Miller, and J. S. Francisco, "Multiple stable isoprene-ozone complexes reveal complex entrance channel dynamics in the Isoprene+ ozone reaction", J. Am. Chem. Soc., 10.1021/jacs.0c02360 (2020).

[205]H. Hao, J. Shee, S. Upadhyay, C. Ataca, K. D. Jordan, and B. M. Rubenstein, "Accurate predictions of electron binding energies of dipole-bound anions via quantum Monte Carlo methods", J. Phys. Chem. Lett. **9**, 6185–6190 (2018).

[206]J. Shee, E. J. Arthur, S. Zhang, D. R. Reichman, and R. A. Friesner, "Singlet–Triplet energy gaps of organic biradicals and polyacenes with auxiliary-field quantum monte carlo", J. Chem. Theory Comput. **15**, 4924–4932 (2019).

[207]J. Shee, E. J. Arthur, S. Zhang, D. R. Reichman, and R. A. Friesner, "Phaseless auxiliary-field quantum Monte Carlo on graphical processing units", J. Chem. Theory Comput. **14**, 4109–4121 (2018).

[208]W. Purwanto, H. Krakauer, Y. Virgus, and S. Zhang, "Assessing weak hydrogen binding on Ca+ centers: An accurate many-body study with large basis sets", J. Chem. Phys. **135**, 164105 (2011).

[209]N. M. O'boyle, A. L. Tenderholt, and K. M. Langner, "Cclib: A library for package-independent computational chemistry algorithms", J. Comput. Chem. **29**, 839–845 (2008).

[210]Q. Sun, T. C. Berkelbach, N. S. Blunt, G. H. Booth, S. Guo, Z. Li, J. Liu, J. D. McClain, E. R. Sayfutyarova, S. Sharma, S. Wouters, and G. K.-L. Chan, "PySCF: the Python-based simulations of chemistry framework", Wiley Interdiscip Rev Comput Mol Sci **8**, e1340 (2018).

[211]Q. Sun, X. Zhang, S. Banerjee, P. Bao, M. Barbry, N. S. Blunt, N. A. Bogdanov, G. H. Booth, J. Chen, Z.-H. Cui, J. J. Eriksen, Y. Gao, S. Guo, J. Hermann, M. R. Hermes, K. Koh, P. Koval, S. Lehtola, Z. Li, J. Liu, N. Mardirossian, J. D. McClain, M. Motta, B. Mussard, H. Q. Pham, A. Pulkin, W. Purwanto, P. J. Robinson, E. Ronca, E. R. Sayfutyarova, M. Scheurer, H. F. Schurkus, J. E. T. Smith, C. Sun, S.-N. Sun, S. Upadhyay, L. K. Wagner, X. Wang, A. White, J. D. Whitfield, M. J. Williamson, S. Wouters, J. Yang, J. M. Yu, T. Zhu, T. C. Berkelbach, S. Sharma, A. Y. Sokolov, and G. K.-L. Chan, "Recent developments in the PySCF program package", J. Chem. Phys. **153**, 024109 (2020).

[212]V. I. Lebedev and D. Laikov, "A quadrature formula for the sphere of the 131st algebraic order of accuracy", Dokl. Math. **59**, 477–481 (1999).

[213]N. Schlömer, N. R. Papior, D. Arnold, M. Ancellin, and R. Zetter, *Nschloe/quadpy v0.15.2*, version v0.15.2, Zenodo, Aug. 2020.

[214]J. Lee, D. W. Small, and M. Head-Gordon, "Excited states via coupled cluster theory without equation-of-motion methods: Seeking higher roots with application to doubly excited states and double core hole states", J. Chem. Phys. **151**, 214103 (2019).

[215]J. Noga and R. J. Bartlett, "The full CCSDT model for molecular electronic structure", J. Chem. Phys. **86**, 7041–7050 (1987).

[216]G. E. Scuseria and H. F. Schaefer, "A new implementation of the full CCSDT model for molecular electronic structure", Chem. Phys. Lett. **152**, 382–386 (1988).

[217]J. D. Watts and R. J. Bartlett, "The coupled-cluster single, double, and triple excitation model for open-shell single reference functions", J. Chem. Phys. **93**, 6104–6105 (1990).

[218]Y. J. Bomble, J. F. Stanton, M. Kállay, and J. Gauss, "Coupled-cluster methods including noniterative corrections for quadruple excitations", J. Chem. Phys. **123**, 054101 (2005).

