# Development and application of statistical and quantum mechanical methods for modelling molecular ensembles

By

Ellen T. Swann

A thesis submitted for the degree of Doctor of Philosophy of the Australian National University

Research School of Chemistry

February, 2018

# Declaration

This thesis is an account of research undertaken between February 2014 and June 2017 as part of a collaboration between CSIRO and ANU.

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university. This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

Ellen T. Swann  
February, 2018

# Acknowledgements

First and foremost I would like to thank my amazing supervisors Michelle Coote, Amanda Barnard and Manolo Per. Your support and guidance over the last three and a half years has been greatly appreciated. This thesis and PhD would not have been possible without you.

I would also like to thank past and present VNLab members, for all they have taught me. Thank you especially to Deidre, for her help with all things QMC related and teaching me SQL and Michael, for teaching me archetypal analysis.

Finally, a massive thank you to Mum, Dad, Hillary, Hugh and Ned for their unwavering love and support. I'd also like to thank my family and friends, for everything they've done to help me get through the last 8 years of study. Lastly, thank you to Han, for keeping me sane despite the puns.

# Abstract

The development of new quantum chemical methods requires extensive benchmarking to establish the accuracy and limitations of a method. Current benchmarking practices in computational chemistry use test sets that are subject to human biases and as such can be fundamentally flawed. This work presents a thorough benchmark of diffusion Monte Carlo (DMC) for a range of systems and properties as well as a novel method for developing new, unbiased test sets using multivariate statistical techniques. Firstly, the hydrogen abstraction of methanol is used as a test system to develop a more efficient protocol that minimises the computational cost of DMC without compromising accuracy. This protocol is then applied to three test sets of reaction energies, including 43 radical stabilisation energies, 14 Diels-Alder reactions and 76 barrier heights of hydrogen and non-hydrogen transfer reactions. The average mean absolute error for all three databases is just 0.9 kcal/mol.

The accuracy of the explicitly correlated trial wavefunction used in DMC is demonstrated using the ionisation potentials and electron affinities of first- and second-row atoms. A multi-determinant trial wavefunction reduces the errors for systems with strong multi-configuration character, as well as for predominantly single-reference systems. It is shown that the use of pseudopotentials in place of all-electron basis sets slightly increases the error for these systems. DMC is then tested with a set of eighteen challenging reactions. Incorporating more determinants in the trial wavefunction reduced the errors for most systems but results are highly dependent on the active space used in the CISD wavefunction. The accuracy of multi-determinant DMC for strongly multi-reference systems is tested for the isomerisation of diazene. In this case no method was capable of reducing the error of the strongly-correlated rotational transition state.

Finally, an improved method for selecting test sets is presented using multivariate statistical techniques. Bias-free test sets are constructed by selecting archetypes and prototypes based on numerical representations of molecules. Descriptors based on the one-, two- and three-dimensional structures of a molecule are tested. These new test sets are then used to benchmark a number of methods.

# Contents

# List of Figures

# List of Tables

Chapter 1

# Introduction

Quantum chemical methods have become an integral part of the chemistry field over the last 50 years. Advances in drug discovery[1,2] and high-throughput material screening[3] would not have been possible without them but their impact is much broader. They allow hypothetical and unobserved systems to be studied in great detail. The foundation of all computational chemistry methods lies in solving the time-independent Schrödinger wave equation (SWE),

$$\hat{H}\Psi = E\Psi \tag{1.1}$$

where $\hat{H}$ is the electronic Hamiltonian operator, $\Psi$ is the wavefunction solution and $E$ is the electronic energy of the system of interest. If the Hamiltonian and wavefunction are known then virtually any property can be calculated. The correlated nature of sub-atomic interactions means the Schrödinger equation can only be solved exactly for systems with one or two electrons. Approximations must be made to solve it for larger systems. Traditional electronic structure methods like Hartree Fock (HF) theory and density functional theory (DFT) are well established and chemically accurate results have been obtained for a range of systems.[4–7] Unfortunately this accuracy comes with a large cost. Post-HF CCSD(T) scales as $N^7$ with respect to system size, $N$. DFT methods are more affordable but their performance is unreliable; accurate DFT results for small systems don't necessarily translate for large systems.[8] A promising alternative to these methods is quantum Monte Carlo (QMC).

QMC uses stochastic integration and has greater freedom in the form of the trial wavefunction compared to *ab initio* wavefunction theory (WFT) or DFT methods. The commonly-used Slater-Jastrow trial wavefunction explicitly accounts for the static and dynamic correlation in a system. By using statistical sampling QMC methods are intrinsically parallelisable and scale nearly perfectly with the number of available cores. They are ideally positioned to take full advantage of the new wave of parallel computers compared to more traditional methods.

The development of new quantum chemical methods requires extensive benchmarking to demonstrate robustness and identify potential weaknesses and shortcomings. Performance of a method is measured by the error, defined as the difference between the calculated value and some reference value that has been obtained from experiment or high-level electronic structure methods. Commonly used metrics include the mean unsigned error (MUE) or mean absolute deviation (MAD) but other metrics like root mean squared error (RMSE) can be used. Smaller errors are desirable and chemical accuracy is defined as an error of 1 kcal/mol or less. For properties where energy differences are expected to be small, like the relative energies of conformers, this accuracy needs to be on the order of 0.1-0.2 kcal/mol. The outcomes of benchmarking guide users towards the most appropriate method for a specific problem and identify the types of systems where a method might fail. Computational chemistry is increasingly being used for simulations where experiment is impossible and it is essential that we can estimate the accuracy of the calculations in these situations.

Benchmarking is a powerful tool for assessing and comparing the accuracy of electronic structure methods but there are serious limitations to the current methodology. The standard practice in computational chemistry is to benchmark using test sets. There are now hundreds, if not thousands, of these sets available in the literature for a vast range of properties. These test sets have been built using 'chemical intuition' and are biased by how we perceive chemical space. This has led to redundancies in test sets and benchmarking results that are highly dependent on the systems studied. Not using a comprehensive test set to evaluate a method can lead to biases in reported error when one class of reaction is over represented.[9–12] The standard practice has become a cumbersome process requiring thousands of calculations to overcome these challenges. Some effort has been directed at finding small, representative subsets but this is limited to only a few test sets for a handful of properties.[13] It also requires thousands of data points from previous benchmarking studies with the biased test sets.

An alternative method for building these test sets is to use multivariate statistics and remove the human bias entirely. Chemical space can be represented by numerical descriptors based on the structure or physical properties of molecules. Techniques like $k$-means clustering and archetypal analysis are routinely used in other fields to find combinations of points that best represent or summarise large data sets. By developing these techniques for computational chemistry, smaller databases can be created for benchmarking that are designed to provide critical tests of methods without redundancy.

The aim of this work is two-fold. Firstly, an extensive benchmarking study is conducted for diffusion Monte Carlo (DMC). Secondly, a machine-learning approach is explored for building new test sets. This thesis begins by introducing some background theory related to electronic structure methods as they pertain to this work. Chapter 3 describes a small benchmarking study of DMC for the reaction between atomic hydrogen and methanol. This study investigated the effect of a number of parameters within the DMC algorithm to improve the efficiency of the calculation but maintain accuracy. Chapter 4 presents a more comprehensive benchmarking of DMC, specifically focusing on reaction barriers for a range of organic systems. In Chapter 5 the effect of pseudopotentials and multi-determinant trial wavefunctions on the correlation energy recovered by DMC is investigated using the ionisation potentials and electron affinities of first- and second-row atoms. In Chapters

6 and 7 the performance of DMC is investigated more rigorously using systems that are known to be challenging for *ab initio* WFT and DFT methods, including a set of 18 'difficult cases' and the isomerisation of diazene. Finally, Chapter 8 presents a novel method for developing test sets using multivariate statistics instead of chemical intuition. These new test sets are then used to benchmark DMC and DFT methods.

## 1.1 List of publications

1. Swann, E.T., Coote, M.L., Barnard, A.S., Per, M.C., 'Efficient protocol for quantum Monte Carlo calculations of hydrogen abstraction barriers : Application to methanol', *Int. J. Quantum Chem*, **2017**, *117*, 1- 7. (Appendix 1)

2. Swann, E.T., Fernandez, M., Coote, M.L., Barnard, A.S., 'Bias-free chemically diverse test sets from machine learning', *submitted*

## 1.2 References

[1] W. L. Jorgensen, *Science* **2004**, *303*, 1813–1818.

[2] W. L. Jorgensen, *Acc. Chem. Res.* **2009**, *42*, 724–733.

[3] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **2013**, *12*, 191–201.

[4] S. M. Bachrach, *Computational Organic Chemistry*, John Wiley & Sons Inc., **2014**.

[5] F. Jensen, *Introduction to Computational Chemistry 2nd ed.*, Wiley, **2007**.

[6] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models 2nd ed.*, John Wiley & Sons Inc., **2004**.

[7] E. G. Lewars, *Computational Chemistry. Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, Kluwer Academic Publishers, **2003**.

[8] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289–320.

[9] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 155–164.

[10] J. J. P. Stewart, *J. Comput. Chem.* **1989**, *10*, 209–220.

[11] G. I. Csonka, A. Ruzsinszky, J. Tao, J. P. Perdew, *Int. J. Quantum. Chem.* **2005**, *101*, 506–511.

[12] S. Grimme, *J. Phys. Chem. A.* **2005**, *109*, 3067–3077.

[13] B. J. Lynch, D. G. Truhlar, *J. Phys. Chem. A* **2003**, *107*, 3898–3906.

Chapter 2

# Theoretical Methods

## 2.1 Introduction

The Schrödinger wave equation:[1]

$$\hat{H}\Psi = E\Psi \tag{2.1}$$

describes the total energy ($E$) of a system as a function of the Hamiltonian ($\hat{H}$) operating on the wavefunction ($\Psi$). If the Hamiltonian and wavefunction are known then virtually any physical or chemical property of a given molecular system can be calculated. The Hamiltonian is an operator describing the observable energy of the system and can be written in terms of kinetic ($\hat{T}$) and potential ($\hat{V}$) energy operators for electrons (e) and nuclei (n):

$$\hat{H} = \hat{T}_{\mathrm{n}} + \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{nn}} \tag{2.2}$$

This definition ignores relativistic effects but provides a good description for the relatively light first- and second-row atoms.[2] The Born-Oppenheimer approximation[3] is commonly used to simplify wavefunction solutions to the Schrödinger wave equation by decoupling the motion of electrons and nuclei. It assumes the nuclei are infinitely heavy relative to the electrons and the electrons move instantaneously in response to the nuclei. The electronic Hamiltonian, $\hat{H}_{\mathrm{elec}}$, can then be written as:

$$\begin{aligned}\hat{H}_{\mathrm{elec}} &= \hat{T}_{\mathrm{e}} + \hat{V}_{\mathrm{ee}} + \hat{V}_{\mathrm{ne}} + \hat{V}_{\mathrm{nn}} \\ &= -\frac{1}{2}\sum_{i}^{n}\nabla_i^2 + \sum_{i<j}^{n}\frac{1}{r_{ij}} - \sum_{I}^{N}\sum_{i}^{n}\frac{Z_I}{r_{Ii}} + \hat{V}_{\mathrm{nn}}\end{aligned} \tag{2.3}$$

for a system of $n$ electrons and $N$ nuclei, where $r_{ij}$ is the distance between electrons $i$ and $j$ and $Z_I$ is the atomic number of nuclei $I$. The Born-Oppenheimer approximation reduces the Schrödinger wave equation to an electronic problem and wavefunction solutions describe the motion of $n$ electrons moving in a field of $N$ fixed nuclei. Electronic structure theory is primarily concerned with finding the electronic wavefunction solutions and their corresponding energy.

The Schrödinger wave equation is a second-order linear differential equation and exact solutions exist for only a small number of systems. Electronic structure methods use different approximations in a trade off between accuracy and computational cost. Hartree-Fock (HF) theory uses the mean-field approximation but fails to account for electron correlation and post-HF *ab initio* wavefunction theory (WFT) attempts to recover this correlation energy. Density functional theory (DFT) reduces the dimensionality of the problem by using the electron density in place of the wavefunction. Quantum Monte Carlo (QMC) methods use stochastic integration and have much greater freedom in the form of the trial wavefunction. An overview of each method is provided below. More detailed information for *ab initio* and DFT methods can be found in Refs. 2,4–6. Detailed information for QMC methods can be found in Refs. 7–10.

## 2.2 *Ab initio* methods

Hartree-Fock (HF) theory[11] is the foundation of *ab initio* wavefunction methods. It approximates an exact $N$-body wavefunction by using single particle functions (orbitals) to describe the distribution of each electron. The non-relativistic electronic Hamiltonian in Equation 2.3 depends only on the spatial coordinates, $r_i$, of each electron but electrons are also characterised by a spin quantum number. The coordinate $\mathbf{x}_i = (r_i, \sigma_i)$ is used instead to define the spin and three spatial coordinates of an electron $i$. The Pauli exclusion principle states that no two electrons can occupy the same point in configuration space[12] and wavefunction solutions to Equation 2.3 must be antisymmetric with respect to the interchange of any two electrons, such that for a system of $N$ electrons:

$$\Psi(\mathbf{x}_1, ..., \mathbf{x}_m, ..., \mathbf{x}_n, ..., \mathbf{x}_N) = -\Psi(\mathbf{x}_1, ..., \mathbf{x}_n, ..., \mathbf{x}_m, ..., \mathbf{x}_N) \tag{2.4}$$

The HF wavefunction is given by a single $N$-electron Slater determinant:[13]

$$\Psi_{\text{HF}} = |\psi_1, \psi_2, ...\psi_N\rangle = \frac{1}{\sqrt{N!}} \begin{vmatrix} \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & ... & \psi_N(\mathbf{x}_1) \\ \psi_1(\mathbf{x}_1) & \psi_2(\mathbf{x}_1) & ... & \psi_N(\mathbf{x}_1) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_1(\mathbf{x}_N) & \psi_2(\mathbf{x}_N) & ... & \psi_N(\mathbf{x}_N) \end{vmatrix} \tag{2.5}$$

The set of functions $\psi_i$ are the individual one-electron wavefunctions, also called molecular orbitals (MOs), that describe the distribution of an electron as a function of its spin and spatial coordinates $\mathbf{x}$. Interchanging any two electrons in the Slater-determinant will change the sign of $\Psi_{\text{HF}}$, satisfying the antisymmetry requirement. The expectation value of the energy is given by $E = \langle\Psi|\hat{H}|\Psi\rangle$ if $\Psi$ is normalised. For the HF wavefunction:

$$E_{\text{HF}} = \sum_i H_i + \frac{1}{2}\sum_{ij}(J_{ij} - K_{ij}) \tag{2.6}$$

where $H_i$ involves one-electron terms arising from the kinetic energy and nuclear attraction of the electrons, $J_{ij}$ is a two-electron term describing the coulomb repulsion between the electrons and $K_{ij}$ is a two-electron term associated with the exchange of electronic coordinates. The variational theorem states that the energy determined by an approximate wavefunction will always be equal to or greater than the energy of the exact wavefunction. The coefficients of the MOs are optimised to minimise $E_{\mathrm{HF}}$ in a process that is carried out iteratively and is known as the *Self-Consistent Field* (SCF) method.[11, 14–18]

There are several variations of HF theory. Restricted Hartree-Fock (RHF) assumes the spin-up ($\alpha$) and spin-down ($\beta$) electrons of an electron pair are energetically degenerate and assigns them to the same spatial MO. This is a reasonable assumption for closed shell species but open shell systems like radicals have uneven numbers of $\alpha$ and $\beta$ electrons. Unrestricted Hartree-Fock (UHF) method allows separate sets of MOs for $\alpha$ and $\beta$ electrons but this can result in spin-contamination and the wavefunctions are no longer eigenfunctions of $\hat{S}^2$.[19] Restricted open-shell Hartree-Fock (ROHF) pairs $\alpha$ and $\beta$ electrons in a similar manner to RHF but allocates separate MOs for unpaired electrons.[6]

The orbitals used in HF theory are independent of the instantaneous motion of other electrons and introduce an intrinsic error known as the 'correlation error'. The correlation energy is defined as:

$$E_{\mathrm{corr}} = E_{\mathrm{exact}} - E_{\mathrm{HF}} \tag{2.7}$$

The HF energy provides an upper limit on the electronic energy of a system and the correlation energy will always be negative. The correlation error increases with system size and the number of electrons. It accounts for only a small percentage of the total electronic energy of a system but is important when energy differences are considered and post-HF methods are chiefly concerned with recovering this energy. Correlation energy is often characterised as either static or dynamic. Static correlation arises from near-degeneracy effects and is important in systems where different orbitals have similar energies such as stretched bonds and low-lying excited states. Dynamic correlation arises from the instantaneous electron-electron interactions. For systems where dynamic correlation is dominant a single determinant is sufficient and HF provides a reasonable description. For systems with significant static correlation more than one reference determinant should be used.

### 2.2.1 Post-HF methods

#### Configuration interaction

The configuration interaction (CI) method allows excitations from occupied orbitals into one or more unoccupied 'virtual' orbitals.[20] Single excitations ($\Psi_i^a$) promote one electron from an occupied orbital, $i$, into a virtual orbital, $a$. Double excitations promote two electrons and so forth. Examples of some of these excitations are shown in Figure 2.1. These different configurations are then mixed together to give a better approximation of the true wavefunction:

$$\Psi_{\mathrm{CI}} = c_0\Psi_0 + \sum_i c_i^a \Psi_i^a + \sum_{ij} c_{ij}^{ab} \Psi_{ij}^{ab} + \sum_{ijk} c_{ijk}^{abc} \Psi_{ijk}^{abc} + .... \tag{2.8}$$

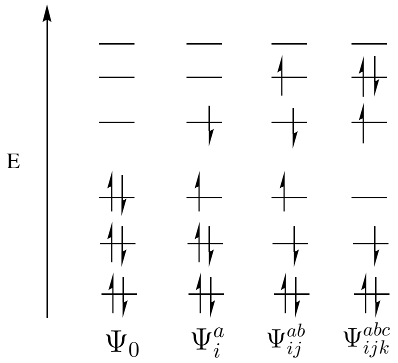


Figure 2.1: Examples of the types of excitations used to generate a configuration interaction wavefunction. Excitations shown include single ($\Psi_i^a$), double ($\Psi_{ij}^{ab}$) and triple ($\Psi_{ijk}^{abc}$) excitations from the Hartree-Fock reference, $\Psi_0$.

where *i, j, k, ...* refer to occupied orbitals and *a, b, c, ...* refer to unoccupied (virtual) levels. The coefficients, $c_i^a$, $c_{ij}^{ab}$, $c_{ijk}^{abc}$, ..., are found using the variational theorem. Configuration state functions (CSFs) are symmetry-adapted linear combinations of Slater determinants and are often used in place of determinants to reduce the number of functions.

Full configuration interaction (FCI) includes all possible configurations for a system with $N$ electrons (up to $N$-fold excitations) and is exact within a given set of basis functions. The cost of FCI increases exponentially and in practice the CI expansion is truncated according to excitation level.[21] In this work CISD (single and double excitations) and CISDTQ (single, double, triple and quadruple excitations) were used. Truncated CI methods are variational but no longer size consistent i.e. the energy of a system with $N$ non-interacting particles is not equal to the sum of the energy of $N$ isolated systems. They recover smaller fractions of the correlation energy as system size increases and can be unreliable for dissociation energies and other energy differences, particularly for larger molecules.[22]

### 2.2.2 Coupled-cluster theory

Coupled-cluster (CC) theory[23–26] defines the true wavefunction as:

$$\Psi_{\mathrm{CC}} = e^{\hat{T}}\Psi_{\mathrm{HF}} \tag{2.9}$$

where $\hat{T}$ is an excitation operator and can be written as a linear combination of excitations, up to $N$-fold excitations for system with $N$ electrons:

$$\hat{T} = \hat{T}_1 + \hat{T}_2 + \hat{T}_3 + ... + \hat{T}_N \tag{2.10}$$

If all excitation levels up to $N$ are included then $\Psi_{\mathrm{CC}}$ is equivalent to $\Psi_{\mathrm{FCI}}$ and is exact within the basis set approximation but the number of excitations is usually truncated at some level. Unlike FCI it includes additional terms to maintain size-consistency but truncated CC is no longer variational. Coupled cluster with single, double and perturbative triple excitations (CCSD(T)[27]) has become the gold-standard in quantum chemical methods and is expected to give results close to the FCI limit for a given basis set. The basis set error is often removed by extrapolating to the complete basis set limit.[2]

### 2.2.3 Multi-reference methods

The methods described above use a single Slater determinant as the reference function. This approach fails for systems where static correlation is important and more than one configuration contributes significantly to the ground state energy. The multi-configuration self consistent field (MCSCF) method begins with linear combination of Slater determinants. It is similar in principle to CI but optimises the MOs used for constructing the determinants as well as the determinant coefficients in an iterative SCF procedure. The Slater determinants included are selected *a priori*, commonly using the complete active space (CAS) method.[28] The CAS wavefunction includes all possible excitations within a set of active orbitals. The near-degeneracies that result in static correlation most often affect the highest occupied and lowest unoccupied orbitals and the active space is usually built using a certain number of these orbitals. The general notation is CAS($n$, $m$), referring to $n$ electrons distributed amongst $m$ active orbitals.

The CASSCF wavefunction does not include dynamic correlation but this can be incorporated with multi-reference configuration interaction (MRCI).[28] MRCI is similar to CI as described above but generates all possible excitations for each determinant in a multi-reference wavefunction. Each determinant is treated equivalently. A cheaper alternative is complete active space second order perturbation theory (CASPT2), where perturbative corrections are made to the CASSCF expansion based on the single and double excitations from every determinant in the active space.[29] Including the perturbations destroys the variationality of MCSCF.

