Quantum Monte Carlo for Accurate Energies and Materials Design

by

Kayahan Saritas

B.S., Sabanci University (2012)

Submitted to the Department of Materials Science and Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Materials Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Author .....
Department of Materials Science and Engineering
May 25, 2017

Certified by .....
Jeffrey C. Grossman
Professor
Thesis Supervisor

Accepted by .....
Donald R. Sadoway
John F. Elliott Professor of Materials Science and Engineering
Chairman, Departmental Committee on Graduate Students

Quantum Monte Carlo for Accurate Energies and Materials Design

by

Kayahan Saritas

Submitted to the Department of Materials Science and Engineering
on May 25, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Materials Science and Engineering

Abstract

Quantum Monte Carlo (QMC) is an electronic structure calculation method that is capable of calculating incredibly accurate solutions of Schrödinger equation of quantum mechanics for real systems. However, QMC is computationally very expensive compared to density functional theory (DFT) method, such that its application has been limited. In addition, QMC is a stochastic (Monte Carlo) method, meaning that the way calculations are initialized, where a lot of user effort is invested, is crucial for getting accurate results. Computational expense can be justified if the data would be used repeatedly, however the lack of automatization is a severe problem, if QMC would be used in materials discovery. In Chapter 4, we show our automated calculation strategy for formation energy of periodic materials using QMC. We show that our method performs almost by an order of a magnitude more accurate, compared to high throughput DFT strategies having empirical corrections. Nevertheless, it would be beneficial to understand when DFT methods fail such that QMC is used only when the computational expense is justified. A single DFT functional rarely performs uniformly accurate accross different materials and properties due to non-systematic errors. In Chapter 5, we investigate one specific example: dihydroazulene ring opening photoisomerization, where different substitutions on the ring opening moiety introduce isomerization enthalpy errors up to 0.8 eV. We show that GGA exchange is the main reason for failure in B3LYP, PBE and TPSSH functionals. However, performing a test, similar to the Chapter 5, on each chemical reaction can be an intimidating task where the benchmark set must be carefully devised by an expert in the field. In the absence of experiments, the DFT functional choice is still often done in heuristic way. In Chapter 6, we demonstrate how we can systematically analyze benchmark sets using machine learning to provide highly accurate reaction energies and provide DFT functional selection for different classes of materials when high accuracy calculations or experiments are not available. Our approach provides probabilities of getting accurate results for a reaction that is investigated using each DFT functional.

Thesis Supervisor: Jeffrey C. Grossman
Title: Professor

Acknowledgments

Pursuing a Ph.D. at MIT has been living a dream for me. I know that it would not be possible, without the people who supported me throughout important milestones in my life and over the course of my Ph.D. Firstly, I would like to express my appreciation and gratitude to my Ph.D. advisor, Prof. Jeffrey. C. Grossman. He certainly has influenced my both academic and professional development in a constructive way. He has been truly supportive of different projects that I had in mind and promoted my growth in all the positive ways. He has always been very enthusiastic discussing research with everyone in the group, patiently answered any questions no matter bright or naive. More importantly, he has been always supporting each member of the group in a very friendly, but commanding manner. It has been a joy for me to be a part of his group, both in the academic and personal sense. I believe that he is a very inspirational character for not just the people in his research group, but anyone who had a chance to interact with him. During the two semesters, which I worked as a teaching assistant for his course, 3.091 Introduction to Solid State Chemistry, Jeff showed me how a professor at MIT really should be. He knew how to produce teamwork to produce very creative and out-of- the-box solutions that one could face when teaching to a class of nearly 250 people. All the goodie bags that he started thinking months before the lecture 1, using air cannons to shoot t-shirts when students are not in the mood are just the few things that made him an exceptional teacher. One day we raided his thermodynamics lecture and invited him to do the "ice bucket challenge", which he used as an example to show that irreversible processes increase the entropy. He has been great to show us that there are always creative ways that make learning fun and rememberable. I am also grateful to my committee members, Prof. Ju Li and Prof. Alexie Kolpak, whose valuable input helped me a lot in shaping my research and thesis. I am very grateful to all members of Grossman Group, who made my 5-year stay in the group an enjoyable one. I like to thank Can Ataca, who helped me a lot helped me grow from a rookie Ph.D. student to an experienced one. I would like 5 to especially thank my office mates, in addition to Can, who always were

eager to discuss my research questions and help me get comfortable in the group: Priyank Kumar (Priku), Rajamani Ragunathan, Jongwon Choi, Michelle Tomasik, Donghun Kim, Jeongyun Kim and later on David Zhitomirsky and Eric Fadel. I am very grateful to meet with Priyank, Rajamani and Can (and their dear wives Tiziana, Manju and Sila) at the beginning of my Ph.D. Often times, we would eat lunch and dinner together, and meet at some of the weekends to spend time together. I will always remember those times happily and gratefully. I am indebted to my housemates over the course of my Ph.D. who helped to make me feel I am returning to my "home" every evening. Ted Golfinopoulos was the first housemate I had at MIT in Ashdown House. It was great to talk to him about basically everything from science to history as he was very knowledgeable. He was also a great cook with an appetite for sharing his food. During my stay in "22 Palermo", Hande Gunduz has been a true friend with whom it was easy to talk about anything. Denizcan Vanli and Omer Karaduman were also great guys with a sense of humor and wit, without whom our place wouldn't turn into such cosy and alive environment as it is. I also have to express my gratitude to my close friends Ece Alpaslan and Mehmet Ali Guney whom, I know, would go above and beyond to support me and make sure that I am doing well.

Lastly, I thank my father, Mustafa Saritas, mother, Hatice Saritas, and brother, Atakan Saritas. I don't think that there are any words which can truly describe now much I am indebted to their unconditional love and support at every moment of my life. I am one of the few lucky people who has the chance to take a family member as role model, my brother. I want to thank my fiancée Yasemin who made my Ph.D. experience and my life in Boston special and memorable for me. She has been just herself, beautiful, kind, loving, supporting and smart. I am grateful to know and have someone like her on my side, at the same time I am grateful to her parents for raising a daughter as valuable as her. I would like to dedicate this thesis to my parents and Yasemin, who have devoted themselves to me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The field of materials informatics and computational materials science is based on the promise of accelerating materials design using the power of computers, rather than performing tedious experiments in the lab based on trial and error. Important chemical and structural trends can be identified very rapidly using computational methods; therefore new insights are developed to perform well guided experiments. Large databases of structures and materials properties have been founded by numerous groups in academia[11, 12]. These databases can serve as a valuable source of reference data in, for example, formation energies, electronic structure band diagrams and evaluating phase stability using pre-calculated total energies of all structures of a compound within its compositional space.

Although having comprehensive databases for multiple materials properties is definitely useful, the accuracy of the predictions based on this database would rely on the quality of the methods that are used to produce such data. In Figure 1-1, computational methods are compared based on their computational cost, time and length scales that are applicable using these methods. Full Configuration Interaction (FCI), is the numerically exact limit for solving the Schroedinger's equation. All the other methods provided in Figure 1-1 include varying degrees of fundamental approximations to the Schroedinger equation. Semi-empirical methods and classical forcefields solve Schroedinger's equation indirectly as modeling the orbital overlaps or forces between the atoms, but their applicability is mostly limited to the systems that they

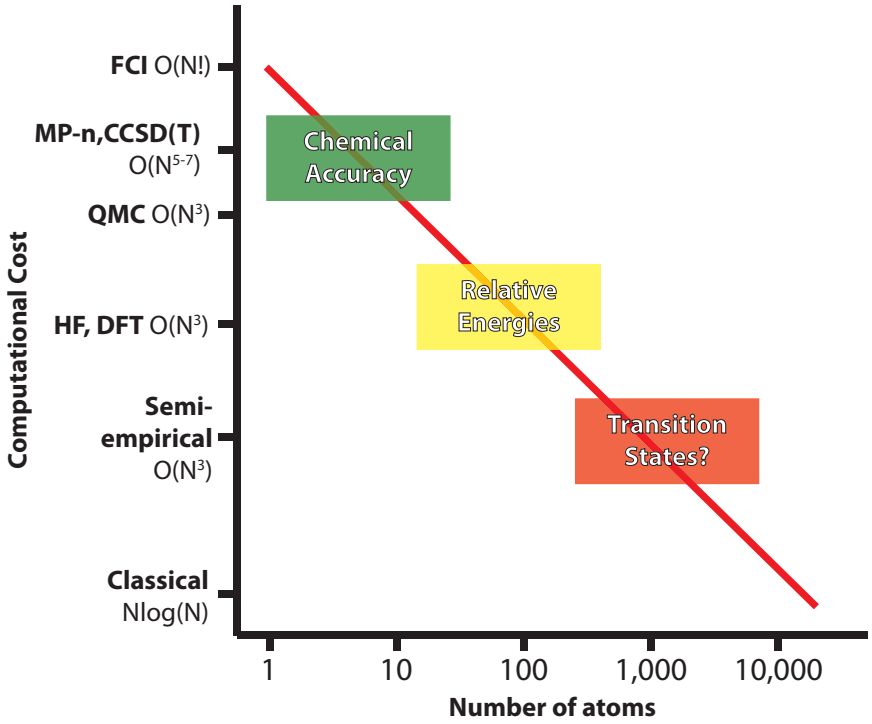


Figure 1-1: Computational cost versus typical sizes of systems that can be calculated with the corresponding method. For each method, the scaling is given with respect to number of electrons in the simulation, N.

are trained. However, electronic strucrure methods such as Hartree-Fock (HF) and Density Functional Theory (DFT)[13, 14] are different from the semi empirical and classical methods in the sense that they are first principles calculations. This means that for a given system no additional parameters are required, except for the geometric descriptions and description of the atoms building the system. Hence, it is expected that first principles calculations can perform accurately in a variety of systems and provide accurate relative energies which can drive further experimental studies. Electronic structure methods also enable calculating various electronic properties such as electronic band structures, magnetic properties, which are not possible with the classical methods. Above the HF method, electronic structure methods are overall listed with respect to their inclusion of the electron correlation. Almost for all systems of interest, solving Schroedinger's equation requires solving the interacting many-body electron problem. These many-body interactions between the electrons can be broken down into exchange and correlation (XC) interactions. Exchange inter-

action for electrons is based on the wavefunction of indistinguishable particles being subject to exchange symmetry, also known as the Pauli exclusion principle. An electron in a solid is an indistinguishable particle, however, its motion is further affected by the movement of other electrons around it by what is called correlation interactions. HF theory possesses the exact exchange but no correlation potential. DFT methods such as the local density approximation[15] (LDA) and generalized gradient approximation[16] (GGA) use local or semi-local descriptions of exchange and correlation approaches. Indeed, the correlation potential functional of the LDA functional was been developed using exact QMC calculations in the three-dimensional homogeneous electron gas[17].

Although DFT methods include crucial approximations to the many-body interaction problem, they have been strikingly useful for many problems of interest to the scientific community. DFT can provide very accurate lattice constants[18] and electronic properties[19] of numerous materials. Due to providing an excellent balance between accuracy and computational cost, DFT has been the main driving method in high throughput materials discovery in many applications as well[11, 20]. This success has been mostly attributed to the fortuitous cancellation of errors between exchange and correlation energies. However, it has been shown that when the chemical environment of the atoms changes significantly such as the oxidation state of an atom in a solid or the hybridization level of an atom in a molecule, cancellation of errors does not work as well. [20–22] In such cases, it is desirable to refer to experimental data, although accurate experimental data may not be available. For example, although there are more than 100,000 material entries in the Inorganic Crystal Structure Database (ICSD), there are fewer than 1,000 experimentally known material enthalpies of formation in the NIST-JANAF thermochemical tables[23].

High throughput materials discovery databases, such as Materials Project[12] and Open Quantum Materials Database[11] often compare the computational materials formation energies to available experimental formation energies, hence providing feedback on the range of errors that can be expected from DFT calculations. For GGA-PBE method[16], the mean absolute deviation (MAD) in predicting the formation

enthalpy of 1386 compounds in Materials Project database is given to be nearly 0.13 eV/atom, in comparison, when the computational recipe of Open Quantum Materials Database is used, this error reduces to 0.108 eV/atom[11]. Especially for complex phase diagrams, this accuracy simply cannot be sufficient to accurately identify all stable compounds on the phase diagram[1]. In Figure 1-2a, we show how the stability analysis is performed on a new phase material M_3O, on the hypothetical M-O binary phase diagram. Considering that formation energies of MO and M_2O_3 compounds are already known, in order to decide whether M_3O exists, a simple line is drawn between M and MO, then depending on where the formation energy of M_3O compound stands (below or above), stability of the compound is concluded (stable or unstable). However, due to non-systematical errors in DFT calculations, some compounds above this line (hull), with the cutoff energy ϵ are also considered as possible structures. In Figure 1-2b, how the choice of ϵ effects the decision making (stable or unstable compound) is given. In this figure, Narayan et. al. [1], studied the false positive and false negative probabilities based on a comparison of their results using high throughput DFT calculations as well as high throughput experimental syntheses. False positive means that a DFT stable but experimentally unstable material probability, whereas false negative means experimentally stable, but DFT unstable materials. According to Figure 1-2b, if $\epsilon = 0$, meaning that DFT calculations are completely trusted, nearly half of the cases, experimentally stable compounds are missed.

To improve the accuracy of the computational data stored in materials databases, researchers have investigated a number of alternatives to DFT/ GGA for high-throughput calculations, including hybrid DFT, DFT+U and GW calculations[24, 25]. The computational expense of these methods can be significantly greater than that of DFT/GGA, but this expense can often be justified if the calculated data is heavily re-used. However, it is important that a method used for the high-throughput calculation of material properties is scalable (both with system size and with the number of computing processors) and generally applicable to a wide variety of materials. With these goals in mind, quantum Monte Carlo (QMC) is a promising method for high-throughput, accurate calculations of material properties[26, 27].

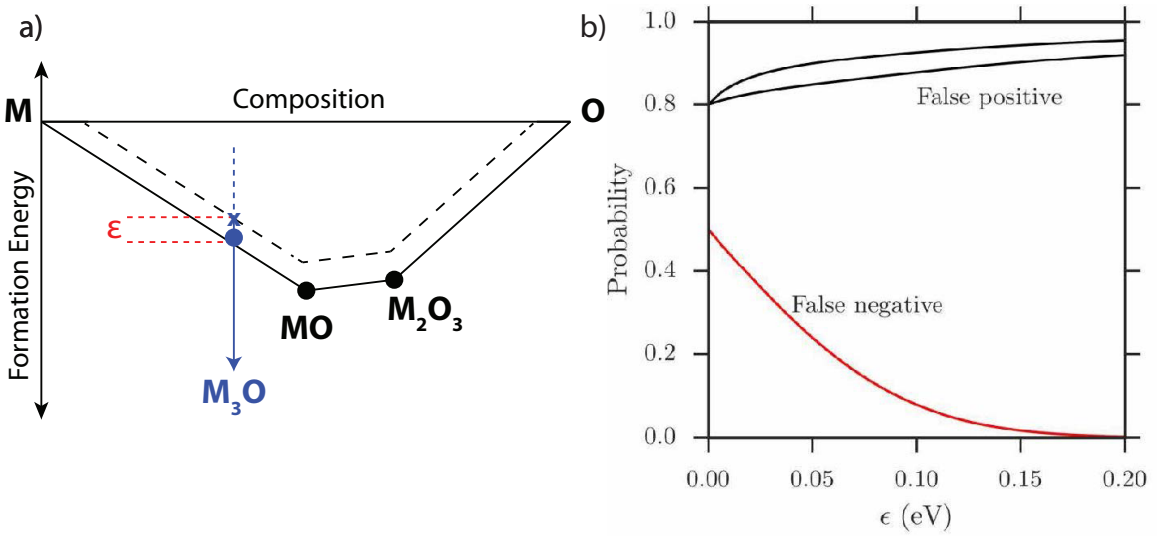


Figure 1-2: a) Stability analysis on a hypothetical M-O binary phase diagram. Assuming that MO and M_2O_3 formation energies are known, a search for M_3O is performed. ϵ can be considered as the trust cutoff on the formation energy of M_3O. b) Using the ϵ in a), false negative and positive probabilities are calculated in the formation energies of various ternary sulfides and selenides in ref. [1]

Although QMC or other "beyond-DFT" calculations, can provide highly accurate energies and other material properties such as energy gaps, it can be beneficial to develop strategies where we can learn from the failures of different DFT approximations in a systematic way, so that higher accuracy predictions can be made without resorting to the expensive electronic structure methods. Hence we develop a machine learning based strategy that combines the results of DFT calculations to make informed decision about DFT functional selection in future calculations. Machine learning approaches[28–30], has been recently applied to the materials science problems very efficiently. A typical application of machine learning applications in DFT for structure prediction is to replace DFT methods, such that instead of spending hours in a workstation to optimize geometries and find minimum energy structures, the same work can efficiently be done using machine learning approaches[31]. When only single DFT approach is used to prepare the reference energies while machine learning approach is being trained, results can be biased or have nonsystemmatic errors. Therefore, it can be beneficial to develop methods that learns from existing DFT benchmark sets to make better-informed decision for choosing the right DFT

functional for a problem at hand. In such an approach, the reference energies can be taken from experiments or high accuracy wave function based methods.

This thesis is organized as follows. First, we introduce the current standard methods with a short description followed by a discussion of the main points of the QMC method. Next, we give a quick summary of the goals of this work. After this, we discuss our main contributions: high throughput QMC calculations, critical benchmarking of DFT results on electrocyclization reactions, machine learning on DFT and QMC methods to obtain chemical accuracy in total energies. Finally, we put this work into the broader context of the currently available methods and the potential impact of QMC for materials discovery and prediction.

Chapter 2

Deterministic Electronic Structure Methods

The formulation of quantum mechanics has revolutionized our understanding of physics in the 20^{th} century. Electronic structure calculations solve the Schroedinger's equation for a system:

$$H = \sum_i \Big[-\frac{1}{2}\nabla_i^2 - \sum_\alpha \frac{Z_\alpha}{r_{i\alpha}} + \sum_{i>j} \frac{1}{r_{ij}} \Big] + E_{Ion-Ion}, \tag{2.1}$$

where H is the Hamiltonian, with the electrons labeled as i=1,2,3,...N, the ions labeled as α and the corresponding differences are given as r. It is assumed that electron movements are typically very fast compared to ion movements, therefore the Born-Oppenheimer approximation, in which ion motion is ignored, is considered. Solving equation 2.1 does not require any parameters other than natural constants such as Planck's constant. In this way, compared to other materials simulations methods, quantum mechanical calculations are distinct in the sense that they are first principles calculations. However, the majority of practical problems pertaining quantum mechanics are typically very hard to solve exactly, hence the use of numerical methods and approximations are adopted. The electron-electron interaction is the main source of the problem associated with solving exact Schroedinger equation for systems having more than two electrons. Therefore, these interactions are simulated

through the use of approximations to the many-body interacting electron problem.

The Hartree-Fock method for example treats the interacting electron problem through the use of Slater determinants, and it is in fact the simplest *ab-initio* treatment for wave function based correlated electron problem. Density functional theory (DFT) calculations, on the other hand, use a density based description for the interacting electron problem, such that the many-body electron problem is mapped to the single body problem, with modification of the electron-electron coulomb repulsion, $(r_{ij})^{-1}$ into what is called as the exchange-correlation potential. Other wavefunction based quantum chemistry methods, "post Hartree-Fock" methods, typically take into account a linear combination of Slater determinants (for configuration interaction methods), excitation operators on Slater determinant (in the case of Coupled Cluster[32] methods) and perturbation theory (for Moller-Plesset-n methods[33]).

2.1 Hartree Fock Theory (HF)

In HF theory one begins with the simplest form for a many-body antisymmetric wavefunction builf from one body orbitals, a Slater determinant:

$$\Psi = \begin{vmatrix} \psi_1(x_1) & \psi_1(x_2) & ... & \psi_1(x_N) \\ \psi_2(x_1) & \psi_2(x_2) & ... & \psi_2(x_N) \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ \psi_N(x_1) & \psi_N(x_2) & ... & \psi_N(x_N) \end{vmatrix}, \tag{2.2}$$

where the x variables includes space and spin. The Slater determinant ensures that the anti-symmetry of the wave function is satisfied, such that the electrons are indistinguishable and the wave function is degenerate when two electrons are exchanged.

2.2 Post-Hartree Fock Theories (Post-HF)

Post Hartree-Fock Methods are grouped into configuration interaction (CI), coupled cluster[32] (CC) and wave function based perturbation theories (MPn)[33]. These approaches are mainly based on expanding the number of determinants representing the wavefunction. A short discussion of these methods are included here, in order to draw comparisons with respect to the QMC method, as given in table 2.1. A more complete review of these methods can be found in [34].

Table 2.1: Comparison of several correlated post-HF methods. Form of the wavefunction, optimized variables and variational properties are listed

method	wavefunction	optimization	variational
CI	$\boldsymbol{\Phi = \sum_k c_k \phi_k}$	$\boldsymbol{c_k}$	yes
MPn	$\boldsymbol{\Phi = \sum_k c_k \phi_k^{(n)}}$	$\boldsymbol{c_k}$	no
CC	$\boldsymbol{\Phi = exp(\hat{T})\phi_{HF}}$	$\boldsymbol{\hat{T}}$	no

The CI method, in simple terms, considers linear combinations of excitations that are made in the Slater determinant. Each excitation is considered as a "configuration", where different electronic states are mixed to include correlation energy of the system. A full CI calculation solves the exact Schroedinger equation in a numerical way, although this would require almost infinite number of determinants. Depending on the number of excitations (single, double, triple...) the wavefunction is typically truncated and then solved in practical calculations. Each configuration is assigned a coefficient, which is then optimized self consistently to obtain energies and eigenstates. Typically, the zeroth degree excitation (the HF solution) provides the largest contribution to the energy, as correlation energies are small. Hence, the $\boldsymbol{c_0}$ is typically close to 1. For these reasons, the convergence of CI with respect to number of configurations can be challenging although the configuration set can be truncated with respect to minimum energies obtained. Although CI is a rigorous way to include correlation effects to the wavefunction, computational costs increase exponentially with the number of electrons. Furthermore, the truncated CI method suffers from size

extensivity problem, since the correlation energy per particle decreases with the size of the system. This means that for large systems or dimers of even small molecules, the results must be inherantly inaccurate.

Moller-Plesset (MPn) perturbation theory, on the other hand, introduces a perturbative potential to the Hamiltonian in the HF method, in order to include correlation effects, using the relation,

$$\hat{\boldsymbol{H}} = \hat{\boldsymbol{H}}_{\boldsymbol{0}} + \boldsymbol{\lambda}\hat{\boldsymbol{V}} \tag{2.3}$$

where $\boldsymbol{H_0}$ is the HF Hamiltonian, $\boldsymbol{\lambda}$ is a dimensionless parameter and $\boldsymbol{V}$ is the perturbative potential. Depending on the order of the perturbation, Moller-Plesset perturbation theory is named as MP2, MP3 ... For the case of MP2 method, the perturbation energy is obtained in the following way and added to the HF energy.

$$\boldsymbol{E_0^{(2)}} = \sum_{\boldsymbol{n}} \frac{|\langle \boldsymbol{0}| \boldsymbol{P} |\boldsymbol{n}\rangle|^{\boldsymbol{2}}}{\boldsymbol{E_0^{(2)}} - \boldsymbol{E_n^{(2)}}} \tag{2.4}$$

MPn theory is not variational, and it does not have to be convergent at the higher orders. Therefore, typically benchmarks are performed at the MP2 level, which is the computationally least demanding variant of MP theory. The accuracy of MP2 is usually between DFT and CC methods, although it can be challening to improve MP methods systematically, due to its non-variational nature.

The CC methods are the computationally most demanding post-Hartree Fock methods (except for full CI) where the CCSD(T) variant is considered to be the *gold standard* in the quantum chemistry community. Similar to MPn theory, the CC method is not variational. However, it is based on using a cluster operator on the Slater Determinant from the HF calculation, where single, double, triple etc. excitation operators are used to generate higher energy configurations. The CC Formalism is based on the second quantization, such that for single and double excitations, it is given as

$$\boldsymbol{T} = \boldsymbol{T_1} + \boldsymbol{T_2} + \boldsymbol{T_3} + \boldsymbol{....} \tag{2.5}$$

such that

$$\boldsymbol{T_1} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{a}} \boldsymbol{t_a^i \hat{a}^a \hat{a}_i}, \tag{2.6}$$

$$\boldsymbol{T_2} = \frac{\boldsymbol{1}}{\boldsymbol{4}} \sum_{\boldsymbol{i,j}} \sum_{\boldsymbol{a,b}} \boldsymbol{t_{ab}^{ij} \hat{a}^a \hat{a}^b \hat{a}_j \hat{a}_i} \tag{2.7}$$

where $\boldsymbol{\hat{a}^a}$ and $\boldsymbol{\hat{a}_i}$ denote creation and annihilation operators respectively. i,j stands for occupied, whereas a and b stands for unoccupied states. The general form of the Hamiltonian is the following:

$$\boldsymbol{|\Phi\rangle = e^T |\Phi_0\rangle} \tag{2.8}$$

where the exponential cluster operator can be expanded using a Taylor expansion

$$\boldsymbol{e^T = 1 + T + \frac{1}{2!} T^2 + \frac{1}{3!} T^3 + ...} \tag{2.9}$$

If the excitation operators were applied without the exponential form, they would correspond to the CI method. However, the use of exponential operators results in creating a larger number of configurations, since it is non-linear, with respect to the CI method. For practical reasons, only double and single excitations are used in the CC method (which is called CCSD with given excitations), however the CCSD method in general does not provide accurate results, therefore fully iterative or approximate (non-iterative) triple excitations are included in the CCSD method (hence CCSD(T) and CCSD[T] respectively). Overall, CCSD(T) calculations have very unfavorable scaling, $\boldsymbol{N^7}$, with repect to the number of electrons, hence these calculations are limited to only small or medium sized molecules with varying basis set quality.

2.3 Density Functional Theory (DFT)

Density Functional Theory is certainly the most successful electronic structure calculation method which, most of the time, provides a significant balance between computational cost and accuracy necessary for materials applications. Hohenberg and Kohn [13] developed the theory showing that for the ground state of a system

there is a unique energy and non-degenerate charge density. Hence, the system can be described using the charge density as a basic variable, since the total energy can be described as a unique function of the density at the ground state. One year later, Kohn and Sham formulated the Hamiltonian using DFT [14]. A comprehensive review for DFT can be found in ref. [35].

The general expression for DFT methods can be given as follows:

$$\boldsymbol{E[n(r)] = T_s[n(r)] + \frac{1}{2}\int\int \frac{n(r)n(r)}{|r' - r|} + E_{XC}[n(r)] + \int n(r)V_{ext}(r)dr} \quad (2.10)$$

here the $\boldsymbol{n(r)}$ denotes the charge density (spin variable omitted for simplicity), first term in the equation, $\boldsymbol{T_s}$, is the kinetic energy of non-interacting electrons, the second term is the electron-electron coulomb repulsion, the last term is the external potential which corresponds to the electron-ion potential energy. The ion-ion potential energy is omitted since it is purely classical. The remaining term, $\boldsymbol{E_{XC}}$, stands for the exchange correlation (XC) functional. DFT is an exact theory, describing the true Schroedinger equation, given the exact form of the exchange correlation functional. Therefore, all terms except for the kinetic energy in equation 2.10, can be written as an effective potential, $\boldsymbol{V_{eff}}$. XC functional reduces the many-body interaction of electrons into a single electron problem with the mean-field interaction:

$$\boldsymbol{\Big(-\frac{1}{2}\nabla_i^2 + V_{eff(r)} - \epsilon_i \Big)\psi_i(r) = 0} \quad (2.11)$$

which is solved numerically to obtain the charge density as given in the Hohenberg-Kohn theorem:

$$\boldsymbol{n(r) = \sum_{i=1}^{N} |\psi_i(r)|^2} \quad (2.12)$$

However, in practice, an exact form of the XC potential is not known, therefore approximations are used. The most simple way to approximate the exchange correlation functional is to create a new functional that is based on the local density. This

approximation can be formulated as:

$$\boldsymbol{E_{xc}[n(r)]} = \int \boldsymbol{n(r)\epsilon_{xc}[n(r)]dr} \tag{2.13}$$

This equation shows that the XC functional ,$\boldsymbol{\epsilon_{xc}}$, depends purely on the local electron density at $\boldsymbol{r}$. This is called as the local density approximation (LDA)[17] which was later found to be a surprisingly good approximation. Parametrization of the XC functional is obtained through QMC and many-body perturbation theory calculations, with their calculations of the homogeneous electron gas. Accuracy of LDA is based on the fact that exchange and correlation effects are dominated by the local environment and hence fluctuations around the point $\boldsymbol{r}$ and a finite volume $\boldsymbol{dr}$, does not affect these interactions very significantly. For these reasons, LDA works better in the case of slowly varying charge density systems, but it still has well known limitations such as underestimating lattice parameters and electronic bandgaps.