[219]T. Yamazaki, S. Matsuura, A. Narimani, A. Saidmuradov, and A. Zaribafiyan, *Towards the practical application of near-term quantum computers in quantum chemistry simulations: A problem decomposition approach*, 2018.

[220]V. Deev and M. A. Collins, "Approximate ab initio energies by systematic molecular fragmentation", J. Chem. Phys. **122**, 154102 (2005).

[221]M. A. Collins and V. A. Deev, "Accuracy and efficiency of electronic energies from systematic molecular fragmentation", J. Chem. Phys. **125**, 10.1063/1.2347710 (2006).

[222]M. A. Addicoat and M. A. Collins, "Accurate treatment of nonbonded interactions within systematic molecular fragmentation", J. Chem. Phys. **131**, 104103 (2009).

[223]M. A. Collins, "Systematic fragmentation of large molecules by annihilation", Phys. Chem. Chem. Phys. **14**, 7744–7751 (2012).

[224]R. Kobayashi, M. A. Addicoat, A. T. Gilbert, R. D. Amos, and M. A. Collins, "The SMFA program for quantum chemistry calculations on large molecules", WIREs Comput. Mol. Sci. **9**, e1413 (2019).

[225]T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, "A polymer dataset for accelerated property prediction and design", Sci. Data **3**, 1–10 (2016).

[226]M. Zheng, N. W. Moriarty, Y. Xu, J. R. Reimers, P. V. Afonine, and M. P. Waller, "Solving the scalability issue in quantum-based refinement: Q|R#1", Acta Crystallogr. Sect. Struct. Biol. **73**, 1020–1028 (2017).

[227]M. Wolter, M. von Looz, H. Meyerhenke, and C. R. Jacob, "Systematic Partitioning of Proteins for Quantum-Chemical Fragmentation Methods Using Graph Algorithms", J. Chem. Theory Comput. **17**, 1355–1367 (2021).

[228]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in python", J. Mach. Learn. Res. **12**, 2825–2830 (2011).

[229]R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method", Comput. J. **16**, 30–34 (1973).

[230]D. Defays, "An efficient algorithm for a complete link method", Comput. J. **20**, 364–366 (1977).

[231]S. Lloyd, "Least squares quantization in PCM", IEEE Trans. Inf. Theory **28**, 129–137 (1982).

[232]H. Steinhaus, "Sur la division des corps matériels en parties", Bull Acad Pol Sci Cl III **4**, 801–804 (1957).

[233]A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: analysis and an algorithm", Adv. Neural Inf. Process. Syst. (2002).

[234]J. H. Ward, "Hierarchical grouping to optimize an objective function", J. Am. Stat. Assoc. **58**, 236–244 (1963).

[235]D. Arthur and S. Vassilvitskii, "How slow is the K-means method?", in Proc. Twenty-Second Annu. Symp. Comput. Geom. SCG '06 (2006), pp. 144–153.

[236]X. Yin, I. Neamtiu, S. Patil, and S. T. Andrews, "Implementation-induced Inconsistency and Nondeterminism in Deterministic Clustering Algorithms", in 2020 IEEE 13th Int. Conf. Softw. Test. Valid. Verification ICST (2020), pp. 231–242.

[237]Yu and Shi, "Multiclass spectral clustering", in Proc. Ninth IEEE Int. Conf. Comput. Vis. (2003), 313–319 vol.1.

[238]J. Shi and J. Malik, "Normalized cuts and image segmentation", IEEE Trans. Pattern Anal. Mach. Intell. **22**, 888–905 (2000).

[239]A. Damle, V. Minden, and L. Ying, "Simple, direct and efficient multi-way spectral clustering", Information and Inference: A Journal of the IMA **8**, 181–203 (2019).

[240]A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm", in Proc. 14th Int. Conf. Neural Inf. Process. Syst. Nat. Synth. NIPS'01 (2001), pp. 849–856.

[241]A. B. Georgescu, P. Ren, A. R. Toland, S. Zhang, K. D. Miller, D. W. Apley, E. A. Olivetti, N. Wagner, and J. M. Rondinelli, "Database, Features, and Machine Learning Model to Identify Thermally Driven Metal–Insulator Transition Compounds", Chem. Mater. **33**, 5591–5605 (2021).

[242]Y. Zhao, Y. Cui, Z. Xiong, J. Jin, Z. Liu, R. Dong, and J. Hu, "Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions", ACS Omega **5**, 3596–3606 (2020).