### 2.2.4 Composite high-level methods

Composite methods attempt to reproduce accurate high-level *ab initio* methods at a fraction of the cost. They use a combination of methods and basis sets via addititivity or extrapolation schemes with theoretical or empirical corrections. Examples include the Gaussian (Gn) procedures,[30–37] the complete basis set (CBS) methods[38–41] and the Weizmann (Wn) procedures.[42–44] This work used a variation of the G4 method, G4MP2-X. Results from high-level methods with small basis sets are combined with results from low-level methods with large basis sets to approximate a high-level energy (CCSD(T)) with a large basis set.[45] The electronic energy calculated with G4MP2-X is defined as:

$$\begin{aligned}\mathbf{E_{G4(MP2)-6X}} =& \mathrm{HF/CBS} + E^{\mathrm{corr}}_{\text{SCS-MP2}}/\mathrm{G3MP2LargeXP} + \\ & \Delta E_{\text{S-CCSD}}/6-31G(\mathrm{d}) + E^{\mathrm{corr}}_{\text{S-(T)}}/6-31G(\mathrm{d}) \\ & + \mathrm{HLC} + E_{\mathrm{SO}}\end{aligned} \tag{2.11}$$

where HLC is a high-level correction and $E_{\mathrm{SO}}$ is a spin-orbit correction. More details can be found in Ref. 45. This method has been shown to deliver chemically accurate results when tested on a set of 526 energies including thermochemical properties, reaction energies and barrier heights and weak interactions. Its empirical nature means that good performance for systems similar to the training set does not translate into reliable results for all systems.[46, 47]

## 2.3 Density functional theory

A popular alternative to *ab initio* methods is density functional theory (DFT). The Hohenberg-Kohn thereom states that the electronic energy of the ground state of a system is determined by the one-electron density, $\rho_0(r)$.[48] While *ab initio* methods become increasingly demanding as the number of atoms increases, the one-electron density is always a function of three variables, independent of the number of atoms. The theorem shows that there exists an energy functional that will return the ground state energy for a given $\rho_0(r)$ but does not give the exact form. Instead, modern DFT methods are based on Kohn-Sham theory.[49] The energy functionals used have the form:

$$E(\rho) = E_{\mathrm{T}} + E_{\mathrm{ne}}(\rho) + E_{\mathrm{ee}}(\rho) + E_{\mathrm{XC}}(\rho) \tag{2.12}$$

The first three terms describe the kinetic energy, nuclei-electron attraction and classical electrostatic repulsion and have well-defined functionals. The last term, $E_{\mathrm{XC}}(\rho)$, is the exchange-correlation functional. Its exact form is unknown and defining this functional is the greatest challenge of DFT methods. The treatment of the exchange-correlation energy determines the accuracy and expense of DFT methods. A hierarchy of the approximate treatments of the exchange-correlation term can be classed in a 'Jacob's Ladder'.[50] The exact functional lies at the top of the ladder and the lower five rungs define a set of assumptions made in approximating the exchange-correlation functional.

At the bottom of the ladder is the local density approximation, where $E_{\mathrm{XC}}$ is assumed to depend only on a locally uniform density. A better approximation assumes the density is not locally uniform and $E_{\mathrm{XC}}$ depends on the density and its derivatives. This is known as the generalised gradient approximation (GGA) and constitutes the second rung. The third rung is meta-GGA functionals and includes a term for the Laplacian of the density ($\nabla^2\rho$) or the orbital kinetic energy density ($\tau$). Hybrid (or hyper-GGA) functionals include a dependence on the exact (HF) exchange and lie on the fourth rung. An example is the B3LYP[51, 52] functional, used in this work to generate trial wavefunctions for QMC methods. It uses a generalised gradient approximate to the electron density and mixes HF in the Becke exchange functional with parameterisations. Double-hybrid functionals incorporate the unoccupied Kohn-Sham orbitals and constitute the fifth rung.

## 2.4 Quantum Monte Carlo

Quantum Monte Carlo methods solve the SWE stochastically rather than analytically. Using a stochastic method like Monte Carlo (MC) integration means there is much greater freedom in the choice of trial wavefunction. Electron correlation effects can be explicitly included, allowing for very accurate calculations of molecular properties. One of the greatest advantages of QMC methods is their favourable scaling with respect to system size, $N$, scaling as $\mathrm{O}(N^{3\text{-}4})$ compared to $\mathrm{O}(N^7)$ for the gold-standard CCSD(T).[7]

The two most common QMC methods are variational quantum Monte Carlo (VMC) and diffusion quantum Monte Carlo (DMC). Both are variational and the calculated energy of the trial wavefunction will always be above the true ground-state.[10]

### 2.4.1 Trial wavefunction

The exact form of the trial wavefunction is not known for most systems. Instead, QMC and many *ab initio* methods construct a trial wavefunction ($\Psi_\mathrm{T}$) as an approximation to the true wavefunction.[7] Most MC methods use a Slater-Jastrow trial wavefunction, such that:

$$\Psi_\mathrm{T} = e^J D^\uparrow D^\downarrow \tag{2.13}$$

where $\Psi_\mathrm{T}$ is the trial wavefunction, $D^\uparrow D^\downarrow$ are single-particle Slater determinants and $J$ is the Jastrow factor. The parameters in the trial wavefunction are optimisable. In this work they are optimised by minimising the total energy at the variational Monte Carlo (VMC) level, using the linear method of Toulouse and Umrigar.[53] Other forms of the trial wavefunction include geminals,[54] backflow-transformed determinants,[55] Pfaffians[56] and multi-determinant expansions.[57]

#### Jastrow factor

The Jastrow factor is a function of inter-electron distances and describes the dynamic electron correlation of the system:

$$J = \sum_{i>j} \sum_{A} [J_\mathrm{ee}(r_{ij}) + J_\mathrm{eN}(r_{iA}) + J_\mathrm{eeN}(r_{iA}, r_{jA}, r_{ij})] \tag{2.14}$$

where *i, j* label electrons and $A$ labels nuclei. The general form includes electron-electron ($ee$), electron-nucleus ($eN$) and electron-electron-nucleus ($eeN$) correlation terms,[58] though there are more extensive versions.[59,60] The third term ($J_{eeN}$) has a very small impact on total energy but a significant contribution to computational time, as shown in Chapter 3. A two-term Jastrow factor is used for all DMC calculations presented here unless stated otherwise. The Jastrow factor enforces the electron-electron Kato cusp conditions, where the local kinetic energy must have an equal and opposite divergence to the Coulomb potential as two charged particles approach.[61]

#### Slater determinant

The Slater determinant describes the nodal surface of the system and enforces the electron-nucleus cusp. The Slater determinant is made up of single particle orbitals that are usually obtained from DFT or HF calculations. Studies have shown Kohn-Sham orbitals taken from DFT calculations perform marginally better than HF orbitals.[62,63] Single-determinant wavefunctions can fail to describe near-degeneracy effects but the inclusion of more determinants can better describe the static correlation of a system. These multi-determinant wavefunctions have been successfully applied to a number of systems.[64–74] More detail is provided in Section 2.4.4

### 2.4.2 Variational Monte Carlo (VMC)

Variational Monte Carlo (VMC) uses the Metropolis algorithm[75] to evaluate a trial wavefunction and calculate molecular properties like the total energy. Random moves are proposed from a standard distribution. Moves to points of higher probability are always accepted but moves to regions of lower probability are rejected according to a particular formula obeying a detailed balance equation. Ergodicity is assumed; any point in the configuration space can be reached in a finite number of moves and the distribution of the moving points will converge to the desired probability distribution after an appropriate period of equilibration. For uncorrelated samples the statistical uncertainty in the integral decreases as the square root of the number of sampling points, independent of the dimensionality of the integral and the result converges much faster than standard grid methods such as the trapezoidal rule. The trial wavefunction can be systematically improved by varying its parameters to minimise the energy estimate.[7] The variational energy of a trial wavefunction $\Psi_{\mathrm{T}}$ can be written as the expectation value of the Hamiltonian:

$$E_{\mathrm{V}} = \frac{\int \Psi_{\mathrm{T}}(\mathbf{R})\hat{H}\Psi_{\mathrm{T}}(\mathbf{R})\mathrm{d}\mathbf{R}}{\int \Psi_{\mathrm{T}}^2(\mathbf{R})\mathrm{d}\mathbf{R}} \geq E_0 \tag{2.15}$$

where $E_{\mathrm{V}}$ is the variational energy, $\hat{H}$ is the many-body Hamiltonian, $\Psi_{\mathrm{T}}$ is the trial wavefunction and $\mathbf{R}$ is a $3N$-dimensional vector of particle coordinates. If $\Psi_{\mathrm{T}}$ has correct symmetry under particle exchange, the first derivative is continuous everywhere except where the potential is finite and $\int \Psi_{\mathrm{T}}^2\mathrm{d}\mathbf{R}$ and $\int \Psi_{\mathrm{T}}\hat{H}\Psi_{\mathrm{T}}\mathrm{d}\mathbf{R}$ exist then $E_{\mathrm{V}}$ will always be greater than the exact ground-state energy, $E_0$, providing an upper bound on the energy.

For stochastic evaluation, Equation 2.15 can be re-written using an importance sampling transform, such that :

$$E_{\mathrm{V}} = \int p(\mathbf{R})E_{\mathrm{L}}(\mathbf{R})\mathrm{d}\mathbf{R} \tag{2.16}$$

where $E_{\mathrm{L}}$ is the local energy, expressed as:

$$E_{\mathrm{L}}(\mathbf{R}) = \Psi_{\mathrm{T}}^{-1}\hat{H}\Psi_{\mathrm{T}} \tag{2.17}$$

and $p$ is a probability distribution, expressed as:

$$p(\mathbf{R}) = \frac{\Psi_{\mathrm{T}}^{2}(\mathbf{R})}{\int \Psi_{\mathrm{T}}^{2}(\mathbf{R}')\mathrm{d}\mathbf{R}'} \tag{2.18}$$

This probability distribution is sampled using the the Metropolis algorithm[75] and the total VMC energy is the local energy averaged over the distribution $p(\mathbf{R})$:

$$E_{\mathrm{V}} = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} E_{\mathrm{L}}(\mathbf{R}_i) \tag{2.19}$$

where $M$ is the number of configurations $\mathbf{R}_i$ that have been generated after equilibrium.

The statistical error introduced by the stochastic MC algorithm is proportional to $\frac{1}{\sqrt{M}}$ for $M$ samples. Configurations are serially correlated and a blocking method is used to give a better estimate of the error. Adjacent data points are averaged together to form block averages.[76] This is performed recursively and the number of data points is halved with each iteration. The calculated value of the standard error increases as a function of the number of blocking transformations until a limiting value is reached.

Selecting the normalised square of the trial wavefunction for the probability distribution (Equation 2.18) simplifies Equation (2.16). $\Psi_{\mathrm{T}}$ is an approximate eigenfunction of the Hamiltonian, i.e $\hat{H}\Psi_{\mathrm{T}} \approx E\Psi_{\mathrm{T}}$, but as $\Psi_{\mathrm{T}}$ approaches the exact eigenfunction the variance of the local energy approaches zero (i.e. $\frac{\hat{H}\Psi_0}{\Psi_0} = E_0$, where $E_0$ is the ground state energy of the system (a constant)). This is the zero-variance property; $E_{\mathrm{L}}$ becomes a smoother function of $\mathbf{R}$ as $\Psi_{\mathrm{T}}$ is improved, reducing the number of sampling points required for an accurate estimate of $E_{\mathrm{v}}$. Unlike other QMC variations VMC is not affected by the fermion sign problem and the accuracy of the VMC method will always be limited by the quality of the trial wavefunction. It can be challenging to ensure equivalent wavefunctions for different systems, leading to inaccurate estimates of energy differences.

### 2.4.3 Diffusion Monte Carlo (DMC)

Diffusion Monte Carlo (DMC) is variational like VMC but its accuracy is not dependent on the form of the trial wavefunction. It propagates the time-dependent Schrödinger wave equation (Equation 2.20) through imaginary time to extract the true ground-state wavefunction, $\Psi_0$:

$$i\hbar\frac{\delta\Psi(\mathbf{R},t)}{\delta t} = \hat{H}\Psi(\mathbf{R},t) \tag{2.20}$$

Substituting $\tau = it/\hbar$ in (Equation 2.20) transforms it into a diffusion equation:

$$\begin{aligned}\frac{\delta\Psi(\mathbf{R},\tau)}{\delta\tau} &= -(\hat{H} - E_{\mathrm{R}})\Psi(\mathbf{R},\tau)\\ &= -(\hat{T} + (\hat{V}(\mathbf{R}) - E_{\mathrm{R}}))\Psi(\mathbf{R},\tau)\\ &= -(\frac{1}{2}\nabla_{\mathbf{R}}^{2} + (\hat{V}(\mathbf{R}) - E_{\mathrm{R}}))\Psi(\mathbf{R},\tau)\end{aligned} \tag{2.21}$$

where $E_\mathrm{R}$ is the reference energy, an arbitrary offset. The electronic Hamiltonian ($\hat{H}$) has been expanded into kinetic and potential energy terms. The wavefunction $\Psi(\mathbf{R}, \tau)$ can be expanded in eigenstates $\psi_i$ of the hamiltonian, such that:

$$\Psi(\mathbf{R}, \tau) = \sum_i e^{-(E_i - E_\mathrm{R})\tau} c_i(0) \psi_i(\mathbf{R}) \tag{2.22}$$

where $E_i$ is an eigenvalue. This will converge on the ground state ($\Psi_0$) in the limit $\tau \rightarrow \infty$ if $E_\mathrm{R} = E_0$ as the excited states have larger $E_i$ values and will decay rapidly. In principle DMC is an exact method but in reality the ground state it converges on is the nodeless bosonic solution. Antisymmetry constraints must be imposed for the solution to converge on the fermionic ground state.

Equation 2.21 can be written in integral form using the Greens function:

$$\Psi(\tau, \mathbf{R}) = \int G(\tau, \mathbf{R}', \mathbf{R}) \Psi(\mathbf{R}') \mathrm{d}\mathbf{R}' \tag{2.23}$$

where $G(\tau, \mathbf{R}', \mathbf{R}) = \left\langle \mathbf{R} | e^{-(\hat{H} - E_\mathrm{R})\tau} | \mathbf{R}' \right\rangle$ is the Green's function describing the propagation from $\mathbf{R}'$ to $\mathbf{R}$ in imaginary time $\tau$ and $\Psi(\mathbf{R}')$ is the initial trial wavefunction. Green's function Monte Carlo (GFMC)[77,78] samples this Green's function directly but the algorithm is too computationally expensive for almost all systems. The Trotter formula[79] can be used to approximate the propagator in terms of kinetic and potential enregy, such that:

$$\begin{aligned} (e^{-(\hat{H} - E_\mathrm{R})\Delta\tau})^N &= (e^{-(\hat{T} + (\hat{V} - E_\mathrm{R}))\Delta\tau})^N \\ &\approx (e^{-\hat{T}\Delta\tau} e^{-(\hat{V} - E_\mathrm{R})\Delta\tau})^N \end{aligned} \tag{2.24}$$

with timestep $\Delta\tau = \tau/N$ for N timesteps, assuming $\Delta\tau$ to be small. Since the kinetic ($\hat{T}$) and potential ($\hat{V}$) energy operators do not commute this approximation introduces a time-step error when $\Delta\tau$ is non-zero. This bias is corrected for by using different values of $\Delta\tau$ and extrapolating to $\Delta\tau \rightarrow 0$. The initial trial wavefunction $\Psi(\mathbf{R}')$ can be taken from an optimised VMC wavefunction but a more efficient solution uses the importance sampling transform and samples from the mixed distribution:

$$f(\mathbf{R}, \tau) = \Psi_0(\mathbf{R}, \tau) \Psi_\mathrm{T}(\mathbf{R}, \tau) \tag{2.25}$$

where $\Psi_\mathrm{T}(\mathbf{R}, \tau)$ is the guiding wavefunction (trial wavefunction).[80] The stochastic realisation of Equation 2.24 models $\hat{T}$ as a diffusion process and $\hat{V}$ as a branching process. Walkers are distributed according to $f(\mathbf{R}, 0) = |\Psi_\mathrm{T}(\mathbf{R})|^2$ with equal weights ($w_k = 1$) before undergoing a drift-diffusion process and the weight is updated according to $w_k(\tau + \Delta\tau) = w_k(\tau) e^{-\Delta\tau(E_\mathrm{L}(\mathbf{R}) - E_\mathrm{R})}$. A stochastic birth-death process is used to control the number of walkers. In the simplest implementation, walkers with weight greater than 1 are duplicated with probability $w_k - 1$ and walkers with weight less than one are killed with probability $1 - w_k$.

The ground state energy can then be calculated:

$$\begin{aligned} E_{\mathrm{DMC}} &= \lim_{\tau\to\infty} \frac{\int \Psi(\mathbf{R},\tau)\hat{H}\Psi_{\mathrm{T}}(\mathbf{R})\mathrm{d}\mathbf{R}}{\int \Psi(\mathbf{R},\tau)\Psi_{\mathrm{T}}(\mathbf{R})\mathrm{d}\mathbf{R}} \\ &= \lim_{\tau\to\infty} \frac{\int f(\mathbf{R},\tau)E_L(\mathbf{R})\mathrm{d}\mathbf{R}}{\int f(\mathbf{R},\tau)\mathrm{d}\mathbf{R}} \\ &= \frac{1}{M}\sum_{M} E_L(\mathbf{R}_M) + \mathcal{O}(1/\sqrt{M}) \end{aligned} \tag{2.26}$$

for M configurations. The statistical uncertainty on the final $E_{\mathrm{DMC}}$ value can be reduced by running for longer periods.

### Fixed-node approximation

DMC is an exact method within statistical error bars but will converge on the bosonic ground state rather than the antisymmetric fermionic solution. The most common solution is the fixed-node approximation (FNA) where DMC solutions are restricted to having the same nodes as the trial wavefunction. These nodes are enforced by rejecting any moves where a walker would cross a node.

The FNA introduces a systematic error when the nodal surface is not exact and is the biggest limitation on the accuracy of fixed-node DMC (FNDMC). Significant effort has been directed towards improving the nodes of the trial wavefunction but the structure of these nodal surfaces is still relatively unknown.[81] Starting orbitals can be generated from canonical HF orbitals, Kohn-Sham orbtials from DFT or natural orbitals from post-HF methods. An obvious solution to the FNA is to simply use better starting orbitals and it has been shown that Kohn-Sham orbitals offer a better starting point.[62,63] Multi-determinant wavefunctions can improve the nodes but only if the coefficients have been reoptimised in the presence of a correlation factor.[81,82]

The nodes can be systematically improved by optimising the Slater determinants in the presence of a correlation function[83] or the nodal surfaces can be optimised directly, as is the case in self-healing DMC.[84,85] Another alternative is the released node method,[86,87] where the fixed-node constraint is relaxed and the exact ground-state energy is estimated by including a factor of -1 for each walker that crosses the nodal surface. The cancellation between the positive and negative contributions to the averages increases with the number of walkers, resulting in rapid growth of the variance as the denominator approaches zero. The rate of error growth increases with the difference in the energy between the fermi and bose ground states and is too large for use, even in small systems.[88] Other wavefunction forms can be used that include correlation more directly than the sums of Slater determinants. Examples are the antisymmetrized geminal power functions,[54] valence-bond,[89] Pfaffian forms[56] and back-flow-transformed determinants.[55] Another alternative is Fermion Monte Carlo (FMC)[90] (also referred to as exact QMC[91]) which is independent of the nodes of the trial function. Instead, cancellation of positive and negative walkers is used to maintain antisymmetry. Although it is an exact method in general it is unstable and at present FMC methods are not practical for large systems.[92]

### 2.4.4 Multi-determinant DMC

The standard trial wavefunction used in DMC calculations has one Slater determinant but the nodal surface has the potential to be improved by including more determinants. In this case, the Slater-Jastrow wavefunction can be written as:

$$\Psi_{\mathrm{T}} = e^{J} \sum_{n} c_n D_n^{\uparrow} D_n^{\downarrow} \tag{2.27}$$

where $c_n$ are coefficients and $D_n^{\uparrow} D_n^{\downarrow}$ are the Slater determinants taken from a multi-determinant wavefunction. These methods generate too many determinants, or configuration state functions (CSFs, spin- and space-symmetry adapted linear combinations of determinants) to be used practically in a DMC methods and the expansions are usually truncated according to some threshold. Traditionally the number of CSFs has been selected by choosing a fixed number of terms[93] or using a threshold value on the CI coefficients.[66, 68, 74, 94, 95]

This work used two different truncation schemes; a weight-based scheme and an energy-based one. The weight-based scheme arranges the CSFs in order of their coefficients, from largest to smallest absolute value. The CSFs are progressively included until sum of the squares of the CSF coefficients (i.e. the norm of the expansion) is equal to some threshold value. This removes a significant number of CSFs with small coefficients that make relatively small contributions to the total wavefunction. The energy-based truncation estimates the contribution each determinant (or CSF) makes to the total energy of the multi-determinant calculation. CSFs are ordered according to their energetic contribution and then summed in a similar manner as the weights until a threshold value is reached.

## 2.5 Basis sets

The molecular orbitals used in *ab initio* and DFT methods are built up using linear combinations of basis functions to describe the electron distribution of atomic orbitals (AOs):

$$\phi = \sum_{i=1}^{N} a_i \psi_i \tag{2.28}$$

where $\phi$ is the molecular orbital and $\psi$ is a basis function describing the atomic orbital with an associated coefficient $a_i$. The set of $N$ basis functions $\psi_i$ is a basis set. Slater-type orbitals (STOs) use functions from the exact solution of the Schrödinger equation for the hydrogen atom and have the form $e^{-\zeta r}$, where $r$ is the distance from the nucleus and $\zeta$ is an exponent. Gaussian-type orbitals (GTOs)[20] use gaussian functions of the form $e^{-\zeta r^2}$ and require more primitives to describe the wavefunction than STO basis sets but they are easier and more efficient to solve. Most calculations use GTO basis sets. The simplest AO representation uses just one function per orbital (which can have multiple primitives), known as a minimal basis set. A double-zeta basis set uses two basis functions to represent each orbital, a triple-zeta uses three and so forth. Larger basis sets give a better description of the AOs but come with a higher computational cost.

Split-valence basis sets use one basis function for the core but two or more basis functions with different exponents for each valence orbital.[96,97] Polarisation functions with higher angular momentum and diffuse basis functions with smaller exponents can also be included. Correlation-consistent basis sets[98–100] are designed to systematically approach the complete basis set (CBS) limit. This work used the Roos augmented triple zeta basis set,[101] which have been constructed using the atomic natural orbital approach.

### 2.5.1 Pseudopotentials

QMC methods scale as approximately $N^{3-4}$ with respect to system size, $N$ but this increases to approximately $Z^{5.5-6.5}$ with respect to atomic number, $Z$.[102,103] The variance in energy of a QMC calculation is determined by the largest energy scale present. For most systems the core electrons have little effect on the valence electronic structure but their large energy fluctuations means the majority of the computational time would be spent sampling the core.[104] Pseudopotentials replace these chemically inert core electrons with an effective potential such that the valence electrons still feel the same electric field but the high-energy core electrons are removed. The total energy of the system is lowered and the energy fluctuations are reduced. Scalar relativistic effects are important beyond third-row atoms (and even some third-row atoms) and these can be included in the pseudopotential.