The generalized graient approximation (GGA)[36] is based on adding the gradient of the charge density as an additional parameter for the exchange correlation functional.

$$\boldsymbol{E_{xc}[n(r)]} = \int \boldsymbol{n(r)\epsilon_{xc}[n(r), \nabla n(r)]dr} \tag{2.14}$$

GGA typically performs better than LDA, provides more accurate band diagrams, lattice constants and thermodynamic quantities such as formation energies. However, further parametrization of the XC functional, such as in the case of meta-GGA[37] where the second derivative of the density is used as a parameter, has not been found to yield significantly better accuracy in most cases.

A more practical method to improve DFT functionals has been to use HF and DFT mixing in exchange and correlation functionals. In hybrid DFT methods, typically some portion of the exact exchange is typically incorporated to the local or semi-local DFT exchange energy. As an example the B3LYP functional [38, 39] is constructed

based on the following mixing of exchange $\boldsymbol{E_x}$ and correlation $\boldsymbol{E_c}$ energies:

$$\boldsymbol{E_{xc}^{B3LYP} = E_x^{LDA} + a_0(E_x^{HF} - E_x^{LDA}) + a_x(E_x^{B88} - E_x^{LDA}) + E_c^{LDA} - a_c(E_c^{LYP} - E_c^{LDA})} \tag{2.15}$$

where $\boldsymbol{a_0}$, $\boldsymbol{a_x}$ and $\boldsymbol{a_c}$ parameters are obtained from fitting to the experimental formation energies of well-studied molecules. Whereas for example, the HSE [40] functional is based on using 0.25 exact exchange, 0.75 GGA exchange and GGA correlation energies.

$$\boldsymbol{E_{xc}^{HSE} = 0.25E_x^{HF} + 0.75E_x^{GGA} + E_c^{GGA}} \tag{2.16}$$

However, ever increasing number of hybrid DFT methods are obtained through relaxing some of the exact constraints such as Lieb-Oxford bounds and spin scaling[41]. Coefficients in such XC formulations are optimized to obtain a set of energy differences that are as close as possible to the empirical target energies possible. However, this can introduce further non-systematic errors, such that performance of this fitted functional may not be transferable. A recent study[2], in Figure 2-1, comparing the performance of hybrid-DFT functionals based on the charge densities generated on a set of neutral and positively charged atoms show that recently developed hybrid DFT functionals show larger deviations from the atomic charge densities. Although for DFT functionals that satisfy the exact constraints at each level (LDA,GGA,meta-GGA, etc.) the error in the charge density decreases at each rung, for the new hybrid-DFT functionals which doesn't satisfy these criteria, the performance is unpredictable.

2.4 Summary

The challenge of the HF and DFT methods is the electron correlation. Although electron correlation comprises a very small part of the total energy (1% typically), in order to reach the accuracies that are required for materials prediction, its accurate treatment is still very important. Still, for systems that have similar properties or slowly varying charge densities, DFT and HF methods can provide reasonable

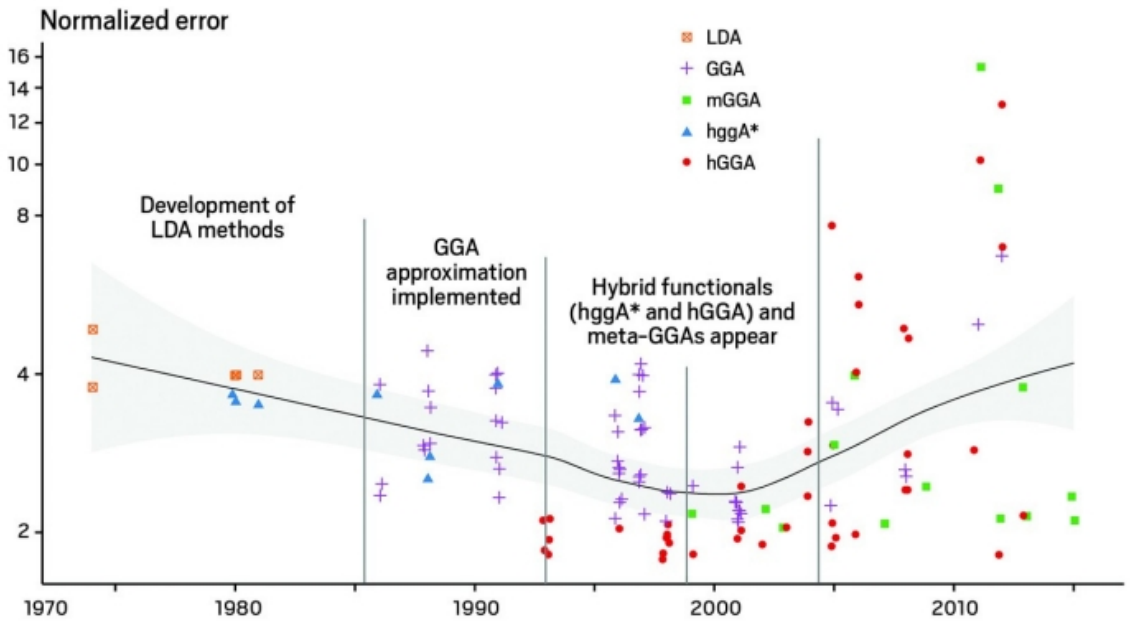


Figure 2-1: Historical trends in maximal deviation of the density produced by various DFT methods from the exact density in several atoms, molecules an cations. Line shows the average deviation and 95% confidence interval. Taken from ref. [2]

trends. Post-HF methods are, on the other hand very accurate, but their application is largely restricted to finite systems and they are very expensive to implement for large molecular systems.

Compared to these approaches, QMC methods provide an important alternative. As will be discussed in next chapter, QMC methods have computational scaling as good as DFT methods, while providing a highly accurate solution to the electron electron interaction. Furthermore, errors in QMC calculations can be improved in a systemmatic way. The accuracy required for a property is also a variable that can be tuned in QMC calculations depending on the length of the calculation.

Chapter 3

Quantum Monte Carlo

3.1 Introduction

Quantum Monte Carlo refers to a family of statistical methods for approximating a solution to the many-body Schrödinger equation in a way that explicitly accounts for both the antisymmetry of the many-body wavefunction and electron correlation. We review the Monte Carlo integration methods, variational Monte Carlo (VMC) and diffusion Monte Carlo (DMC) methods specifically. In chapter 4, we discuss efficient ways to implement QMC calculations in an high throughput environment.

3.2 Monte Carlo Method

3.2.1 Monte Carlo Integration

The integral of a function, $\boldsymbol{f(x)}$,can be written as follows:

$$\boldsymbol{F = \int_a^b f(x)dx = \lim_{N\to\infty} \frac{b-a}{N} \sum_{i=1}^{N} f(X_i),} \tag{3.1}$$

The first expression in eqn. 3.1 is the integration for the continuous variable $\boldsymbol{x}$, whereas the second expression stands for the Monte Carlo integration, using the discrete variable $\boldsymbol{X_i}$. A classical Riemann sum of the $\boldsymbol{X_i}$ variables would require these

variables to be uniformly distributed in the $[a, b]$ interval. However, when random variables are used, then a set of sums with a Gaussian distribution is obtained.

For higher dimensional integration, using uniform grid of random variables can be an inefficient strategy, as when d dimensions are used, then N^d number of grids are required for integration. Therefore a probability distribution, w_i, for the grid density is assigned to achieve efficient sampling of the integration grid:

$$F = \int_a^b f(x)dx = \lim_{N\to\infty} \frac{b-a}{N} \frac{\sum_{i=1}^{N} w_i f(X_i)}{\sum_{i=1}^{N} w_i}, \tag{3.2}$$

3.2.2 Importance Sampling

Equation 3.2 shows the general way to assign a probability distribution over integration grid. w_i in equation 3.2 can be modified using $p(x) = g(x)/\int_a^b g(x)dx$, where $g(x) \approx f(x)$. Many functions require more sampling near critical or density rich regions in order to provide a more efficient integration using a smaller grid. For example in gaussians, after three multiples of the variance, the value of the distribution decays almost to zero, therefore using a uniform grid in these ares for integration can be a waste of computational effort.

In Figure 3-1, we explain this strategy with an example using two gaussians as the starting probability density distribution, $f(x)$. Then randomly distributed uniform grid of points are chosen to perform the integration. However, with the importance sampling, the points are sampled among the initial distribution giving larger weights, w_i to the ones that are in large density regions. Therefore, thanks to the use of importance sampling, the standard deviation on the accuracy of the total energy decreases significantly:

$$\sigma_N = \sqrt{\frac{\frac{b-a}{N}\sum_{i=1}^{N} f^2(x_i) - E_N^2}{N-1}} \tag{3.3}$$

However, even though importance sampling significantly reduces the effort required for integration, the variance of the total energy, σ_N still scales with $1/\sqrt{N}$ as given in the denominator of eqn. 3.3.

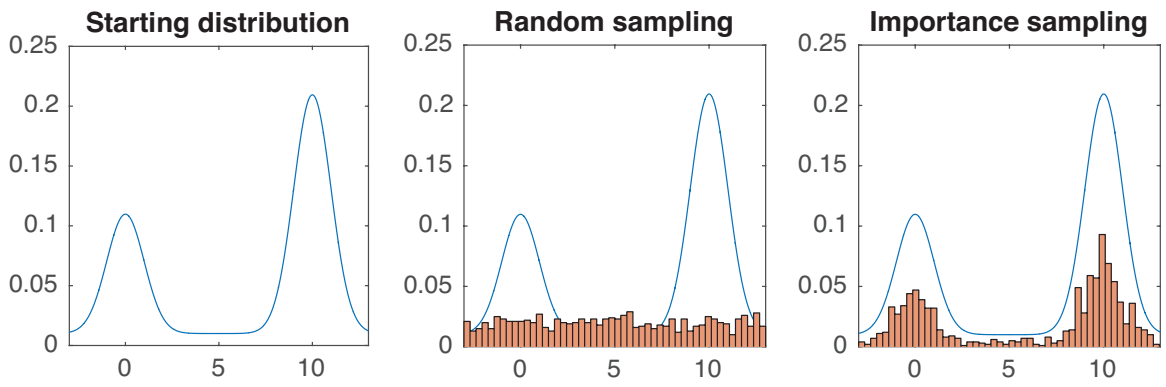


Figure 3-1: First image shows the starting distribution, $\boldsymbol{f(x)}$, with two gaussians centered at x=0 and 10 with variances of unity. The histograms, $\boldsymbol{p(x)}$, in the second image show the distribution of the uniform random grid using 10^3 points used to integrate the area under the gaussians. In the third image, however, the integration grid is importance sampled such that less grid points are used in areas where starting density is low.

3.2.3 Metropolis Algorithm

Importance sampling requires that a probability distribution function, $\boldsymbol{f(x)}$, must be defined in order to sample the integration grid. The Metropolis algorithm [42], in this work, enables one to obtain such a probability distribution function in the Monte Carlo simulation. From a starting phase or position in the configuration space describing the $\boldsymbol{f(x)}$, a move is generated and then rejected or accepthed through Metropolis algorithm. Hence the aim is to arrive at a stationary distribution of $\boldsymbol{X_i}$ as given in eqn. 3.3. With a move in Metropolis algorithm, the importance sampling of the $\boldsymbol{f(x)}$ ($\boldsymbol{X_i}$) is updated. Metropolis algorith aims to explore all possible configuration space through making a number of configuration moves. A move from a starting distribution of $\boldsymbol{X}$ to a new configuration $\boldsymbol{Y}$ must therefore satisfy:

$$\boldsymbol{p(X)K(X \rightarrow Y) = p(Y)K(Y \rightarrow X)} \tag{3.4}$$

$\boldsymbol{K(X \rightarrow Y)}$ is the conditional probability of moving from $\boldsymbol{X}$ to $\boldsymbol{Y}$, hence right and left sides of the equation 3.4 satisfies equal fluxes on both sides of the equation (detailed balance), meaning that random walk is ergodic such that all the points in the configuration space are reachable from one another. Furthermore, a sufficient

number of steps must be performed and sufficiently large $\boldsymbol{X_i}$ must be used to make sure that this approach is valid.

Since the conditional probability of moving from step $\boldsymbol{X}$ to $\boldsymbol{Y}$ is defined as $\boldsymbol{K(X \rightarrow Y)}$, this expression can be divided in two parts as the probability of acceptance and an arbitrary starting distribution. Hence $\boldsymbol{K(X \rightarrow Y)}$ can be written as:

$$\boldsymbol{p(X)A(X \rightarrow Y)T(X \rightarrow Y) = p(Y)A(Y \rightarrow X)T(Y \rightarrow X)} \tag{3.5}$$

where $\boldsymbol{T}$ is the arbitrary starting distribution and $\boldsymbol{A}$ is the acceptance distribution which is defined by the Metropolis Algorithm:

$$\boldsymbol{A(X \rightarrow Y) = \mathbf{min}\left[1, \frac{T(X \rightarrow Y)p(X)}{T(Y \rightarrow X)p(Y)}\right]} \tag{3.6}$$

In simple terms, the algorithm proceeds in the following way:

- given a starting position, $\boldsymbol{X_i}$, attempt to move to a new position, $\boldsymbol{Y}$, using $\boldsymbol{T(X_i \rightarrow Y)}$
- calculate the acceptance probability of the new move, $\boldsymbol{A(X \rightarrow Y)}$
- if the new move is accepted, then $\boldsymbol{X_{i+1} = Y}$, if not then $\boldsymbol{X_{i+1} = X_i}$

3.3 Variational Monte Carlo

Variational Monte Carlo (VMC) method simply aims to calculate multidimensional integral of Schroedinger's equation with Monte Carlo integration techniques. The expectation value of the energy, using the trial function is given as follows:

$$\boldsymbol{< \hat{H} >= \frac{\int E_L(R)|\Psi(R)|^2 dR}{\int |\Psi(R)|^2 dR}} \tag{3.7}$$

where $\boldsymbol{E_L = \Psi^{-1}\hat{H}(R)\Psi(R)}$ is given as the local energy. $\boldsymbol{H}$ is the Schroedinger Hamiltonian and $\boldsymbol{\Psi(R)}$ is the wavefunction of a given system. Within the VMC method, this expectation value is evaluated using the Metropolis algorithm, as given in

equation 3.8 (in discrete form), while generating a sequence of $\boldsymbol{R}$ (walkers) distributed according to $|\boldsymbol{\Psi(R)}|^{\boldsymbol{2}}$ and averaging the corresponding local energies.

$$< \boldsymbol{E_L} >= \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{k=1}}^{\boldsymbol{N}} \boldsymbol{E_L(R_k)} \tag{3.8}$$

VMC is a variational method, such that it can estimate an upper limit (variational limit) to the energy of a system. The variational energy that is obtained from VMC method has two types of errors: 1) a systematic error due to the use of an approximate $\boldsymbol{\Psi(R)}$ and 2) statistical unceratinty due to sampling of the local energy as in eqns. 3.3 and 3.8. The systematic error due to using approximate wavefunction can only be determined only after the Hamiltonian is solved exactly, meaning after diffusion Monte Carlo method is applied. If $\boldsymbol{\Psi(R)}$ is an exact function of the Hamiltonian, it would mean that the local energy $\boldsymbol{E_L}$ is also exact, such that given there are sufficient number of samples, the variance decays to zero. Although the statistical uncertainty depends on the number of samples, or the system size for a fixed number of samples used in Monte Carlo integration, the statistical uncertainty can also be reduced if the variance of the local energy is small. In this limit again, the difference between the exact and variation energies decrease to zero and hence the variance of the local energy also converges to zero as well.

In other words, for a fixed eigenstate of the wavefunction the local energy is constant; however, for the trial wavefunction it is a series of normally distributed values. Hence, the efficiency of the VMC method relies on the variance of this distribution. Here, efficiency of VMC method can be determined through the percentage of the correlation energy recovered through VMC method. Typically, wavefunction optimization through VMC method recovers around 90% of the correlation energy. The remaining correlation energy, which remains as the variational bias in VMC calculations, is captured through diffusion Monte Carlo (DMC) method which is discussed in the next section. In practical calculations, a well optimized wavefunction through VMC calculations would mean that the energy differences between VMC and diffusion Monte Carlo (DMC) calculations, $\boldsymbol{E_{DMC} - E_{VMC}}$ are small. This is again related

to the variance of the energies obtained from VMC method, such that the set of $\boldsymbol{R}$, $\boldsymbol{R}$, converges faster to equilibrium positions, meaning the VMC calculations yield smaller standard deviations. Therefore, smaller number of statistical accumulation steps become sufficient to obtain well converged results, meaning that smaller computational effort is spent in the computationally much expensive DMC calculations. For finite size calculations, feasible options for creating suitable trial wavefunction is much more extensive, e.g. multiple determinants can be used to generate the trial wavefunction. However, for solids, one is limited to using density functionals, due to the complexity of performing wavefunction based correlated calculations in periodic systems. A typical strategy to obtain a suitable trial wavefunction for solids is to use a hybrid scheme and varying the exact exchange ratio of the functional, with respect to local (or semi-local) exchange.

The VMC algorithm is composed of two main steps, where initially the random set of walkers are *equilibrated* according to a Metropolis algorithm based on acceptance and rejection. In the second step, statistical accumulation of the walkers is applied where they are weighted using the Metropolis acceptance probabilities (eqn 3.8) used to obtain a distribution of walkers, electron-by-electron or configuration-by-configuration. In electron-by-electron sampling, each step consists of proposing individual moves for each of the electrons and each move is subject to individual acceptance/rejection probabilities. Configuration-by-configuration sampling involves a full configuration move (i.e. all electrons) per step which is similarly evaluated via Metropolis algorithm. Configuration-by-configuration typically suffers from long correlation times, such that longer VMC simulations would be necessary to equilibrate and accumulate VMC calculations. Eliminating the correlation in configuration-by-configuration sampling requires larger computational time overall, hence electron-by-electron sampling is used in all cases we investigate in this work.

3.4 Diffusion Monte Carlo

The DMC method solves the time dependent Schroedinger equation in a stochastic manner. It is based on the imaginary time projector $\boldsymbol{exp(-\tau H)}$, where $\boldsymbol{\tau = it}$ and when $\boldsymbol{\tau \to \infty}$, higher energy eigenstates are filtered out making the solution effectively reaching the unique ground state. The imaginary time solution is:

$$\frac{\boldsymbol{\partial} |\boldsymbol{\Phi}\rangle}{\boldsymbol{\partial \tau}} = -\hat{\boldsymbol{H}} |\boldsymbol{\Phi}\rangle \tag{3.9}$$

Hence:

$$\hat{\boldsymbol{H}} |\boldsymbol{\Phi}\rangle = \hat{\boldsymbol{H}} \sum_{\boldsymbol{i=0}}^{\boldsymbol{\infty}} \boldsymbol{c_i} |\boldsymbol{\psi_i}\rangle = \sum_{\boldsymbol{i=0}}^{\boldsymbol{\infty}} \boldsymbol{\epsilon_i c_i} |\boldsymbol{\psi_i}\rangle \tag{3.10}$$

The imaginary time propagated Schroedinger equation follows:

$$|\boldsymbol{\Phi(\tau + \delta\tau)}\rangle = \boldsymbol{e^{-\hat{H}\delta\tau}} |\boldsymbol{\Phi(\tau)}\rangle \tag{3.11}$$

Therefore, the equation 3.10 can be modified into:

$$\hat{\boldsymbol{H}} |\boldsymbol{\Phi(\delta\tau)}\rangle = \sum_{\boldsymbol{i=0}}^{\boldsymbol{\infty}} \boldsymbol{c_i e^{-\epsilon_i \delta\tau}} |\boldsymbol{\psi_i(\tau)}\rangle \tag{3.12}$$

Following equation 3.12, we can say that after sufficient number of time propagations, high energy eigenstates decay with the exponential term in the Hamiltonian. However, the exponential term requires a reference energy, $\boldsymbol{E_T}$ in order to obtain a finite ground state energy:

$$\boldsymbol{\partial_t} |\boldsymbol{\Phi}\rangle = -(\boldsymbol{H - E_t}) |\boldsymbol{\Phi}\rangle = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \frac{\hbar}{\boldsymbol{2m_i}} \boldsymbol{\nabla_i^2} |\boldsymbol{\Phi}\rangle - [\boldsymbol{V(R) - E}] |\boldsymbol{\Phi}\rangle \tag{3.13}$$

$|\boldsymbol{\Phi}\rangle$ here is the DMC wavefunction which uses the importance sampling as described before. However, if this equation is not provided and boundary conditions, it converges to a bosonic ground state. Many-electron problem is anti-symmetric with the boundary conditions which is defined as the nodal surface of the trial wave function such that $\boldsymbol{\Psi_T(R) = 0}$. Hence the wavefunction in equation 3.13 modified

via importance sampling transformation, $\boldsymbol{f(R,t) = |\Phi(R,t)\rangle\,|\Psi_T(R)\rangle}$, such that unknown $|\Phi\rangle$ is multipled with $|\Psi_T\rangle$. This also solves the fermion sign problem as, the new distribution $\boldsymbol{f(R,t)}$ can be non-negative for all $\boldsymbol{R}$ and $\boldsymbol{t}$. $\boldsymbol{f(R,t)}$ is then sampled using a Green's function:

$$\boldsymbol{f(R,t) = \int G(R \leftarrow R', t-t')f(R',t')dR'} \tag{3.14}$$

where the Green's function can be separated into drift-diffusion and branching probabilities:

$$\begin{aligned} \boldsymbol{G(R \leftarrow R', \delta\tau) = G_D(R \leftarrow R', \delta\tau)G_B(R \leftarrow R', \delta\tau)} \\ \boldsymbol{G_D(R \leftarrow R', \delta\tau) = \frac{1}{(2\pi\delta\tau)^{\frac{3N}{2}}} exp\Big(- \frac{(R - R' - \delta\tau V(R'))^2}{2\tau}\Big)} \\ \boldsymbol{G_B(R \leftarrow R', \delta\tau) = exp(-\frac{\delta\tau}{2}[E_L(R) + E_L(R') - 2E_T)])} \end{aligned} \tag{3.15}$$

Separation of drift-diffusion and branching parts of the propagator is valid for sufficiently small $\boldsymbol{\delta\tau}$ values only. Therefore $\boldsymbol{\delta\tau}$, the time step, should be chosen in a way that will not limit the speed of the calculations, as well as not causing any breakdown of the approximation made for the decomposition of the Green's function. Typically 0.01 H^{-1} can be a suitable value for the time step[43], but errors must be checked to make sure there is no significant bias introduced. A more practical way to check for the ideal time step is to observe the acceptance ratio of the DMC steps. Typically a sufficiently small time step $\boldsymbol{\delta\tau}$ must provide acceptance ratios more than 99%. Acceptance ratios are general guides for choosing a suitable timesteps for a simulation, but if one aims to perform highly accurate calculations, then multiple timesteps must be tested.

3.5 Implementation of QMC calculations

Although theoretical details of the VMC and DMC calculations are well described, further efforts must be spent in order to apply these principles to practical calcula-

tions. Overall, these details are related to the compactness of the wavefunctions and efficient ways of wavefunction optimization; removing chemically inactive core electrons via pseudopotentials, and eliminating of the spurious effects of using periodic potentials for integrating the Schroedinger equations via eliminating finite size errors and performing supercell calculations.

3.5.1 Wavefunctions

Jastrow-Slater[44] and Jastrow-Pfaffian[45] or geminal Jastrow[46] are the primary forms wavefunctions that enable performing many-body electron-electron interactions in QMC methods. While Jastrow-Slater form is typically used in one particle forms (mean field), Jastrow-Pfaffian or geminal Jastrow forms are used in two particle (singlet pairs) ground state wavefunctions. Given infinite amount of computational power, QMC results do not have to rely on VMC calculations and a trial wavefunction optimized through VMC calculations, but as explained in the previous section, an exact density of the Hamiltonian yields exact energies with less compuattional power and smaller variances. QMC calculations are not restricted by the form of trial wavefunction that can be used and in fact a very large number of Slater-jastrow forms are possible [47] , each having the general form of:

$$\boldsymbol{\Psi_T(R)} = \sum_{n} \boldsymbol{a_n D_n^{\uparrow}}\ (\boldsymbol{r_1^{\uparrow}}\ , ..., \boldsymbol{r_n^{\uparrow}}\ )\boldsymbol{D_n^{\downarrow}(r_1^{\downarrow}, ..., r_n^{\downarrow})exp|J|} \tag{3.16}$$

$\boldsymbol{D_n}$ is the Slater determinant as defined in equation 2.2, with up and down spins separately defined using the superscript. The Jastrow factor $\boldsymbol{|J|}$ on the other hand is a function of electron and atom positions. As mentioned before, the $\boldsymbol{D_n}$ is constructed from DFT or HF calculations. The general definition of the $\boldsymbol{\Psi_T(R)}$ includes $\boldsymbol{a_n}$, such that multiple determinants can be optimized simultaneously. However, as mentioned previously, this is either necessary only for very accurate calculations or already computationally too prohibitive.

The Jastrow factors we use include isotropic electron-electron terms $\boldsymbol{u}$, electron-electron-electron term $\boldsymbol{W}$, electron-nucleus terms $\boldsymbol{\chi}$, electron-electron-nucleus terms

$\boldsymbol{f}$ and also for periodic systems, plane-wave expansion of electron-electron separation and electron position, $\boldsymbol{p}$ and $\boldsymbol{q}$:

$$\boldsymbol{J}(\boldsymbol{r_i}, \boldsymbol{r_j}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{u}(\boldsymbol{r_{ij}}) + \boldsymbol{W}(\boldsymbol{r_{ij}}) + \sum_{I=1}^{N_{ions}} \sum_{i=1}^{N} \boldsymbol{\chi_I}(\boldsymbol{r_{iI}}) + \sum_{I=1}^{N_{ions}} \sum_{i=1}^{N} \sum_{j=i+1}^{N} \boldsymbol{f_I}(\boldsymbol{r_{iI}}, \boldsymbol{r_{jI}}, \boldsymbol{r_{ij}}) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{p}(\boldsymbol{r_{ij}}) + \sum_{i=1}^{N} \boldsymbol{q}(\boldsymbol{r_i}) \quad (3.17)$$

where $\boldsymbol{N}$ is the number of electrons, $\boldsymbol{N_{ions}}$ is the number of ions; $\boldsymbol{r_{ij}}$, $\boldsymbol{r_{iI}}$ are the electron-electron and electron-ion distances respectively. Multiple forms of functions describing these interactions are available, however throughout our work we used the Jastrow factors as used by Drummond et. al. [48], because it uses simple power series in interparticle distances, which is much easier to evaluate compared to Chebyshev polynomials[49], exponentials[50] or scaled variables[51]. The electron-electron interaction is described in the following way:

$$\boldsymbol{u}(\boldsymbol{r_{ij}}) = (\boldsymbol{r_{ij}} - \boldsymbol{L_U})^{\boldsymbol{C}} \times \boldsymbol{\theta}(\boldsymbol{L_U} - \boldsymbol{r_{ij}}) \times \Big(\boldsymbol{\alpha_0}\Big[\frac{\boldsymbol{\Gamma_{ij}}}{(-\boldsymbol{L_U})^{\boldsymbol{C}}} + \frac{\boldsymbol{\alpha_0 C}}{\boldsymbol{L_U}}\Big]\boldsymbol{r_{ij}} + \sum_{l=2}^{N_u} \boldsymbol{\alpha_l} \boldsymbol{r_{ij}^l}\Big) \quad (3.18)$$

where the interaction has a cutoff at $\boldsymbol{L_U}$ and has $\boldsymbol{C-1}$ continuous derivatives at $\boldsymbol{L_U}$. $\boldsymbol{\theta}$ is the Heaviside function, Kato cusp conditions[48, 52] impose $\boldsymbol{\Gamma_{ij} = 1/2}$ if the i and j have opposite spins, $\boldsymbol{\Gamma_{ij} = 1/4}$ if they have the same spin. C describes the behavior of the gradient of $\boldsymbol{u}$ at the cutoff length. If $\boldsymbol{C = 2}$ the second derivative of u, hence the local energy, is discontinuous, whereas if $\boldsymbol{C = 3}$, then the third derivative of u is discontinuous. However, in an inhomogeneous system, it can distort the charge density. For the electron-ion interaction, the expression above is slightly modified:

$$\boldsymbol{\chi I}(\boldsymbol{r_{ij}}) = (\boldsymbol{r_{iI}} - \boldsymbol{L_{\chi I}})^{\boldsymbol{C}} \times \boldsymbol{\theta}(\boldsymbol{L_{\chi I}} - \boldsymbol{r_{ij}}) \times \Big(\boldsymbol{\beta_0}\Big[\frac{\boldsymbol{Z_I}}{(-\boldsymbol{L_{\chi I}})^{\boldsymbol{C}}} + \frac{\boldsymbol{\beta_{0I} C}}{\boldsymbol{L_{\chi I}}}\Big]\boldsymbol{r_{iI}} + \sum_{l=2}^{N_u} \boldsymbol{\alpha_l} \boldsymbol{r_{ij}^l}\Big) \quad (3.19)$$

where, for example the $\boldsymbol{\Gamma}$ is replaced by the ionic charge $\boldsymbol{Z_I}$ due to electron nucleus cusp conditions. Compared to $\boldsymbol{u}$, $\boldsymbol{\chi}$ undoes the effect of $\boldsymbol{u}$ on the HF (or DFT) charge

density.