[243]B. Cordero, V. Gómez, A. E. Platero-Prats, M. Revés, J. Echeverría, E. Cremades, F. Barragán, and S. Alvarez, "Covalent radii revisited", Dalton Trans., 2832 (2008).

[244]D. G. A. Smith, D. Altarawy, L. A. Burns, M. Welborn, L. N. Naden, L. Ward, S. Ellis, B. P. Pritchard, and T. D. Crawford, "The MolSSI QCArchive project: An open-source platform to compute, organize, and share quantum chemistry data", WIREs Comput. Mol. Sci. **11**, e1491 (2021).

[245]I. Mayer, "Bond order and valence indices: A personal account", J. Comput. Chem. **28**, 18 (2006).

[246]L. Hubert and P. Arabie, "Comparing partitions", J. Classif. **2**, 193–218 (1985).

[247]M. Meilă, "Comparing clusterings–an information based distance", J. Multivar. Anal. **98**, 873–895 (2007).

[248]D. J. Wales, J. P. K. Doye, A. Dullweber, M. P. Hodges, F. Y. Naumkin, F. Calvo, J. Hernández-Rojas, and T. F. Middleton, *The Cambridge Cluster Database*.

[249]W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water", J. Chem. Phys. **79**, 926–935 (1983).

[250]N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open babel: An open chemical toolbox", J. Cheminformatics **3**, 1–14 (2011).

[251]J. D. Chai and M. Head-Gordon, "Long-range corrected double-hybrid density functionals", J. Chem. Phys. **131**, 1–13 (2009).

[252]J. D. Chai and M. Head-Gordon, "Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections", Phys. Chem. Chem. Phys. **10**, 6615–6620 (2008).

[253]Y. Xiang, D. J. Burrill, K. K. Bullard, B. J. Albrecht, L. E. Tragesser, J. McCaffrey, D. S. Lambrecht, and E. Pentzer, "Polymerization of silyl ketenes using alkoxide initiators: A combined computational and experimental study", Polym. Chem. **8**, 5381–5387 (2017).

[254]N. Mardirossian and M. Head-Gordon, "ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation", J. Chem. Phys. **144**, 214110 (2016).

[255]W. J. Hehre, R. F. Stewart, and J. A. Pople, "Self-consistent molecular-orbital methods. I. Use of gaussian expansions of slater-type atomic orbitals", J. Chem. Phys. **51**, 2657–2664 (1969).

[256]W. J. Hehre, R. Ditchfield, R. F. Stewart, and J. A. Pople, "Self-consistent molecular orbital methods. IV. Use of gaussian expansions of slater-type orbitals. Extension to second-row molecules", J. Chem. Phys. **52**, 2769–2773 (1970).

[257]E. D. Glendening and D. Feller, "Cation-water interactions: The M+(H2O)n clusters for alkali metals, M = li, na, K, rb, and cs", J. Phys. Chem. **99**, 3060–3067 (1995).

[258]J. Andzelm, S. Huzinaga, M. Klobukowski, E. Radzio-Andzelm, Y. Sakai, and T. Hiroshi, "Gaussian basis sets for molecular calculations", Phys. Sci. Data **16**, edited by e. S. Huzinaga, 27–426 (1984).

[259]T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen", J. Chem. Phys. **90**, 1007–1023 (1989).

[260]D. E. Woon and T. H. Dunning, "Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon", J. Chem. Phys. **98**, 1358–1371 (1993).

[261]K. R. Brorsen, Y. Yang, M. V. Pak, and S. Hammes-Schiffer, "Is the Accuracy of Density Functional Theory for Atomization Energies and Densities in Bonding Regions Correlated?", J. Phys. Chem. Lett. **8**, 2076–2081 (2017).

[262]B. Champagne, E. A. Perpète, S. J. A. van Gisbergen, E.-J. Baerends, J. G. Snijders, C. Soubra-Ghaoui, K. A. Robins, and B. Kirtman, "Assessment of conventional density functional schemes for computing the polarizabilities and hyperpolarizabilities of conjugated oligomers: An ab initio investigation of polyacetylene chains", J. Chem. Phys. **109**, 10489–10498 (1998).

[263]Y. Mei, N. Yang, and W. Yang, "Describing polymer polarizability with localized orbital scaling correction in density functional theory", J. Chem. Phys. **154**, 054302 (2021).