The true Hamiltonian, $\hat{H}$, is replaced with an effective Hamiltonian, $\hat{H}^{\text{eff}}$, of the form:

$$\hat{H}^{\text{eff}} = K + V_{\text{loc}} + V_{\text{non-loc}} \tag{2.29}$$

where $K$ is the kinetic energy. The local potential, $V_{\text{loc}}$, depends only on the distance of the electron from the nucleus but the non-local potential, $V_{\text{non-loc}}$, is different for each angular-momentum. The use of pseudopotentials in VMC is quite straightforward[105,106] but their non-local component is incompatible with DMC. A locality approximation can be made where the non-local part of pseudopotential is taken to act on the trial wavefunction rather than the DMC wavefunction, introducing singularities in the nodal regions of the trial wavefunctions.[107] This approximation destroys the variational property of the algorithm. An alternative semi-localisation scheme, known as 'T-moves', essentially 'pushes' walkers away from divergences in the non-local pseudopotential. This scheme restores the variational property and has better numerical stability than the locality approximation.[108]

The standard pseudopotentials used in DFT and HF methods have a singularity at the nucleus and can result in large time-step errors in QMC. QMC-specific pseudopotentials have been designed that are soft and have no singularities at the nucleus. The non-local potential of these QMC-specific pseudopotentials decays quickly away from the nucleus to reduce the cost of the numerical integration. These pseudopotentials cannot be generated within a QMC framework and are created from external sources, generally using HF or Dirac-Fock calculations. The first QMC-specific pseudopotentials were generated for the carbon atom from a HF starting point by Greeff et al.[109] This procedure was then extended to all first- and second-row elements.[110] Trail and Needs introduced singularity-free relativistic pseudopotentials for most of the periodic table based on Dirac-Fock calculations,[111,112] however these spin-orbit pseudopotentials only have basis sets for hydrogen and the atoms B to Ne.[113] Burkatzki et al. have also created non-singular energy-consistent scalar-relatvisitic HF pseudopotentials and basis sets for main group elements[114] as well as $3d$-transition metals.[115] The accuracy of these pseudopotentials have been demonstrated for a number of systems.[112,114–118] This work made use of Burkatzki-Fillipi-Dolg[114,115] (BFD) and Trail-Needs[111] (TN) pseudopotentials.

## 2.6 Summary

As evidenced above there are a range of different approaches for solving the Schrödinger wave equation for electrons. These methods vary significantly in cost and accuracy and the most appropriate method often depends on the nature of the chemical systems studied. An aim of this thesis is to test the accuracy of these methods and develop optimal protocols for applying them.

## 2.7 References

[1] E. Schrödinger, *Phys. Rev.* **1926**, *28*, 1049–1070.

[2] F. Jensen, *Introduction to Computational Chemistry 2nd ed.*, Wiley, **2007**.

[3] M. Born, R. Oppenheimer, *Annalen der Physik* **1927**, *389*, 457–484.

[4] E. G. Lewars, *Computational Chemistry. Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, Kluwer Academic Publishers, **2003**.

[5] C. J. Cramer, *Essentials of Computational Chemistry: Theories and Models 2nd ed.*, John Wiley & Sons Inc., **2004**.

[6] S. M. Bachrach, *Computational Organic Chemistry*, John Wiley & Sons Inc., **2014**.

[7] B. Austin, D. Y. Zubarev, W. A. Lester, *Chem. Rev.* **2012**, *112*, 263–88.

[8] M. Bajdich, L. Mitáš, *Acta Physica Slovaca* **2009**, *59*, 81–168.

[9] A. Lüchow, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 388–402.

[10] R. J. Needs, M. D. Towler, N. D. Drummond, P. López Ríos, *J. Phys. Condens. Matter* **2010**, *22*, 023201.

[11] D. R. Hartree, *Proc. Cambridge. Phil. Soc.* **1928**, *24*, 89–110.

[12] W. Pauli, *Zeitschrift für Physik A Hadrons and Nuclei* **1925**, *31*, 765–783.

[13] J. C. Slater, *Phys. Rev.* **1929**, *34*, 1293–1322.

[14] D. R. Hartree, *Proc. Cambridge. Phil. Soc.* **1928**, *24*, 111–132.

[15] D. R. Hartree, *Proc. Cambridge. Phil. Soc.* **1928**, *24*, 426–437.

[16] V. Fock, *Z. Physik* **1930**, *62*, 126–148.

[17] C. C. J. Roothaan, *Rev. Mod. Phys.* **1951**, *23*, 69–89.

[18] J. A. Pople, R. K. Nesbet, *J. Chem. Phys.* **1954**, *22*, 571–572.

[19] P. J. Knowles, N. C. Handy, *J. Phys. Chem.* **1988**, *92*, 3097–3100.

[20] S. Boys, *Proc. R. Soc. Lond.* **1950**, *201*, 125–137.

[21] R. Krishnan, H. B. Schlegel, J. A. Pople, *J. Chem. Phys* **1980**, *72*, 4654–4655.

[22] I. Shavitt, *Mol. Phys.* **1998**, *94*, 3–17.

[23] G. E. Scuseria, H. F. Schaefer III, *J. Chem. Phys.* **1989**, *90*, 3700–3703.

[24] G. E. Scuseria, C. L. Janssen, H. F. Schaefer III, *J. Chem. Phys.* **1988**, *89*, 7382–7387.

[25] G. D. Purvis III, R. J. Bartlett, *J. Chem. Phys.* **1982**, *76*, 1910–1918.

[26] J. Cizek, *Adv. Chem. Phys* **1969**, *14*, 35–89.

[27] K. Raghavachari, G. W. Trucks, J. A. Pople, M. Head-Gordon, *Chem. Phys. Lett.* **1989**, *157*, 479–483.

[28] B. O. Roos, P. R. Taylor, *Chem. Phys.* **1980**, *48*, 157–173.

[29] K. Andersson, P.-Å. Malmqvist, B. O. Roos, *J. Chem. Phys.* **1992**, *96*, 1218–1226.

[30] L. A. Curtiss, P. C. Redfern, K. Raghavachari, *J. Chem. Phys.* **2007**, *127*, 124105.

[31] L. A. Curtiss, P. C. Redfern, K. Raghavachari, *J. Chem. Phys.* **2007**, *126*, 084108.

[32] L. A. Curtiss, P. C. Redfern, K. Raghavachari, J. A. Pople, *J. Chem. Phys.* **2001**, *114*, 108–117.

[33] L. A. Curtiss, P. C. Redfern, K. Raghavachari, V. Rassolov, J. A. Pople, *J. Chem. Phys.* **1999**, *110*, 4703–4709.

[34] L. A. Curtiss, K. Raghavachari, P. C. Redfern, V. Rassolov, J. A. Pople, *J. Chem. Phys.* **1998**, *109*, 7764–7776.

[35] L. A. Curtiss, K. Raghavachari, *J. Chem. Phys.* **1991**, *94*, 7221–7230.

[36] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari, L. A. Curtiss, *J. Chem. Phys.* **1989**, *90*, 5622–5629.

[37] L. A. Curtiss, C. Jones, G. W. Trucks, K. Raghavachari, J. A. Pople, *J. Chem. Phys.* **1990**, *93*, 2537–2545.

[38] G. P. F. Wood, L. Radom, G. A. Petersson, E. C. Barnes, M. J. Frisch, J. A. Montgomery Jr., *J. Chem. Phys.* **2006**, *125*, 094106.

[39] J. A. Montgomery Jr., M. J. Frisch, J. W. Ochterski, G. A. Petersson, *J. Chem. Phys.* **2000**, *112*, 6532–6542.

[40] J. A. Montgomery Jr., M. J. Frisch, J. W. Ochterski, G. A. Petersson, *J. Chem. Phys.* **1999**, *110*, 2822–2827.

[41] J. W. Ochterski, G. A. Petersson, J. A. Montgomery Jr., *J. Chem. Phys.* **1996**, *104*, 2598–2619.

[42] J. M. L. Martin, G. de Oliveira, *J. Chem. Phys.* **1999**, *111*, 1843–1856.

[43] A. D. Boese, M. Oren, O. Atasoylu, J. M. L. Martin, M. Kállay, J. Gauss, *J. Chem. Phys.* **2004**, *120*, 4129–4141.

[44] A. Karton, E. Rabinovich, J. M. L. Martin, B. Ruscic, *J. Chem. Phys.* **2006**, *125*, 144108.

[45] B. Chan, J. Deng, L. Radom, *J. Chem. Theory Comput.* **2011**, *7*, 112–120.

[46] N. J. Mayhall, K. Raghavachari, P. C. Redfern, L. A. Curtiss, *J. Phys. Chem. A* **2009**, *113*, 5170–5175.

[47] A. Karton, R. J. O'Reilly, L. Radom, *J. Phys. Chem. A* **2012**, *116*, 4211–4221.

[48] P. Hohenberg, W. Kohn, *Phys. Rev.* **1964**, *136*, B864.

[49] W. Kohn, L. J. Sham, *Phys. Rev.* **1965**, *140*, A1133.

[50] J. P. Perdew, K. Schmidt, *AIP Conference Proceedings* **2001**, *577*, 1–20.

[51] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[52] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.

[53] J. Toulouse, R. Assaraf, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 244112.

[54] M. Casula, S. Sorella, *J. Chem. Phys.* **2003**, *119*, 6500.

[55] P. López Ríos, A. Ma, N. D. Drummond, M. D. Towler, R. J. Needs, *Phys. Rev. E* **2006**, *74*, 066701.

[56] M. Bajdich, L. Mitáš, G. Drobný, L. Wagner, K. Schmidt, *Phys. Rev. Lett.* **2006**, *96*, 130201.

[57] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **2012**, *8*, 2181–2188.

[58] N. D. Drummond, M. D. Towler, R. J. Needs, *Phys. Rev. B* **2004**, *70*, 235119.

[59] M. C. Per, S. P. Russo, I. K. Snook, *J. Chem. Phys.* **2009**, *130*, 134103.

[60] P. López Ríos, P. Seth, N. D. Drummond, R. J. Needs, *Phys. Rev. E* **2012**, *86*, 036703.

[61] T. Kato, *Comm. Pure Appl. Math.* **1957**, *10*, 151–177.

[62] M. C. Per, K. Walker, S. Russo, *J. Chem. Theory Comput.* **2012**, *8*, 2255–2259.

[63] R. C. Clay III, M. A. Morales, *J. Chem. Phys.* **2015**, *142*, 234103.

[64] F. Schautz, F. Buda, C. Filippi, *J. Chem. Phys.* **2004**, *121*, 5836–5844.

[65] F. Schautz, C. Filippi, *J. Chem. Phys.* **2004**, *120*, 10931–10941.

[66] W. A. Al-Saidi, C. J. Umrigar, *J. Chem. Phys.* **2008**, *128*, 154324.

[67] C. Filippi, M. Zaccheddu, F. Buda, *J. Chem. Theory Comput.* **2009**, *5*, 2074–2087.

[68] P. M. Zimmerman, J. Toulouse, Z. Zhang, C. B. Musgrave, C. J. Umrigar, *J. Chem. Phys.* **2009**, *131*, 124103.

[69] R. Berner, A. Lüchow, *J. Phys. Chem. A* **2010**, *2*, 13222–13227.

[70] M. Dubecký, R. Derian, L. Mitáš, I. Štich, *J. Chem. Phys.* **2010**, *133*, 244301.

[71] M. Dubecký, R. Derian, L. Horváthová, M. Allan, I. Štich, *Phys. Chem. Chem. Phys.* **2011**, *13*, 20939–20945.

[72] B. K. Clark, M. A. Morales, J. McMinis, J. Kim, G. E. Scuseria, *J. Chem. Phys.* **2011**, *135*, 244105.

[73] F. R. Petruzielo, J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2012**, *136*, 124116.

[74] L. Koziol, M. Q. Morales, *J. Chem. Phys.* **2014**, *140*, 224316.

[75] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[76] H. Flyvbjerg, H. G. Petersen, *J. Chem. Phys.* **1989**, *91*, 461–466.

[77] M. H. Kalos, *Phys. Rev.* **1962**, *128*, 1791–1795.

[78] M. H. Kalos, D. Levesque, L. Verlet, *Phys. Rev. A* **1974**, *9*, 2178–2195.

[79] H. F. Trotter, *Proc. Am. Math. Soc.* **1959**, *10*, 545–551.

[80] R. C. Grimm, R. G. Storer, *J. Comput. Phys.* **1971**, *7*, 134–156.

[81] D. Bressanini, G. Morosi, *J. Chem. Phys.* **2008**, *129*, 054103.

[82] J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2008**, *128*, 174101.

[83] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, R. G. Hennig, *Phys. Rev. Lett.* **2007**, *98*, 110201.

[84] F. A. Reboredo, *J. Chem. Phys.* **2012**, *136*, 204101.

[85] F. A. Reboredo, J. Kim, *J. Chem. Phys.* **2014**, *140*, 074103.

[86] D. M. Ceperley, A. B. J, *J Chem Phys* **1984**, *81*, 5833.

[87] F. Alet, S. Capponi, N. Laflorencie, M. Mambrini, *Phys. Rev. Lett.* **2007**, *99*, 117204.

[88] N. M. Tubman, J. L. DuBois, R. Q. Hood, B. J. Alder, *J. Chem. Phys.* **2011**, *135*, 184109.

[89] D. Domin, B. Braïda, W. A. Lester, *J. Phys. Chem. A* **2008**, *112*, 8964–8969.

[90] D. M. Arnow, *J. Chem. Phys.* **1982**, *77*, 5562–5572.

[91] M. Kalos, F. Pederiva, *Phys. Rev. Lett.* **2000**, *85*, 3547–3551.

[92] R. Assaraf, M. Caffarel, A. Khelif, *Journal of Physics A: Mathematical and Theoretical* **2007**, *40*, 1181.

[93] P. Seth, P. López Ríos, R. J. Needs, *J. Chem. Phys.* **2011**, *134*, 084105.

[94] O. Valsson, C. Filippi, *J. Chem. Theory Comput.* **2010**, *6*, 1275–1292.

[95] D. M. Cleland, M. C. Per, *J. Chem. Phys.* **2016**, *144*, 124108.

[96] R. Ditchfield, W. J. Hehre, J. A. Pople, *J. Chem. Phys.* **1971**, *54*, 724–728.

[97] W. J. Hehre, R. Ditchfield, J. A. Pople, *J. Chem. Phys.* **1972**, *56*, 2257–2261.

[98] T. H. Dunning Jr, *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[99] R. A. Kendall, T. H. Dunning Jr, R. J. Harrison, *J. Chem. Phys.* **1992**, *96*, 6796–6806.

[100] D. E. Woon, T. H. Dunning Jr, *J. Chem. Phys.* **1993**, *98*, 1358–1371.

[101] P.-O. Widmark, P.-A. Malmqvist, B. O. Roos, *Theor. Chim. Acta* **1990**, *77*, 291–306.

[102] D. Ceperley, *J. Stat. Phys.* **1986**, *43*, 815–826.

[103] A. Ma, N. Drummond, M. D. Towler, R. J. Needs, *Phys. Rev. E* **2005**, *71*, 066704.

[104] J. B. Anderson, *J. Chem. Phys.* **1975**, *63*, 1499–1503.

[105] S. Fahy, X. W. Wang, S. G. Louie, *Phys. Rev. Lett.* **1988**, *61*, 1631–1634.

[106] S. Fahy, X. W. Wang, S. G. Louie, *Phys. Rev. B* **1990**, *42*, 3503–3522.

[107] L. Mitáš, E. L. Shirley, D. M. Ceperley, *J. Chem. Phys.* **1991**, *95*, 3467–3475.

[108] M. Casula, *Phys. Rev. B* **2006**, *74*, 161102.

[109] C. W. Greeff, W. A. Lester, *J. Chem. Phys.* **1998**, *109*, 1607–1612.

[110] I. Ovcharenko, A. Aspuru-Guzik, W. A. Lester, *J. Chem. Phys.* **2001**, *114*, 7790–7794.

[111] J. R. Trail, R. J. Needs, *J. Chem. Phys.* **2005**, *122*, 014112.

[112] J. R. Trail, R. J. Needs, *J. Chem. Phys.* **2005**, *122*, 174109.

[113] J. Xu, M. Deible, K. A. Peterson, K. Jordan, *J. Chem. Theory Comput.* **2013**, *9*, 2170–2178.

[114] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[115] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2008**, *129*, 164115.

[116] M. Dubecký, R. Derian, P. Jurečka, L. Mitáš, P. Hobza, M. Otyepka, *Phys. Chem. Chem. Phys.* **2014**, *16*, 20915.

[117] J. R. Trail, R. J. Needs, *J. Chem. Phys.* **2013**, *139*, 014101.

[118] J. R. Trail, R. J. Needs, *J. Chem. Theory Comput.* **2014**, *10*, 2049–2053.

Chapter 3

# An efficient protocol for diffusion Monte Carlo calculations

## 3.1 Introduction

Hydrogen-abstraction reactions have an important role in many fields of chemistry, including biology, combustion, autoxidation, atmospheric chemistry, polymerisation and many other synthetic processes.[1,2] Transition states are challenging structures and often high-level methods are necessary for a reasonable description of electronic correlation. Studies have shown that DFT methods are capable of producing accurate geometries and frequencies for this reaction but underestimate barrier heights. Functionals specifically parameterised for kinetic reactions (i.e. BMK[3]) perform better but their accuracy cannot be guaranteed for reactions not included in the training set.[4] In general expensive, high-level *ab initio* or composite methods are necessary for correct barrier heights.[5–8]

Reliable kinetic models need accurate barrier heights but chemically reliable results often require a high-level treatment of electron correlation effects. This is illustrated by the H abstraction of methanol by an H atom. The two main reaction pathways are

$$\begin{aligned} \mathrm{CH_3OH + H} &\rightarrow \mathrm{CH_2OH + H_2} \qquad (3.1) \\ &\rightarrow \mathrm{CH_3O + H_2} \qquad (3.2) \end{aligned}$$

with $CH_2OH$ is the dominant product. These pathways are illustrated in Figure 3.1 These reactions are known to be important in the combustion of methanol under fuel-rich conditions,[9] and have been studied using a wide range of theoretical methods.[5–8,10,11] Despite the apparent simplicity of this system, studies have shown that accurate calculations of the barrier heights require methods that scale as a large power of the system size,[8] and that there are large discrepancies between methods. The performance of both MP2 theory and the popular B3LYP density functional[12] are particularly poor for this system.

QMC methods have been shown to be highly accurate for energetic[13,14] and structural properties.[15] The main advantages of QMC methods over more widely-used alternatives such as coupled cluster theory are their low scaling with system size ($N^{3-4}$), and their immense parallelisability.[16] Despite these advantages there have been relatively few QMC calculations of H abstraction barrier heights. The earliest example, of the reaction

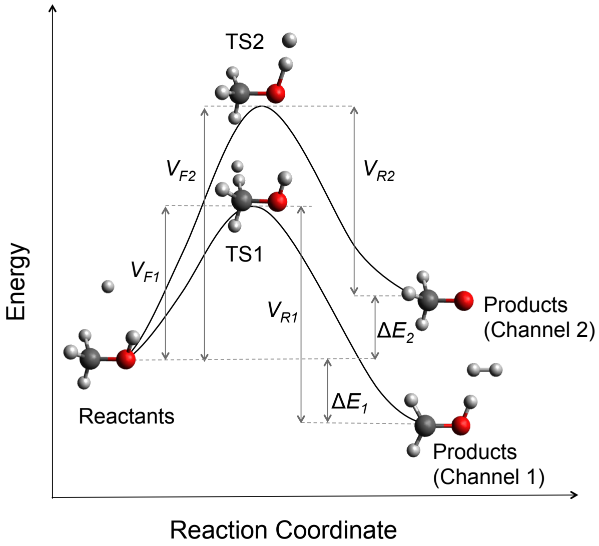


Figure 3.1: The potential energy surface for the reaction between atomic hydrogen and methanol.

$H_2 + H \rightarrow H + H_2$,[17,18] has very recently been revised to even higher accuracy.[19] Other examples include the reaction $OH + H_2 \rightarrow H_2O + H$,[20] and a study by Kollias et al. of the H abstraction of methanol by a Cl atom, which showed agreement with MP2 calculations.[21] More recent examples include the H abstraction by styrene of the H-terminated Si(001) surface,[22] and calculations of the barrier heights of three H-transfer reactions involving small molecules.[23]

In order to achieve resolutions of chemical accuracy in the barrier heights, statistical uncertainties in the stochastic QMC energies need to be on the order of fractions of a kcal/mol. Even though QMC scales well, this need for small uncertainties makes the calculations computationally expensive. As with other electronic-structure theories, efficient use of QMC methods requires a number of methodological choices to be made, including the choice of trial wavefunction and treatment of non-local pseudopotentials. Wavefunction choice is often discussed in reports of QMC calculations, but the effects of the parameters governing the treatment of pseudopotentials, including quadrature grids and cutoffs, are rarely mentioned. The impact of these choices, and their mutual interactions, are investigated in this chapter by performing a detailed study of the barrier heights of H abstraction in methanol by an H atom.

## 3.2 Trial wavefunction

Practical QMC calculations require user-defined trial wavefunctions. The complexity of these wavefunctions strongly influence the computational cost of the calculations. Complicated wavefunctions are more expensive to optimise and evaluate at each Monte Carlo step, but more accurate wavefunctions lower the variance of the energy and therefore require fewer Monte Carlo steps to obtain a given statistical accuracy. In addition, the nodal surface of the wavefunction (the hypersurface on which it equals zero, and across which it changes sign) determines the systematic errors in fixed-node DMC calculations. The trial wavefunctions employed here have the Slater-Jastrow form,

$$\Psi_T = e^J D^{\uparrow} D^{\downarrow} \tag{3.3}$$

where the $D^{\uparrow,\downarrow}$ are Slater determinants constructed from single-particle orbitals, and $J$ is a Jastrow factor containing explicit electron correlation terms.