Electron-electron-ion also uses a similar expression compared to $\boldsymbol{u}$ and $\boldsymbol{\chi}$. However, it is relatively more costly to evaluate these parameters, hence not used in the rest of our work. For periodic calculations, we used the plane-wave expansions of electron-electron separation, $\boldsymbol{p}$:

$$\boldsymbol{p}(\boldsymbol{r_{ij}}) = \sum_{\boldsymbol{A}} \boldsymbol{a_A} \sum_{\boldsymbol{\{G^+\}}} \boldsymbol{cos}(\boldsymbol{G} \cdot \boldsymbol{r_{ij}}) \tag{3.20}$$

where $\boldsymbol{G}$ is the set of the reciprocal lattice vectors that are independent under the imposed symmetry. Here, '+' means that if the $+\boldsymbol{G}$ is included in the sum, then $-\boldsymbol{G}$ is excluded.

The efficiency of using multiple Jastrow factors for each interaction type can be investigted comparing the energies (and variances) obtained from these variables. Typically a larger number of Jastrow factors can recover a larger portion of the correlation energy, hence provide lower energies, and tighter variances; however one type of interaction may be more important compared to others. Furthermore, an increased number of Jastrow parameters means wavefunction optimization will be more difficult, which effectively can be limited by memory or CPU. Therefore, the many-body trial wavefunction obtained from VMC calculations should be as simple as possible without losing much of the accuracy.

Furthermore, the choice of the representation of the trial wavefunction can substatially effect the scaling performance of the calculations. In DMC, each time a single electron is moved in the configuration, O(N) orbitals must be evaluated, where N is the number of electrons. However, all electrons are updated within a move, such that total cost of performing a single step is $O(N^2)$. After each configuration move, basis set coefficients for the DMC wawvefunctions must be updated as well. This is where the choice of wavefunction matters, such that when plane-wave basis is used, due to its non-local representation, the plane wave coeffieients must be updated for each electron move, meaning that the cost is O(N). However, if a local basis is used (e.g. blip splines), then this evaluation is O(1) due to using a local fixed cutoff length

for electron-electron interactions. Finally, O(N) steps must be performed to reduce the statistical error bar to a desirable level, (e.g. 3.3). Therefore, DMC scaling can change between $O(N^{3-4})$ depending on using non-local (plane-wave) or local basis.

3.5.2 Wavefunction optimization

Optimization of the wavefunction essentially means the 1) optimization of the Jastrow factors in the many-body wavefunction using VMC calculations and 2) optimization of the nodal surface of the trial wavefunction. On the nodal surface of a wave function, the wavefunction equals to zero meaning that probability of finding an electron in these regions is zero as well. Therefore in fixed node DMC calculations, nodal surface of the trial wavefunction is not optimized during the calculations. Without this constraint, the equation in 3.13 suffers from the Fermion sign problem[53], meaning that the ground state of the many-body electron system would be of bosonic character. Therefore, non-exact nodal surfaces would introduce what is called as fixed node error. There still ways to optimize the nodal surface of the trial wavefunction using backflow transformations or using multiple determinants from post-Hartree-Fock methods such as Configuration Interaction. However, optimization of the nodal surface of the trial wavefunction is more challenging due to spanning a larger number of optimization parameters as well as the computational expense of performing these calculations. Overall effect of optimizing the trial wavefunction optimizing Jastrow factors only and backflow transformations is shown in Figure A well optimized wave function can recover a larger portion of the correlation energy and yield smaller variances in the VMC calculations. The amount of work that needs to be carried out in the DMC calculations based on such wavefunctions can be decreased.

Typically, tens of parameters are used for each variable of the Jastrow factor and the minimum is found using unconstrained minimization, which does not require calculation of the derivatives. As mentioned in equation 3.3, variance is an important quantity that can be minimized to obtain accurate trial wavefunctions. We consider a trial wavefunction, $\boldsymbol{\Psi^{\{\alpha\}}(R)}$, where $\boldsymbol{\{\alpha\}}$ is the set of Jastrow parameters. For a set of $\boldsymbol{N_C}$ configurations, $\boldsymbol{R}$ is distributed according to $\boldsymbol{|\Phi^{\{\alpha_0\}}(R)|^2}$ for some fixed

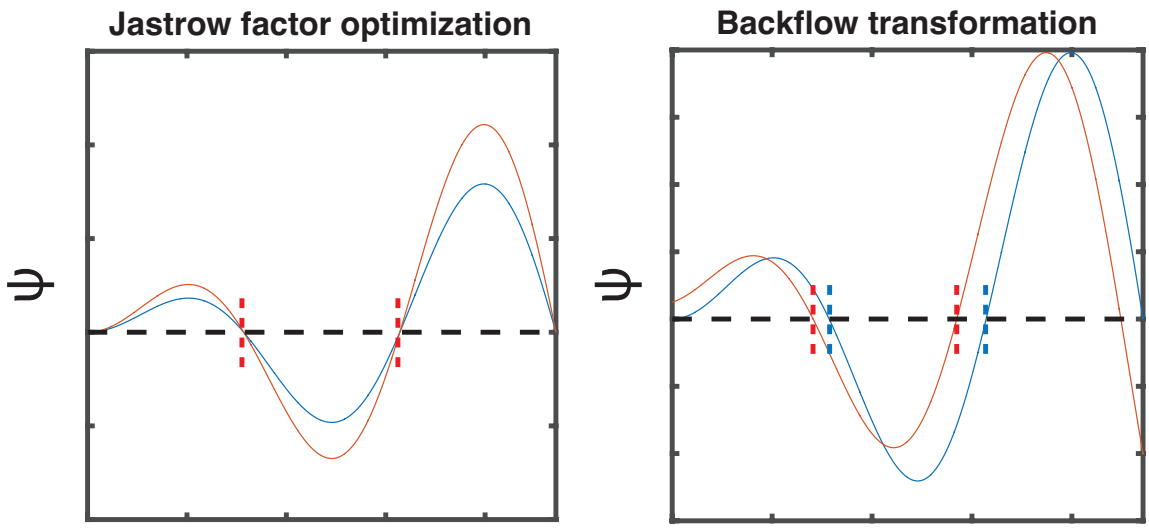


Figure 3-2: Effect of Jastrow factor optimization and backflow transformation on the cross section of a wave function and its nodal surface. In both graphs, dashed horizontal lines denote the multidimensional plane where $\boldsymbol{\psi = 0}$. In both cases, the blue colored wavefunction is the starting wavefunction for optimization. Adapted from [3].

parameter set $\boldsymbol{\{\alpha_0\}}$. When VMC wavefunction optimization involves optimization of the trial wavefunction, then it is possible that Jastrow parameters may diverge, so that the reweighted variance, which assigns weights to each configuration sampled, could be useful to obtain the variance[54]. However, throughout this work, we use the unreweighted variance because it provides more stable optimization of the variance[54]:

$$\boldsymbol{\sigma^2 = \frac{1}{N_C - 1} \sum_{R} \left| E_L^{\{\alpha\}}(R) - \bar{E}_u \right|^2} \tag{3.21}$$

where $\boldsymbol{E_L^{\{\alpha\}}(R)}$ is the local energy using Jastrow parameters $\boldsymbol{\{\alpha\}}$, and unreweighted energy is:

$$\boldsymbol{\bar{E}_u = \frac{1}{N_C} \sum_{R} Re\left(E_L^{\{\alpha\}}(R) \right)} \tag{3.22}$$

In addition to the variance minimization, another quantity which is useful for optimization is the mean average deviation (MAD):

$$\boldsymbol{MAD = \frac{1}{N_C} \sum_{R} \left| E_L^{\{\alpha\}}(R) - \bar{E}_u \right|} \tag{3.23}$$

MAD is rather more robust compared to variance minimization as it assigns less weight on the tails of the distribution which can lead to numerical errors in variance optimization. For much of our work, we used MAD minimization in the wavefunction optimization.

Although these two optimization methods may prove to be very useful, optimizing the energy may still be more desirable. It is very likely that energy minimization would yield lower energies compared to variance minimization. Energy optimized wavefunctions typically give better estimates in DMC energies, and also the variance of the DMC wave function is proportional to the energy difference between VMC and the ground state DMC energy[55]. On the other hand, energy minimization is typically computationally more costly that performing variance (or MAD) minimization especially for the cutoffs of the Jastrow parameters. Therefore, Jastrow parameter cutoffs can be optimized within variance minimization calculations, and then rest of the Jastrow parameters can be optimized using energy minimization.

3.5.3 Pseudopotentials

It is possible to perform all-electron calculations within the QMC method. However, especially for heavier atoms, this is largely inefficient. Core electrons are more closely packed than valence electrons, therefore treating these accurately means that more stringent sampling techniques or much smaller time steps need to be used to prevent population explosion errors in DMC calculations. Population explosion occurs due to accumulating too many number of walkers in the low local energy regions (due to importance sampling), hence biasing the calculations[43]. Since the core electrons are chemically less relevant in bonding, it is beneficial to replace them with some effective interaction without losing the accuracy.

There are multiple ways of constructing pseudopotentials (PP) for QMC or other correlated wave function calculations. The two most well-known and utilized approaches are norm-conserving[56] and energy consistent pseudopotentials[57]. Pseudopotential optimization is an on-going research with more than 50 years of development especially for DFT calculations. [56, 58–63] All pseudopotentials that are

generated for use in QMC calculations are originally optimized using HF, LDA or PBE methods, such that to our knowledge no pseudopotential that exists being optimized solely within QMC method. Norm conserving pseudopotentials reproduce the scattering properties of a free standing atom, such that beyond the core region the pseudo wave function and all-electron wave function have the exact same properties. For the energy consistent PP, only the excitation energies (e.g. from QMC or CCSD(T)) are compared to experimental energies to obtain the best functional form of the pseudo-wavefunction[57], constraints on the scattering properties are more relaxed. Although, there have been numerous efforts in generating pseudopotentials for low atomic mass atoms in QMC, pseudopotentials for higher mass atoms have been challening[64]. Norm conserving pseudopotentials are typically used in periodic electronic calculations, whereas effective core potentials (ECP) are made of localized basis functions, hence they are ideal for calculations in finite systems. There has been significant efforts in developing ECP type QMC pseudopotentials [65, 66] However, their conversion to planewave expansions may require significant care, as the conversion may result in "ghost states", i.e. having inaccurate number of nodes for an angular momentum component of the pseudopotential[67]

Pseudopotentials have two types of interactions where, in the short range each orbital imposes another potential on the ion core, while in the long range only the effective coulomb interaction is observed, $\hat{\boldsymbol{V}}^{\boldsymbol{loc}} = -\boldsymbol{Z_{eff}}/\boldsymbol{r}$, where the effective ion core $\boldsymbol{Z_{eff}} = \boldsymbol{Z} - \boldsymbol{Z_{pseudo}}$ is the original charge of the ion core minus the pseudized charge. The long range charge depends only on the position of the electron hence it is local, whereas the short range (non-local)potential is composed of Legendre polynomials for each radial operator:

$$\hat{\boldsymbol{V}}^{\boldsymbol{nl}}\boldsymbol{\Phi} = \sum_{\boldsymbol{l}} \boldsymbol{V_l}(\boldsymbol{r}) \Big[\sum_{\boldsymbol{m}} \int \boldsymbol{Y^*_{lm}}(\boldsymbol{\Omega'}) \boldsymbol{\Phi}(\boldsymbol{r}, \boldsymbol{\Omega'}) \boldsymbol{d\Omega'} \Big] \boldsymbol{Y_{lm}} \tag{3.24}$$

For a simple atom at the origin and electron $\boldsymbol{i}$, using QMC, this can be written as:

$$\hat{\boldsymbol{V}}^{\boldsymbol{nl,i}} = \sum_{\boldsymbol{i}} \boldsymbol{V^{ps}_{nl,i}}(\boldsymbol{r_i}) + \frac{\boldsymbol{2l+1}}{\boldsymbol{4\pi}} \int \boldsymbol{Y_l}[\boldsymbol{cos}(\boldsymbol{\theta'_i})] \frac{\boldsymbol{\Phi}(\boldsymbol{r_1} \ldots, \boldsymbol{r'_i}, \ldots, \boldsymbol{r_N})}{\boldsymbol{\Phi}(\boldsymbol{r_1} \ldots, \boldsymbol{r_i}, \ldots, \boldsymbol{r_N})} \boldsymbol{d\Omega'_i} \tag{3.25}$$

Non-local pseudopotentials require integration over the surface of a sphere, hence a quadrature rule is used, which integrates products of spherical harmonics up to a maximum angular momentum chosen. VMC calculations require much smaller quadrature grid as errors can cancel out over many number of steps, but DMC calculations require a larger grid density in comparison, due to the higher precision achieved in DMC calculations. Insufficient quadrature grid density can introduce bias in the calculations that is undesirable in DMC calculations . However, it is known that matrix elements of the non-local operator (for integrating the density) in imaginary time diffusion are negative [68], which creates a problem similar to the fermion sign problem. Two main approaches have been developed to avoid such an error: the localization approximation [69] and T-moves [70, 71]. The localization approximation is not very sensitive to the details of the trial wavefunction, but yields a non-variational DMC energy. Whereas the T-moves scheme preserves variational property of DMC calculations, although it typically requires smaller time steps.

3.5.4 Finite size errors

QMC calculations of condensed matter systems are subject to periodic boundary conditions, where the energies of finite sized simulation cell are used to extrapolate the energy of infinitely large system. However, performing these calculations in finite simulation cells means that 1) the calculations are to be performed in reciprocal space to retain the periodicity of the simulation cell 2) and a specific periodic interaction method needs to be developed to obtain periodic many body electron-electron interaction terms. The first constraint here means that Brillouin zone of the simulation must be sampled using sufficient number of "k-points" to ensure convergence in energies in each finite simulation cell. Reciprocal representation, which is also called as Bloch form $\boldsymbol{\Psi_k(r) = exp(ik \cdot r)u_k(r)}$, where $\boldsymbol{u_k}$ is the periodicity of the primitive cell and $\boldsymbol{k}$ is a reciprocal lattice vector of the simulation cell, or the k-points. Reciprocal simulation cell is integrated at every k-point, which are also called as twists, on the reciprocal grid. This property has already been used in periodic DFT calculations. In periodic DMC calculations, this process is called as twist averaging, which simply

means taking the average of energies evaluated at each k-point of the reciprocal grid. However, performing DMC calculations with a dense grid can be computationally expensive. A judicious choice of the k-point grid can facilitate much faster convergence of the energy with respect to the size of the grid density (e.g. Baldereschi points [72]). However, in cases where this is not possible, such as low symmetry simulation cells, performing twist averaging with relatively smaller number of twists may result in finite size errors, as mentioned. Hence, in order to eliminate these errors a twist averaging method with DFT based correction has been used:

$$\boldsymbol{E_{DMC}} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}}^{\boldsymbol{n}} [\boldsymbol{E_{DMC,i}} + (\boldsymbol{E_{DFT,\infty}} - \boldsymbol{E_{DFT,i}})] \tag{3.26}$$

where $\boldsymbol{E_{DMC}}$ is the final energy, $\boldsymbol{i}$ is the index of each twist on the k-poont grid, $\boldsymbol{n}$ is the total number of twists on the k-point grid, $\boldsymbol{E_{DMC,i}}$ is the DMC energy evaluated on twist $\boldsymbol{i}$, $\boldsymbol{E_{DFT,\infty}}$ is the fully converged DFT calculation with a large k-point grid and $\boldsymbol{E_{DFT,i}}$ is the DFT energy evaluated on the same twist $\boldsymbol{i}$. The eqn. 3.26, mostly accounts for the errors in the kinetic energy of the Hamiltonian which make up the more significant portion of the total finite size errors.

The second error accounts for the errors in many-body electron-electron interaction of the periodic cells. Electrons interact with a coulombic term that scales with $\mathbf{1/r}$. However, when particles are subjected to periodic interactions, it is not possible to use this simple term, because its periodic integration creates an alternating series which does not converge with the simulation cell size. Hence these interactions are calculated using Ewald interaction or model periodic Coulomb (MPC) interactions. The electron electron interaction energy can be separated into two: Hartree and the XC energies. Hartree energy is based only on the electron density, therefore it accounts for a classical, purely coulombic interaction like electron ion interaction. It has the periodicity of the simulation cell, therefore rapidly convergent with respect to system sizes. However, the energy contributions to the non-classical interactions between electrons, XC energy, converges rather slowly with respect to system size. XC interactions keep electrons further apart, beyond classical interactions (e.g. Pauli

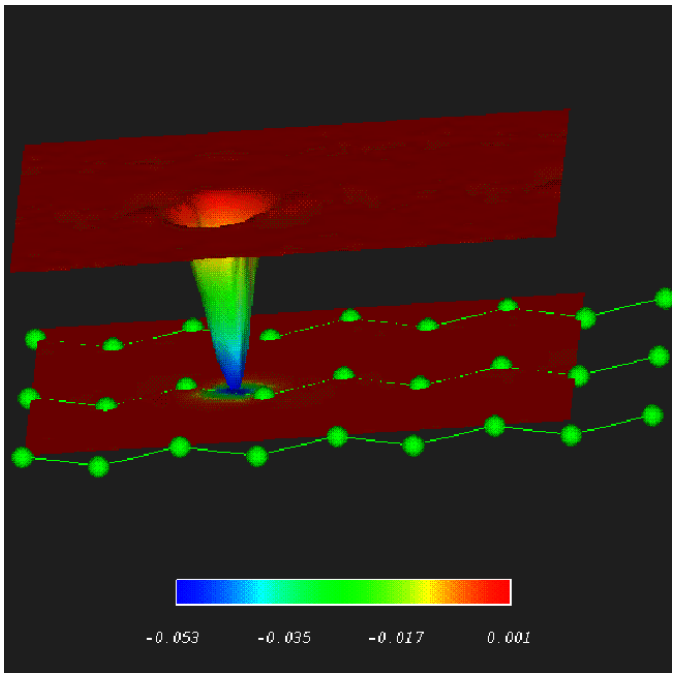


Figure 3-3: Surrounding every electron in a solid there is an exclusion zone, called the exchange-correlation (XC) hole, into which other electrons rarely venture. This is the XC hole around an electron near the centre of a bond in silicon. Taken from ref. [4]

exclusion principle), hence creating an exclusion zone around each electron and also decrease the total energy of the system, hence always negative. Therefore, an electron's contribution to XC energy can be explained using "XC hole". An example of its representation is given for an electron in the bonding region of bulk Silicon in Figure 3-3. Size of the exchange correlation hole may depend on the simulation cell size, hence its shape can be distorted. Hence, the interaction between periodic XC holes become larger than $\mathbf{1/r}$. In order to achieve an interaction that scales with $\mathbf{1/r}$, larger simulation cells must be used to extrapolate to the infinitely sized supercell. Otherwise, the order of error due to XC hole is typically lower polynomial order with respect to inverse size of the simulation cell. Hence, in order to obtain faster convergence with respect to the supercell size, each supercell should be selected to maximize the minimum periodic distances. Although, an extrapolation to the infinitely sized supercell would still be necessary, using MPC interaction for the integration of the XC energy can allevite this problem to certain degree, since it uses a correction term

on the Ewald interaction to have $1/r$ interaction between an electron and its XC hole.

When the finite size errors due to these two factors are controlled, the energies are extrapolated to the infinitely sized system limit. Extrapolation is performed using the energies at inverse the sizes of the supercells, $1/N$, where N is the supercell size. Using the energies obtained at multiple supercell sizes, a function, $F(1/N)$, is obtained which allows to obtain energies at $1/N = 0$, such that $N = \infty$. In our work, we consider $F(1/N)$ to be linear, and use least squares method to obtain the linear fit. This scheme can be prone to error if the slope of the extrapolation is large, since only a few points can be used in QMC calculations due to computational limitations. An example of this method is given in Figure 3-4, where extrapolation of pure DMC and DMC with DFT corrected energies (Eqn. 3.26) are compared.

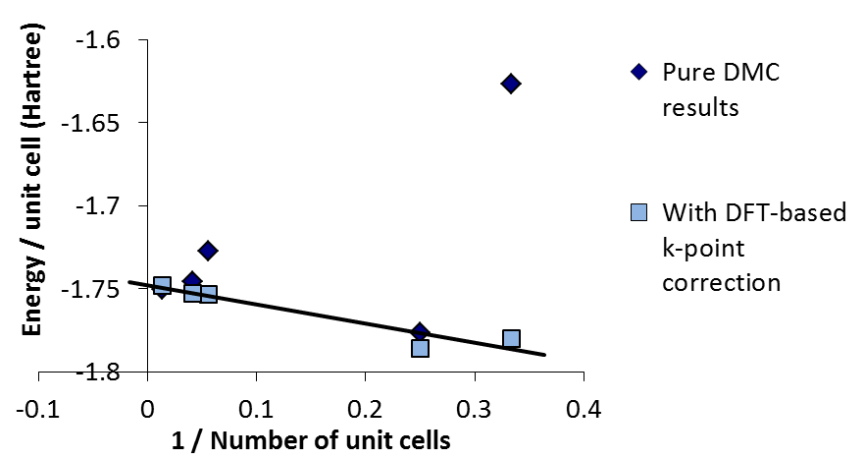


Figure 3-4: Comparison the the finite size extrapolation schemes with pure DMC energies and DMC energies with DFT based correction.

3.6 Summary

In this section, the VMC and DMC methods were described with specific discussions for several areas where application of DMC method is prone to error.

The VMC method is the intermediate step between DFT and DMC calculations where, typically, a single determinant from DFT calculations is converted into a many-

body electron wavefunction using Jastrow parameters. These Jastrow parameters are optimized using VMC calculations with several quantities as the optimization metric such as variance, mean absolute deviation of the total energy or the total energy itself. After the Jastrow parameters are optimized, VMC calculations can typically recover 90% of the correlation energy, however typically this does not provide sufficient accuracy beyond DFT calculations. Therefore DMC calculations are performed.

Fixed node DMC calculations typically recover more than 90% of the valence correlation energy, hence they can provide very accurate total energies and related properties. Although using fixed nodes introduces some additional error to the exact energies, typically these errors are cancelled when energy differences are required, hence fixed node DMC provides a very good approximation. In this perspective, pseudopotential and finite-size errors are much more serious errors for practical calculations. In the next chapter we discuss how DMC performs in material formation energies when one makes a fixed set of decisions regarding all the steps of VMC and DMC calculations.

Chapter 4

Investigation of High Throughput QMC Calculations

4.1 Introduction

With the rapid advances in the field of materials informatics the design and development of new materials can be accelerated through the creation and analysis of large databases of material properties. Rather than using trial and error via experimental methods to identify promising materials for new applications, these databases can provide important chemical and structural trends to provide more clear insight. For example, databases of material energies may be used to estimate the thermodynamic stability of a new material by enabling the rapid comparison of its calculated energy to the pre-calculated energies of all known possible decomposition products[11, 12]. This can be considered as the most basic step to identify whether a proposed structure is stable under some given set of conditions.

The exponential increase of the clock speed of computers has made it possible to populate material property databases using high-throughput calculations. However the quality of the predictions inevitably depends on the accuracy of the methods that are used to generate the databases. DFT[13, 14] within the GGA[36] approximation has been the most effective method to date, offering a favorable compromise between accuracy and speed of performing electronic structure calculations. Typically, DFT

methods benefit from error cancellations[73] which is the most effective for energy differences between structures where chemical environments of the atoms or molecules do not change significantly. In some cases, such as when the oxidation state or local chemical environment of an element changes, cancellation of errors can break down and the error in DFT / GGA can be significant[20–22]. Where these errors are not controllable within DFT methods, empirical corrections are added to obtain reliable thermodynamic or electronic properties[22].

In order to obtain more accurate formation energies than DFT, we focus on a rather straightforward way to implement QMC calculations with limited user effort, where the calculations can be performed in a high-throughput environment. The use of QMC in automated, high-throughput calculations is an emerging area of research. Shulenburger and Mattson[26] investigated equilibrium lattice constants and bulk moduli of several solids. Krogel[74] also developed an automated code for QMC workflows that were used to calculate formation energies and lattice constants of several binary oxides[75]. However, due to the computational challenges in performing DMC calculations, there has been very limited improvements to date. Unlike DFT, QMC is lacking a standard set of pseudopotentials where extensive benchmark calculations are performed using each set for different compounds. Only until recently, due to the computational expense of performing these calculations, almost each work in the literature mostly has been focusing on a single material or its polymorphs. Furthermore, wavefunction optimization and finite size extrapolations can be very labor intensive as performing these calculations require large number of user supplied settings to perform these calculations in an efficient way. We are particularly interested in formation energies of solid compunds, as performing DMC calculations on finite systems is relatively more simple, as it requires much less user controlled steps in the implementation. Although DMC calculations are much more expensive compared to DFT calculations, formation energies are typically reused frequently justifying the additional computational expense.

4.2 Test set

To evaluate the accuracy of DMC for calculating the formation energies of materials, it is necessary to have a set of materials for which highly accurate 0K experimental data exist. The Committee on Data for Science and Technology (CODATA)[76] has generated accurate thermodynamic data for 151 different substances, including 51 crystalline materials whose constituents are selected from different blocks of the periodic table which have similar properties, following the "Standard Order of Arrangement" procedure. From these data it is possible to calculate experimental 0 K enthalpies of formation for 26 different materials. For the present study, we have eliminated one water-containing compound, $CdSO_4 \cdot 8/3H_2O$ due to the difficulty in experimentally determining the structure of water inside the material. We have also eliminated three uranium-containing materials and one thorium-containing material due to the challenges in running DFT calculations for such heavy elements. The enthalpies of formation for the remaining 21 materials were used to benchmark the automated DMC calculations. To calculate these enthalpies of formation, it was necessary to use DMC to calculate the energies of 39 materials and molecules.

4.3 High-throughput framework for DMC

In order to automate the DMC calculation process in solids several decisions need to be made related to core electrons, trial wavefunctions, simulation steps and stop conditions. Furthermore, how the supercell selection is made and the number of supercells to be calculated for finite size error extrapolation also must be decided automatically without user interaction. In Figure 4-1, we show a schematic summary of the high-throughput framework for DMC calculations.

Initially an input structure is taken from the ICSD database, and supercells are generated using a general matrix form with increasing sizes. Then DFT calculations are performed on these structures to obtain single particle orbitals with norm-conserving pseudopotentials (NCPP). After this step, Jastrow parameters are opti-

Calculation/ Method	Software
DFT	PWSCF/NCPP[5] (FNCPP), CRYSTAL/ECPP[79]
QMC	CASINO[6]
Phonons	VASP-Phonopy[78]
Workflow	Custom scripts

Table 4.1: List of software that is used for each distinct parts of the calculations

mized using VMC calculations. These calculations are done on a single twist only. Since energies integrated on all twists (later with the DMC calculations) are averaged to obtain total energies, using a single set of Jastrow parameters on all twists yields statistically almost the same results as optimizing the Jastrow parameters on all twists separately. Similar calculations are repeated for every supercell. However, we find that using the pre-converged Jastrow parameters from smaller unit cell to obtain new optimized parameters in a supercell calculation provides faster convergence. Therefore we perform each supercell calculation in a sequence of large supercells, following the smaller ones. Finally, after the finite size extrapolation is performed to obtain a DMC total energy, $\boldsymbol{E_{DMC}(n = \infty)}$, where $\boldsymbol{n}$ is the supercell size, we add the zero point energy obtained from GGA calculations in VASP.