[264]S. J. A. van Gisbergen, P. R. T. Schipper, O. V. Gritsenko, E. J. Baerends, J. G. Snijders, B. Champagne, and B. Kirtman, "Electric field dependence of the exchange-correlation potential in molecular chains", Phys. Rev. Lett. **83**, 694–697 (1999).

[265]E. Sim, S. Song, S. Vuckovic, and K. Burke, "Improving Results by Improving Densities: Density-Corrected Density Functional Theory", J. Am. Chem. Soc. **144**, 6625–6639 (2022).

[266]E. Sim, S. Song, and K. Burke, "Quantifying Density Errors in DFT", J. Phys. Chem. Lett. **9**, 6385–6392 (2018).

[267]S. Boixo, S. V. Isakov, V. N. Smelyanskiy, R. Babbush, N. Ding, Z. Jiang, M. J. Bremner, J. M. Martinis, and H. Neven, "Characterizing quantum supremacy in near-term devices", Nat. Phys. **14**, 595–600 (2018).

[268]T. Monz, P. Schindler, J. T. Barreiro, M. Chwalla, D. Nigg, W. A. Coish, M. Harlander, W. Hänsel, M. Hennrich, and R. Blatt, "14-qubit entanglement: Creation and coherence", Phys. Rev. Lett. **106**, 130506 (2011).

[269]H.-S. Zhong, H. Wang, Y.-H. Deng, M.-C. Chen, L.-C. Peng, Y.-H. Luo, J. Qin, D. Wu, X. Ding, Y. Hu, et al., "Quantum computational advantage using photons", Science **370**, 1460–1463 (2020).

[270]IBM, *IBM's Roadmap For Scaling Quantum Technology*, (2020) `https://www.ibm.com/blogs/research/2020/09/ibm-quantum-roadmap/`.

[271]J. Preskill, "Quantum Computing in the NISQ era and beyond", Quantum **2**, 79 (2018).

[272]A. Fatima and I. L. Markov, "Faster Schrödinger-style simulation of quantum circuits", 2020.

[273]I. L. Markov, A. Fatima, S. V. Isakov, and S. Boixo, "Massively parallel approximate simulation of hard quantum circuits", in 2020 57th ACMIEEE Des. Autom. Conf. DAC (2020), pp. 1–6.

[274]Y. Ding and F. T. Chong, "Quantum computer systems: Research for noisy intermediate-scale quantum computers", Synth. Lect. Comput. Archit. **15**, 1–227 (2020).

[275]A. Li, O. Subasi, X. Yang, and S. Krishnamoorthy, "Density matrix quantum circuit simulation via the BSP machine on modern GPU clusters", in Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal. (2020), pp. 1–15.

[276]Z. Li and J. Yuan, "Quantum computer simulation on gpu cluster incorporating data locality", in Int. Conf. Cloud Comput. Secur. (2017), pp. 85–97.

[277]T. Häner and D. S. Steiger, "0.5 Petabyte Simulation of a 45-Qubit Quantum Circuit", in Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal. SC '17 (2017).

[278]X.-C. Wu, S. Di, E. M. Dasgupta, F. Cappello, H. Finkel, Y. Alexeev, and F. T. Chong, "Full-state quantum circuit simulation by using data compression", in Proc. Int. Conf. High Perform. Comput. Netw. Storage Anal. (2019), pp. 1–24.

[279]E. Pednault, J. A. Gunnels, G. Nannicini, L. Horesh, T. Magerlein, E. Solomonik, and R. Wisnieff, "Breaking the 49-qubit barrier in the simulation of quantum circuits", 2017.

[280]J. Doi, H. Takahashi, R. Raymond, T. Imamichi, and H. Horii, "Quantum computing simulator on a heterogenous hpc system", in Proc. 16th ACM Int. Conf. Comput. Front. (2019), pp. 85–93.