The Jastrow factor used is a sum of electron-electron $(ee)$, electron-nucleus $(eN)$, and electron-electron-nucleus $(eeN)$ terms,

$$J = \sum_{i>j} \sum_{A} \left[ J_{ee}(r_{ij}) + J_{eN}(r_{iA}) + J_{eeN}(r_{iA}, r_{jA}, r_{ij}) \right] \tag{3.4}$$

where $i, j$ label electrons, and $A$ labels nuclei. These terms were constructed as compactly-supported natural polynomial expansions in the electron-electron and electron-nucleus distances,

$$J_{ee}(r_{ij}) = f(r_{ij}; L^{ee}) \sum_{l=0}^{N_{ee}} \alpha_l \, r_{ij}^l \tag{3.5}$$

$$J_{eN}(r_{iA}) = f(r_{iA}; L^{eN}) \sum_{l=0}^{N_{eN}} \beta_{l;A} \, r_{iA}^l \tag{3.6}$$

$$J_{eeN}(r_{iA}, r_{jA}, r_{ij}) = f(r_{iA}; L^{eeN}) f(r_{jA}; L^{eeN}) \sum_{l,m,n=0}^{N_{eeN}} \gamma_{lmn;A} \, r_{iA}^l r_{jA}^m r_{ij}^n \tag{3.7}$$

where $L$ is the cutoff range, and $\{\alpha, \beta, \gamma\}$ are optimisable parameters. The cutoff function $f(r; L)$ is a $C^2$-smooth Wendland function[24] which goes to zero at $L$,

$$f(r; L) = \begin{cases} \left(1 - \frac{r}{L}\right)^4 \left(1 + 4\frac{r}{L}\right) & 0 \leq r \leq L \\ 0 & r > L \end{cases} \tag{3.8}$$

All the calculations presented here used fixed ranges of $L = 5$ Bohr. The electron-electron cusp condition and the electron-nucleus no-cusp conditions were enforced by constraining the optimisable parameters in the Jastrow factor. The method described in the appendix of Ref. 25 is used for the more complicated $eeN$ term. The free parameters in the Jastrow factor were optimised by minimising the total energy at the Variational Monte Carlo (VMC) level, using the linear method of Toulouse and Umrigar.[26]

In addition to all-electron calculations, non-local pseudopotentials were used to represent the ionic cores. The use of pseudopotentials can greatly reduce the cost of QMC calculations, as the removal of the chemically inert core electrons reduces the fluctuations in the local energy. Evaluation of the local energy requires the non-local potential to be projected onto the trial wavefunction. Following Mitas et al.,[27] for each ion the contribution to this projection from an electron labelled $i$ can be written as a sum over angular momenta $l$,

$$\left(\frac{\hat{V}_{\rm nl}\Psi_T}{\Psi_T}\right)_i = \sum_l \frac{(2l+1)}{4\pi} v_l(r_i) \int_{4\pi} P_l[\cos\theta_{i'}] \frac{\Psi_T(...,\mathbf{r}'_i,...)}{\Psi_T(...,\mathbf{r}_i,...)} d\Omega'_i \tag{3.9}$$

where $v_l$ is the angular-momentum dependent radial potential, and the integral is over the surface of a sphere of radius $r_i$ centred on the ion. In practice the integral is evaluated using a deterministic approach,

$$\int_{4\pi} f(\mathbf{r}'_i) d\Omega'_i \approx \sum_k^{N_Q} w_k f(\mathbf{r}_k) \tag{3.10}$$

where the $N_Q$ weights $w_k$ and points $\mathbf{r}_k$ are chosen according to a Gaussian quadrature rule, with values taken from Ref.[27] This projection must be evaluated for each electron within range of each ion, at each step of the DMC calculation, so the number of quadrature points and the range of the radial potentials can have a large impact on the cost of the QMC calculation.

## 3.3 Computational details

The forward (F) and reverse (R) barrier heights of the reactions shown in Equations (3.1,3.2) are defined as the total energy differences

$$V_{\rm F1} = E(\text{TS1}) - \text{E}(\text{CH}_3\text{OH}) - \text{E}(\text{H}) \tag{3.11}$$

$$V_{\rm R1} = E(\text{TS1}) - \text{E}(\text{CH}_2\text{OH}) - \text{E}(\text{H}_2) \tag{3.12}$$

$$V_{\rm F2} = E(\text{TS2}) - \text{E}(\text{CH}_3\text{OH}) - \text{E}(\text{H}) \tag{3.13}$$

$$V_{\rm R2} = E(\text{TS2}) - \text{E}(\text{CH}_3\text{O}) - \text{E}(\text{H}_2) \tag{3.14}$$

where TS1 and TS2 are the transition-state structures for the reactions. The molecular geometries were obtained from B3LYP calculations with the Roos augmented triple-zeta (ATZ) basis set,[28] using Gaussian09.[29] The orbitals used in the Slater determinants were taken from B3LYP calculations. Although the orbitals themselves contain no description of electron correlation, it has been shown that using orbitals from a correlated method such as B3LYP results in better QMC energies than using Hartree-Fock orbitals.[30] For all-electron calculations using the full electron-ion Coulomb potential, the orbitals were expanded in the Gaussian-type Roos ATZ basis set,[28] and cusp-corrected using a standard approach.[31]

The effect of using two different pseudopotentials was compared. Both are Hartree-Fock pseudopotentials including scalar relativistic effects and were explicitly constructed for use in QMC calculations. The Trail-Needs (TN)[32] potentials are shape-consistent, whereas the Burkatzki-Filippi-Dolg (BFD)[33] potentials are energy-consistent. The BFD calculations used the associated valence triple-zeta (VTZ) basis sets, and an improved H-atom potential.[34] Calculations with the TN potentials used the aug-cc-pVTZ-CDF basis set from Ref.[35]

Non-local pseudopotentials were treated beyond the locality approximation in DMC using the size-consistent T-moves approach.[36] Imaginary time-step sizes of $\tau$=(0.04, 0.02, 0.01, 0.005) a.u. were used for pseudopotential calculations. Smaller time-step sizes are required for all-electron calculations, and in this case values of $\tau = (0.02, 0.01, 0.005, 0.001)$ a.u. were used. All DMC energies were extrapolated to $\tau = 0$ using quadratic fits. Target population sizes of 8000 walkers were used in all DMC calculations. DMC results have a small bias from using a finite population size. This bias is smaller for better trial wave functions. For small systems like those studied here this bias is very small for populations over 1000. The number of walkers used in the calculations is determined by the computational efficiency for parallel codes on a large number of cores. The walker population is redistributed at each step but this is a slow process. Using larger numbers of walkers increases the ratio of compute to communcation and avoids wasting time. Therefore larger populations are used for larger numbers of cores. For further information please see Ref. 37. All the QMC calculations were performed using the CSIRO Quantum Monte Carlo code.[38]

Forward and reverse barrier heights are not directly available from experiment. The highest level theoretical results available in the literature use coupled cluster methods. To compare with these coupled cluster calculations using the CCSD(T) approach were also performed using Molpro[39] with Dunning's aug-cc-pVQZ basis set[40] and an unrestricted Hartree-Fock reference state.

Finally, the accuracy of density functional theory is evaluated using 12 different exchange-correlation functionals. The types of functionals chosen were the local density approximation (LDA[41] ), generalised gradient approximation (GGA) functionals (BLYP,[42,43] PBE,[44] B97D3[45]), meta-GGAs (TPSS,[46] M06L[47]), hybrid GGAs (B3LYP,[12] PBE0[48]), hybrid meta-GGAs (B1B95,[49] MPW1B95[50]), and double hybrids (B2PLYP,[51] mPW2PLYP[52]).

## 3.4 Results and discussion

Barrier heights calculated using density functional, CCSD(T), and QMC methods are compared against results from the literature in Table 3.1.

The CCSD(T) results presented here were calculated using the B3LYP geometries and agree closely with previous calculations. When compared against the results of Carvalho et al.,[6] who used CCSD(T)/cc-pVTZ geometries, the largest deviation seen is only 0.3 kcal/mol. This demonstrates the accuracy of the B3LYP geometries, despite the inability of that level of theory to predict accurate barrier heights.

None of the exchange-correlation functionals used are able to recover all the barrier heights to within chemical accuracy of the CCSD(T) reference values. The most accurate functionals are the double-hybrids. They perform well for the forward barriers but the

Table 3.1: Reaction barrier heights for the H abstraction of methanol by an H atom (in kcal/mol) using different methods. Statistical uncertainties in the last digit of the DMC results are shown in parentheses. All DFT calculations used the Roos-ATZ basis, unless stated otherwise.

| Method | $V_{\mathrm{F1}}$ | $V_{\mathrm{R1}}$ | $V_{\mathrm{F2}}$ | $V_{\mathrm{R2}}$ |
|---|---|---|---|---|
| LDA | -3.5 | -0.1 | 1.8 | -7.7 |
| BLYP | 1.1 | 12.0 | 3.3 | 6.5 |
| PBE | 2.0 | 8.3 | 5.7 | 3.2 |
| B97D3 | 0.8 | 12.5 | 3.9 | 7.3 |
| TPSS | -0.9 | 11.4 | 1.1 | 7.3 |
| M06L | 7.0 | 10.1 | 7.8 | 8.2 |
| B3LYP | 3.6 | 13.3 | 6.9 | 9.2 |
| B3LYP/6-31+G(d,p)[8] | 3.2 | 12.9 | - | - |
| PBE0 | 5.5 | 11.0 | 10.4 | 8.0 |
| B1B95 | 6.8 | 13.5 | 11.3 | 10.5 |
| B1B95/MG3S[8] | 7.0 | 13.5 | - | - |
| mPW1B95 | 6.9 | 12.8 | 11.5 | 10.0 |
| mPW1B95/MG3S[8] | 7.1 | 12.9 | - | - |
| B2PLYP | 10.0 | 17.6 | 14.0 | 17.2 |
| mPW2PLYP | 9.8 | 17.0 | 13.9 | 16.7 |
| MP2/6-31+G(d,p)[8] | 16.8 | 18.0 | - | - |
| CCSD(T)/aug-cc-pVQZ[8] | 9.6 | 15.6 | - | - |
| CCSD(T)/cc-pVQZ [6] | 9.8 | 15.8 | 15.1 | 12.0 |
| CCSD(T)/aug-cc-pVQZ | 9.5 | 15.5 | 15.1 | 11.7 |
| DMC (All-electron) | 9.9(1) | 15.5(1) | 16.3(1) | 12.1(1) |
| DMC (BFD) | 9.8(2) | 15.0(2) | 15.9(2) | 12.1(2) |
| DMC (TN) | 9.9(2) | 15.6(2) | 15.7(2) | 12.2(2) |

reverse barrier $V_{\mathrm{R}2}$ deviates from the reference CCSD(T) value by over 5 kcal/mol. There is significant variation of barrier heights, even within the same class of functionals. For example, the forward barrier heights obtained using the meta-GGA functionals TPSS and M06L differ by around 7 kcal/mol. The extreme variability in the accuracy of the different density functionals emphasises the need for extensive benchmarking using higher level methods.

All-electron DMC calculations were performed using the complete Jastrow factor shown in Equation 3.4, including the $eeN$ terms. The barrier heights obtained using this method agree closely with the CCSD(T) results for barriers except for $V_{\mathrm{F}2}$, where there is a difference of just over 1 kcal/mol. This disagreement is potentially due to the presence of non-dynamical correlation effects. The $T_1$ diagnostic[53] is a widely-used indicator of non-dynamical effects in coupled-cluster calculations. Typically, $T_1$ values of 0.02 or greater are taken as an indication that a single determinant reference state is insufficient, though some researchers suggest this value should be higher for open-shell systems.[54] Nearly all of the structures have small $T_1$ values, the exceptions being $CH_3O$ (0.021) and TS2 (0.032).

Reference pseudopotential QMC results were obtained using both BFD and TN pseudopotentials, with the same sized Jastrow factor used in the all-electron calculations. The barrier heights are all within two standard deviations of the all-electron results. These pseudopotential calculations are actually more expensive per Monte Carlo step than the all-electron approach (1.54x, for the DMC calculation of $CH_3OH$), due to the need to repeatedly evaluate the integral in Equation 3.9. The cost benefits of the pseudopotential approach come from the ability to use larger imaginary time-step sizes, and the reduced variance of the energy. As an example, for a fixed time-step size of $\tau = 0.005$ a.u., the BFD calculation of $CH_3OH$ is over 20% faster than the all-electron calculation, to obtain the same statistical accuracy. This reduction in cost will be significantly larger for systems containing heavier elements.

### 3.4.1 Approximations

To further reduce the computational cost of the DMC calculations, the effect of reducing both the complexity of the Jastrow factor and the treatment of the pseudopotentials was investigated. Dubecký et al.[55] have shown that for noncovalent interactions a two-body Jastrow factor is sufficient and the same modification is considered here, removing the most expensive $eeN$ terms. For the pseudopotentials, the number of quadrature points in the evaluation of Equation 3.9 and the range of both the local and non-local radial potentials was reduced. The notation used to define these settings is $x.y.z$. Here $x$ denotes the size of the Jastrow factor, which is either 2J (indicating use of $ee$ and $eN$ terms) or 3J (indicating use of $ee$, $eN$, and $eeN$ terms). The number of points used in the quadrature grid ($N_Q$ in Equation 3.10) is given by $y$.

Points were distributed tetrahedrally ($y = 4$) or icosahedrally ($y = 12$), with the locations of the points on the unit sphere taken from Ref.[27] Finally, $z$ is a parameter which determines the ranges of the radial parts of both the local and non-local pseudopotentials.

The ranges $r$ of the local potentials are chosen such that $r$ is the point furthest from the nucleus which deviates by more than $10^{-z}$ from the bare Coulomb potential. Similarly, the range of the radial part of a non-local potential is defined as the point furthest from the nucleus which deviates by more than $10^{-z}$ from zero. The potentials are set to zero outside these ranges. This method of defining the ranges leads to different values for each element and for both types of pseudopotential used. For the elements considered here, the BFD potentials are shorter ranged than the TN potentials for each of our choices of $z = 8, 5, 3$. Using this notation, the settings used for the reference pseudopotential calculations in Table 3.1 are 3J.12.8.

### 3.4.2 Accuracy

As shown in Figure 3.2, the different settings used for the Jastrow factor and pseudopotentials have very little effect on the predicted barrier heights. The majority of settings result in no statistically significant change, and the largest changes are less than 1 kcal/mol. Overall the deviations from the reference values are statistically equivalent for both types of pseudopotential considered.

### 3.4.3 Cost

Despite the relatively insignificant changes in the barrier heights, the reduced settings can have a very strong effect on the computational cost of the QMC calculations. The timings for a complete DMC calculation of $CH_3OH$, relative to the reference settings using the BFD pseudopotentials, are shown in Figure 3.3.

The largest time saving comes from eliminating the *eeN* terms in the Jastrow factor, which makes the DMC calculation of $CH_3OH$ 3x faster per Monte Carlo step. This simplified Jastrow factor increases the variance of the local energy (see Sec. 3.4.4), but even when this effect is taken into account, a speedup of around 2.5x is still obtained for the time to achieve a fixed statistical accuracy in the total energy.

The next most important speedup comes from reducing the number of points in the quadrature grids. Evaluating the projection of the non-local operator onto the trial wavefunction requires multiple evaluations of the wavefunction ratio with the position of one electron moved. Even when using efficient methods for calculating this ratio, reducing the number of quadrature points from 12 to 4 results in a speedup of 1.7x when using the simpler 2J Jastrow factor. Reducing the ranges of the local and non-local parts of the pseudopotentials also reduces the cost of the calculations, but the improvement obtained is much smaller than when simplifying the Jastrow factor or reducing the number of quadrature points. When combined, all three measures provide a speedup greater than 5x, with no reduction in the quality of the barrier heights.

Calculations using the TN pseudopotentials were always more expensive than when using the BFD potentials, as shown in Figure 3.3. There are a number of reasons for this, the most important being simply the size of the one-electron basis set used to construct the B3LYP orbitals in the trial wavefunction. When expressed in spherical harmonic (as

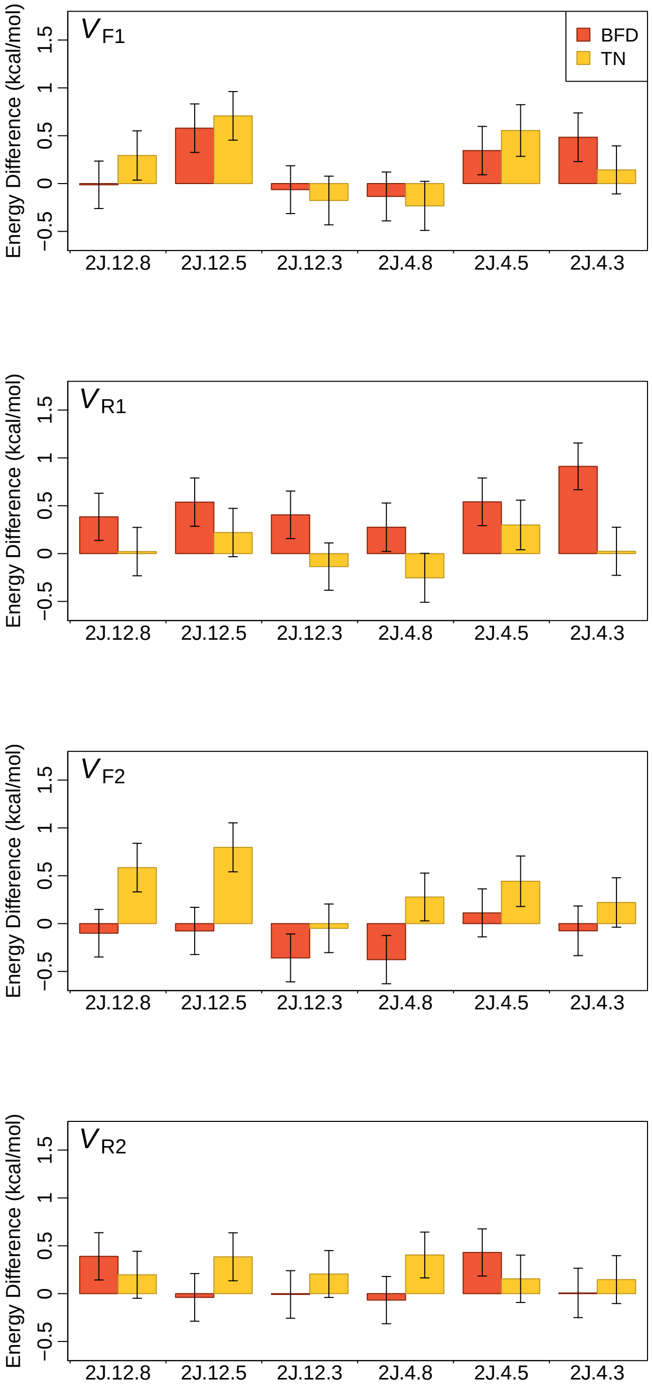


Figure 3.2: Deviations of DMC barrier heights from the 3J.12.8 reference values, for different Jastrow factor and pseudopotential settings. The settings use the notation $x.y.z$, where $x$ denotes the size of the Jastrow factor, $y$ is the number of points used in the quadrature grid for evaluating non-local pseudopotentials, and $z$ is a measure of the cutoff applied to the radial parts of the pseudopotentials. For more details see Sec. 3.4.

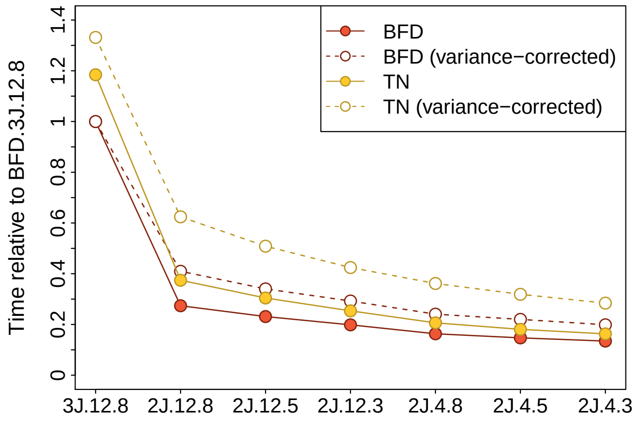


Figure 3.3: DMC timings for $CH_3OH$ relative to BFD.3J.12.8 settings. Solid points indicate relative times for a fixed number of Monte Carlo steps. Open points indicate relative times to achieve a fixed statistical uncertainty.

opposed to Cartesian) Gaussians, the CDF basis set used with the TN potentials has 35% more primitives than the BFD basis set. There is also some contribution from the fact that the TN potentials contain a non-local term for H. The TN potentials contain $s$, $p$, and $d$ channels for each element used in this work. The BFD potentials contain only $s$ and $p$ channels for C and O species, and only a local component for H. Finally, in the Gaussian representation of the pseudopotentials used, the TN potentials contain many more terms than the BFD potentials. This has a very small effect on the cost, but it could be eliminated by representing both potentials on a radial grid.

### 3.4.4 Variance

Reducing the complexity of the trial wavefunction by removing the $eeN$ terms from the Jastrow factor leads to an increase in the variance of the local energy, shown in Figure 3.4. The results shown here are for energy-optimised trial wavefunctions. It should be possible to obtain lower variances by explicitly minimising the variance of the local energy rather than the total energy, though the gains are likely to be small.

Reducing the number of points in the quadrature grid has no effect on the variance when using the BFD pseudopotentials, but results in a 5% increase in the variance when using the TN potentials. This is likely due to the different angular momenta used in the construction of the different potentials. As mentioned above, the TN pseudopotentials use higher angular momentum terms than the BFD potentials, and so one would expect them to require a higher order quadrature rule. However, the effect is small, and does not translate in any statistically significant way to the quality of the energy barriers as shown in Figure 3.2. Reducing the ranges of the local and non-local radial potentials has no noticeable effect on the variance for either type of pseudopotential.

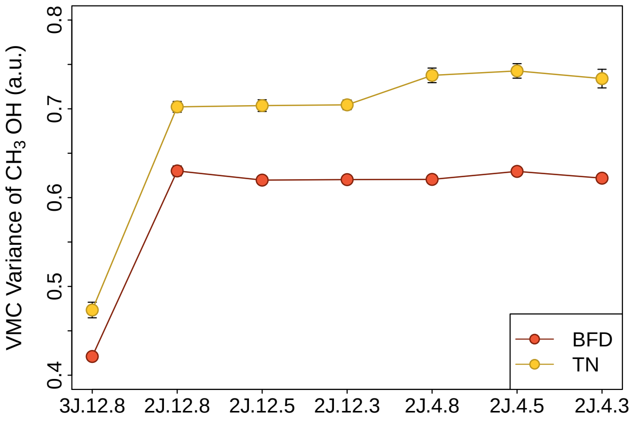


Figure 3.4: VMC variance of the local energy for $CH_3OH$ using different settings for the Jastrow factor and pseudopotentials.