All high-throughput quantum Monte Carlo calculations were performed using the CASINO software package[6], and density functional theory calculations were performed using Quantum ESPRESSO (QE)[5] with norm-conserving pseudopotentials (NCPP) for all structures and energy consistent pseudopotentials (ECPP)[77] for a few select structures. In order to compare to the experimental formation energies found in the CODATA database, phonon contributions (zero point energies) must be calculated. Anisotropic contributions are neglected using a quasiharmonic approach calculated using DFT, with VASP[7]/Projector Augmented Wave (PAW) pseudopotentials[8] and Phonopy[78] software packages. All the software packages used to obtain the total formation energies are listed in Table 4.1.

Although Figure 4-1 shows the general workflow to perform automated QMC calculations, there are still remaining options that can be made to optimize the calculation process. In Table 4.3, we list all the finer details or the options available

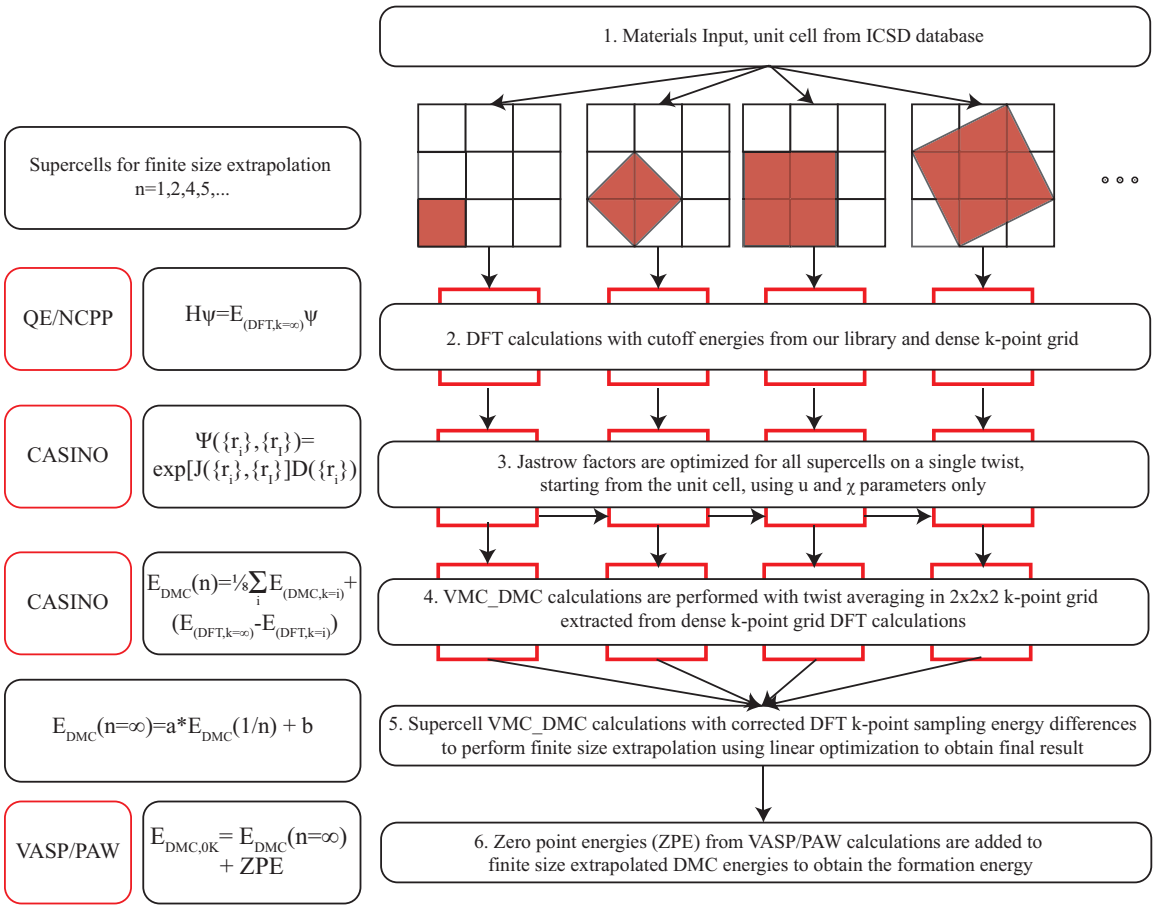


Figure 4-1: High throughput DMC calculation scheme. Supercell sizes are only shown as representative. For the equations on the left side of the figure, n corresponds to an arbitrary size of supercell used for finite size extrapolation, whereas k corresponds to the collection of grid points used for that calculation. In this respect, $\boldsymbol{k = i}$ stands for one of the eight reciprocal cell grid points used in the DMC integration and $\boldsymbol{E_{(DMC,k=i)}}$ is the DMC energy of a structure integrated at single k-point $\boldsymbol{i}$. The same notation has also been used for DFT calculations. QE[5]/NCPP, CASINO[6] and VASP[7]/PAW[8] indicate the software used in the calculation and the pseudopotential used. NCPP and PAW stands for norm-conserving pseudopotential and projector augmented wave methods respectively. $\boldsymbol{J[(r_i, r_j)]}$ is the Slater-Jastrow factor where $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$ represents electron and ion coordinates. $\boldsymbol{E_{DMC}(n = \infty)}$ is the finite size extrapolated DMC energy, which is obtained by performing linear fitting to $\boldsymbol{E_{DMC}(n)}$ values at the reciprocal of the supercell size, $\boldsymbol{1/n}$. Finally, $\boldsymbol{E_{DMC,0K}}$corresponds to the formation energy of the structure at 0 Kelvin.

within the workflow discussed in Figure 4-1. All DFT calculations were run allowing both up and down spins of the electrons in the system being optimized. Magnetic moment of an atom means the difference between number of up and down spin electrons on the atom. Therefore in order to break the spin symmetry of the starting charge density of the simulation cell, the initial magnetic moment was set to 70% up and 30% down spin on all atoms of a randomly selected element. The number of spin-up and spin-down electrons to be included in the QMC calculations was determined by sorting the DFT eigenvalues from lowest to highest and selecting the first N_{elect} orbitals, where N_{elect} is the total number of valence electrons in the simulation. For simulations in supercells, the net spin (# of up electrons - # of down electrons) was always rounded to a multiple of the number of primitive cells in the supercell. This resulted in a consistent net spin per unit cell across all supercell sizes. Of the compounds included in our test set, only molecular oxygen and $CuSO_4$ were found to have non-zero net spin. DFT calculations were performed using the same norm-conserving pseudopotentials (NCPP) used for DMC. Detailed information regarding pseudopotential selection can be found in section 4.4.2.

We used a 2x2x2 k-point grid for our calculations to maximize the sampling of the Brillouin zone without using complex arithmetic. However, in order to eliminate any potential undersampling of the Brillouin zone, we used the method similar to that proposed by Rajagopal et.al [80]. DMC energy calculated using a 2x2x2 k-point grid is corrected by the difference between a well-converged DFT energy and the DFT energy calculated using a 2x2x2 grid. The well-converged DFT energies were calculated using a k-point grid with density of at least 8000 k-points per $Å^{-3}$. However, the 2x2x2 k-point grid is generally not sufficient to ensure reliable convergence of the electronic self-consistency loop in DFT, therefore the DFT energy for the 2x2x2 grid was calculated by extracting the 2x2x2 subset of k-points which is used to perform twist averaging in DMC calculations.

The parameters in the Jastrow factor were optimized using VMC calculations using standard routines available in CASINO. In each of these calculations, the electronic configurations were propagated for 10,000 equilibration steps, and then an

additional 150,000 steps were run to generate a sample of 10,000 random configurations. The VMC timestep is internally optimized at every step during equilibration, aiming for 50% acceptance ratio in the Metropolis algorithm. The high acceptance ratios would mean that time steps are small, therefore electrons only move to very short distances and causing serial correlation among the results. In this case, the calculations would need to be run for longer number of steps to reduce the correlation and sample the configuration space more thoroughly. If very large time steps are used, each configuration sampled would be significantly different from each other, causing lower probability configurations to be sampled more frequently, hence reducing the Metropolis acceptance rates. Therefore keeping the acceptance ratio around 50% is optimal balance to sample configuration space more efficiently. Furthermore, the samples were taken every 15 steps to reduce serial correlation.

For large supercell sizes, we found that occasionally the optimization of the Jastrow factor would converge poorly (e.g. resulting in an anomalously high mean absolute deviation of the local energy), while for the smallest supercell sizes for each material the Jastrow factor converged reliably in all cases. Reliable convergence of the Jastrow factor at larger supercell sizes was achieved by initializing the Jastrow parameters for each supercell with the converged Jastrow parameters for the next-smallest supercell. As already discussed in section It was found that the same Jastrow parameters worked nearly equally well at all k-points, so in our automated formulation the Jastrow parameters were optimized at a single k-point and used for the remaining seven k-points on the 2x2x2 k-point grid.

4.4 Results

4.4.1 Tests with Rappe-Bennett PP

In Figure 4-2, we show the results of our DMC calculations using the Rappe-Bennett pseudopotentials[82] (RB-PP) for all compounds. We use these results as our reference and first attempt to utilize our DMC recipe. Up to Hg_2Cl_2 on the x-axis,

Element	Options
DFT	• Spin polarized GGA (PBE) • 0.02 Ry smearing, 8000 k-points per $Å^{-3}$ • 0.7 initial magnetic moment on an element • Supercell spin is rounded to the multiple of number of primitive cells • Even number of electrons in the unit cell
Basis Sets	**1. Plane waves** ✓ Natural for periodic systems ✓ Single parameter convergence × Non-local, expensive to evaluate **2. Blips (splines)** ✓ Local, cheap to evaluate, flexible × Requires very large memory
Pseudopotentials	**1. FHI-PP[81]** ✓ Covers most of the periodic table × Mostly soft core **2. Rappe-Bennett PP[82]** ✓ An alternative to ABINIT × Only available for select elements **3. BFD-PP[77]** ✓ Energy consistent optimization ✓ Designed for correlated wave function based calculations **4. OPT-PP[64]** ✓ Optimized for QMC calculations × Currently only available for transition metals × Requires very high cutoff energy (280 Ry)
Finite size effects	• Supercells maximize the minimum image distances • Ceperley's formula[83] is applied to eliminate finite size effects $\boldsymbol{E'_{DMC}} = \boldsymbol{E_{DMC,k_i}} + [\boldsymbol{E_{DFT,k_\infty}} - \boldsymbol{E_{DFT,k_i}}]$
Twist averaging	• 2x2x2 grid • 2500 statistics accumulation at each point
Wave function Optimization	• 3-body terms not included, only $\boldsymbol{u}$, $\boldsymbol{\chi}$ and $\boldsymbol{p}$ in eq. 3.17. • All twists use the same Jastrow factor, $\lvert\boldsymbol{J}\rvert$ • Supercell $\lvert\boldsymbol{J}\rvert$ are initialized from smaller unit cells • MAD minimization, eq. 3.23, is used to optimize $\lvert\boldsymbol{J}\rvert$

Table 4.2: List of options and settings that are made for the High throughput DMC method we develop. Bullet points, •, show the various fixed options that are applied without any changes for all calculations. Enumerated lists show the various options for a given element. Check and cross marks, ✓and ×, are for informative purposes only, listing pros and cons of choosing one option for an element in the calculation that is left up to the user. However, for basis sets only, Blips are always preferred over plane waves as blips have more favorable scaling in evaluating the Slater determinant.

DMC calculations are able to provide chemical accuracy, meaning that the results are within 1kcal/mol/atom (4.12 kJ/mol/atom) of the experimental formation energies. These results are substantially more accurate than the corresponding DFT with errors ranging from 5-60 kJ/mol/atom for these materials. Up to MgF_2, in Figure 4-2, DMC calculations enabled better accuracy compared to QE/NCPP calculations. However, for HgO, AgCl, Hg_2SO_4 and ZnO, DMC formation energy errors are larger than their QE/NCPP counterparts and in some cases the DMC error is extremely large.

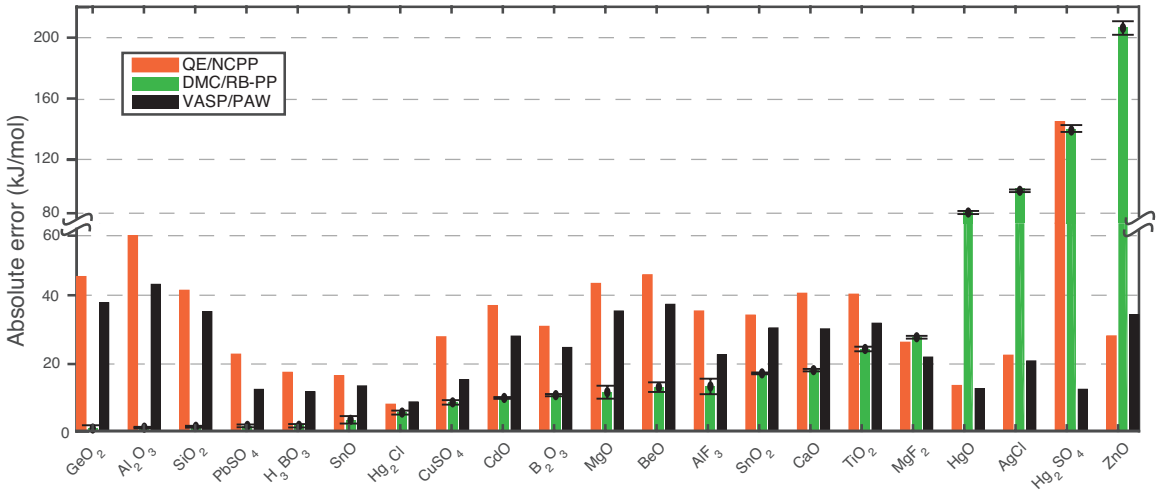


Figure 4-2: Absolute error per atom with respect to experimental formation energies for the compounds in the benchmark set using RB-PP. The QE/NCPP results are shown with orange and DMC results are shown with blue bar histograms. Black error bar lines on DMC/RB-PP results represent the statistical error that results from the Monte Carlo algorithm. VASP/PAW results are shown with the black bar histograms. On the y-axis a break is placed between 60-80 kJ/mol and upper half of the y-axis has larger intervals for better representation.

4.4.2 Tests using multiple PP for problematic cases

In order to understand these cases with large DMC errors we investigated several possible causes. [84] First was the use of asymmetric Casula T-move branching factors, which make the DMC energy variational and prevent population explosion errors when non-local pseudopotentials are used. The T-move scheme may also increase the localization error due to the pseudopotentials and yield a slightly larger time-step

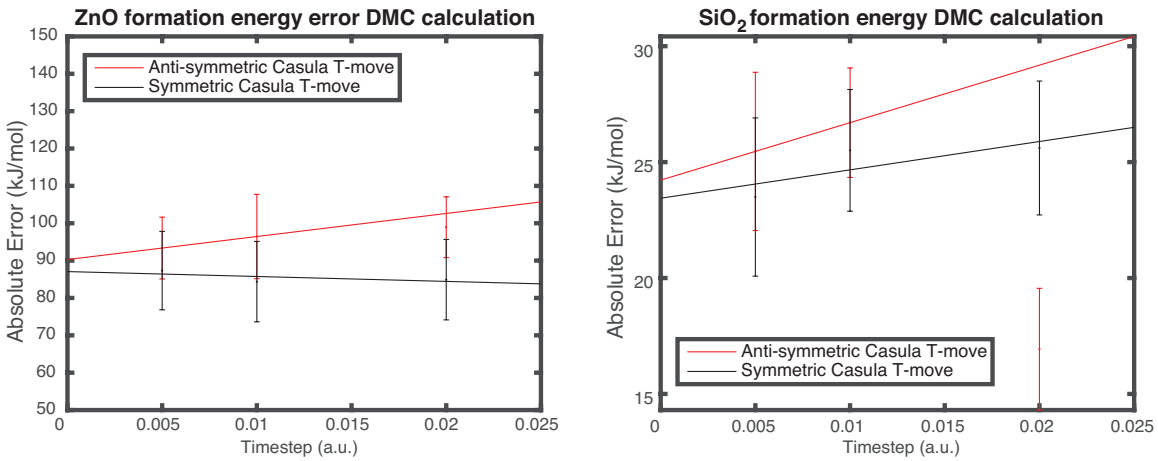


Figure 4-3: Timestep extrapolation for the formation energy of ZnO and SiO_2 using RB-PP at 3 different time steps, 0.005, 0.01 and 0.02 a.u. Two different implementation of the Casula T-move scheme has been applied. Calculations are performed on ZnO and SiO_2 unitcells. The Y-axis represents the errors in the formation energies per atom with respect to the experimental formation energy. For SiO_2 antisymmetric Casula T-move scheme is extrapolated at time step of 0.005 and 0.01 a.u. due to possible error using this timestep.

bias. An alternative is to use symmetric T-move branching factors, which have been shown to decrease the time-step bias further to some extent[70]. To evaluate whether large errors originate from timestep extrapolation or our choice of T-move scheme, we compared the performance of the two Casula T-move schemes for SiO_2 and ZnO structures and perform time-step extrapolation to zero for total energies and formation energies of each compound. We show that the large error in the formation energy is not a result of the difference in Casula T-move implementation or choice of the time step, as the difference in formation energies at different time steps is within the error bar of the DMC results at 0.01 a.u as shown in Figures 4-3.

Secondly, we investigate the pseudopotential errors in the Hamiltonian, as they are found to be rather large and non-systematic especially for heavy elements[85]. Pseudopotentials can also affect the nodal surface and shape of the trial wavefunction, leading to second order interactions between two types of errors, which makes it challenging to characterize and isolate the source of error being either the pseudopotential itself or the nodal surface of the trial wavefunction. Therefore, to separate between errors in finite size and the pseudopotential, we initially performed a litera-

ture search for possibly more accurate pseudopotentials for the compounds that have large formation error when RB-PP is used in DMC calculations and re-calculated the formation energies of select compounds using these pseudopotentials.

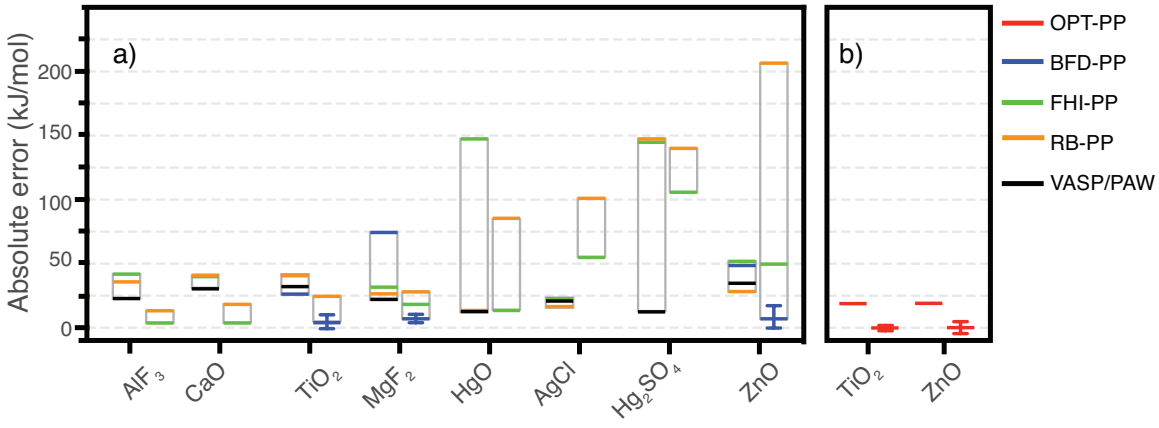


Figure 4-4: Absolute error per atom for the compounds which are identified to be problematic when RB-PP is used. Figures a) and b) use the same scale on the y-axis. However, they are separated as all the calculations in a) use PBE method in DFT and orbital generation for QMC, whereas calculations in b) use LDA. Within each figure there are two groups of data for each compound, each enclosed with a box if results of more than one pseudopotentials are compared. The first group, on the left, represents the DFT calculations, whereas the second group represents DMC calculations. Each color given in the legend shows the pseudopotential used in performing respective calculations. Error bars in DMC calculations are smaller than thickness of the associated lines, if not shown explicitly. Tabulated representation can be found in the SI.

Among the test set materials considered in this work, compounds that include Zn, Hg, Ag, F and Ti atoms, tend to have the largest discrepancy between DMC and experimental results (Fig. 4-2). To test the BFD-PP and OPT-PP pseudopotentials, we needed to make slight changes to our recipe. DFT calculations using BFD-PP were performed using the CRYSTAL code[79] at the PBE level using a double zeta Gaussian basis. As OPT-PP are available for only first row transition metals and they are optimized for the local density approximation (LDA)[86], whenever there is a p-block element in a compound to be simulated, we combine it with Fritz-Haber Institute pseudopotentials optimized at LDA level and perform DFT calculations to prepare trial wavefunctions using LDA. Therefore, only $\mathbf{TiO_2}$ and ZnO in our test set could be simulated using OPT-PP.

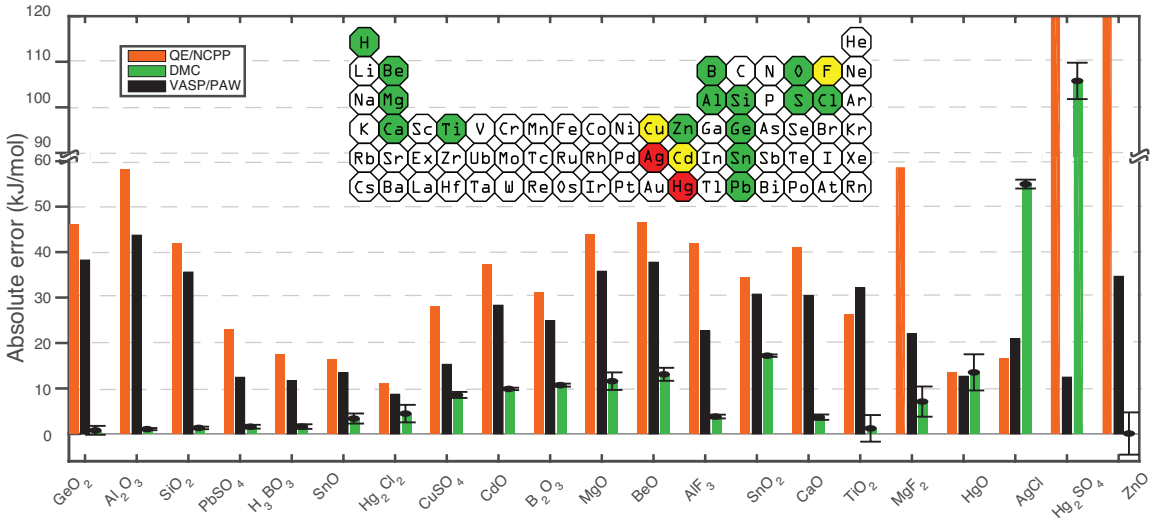


Figure 4-5: Absolute error per atom for the benchmark set using RB-PP for all atoms except for the compounds containing F, Ca, Ti, Hg, Ag and Zn. For the compounds which contain these atoms, results here are taken from the best DMC calculation in Figure 4-4. Bar histograms are represented in the same way as Figure 4-2. The periodic table in the inset represents the atoms that perform with desirable accuracy in green, with slightly worse accuracy in yellow and atoms whose pseudopotentials need improvement in red. On the y-axis a break is placed between 60-90 kJ/mol for better representation. Similarly, QE/NCPP values for Hg_2SO_4 and ZnO are not shown as the graph is truncated at 120 kJ/mol. These values are 147.41 and 237.78 kJ/mol respectively. Tabulated representation can be found in the SI.

We initially calculated the formation energy of given compounds using RB-PP as shown in Figure 4-2. Then in Figure 4-4, for the problematic cases of Figure 4-2, we show that using FHI-PP leads to improved DMC formation enthalpies for most compounds compared to using RB-PP. For example, AlF_3 and CaO have errors of less than 5 kJ/mol atoms in the DMC/FHI-PP calculations, compared to 13.3 $\pm$ 2.3 and 18.1 $\pm$ 0.3 kJ/mol atoms for the DMC/RB-PP results. For transition metal containing compounds and MgF_2 however, still DMC/FHI-PP results are not significantly better than VASP/PAW calculations. For TiO_2, both DMC/OPT-PP and DMC/BFD-PP perform substantially better compared to DMC/RB-PP and VASP/PAW. For MgF_2, we could only perform DMC/FHI, DMC/BFD and DMC/RB calculations, since the pseudopotential for Mg does not exist in the OPT-PP set. For ZnO however, DMC calculations substantially improve when OPT-PP and BFD-PP are used, resulting in errors of less than 7.1 kJ/mol per atom. Eliminating finite size errors in metal-

lic materials can be especially challenging due to the complex shape of the Fermi surface, which may require denser k-point sampling. However, the accurate results obtained for these compounds, using BFD and OPT-PP show that for the cases considered here, pseudopotentials are the largest source of error and our recipe yields transferable performance given the use of suitable pseudopotentials.

4.5 Conclusions

For a test set of 21 compounds with experimentally known formation energies, chemical accuracy was obtained with our automated DMC recipe for 11 structures. DFT calculations using either the QE/NCPP or VASP/DFT methods were not able to provide chemical accuracy in any of the structures investigated. Overall, for 18 of the 21 compounds, our DMC recipe provides results with significantly improved accuracy compared to VASP/PAW. We find that for the three remaining cases, formation energies of AgCl, HgO and Hg_2SO_4, DMC performs worse than VASP/PAW. Among these, the DMC errors are anomalously high for two of compounds: AgCl and Hg2SO4. Because there are currently no BFD or OPT pseudopotentials available for Ag or Hg, were were only able to generate results for these compounds using the RB and FHI pseudopotentials. Based on our tests on ZnO, TiO2, and MgF2, we believe the RB and FHI pseudopotentials for Ag and Hg are likely the source of the anomalously high error.

Although our results show that performing DMC calculations in a pre-set scheme on a range of materials can provide significantly improved accuracy over DFT calculations, pseudopotential errors are non-systematic in DMC and can, for a small number of cases, lead to serious errors. Given that an improvement over DFT was obtained in 85% of the cases tested here, the simple, automated approach we present may be sufficient for a host of applications. We find that pseudopotentials that perform accurately in DFT do not necessarily perform as well in QMC. In Table 4.5, we show the differences between the tested pseudopotentials for the two transition metals we investigated in detail, Zn and Ti. Both BFD-PP and OPT-PP use a Ne

	OPT-PP	BFD-PP	FHI-PP	RB-PP
Number of valence electrons (Z_{eff})				
Zn	20	20	12	12
Ti	12	12	4	12
d-orbital core radii (in a.u.)				
Zn	0.80	1.16	2.37	1.97
Ti	0.80	1.60	2.71	1.70
Local and highest l channels				
Zn	p,d	d,d	s,d	s,d
Ti	p,d	d,d	p,d	s,d

Table 4.3: Comparison between all investigated pseudopotentials for Zn and Ti for valence electrons, d-orbital core radii, local and highest angular momentum ($\boldsymbol{l}$) channels

core for first row transition metals as it was found that semicore effects can be rather crucial for these compounds[64]. These PP also have relatively smaller core radii on the d orbitals. Additionally, BFD-PP uses an energy consistent scheme rather than norm-conserving, as they found that energy consistent pseudopotentials can give more accurate results in MP2, CCSD(T) and DMC calculations[87]. On the other hand, OPT-PP uses very small non-local radii, < 1 a.u., to increase its transferability. It has been suggested that inclusion of higher angular momentum channels and using them as the local channel for a pseudopotential can lead to improved energies in QMC calculations[88]. However, in our comparisons between different pseudopotentials, we see the smaller core radius and larger number of valence electrons as an important indicator for more accurate pseudopotentials.