[281]G. Aleksandrowicz, T. Alexander, P. Barkoutsos, L. Bello, Y. Ben-Haim, D. Bucher, F. J. Cabrera-Hernández, J. Carballo-Franquis, A. Chen, C.-F. Chen, J. M. Chow, A. D. Córcoles-Gonzales, A. J. Cross, A. Cross, J. Cruz-Benito, C. Culver, S. D. L. P. González, E. D. L. Torre, D. Ding, E. Dumitrescu, I. Duran, P. Eendebak, M. Everitt, I. F. Sertage, A. Frisch, A. Fuhrer, J. Gambetta, B. G. Gago, J. Gomez-Mosquera, D. Greenberg, I. Hamamura, V. Havlicek, J. Hellmers, Ł. Herok, H. Horii, S. Hu, T. Imamichi, T. Itoko, A. Javadi-Abhari, N. Kanazawa, A. Karazeev, K. Krsulich, P. Liu, Y. Luh, Y. Maeng, M. Marques, F. J. Martín-Fernández, D. T. McClure, D. McKay, S. Meesala, A. Mezzacapo, N. Moll, D. M. Rodríguez, G. Nannicini, P. Nation, P. Ollitrault, L. J. O'Riordan, H. Paik, J. Pérez, A. Phan, M. Pistoia, V. Prutyanov, M. Reuter, J. Rice, A. R. Davila, R. H. P. Rudy, M. Ryu, N. Sathaye, C. Schnabel, E. Schoute, K. Setia, Y. Shi, A. Silva, Y. Siraichi, S. Sivarajah, J. A. Smolin, M. Soeken, H. Takahashi, I. Tavernelli, C. Taylor, P. Taylour, K. Trabing, M. Treinish, W. Turner, D. Vogt-Lee, C. Vuillot, J. A. Wildstrom, J. Wilson, E. Winston, C. Wood, S. Wood, S. Wörner, I. Y. Akhalwaya, and C. Zoufal, *Qiskit: An Open-source Framework for Quantum Computing*, version 0.7.2, Zenodo, Jan. 2019.

[282]Q. A. team and collaborators, *Qsim*, Zenodo, Sept. 2020.

[283]Microsoft, *QDK*.

[284]M. A. Nielsen and I. Chuang, *Quantum computation and quantum information*, 2002.

[285]K. De Raedt, K. Michielsen, H. De Raedt, B. Trieu, G. Arnold, M. Richter, T. Lippert, H. Watanabe, and N. Ito, "Massively parallel quantum computer simulator", Comput. Phys. Commun. **176**, 121–136 (2007).

[286]S. Aaronson and D. Gottesman, "Improved simulation of stabilizer circuits", Phys. Rev. A **70**, 052328 (2004).

[287]A. McCaskey, E. Dumitrescu, M. Chen, D. Lyakh, and T. Humble, "Validating quantum-classical programming models with tensor network simulations", PloS one **13**, e0206704 (2018).

[288]C. Huang, F. Zhang, M. Newman, J. Cai, X. Gao, Z. Tian, J. Wu, H. Xu, H. Yu, B. Yuan, et al., "Classical simulation of quantum supremacy circuits", 2020.

[289]D. Lykov, R. Schutski, A. Galda, V. Vinokur, and Y. Alexeev, "Tensor network quantum simulator with step-dependent parallelization", 2020.

[290]I. L. Markov and Y. Shi, "Simulating quantum computation by contracting tensor networks", SIAM J. Comput. **38**, 963–981 (2008).

[291]E. Pednault, J. A. Gunnels, G. Nannicini, L. Horesh, and R. Wisnieff, "Leveraging secondary storage to simulate deep 54-qubit sycamore circuits", 2019.

[292]M. Smelyanskiy, N. P. Sawaya, and A. Aspuru-Guzik, "qHiPSTER: The quantum high performance software testing environment", 2016.

[293]R. Li, B. Wu, M. Ying, X. Sun, and G. Yang, "Quantum supremacy circuit simulation on Sunway TaihuLight", IEEE Trans. Parallel Distrib. Syst. **31**, 805–816 (2019).

[294]A. Amariutei and S. Caraiman, "Parallel quantum computer simulation on the GPU", in 15th Int. Conf. Syst. Theory Control Comput. (2011), pp. 1–6.

[295]N. H. Stair, R. Huang, and F. A. Evangelista, "A Multireference Quantum Krylov Algorithm for Strongly Correlated Electrons", J. Chem. Theory Comput. **16**, 2236–2245 (2020).

[296]E. Farhi, J. Goldstone, and S. Gutmann, "A quantum approximate optimization algorithm", 2014.

[297]D. E. Koh, "Further extensions of Clifford circuits and their classical simulation complexities", 2015.

[298]M. Hein, W. Dür, J. Eisert, R. Raussendorf, M. Nest, and H.-J. Briegel, "Entanglement in graph states and its applications", 2006.

[299]S. Bravyi, D. Gosset, and R. König, "Quantum advantage with shallow circuits", Science **362**, 308–311 (2018).