### 3.4.5 Time-step error

The changes to the Jastrow factor and treatment of pseudopotentials also have an effect on the time-step error in DMC, as shown in Figure 3.5. A non-symmetric branching factor was used in DMC calculations with T-moves, which results in large time-step ($\tau$) errors, but with a predominantly linear behaviour that is easily extrapolated to $\tau = 0$. Using a symmetric branching factor does result in a smaller error for a given value of $\tau$, but in practice it has been frequently observed that the increased curvature means that reliable extrapolation to $\tau = 0$ still requires relatively small values of $\tau$. Using symmetric branching with T-moves is slightly more expensive than non-symmetric branching, as it requires a second evaluation of the local energy at each DMC step if a T-move is accepted. Our current approach is to use non-symmetric branching if performing a full extrapolation to $\tau = 0$, and to use symmetric branching if a single small value of $\tau$ is used.

As with the variance, the largest effect on the time-step error is the quality of the trial wavefunction. Using the larger 3J Jastrow factor results in smaller time-step errors than when using the 2J form. The quality of the Jastrow factor also has a small effect on the final $\tau = 0$ DMC energy, which comes from the projection of the non-local pseudopotential onto the trial wavefunction (Equation 3.9).

The majority of the quadrature grid and cutoff settings result in time-step errors that are mutually indistinguishable. Using the simpler 2J Jastrow factor, the use of a short range in the pseudopotentials has a larger effect than the number of quadrature points. Using the shortest range (corresponding to $z = 3$) produces noticeably higher energies, regardless of the number of quadrature points used. As the difference in cost between using ranges corresponding to $z = 3$ and $z = 5$ is so small, it is safer to use the larger value, which has no visible effect on the time-step error.

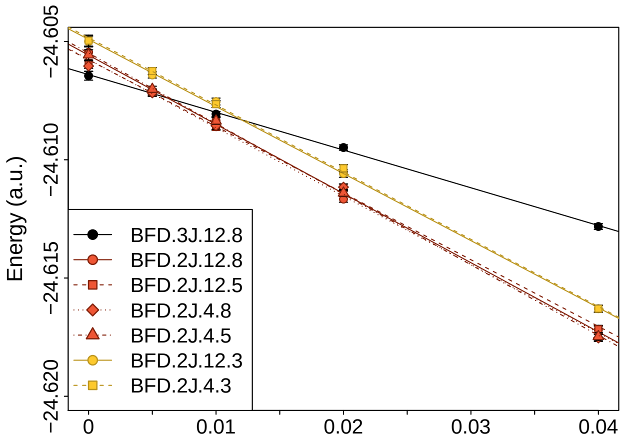


Figure 3.5: DMC energies as a function of imaginary time-step for the TS2 geometry, using the BFD pseudopotentials.

## 3.5 Summary

DMC has been used to calculate the reaction barrier heights of the two main channels for H abstraction of methanol by an H atom, a problem that requires a high-level treatment of electron correlation effects. The combination of B3LYP geometries and QMC energies predicts barrier heights that agree with CCSD(T) values to within chemical accuracy.

The cost of the DMC calculations can be minimised by simplifying the trial wavefunction and the treatment of non-local pseudopotentials. The largest cost saving can be achieved by using a simple Jastrow factor that includes only two-body correlation effects. By combining this simplified trial wavefunction with a sparse quadrature grid in the projection of the non-local pseudopotential, and applying cutoffs to the ranges of these potentials, the cost of DMC calculations was reduced by a factor of 5x over reference calculations, with no loss in accuracy.

In the notation defined in Sec. 3.4, our recommended protocol is 2J.4.5, using the BFD pseudopotentials. However, a caveat is that one should be careful with the choice of integration grids for systems containing much heavier elements. These cases are likely to be more sensitive to the number of quadrature points due to the importance of larger angular momenta in the pseudopotentials.

Together with these cost-reducing measures, the accuracy, favourable scaling, and low memory requirements of QMC methods indicate this is a practical route to tackle H abstraction in much larger systems.

## 3.6 References

[1] S. M. Sarathy, P. Oßwald, N. Hansen, K. Kohse-Höinghaus, *Prog. Energy Combust. Sci.* **2014**, *44*, 40–102.

[2] L. Pardo, J. R. Banfelder, R. Osman, *J. Am. Chem. Soc.* **1992**, *114*, 2382–2390.

[3] A. D. Boese, J. M. L. Martin, *J. Chem. Phys.* **2004**, *121*, 3405–3416.

[4] E. I. Izgorodina, D. R. B. Brittain, J. L. Hodgson, E. H. Krenske, C. Y. Lin, M. Namazian, M. L. Coote, *J. Phys. Chem. A* **2007**, *111*, 10754–10768.

[5] R. Meana-Pañeda, D. G. Truhlar, A. Fernández-Ramos, *J. Chem. Phys.* **2011**, *134*, 094302.

[6] E. Carvalho, A. N. Barauna, F. B. Machado, O. Roberto-Neto, *Chem. Phys. Lett.* **2008**, *463*, 33–37.

[7] E. Carvalho, A. N. Barauna, F. B. Machado, O. Roberto-Neto, *Int. J. Quantum Chem.* **2008**, *108*, 2476–2485.

[8] J. Pu, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 773–778.

[9] H.-H. Grotheer, S. Kelm, H. S. T. Driver, R. J. Hutcheon, R. D. Lockett, G. N. Robertson, *Ber Bunsenges Phys. Chem.* **1992**, *96*, 1360–1376.

[10] J. T. Jodkowski, M.-T. Rayez, J.-C. Rayez, T. Bérces, S. Dóbé, *J. Phys. Chem. A* **1999**, *103*, 3750–3765.

[11] Y.-Y. Chuang, M. L. Radhakrishnan, P. L. Fast, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. A* **1999**, *103*, 4893–4909.

[12] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[13] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory. Comput.* **2012**, *8*, 2181–2188.

[14] F. R. Petruzielo, J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2012**, *136*, 124116.

[15] D. M. Cleland, M. C. Per, *J. Chem. Phys.* **2016**, *144*, 124108–9.

[16] J. Kim, K. P. Esler, J. McMinis, D. M. Ceperley in *Proceedings of the Scientific Discovery through Advanced Computing SciDac Conference*, Chattanooga, Tennessee.

[17] F. Mentch, J. B. Anderson, *J. Chem. Phys.* **1984**, *80*, 2675–7.

[18] D. L. Diedrich, J. B. Anderson, *Science* **1992**, *258*, 786–788.

[19] J. B. Anderson, *J. Chem. Phys.* **2016**, *144*, 166101–3.

[20] J. C. Grossman, L. Mitas, *Phys. Rev. Lett.* **1997**, *79*, 4353–4356.

[21] A. C. Kollias, O. Couronne, W. A. Lester, *J. Chem. Phys.* **2004**, *121*, 1357–8.

[22] Y. Kanai, N. Takeuchi, *J. Chem. Phys.* **2009**, *131*, 214708–5.

[23] F. Fracchia, C. Filippi, C. Amovilli, *J. Chem. Theory. Comput.* **2013**, *9*, 3453–3462.

[24] H. Wendland, *Adv. Comput. Math.* **1995**, *4*, 389–396.

[25] N. D. Drummond, M. D. Towler, R. J. Needs, *Phys. Rev. B* **2004**, *70*, 235119.

[26] J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 084102.

[27] L. Mitas, E. L. Shirley, D. M. Ceperley, *J. Chem. Phys.* **1991**, *95*, 3467.

[28] P. O. Widmark, P. A. Malmqvist, B. O. Roos, *Theor. Chim. Acta* **1990**, *77*, 291–306.

[29] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr, J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, N. J. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, *Gaussian 09*, Gaussian, Inc., **2009**.

[30] M. C. Per, K. A. Walker, S. P. Russo, *J. Chem. Theory. Comput.* **2012**, *8*, 2255–2259.

[31] M. C. Per, S. P. Russo, I. K. Snook, *J. Chem. Phys.* **2008**, *128*, 114106.

[32] J. R. Trail, R. J. Needs, *J. Chem. Phys.* **2005**, *122*, 174109.

[33] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[34] M. Dolg, C. Filippi, *private communication* **2014**.

[35] J. Xu, M. J. Deible, K. A. Peterson, K. D. Jordan, *J. Chem. Theory. Comput.* **2013**, *9*, 2170–2178.

[36] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[37] M. Gillan, M. Towler, D. Alfe, *Psik Newsletter* **2011**, *103*, 32.

[38] M. C. Per, *CSIRO Quantum Monte Carlo software package*, **2017**.

[39] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, *WIRES Comput. Mol. Sci.* **2011**, *2*, 242–253.

[40] T. H. Dunning, *J. Chem. Phys.* **1989**, *90*, 1007–18.

[41] S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* **1980**, *58*, 1200–1211.

[42] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098–3100.

[43] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.

[44] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[45] S. Grimme, S. Ehrlich, L. Goerigk, *J. Comput. Chem.* **2011**, *32*, 1456–1465.

[46] J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, *Phys. Rev. Lett.* **2003**, *91*, 146401–4.

[47] Y. Zhao, D. G. Truhlar, *J. Chem. Phys.* **2006**, *125*, 194101–19.

[48] C. Adamo, V. Barone, *J. Chem. Phys.* **1999**, *110*, 6158–14.

[49] A. D. Becke, *J. Chem. Phys.* **1996**, *104*, 1040–8.

[50] Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2004**, *108*, 6908–6918.

[51] S. Grimme, *J. Chem. Phys.* **2006**, *124*, 034108–16.

[52] T. Schwabe, S. Grimme, *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4.

[53] T. J. Lee, P. R. Taylor, *Int. J. Quant. Chem.* **1989**, *36*, 199–207.

[54] J. C. Rienstra-Kiracofe, W. D. Allen, H. F. Schaefer, *J. Phys. Chem. A* **2000**, *104*, 9823–9840.

[55] M. Dubecký, R. Derian, P. Jurečka, L. Mitas, P. Hobza, M. Otyepka, *Phys. Chem. Chem. Phys.* **2014**, *16*, 20915–20923.

Chapter 4

# Calculating barrier heights with quantum Monte Carlo

## 4.1 Introduction

Calculating barrier heights can give detailed insight into reaction mechanisms that is unobtainable through experiment alone. This information can be used to build kinetic models of the atmosphere,[1] investigate which side reactions might compromise polymerisation without wasting material[2] and gain a better understanding of enzyme catalysis.[3] These applications require accurate barrier heights but different classes of reactions pose unique challenges for quantum chemistry methods.

DFT methods are particularly unreliable for reactions. They underestimate barrier heights in an erratic manner[4,5] and often fail to locate transition states.[6] Relative energies of barrier heights are also sensitive to the basis set used.[7,8] Functionals like mPW1K[9] and KMLYP[10] have been specifically parameterised using barrier heights but improved performance for kinetics comes at the cost of poor results for ground state properties like atomisation energies and geometries. Newer functionals like BMK[11] are parameterised with larger training sets and perform better for these ground state properties but still perform worse than B3LYP[12,13] for geometries and vibrational frequencies. Chemically reliable results are often only obtainable with high-level quantum chemistry methods but these are computationally expensive and scale poorly with system size.

QMC methods[14–18] are a promising alternative to traditional high-level electronic structure methods for these types of problems. By using stochastic integration they have greater freedom in the choice of trial wavefunction and give a better description of static and dynamic correlation in the system. DMC also has much smaller basis set truncation and basis set superposition errors compared to other electronic structure methods.[14]

The accuracy and reliability of QMC methods have been extensively demonstrated for basic atomic properties like total energies, ionisation potentials and electron affinities.[19–21] They have been benchmarked for the atomization energies of small molecules with the G1 and G2 test set.[22–25] Using multi-determinant trial wavefunctions in DMC reduced the mean absolute deviation from experimental values of the G2 test set to just 1.2 kcal/mol.[25] QMC methods describe both static and dynamic correlation effectively and have exceptional performance for noncovalent interactions, with a mean absolute deviation of 0.68 kcal/mol for the S22 test set[26] and just 0.15 kcal/mol for the A24 test set.[27]

Until very recently, QMC benchmarking for barrier heights had been carried out in a scattered manner. The simple exchange reaction barrier $H + H_2 \rightarrow H_2 + H$[28–30] was one of the first systems studied and has recently been revised to even higher accuracy.[31] Other simple reactions studied include five prototypical chemical reactions, including three hydrogen-exchange, one heavy atom exchange and one association reaction.[32] More challenging organic reactions have also been studied.[33–35] Chapter 3 presented a small study of the reaction barrier for hydrogen abstraction from methanol,[36] a simple reaction yet accurate barrier heights are typically only obtainable with the highest levels of quantum chemistry. DMC has performed well in all cases, with errors close to or less than the chemical accuracy standard of 1.0 kcal/mol. A recent study investigated the performance of DMC methods for a set of 19 hydrogen abstraction reactions. They compared the performance of DMC with all-electron basis-sets and pseudopotentials and in both cases the the MAD was 1.0 kcal/mol.[37] Another study looked at 19 non-hydrogen transfer reactions and the DMC error was 1.5 ± 0.5 kcal/mol.[38] This chapter presents a more systematic benchmarking study for a range of reaction classes, including radical stabilisation energies, Diels-Alder reactions and hydrogen and non-hydrogen transfer reactions.

## 4.2 Methods

Geometries and reference values for the test sets were taken from previous work, as outlined below. Single-determinant DMC calculations used B3LYP trial wavefunctions. Burkatzki-Filippi-Dolg (BFD) pseudopotentials[39] with the associated triple-zeta (VTZ) basis sets were used, with an improved H-atom potential.[40] Nonlocal pseudopotentials use size-consistent T-moves with a symmetric branching term.[41] Fixed-node Diffusion Monte Carlo calculations were performed using the CSIRO QMC code[42] with a target population size of 6400 walkers and an imaginary time step size of 0.01 a.u.

Configuration interaction (CI) calculations used Kohn-Sham orbitals from B3LYP density functional calculations incorporating single and double excitations (CISD). The use of Kohn-Sham orbitals in CI wavefunctions has been shown to give better DMC nodal surfaces than Hartree-Fock orbitals.[43] The CSFs were taken from the natural orbitals of the CISD wavefunction and were selected using the recently developed energy truncation method.[44] The variable parameters in the Jastrow factor and the CSF coefficients were optimised by minimising the variational energy using an approach based on the linear method.[45] B3LYP and CI calculations were performed with GAMESS.[46,47]

## 4.3 Radical stabilisation energy

Radical stabilisation energy (RSE) is defined as the reaction enthalpy ($\Delta H_{\text{rxn}}$) for the hydrogen transfer reaction:

$$\mathrm{X}\cdot + \mathrm{R{-}H} \xrightarrow{\Delta H_{\mathrm{rxn}}} \mathrm{X{-}H} + \mathrm{R}\cdot \tag{4.1}$$

For carbon-centred radicals, RSE is defined in terms of the methyl radical (i.e. $\mathrm{X}\cdot$ is replaced by the methyl radical, $\mathrm{CH_3}\cdot$ and X-H is its closed shell parent compound methane, $\mathrm{CH_4}$). Equivalent isodesmic reactions can be written for oxygen- or nitrogen-centred radicals, where $\mathrm{X}\cdot$ is $\mathrm{NH_2}\cdot$ or $\mathrm{OH}\cdot$ and the closed shell compound is ammonia or water.[48] Eq. 4.1 is essentially the difference in bond dissociation energies (BDE) of R−H and X−H. Absolute BDE calculations can challenging for electronic structure methods but radical stabilisation energies are defined by an isodesmic reaction that gives favourable error cancellation for most methods. Even moderate DFT methods will sometimes give reasonable values,[48] provided the percentage of exact exchange is high.[49,50]

The theoretical prediction of thermodynamic stabilities allows for quantitative comparison of substituent effects for radicals of varying structure and electronic characteristics. Accurate RSEs are needed to understand, and subsequently design, complex radical reactions. For example, RSEs have been used to identify reactions that involve endothermic hydrogen atom transfer steps.[51] They have also been used to identify hydrogen atom donors and acceptors that will accelerate homolytic hydrogen transfer in polarity reversal catalysis[52] and to distinguish between toxic and non-toxic general anaesthetics.[53] These types of applications require methods that are not only accurate but also consistent across a range of radicals and substituents. The performance of DMC for radical stabilisation energies is assessed here using RSE43, a collection of 43 H-abstraction energies for the reactions of hydrocarbons with a methyl radical ($\mathrm{R{-}H} + \mathrm{CH_3}\cdot \rightarrow \mathrm{R}\cdot + \mathrm{CH_4}$).[54] Reference CCSD(T)/CBS energies (ignoring ZPVE and thermal effects) using B3LYP/TZVP geometries were taken from previous work.[55] The reference value for reaction 4 was recalculated in this work at the CCSD(T)/CBS level using the standard two-point extrapolation scheme of Helgaker, Klopper and co-workers[56,57] with the Dunning basis sets cc-pVDZ and cc-pVTZ.[58,59] These calculations were carried out using MOLPRO 2012.[60]

DMC deviations from reference values are reported in Table 4.1 alongside results from some popular DFT and WFT methods for comparison.[54] DMC shows excellent agreement with CCSD(T)/CBS reference values for all reactions, with an overall MAD of 0.3 kcal/mol. The two main sources of errors in a DMC calculation come from the fixed-node approximation and the time-step bias but these cancel out for energy differences.[26,63–67] The isodesmic reactions in RSE43 also result in favourable error cancellation for other methods and this is seen in the generally low MADs for all methods. Although the DFT and WFT methods in Table 4.1 have small MADs, the overall error for DMC is significantly lower and surpasses chemical accuracy.

Closer inspection of Table 4.1 shows that despite the methods having small MADs there is considerable variability in the quality of results for individual reactions. The WFT methods MP2 and SCS-MP2 generally perform well and have an overall MAD of 3.1 and 3.5 kcal/mol respectively but fail catastrophically for reactions 1 and 2, where errors are

Table 4.1: Deviations of radical stabilisation energies from reference values for the RSE43 test set. Statistical uncertainties in the last of digit of the DMC results are shown in parentheses.