When compared to existing high throughput recipes using DFT calculations, advantages of using QMC over these methods can be better understood. However, it must be kept in mind that the high-throughput DFT recipes make use of empirical fitting schemes for elemental energies, whereas our QMC calculations use no empirical fitting parameters. Saal et.al. [11] compared the performance of High throughput DFT recipes over two of the existing databases: Materials Project [20, 84] and Open Quantum Materials Database (OQMD) [11]. The formation energies of 1386 compounds found in the Materials Project are compared with respect to experimental

energies when two recipes are employed yielding 0.133 eV/atom and 0.108 eV/atom mean absolute errors (MAE). Within OQMD recipe, when only binary compounds are investigated, the resulting MAE is 0.119 eV/atom. In comparison, when our QMC recipe yields a MAE of 0.058(6) eV/atom, but then the compounds that rely on Ag and Hg pseudopotentials are excluded, (AgCl, Hg_2SO_4 and HgO) our QMC recipe yields a MAE of 0.028(5) eV/atom, whereas pure DFT-GGA calculations we performed with VASP/PAW, yields a MAE of 0.276 ev/atom over the same set.

Chapter 5

QMC applied to molecules: accuracy of DFT calculations in electrocyclization reactions towards Solar Thermal Fuel applications

5.1 Introduction

In Chapter 4, an automated recipe for performing QMC calculations in periodic solids has been discussed. However, QMC method can also be used to calculate energies and other relevant properties in molecules. Performing calculations on a gas phase molecule at 0 Kelvin represents a finite system where calculating periodic interactions is not required. Therefore, for QMC calculations, integration at a single k-point is sufficient to obtain total energies for a system. This would mean that 1) finite size errors do not exist in while calculating the properties of molecules (or atomic clusters) and 2) twist averaging is also not required since single k-point is enough to calculate the total energy. In finite systems, QMC is also known to yield very accurate energies[89], such that in cases where DFT methods yield inconsistent energies, QMC can be used as a benchmark method to assess the accuracy of each DFT method.

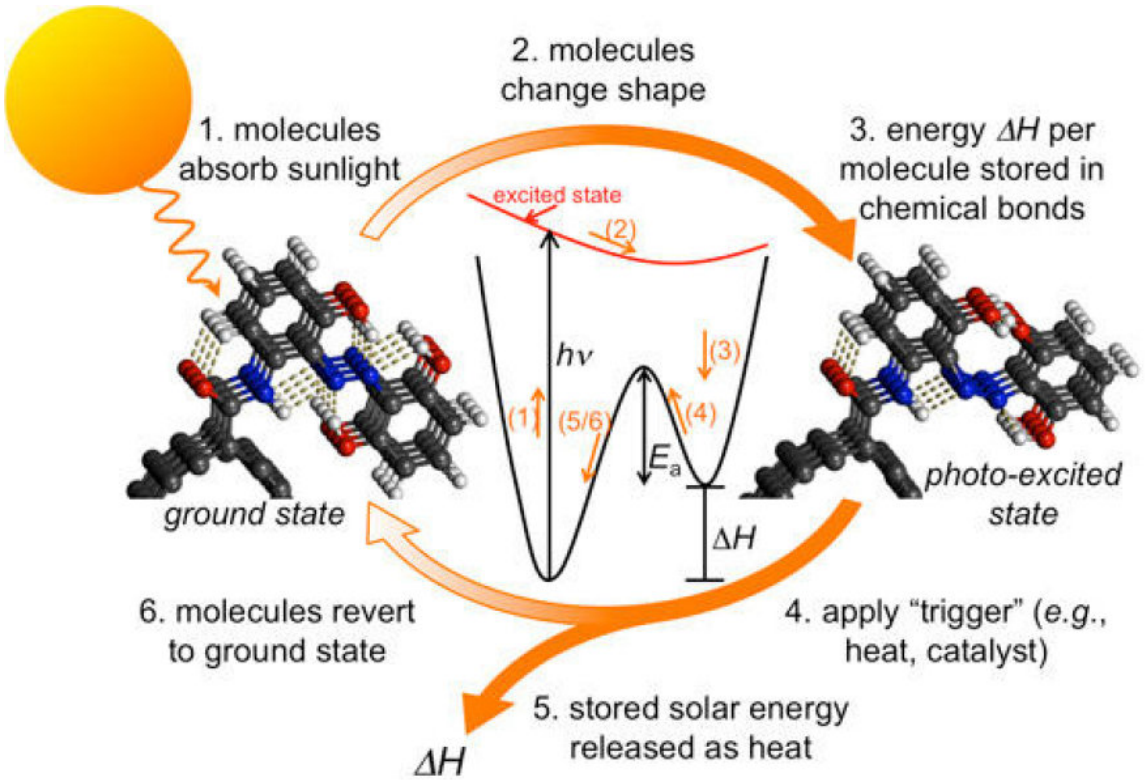


Figure 5-1: Azobenzene as a solar thermal fuel. Full cycle of photoexcitation and energy release in the form of heat is given with the intermediate steps along the reactions. Taken from

Solar thermal fuels are the materials that can convert between at least two states where conversion from one state to another is initiated via absorbing sunlight. Then, the absorbed solar energy is stored in the form of chemical energy, which later can be transformed into heat when needed. that can convert sunlight into chemical energy, and then release this chemical energy in the form of heat, when needed. Storing the solar energy and then releasing the chemical energy makes up one cycle, which can ideally be repeated multiple times. Therefore, molecular photoswitches can be ideal to perform these energy conversions. An example of a photoswithch, azobenzene[90–95], is given in Figure 5-1, where ground state *cis* conformation is converted into *trans* conformation, while storing 0.34 eV energy due the chemical bond transformation. However, azobenzene is not the only example performing this task, other molecules such as norbornadiene[96–98] has also recently gained renewed interest as a renew-

able energy storage material, although nearly any photoswitchable molecule can be considered as a potential solar thermal storage material, as long as it possesses large isomerization enthalpy, small volume (or molecular weight), well separated absorption bands for the isomers, and high fatigue resistance for photoswitching[90]. Among the well studied photoswitches[99], 1,8a-Dihydroazulene-1,1-dicarbonitrile (DHA, **1a**) derivatives have also shown great promise for its application as a solar thermal energy storage material. As given in Figure 5-2, DHA, **1a**, undergoes ring opening reaction to form cis-vinylheptafulvalene (cis-VHF, **1c**). Then cis-VHF converts to trans-vinylheptafulvalene (trans-VHF, **1e**) via cis-trans isomerization under thermal conditions. One particular reason why the DHA/VHF couple extensively studied is its one-way photochromism. The conversion from DHA to trans-VHF can be photoinduced while the reverse reaction proceeds only with thermal activation[100]. This strictly one-way photochromism is achieved through a conical intersection near the cis-VHF conformation between the ground and first excited states [101] and it allows a high fatigue resistance for photocyclability[102], meaning that the conversion between DHA and trans-VHF can be repeated with a reduced side product formation. The forward, ring opening, reaction can occur with a high quantum yield measured from 0.1 to 0.6 at room temperature, whereas the maximum optical absorption is around 350 nm for DHA, corresponding to a high intensity solar flux[103]. In comparison, trans-VHF has the maximum optical absorption around 470 nm, although the photochemical back reaction is inactive[100]. Experimental[100, 103–108] and computational[109–112] efforts have focused on tuning switching properties of DHA/trans-VHF system with the attachment of electron donating and withdrawing groups. Although the absorption properties of both isomers, reverse activation energies of the metastable isomers and fatigue resistances of DHA derivatives have been studied experimentally, to our knowledge there has been no effort to characterize their isomerization enthalpies using experimental techniques.

DFT is widely used to predict the optical and thermodynamical properties of molecular photoswitches. Using DFT and implicit solvation method, Olsen et. al. [110] showed that using polar solvents, for example acetonitrile, increases the stability

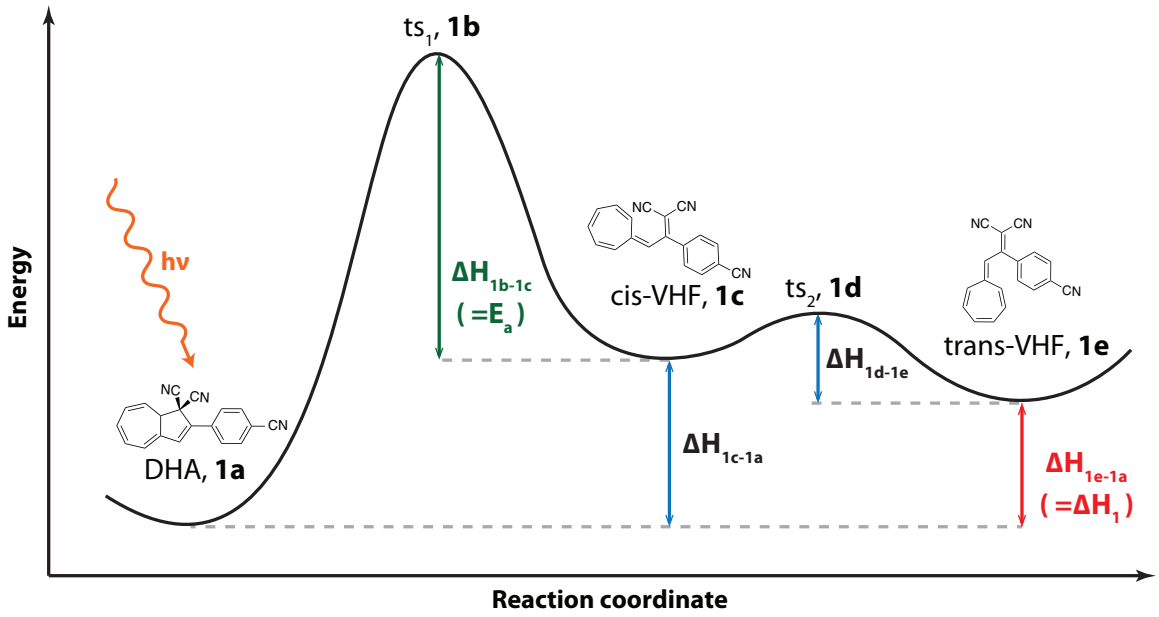


Figure 5-2: Ground state potential energy surface for the DHA/VHF photoswitch system. Ground state, metastable state and transition state structures are enumerated from **1a** to **1e**. $\mathbf{\Delta H_1}$ is the energy difference between the ground state and lowest energy metastable state, DHA and trans-VHF. $\mathbf{\Delta E}_a$ is the back reaction activation barrier.

of trans-VHF, compared to DHA[110], hence decreasing the isomerization enthalpy. Therefore, in vacuum, the calculated isomerization enthalpy should be even higher than in solvent conditions. Their analysis showed that, among the DFT functionals they investigated, M06-2X[114] is the only functional that provides qualitatively correct results. They also showed that the B3LYP[38] functional yields a qualitatively inaccurate ordering, even in vacuum, as also noted in other work related to different DHA derivatives[105]. B3LYP calculations in our benchmark fo DHA/VHF isomerization enthalpies, Table 5.1, confirm their findings. Furthermore, we find that the PBE[16] functional predicts DHA and trans-VHF to be nearly iso-energetic, further calling into question the accuracy of DFT calculations for these compounds. Beyond determining which functionals provide quantitative or qualitative accuracy, it is also of interest to understand why a given functional is predictive or not for these isomerization enthalpies. For benchmarking energies of finite systems in the absence of experimental results, we already mentioned that QMC is a powerful alternative, but other correlated wavefunction based approaches such as CCSD(T), which is also

Table 5.1: Energy differences in vacuum at 0 Kelvin for the structures on the ground state potential energy surface of DHA/VHF isomerization.

	ΔH_{1a-1c}	ΔH_{1c-1e}	$\Delta H_{1a-1e} (= \Delta H_1)$	ΔH_{1b-1c}
HF	0.45	-0.11	0.34	1.48
LDA	0.37	-0.03	0.34	0.83
PBE	0.09	-0.07	0.02	0.96
B3LYP	0.02	-0.07	-0.05	1.11
PBE0	0.33	-0.07	0.26	1.10
M06	0.24	-0.06	0.18	1.15
CAM-B3LYP	0.33	-0.06	0.27	1.25
wB97XD	0.43	-0.07	0.36	1.24
DMC	0.41(6)	-0.09(6)	0.32(6)	1.31(6)
Exp.	-	-	-	1.39[113]

known as the gold standard method in the quantum chemistry community, are also feasible.

Therefore, in this chapter we describe our work on unerstanding this system, specifically:

- We investigate the photoisomerization of 21 DHA derivatives, using QMC and CCSD(T) calculations to benchmark and analyze several DFT approximations.

- We find that the DFT errors in ring opening isomerization energies of DHA/trans-VHF are correlated to ring opening reactions of simpler compounds such as cyclobutene and cyclohexene.

- Finally, we investigate the breakdown of exchange and correlation energies on the reaction path of cyclobutene and cyclohexene to show that incorrect description of the GGA exchange in B3LYP and the description of correlation in PBE and TPSSH[37] functionals are largely responsible for the inaccurate predictions obtained from these methods compared to exact calculation of correlation energy in QMC calculations.

5.2 Computational Methods

In chapter 3, we performed DFT calculations using plane-wave DFT codes. Wavefunction has the periodicity of the simulation cell, therefore using the Block theorem, it is straight forward to think of the wave function as an expansion of plane-waves. However, for finite systems, such periodicity is not observed, hence local basis functions, such as gaussian expansions, can be used. Therefore, for the systems discussed in this chapter, DFT calculations are performed as implemented in Gaussian 09[115] code using all electron gaussian basis sets. All electron basis sets would mean that unlike the pseudopotentials discussed in Chapter 3, core electrons are not pseudized. We report DFT, QMC and CCSD(T) energies at 0K in vacuum, without zero point energy contributions. Geometries along the potential energy surface, shown in Figure 5-2, have been optimized starting from the experimental coordinates obtained from X-ray diffraction[116], then a using 6-31+G* basis and subsequent single point calculations are performed with a 6-311++G** level Pople basis sets. A transition state search to optimize the geometries of transition state compounds (e.g. **1b** and **1d** in Figure 5-2) is performed using Berny algorithm[117]. Intrinsic reaction coordinate calculations are performed, starting from the transition states, to confirm that the reaction path starting from the transition state leads to both reactants and products.

CCSD(T) calculations are performed with the frozen core approximation. Frozen core approximation ensures that all core electrons are doubly occupied. Correlation consistent Dunning type aug-cc-pvdz basis sets are used for the CCSD(T) calculations in this chapter. Aug- stands for being augmented where extra diffuse basis sets are included for each atom to capture weak interactions, cc stands for being correlation consistent, -p for polarized and vdz means valence only double zeta basis. Larger basis sets with vtz, vqz and so on (triple zeta and quadruple zeta respectively) would mean that larger shells of valence orbitals (polarizations) are incorporated to the basis set. Therefore, with each basis set where extra sets of valence orbitals are included in the basis set, correlation consistent basis sets ensure that infinite basis set limit can be attained. However, for many systems of interest in this chapter, performing

CCSD(T) calculations involving beyond double zeta basis has been computationally too expensive, hence an MP2[118] aided extrapolation recipe by Truhlar[119] is used. CCSD(T) and MP2 calculations are performed with the NWCHEM[120] package.

Before performing any DMC calculations, initially, molecular geometries, $\{\boldsymbol{R_i}\}$, are optimized at the B3LYP level, as evaluating interatomic forces via DMC is computationally challenging[121]. Compared to the solids discussed in Chapter 4, an accurate nodal surface of the wavefunction in finite systems can be of multiconfigurational nature, hence it could benefit from larger number of Slater determinants. However, this would require much larger computational resources, which scales with number of determinants, and the level of truncation would need to be optimized at each calculation. It is clear that a trial wavefunction prepared using CI method with multiple determinants can be describe the true nodal surface of the wavefunction better than a single determinant obtained from HF method, but comparison of CI methods to DFT in generating accurate trial wavefunctions may not be as straightforward[122]. Per et.al [122] showed that in finite systems, trial wavefunctions, $\boldsymbol{\Psi_T(R)}$, prepared using LDA and PBE methods can yield much lower variances in the total energies compared to using $\boldsymbol{\Psi_T(R)}$ from HF. Therefore, we generate $\boldsymbol{\Psi_T(R)}$ using PBE approximation with BFD-PP, and benchmark the DMC results against CCSD(T) calculations to check for discrepancies due to the fixed node error. DMC calculations are performed using a 0.01 a.u time step and Casula T-move scheme with symmetric branching algorithm. Target precision on the error bar of the calculations have been set to 0.001 Ha (0.027 eV) for each calculation, such that each statistical DMC run typically takes slightly longer than 50000 steps with 2400 walkers for DHA derivatives.

5.3 Results and Discussion

In Table 5.1, we show the results of the first set of benchmark calculations we performed on the DHA molecule given in Figure 5-2. The structures **1a-e** are ordered from left to the right on the reaction coordinate in Figure 5-2. For practical applications, the most significant quantities are the trans-VHF/DHA isomerization en-

thalpy, $\mathbf{\Delta H_{1e-1a}} = \boldsymbol{H_{1e}} - \boldsymbol{H_{1a}}$ (=$\mathbf{\Delta H_1}$), and the back reaction barrier $\mathbf{\Delta H_{1b-1c}}$ ($\mathbf{\Delta E_a}$). According to the experimental results, $\mathbf{\Delta H_1}$ is inaccurately predicted by B3LYP and PBE functionals since the B3LYP result yields negative $\mathbf{\Delta H_1}$ and PBE result yields DHA and trans-VHF almost iso-energetic. Using DFT and implicit solvation method, Olsen et.al. [110] showed that with increasing solvent polarity, the trans-VHF molecule becomes more stable, decreasing the isomerization enthalpy between DHA and trans-VHF. Experimentally, under acetonitrile solvent, DHA is found to be more stable compared to trans-VHF, hence under vacuum conditions, the predicted DHA/trans-VHF isomerization enthalpy should be larger than in solvent conditions. Given the significant errors of B3LYP and PBE functionals, even in vacuum conditions, it can be useful to investigate the energy differences between the isomers on the ground state potential energy surface to understand which chemical change during the isomerization leads to variations among DFT results.

As given in Figure 5-2, we mentioned that DHA undergoes two chemical reactions to convert to trans-VHF: ring-opening and cis-trans isomerizations. Therefore, in Table 5.1, we investigate the stability of DHA relative to cis-VHF, $\mathbf{\Delta H_{1a-1c}}$, stability of cis-VHF relative to trans-VHF, $\mathbf{\Delta H_{1c-1e}}$ and also the back reaction barrier from cis-VHF to DHA, $\mathbf{\Delta H_{1b-1c}}$. We show that the variation among the DFT methods is less than 0.08 eV when $\mathbf{\Delta H_{1c-1e}}$ is investigated. However, when HF and LDA results are disregarded, this variation reduces to 0.01 eV. All the DFT functionals yield trans-VHF more stable compared to cis-VHF. However, for $\mathbf{\Delta H_{1a-1c}}$ the DFT results vary from 0.02 to 0.45 eV. When HF and LDA results are disregarded, this outcome does not change significantly, as the range becomes 0.02 to 0.43 eV. This suggests that for DHA/VHF isomerization, DFT functionals yield larger variations during the ring opening reaction rather than cis-trans isomerization of VHF.

Although benchmarking different DFT calculations for a particular reaction, DHA/ VHF isomerization in this case, if this system is of interest, it is also important to identify how the DFT errors present in one chemical reaction may translate (or not) into another one. We are especially interested in understanding what components of the exchange correlation functionals contribute most to the error in total energies.

Since we identified that the ring opening reactions lead to largest deviations among DFT results, we study prototype ring opening reactions on the simplest systems possible that undergo similar chemical changes as DHA/VHF isomerization. Ring opening isomerizations can be divided in to sub-groups that depend on the number of carbon atoms involved in the cyclization[123]. Typically, it is found that these systems involve 4 or 6 carbon atoms (4π-6π cyclization).

Several synthetic[99] and biological[123] compounds can be found to fall under these classifications undergoing ring opening isomerization, i.e., Vitamin D which is one of the most intensely studied cases. Instead of studying these large molecules to look for patterns of DFT errors, we examine the ring opening isomerization in cyclobutene and 1,3-cyclohexadiene (Fig. 5-3) as prototype systems and compare the errors of the DFT functionals in these reactions to the ring opening isomerization in the DHA/VHF isomerization. Using these structures as prototype is advantageous, as it removes any possible substitution effects on the isomerization enthalpies. Cylobutene and 1,3-cyclohexadiene are much extensively studied experimentally for isomerization enthalpies and reaction barriers in both directions, hence making it more straightforward to evaluate the performance of the computational methods.

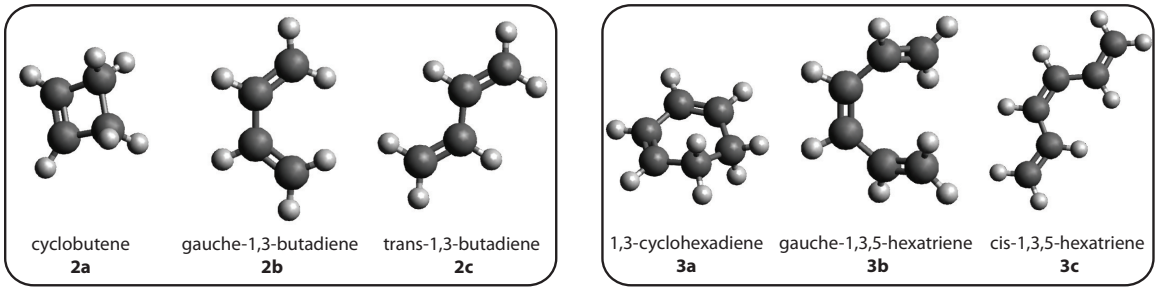


Figure 5-3: Cyclobutene, **2a-c**, and 1,3-cyclohexadiene, **3a-c** isomers studied in this work. Through ring opening isomerization reaction, cyclobutene converts into trans-1,3-butadiene, whereas 1,3-cyclohexadiene converts into cis-1,3,5-hexatriene.

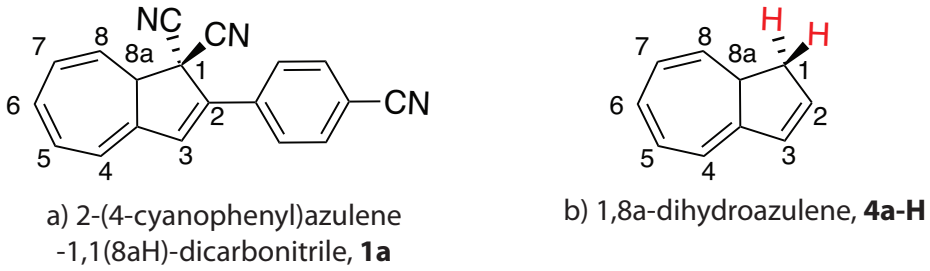
In Table 5.2, we show the isomerization enthalpies obtained for several isomers of cyclobutene and cyclohexane (given in Figure 5-3) using DFT functionals. For both cyclobutene and cyclohexane, the ring opening isomerization occurs between ring (cyclobutene and 1,3-cyclohexadiene) and gauche (gauche-1,3-butadiene, **2b** and

Table 5.2: Energy differences between the isomers of cyclobutene and 1,3-cyclohexadiene on the ground state potential energy surface of ring opening isomerizations.

	HF	LDA	PBE	B3LYP	PBE0	M06	CAM-B3LYP	TPSSH	DMC	CCSD(T)	Exp.
Cyclobutene, **2a**											
ΔH_{2c-2b}	0.138	0.151	0.157	0.153	0.147	0.128	0.132	0.154	0.14(6)		
ΔH_{2c-2a}	0.636	0.199	0.418	0.616	0.326	0.392	0.494	0.427	0.46(5), 0.53(1)[124]	0.482, 0.479 [125]	0.458(6)[124]
Cyclohexene, **3a**											
ΔH_{3c-3b}	0.37	0.43	0.439	0.432	0.412	0.363	0.372	0.435	0.32(6)		
ΔH_{3c-3a}	-0.556	-1.000	-0.626	-0.483	-0.800	-0.708	-0.684	-0.584	-0.73(8)	-0.768, -0.746[125]	-0.70(13)[126]

gauche-1,3,5-hexatriene, **3b**) conformations. However, experimental isomerization energies for cyclobutene are only available between trans-1,3-butadiene, **2c**, and cyclobutene, whereas for cyclohexane, the experimental isomerization energies are also available between 1,3-cyclohexadiene (**3a**), cis-1,3,5-hexatriene (**3c**) and trans-1,3,5-hexatriene[124, 126]. Therefore, even though the ring opening occurs between the ring and gauche conformations, in order to compare to the experiments, we additionally study the trans (or cis) conformations. In Table 5.2, we show that isomerization energies between gauche and trans (or cis) conformations are predicted using different DFT functionals with almost uniform accuracy. For cyclobutene, the trans-gauche isomerization enthalpy, ΔH_{2c-2b}, is predicted to be between 0.128-0.157 eV, where the DMC prediction is 0.14(6) eV. Similarly for cyclohexane, the cis-gauche isomerization enthalpy, ΔH_{3c-3b}, is predicted to be between 0.363-0.439 eV although in this case QMC prediction is 0.32(6) eV. For the trans-ring isomerization in cyclobutene, ΔH_{2c-2a}, we find that the variation among DFT results is much larger, 0.199-0.636 eV. Similarly for cyclohexane, the cis-ring isomerization enthalpy, ΔH_{3c-3a}, has a large variation among DFT results, between -0.570 to -1.084 eV. Thus, similar to the ring opening in the DHA-VHF isomerization, the ring opening reactions in cyclobutene and cyclohexane lead a larger deviation between DFT results compared to gauche-cis or gauche-trans isomerizations.

We next investigate how the DFT errors in the DHA-VHF ring opening isomerization correlate with the ring opening isomerization in cyclobutene and cyclohexane. However, it is known that different functional groups on the DHA molecule can mod-



a) 2-(4-cyanophenyl)azulene -1,1(8aH)-dicarbonitrile, **1a**

b) 1,8a-dihydroazulene, **4a-H**

Figure 5-4: DHA derivatives that are used in this study, shown with the atom numbering on the DHA backbone. In (b), -H groups colored in red indicate the site of substitutions using the functional groups listed in Figure 5-6.

ify the isomerization enthalpies[103, 110, 111, 127]. We are especially interested in the substitutions on the carbon atoms where the sigma bond in between is broken during the ring opening isomerization. In order to compare to the isomerization enthalpies of cyclobutene and 1,3-cyclohexadiene, we examine a derivative of DHA, **4a-H**, where the two cyanide (-CN) groups on these carbon atoms are substituted with -H and the benzonitrile group is removed on the five membered carbon ring. The overall substitution scheme is shown in Figure 5-4a-b. We use the notation **4a-R**, where **-R** is the functional group used to substitute the -H atoms represented in red in Figure 5-4b. There are two main reasons for doing this: (1) we show that (in the SI), removing the benzonitrile group has a very small effect on the isomerization enthalpy, such that the isomerization enthalpy between DHA and trans-VHF conformations change less than 0.05 eV upon removing the benzonitrile functional group. (2) Performing CCSD(T) calculations on the DHA derivatives with benzonitrile functional group takes significanly longer amount of time (around x20 times longer, assuming no memory limitations), due to its scaling $\sim \boldsymbol{N^7}$. Although DMC calculations does not suffer from such problems, in the absence of experimental isomerization energies, relying on a single theoretical method as the benchmark can be misleading. Therefore, any possible agreement between QMC and CCSD(T) methods can give us reasonable predictions on the experimental isomerization enthalpies

In Figure 5-5, all the values on x- and y-axes are the errors of the DFT functionals for **4a-H** with respect to the CCSD(T) calculations. Figure 5-5 shows that there is a good correlation between the DFT functionals in the isomerization enthalpy for all

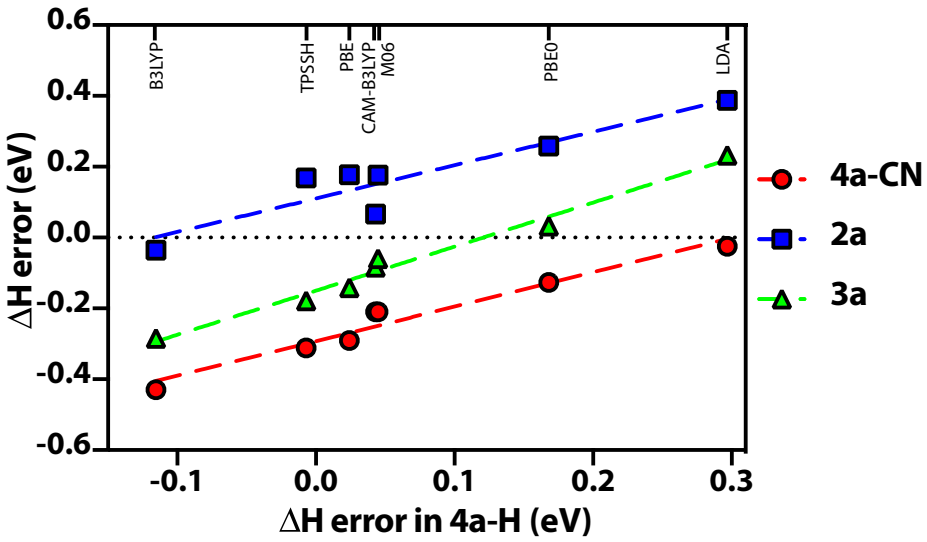


Figure 5-5: DFT errors in the isomerization enthalpy of -CN substituted DHA (**4a-CN**), cyclobutene and 1,3-cyclohexadiene, compared to the DFT errors in -H substituted DHA,(**4a-H**). Errors in DFT calculations are compared to the CCSD(T) method. All results are given in eV.

the ring opening isomerization reactions considered here: the two DHA derivatives (**1a** and **4a-H**), cyclobutene and 1,3-cyclohexadiene. However, although the trends are very similar, all the given errors have relative shifts with respect to each other. Between **4a-H** and cyclo-1,3-hexadiene, **3a**, this shift is very small. However, in **4a-CN** and cyclobutene, **2a**, this shift is relatively larger. Therefore, we can see that depending on the carbon ring size and the substitutions on the ring opening carbon atoms, the error introduced via DFT methods may vary up to $\sim$0.4 eV. While making comparions between such materials, DFT results must be benchmarked for each substitution on the ring opening carbon atoms, such that the scale of errors in DFT functionals may change depending on the substitution that is performed or size of the carbon ring.