[300]A. JavadiAbhari, S. Patil, D. Kudrow, J. Heckey, A. Lvov, F. T. Chong, and M. Martonosi, "ScaffCC: Scalable compilation and analysis of quantum programs", Parallel Comput. **45**, 2–17 (2015).

[301]M. J. Bremner, R. Jozsa, and D. J. Shepherd, "Classical simulation of commuting quantum computations implies collapse of the polynomial hierarchy", Proc. R. Soc. Math. Phys. Eng. Sci. **467**, 459–472 (2011).

[302]M. J. Bremner, A. Montanaro, and D. J. Shepherd, "Average-case complexity versus approximate simulation of commuting quantum computations", Phys. Rev. Lett. **117**, 080501 (2016).

[303]A. Gilliam, S. Woerner, and C. Gonciulea, "Grover adaptive search for constrained polynomial binary optimization", 2019.

[304]M. Motta, C. Genovese, F. Ma, Z.-H. Cui, R. Sawaya, G. K.-L. Chan, N. Chepiga, P. Helms, C. Jiménez-Hoyos, A. J. Millis, et al., "Ground-state properties of the hydrogen chain: dimerization, insulator-to-metal transition, and magnetic phases", Phys. Rev. X **10**, 031058 (2020).

[305]W. A. Al-Saidi, S. Zhang, and H. Krakauer, "Bond breaking with auxiliary-field quantum Monte Carlo", J. Chem. Phys. **127**, 144101 (2007).

[306]J. Hachmann, W. Cardoen, and G. K.-L. Chan, "Multireference correlation in long molecules with the quadratic scaling density matrix renormalization group", J. Chem. Phys. **125**, 144101 (2006).

[307]A. Baiardi and M. Reiher, "The density matrix renormalization group in chemistry and molecular physics: Recent developments and new challenges", J. Chem. Phys. **152**, 040903 (2020).

[308]Y. Shen, X. Zhang, S. Zhang, J.-N. Zhang, M.-H. Yung, and K. Kim, "Quantum implementation of the unitary coupled cluster for simulating molecular electronic structure", Phys. Rev. A **95**, 020501 (2017).

[309]A. Bouland, B. Fefferman, C. Nirkhe, and U. Vazirani, "On the complexity and verification of quantum random circuit sampling", Nat. Phys. **15**, 159–163 (2019).

[310]H. Aschauer, W. Dür, and H.-J. Briegel, "Multiparticle entanglement purification for two-colorable graph states", Phys. Rev. A **71**, 012319 (2005).

[311]D. A. Lidar and T. A. Brun, *Quantum error correction* (Cambridge university press, 2013).

[312]P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer", SIAM Rev. **41**, 303–332 (1999).

[313]G. Li, Y. Ding, and Y. Xie, "Tackling the qubit mapping problem for NISQ-era quantum devices", in Proc. Twenty-Fourth Int. Conf. Archit. Support Program. Lang. Oper. Syst. (2019), pp. 1001–1014.

[314]M. A. O'Neil and M. Burtscher, "Floating-point data compression at 75 Gb/s on a GPU", in Proc. Fourth Workshop Gen. Purp. Process. Graph. Process. Units (2011), pp. 1–7.

[315]A. W. Cross, L. S. Bishop, J. A. Smolin, and J. M. Gambetta, "Open quantum assembly language", 2017.

[316]A. Zulehner and R. Wille, "Advanced simulation of quantum computations", IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. **38**, 848–859 (2018).

[317]Y. Shi, N. Leung, P. Gokhale, Z. Rossi, D. I. Schuster, H. Hoffmann, and F. T. Chong, "Optimized compilation of aggregated instructions for realistic quantum computers", in Proc. Twenty-Fourth Int. Conf. Archit. Support Program. Lang. Oper. Syst. (2019), pp. 1031–1044.

[318]E. Gutierrez, S. Romero, M. A. Trenas, and E. L. Zapata, "Simulation of quantum gates on a novel GPU architecture", in Int. Conf. Syst. Theory Sci. Comput. (2007).

[319]P. Zhang, J. Yuan, and X. Lu, "Quantum computer simulation on multi-GPU incorporating data locality", in Int. Conf. Algorithms Archit. Parallel Process. (2015), pp. 241–256.

[320]A. Avila, A. Maron, R. Reiser, M. Pilla, and A. Yamin, "GPU-aware distributed quantum simulation", in Proc. 29th Annu. ACM Symp. Appl. Comput. (2014), pp. 860–865.