| | | | Deviation (kcal/mol) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Radical | Ref[a] | DMC | MP2[b] | SCS-MP2[b] | B1B95[b] | PWPB95[b] | B3LYP[b] | PBE0[b] | M05-2X[b] |
| 1 | $^{\bullet}CH_2-C_6H_5$ | -15.2 | 0.3(2) | -24.6 | -23.1 | 2.2 | 0.6 | 2.1 | 1.2 | 1.0 |
| 2 | $CH_2=C^{\bullet}-CN$ | 1.9 | -0.7(2) | -27.8 | -30.7 | 5.1 | 1.8 | 4.6 | 4.0 | 0.6 |
| 3 | $^{\bullet}CF=CH_2$ | 6.8 | 0.6(2) | -6.4 | -8.2 | 2.3 | 0.9 | 0.9 | 1.5 | -0.5 |
| 4 | $^{\bullet}CH_2-CCl_3$ | -0.9[c] | -0.4(2) | -1.5 | -1.5 | 2.1 | 1.1 | 2.5 | 1.6 | 0.9 |
| 5 | $^{\bullet}CH_2-CF_2-CH_3$ | 0.1 | 0.1(2) | -0.5 | -0.5 | 1.4 | 0.8 | 1.5 | 1.2 | 0.6 |
| 6 | $^{\bullet}CH_2-CF_3$ | 1.4 | 0.2(2) | -0.5 | -0.5 | 1.2 | 0.7 | 1.4 | 1.1 | 0.4 |
| 7 | $^{\bullet}CH_2-CH_2-Cl$ | -3.2 | -0.5(2) | -1.3 | -1.4 | 1.7 | 0.9 | 2.2 | 1.1 | 1.1 |
| 8 | $^{\bullet}CH_2-CH_2-F$ | -1.3 | 0.2(2) | -0.2 | -0.2 | 1.3 | 0.8 | 1.4 | 1.1 | 0.7 |
| 9 | $^{\bullet}CH_2-CH_2-OH$ | -1.8 | 0.1(2) | -0.4 | -0.5 | 1.6 | 1.0 | 1.4 | 1.4 | 0.7 |
| 10 | $^{\bullet}CH_2-CH=CH_2$ | -17.5 | 0.4(2) | -6.0 | -6.8 | 2.4 | 0.9 | 2.4 | 1.8 | 1.8 |
| 11 | $^{\bullet}CH_2-CHO$ | -10 | 0.0(2) | -9.5 | -9.7 | 2.2 | 0.7 | 2.1 | 1.4 | 1.4 |
| 12 | $^{\bullet}CH_2-CN$ | -8.6 | -0.6(2) | -11.1 | -11.8 | 2.4 | 0.8 | 2.5 | 1.7 | 0.5 |
| 13 | $^{\bullet}CH_2-CO-CH_3$ | -6.3 | 0.0(2) | -1.7 | -1.8 | 1.7 | 1.0 | 1.7 | 1.1 | 0.9 |
| 14 | $^{\bullet}CH_2-CO-NH_2$ | -6.3 | 0.3(2) | -1.5 | -1.6 | 1.8 | 1.1 | 1.8 | 1.2 | 0.9 |
| 15 | $^{\bullet}CH_2-CO-NH-CH_3$ | -6.6 | 0.4(2) | -2.4 | -2.6 | 1.7 | 0.9 | 1.7 | 1.2 | 0.9 |
| 16 | $^{\bullet}CH_2-CO-O-CH_3$ | -6.4 | -0.2(2) | -2.5 | -2.6 | 1.8 | 0.9 | 1.8 | 1.2 | 1.0 |
| 17 | $^{\bullet}CH_2-CO-OH$ | -3.0 | 0.2(2) | -0.5 | -0.6 | 1.5 | 0.8 | 1.4 | 1.2 | 0.6 |
| 18 | $^{\bullet}CH_2-CH(-CH_2)_2$ | -3.9 | 0.0(2) | -0.4 | -1.2 | 1.7 | 0.8 | 0.8 | 1.2 | 0.1 |
| 19 | $^{\bullet}CH_2-F$ | -12 | 0.4(2) | 0.1 | -0.7 | 3.1 | 2.0 | 2.3 | 1.9 | 1.6 |
| 20 | $^{\bullet}CH_2-NH_2$ | 4.7 | 0.0(2) | -0.1 | -0.5 | 1.1 | 0.6 | 0.8 | 1.1 | 0.1 |
| 21 | $^{\bullet}CH_2-NH_3{}^{+}$ | -12.6 | 0.1(2) | -0.1 | -0.8 | 3.4 | 2.2 | 2.6 | 2.0 | 1.5 |
| 22 | $^{\bullet}CH_2-NH-CH_3$ | -11.1 | -0.3(2) | -1.0 | -1.7 | 2.8 | 1.5 | 2.2 | 1.7 | 1.0 |
| 23 | $^{\bullet}CH_2-NH-CHO$ | -8.6 | 0.4(2) | -0.8 | -1.7 | 3.6 | 2.1 | 3.0 | 2.4 | 1.4 |
| 24 | $^{\bullet}CH_2-NH-OH$ | -12.8 | 0.1(2) | -0.5 | -1.0 | 3.6 | 2.2 | 2.8 | 2.0 | 1.2 |
| 25 | $^{\bullet}CH_2-N(-CH_3)_2$ | -3.3 | -0.1(2) | -1.5 | -2.2 | 2.1 | 1.0 | 2.0 | 1.6 | 0.9 |
| 26 | $^{\bullet}CH_2-NO_2$ | -3.9 | -0.2(2) | -0.7 | -1.4 | 2.0 | 1.0 | 1.2 | 1.4 | 0.2 |
| 27 | $^{\bullet}CH_2-O-CH_3$ | -2.7 | 0.0(2) | -0.4 | -0.8 | 2.2 | 1.3 | 1.8 | 1.7 | 1.2 |
| 28 | $^{\bullet}CH_2-O-CHO$ | -5.9 | -0.5(2) | -0.6 | -1.3 | 2.4 | 1.3 | 1.7 | 1.6 | 0.6 |
| 29 | $^{\bullet}CH_2-CO-O-CH_3$ | -6.2 | -0.4(2) | -0.7 | -1.4 | 2.8 | 1.5 | 2.1 | 1.9 | 0.8 |
| 30 | $^{\bullet}CH_2-OH$ | -4.2 | 0.1(2) | -0.2 | -0.8 | 2.0 | 1.1 | 1.4 | 1.6 | 1.1 |
| 31 | $^{\bullet}CH_2-PH_3{}^{+}$ | 0.7 | 0.2(2) | -0.5 | -0.2 | 1.0 | 0.6 | 1.0 | 0.6 | 0.2 |
| 32 | $^{\bullet}CH_2-S-CH_3$ | -10.8 | 0.4(2) | -0.7 | -1.3 | 3.0 | 1.9 | 2.1 | 1.8 | 1.4 |
| 33 | $^{\bullet}CH_2-S-CHO$ | -8.4 | 0.3(2) | -1.0 | -1.6 | 2.9 | 1.7 | 2.1 | 1.9 | 1.0 |
| 34 | $^{\bullet}CH_2-SH_2{}^{+}$ | 2.7 | 0.3(2) | -0.2 | -0.5 | 1.5 | 0.9 | 1.1 | 1.3 | 0.4 |
| 35 | $^{\bullet}CH_2-SH$ | -9.4 | 0.6(2) | -0.5 | -1.1 | 2.8 | 1.8 | 1.9 | 1.7 | 1.3 |
| 36 | $^{\bullet}CH_2-SO-O-CH_3$ | 0.0 | 0.1(2) | -0.9 | -0.9 | 1.8 | 1.1 | 1.6 | 1.3 | 0.1 |
| 37 | $^{\bullet}CH_2-SO-CH_3$ | -2.9 | 0.6(2) | -1.0 | -1.2 | 2.7 | 1.6 | 2.1 | 1.8 | 0.7 |
| 38 | $NH2\text{-}CH^{\bullet}-CN$ | -22.5 | -0.2(2) | -8.8 | -10.3 | 6.2 | 3.5 | 5.2 | 4.6 | 2.8 |
| 39 | $NH2\text{-}CH^{\bullet}-CO-NH_2$ | -24.1 | 0.5(2) | 0.3 | -0.8 | 5.9 | 3.8 | 4.8 | 4.1 | 3.4 |
| 40 | $NH2\text{-}CH^{\bullet}-CO-OH$ | -25.4 | 0.7(2) | 0.3 | -1.0 | 6.5 | 4.1 | 5.4 | 4.5 | 4.0 |
| 41 | $^{\bullet}CH_2-C\equiv CH$ | -13.1 | -0.6(2) | -9.6 | -10.8 | 3.1 | 1.2 | 3.1 | 2.4 | 1.4 |
| 42 | $^{\bullet}C(-CH_3)_3$ | -6.4 | -0.1(2) | -1.2 | -1.4 | 4.3 | 2.4 | 3.9 | 3.7 | 1.5 |
| 43 | $^{\bullet}CH_2-C(-CH_3)_3$ | -2.3 | 0.3(2) | -0.7 | -0.7 | 1.4 | 0.8 | 1.4 | 1.2 | 0.6 |
| | MAD | | 0.3(2) | 3.1 | 3.5 | 2.6 | 1.4 | 2.2 | 1.8 | 1.0 |

[a] CCSD(T)/CBS[55]

[b] Values taken from Ref. 61 , calculated using the large Ahlrichs' type quadruple-$\zeta$ basis sets def2-QZVP[62]

[c] Calculated in this work

greater than 20 kcal/mol. Previous studies have shown that MP2 and its variants struggle with reactions that involve highly delocalised radicals.[49,50,68] Reactions 38, 39 and 40 are particularly problematic for DFT methods and have errors more than twice the average error for the respective method. The best performing DFT method is M05-2X[69] with a MAD of just 1.0 kcal/mol. The results are generally consistent for all reactions but it was specifically parameterised with a set of carbon-carbon bond dissociation energies.[68] In contrast to the other methods presented here, DMC performs consistently well and deviations from reference values are less than 1 kcal/mol for all substituents. It was through this consistent performance that an error in the reference value for reaction 4 was found. The original reference value was -7.0 kcal/mol and DMC had an error of -7.4 kcal/mol. While an error this large wouldn't seem unusual for DFT methods it was in stark contrast with every other result for DMC. The reference value was recalculated using the same CCSD(T)/CBS method and found to be -0.9 kcal/mol, in much better agreement with the DMC value.

## 4.4 Diels-Alder reactions

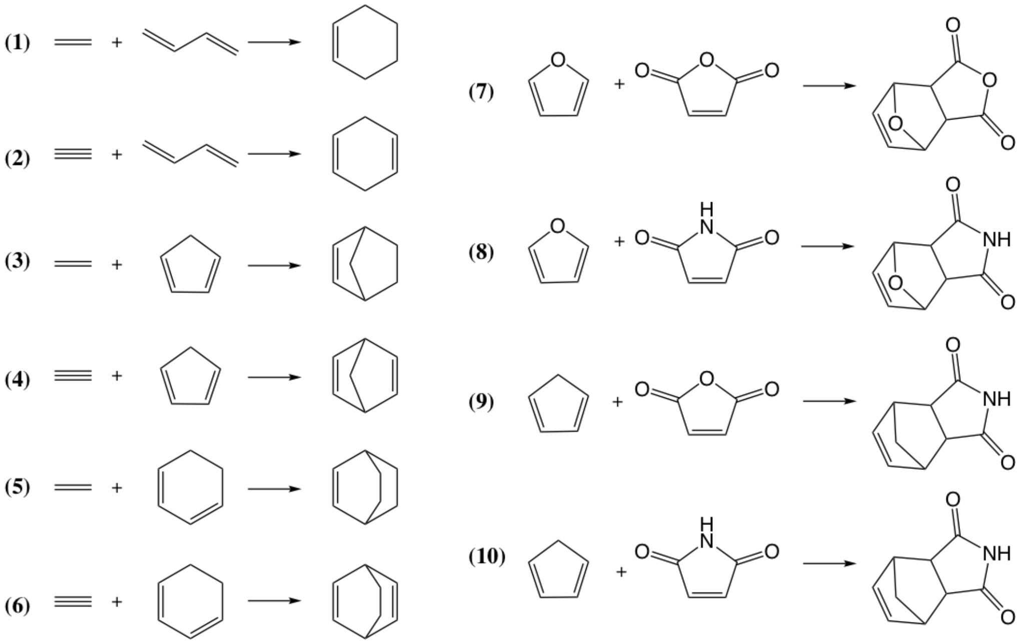


Figure 4.1: The Diels-Alder reactions included in the DARC test set.[54] The test set includes the endo and exo forms of the products of reactions 7-10. Reference energies are listed in Table 4.2.

DFT methods fail to treat fractional charges and distributes the electron densities artificially. The discrete nature of electrons means the exact energy of an atom with a fractional electron charge is a straight-line interpolation between the integers but approximate functionals are convex between the integers.[70] This is known as the delocalisation error. This error leads to DFT methods underbinding reactions that involve the formation of cyclic or bicyclic compounds, overstabilising charge-transfer complexes, overestimat-

Table 4.2: Deviations of reaction energies from reference values for the DARC test set. Statistical uncertainties in the last digit of the DMC results are shown in parentheses.

| | | Deviation (kcal/mol) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| No. | Ref[a] | DMC | MP2[b] | SCS-MP2[b] | B1B95[b] | PWPB95[b] | B3LYP[b] | PBE0[b] | M05-2X[b] |
| 1 | -43.8 | 3.6(2) | 6.1 | 1.4 | -1.0 | -0.4 | -10.1 | 5.1 | 2.9 |
| 2 | -59.3 | 1.7(2) | 3.0 | -0.8 | 2.1 | 0.9 | -6.2 | 8.6 | 4.2 |
| 3 | -30.0 | 1.8(2) | 4.7 | 0.2 | -3.4 | -2.1 | -14.9 | 0.6 | -0.9 |
| 4 | -33.1 | -0.3(2) | 3.3 | -0.8 | -0.1 | -0.2 | -11.5 | 3.6 | -0.1 |
| 5 | -36.5 | 2.9(2) | 6.7 | 1.4 | -2.8 | -1.4 | -14.1 | 1.7 | 1.5 |
| 6 | -48.2 | 0.4(2) | 4.9 | -0.4 | 0.1 | -0.1 | -11.3 | 4.4 | 2.3 |
| 7 | -14.4 | 0.5(3) | 1.4 | -1.5 | -8.2 | -5.1 | -18.6 | -4.1 | -2.4 |
| 8 | -16.2 | 0.7(3) | 1.1 | -1.6 | -8.1 | -5.0 | -18.0 | -3.7 | -2.3 |
| 9 | -17.2 | 0.4(3) | 1.6 | -1.5 | -8.3 | -5.2 | -18.8 | -4.1 | -2.3 |
| 10 | -19.2 | 0.9(3) | 1.3 | -1.6 | -8.1 | -5.1 | -18.1 | -3.6 | -2.2 |
| 11 | -31.6 | 1.2(3) | 5.3 | 0.2 | -6.6 | -3.8 | -18.7 | -2.7 | -0.7 |
| 12 | -32.1 | 1.1(3) | 5.1 | 0.0 | -6.4 | -3.7 | -18.4 | -2.4 | -0.7 |
| 13 | -34.1 | 1.1(3) | 5.5 | 0.2 | -6.7 | -3.9 | -18.9 | -2.7 | -0.7 |
| 14 | -34.4 | 1.2(3) | 5.3 | 0.1 | -6.3 | -3.7 | -18.4 | -2.3 | -0.6 |
| MAD | | 1.2(3) | 3.9 | 0.8 | 4.9 | 2.9 | 15.4 | 3.5 | 1.7 |

[a] CCSD(T)/CBS[54]

[b] Values taken from Ref. 61, calculated using the large Ahlrichs' type quadruple-$\zeta$ basis sets def2-QZVP[62]

ing the polarisabilities of extended polymers and underestimating HOMO-LUMO gaps.[71] Functionals that include a relatively large fraction of exact exchange can minimise this error but in general DFT methods perform poorly for these systems. Kohn-Sham orbitals give better nodal surfaces for DMC calculations[43, 72] and they have become the orbitals of choice for most calculations. QMC methods do not suffer from the delocalisation error but whether that error translates from DFT methods to poor nodal surfaces in the trial wavefunction warrants investigation. The performance of DMC is assessed here using a collection of 14 Diels-Alder reactions (DARC). Diels-Alder reactions involve the cycloaddition of a substituted alkene to a conjugated diene. This class of reaction is known to be problematic for DFT methods and the unexpectedly large errors are attributed to the delocalisation error.[71, 73] The DARC test set was originally studied by Johnson et al.[71] and includes the reactions of butadiene, cyclopentadiene, cyclohexadiene and furane with ethene, ethyne, maleine and maleinimide acting as dienophiles. Structures are shown in Figure 4.1. CCSD(T)/ CBS reference values using B3LYP/6-31G(*2df, p*) geometries were taken from previous work.[54]

DMC deviations are reported in Table 4.2. The DMC results are in good agreement with the CCSD(T)/CBS reference values despite the use of Kohn-Sham orbitals in the trial wavefunction. DMC relies on trial wavefunctions built using orbitals from other methods and reactions involving cycloaddition of substituted alkenes, like those in DARC, suffer from the delocalisation error when treated with DFT methods. Interestingly, the DMC/B3LYP results here consistently overestimate the reaction exothermicity, whereas B3LYP underestimates it.[71]

The largest errors are seen for reactions 1, 3 and 5, where the dienophile is ethene ($C_2H_4$). Much smaller errors are seen for their equivalent reactions with ethyne ($C_2H_2$) as the dienophile (reactions 2, 4 and 6). Ethene and ethyne are both part of the G2 test set,[74] a collection of atomisation energies of small molecules. DMC calculations show ethene has much larger errors compared to ethyne, reflecting the trend seen here. Accurate reaction energies rely on a balanced treatment of electron correlation effects for all species. The

Table 4.3: The 38 hydrogen-transfer (HTBH) and 38 non-hydrogen transfer (NHTBH) barrier heights in the BH76[54] test set.

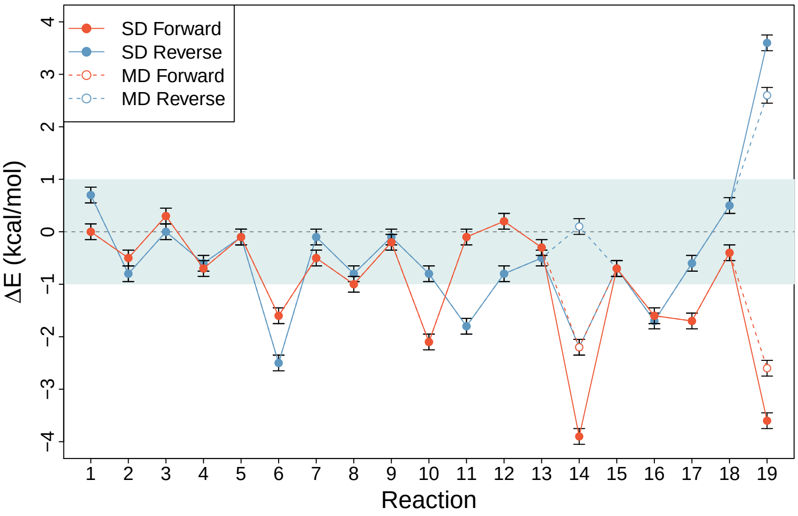
| ID | HTBH | NHTBH |
|---|---|---|
| 1 | $H + HCl \rightarrow H_2 + Cl$ | $H + N_2O \rightarrow OH + N_2$ |
| 2 | $OH + H_2 \rightarrow H + H_2O$ | $H + FH \rightarrow HF + H$ |
| 3 | $CH_3 + H_2 \rightarrow CH_4 + H$ | $H + ClH \rightarrow HCl + H$ |
| 4 | $OH + CH_4 \rightarrow CH_3 + H_2O$ | $H + FCH_3 \rightarrow HF + CH_3$ |
| 5 | $H + H_2 \rightarrow H + H_2$ | $H + F_2 \rightarrow HF + F$ |
| 6 | $OH + NH_3 \rightarrow H_2O + NH_2$ | $CH_3 + FCl \rightarrow CH_3F + Cl$ |
| 7 | $HCl + CH_3 \rightarrow Cl + CH_4$ | $F^- + CH_3F \rightarrow FCH_3 + F^-$ |
| 8 | $OH + C_2H_6 \rightarrow C_2H_5 + H_2O$ | $F^-...CH_3F \rightarrow FCH_3...F^-$ |
| 9 | $F + H_2 \rightarrow HF + F$ | $Cl^- + CH_3Cl \rightarrow ClCH_3 + Cl^-$ |
| 10 | $O + CH_4 \rightarrow OH + CH_3$ | $Cl^-...CH_3Cl \rightarrow ClCH_3...Cl^-$ |
| 11 | $H + PH_3 \rightarrow H_2 + PH_2$ | $F^- + CH_3Cl \rightarrow FCH_3 + Cl^-$ |
| 12 | $H + OH \rightarrow H_2 + O$ | $F^-...CH_3Cl \rightarrow FCH_3...Cl^-$ |
| 13 | $H + H_2S \rightarrow H_2 + HS$ | $OH^- + CH_3F \rightarrow HOCH_3 + F^-$ |
| 14 | $O + HCl \rightarrow OH + Cl$ | $OH^-...CH_3F \rightarrow HOCH_3...F^-$ |
| 15 | $NH_2 + CH_3 \rightarrow CH_4 + NH$ | $H + N_2 \rightarrow HN_2$ |
| 16 | $NH_2 + C_2H_5 \rightarrow C_2H_6 + NH$ | $H + CO \rightarrow HCO$ |
| 17 | $C_2H_6 + NH_2 \rightarrow NH_3 + C_2H_5$ | $H + C_2H_4 \rightarrow CH_3CH_2$ |
| 18 | $NH_2 + CH_4 \rightarrow NH_3 + CH_3$ | $CH_3 + C_2H_4 \rightarrow CH_3CH_2CH_2$ |
| 19 | $C_5H_8 \rightarrow TS$ | $HCN \rightarrow HNC$ |

discrepancies seen with the different dienophiles suggests a problem with the description of the electron correlation in ethene and ethyne compared to their respective products. For the larger systems there is a more favourable cancellation of errors. DMC performs well for this set of reactions, with a MAD of just 1.2 kcal/mol.

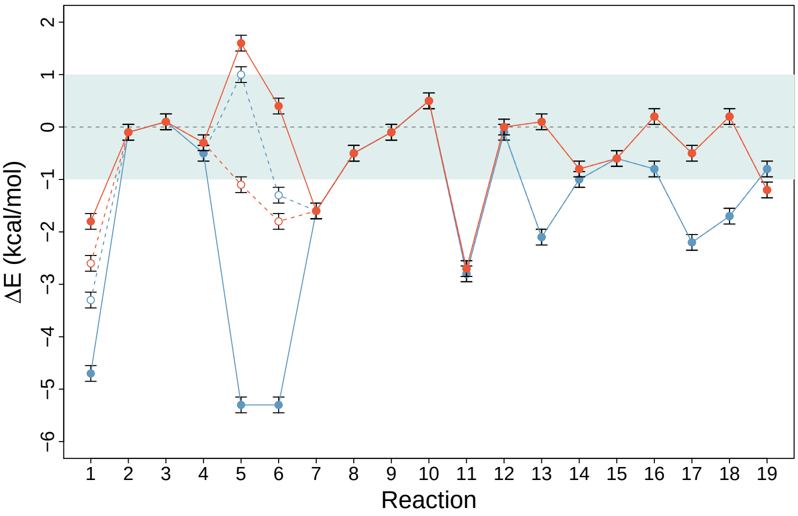
## 4.5 Hydrogen and non-hydrogen transfer barrier heights

The RSE43 and DARC test sets focus on specific properties and only reflect a small part of chemical space. DMC performs extremely well for these problems but this does not guarantee the errors will be small for other systems. The BH76 test set combines two sets of transition state barrier heights for 38 hydrogen transfer (HTBH) and 38 non-hydrogen transfer (NHTBH) reactions. The NHTBH database consists of 12 heavy-atom transfer barrier heights, 16 nucleophilic substitution barrier heights and 10 unimolecular or association reaction barrier heights. The best estimates for the barrier heights are taken from previous work and were obtained from a combination of experimental and theoretical data using QCISD/MG3 geometries.[75,76]

DMC errors for BH76 are shown in Figure 4.2. Overall the DMC values agree closely with the reference values and the mean absolute deviation (MAD) is 1.1 ± 0.2 kcal/mol. Hydrogen transfer barrier heights (Figure 4.2a) show better agreement with reference values than non-hydrogen transfer barrier heights (Figure 4.2b). Heavy-atom transfer reactions have a MAD of 1.7 ± 0.2 kcal/mol compared to 0.9 ± 0.2 kcal/mol for nucle-



(a)



(b)

Figure 4.2: Deviation of DMC from reference values for the a) hydrogen transfer (HTBH) and b) non-hydrogen transfer (NHTBH) subsets of BH76. The reactions are defined in Table 4.3. SD refers to single determinant DMC, MD refers to multi-determinant DMC. Full details can be found in the text. The shaded region denotes nominal chemical accuracy, ± 1 kcal/mol. The lines between points are simply drawn to guide the eye.

ophilic substitution and 1.0 ± 0.2 kcal/mol for unimolecular and association reactions. Previous studies have shown that these heavy atom transfer reactions are generally the most challenging for *ab initio* WFT and hybrid-*meta* DFT compared to the other reaction classes.[76]

Some of the reactions in this test set have previously been studied using QMC methods, including HTBH reactions 2,[33] 4, 6, and 18[32] and NHTBH reactions 1, 5, 8 and 15.[6,32] In all cases the single-determinant reaction barrier heights presented here are in good agreement with previous results. The entire HTBH test set was recently used to investigate the effect pseudopotentials have on these types of reactions.[37] The BFD results reported here agree closely with theirs and the overall MAD is the same (1.0 kcal/mol). The NHTBH test set has also been studied with DMC and the results are statistically equivalent.[38]

### Multi-determinant study

Single-determinant DMC with B3LYP orbitals in the trial wavefunction had errors less than 1.0 kcal/mol for most but not all barrier heights in BH76. Reactions 14 and 19 from HTBH and 1, 5, and 6 from NHTBH had particularly large errors. These reactions have problematic species like atomic oxygen and NHTBH reaction 1 is known to have multireference character.[6] It has been demonstrated with other systems that including more determinants in the trial wavefunction improves the reaction barrier results.[32] More specifically, a DMC study of NHTBH Reaction 5 used a complete active space wavefunction (CASSCF(3,3)) which was shown to significantly reduce the error.[6]

To investigate this further, multi-determinant expansions were built for HT14 and NHT1, NHT5 and NHT6 with CISD wavefunctions using an active space incorporating all valence electrons and sufficient virtual levels to close the n=3 shell (13 levels per atom) for heavy atoms. One virtual level was included for each H atom. For HTBH reaction 19 the *d*-orbitals were omitted (8 levels per heavy atom).