In order to identify any trends for the DHA/VHF isomerization enthalpy error based on substitutional group on the ring opening moiety, we perform 22 different substitutions on the ring opening carbon atoms, given in Figure 5-6a-b. We use the substitution scheme explained in Figure 5-4b. On the x-axis of Figure 5-6a and b, the compounds are listed with respect to isomerization enthalpy errors obtained

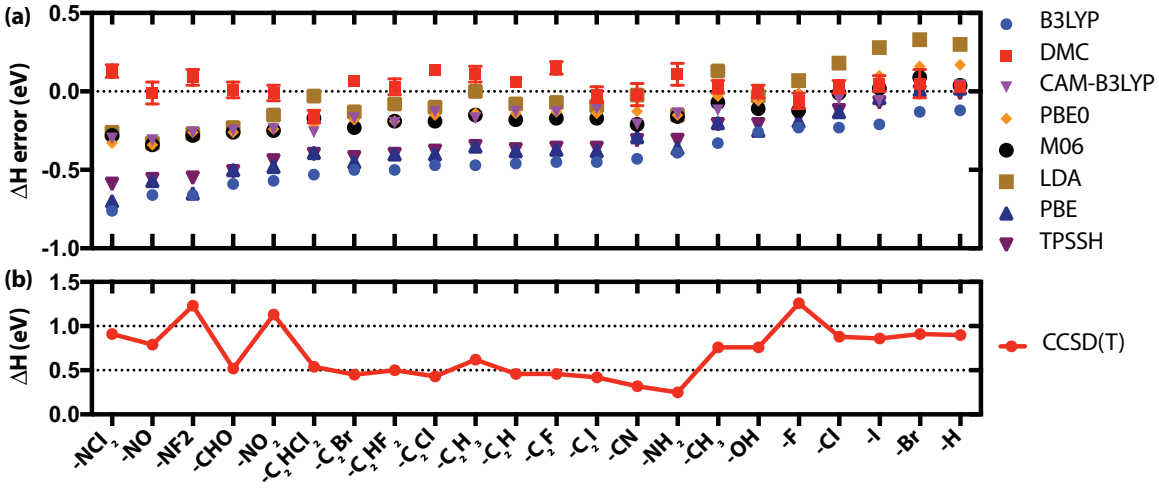


Figure 5-6: (a) Isomerization enthalpy, $\boldsymbol{\Delta H}$, error of DHA/VHF with substitutions on Carbon 1 site in Figure 5-4b. (b) Isomerization enthalpy using CCSD(T) method. All energies are given in eV.

using B3LYP functional from left to the right. Hence, in Figure 5-6a, it shows that all DFT functionals follow a similar trend in this substitutions vs. isomerization enthalpy error analysis, but only the DMC method does not follow any significant trend similar to the DFT functionals. Figures 5-6a and b do not also show any significant trend between each other, meaning that large DFT errors are not correlated with large isomerization enthalpies predicted using CCSD(T) method. However, the similarity between the errors in B3LYP, PBE and TPSSH in Figure 5-6a is interesting to note, as these methods systematically underestimate the isomerization enthalpy. Therefore, comparison of the breakdown of total energies from these functionals can help us understand which component of the DFT Hamiltonian is responsible from these errors.

A practical DFT calculation involves subsequent optimizations of charge density and geometric coordinates of the atoms which is based on the total energies that are calculated using the DFT Hamiltonian. Therefore, each DFT approximation yields different charge densities and geometric coordinates of atoms. In order to account for the contribution of each of these elements to the total error, $\boldsymbol{\Delta E}$, we write:

$$\boldsymbol{\Delta E = \Delta E_F + \Delta E_D + \Delta E_G} \tag{5.1}$$

where $\boldsymbol{\Delta E_F}$ is the error due to the functional, $\boldsymbol{\Delta E_D}$ is error due to the density and $\boldsymbol{\Delta E_G}$ is the error due to the geometry[128]. In order to understand the extent that these elements contribute to the error in DFT calculations, we compare the DFT isomerization enthalpies for two cases: (1) with geometry and charge density optimized within each functional separately and (2) geometry and charge density optimized using the B3LYP functional only while they are calculated non-self consistently for the otherfunctionals. We find that using B3LYP geometries and charge densities compared to optimizing each at every point on the reaction coordinate yields no more than 0.039 eV (0.005 eV/atom) difference (see SI for further details). Compared to the variations of up to ∼0.6 eV in the DFT total energies (given in Figures 5-5 and 5-6a), $\boldsymbol{\Delta E_D}$ and $\boldsymbol{\Delta E_G}$ is rather small, therefore $\boldsymbol{\Delta E_F}$ makes the largest contribution to $\boldsymbol{\Delta E}$. It has been discussed previously that DFT approximations yield nearly very close charge densities, as exchange and correlation make up only a small part of the total energy composition, hence their effect on the total charge density can be minor in most of the cases[128, 129].

In Figure 5-7, we show how the exchange and correlation energies change upon isomerization from ring conformation to trans conformation in **4a-CN** and **4a-H**. There are two main conclusions that can be drawn from Figure 5-7: (1) the change in the correlation energy upon isomerization, $\boldsymbol{\Delta E_c}$, stays unchanged between the isomerization reactions of **4a-CN** and **4a-H** for all DFT functionals, except M06. However, in MP2, CCSD(T) and DMC calculations, we find that $\boldsymbol{\Delta E_c}$ is larger for **4a-CN** compared to **4a-H**, indicating that correlation functionals of almost all DFT functionals must be inaccurate. (2) In **4a-CN** isomerization, exact exchange energy difference, $\boldsymbol{\Delta E_x}$, is substatially larger than other DFT exchange energy differences, indicating that -CN substitution may favor larger amount of exact exchange contribution. However, local LDA exchange yields very similar $\boldsymbol{\Delta E_x}$ as in HF for both in **4a-CN** and **4a-H**. PBE, B3LYP and TPSSH exchange functionals on the other hand yield lower $\boldsymbol{\Delta E_x}$ in both cases. The deficiency in the exchange functional of these functionals can also explain the systematical behavior shown in Figure 5-6. PBE exchange functional is based on LDA exchange with an enhancement factor

Where $\boldsymbol{E_x^{LDA}}$ is the local LDA exchange, $\boldsymbol{E_x^{HF}}$ is the exact exchange, $\boldsymbol{E_x^{B88}}$ is the Becke 88 exchange mixing, whereas $\boldsymbol{E_c^{LDA}}$ is LDA correlation and $\boldsymbol{E_c^{LYP}}$ is the Lee-Par-Yang correlation functional. $\boldsymbol{a_0}$, $\boldsymbol{a_x}$ and $\boldsymbol{a_c}$ represent mixing parameters, where for the B3LYP functionals they are know to be 0.2, 0.72 and 0.81 respectively. In Figure 5-7, it is shown that $\boldsymbol{\Delta E_x}$ for both HF and LDA is larger than B3LYP in the given DHA derivatives. Considering the B3LYP exchange is composed of HF, LDA and GGA exchange (B88), it can be expected that the GGA corrections should be responsible for lowering the $\boldsymbol{\Delta E_x}$ in both reactions, hence lowering the $\boldsymbol{\Delta E_x}$ in both reactions.

Therefore for B3LYP, we can suggest that the description of the GGA mixing to exchange energy must be adjusted to obtain accurate energies in ring opening reactions. In Figure 5-8, therefore we perform several calculations varying the Becke88 exchange mixing, $\boldsymbol{a_x}$, where $\boldsymbol{a_x = 0.72}$ corresponds to the original B3LYP formulation. Throughout these calculations, we keep the exact exchange mixing constant, 0.2, therefore $\boldsymbol{a_x = 0.0}$ means LDA exchange is 0.8, therefore there is no Becke 88 exchange contribution. For both cyclobutene and cyclohexane we find that a smaller amount of Becke 88 exchange, around $\boldsymbol{a_x = 0.6}$ should suffice to find accurate isomerization enthalpies for these isomerizations. However, for **4a-CN**, errors are more severe, therefore even smaller amount of Becke exchange must be used to find accurate isomerization enthalpies and the optimal Becke 88 exchange mixing is found to be $\boldsymbol{a_x \approx 0.2 - 0.3}$. In comparison to B3LYP exchange, CAM-B3LYP uses 0.19 HF

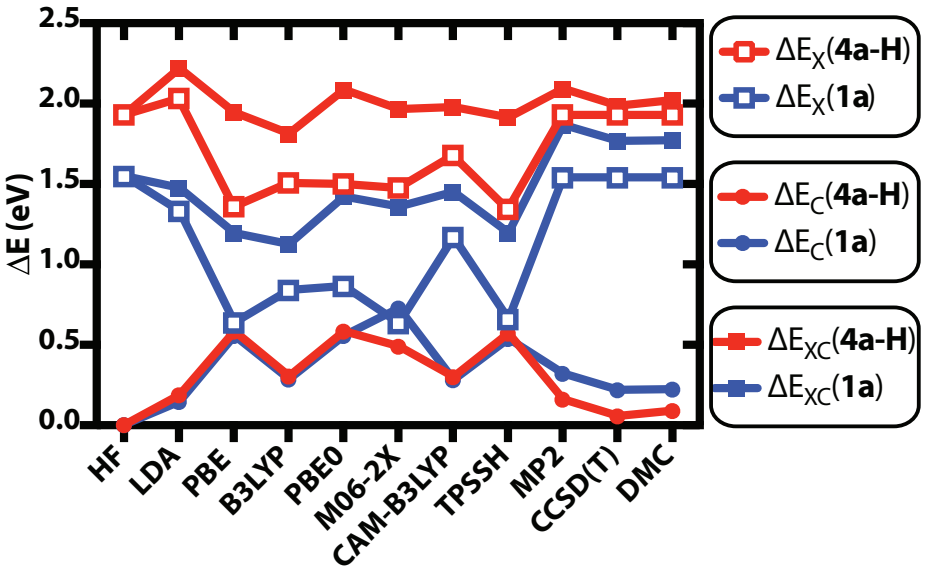


Figure 5-7: Changes in the exchange, $\boldsymbol{\Delta E_x}$, correlation, $\boldsymbol{\Delta E_c}$ and exchange-correlation, $\boldsymbol{\Delta E_{xc}}$, energies upon ring opening isomerization from ring conformation to trans conformation in -H and -CN substituted DHA derivatives, **4a-H** and **4a-CN** respectively.

and 0.81 B88 exchange for the short range, whereas 0.65 HF and 0.35 B88 exchange for the long range interactions. Therefore, using a larger portion of exact exchange for the weaker, long range interactions could help CAM-B3LYP functional to yield a $\boldsymbol{\Delta E_x}$ closer to exact $\boldsymbol{\Delta E_x}$ in HF. Having the same correlation energy as B3LYP, the difference in the CAM-B3LYP exchange functional leads to better total energies in Figure 5-7.

The substitutions performed on DHA molcule given in Figure 5-6a-b, not only help us identify the trend of errors in DFT calculations, depending in the substitutions, but it can also help us identify accurate estimations of the isomerization enthalpies. Therefore, we can use these results to calculate gravimetric energy density of these molecules, and then evaluate their potential as a solar thermal energy storage material. For all the substitutions considered, our QMC results show that the isomerization enthalpy of DHA/trans-VHF varies between 0.38 to 1.32 eV. For **4a-NF$_2$** derivative, 1.32 eV/reaction enthalpy is larger than the isomerization enthalpy of norbonadiene, 1.14 eV[98], whereas -F substitution yields an isomerization enthalpy of 1.20 eV. However, gravimetric energy storage capacity is the highest us-

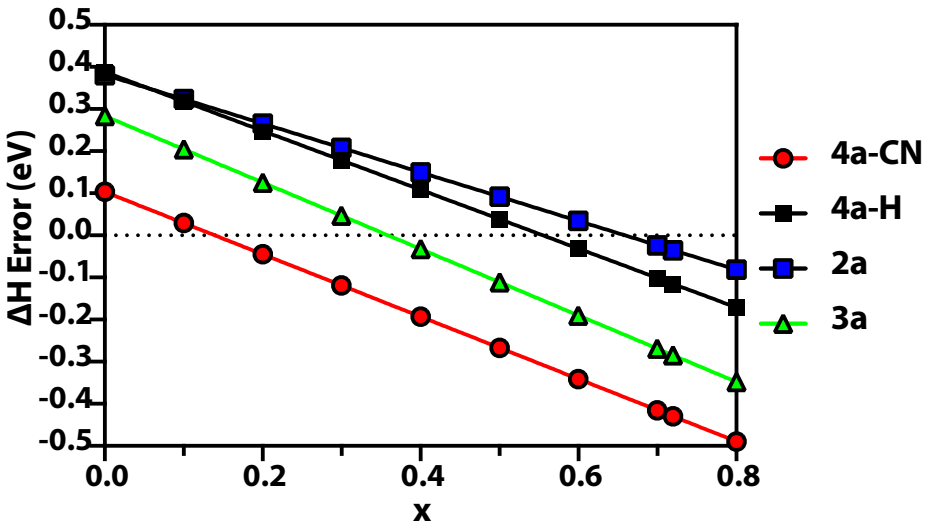


Figure 5-8: B3LYP errors in the isomerization enthalpy of -CN substituted DHA, -H substituted DHA, cyclobutene and 1,3-cyclohexadiene as a function of Becke 88 exchange mixing parameter. Errors in DFT calculations are compared to the DMC isomerization energies. All results are given in eV.

ing -F substituted DHA molecule, **4a-F**, that is 732 kJ/kg, while the second highest storage capacity is obtained with **4a-H**, with 667 kJ/kg. We find that both these compounds can achieve higher gravimetric energy density than highest energy density norbonadiene derivative according to our knowledge, 636 kJ/kg[90, 98].

The substitutions performed on DHA molcule given in Figure 5-6a-b, not only help us identify the trend of errors in DFT calculations, depending in the substitutions, but it can also help us identify accurate estimations of the isomerization enthalpies. Therefore, we can use these results to calculate gravimetric energy density of these molecules, and then evaluate their potential as a solar thermal energy storage material. For all the substitutions considered, our QMC results show that the isomerization enthalpy of DHA/trans-VHF varies between 0.38 to 1.32 eV. For **4a-NF_2** derivative, 1.32 eV/reaction enthalpy is larger than the isomerization enthalpy of norbonadiene, 1.14 eV[98], whereas -F substitution yields an isomerization enthalpy of 1.20 eV. However, gravimetric energy storage capacity is the highest using -F substituted DHA molecule, **4a-F**, that is 732 kJ/kg, while the second highest storage capacity is obtained with **4a-H**, with 667 kJ/kg. We find that both these compounds can achieve higher gravimetric energy density than highest energy density

norbonadiene derivative according to our knowledge, 636 kJ/kg[90, 98].

5.4 Conclusions

In this chapter, we are able to show that DFT approximations may fail to describe thermochemistry of DHA/VHF ring opening isomerization reaction. We show that these are not only specific to DHA/VHF couple, very similar error patterns can also be found for cyclobutene and 1,3-cyclohexadiene ring opening isomerizations. We show that particularly B3LYP and PBE functionals predict qualitatively inaccurate relative stabilities in DHA-VHF isomerization. We show that mainly GGA exchange functionals in both functionals are responsible for these results. However, correct behavior of the correlation functional is equally important, as almost none of the DFT functionals investigated yield the similar qualitative changes in the correlation energy upon the isomerization of DHA/VHF couple as in CCSD(T) and QMC calculations.

We find that DHA and VHF derivatives are promising alternatives for solar thermal energy storage applications, as the gravimetric energy density of the investigated molecules varies between 94-732 kJ/kg, which can be larger than the norbornadiene derivative that has the largest gravimetric energy density until now, that can be harnessed in solar thermal energy storage applications. For all the DHA/VHF derivatives that are investigated, we show that QMC and CCSD(T) isomerization energies agree with each other with a mean absolute error (MAE) of 0.07(1) eV/reaction. In the absence of experimental isomerization enthalpiy values, this agreement between two highly accurate methods is highly suggestive of the actual experimental values

Chapter 6

Genetic algorithm combined with QMC for accurate atomization energies of simple molecules and isomerization energies of electrocyclization reactions

6.1 Introduction

In Chapter 5, we discussed how DFT methods may provide sometimes qualitatively varying results in ring opening isomerization reactions. In such cases, where DFT functionals yield inconsistent results, one can resort to higher accuracy methods such as post-Hartree-Fock or QMC methods. Typical approach of using DFT in such cases, as we did in Chapter 5, is that DFT benchmarks are compared to a reference computational method or experiments for some subset of calculations, and then best performing DFT method is chosen for the remainder of the investigation. However, we also showed that this approach can fail substantially to calculate the isomerization enthalpy of ring opening reaction in DHA/VHF photoswitches when different

substitutions are considered.

Some of the DFT functionals investigated in Chapter 5 have well known deficiencies that are already learned via large sets of benchmark calculations[35, 131]. However, when DFT calculations are performed on an unexplored problem, the choice of functional is often left to the expertise of the person performing the calculations. In order to overcome this challenge, machine learning approaches have been proposed to systematically improve the accuracy of DFT functionals[28, 29] or to replace them with classical potentials[132]. Machine learning algorithms are basically form of functions which are obtained through regression strategies using input parameters to predict target values. For materials science problems, the input parameters can be geometric coordinates, such as the distances between the atoms or constituents of the system such as number of elements of each kind. The target values however, can be any physically interesting property such as energies, electronic or magnetic properties.

Although it would be desirable for these efforts in machine learning to yield a description of an improved functional, they make only minimal use of existing benchmark data except for a single reference method. This is acceptable if one chooses the reference method to be experimental results, but typically in order to generate large sets of reference data, computational methods, such as DFT, are chosen. However, this would mean that any non-systematic error that might be found in that DFT method can introduce bias to the machine learning results as well. Therefore, if one aims to reproduce chemical accuracy results using machine learning, then the reference method should be either the experimental results, or a computational method that is capable of producing chemical accuracy results. One can choose QMC as the computational reference method that can yield energies with chemical accuracy, but performing QMC calculations in large sets would again be no easy task due to its computational expense. However, in this case the deficiency of each DFT functional can be exploited in order to obtain as accurate as QMC method. Recent work demonstrates that machine learning approaches can outperform traditional human strategies in predicting new reaction successes from previously "failed" experiments[133]. Simi-

larly, each inaccurate DFT method, for a given problem, can be regarded as a "failed experiment". Therefore, a systematic way to learn from these "failures" of DFT methods could provide, without human intervention, progressively more accurate guidance on functional selection with increasingly available benchmark and functional data. In this chapter, we present a machine learning scheme to benchmark and combine an ensemble of DFT functionals for increased predictive capability and systematic DFT functional selection. We use a genetic algorithm (GA) based machine learning approach and a range of molecular descriptors based on composition and topology. We take both an existing standard benchmark set as well as a set of chemical reactions we developed as test beds. We show that our approach can: 1) help make decisions related to functional accuracy, 2) be applied to different materials and properties, 3) be performed over any number of DFT functionals and 4) perform even with low accuracy DFT results as a starting point.

6.2 Genetic Algorithm

A genetic algorithm is a type of machine learning algorithm which belongs to the larger class for evolutionary algorithms[134, 135]. The genetic algorithm performs symbolic regression, hence a set of variables are combined with another set of operators to form a solution space. It is particularly useful when there is fairly limited information about the mathematical formulation that the fitting function can possess. A priori information about likely types of operators or variables that may form the fitted function can be useful, but the genetic algorithm can yield a reasonable solution even in the absence of such information. It has been used in a variety of fields demonstrating its versatility for different problems such as finding optical gaps in amorphous silicon structures[10], self-assembly of molecular materials [136] and and materials design [137]. Genetic algorithm is inspired by evolutionary trees where functions are typically represented in a tree-like structure; input variables and constant values are represented by the leaf nodes and the remaining nodes are mathematical operators or simple functions[9], as shown in figure 6-1. Then as given in Figure 6-2, new functions

are created using crossover and mutation operators to search the configuration space of all possible functions, meaning either nodes of two possible functions are swapped between each other generating a new function (crossover), or one node of the function is randomly changed (mutation). Then the functions which show the least error to the target values are obtained using successive such operations. From the perspective of materials research, using a set of descriptors as variables, a population of functions can be created by calculating the correlation between the output values (due to the input descriptors) and observables from ab-initio electron structure calculations.

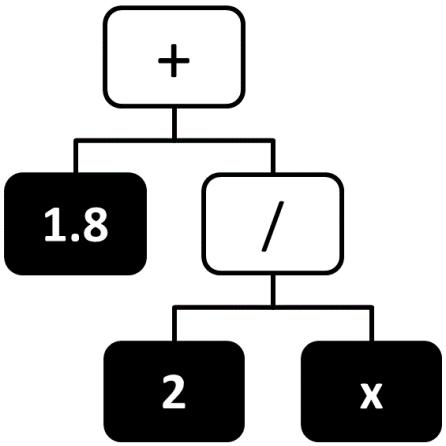


Figure 6-1: A tree-like representation of the expression $\mathbf{1.8} + \frac{\mathbf{2}}{\boldsymbol{x}}$. Operators are shown in boxes with white backgroun whereas variables have black background. Description of how Taken from ref. [9]

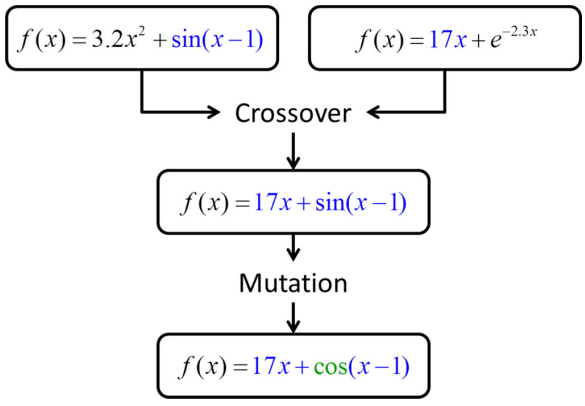


Figure 6-2: Examples of possible crossover and mutation steps in genetic programming. Taken from ref. [10]

It is easy to see that using larger number of operators and variables, one can obtain

better fitting functions. This would mean that the training set is overparametrized or overfitted. In order to control overfitting the data, complexity degrees must be assigned to each function/operator (or "building block") used to train the genetic algorithm. Complexity of each trained function is assigned as simply the sum of complexity values given to each building block. Complexity values are simple integer numbers that are assigned to each operator and variable that are allowed in genetic algorithm training. Total complexity value of a function would be summation of the complexity values of all variables and operators in its nodes and leaves respectively. For example, if complexity value of summation operator is set to 2, division set to 3, integer variables set to 1 and x variable set to 1 as well, then the complexity value of the function in Figure 6-1 would be 8. In this chapter, we used complexity values between 1-5, where simple operators such as addition, multiplication would get the complexity value of 1, whereas power and square root functions have the complexity value of 5. Hence, larger complexity values would mean that, overall, larger number of variables or operators are used to trained the target values. Therefore, each solution from the genetic algorithm can be grouped with respect to its complexity value. Then, Best fitting function with a given complexity value can be regarded as the best fitting function at each complexity value. Configuration space of these functions are also called as Pareto frontier, given in Figure 6-3, shows these best trained functions with a given complexity value with the red dots, connected with the red dashes. Genetic algorithm yields a distribution of solutions to the fitting problem, as shown on the Pareto front, and it is up to the end user to select one of these functions, making a compromise between complexity and fitness. The most complex function would always yield the best fit function, however this would also increase the obtain overfitted functions. Hence, sudden drops on the Pareto frontier must be observed, since they represent the instances when a new information is learned through the fitting. Throughout this work, we used the GA functions that have the complexity values of no more than 10 and having the largest drop in the unfitness close to this cutoff value. We use the genetic algorithm method as implemented in EUREQA software package [138] and the default algorithm in EUREQA was used to split the samples

in training and validation set. Then, best fit functions having the same complexity values are observed on the Pareto front, the least fit functions are discarded.

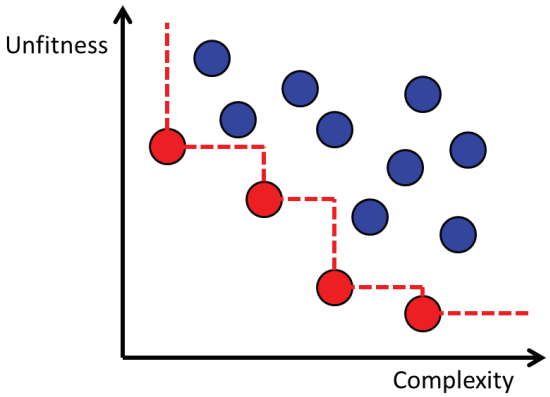


Figure 6-3: Pareto frontier for a set of functions with different levels of fitness and complexity. Each dot represents a function, and the red dots (connected with dashed line) are the ones on the Pareto frontier. Taken from ref. [9]

6.3 Computational Methods

Our GA approach can be compared to other machine learning based[132] or analytical[139] forcefield approaches where a single kernel is trained with respect to the reference. In these methods however, reference data is usually produced using a single DFT approximation. The choice of the DFT approximation is justified through its general performance in the literature, without investigating whether that DFT functional provides the most accurate solution for each compound in the reference data. Therefore, these forcefield approaches are also likely to yield inaccurate results when it is not certain if the computational reference method is accurate for the training set. Further, it is not possible to provide an estimate of error on the prediction, unless an ensemble of such models are trained. Our approach shares similar qualities to these approaches to increase its accuracy, e.g., it can be trained with respect to a larger set of benchmark data, the functional complexity can be increased (e.g. higher order terms for a physical interaction) or new terms can be added to the model to account for previously unrepresented physical interactions.

Since our aim is to increase the accuracy of the DFT calculations through systematic machine learning applications, we focus on examples of finite systems where DFT methods are found to yield large deviations in the ground state energy differences such as simple hydrocarbon reactions[140], allene-propyne isomerizations and electrocyclization reactions[141]. These systems pose a challenge to DFT due to significant local changes in the molecular structure such as the change in hybridization or complex rearrangement of atoms. Therefore, the atomization energy of molecules can also yield large deviations between different DFT methods, as it leads to the largest change in the local environment of each atom. We first evaluate the potential of our approach using atomization energies of well known benchmark sets, such as G2/97[142] and W4-08 [143] using them as our test bed, and then on a set of electrocyclic isomerization reactions.

The following steps are performed to construct our training set for the atomization or isomerization energy of a molecule: a training set is initially benchmarked using DFT methods and higher accuracy theoretical calculations as the reference. The error of each DFT method, $\boldsymbol{\epsilon^i = E^i - E^{ref}}$, where i is the DFT method considered and $\boldsymbol{E^{ref}}$ is the reference energy.$\boldsymbol{\epsilon^i}$ is then trained with respect to the molecular descriptors[144], $\boldsymbol{\theta}$. This training process yields $\boldsymbol{\epsilon^i(\theta)}$, which is the algebraic function for DFT errors which uses descriptors as input. The role of using suitable descriptors for a given property is found to be critical[31]. We use a large number of molecular descriptors such as the autocorrelation functions of different quantities (i.e. mass, charge, electronegativities) between atoms in the molecular network. The genetic programming algorithm in EUREQA evaluates the stability of each variable using cross validation, optimizing the mean absolute error of the fit while minimizing the number of variables that are used. We found that no more than 15 variable solutions are obtained for each case and using more than 5 variables introduced only relatively minor improvement on the quality of the fit. Then GA assisted DFT energies, $\boldsymbol{E_i^{GA}}$ are obtained simply by summing the DFT energy, $\boldsymbol{E^i}$, and the GA correction, $\boldsymbol{\epsilon^i(\theta)}$. This process can be repeated over an unlimited number of DFT approximations, $\mathrm{N}_{\boldsymbol{DFT}}$. $\mathrm{N}_{\boldsymbol{DFT}}$ independent models are then combined to solve a constrained linear

least squares fit to obtain a single observable,

$$\boldsymbol{E^{GA}} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_{DFT}}} \boldsymbol{w_i E_i^{GA}} \tag{6.1}$$

as well as an error estimate,

$$\boldsymbol{\sigma} = \sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N_{DFT}}} \boldsymbol{w_i (E_i^{GA} - E^{GA})^2}} \tag{6.2}$$

where w_i is the associated weight for each GA assisted DFT model. Following a simple Bayesian analysis, we expect that the larger the error estimate, $\boldsymbol{\sigma}$, the larger the error of the GA method compared to the reference, $\boldsymbol{\epsilon^{GA} = E^{GA} - E^{ref}}$. In Figure 6-4a, we show for an example case, what these variables correspond to. In the $\boldsymbol{E - GA^i}$ versus $\boldsymbol{E^{QMC}}$ plot, we show the corrected energies from each DFT approximation $\boldsymbol{i}$, where, $\boldsymbol{E^{QMC}}$ is the reference energy method that we used to benchmark our DFT calculations. Then, in Figure 6-4b, we combine all different $\boldsymbol{E_i^{GA}}$ results using equation 6.1. The spread of $\boldsymbol{E_i^{GA}}$ energies is also used as the standard deviation, $\boldsymbol{\sigma}$, using equation 6.2. Our approach similar to that implemented in ref 145, where uncertainty measure is proportional to the cost function. Reference calculations are obtained using both QMC and in some suitable cases coupled cluster with single, double and perturbationally triple excitations, CCSD(T) are also used in addition to QMC method four double checking.

6.4 Results and Discussion

As an initial test of our approach, for the training set we use the atomization energies of neutral molecules in the G2/97 set[142] at 0K without zero-point energies. In order to test transferability of the model we then apply it to the test set, W4-08 [143], without modifying $\boldsymbol{\epsilon^i(\theta)}$ and w_i. Duplications among the two sets are omitted leaving a modified W4-08 set, W4-08*, with 60 molecules. In Figure 6-5a and b, we compare the performance of the three best performing DFT functionals based on root mean square

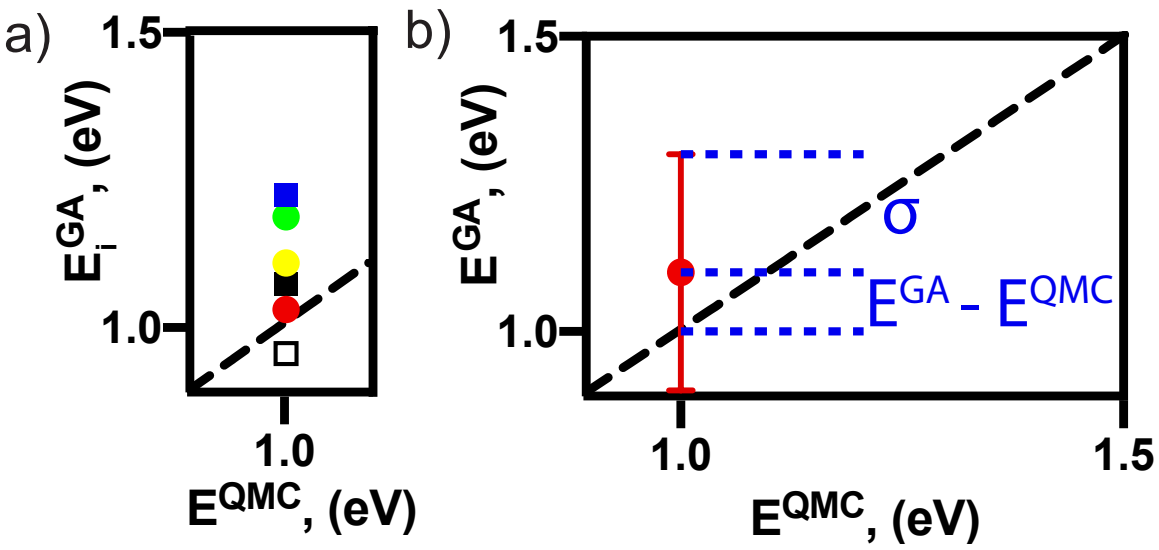


Figure 6-4: a) $\boldsymbol{E_i^{GA}}$ energies with respect to the reference energy from QMC calculations for a hypothetical compound with $\boldsymbol{E^{QMC}}$ =1.0 eV. $\boldsymbol{i}$ denotes each DFT approximation used in the training. b) Combination of the results in a) using equations 6.1 and 6.2.

deviations (RMSD) with respect to CCSD(T) calculations in the G2/97 and W4-08* sets respectively. We have benchmarked 8 other functionals[15, 39, 146–152] and use all of the DFT results to perform the GA training. CCSD(T) calculations are used as the reference method to evaluate errors of all the functionals for the atomization energies. As has already been noted in the literature, hybrid functionals typically yield the most accurate atomization energies. Atomization energy benchmark sets such as G2/97 [142] and W4-08 [143] yield B3LYP as the best performing functional (i.e. G2/97: RMSD in B3LYP 0.11, W4-08*: RMSD in B3LYP 0.15 eV/atom). However, the GA algorithm implemented in this work yields the smallest RMSD in both G2/97 and W4-08* sets (0.03 and 0.11 eV/atom, respectively) compared to DFT with the distribution given in Figure 6-5a and b.

In order to understand how prediction accuracy increases with the subset of the ensemble of energies used in the training set, we use a total of 11 different DFT functionals and train our GA algorithm on every possible combination. In Figure 6-5c, we see that the RMSD errors for the GA approach throughout the G2/97 and W4-08* sets decrease as more DFT functionals are used in the GA training, where

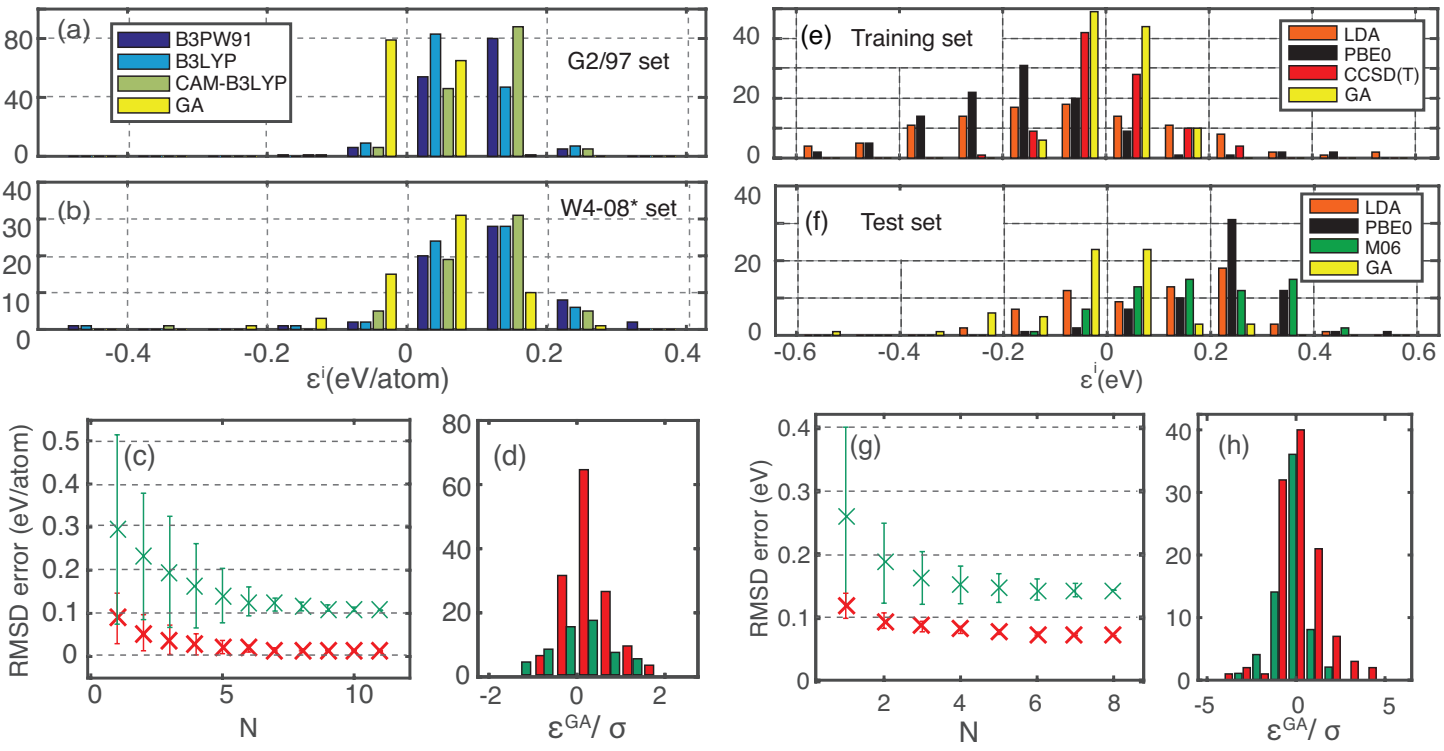


Figure 6-5: (a) Histogram of errors atomization energy per atom, $\boldsymbol{\epsilon^i}$ for each DFT method, $\boldsymbol{i}$, in training set (G2/97) (b) and test set W4-08* (in eV). The three best performing DFT functionals and our GA approach relative to CCSD(T) calculations are given in both plots. (c) Root mean square deviation (RMSD) for the GA model as a function of the size (k) of subset of DFT functionals (see text), where error bars here represent the spread of $\boldsymbol{\sigma}$ across different combinations, (d) actual error of the GA estimate $\boldsymbol{\epsilon^{GA}}$ relative to the error bar determined by the GA approach. In both (c) and (d), red and green colors represent training and test sets, respectively. For (e-h), the same order follows from (a-d), but the analysis is performed for electrocyclic reactions.

$\boldsymbol{N}$ is the number of DFT functionals used in the training. Error bars in the same graph represent the spread of the RMSD error when different combinations of DFT functionals, with the same size $\boldsymbol{N}$, is used to train our model. In this case even a standalone functional with low accuracy (i.e., LDA and HF for this set) can improve the accuracy of the prediction once the GA based correction is applied.

In Figure 6-5d, we show the ratio of actual error, $\boldsymbol{\epsilon}$, versus the error estimate, $\boldsymbol{\sigma}$. This ratio produces a unitless parameter that can be helpful to identify the reliability of the error estimates from a statistical perspective. Large values of $\boldsymbol{\sigma}$ in the test set, W4-08* can help identify new systems that should be added to the

training set for increased transferability. For example, oxygen fluoride (OF), dioxygen difluoride (FOOF),tetra phosphorus (P_4), and sulfur trioxide (SO_3) have the largest $\boldsymbol{\sigma}$ in the W4-08* set (>0.3 eV/atom), which is perhaps unsurprising since the G2/97 set contains no sulfur atoms with sp^3d^2 hybridization, no P atom and only one example of a molecule (OF_2) where oxygen is bonded to fluoride.

The atomization energies of the G2/97 training and W4-08* test sets provide an excellent starting point to gauge the accuracy of any total energy method; however, understanding the transferability of an approach not just to other materials but also other properties is of crucial importance. Therefore, we assess the performance of our approach for a more challenging problem in DFT, namely electrocyclic reactions. The electrocyclic reaction is a type of pericyclic isomerization reaction where the transition state of the molecule has a cyclic geometry and the net result is one sigma bond converted to one pi-bond or vice versa[153]. Application of standard DFT functionals to such materials has been shown to yield inaccurate thermochemistry predictions, e.g. between isomers in pericyclic reactions such as Diels-Alder reactions [154] and bond rearrangement energies in ring and cage-like molecules[140, 155] and extended conjugation systems[141]. A further challenge lies in the fact that experiments typically only provide activation barriers in single direction only, as such, they are not sufficient to identify isomerization enthalpies. Molecules which could be of interest to engineering applications typically contain 100's of valence electrons, rendering wavefunction based electronic structure calculations such as CCSD(T) computationally prohibitive. Therefore, following a previous work[156], here we make use of small model molecules to benchmark DFT functionals and construct our GA model. We construct this training set with the idea that these molecules undergo similar chemical changes in their structure with respect to our test set, which is composed of molecules of engineering interest. As with any empirical approach, our aim is to learn from smaller systems where performing high quality calculations is feasible and use this understanding to predict energies of more complex molecules spanning a similar chemical space. We use the QMC as our reference method for both the training and validation sets, due to its efficient scaling with respect to number of valence electrons.

Otherwise, the same steps are performed as in training and validation of G2/97 and W4-08* sets.

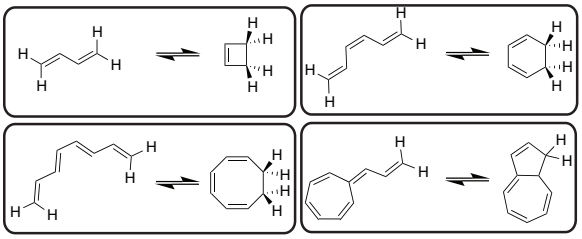


Figure 6-6: Set of molecules that are used in the photocyclization training set. Substitutions are performed on each explicit H atom given in the figure.

The electrocyclization training set (Fig. 6-6) consists of even-numbered carbon rings of four different sizes, up to 10 carbon atoms in the ring, with several substitutions, 27 for each ring size. All substitutions considered in our calculations are applied at the same sites as the explicit hydrogen atoms shown in figure 6-5e, and are listed in the SI. Different substitutions provide tunability in electron donating/withdrawing power of substitutions, hence changes the isomerization enthalpy. The test set in this case is composed of 65 photochromic molecules which also undergo electrocyclization reaction such as dihydroazulenes(DHA), diarylethenes (DTE), fulgides (FLG) and spiropyrans (SPR). In Figure 6-5f and c, we present our benchmarking results of DFT functionals with lowest RMSD, GA and the CCSD(T) method with respect to QMC calculations. We observe that across all investigated functionals (which may not be shown in Fig. 6-5) results may vary as much as 2 eV for the isomerization enthalpy of a reaction. In the training set, DFT functionals typically underestimate the isomerization enthalpy except for, perhaps quite surprisingly, LDA which has the smallest ME of -0.08 eV. In contrast, B3LYP dramatically underestimates the isomerization enthalpy with a ME of -0.62 eV. The long range attenuation factor in the CAM-B3LYP functional, with a ME of -0.28 eV, improves the results with respect to the B3LYP functional while the QMC and CCSD(T) methods are in very good agreement (ME 0.003 eV). Our method performs exceedingly well for the test set where the majority of the results are 0.1 eV off from the QMC result, as shown in Fig. 6-5f. Fig. 6-5g-h however provides similar results as we showed in 6-5c-d.

For cases where either experimental or higher accuracy ("beyond DFT") results are unavailable, the GA approach can be utilized to predict the best performing DFT functional for different classes of chemistries. In our example, the training set is composed of four different classes of molecules, so we may expect that the performance of a DFT functional across these types of electrocyclic molecules may not be transferable. In order to quantify the selection criteria as probability of choosing one DFT functional, we use discrete choice analysis [157], such that the probability can be expressed as

$$\boldsymbol{p_i} = \frac{\boldsymbol{e^{\mu V_i}}}{\sum_{\boldsymbol{i}} \boldsymbol{e^{\mu V_i}}}, \tag{6.3}$$

where the utility function, V_i, is $-(\boldsymbol{E^i} - \boldsymbol{E^{ref}})^{\boldsymbol{2}}$ in our case and $\boldsymbol{\mu} = (\boldsymbol{2\sigma^2})^{\boldsymbol{-1}}$, which is inversely proportional to the variance of the reference method. If GA is used as the reference method to determine the best performing DFT functional for that reaction, we use the $\boldsymbol{\sigma^2}$ as determined in eqn. 6.2, but if a high accuracy reference method or experiments are used, then we choose it to be equal to "chemical accuracy," or 1 kcal/mol (0.043 eV), assuming that a high accuracy calculation can actually substitute for the experiment.

In Figure 6-7a and b, each group of bars shows the $\boldsymbol{p_i}$ of each DFT functional within a molecular group given in the electrocyclic test set, shown on the x-axis, when QMC and our method is used as the reference, respectively. Whereas, RMSD errors in Fig. 6-7c are calculated using QMC as reference, therefore they are inversely related to the $\boldsymbol{p_i^{QMC}}$ in Fig. 6-7a, due to the equation 6.3. Fig. 6-7c shows that for DHA derivatives, the HF method yields the best performance. For SPR derivatives, although CAM-B3LYP, HF and LDA have very similar RMSD values, the mean absolute error of HF is 0.04 eV smaller the next best DFT functional, therefore yields slightly higher probability in Fig 6-7a. Comparing 6-7a and 6-7b, the relative ordering of the $\boldsymbol{p_i}$ among DFT functionals is almost the same with very few exceptions where $\boldsymbol{p_i}$ values of two functionals are very close. Using our GA approach to provide such ordering of probabilities, which is invariant with respect to the choice of $\boldsymbol{\mu}$, could prove useful to identify best performing DFT functional without using much expensive

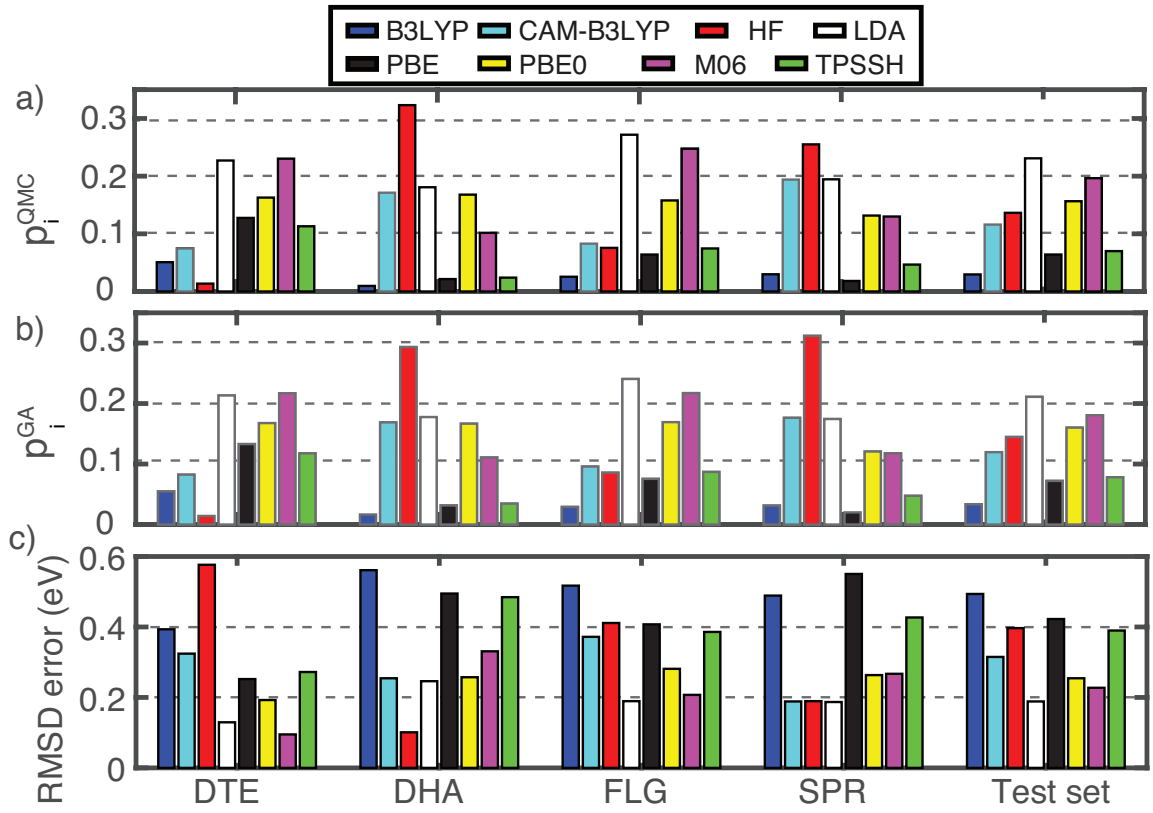


Figure 6-7: Probability of choosing a DFT functional in the electrocyclic reactions test set, when (a) QMC energies or (b) GA energies are used as reference. Each figure is divided into five groups according to molecular classes given on the x-axis, where the last group covers the whole test set. (c) shows the RMSD errors of each DFT functional with respect to QMC calculations, which is conversely related to (a).

computational approaches, such as QMC in this case.

6.5 Conclusion

To summarize, we present a simple way to build large scale, benchmark quality data sets using molecular descriptors and DFT, combined with high-accuracy methods such as QMC or CCSD(T). Our approach can easily be applied to existing benchmark sets and provide increased predictive ability for new DFT calculations. We note that our approach could be modified for periodic solids or other observables provided that suitable topological or structural descriptors for are available. In practical calculations the error estimates may depend on both the choice of the database and basis set size,

as well as the pseudopotentials or DFT functionals used in the study. However, the use of multiple DFT approximations helps to obtain better statistics and the size of the error bars determines the confidence in a given result.

Chapter 7

Conclusions and Outlook

In this thesis, we used a variety of computational tools and strategies to achieve very accurate formation, atomization and isomerization energies, as well as intercalation potentials for lithium atoms in layered transition metal oxide cathode materials. Mainly, we showed that quantum monte Carlo (QMC) method can be applicable to variety of materials problems and can also be used to refine existing DFT calculations to make higher accuracy predictions using a machine learning approach. We showed that 1) QMC calculations can be applied to periodic solids, using a well defined recipe that limits the user effort that is put into optimizing the many body wave function and minimizing finite size errors due to errors in coulombic summation and inadequate k-point sampling especially for metallic systems. 2) QMC calculations can be used together with coupled cluster methods (CCSD(T)) to identify the shortcomings of the DFT functionals for molecular reactions. These results can also be applied in an automated scheme where DFT functionals are combined to predict QMC and CCSD(T) results, in the absence of these calculations.

First, performed DMC calculations on the formation energies of 21 compounds, whose formation energies are well-known experimentally. Out of these 21 structures, we obtained chemical accuracy in 11 structures, where as for the compounds that include Ag and Hg, (which are AgCl, HgO and Hg_2SO_4), DMC formation enthalpy errors were significantly larger than DFT (VASP/PAW) calculations. We investigated several sources of error that may lead to this result, but found that the pseudopoten-

tials are the main factor that is responsible from such non-systematic errors. We found that BFD and OPT pseudopotentials are able to provide very accurate formation energies for the transition metal oxides that we investigated, but such pseudopotentials are not available for below first fow transition metals. Therefore, based on our tests on ZnO, TiO2, and MgF2, we believe the RB and FHI pseudopotentials for Ag and Hg are likely the source of the anomalously high error. Showing that a simple approach that we applied here is able to provide formation energies more accurate than DFT calculations in 85% of the cases, the DMC protocol investigated here can be a host for many different applications. Our critical analysis of the pseudopotentials show that the developing accurate pseudopotentials form DMC applications is currently the main obstacle for the wider applicability of the method. Compared to DFT calculations, we find that very small non local radii should be chosen in pseudopotentials that could be applicable for DMC calculations. However, typically these pseudopotentials are avoided in DFT calculations due to very high cutoff energies associated with the small radii.

Secondly, we show that DFT applications can be completely misleading in giving qualitatively correct isomerization energies of electrocyclization reactions such as dihydroazulene and vinylheptafulvalene (DHA/VHF) isomerization. We identified that GGA exchange in PBE, TPSSH and B3LYP functionals are mainly responsible from the qualitatively inaccurate results in these functionals. We also find that except for the M06 functional, none of the functionals were able to yield accurate correlation energies for the isomerization. We investigated the electrocyclic isomerizations in cyclobutene and cyclohexene as well, finding that DFT functionals exhibit the same error trends in these cases, signifying a deeper problem in these functionals. However, using genetic algorithm for atomization energies and for these electrocyclization enthalpies, we devised a strategy where it is possible to obtain accurate isomerization enthalpies solely from DFT calculations. We presented a simple way to build large scale, benchmark quality data sets using molecular descriptors and DFT, combined with high-accuracy methods such as QMC or CCSD(T). Our approach can easily be applied to existing benchmark sets and provide increased predictive ability for new

DFT calculations.

We believe that this work opened a new direction in the field of quantum Monte Carlo calculations, showing that QMC methods can be used with a limited user intervention to provide high accuracy energies for both periodic and molecular structures. QMC accuracies are comparable to experimental results or CCSD(T) calculations. High accuracy data is critical in obtaining quality benchmark sets that will further enhance the existing efforts for building machine learning approaches which can be constructed using thousands of structures, for rapid evaluation of structural parameters as well as total energies. This thesis shows that our work can make an important impact in these areas, but with the increasing computational resources, computational cost of performing QMC calculations would be reduced and larger databases would be constructed to make greater impact.

Bibliography

[1] Awadhesh Narayan, Ankita Bhutani, Samantha Rubeck, James N Eckstein, Daniel P Shoemaker, and Lucas K Wagner. Computational and experimental investigation for new transition metal selenides and sulfides: The importance of experimental verification for stability. *Phys. Rev. B - Condens. Matter Mater. Phys.*, 94(4):045105, 2016.

[2] Michael G. Medvedev, Ivan S. Bushmarinov, Jianwei Sun, John P. Perdew, and Konstantin A. Lyssenko. Density functional theory is straying from the path toward the exact functional. *Science*, 355(6320):49–52, 2017.

[3] Backflow corrections in qmc. https://vallico.net/tti/qmcitaa_05/talks/lopezrios/backflow_ESDG_01June05.pdf. Accessed: 05-08-2017.

[4] Quantum monte carlo methods. http://www.cmth.ph.ic.ac.uk/people/m.foulkes/qmc.html. Accessed: 05-08-2017.

[5] Paolo Giannozzi et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *J. Phys.: Condens. Matter*, 21(39):395502, 2009.

[6] Mike D. Towler. Casino manual 2015. Accessed: 2016-10-17.

[7] G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B*, 47:558–561, Jan 1993.

[8] G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B*, 59:1758–1775, Jan 1999.

[9] Tim Mueller, Aaron Gilad Kusne, and Rampi Ramprasad. *Machine Learning in Materials Science*, pages 186–273. John Wiley and Sons, Inc, 2016.

[10] Tim Mueller, Eric Johlin, and Jeffrey C. Grossman. Origins of hole traps in hydrogenated nanocrystalline and amorphous silicon revealed through machine learning. *Phys. Rev. B*, 89:115202, Mar 2014.

[11] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *Jom*, 65(11):1501–1509, 2013.

[12] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1), 2013.

[13] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[14] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[15] W. Kohn and L. Sham. Quantum Density Oscillations in an Inhomogeneous Electron Gas. *Phys. Rev.*, 137(6A):A1697–A1705, March 1965.

[16] Jp Perdew, K Burke, and M Ernzerhof. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.*, 77(18):3865–3868, October 1996.

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

[18] Alexander Khein, D. J. Singh, and C. J. Umrigar. All-electron study of gradient corrections to the local-density functional in metallic systems. *Phys. Rev. B*, 51:4105–4109, Feb 1995.

[19] Alejandro J. Garza and Gustavo E. Scuseria. Predicting band gaps with hybrid density functionals. *J. Phy. Chem. Lett.*, 7(20):4165–4170, 2016.