The number of configuration state functions (CSFs) produced by multi-reference methods is usually too large for DMC methods and they are truncated according to some threshold. Traditionally CSFs are truncated to keep a fixed number of terms[21] or using a cutoff on the CI expansion coefficients.[19,77–79] An alternative solution uses the energetic contribution each CSF makes to the total energy.[44] CSFs are ordered according to their energetic contributions to the total CI energy before being summed in increasing order until some cut-off threshold is reached, $E_{\text{trunc}}$.

Three different variations of this energy-based truncation were used here. The first, hereafter referred to as single-point truncation, uses the same threshold for all species involved in a reaction. The second scheme uses different cut-offs for transition states and products (or reactants) to maximise error cancellation. This approach is referred to as 'sum truncation', and $E_{\text{trunc}}$ is selected according to Equation 4.2:

$$E^{TS}_{\text{trunc}} = E^{P1}_{\text{trunc}} + E^{P2}_{\text{trunc}} \tag{4.2}$$

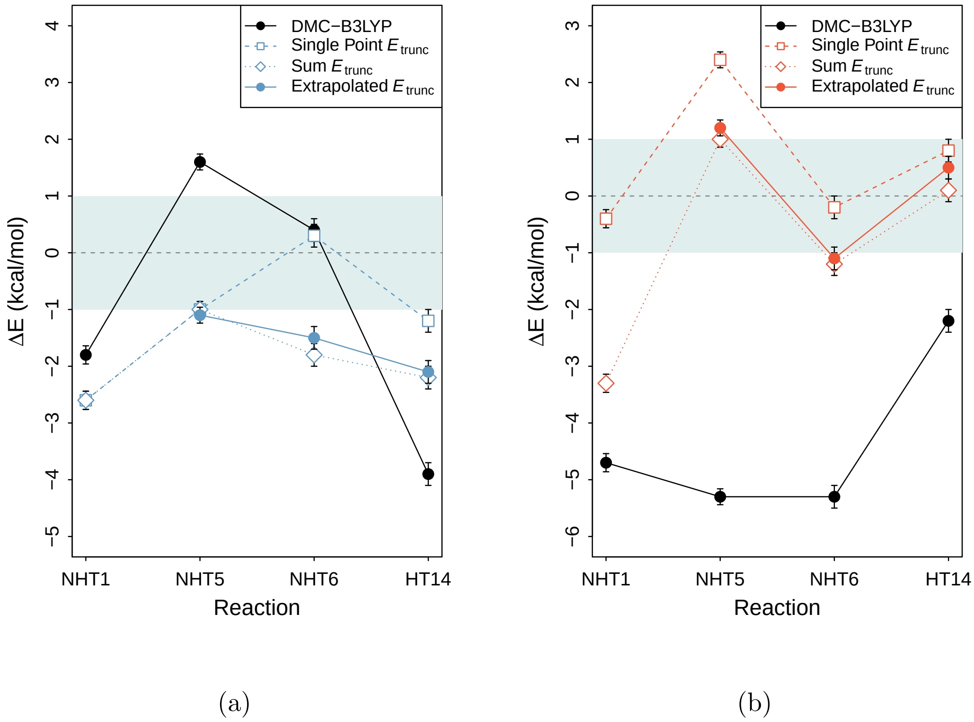


Figure 4.3: Deviation of DMC reaction energies from reference values for the (a) forward and (b) reverse barrier heights of four reactions taken from BH76. The reactions are defined in Table 4.3. 'Single-point $E_{\text{trunc}}$' uses the largest truncation value for each molecule shown in Table 4.4, 'Sum $E_{\text{trunc}}$' uses the largest truncation value on the transition states and a smaller value for the products (or reactants). 'Extrapolated $E_{\text{trunc}}$' extrapolates the energy to the $E_{\text{trunc}} = 0$ limit. NHT1 could not be extrapolated due to the system size. The shaded region denotes chemical accuracy, ± 1 kcal/mol. The lines between points are merely drawn to guide the eye.

where $P1$ and $P2$ are the products (or reactants) and $TS$ is the transition state. For reactions where H is a product or reactant $E^{TS}_{\text{trunc}} = E^{P1}_{\text{trunc}}$. This approach has the benefit of allowing larger cutoffs (and fewer CSFs) for transition states and larger systems. The third scheme extrapolates the individual energies to the $E_{\text{trunc}} = 0$ limit. A comparison of the results from the truncation schemes is shown in Figure 4.3. The number of CSFs included in the DMC calculation after the truncation was applied are shown in Table 4.4.

In general, the use of a multi-determinant wavefunction is better than a single-determinant for these problematic species. The sum-truncation method gives better agreement with the extrapolated values compared to the single point method. In some cases (i.e. NHT6 and NHT14 forward barriers) the single-point energy agrees better with the reference value and the extrapolated energy is worse than the single reference value. This is a side-effect of an inadequate active space for the extrapolation and a fortuitous cancellation of errors for the single-point truncation. The results could be further improved by using a larger active space but this simple active space is adequate for most barriers considered here. The transition state for NHT1 was too large to extrapolate to the CSF limit but based on

Table 4.4: Summary of the final number of CSFs used in the DMC calculation of each molecule for different $E_{\text{trunc}}$ values. 'Single-point $E_{\text{trunc}}$' calculations used the largest $E_{\text{trunc}}$ value shown. 'Sum $E_{\text{trunc}}$' calculations used the larger value on the transition state (TS) but smaller values on other molecules.

(a) NHT1

| Mol | $E_{\text{trunc}} = 0.07$ | $E_{\text{trunc}} = 0.035$ |
|---|---:|---:|
| $N_2O$ | 576 | 1141 |
| OH | 1 | 7 |
| $N_2$ | 45 | 110 |
| TS1 | 1053 | 2430 |

(b) NHT5

| Mol | $E_{\text{trunc}} = 0.02$ | $E_{\text{trunc}} = 0.01$ |
|---|---:|---:|
| $F_2$ | 299 | 422 |
| HF | 11 | 20 |
| F | 35 | 49 |
| TS5 | 274 | 509 |

(c) NHT5

| Mol | $E_{\text{trunc}} = 0.02$ | $E_{\text{trunc}} = 0.01$ |
|---|---:|---:|
| $CH_3$ | 14 | 36 |
| FCl | 326 | 495 |
| CH3F | 99 | 236 |
| Cl | 35 | 50 |
| TS6 | 923 | 1843 |

(d) HT14

| Mol | $E_{\text{trunc}} = 0.02$ | $E_{\text{trunc}} = 0.01$ |
|---|---:|---:|
| O | 29 | 42 |
| HCl | 29 | 51 |
| OH | 36 | 67 |
| Cl | 35 | 50 |
| TS14 | 1072 | 1925 |

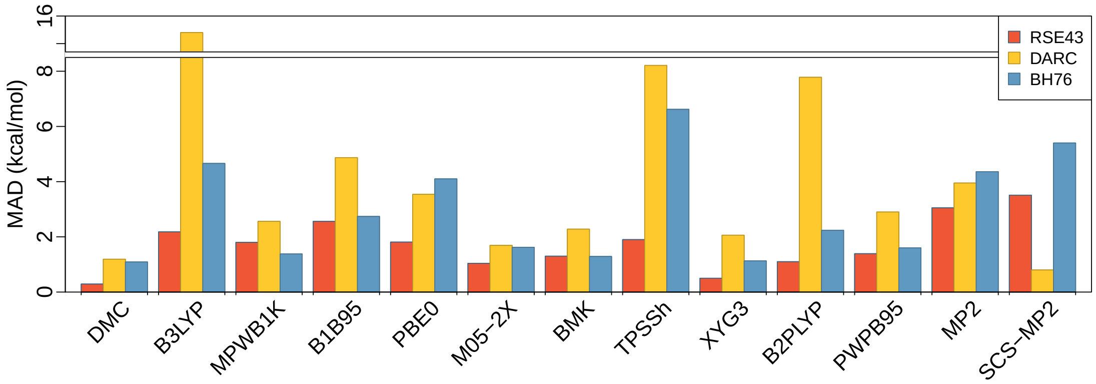


Figure 4.4: Mean absolute deviation (kcal/mol) for the three databases. Methods shown include *ab initio* wavefunction, hybrid functionals and double hybrid functionals (values taken from Ref. 54).

results here the sum-truncation results gives the same answer as the extrapolated limit.

Based on the performance of the multi-determinant wavefunctions and truncation schemes, the final sum-truncation multi-determinant values are included in Figure 4.2. For HTBH reaction 19, the active space incorporated all valence electrons and up to 8 virtual levels for each heavy atom. This reduced to error by approximately 1 kcal/mol but there is still a large deviation from experimental values. Using these improved wavefunctions gives better agreement with the reference values and reduces the mean absolute deviation to 0.9 kcal/mol for both databases.

## 4.6 Comparison to other methods

Finally, the DMC results are compared to some popular DFT and WFT methods in Figure 4.4. DMC is accurate for all three test sets but the performance of other methods is much less consistent. In general, the DFT methods perform well for RSE43 but this is largely due to the error cancellation in the isodesmic reactions. XYG3[80] is the best performing functional for RSE43. DARC is the most challenging for DFT methods and this is attributed to the delocalisation error. SCS-MP2 is an *ab initio* wavefunction method not affected by the delocalisation error and has the smallest error for DARC. Functionals like MPWB1K[81] and BMK[11] are specifically parameterised for reaction energies and perform well for the BH76 dataset.

## 4.7 Summary

DMC has been benchmarked using three test sets to investigate performance for radical stabilisation energies, Diels-Alder reactions and barrier heights of hydrogen and non-hydrogen transfer reactions. DMC had consistently low MADs for all three reactions classes. For reactions with large errors DMC could be systematically improved by using a multi-determinant wavefunction. Although the best DFT methods for individual test sets are competitive with DMC the quality varies dramatically with choice of exchange-correlation functional. There is no *a priori* way of making a good choice of DFT method. The exceptional performance of DMC, coupled with its favourable scaling, makes it an ideal method for providing reference results for reaction barrier heights, which can be used to benchmark lower level methods such as DFT.

## 4.8 References

[1] L. Vereecken, J. Francisco, *Chem. Soc. Rev.* **2012**, *41*, 6259–6293.

[2] G. Gryn'ova, C. Y. Lin, M. L. Coote, *Polym. Chem.* **2013**, *4*, 3744–3754.

[3] J. Chalupský, T. A. Rokob, Y. Kurashige, T. Yanai, E. I. Solomon, L. Rulíšek, M. Srnec, *J. Am. Chem. Soc.* **2014**, *136*, 15977–15991.

[4] Y. Feng, L. Liu, J.-T. Wang, H. Huang, Q.-X. Guo, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2005–2013.

[5] A. J. Cohen, P. Mori-Sanchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289–320.

[6] S. Saccani, C. Filippi, S. Moroni, *J. Chem. Phys.* **2013**, *138*, 084109.

[7] J. L. Durant, *Chem. Phys. Lett.* **1996**, *256*, 595 – 602.

[8] S. Parthiban, D. Oliveira, J. M. L. Martin, *J. Phys. Chem. A* **2001**, *105*, 895–904.

[9] B. J. Lynch, P. L. Fast, M. Harris, D. G. Truhlar, *J. Phys. Chem. A* **2000**, *104*, 4811–4815.

[10] J. K. Kang, C. B. Musgrave, *J. Chem. Phys.* **2001**, *115*, 11040–11051.

[11] A. D. Boese, J. M. L. Martin, *J. Chem. Phys.* **2004**, *121*, 3405–3416.

[12] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[13] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785–789.

[14] B. Austin, D. Y. Zubarev, W. A. Lester, *Chem. Rev.* **2012**, *112*, 263–88.

[15] M. Bajdich, L. Mitáš, *Acta Physica Slovaca* **2009**, *59*, 81–168.

[16] A. Lüchow, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 388–402.

[17] R. J. Needs, M. D. Towler, N. D. Drummond, P. López Ríos, *J. Phys. Condens. Matter* **2010**, *22*, 023201.

[18] T. Pang, *Am. J. Phys.* **2014**, *82*, 980–988.

[19] W. A. Al-Saidi, *J. Chem. Phys.* **2008**, *129*, 064316.

[20] P. Maldonado, A. Sarsa, E. Buenda, F. J. Glvez, *J. Chem. Phys.* **2010**, *133*, 064102.

[21] P. Seth, P. López Ríos, R. J. Needs, *J. Chem. Phys.* **2011**, *134*, 084105.

[22] J. Grossman, *J. Chem. Phys.* **2002**, *117*, 1434–1440.

[23] N. Nemec, M. D. Towler, R. J. Needs, *J. Chem. Phys.* **2010**, *132*, 034111.

[24] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **2012**, *8*, 2181–2188.

[25] F. R. Petruzielo, J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2012**, *136*, 124116.

[26] M. Korth, A. Lüchow, S. Grimme, *J. Phys. Chem. A* **2008**, *112*, 2104–2109.

[27] M. Dubecký, R. Derian, P. Jurečka, L. Mitáš, P. Hobza, M. Otyepka, *Phys. Chem. Chem. Phys.* **2014**, *16*, 20915.

[28] R. N. Barnett, P. J. Reynolds, W. A. Lester, *J. Chem. Phys.* **1985**, *82*, 2700–2707.

[29] D. L. Diedrich, J. B. Anderson, *Science* **1992**, *258*, 786–788.

[30] D. L. Diedrich, J. B. Anderson, *J. Chem. Phys.* **1994**, *100*, 8089–8095.

[31] J. B. Anderson, *J. Chem. Phys.* **2016**, *144*, 166101.

[32] F. Fracchia, C. Filippi, C. Amovilli, *J. Chem. Theory Comput.* **2013**, *9*, 3453–3462.

[33] J. Grossman, L. Mitáš, *Phys. Rev. Lett.* **1997**, *79*, 4353–4356.

[34] C. Filippi, S. B. Healy, P. Kratzer, E. Pehlke, M. Scheffler, *Phys. Rev. Lett.* **2002**, *89*, 166102.

[35] M. Barborini, L. Guidoni, *J. Chem. Phys.* **2012**, *137*, 224309.

[36] E. T. Swann, M. L. Coote, A. S. Barnard, M. C. Per, *Int. J. Quantum Chem.* **2017**, e25361.

[37] X. Zhou, F. Wang, *J. Comput. Chem.* **2017**, *38*, 798–806.

[38] K. Krongchon, B. Busemeyer, L. Wagner, *J. Chem. Phys.* **2017**, *146*, 124129.

[39] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[40] M. Dolg, C. Filippi, *private communication* **2014**.

[41] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[42] M. C. Per, *CSIRO Quantum Monte Carlo software package*, **2017**.

[43] R. C. Clay III, M. A. Morales, *J. Chem. Phys.* **2015**, *142*, 234103.

[44] M. C. Per, D. M. Cleland, *J. Chem. Phys.* **2017**, *146*, 164101.

[45] J. Toulouse, R. Assaraf, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 244112.

[46] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[47] M. Gordon, M. Schmidt in *Theory and Applications of Computational Chemistry: the first forty years*, C. Dykstra, G. Frenking, K. Kim, G. Scuseria (Eds.), Elsevier, Amsterdam, **2005**, pp. 1167–1189.

[48] H. Zipse, *Radicals Synth. I* **2006**, *263*, 163–189.

[49] E. I. Izgorodina, M. L. Coote, *J. Phys. Chem. A* **2006**, *110*, 2486–2492.

[50] E. I. Izgorodina, D. R. B. Brittain, J. L. Hodgson, E. H. Krenske, C. Y. Lin, M. Namazian, M. L. Coote, *J. Phys. Chem. A* **2007**, *111*, 10754–10768.

[51] D. Šakić, H. Zipse, *Adv. Synth. Cat.* **2016**, *358*, 3983–3991.

[52] J. Hioe, H. Zipse, *Org. Biomol. Chem.* **2010**, *8*, 3609–3617.

[53] S. Bhatia, V. Dixit, H. Jangra, P. Bharatam, *Drug Metab. Lett.* **2013**, *6*, 221–234.

[54] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2010**, *6*, 107–126.

[55] F. Neese, T. Schwabe, S. Kossmann, B. Schirmer, S. Grimme, *J. Chem. Theory Comput.* **2009**, *5*, 3060–3073.

[56] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, A. K. Wilson, *Chem. Phys. Lett.* **1998**, *286*, 243–252.

[57] T. Helgaker, W. Klopper, H. Koch, J. Noga, *J. Chem. Phys.* **1997**, *106*, 9639–9646.

[58] D. E. Woon, T. H. Dunning, *J. Chem. Phys.* **1995**, *103*, 4572–4585.

[59] T. H. Dunning Jr, *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[60] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, Y. Liu, A. W. Lloyd, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, M. Wang, *MOLPRO, version 2015.1, a package of ab initio programs*, **2015**, see.

[61] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

[62] F. Weigend, R. Ahlrichs, *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.

[63] C. Diedrich, A. Lüchow, S. Grimme, *J. Chem. Phys.* **2005**, *123*, 184106.

[64] N. D. Drummond, P. López Ríos, A. Ma, J. R. Trail, G. G. Spink, M. D. Towler, R. J. Needs, *J. Chem. Phys.* **2006**, *124*, 224104.

[65] I. G. Gurtubay, R. J. Needs, *J. Chem. Phys.* **2007**, *127*, 124306.

[66] M. D. Towler in *Comput. Methods Large Syst. Electron. Struct. Approaches Biotechnol. Nanotechnol.*, J. R. Reimers (Ed.), **2011**.

[67] M. Dubecký, P. Jurečka, R. Derian, P. Hobza, M. Otyepka, L. Mit, *J. Chem. Theory Comput.* **2013**, *9*, 4287–4292.

[68] E. I. Izgorodina, M. L. Coote, L. Radom, *J. Phys. Chem. A* **2005**, *109*, 7558–7566.

[69] Y. Zhao, N. E. Schultz, D. G. Truhlar, *J. Chem. Theory Comput.* **2006**, *2*, 364–382.

[70] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Science* **2008**, *321*, 792–794.

[71] E. R. Johnson, P. Mori-Sánchez, A. J. Cohen, W. Yang, *J. Chem. Phys.* **2008**, *129*, 204112.

[72] M. C. Per, K. Walker, S. P. Russo, *J. Chem. Theory Comput.* **2012**, *8*, 2255–2259.

[73] V. Guner, K. S. Khuong, A. G. Leach, P. S. Lee, M. D. Bartberger, K. N. Houk, *J. Phys. Chem. A* **2003**, *107*, 11445–11459.

[74] L. A. Curtiss, K. Raghavachari, *J. Chem. Phys.* **1991**, *94*, 7221–7230.

[75] Y. Zhao, B. J. Lynch, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2005**, *7*, 43–52.

[76] Y. Zhao, N. González-García, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 2012–2018.

[77] P. M. Zimmerman, J. Toulouse, Z. Zhang, C. B. Musgrave, C. J. Umrigar, *J. Chem. Phys.* **2009**, *131*, 124103.

[78] O. Valsson, C. Filippi, *J. Chem. Theory Comput.* **2010**, *6*, 1275–1292.

[79] L. Koziol, M. A. Morales, *J. Chem. Phys.* **2014**, *140*, 224316.

[80] Y. Zhang, X. Xu, W. A. Goddard, *Proc. Nat. Acad. Sci.* **2009**, *106*, 4963–4968.

[81] Y. Zhao, B. J. Lynch, D. G. Truhlar, *J. Phys. Chem. A* **2004**, *108*, 2715–2719.

Chapter 5

# Ionisation potentials and electron affinities of first- and second-row atoms

## 5.1 Introduction

Deriving accurate physical properties from electronic wavefunctions is only possible when there is a good description of the dynamic and non-dynamic electron correlations for all species. Hartree-Fock theory does not take into account the correlation arising from interacting electrons and recovering this energy is one of the greatest challenges for quantum chemical methods. The correlation energy is only a small component of the total energy of a system but it is extremely important for energy differences, especially when the energy difference is small. Finding an equivalent description of electron correlation on charged and neutral species with the same atomic number is challenging and the electron affinity and ionisation potential can be used to measure a method's performance. The electron affinity of an atom is defined as the difference in energy between the neutral and anionic species. Similarly, the ionisation potential is defined as the energy difference between the neutral and cationic form.

QMC methods recover a significant amount of this correlation energy and are well suited for these types of problems. First-row atoms and ions have been thoroughly studied.[1–15] Early work looked at the electron affinity of the fluorine atom. Using DMC with a small, double-zeta basis set recovered over 90% of the correlation energy for the neutral atom and anion. The calculated electron affinity was 3.45±0.11 eV agreed closely with the experimental value of 3.40 eV.[15] Another study looked at the VMC energies of first-row cations $Li^+$ through $Ne^+$ and anions $B^-$ through $F^-$. The overall mean absolute deviation was 110 meV for ionisation potentials and 70 meV for electron affinities.[13] Subsequent work showed that using trial wavefunctions that included more configurations for species like Be, B and C to account for the $2s - 2p$ near degeneracy recovered more correlation energy.[7,8] QMC methods also recover a significant amount of correlation energy for second row atoms.[1,8,12] Other systems studied include post-d group 13-17 elements[16] and third-row transition metals iron[17] and copper.[18] Thus far chemically accurate results, defined as an error less than 43 meV, are only consistently attainable for first-row atoms with multi-determinant wavefunctions[1,7] or more expensive methods like full-configuration interaction QMC.[4,6]

The aims of this chapter are twofold. The first is to establish the accuracy of the CSIRO QMC code (CMQMC[19]) for these atomic properties. QMC programs have the same underlying formula but the actual implementation of the algorithms can differ. A significant advantage of *in silico* experiments over physical experiments is the reproducibility and consistency of results but this is dependent on the parameters and algorithms within the programs. This is a well known problem for DFT methods where differences in the integration grid size can affect the precision of the final result.[20–23] In the case of the G2/97 test the B3LYP total energy was different by as much as 0.5 kcal/mol for some programs.[22] For newer meta-GGA functionals these errors were as large as 3.21 kcal/mol for M06HF for reaction energies.[23] As seen in Chapter 3 there are many parameters that can be adjusted in QMC methods that can affect the cost and accuracy of calculations. The sensitivity of ionisation potentials and electron affinities of first- and second-row atoms to electronic correlation treatment make them a good test set to validate a new code.