[20] Anubhav Jain, Geoffroy Hautier, Charles J. Moore, Shyue Ping Ong, Christopher C. Fischer, Tim Mueller, Kristin A. Persson, and Gerbrand Ceder. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.*, 50(8):2295 – 2310, 2011.

[21] F. Zhou, M. Cococcioni, C. A. Marianetti, D. Morgan, and G. Ceder. First-principles prediction of redox potentials in transition-metal compounds with lda+u. *Phys. Rev. B*, 70:235121, Dec 2004.

[22] Lei Wang, Thomas Maxisch, and Gerbrand Ceder. Oxidation energies of transition metal oxides within the gga+u framework. *Phys. Rev. B*, 73:195107, May 2006.

[23] Malcolm W. J. Chase. *NIST-JANAF Thermochemical Tables, 4th Edition.* American Institute of Physics, New York, 1998.

[24] John P. Perdew, Matthias Ernzerhof, and Kieron Burke. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.*, 105(22):9982–9985, 1996.

[25] Lars Hedin. New method for calculating the one-particle green's function with application to the electron-gas problem. *Phys. Rev.*, 139:A796–A823, Aug 1965.

[26] Luke Shulenburger and Thomas R. Mattsson. Quantum monte carlo applied to solids. *Phys. Rev. B*, 88:245117, Dec 2013.

[27] Awadhesh Narayan, Ankita Bhutani, Samantha Rubeck, James N. Eckstein, Daniel P. Shoemaker, and Lucas K. Wagner. Computational and experimental investigation for new transition metal selenides and sulfides: The importance of experimental verification for stability. *Phys. Rev. B*, 94:045105, Jul 2016.

[28] J. Mortensen et al. Bayesian error estimation in density-functional theory. *Phys. Rev. Lett.*, 95:216401, Nov 2005.

[29] Manuel Aldegunde et al. Development of an exchangeâĂŞcorrelation functional with uncertainty quantification capabilities for density functional theory. *J. Comp. Phys.*, 311:173 – 195, 2016.

[30] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007.

[31] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Phys. Rev. Lett.*, 114:105503, Mar 2015.

[32] Rika Kobayashi and Alistair P. Rendell. A direct coupled cluster algorithm for massively parallel computers. *Chemical Physics Letters*, 265(1):1 – 11, 1997.

[33] Chr. Møller and M. S. Plesset. Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618–622, Oct 1934.

[34] Rodney J. Bartlett and John F. Stanton. *Applications of Post-Hartree-Fock Methods: A Tutorial*, pages 65–169. John Wiley & Sons, Inc., 2007.

[35] Aron J. Cohen, Paula Mori-SÃąnchez, and Weitao Yang. Challenges for density functional theory. *Chemical Reviews*, 112(1):289–320, 2012. PMID: 22191548.

[36] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996.

[37] Viktor N. Staroverov, Gustavo E. Scuseria, Jianmin Tao, and John P. Perdew. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *The Journal of Chemical Physics*, 119(23):12129–12137, 2003.

[38] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.*, 98(7):5648, 1993.

[39] Axel D. Becke. A new mixing of hartreeâĂŞfock and local densityâĂŘfunctional theories. *The Journal of Chemical Physics*, 98(2):1372–1377, 1993.

[40] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.*, 118(18):8207, 2003.

[41] Jianwei Sun, Richard C Remsing, Yubo Zhang, Zhaoru Sun, Adrienn Ruzsinszky, Haowei Peng, Zenghui Yang, Arpita Paul, Umesh Waghmare, Xifan Wu, Michael L Klein, and John P Perdew. Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional. *Nat Chem*, 8(9):831–836, sep 2016.

[42] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[43] C. J. Umrigar, M. P. Nightingale, and K. J. Runge. A diffusion monte carlo algorithm with very small timeâĂŘstep errors. *The Journal of Chemical Physics*, 99(4):2865–2890, 1993.

[44] C. J. Umrigar, K. G. Wilson, and J. W. Wilkins. Optimized trial wave functions for quantum monte carlo calculations. *Phys. Rev. Lett.*, 60:1719–1722, Apr 1988.

[45] M. Bajdich, L. Mitas, L. K. Wagner, and K. E. Schmidt. Pfaffian pairing and backflow wavefunctions for electronic structure quantum monte carlo methods. *Phys. Rev. B*, 77:115112, Mar 2008.

[46] Michele Casula and Sandro Sorella. Geminal wave functions with jastrow correlation: A first application to atoms. *The Journal of Chemical Physics*, 119(13):6500–6511, 2003.

[47] S. A. Alexander and R. L. Coldwell. Atomic wave function forms. *International Journal of Quantum Chemistry*, 63(5):1001–1022, 1997.

[48] N. D. Drummond, M. D. Towler, and R. J. Needs. Jastrow correlation factor for atoms, molecules, and solids. *Phys. Rev. B*, 70:235119, Dec 2004.

[49] A. J. Williamson, S. D. Kenny, G. Rajagopal, A. J. James, R. J. Needs, L. M. Fraser, W. M. C. Foulkes, and P. Maccullum. Optimized wave functions for quantum monte carlo studies of atoms and solids. *Phys. Rev. B*, 53:9640–9648, Apr 1996.

[50] Jeffrey C. Grossman, Lubos Mitas, and Krishnan Raghavachari. Structure and stability of molecular carbon: Importance of electron correlation. *Phys. Rev. Lett.*, 75:3870–3873, Nov 1995.

[51] S. F. Boys and N. C. Handy. A calculation for the energies and wavefunctions for states of neon with full electronic correlation accuracy. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 310(1500):63–78, 1969.

[52] Tosio Kato. On the eigenfunctions of many-particle systems in quantum mechanics. *Communications on Pure and Applied Mathematics*, 10(2):151–177, 1957.

[53] Matthias Troyer and Uwe-Jens Wiese. Computational complexity and fundamental limitations to fermionic quantum monte carlo simulations. *Phys. Rev. Lett.*, 94:170201, May 2005.

[54] P. R. C. Kent, R. J. Needs, and G. Rajagopal. Monte carlo energy and variance-minimization techniques for optimizing many-body wave functions. *Phys. Rev. B*, 59:12344–12351, May 1999.

[55] Paul RC Kent. *Techniques and Applications of Quantum Monte Carlo*. PhD thesis, Univeristy of Cambridge, 1999.

[56] N. Troullier and José Luís Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43:1993–2006, Jan 1991.

[57] M. Burkatzki, C. Filippi, and M. Dolg. Energy-consistent pseudopotentials for quantum monte carlo calculations. *The Journal of Chemical Physics*, 126(23):234105, 2007.

[58] D. R. Hamann, M. Schlüter, and C. Chiang. Norm-conserving pseudopotentials. *Phys. Rev. Lett.*, 43:1494–1497, Nov 1979.

[59] G. B. Bachelet, D. R. Hamann, and M. Schlüter. Pseudopotentials that work: From h to pu. *Phys. Rev. B*, 26:4199–4228, Oct 1982.

[60] G Kresse, J Hafner, and R J Needs. Optimized norm-conserving pseudopotentials. *J. Phys. Condens. Matter*, 4(36):7451, 1992.

[61] David Vanderbilt. Soft self-consistent pseudopotentials in a generalized eigenvalue formalism. *Phys. Rev. B*, 41:7892–7895, Apr 1990.

[62] Kevin F Garrity, Joseph W Bennett, Karin M Rabe, and David Vanderbilt. Pseudopotentials for high-throughput {DFT} calculations. *Comput. Mater. Sci.*, 81:446–452, 2014.

[63] P. E. Blöchl. Projector augmented-wave method. *Phys. Rev. B*, 50:17953–17979, Dec 1994.

[64] Jaron T. Krogel, Juan A. Santana, and Fernando A. Reboredo. Pseudopotentials for quantum monte carlo studies of transition metal oxides. *Phys. Rev. B*, 93:075143, Feb 2016.

[65] J. R. Trail and R. J. Needs. Smooth relativistic hartree fock pseudopotentials for h to ba and lu to hg. *J. Chem. Phys.*, 122(17):174109, 2005.

[66] Jiawei Xu, Michael J. Deible, Kirk A. Peterson, and Kenneth D. Jordan. Correlation consistent gaussian basis sets for h, bâĂŞne with diracâĂŞfock arep pseudopotentials: Applications in quantum monte carlo calculations. *Journal of Chemical Theory and Computation*, 9(5):2170–2178, 2013. PMID: 26583711.

[67] N. D. Drummond, J. R. Trail, and R. J. Needs. Trail-needs pseudopotentials in quantum monte carlo calculations with plane wave and blip basis sets. *Phys. Rev. B*, 94:165170, Oct 2016.

[68] Andrea Bosin, Vincenzo Fiorentini, Andrea Lastri, and Giovanni B. Bachelet. Local norm-conserving pseudo-hamiltonians. *Phys. Rev. A*, 52:236–257, Jul 1995.

[69] LuboÅą MitÃąÅą, Eric L. Shirley, and David M. Ceperley. Nonlocal pseudopotentials and diffusion monte carlo. *The Journal of Chemical Physics*, 95(5):3467–3475, 1991.

[70] Michele Casula, Saverio Moroni, Sandro Sorella, and Claudia Filippi. Size-consistent variational approaches to nonlocal pseudopotentials: Standard and lattice regularized diffusion monte carlo methods revisited. *J. Chem. Phys.*, 132(15):154113, 2010.

[71] Michele Casula, Saverio Moroni, Sandro Sorella, and Claudia Filippi. Size-consistent variational approaches to nonlocal pseudopotentials: Standard and lattice regularized diffusion monte carlo methods revisited. *The Journal of Chemical Physics*, 132(15):154113, 2010.

[72] A. Baldereschi. Mean-value point in the brillouin zone. *Phys. Rev. B*, 7:5212–5215, Jun 1973.

[73] Randolph Q. Hood, M. Y. Chou, A. J. Williamson, G. Rajagopal, and R. J. Needs. Exchange and correlation in silicon. *Phys. Rev. B*, 57:8972–8982, Apr 1998.

[74] Jaron T. Krogel. Nexus: A modular workflow management system for quantum simulation codes. *Comp. Phys. Comm.*, 198:154 – 168, 2016.

[75] Juan A. Santana, Jaron T. Krogel, Paul R. C. Kent, and Fernando A. Reboredo. Cohesive energy and structural parameters of binary oxides of groups iia and iiib from diffusion quantum monte carlo. *J. Chem. Phys.*, 144(17):174707, 2016.

[76] E. U. Franck. J. d. cox, d. d. wagman, v. a. medvedev: Codata key values for thermodynamics, aus der reihe: Codata, series on thermodynamic properties. hemisphere publishing corporation, new york, washington, philadelphia, london 1989. 271 seiten, preis: Âč 28.00. *Ber. der Bunsenges. Phys. Chem.*, 94(1):93–93, 1990.

[77] M Burkatzki, C Filippi, and M Dolg. Energy-consistent pseudopotentials for quantum Monte Carlo calculations. *J. Chem. Phys.*, 126(23):234105, July 2007.

[78] A Togo and I Tanaka. First principles phonon calculations in materials science. *Scr. Mater.*, 108:1–5, Nov 2015.

[79] R Dovesi, R Orlando, B Civalleri, C Roetti, V. R. Saunders, and C. M. Zicovich-Wilson. Crystal: a computational tool for the ab-initio study of the electronic properties of crystals. *Z. Kristallogr.*, 220, 2005.

[80] G. Rajagopal, R. J. Needs, S. Kenny, W. M. C. Foulkes, and A. James. Quantum monte carlo calculations for solids using special $\boldsymbol{k}$ points methods. *Phys. Rev. Lett.*, 73:1959–1962, Oct 1994.

[81] Martin Fuchs and Matthias Scheffler. Ab initio pseudopotentials for electronic structure calculations of polyatomic systems using density functional theory. *Comp. Phys. Comm.*, 119(1):67 – 98, 1999.

[82] Rappe-bennett pseudopotentials. http://www.sas.upenn.edu/rappegroup/research/pseudo-potential-gga.html. Accessed: 2016-10-11.

[83] DM Ceperley and BJ Alder. Ground state of the electron gas by a stochastic method. *Phys. Rev. Lett.*, 45(7):566–569, 1980.

[84] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1):011002, 2013.

[85] R. Nazarov, L. Shulenburger, M. Morales, and Randolph Q. Hood. Benchmarking the pseudopotential and fixed node approximations in diffusion monte carlo calculations of molecules and solids. *Phys. Rev. B*, 93:094111, Mar 2016.

[86] J. P. Perdew and Alex Zunger. Self interaction correction to density functional approximations for many electron systems. *Phys. Rev. B*, 23:5048–5079, May 1981.

[87] M. Burkatzki, C. Filippi, and M. Dolg. Energy consistent pseudopotentials for quantum monte carlo calculations. *J. Chem. Phys.*, 126(23):234105, 2007.

[88] William W. Tipton, Neil D. Drummond, and Richard G. Hennig. Importance of high angular momentum channels in pseudopotentials for quantum monte carlo. *Phys. Rev. B*, 90:125110, Sep 2014.

[89] Norbert Nemec, Michael D. Towler, and R. J. Needs. Benchmark all-electron ab initio quantum monte carlo calculations for small molecules. *The Journal of Chemical Physics*, 132(3):034111, 2010.

[90] Timothy J. Kucharski, Yancong Tian, Sergey Akbulatov, and Roman Boulatov. Chemical solutions for the closed-cycle storage of solar energy. *Energy Environ. Sci.*, 4:4449–4472, 2011.

[91] T. J. Kucharski, N. Ferralis, A. M. Kolpak, J. O. Zheng, D. G. Nocera, and J. C. Grossman. Templated assembly of photoswitches significantly increases the energy-storage capacity of solar thermal fuels. *Nature Chemistry*, 6:441–447, May 2014.

[92] Alexie M. Kolpak and Jeffrey C. Grossman. Azobenzene-functionalized carbon nanotubes as high-energy density solar thermal fuels. *Nano Letters*, 11(8):3156–3162, 2011. PMID: 21688811.

[93] Alexie M. Kolpak and Jeffrey C. Grossman. Azobenzene-functionalized carbon nanotubes as high-energy density solar thermal fuels. *Nano Letters*, 11(8):3156–3162, 2011. PMID: 21688811.

[94] David Zhitomirsky and Jeffrey C. Grossman. Conformal electroplating of azobenzene-based solar thermal fuels onto large-area and fiber geometries. *ACS Applied Materials & Interfaces*, 8(39):26319–26325, 2016. PMID: 27611884.

[95] Eugene N. Cho, David Zhitomirsky, Grace G. D. Han, Yun Liu, and Jeffrey C. Grossman. Molecularly engineered azobenzene derivatives for high energy density solid-state solar thermal fuels. *ACS Applied Materials & Interfaces*, 9(10):8679–8687, 2017. PMID: 28234453.

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

[97] Mikael J. Kuisma, Angelica M. Lundin, Kasper Moth-Poulsen, Per Hyldgaard, and Paul Erhart. Comparative ab-initio study of substituted norbornadiene-quadricyclane compounds for solar thermal storage. *The Journal of Physical Chemistry C*, 120(7):3635–3645, 2016. PMID: 26966476.

[98] Maria Quant, Anders Lennartson, Ambra Dreos, Mikael Kuisma, Paul Erhart, Karl BÃűrjesson, and Kasper Moth-Poulsen. Low molecular weight norbornadiene derivatives for molecular solar-thermal energy storage. *Chemistry âĂŞ A European Journal*, 22(37):13265–13274, 2016.

[99] Ben Feringa. *Molecular Switches*, chapter Optoelectronic Molecular Switches Based on DHA-VHF, pages 63–106. Wiley-VCH Verlag GmbH & Co. KGaA, 2011.

[100] H Goerner, Christian Fischer, S Gierisch, and Jorg Daub. Vinylheptafulvene photochromism: effects of substituents, solvent, and temperature in the photorearrangement of dihydroazulenes to vinylheptafulvenes. *J. Phys. Chem.*, 97:4110–4117, 1993.

[101] Martial Boggio-Pasqua, Michael J Bearpark, Patricia a Hunt, and Michael a Robb. Dihydroazulene/vinylheptafulvene photochromism: a model for one-way photochemistry via a conical intersection. *J. Am. Chem. Soc.*, 124(7):1456–70, February 2002.

[102] Vincent De Waele, Uli Schmidhammer, Thomas Mrozek, Jorg Daub, and Eberhard Riedle. Ultrafast bidirectional dihydroazulene/vinylheptafulvene (DHA/VHF) molecular switches: photochemical ring closure of vinylheptafulvene proven by a two-pulse experiment. *J. Am. Chem. Soc.*, 124(11):2438–9, March 2002.

[103] Soren Lindbaek Broman and Mogens Brondsted Nielsen. Dihydroazulene: from controlling photochromism to molecular electronics devices. *Phys. Chem. Chem. Phys.*, 16:21172–21182, 2014.

[104] H Goerner and Christian Fischer. Dihydroazulene/vinylheptafulvene photochromism: effects of substituents, solvent, and temperature in the photorearrangement of dihydroazulenes to vinylheptafulvenes. *J. Phys. Chem.*, 97:4110–4117, 1993.

[105] Søren Lindbaek Broman, Michael Axman Petersen, Christian G Tortzen, Anders Kadziola, Kristine Kilså, and Mogens Brø ndsted Nielsen. Arylethynyl derivatives of the dihydroazulene/vinylheptafulvene photo/thermoswitch: tuning the switching event. *J. Am. Chem. Soc.*, 132(26):9165–74, July 2010.

[106] Soren Lindbaek Broman, Anne Ugleholdt Petersen, Christian Gregers Tortzen, Johan Vibenholt, Andrew D Bond, and Mogens Brondsted Nielsen. A Bis (heptafulvenyl) -dicyanoethylene Thermoswitch with Two Sites for Ring Closure. *Org. Lett.*, 14(1):318–321, 2012.

[107] Oliver Schalk, Søren L Broman, Michael ÅPetersen, Dmitry V Khakhulin, Rasmus Y Brogaard, Mogens Brø ndsted Nielsen, Andrey E Boguslavskiy, Albert Stolow, and Theis I Sø lling. On the condensed phase ring-closure of vinylheptafulvalene and ring-opening of gaseous dihydroazulene. *J. Phys. Chem. A*, 117(16):3340–7, April 2013.

[108] Anne S. Hansen, Kasper Mackeprang, SÃÿren L. Broman, Mia Harring Hansen, Anders S. Gertsen, Jens V. Kildgaard, Ole Faurskov Nielsen, Kurt V. Mikkelsen, Mogens BrÃÿndsted Nielsen, and Henrik G. Kjaergaard. Characterisation of dihydroazulene and vinylheptafulvene derivatives using raman spectroscopy: The cn-stretching region. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 161:70 – 76, 2016.

[109] Anders S. Gertsen, Stine T. Olsen, SÃÿren L. Broman, Mogens BrÃÿndsted Nielsen, and Kurt V. Mikkelsen. A dft study of multimode switching in a combined dha/vhf-dte/dhb system for use in solar heat batteries. *The Journal of Physical Chemistry C*, 121(1):195–201, 2017.

[110] Stine T. Olsen, Jonas Elm, Freja EilsÃÿ Storm, Aske NÃÿrskov Gejl, Anne S. Hansen, Mia Harring Hansen, Jens Rix Nikolajsen, Mogens BrÃÿndsted Nielsen, Henrik G. Kjaergaard, and Kurt V. Mikkelsen. Computational methodology study of the optical and thermochemical properties of a molecular photoswitch. *J. Phys. Chem. A*, 119(5):896–904, 2015. PMID: 25569127.

[111] Mia Harring Hansen, Jonas Elm, Stine T. Olsen, Aske NÃÿrskov Gejl, Freja E. Storm, Benjamin N. Frandsen, Anders B. Skov, Mogens BrÃÿndsted Nielsen, Henrik G. Kjaergaard, and Kurt V. Mikkelsen. Theoretical investigation of substituent effects on the dihydroazulene/vinylheptafulvene photoswitch: Increasing the energy storage capacity. *The Journal of Physical Chemistry A*, 120(49):9782–9793, 2016. PMID: 27973809.

[112] Martina Cacciarini, Anders B. Skov, Martyn Jevric, Anne S. Hansen, Jonas Elm, Henrik G. Kjaergaard, Kurt V. Mikkelsen, and Mogens BrÃÿndstedâĂĚNielsen. Towards solar energy storage in the photochromic dihydroazuleneâĂŞvinylheptafulvene system. *Chem. Eur. J.*, 21(20):7454–7461, 2015.

[113] Jorg Daub, Thomas Knochel, and Albrecht Mannschreck. Photosensitive dihydroazulenes with chromogenic properties. *Angew. Chem.*, 94(1972):960–961, 1984.

[114] Yan Zhao and Donald G. Truhlar. The m06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four m06-class functionals and 12 other functionals. *Theor. Chem. Acc.*, 120(1):215–241, 2008.

[115] M. J. Frisch et al. Gaussian-09 Revision D.01. Gaussian Inc. Wallingford CT 2009.

[116] JÃűrg Daub, Sebastian Gierisch, Ulrich Klement, Thomas KnÃűchel, Gerhard Maas, and Ulrich Seitz. Lichtinduzierte reversible reaktionen: Synthesen und eigenschaften photochromer 1,1-dicyan-1,8a-dihydroazulene und thermochromer 8-(2,2-dicyanvinyl)heptafulvene. *Chemische Berichte*, 119(8):2631–2646, 1986.

[117] Chunyang Peng, Philippe Y. Ayala, H. Bernhard Schlegel, and Michael J. Frisch. Using redundant internal coordinates to optimize equilibrium geometries and transition states. *Journal of Computational Chemistry*, 17(1):49–56, 1996.

[118] Chr. Møller and M. S. Plesset. Note on an approximation treatment for many-electron systems. *Phys. Rev.*, 46:618–622, Oct 1934.

[119] Donald G. Truhlar. Basis set extrapolation. *Chem. Phys. Lett.*, 294:45–48, 1998.

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

[121] A. Badinski and R. J. Needs. Total forces in the diffusion monte carlo method with nonlocal pseudopotentials. *Phys. Rev. B*, 78:035134, Jul 2008.

[122] Manolo C. Per, Kelly A. Walker, and Salvy P. Russo. How important is orbital choice in single-determinant diffusion quantum monte carlo calculations? *J. Chem. Theory Comput.*, 8(7):2255–2259, 2012. PMID: 26588958.

[123] Christopher M. Beaudry, Jeremiah P. Malerich, and Dirk Trauner. Biosynthetic and biomimetic electrocyclizations. *Chemical Reviews*, 105(12):4757–4778, 2005. PMID: 16351061.

[124] Matteo Barborini and Leonardo Guidoni. Reaction pathways by quantum monte carlo: Insight on the torsion barrier of 1,3-butadiene, and the conrotatory ring opening of cyclobutene. *The Journal of Chemical Physics*, 137(22):224309, 2012.

[125] Data from nist computational chemistry comparison and benchmark database. ccsd(t) calculations are performed at ccsd(t)// mp2fc/6-31g* level. http://cccbdb.nist.gov/.

[126] Estimated using heat of formations of 1,3-butadiene (h_f=26.00 $\pm$ 0.19 kcal/mol and 26.75 $\pm$ 0.23 kcal/mol), ethylene (h_f=12.54 kcal/ mol), and cyclohexene (h_f= -1.03 $\pm$ 0.23 kcal/mol). data from nist standard reference database. http://webbook.nist.gov/chemistry.

[127] Nasir Shahzad, Riffat Un Nisa, and Khurshid Ayub. Substituents effect on thermal electrocyclic reaction of dihydroazulene–vinylheptafulvene photoswitch: a dft study to improve the photoswitch. *Struct. Chem.*, 24(6):2115–2126, 2013.

[128] Min-Cheol Kim, Eunji Sim, and Kieron Burke. Understanding and reducing errors in density functional calculations. *Phys. Rev. Lett.*, 111:073003, Aug 2013.

[129] Aron J. Cohen, Paula Mori-Sanchez, and Weitao Yang. Challenges for density functional theory. *Chem. Rev.*, 112(1):289–320, jan 2012.

[130] Jianmin Tao, John P. Perdew, Viktor N. Staroverov, and Gustavo E. Scuseria. Climbing the density functional ladder: Nonempirical metageneralized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.*, 91:146401, Sep 2003.

[131] Joachim Paier, Martijn Marsman, and Georg Kresse. Why does the b3lyp hybrid functional fail for metals? *J. Chem. Phys.*, 127(2), 2007.

[132] Matthias Rupp, Tkatchenko, et al. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012.

[133] Paul Raccuglia, Katherine C. Elbert, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601):73–76, may 2016.

[134] John R. Koza. Hierarchical genetic algorithms operating on populations of computer programs. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'89, pages 768–774, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[135] John R. Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and Computing*, 4(2):87–112, 1994.

[136] Milan Keser and Samuel I Stupp. Genetic algorithms in computational materials science and engineering: simulation and design of self-assembling materials . *Computer Methods in Applied Mechanics and Engineering*, 186(2âĂŞ4):373 – 385, 2000.

[137] Wojciech Paszkowicz. Genetic algorithms, a nature-inspired tool: A survey of applications in materials science and related fields: Part ii. *Materials and Manufacturing Processes*, 28(7):708–725, 2013.

[138] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, 324(5923):81–85, 2009.

[139] Adri C. T. van Duin, Siddharth Dasgupta, Francois Lorant, and William A. Goddard. ReaxFF A Reactive Force Field for Hydrocarbons. *The Journal of Physical Chemistry A*, 105(41):9396–9409, 2001.

[140] Catherine E Check and Thomas M Gilbert. Progressive systematic underestimation of reaction energies by the B3LYP model as the number of C-C bonds increases: why organic chemists should use multiple DFT models for calculations involving polycarbon hydrocarbons. *J. Org. Chem.*, 70(24):9828–34, November 2005.

[141] H. Lee Woodcock, Henry F. Schaefer, and Peter R. Schreiner. Problematic Energy Differences between Cumulenes and Poly-ynes: Does This Point to a Systematic Improvement of Density Functional Theory? *J. Phys. Chem. A*, 106(49):11923–11931, December 2002.

[142] Larry A. Curtiss et al. Assessment of gaussian-2 and density functional theories for the computation of enthalpies of formation. *J. Chem. Phys.*, 106(3):1063–1079, 1997.

[143] Amir Karton et al. Highly accurate first-principles benchmark data sets for the parametrization and validation of density functional and other approximate methods. derivation of a robust, generally applicable, double-hybrid functional for thermochemistry and thermochemical kinetics. *J. Phys. Chem. A*, 112(50):12868–12886, 2008.

[144] Roberto Todeschini and Viviana Consonni. *Handbook of Molecular Descriptors*. Wiley-VCH Verlag GmbH, 2008.

[145] Kevin S. Brown et al. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E*, 68:021904, Aug 2003.

[146] W Kohn and LJ Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 385(1951), 1965.

[147] John P. Perdew et al. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.*, 105(22):9982, 1996.

[148] Viktor Staroverov et al. Comparative assessment of a new nonempirical density functional: Molecules and hydrogen-bonded complexes. *J. Chem. Phys.*, 119(23):12129, December 2003.

[149] Takeshi Yanai et al. A new hybrid exchange correlation functional using the Coulomb attenuating method (CAM-B3LYP). *Chem. Phys. Lett.*, 393(1-3):51–57, July 2004.

[150] Yan Zhao et al. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other function. *Theor. Chem. Acc*, 120(1-3):215–241, July 2007.

[151] Carlo Adamo et al. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.*, 110(13):6158, 1999.

[152] Jianmin Tao et al. Climbing the Density Functional Ladder: Nonempirical Meta Generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.*, 91(14):146401, September 2003.

[153] RB Woodward and R Hoffmann. Stereochemistry of Electrocyclic Reactions Sir. *J. Am. Chem. Soc.*, 1(1959):395–397, 1964.

[154] Erin R Johnson, Paula Mori-Sánchez, Aron J Cohen, and Weitao Yang. Delocalization errors in density functionals and implications for main-group thermochemistry. *J. Chem. Phys.*, 129(20):204112, November 2008.

[155] Yan Zhao and Donald G Truhlar. A density functional that accounts for medium-range correlation energies in organic chemistry. *Org. Lett.*, 8(25):5753–5, December 2006.

[156] Martin Korth et al. "Mindless" DFT Benchmarking. *J. Chem. Theory Comput.*, 5(4):993–1003, 2009.

[157] Moshe Ben-Akiva and Steven Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, USA, 1985.