The second aim of this chapter is to test the performance of the energy-consistent Burkatzki-Fillipi-Dolg (BFD) pseudopotential[27] for these systems where electron correlation is important. QMC methods scale as approximately $N^{3-4}$ with respect to system size, $N$, but the large energy fluctuations of core electrons increases the scaling to approximately $Z^{5.5-6.5}$ with nuclear charge, $Z$.[24,25] Pseudopotentials are routinely used in QMC calculations to replace core electrons with an effective potential. The chemically active valence electrons still feel the same electronic field but the large energy and small length scale associated with the chemically inert core electrons is removed and the computational cost is significantly reduced.[26] Most applications of DMC would be impossible without pseudopotentials but approximating the core electrons with an effective potential introduces a systematic error. The BFD pseudopotential is constructed to reproduce HF energies of the ground state and some excited states of an atom. It is a popular pseudopotential and is routinely used in QMC calculations. The errors introduced by the approximation are usually small and cancel out for energy differences. Benchmarking studies have shown good performance in QMC relative to all-electron results.[27–29] A study on post-d group 13-17 elements showed single-determinant DMC had an error of 62.8±0.6 meV for ionisation potentials and 50.00 ± 0.04 meV for electron affinities (third-row elements were omitted because of their large spin-orbit effects).[16] The BFD pseudopotential was built specifically for QMC but it can be used with other methods. A study of the ionisation potential and electron affinity of H and Li atoms with RCCSD(T) showed results differ from experiment by 50 meV for IP and 10 meV for EA.[30] No benchmarking study has investigated the effect of pseudopotentials on QMC IPs and EAs of first- and second-row atoms.

## 5.2 Methods

Single-determinant DMC calculations used B3LYP orbitals in the trial wavefunctions. Configuration interaction (CI) calculations used Kohn-Sham orbitals from B3LYP density functional calculations incorporating single and double excitations (CISD). The use of Kohn-Sham orbitals in CI wavefunctions has been shown to give better DMC nodal surfaces than Hartree-Fock orbitals.[31] The CSFs were selected using the recently developed energy truncation method.[32] All molecules used an $E_{\text{trunc}}$ value of 0.01, except for CISDTQ/AE wavefunctions for O and F (and their respective ions) which use $E_{\text{trunc}}$=0.015. It has been shown[32] that there is usually a linear relationship between the DMC energy and trunction, with a gradient of approximately 0.1. For the truncation values used here it is expected that the total energies are converged to within 0.001 and 0.0015 Hartree respectively (0.027-0.04 eV). The variable parameters in the Jastrow factor and the CSF coefficients were optimised by minimising the variational energy using an approach based on the linear method.[33]

All-electron calculations used the Roos ATZ basis sets[34] and cusp-corrected orbitals.[35] Pseudopotential calculations replace the core electrons with an effective potential. Here the BFD pseudopotentials[27] with the associated triple-zeta (VTZ) basis sets were used, with an improved H-atom potential.[36] In DMC calculations involving these non-local pseudopotential the size-consistent T-moves scheme was used with a symmetric branching term.[37] B3LYP and CI calculations were performed with GAMESS.[38,39] Fixed-node diffusion Monte Carlo calculations were performed using the CSIRO QMC code[19] with a target population size of 6400 walkers. An imaginary time step size of 0.01 a.u. was used for pseudopotential calculations. Smaller time-steps are required for all-electron calculations and a time-step of 0.001 a.u. was used.

## 5.3 Single-determinant DMC

Ionisation potentials and electron affinities were calculated for all first- and second-row atoms except for the electron affinities of Be, N, Ne and Mg, which do not form stable anions. Experimental reference values are taken from Ref 7. Scalar relativistic effects are built into the BFD pseudopotential but are not included in the all-electron results. Accordingly, BFD errors are calculated using reference values that include relativistic effects and all-electron errors use reference values with these effects removed.

All-electron (AE) ionisation potentials and electron affinities for the atoms Li through Ar are shown in Tables 5.1 and 5.2. Deviations from (nonrelativistic) experimental reference values are shown in Figure 5.1. The DMC-B3LYP/AE electron affinities agree closely with experimental reference values and have a mean absolute deviation (MAD) of just 34 ± 6 meV. The quality of ionisation potentials is lower and the overall MAD is 77 ± 6 meV. For ionisation potentials and electron affinities to be accurate there must be a balanced description of electronic correlation on all species. This is challenging for neutral and charged species with the same atomic number but it is especially difficult for systems

Table 5.1: All-electron (AE) ionisation potentials (eV) for the atoms Li through Ar calculated with DMC and different trial wavefunctions. Mean absolute deviation (MAD) is reported in meV. Uncertainty in last digit of DMC results is shown in parentheses.

| | Ref.[a] | B3LYP/AE | CISD/AE | CISDTQ/AE | POEP-SC[a] | POEP-MC[a] | POEP-2014[b] |
|---|---|---|---|---|---|---|---|
| Li | 5.3197 | 5.382(6) | 5.393(4) | 5.389(5) | 5.391(1) | | 5.391(1) |
| Be | 9.3227 | 9.043(6) | 9.326(6) | 9.322(5) | 9.050(2) | 9.320(1) | 9.314(2) |
| B | 8.298 | 8.467(6) | 8.270(6) | 8.282(6) | 8.449(2) | 8.249(2) | 8.137(4) |
| C | 11.2603 | 11.418(6) | 11.264(6) | 11.248(6) | 11.410(6) | 11.203(2) | 11.167(4) |
| N | 14.5551 | 14.698(6) | 14.551(6) | 14.551(6) | 14.713(3) | 14.527(2) | 14.499(4) |
| O | 13.6181 | 13.592(6) | 13.601(6) | 13.579(6) | 13.620(10) | | 13.589(4) |
| F | 17.446 | 17.450(6) | 17.437(6) | 17.431(6) | 17.432(5) | | 17.413(3) |
| Ne | 21.6239 | 21.656(6) | 21.609(6) | 21.618(6) | 21.660(3) | | 21.658(7) |
| Na | 5.1391 | 5.124(6) | 5.199(6) | 5.147(6) | 5.159(8) | | 5.112(2) |
| Mg | 7.6368 | 7.495(6) | 7.605(6) | 7.573(6) | 7.510(20) | 7.620(10) | 7.589(4) |
| Al | 5.9858 | 5.964(6) | 5.975(6) | 5.875(6) | 5.880(20) | 5.930(20) | 5.784(5) |
| Si | 8.169 | 8.156(6) | 8.171(6) | 8.094(6) | 8.210(20) | 8.080(30) | 8.075(7) |
| P | 10.5379 | 10.550(6) | 10.611(6) | 10.360(6) | 10.540(30) | 10.460(10) | 10.488(9) |
| S | 10.36 | 10.280(6) | 10.320(6) | 10.095(6) | 10.320(30) | | 10.241(11) |
| Cl | 12.9939 | 12.942(6) | 12.901(6) | 12.762(6) | 12.910(20) | | 12.934(6) |
| Ar | 15.8407 | 15.818(6) | 5.827(8) | 15.660(8) | 15.830(30) | | 15.806(14) |
| MAD | | 77(6) | 30(6) | 80(6) | 80(13) | 54(10) | 70(5) |

[a] Reference 7

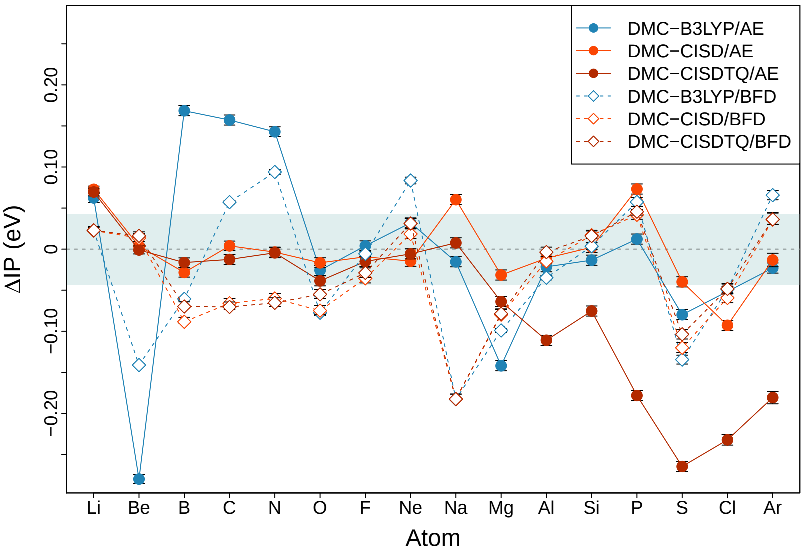
[b] Reference 1

Table 5.2: All-electron (AE) electron affinities (eV) for the atoms Li through Ar calculated with DMC and different trial wavefunctions. Mean absolute deviation (MAD) is reported in meV. Uncertainty in last digit of DMC results is shown in parentheses.

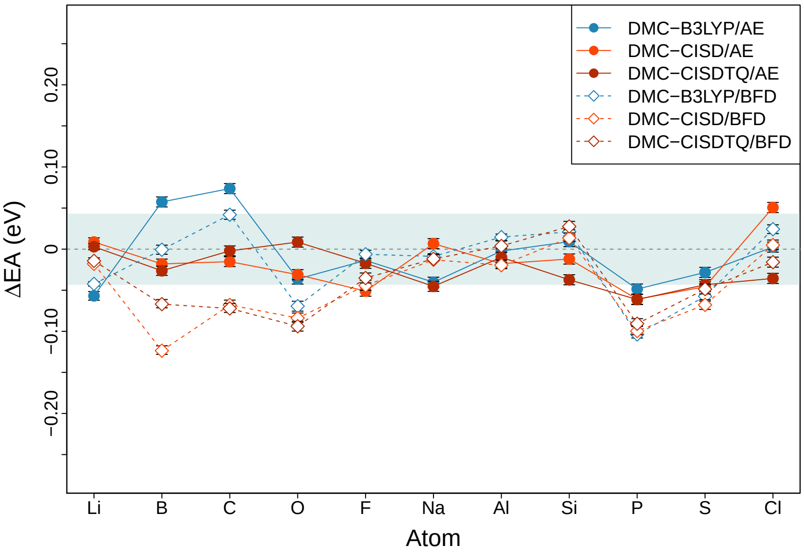
| | Ref.[a] | B3LYP/AE | CISD/AE | CISDTQ/AE | POEP-SC[a] | POEP-MC[a] | POEP-2014[b] |
|---|---|---|---|---|---|---|---|
| Li | 0.618 | 0.561(5) | 0.627(5) | 0.621(4) | 0.559(2) | 0.619(1) | 0.593(1) |
| B | 0.2797 | 0.337(6) | 0.262(6) | 0.253(6) | 0.340(2) | 0.158(3) | 0.157(4) |
| C | 1.2621 | 1.336(6) | 1.247(6) | 1.260(6) | 1.342(6) | 1.161(2) | 1.206(6) |
| O | 1.4611 | 1.425(6) | 1.430(6) | 1.470(6) | 1.370(20) | | 1.399(4) |
| F | 3.4325 | 3.419(6) | 3.381(6) | 3.415(6) | 3.445(8) | | 3.452(3) |
| Na | 0.5479 | 0.508(6) | 0.555(6) | 0.503(6) | 0.480(10) | 0.570(10) | 0.558(3) |
| Al | 0.4414 | 0.439(6) | 0.424(6) | 0.431(6) | 0.500(30) | 0.380(20) | 0.447(7) |
| Si | 1.4155 | 1.425(6) | 1.403(6) | 1.378(6) | 1.400(30) | 1.340(30) | 1.421(5) |
| P | 0.7465 | 0.698(6) | 0.686(6) | 0.685(6) | 0.690(30) | | 0.717(12) |
| S | 2.096 | 2.068(6) | 2.050(6) | 2.053(6) | 2.050(40) | | 2.030(10) |
| Cl | 3.667 | 3.669(6) | 3.718(6) | 3.631(6) | 3.760(20) | | 3.696(5) |
| MAD | | 34(6) | 29(6) | 26(6) | 58(18) | 64(11) | 39(5) |

[a] Reference 7

[b] Reference 1



(a) Ionisation potential



(b) Electron affinity

Figure 5.1: Deviation from experimental ionisation potentials, ΔIP, (upper plot) and electron affinities, ΔEA, (bottom plot) for the atoms Li through Ar obtained with single- and multi-determinant DMC with all-electron (AE) basis sets or the BFD pseudopotential. The shaded region denotes chemical accuracy. Experimental reference values taken from Ref. 7. Lines are drawn between points to guide the eye.

with orbital degeneracies.

From Figure 5.1 it can be seen that the largest errors for ionisation potentials come from Be, B, C and N. The errors for second-row atoms are much smaller. The single-particle energies of the $2s$ and $2p$ orbitals of the atoms Be, B and C and the ions Li$^-$, B$^+$, C$^+$, B$^-$ and N$^+$ are relatively close and the $2s^2$ pair can easily be promoted to a $2p^2$ pair. This is known as the $2s2p$ near degeneracy. For second row atoms the $3s$, $3p$ and $3d$ subshells are sufficiently close in energy and single or pair excitations are important. These near-degeneracies affect Mg, Al and Si as well as the ions Na$^-$, Al$^+$, Al$^-$, Si$^+$, and P$^+$ but the effect of these degenerate orbitals less important compared to first-row atoms[7] and the errors for second-row atoms are smaller. These degeneracies are also less pronounced for anions compared to cations and this is reflected in the smaller errors for electron affinities for both first- and second-row atoms relative to the ionisation potentials.

DMC-B3LYP/AE results are compared with the most comprehensive set of single-determinant DMC results in the literature in Tables 5.1 and 5.2. The DMC-B3LYP/AE ionisation potentials of Na, Al and Si and the electron affinities of O, Al and Cl have smaller errors but overall both methods are in agreement. The quality of the nodes of the trial wavefunction is the biggest limitation for the accuracy of DMC under the fixed-node approximation. In Ref. 7 the DMC-SC trial wavefunctions were obtained within the parameterised optimised effective potential (POEP) approximation where the potential is parameterised as well as the radial parts of the orbital. POEP orbitals are a good approximation to Hartree-Fock theory but Kohn-Sham orbitals are known to give better nodal surfaces for DMC.[40]

DMC-B3LYP/AE results presented here are in good agreement with experimental reference values. They come close to chemical accuracy, defined as an error less than 43 meV, but the orbital degeneracies of some atoms and cations resulted in significant errors for these species. The inclusion of additional determinants in the trial wavefunction can recover some of the correlation energy associated with these degeneracies and decrease these errors.

## 5.4 Multi-determinant DMC

Correlation energy is defined as the difference between the energy in the Hartree-Fock limit and the exact (nonrelativistic) energy of a system. This correlation energy arises from the interaction of electrons (dynamic correlation) and orbital degeneracies (static correlation). Single-determinant DMC recovers a significant portion of the correlation energy but including more determinants in the trial wavefunction can systematically improve the nodal surface and recover the remaining correlation energy. This is especially important for systems with known degeneracies but it can also improve the nodal surface of systems that are predominantly single-reference in character. The effect of these additional configurations on the total energy can be described by the nodal correlation energy, defined here as the difference between the single-determinant DMC-B3LYP and multi-determinant DMC-CI energies:

$$\Delta E_{\text{corr}} = E_{\text{DMC-CI}} - E_{\text{DMC-B3LYP}}$$

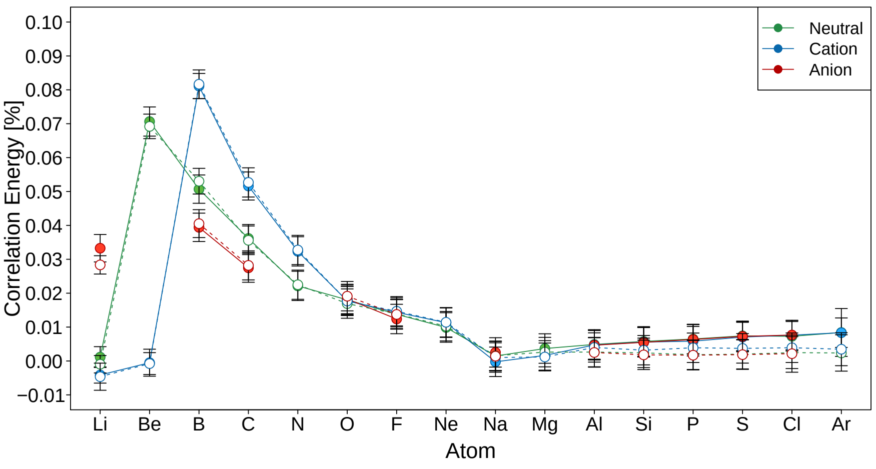
Multi-determinant wavefunctions were built using configuration interaction (CI) with single and double excitations (CISD) and CI with single, double, triple and quadruple excitations (CISDTQ) from a B3LYP reference wavefunction. The active space used here for all-electron CISD and first-row CISDTQ calculations is the entire orbital space as defined by the Roos ATZ basis set[34] (46 levels per atom). All-electron CISDTQ calculations with the complete active space were not feasible for second-row atoms and the active space was restricted to 13 levels per atom. The pseudopotential calculations (DMC-CISD/BFD and DMC-CISDTQ/BFD) used all levels in the active space as defined by the BFD triple-zeta basis set associated with the pseudopotential. For Li, Be, Na and Mg this was 15 levels. For all other atoms it was 29 levels.

CISD and CISDTQ all-electron ionisation potentials and electron affinities for first- and second-row atoms are shown in Tables 5.1 and 5.2 and deviations from experimental reference values are included in Figure 5.1. Relative nodal correlation energies are shown in Figure 5.2a.

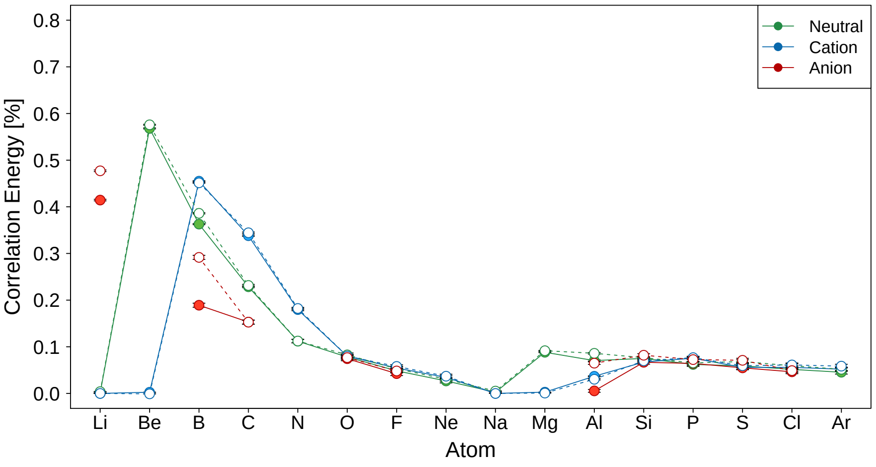
For the atomic calculation performed here, the DMC wavefunctions built with CISD and CISDTQ determinants recover the same amount of correlation energy when the same active space is used, as seen in Figure 5.2a. Using all virtual orbitals in the active space was not computationally feasible for second-row CISDTQ calculations and a restricted active space with fewer levels was used instead. This smaller active space recovers less correlation energy compared to CISD with a complete active space. These results indicate triple and quadruple excitations have little contribution to the nodal surface but single and double excitations into higher virtual orbitals are important. By definition, no correlation energy is recovered for one electron systems when using pseudopotentials.

DMC-CISD/AE calculations reduced the MAD for ionisation potentials to just 30 ± 6 meV, compared to 77 ± 6 meV for single-determinant DMC. The MAD for electron affinities was reduced to 29 ± 6 meV. Multi-determinant wavefunctions recovered a significant amount of correlation energy for near-degenerate systems, namely Be, B, C and N, and this improved the quality of the ionisation potentials and electron affinities of these atoms. While there is an overall reduction in the error, the multi-determinant DMC-CISD/AE ionisation potentials for Na, P and Cl and the electron affinity for Cl are worse than the single-determinant result. The Na and P neutral atoms recover more nodal correlation energy relative to the DMC-B3LYP/AE energy compared to the cations but the Cl atom recovers less than the charged species. The accuracy of the final energy difference is ultimately governed by the convergence of the energy with respect to the multi-determinant expansion. A balanced description of correlation cannot be guaranteed for all species, even when a complete active space is used. When one species converges faster it can lead to results worse than the single-determinant.

The DMC-CISD/AE and DMC-CISDTQ/AE ionisation potentials and electron affini-



(a) All-electron



(b) BFD pseudopotential

Figure 5.2: DMC nodal correlation energy as a fraction of the total DMC-B3LYP energy for CISD (closed circles) and CISDTQ (open circles) trial wavefunctions.

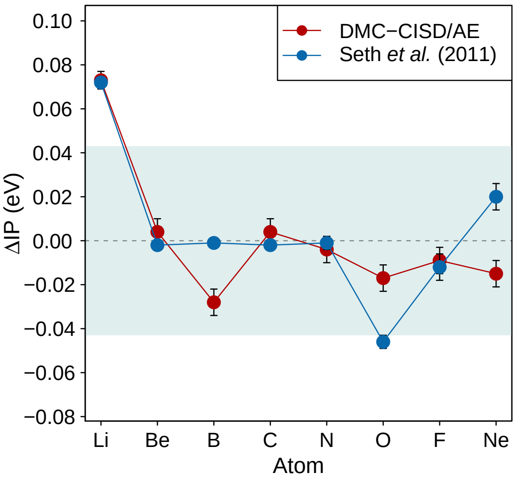


Figure 5.3: Deviation from experimental ionisation potentials (ΔIP) for the first row atoms obtained with DMC-CISD/AE compared to DMC-CISD/HF results taken from Ref. 3. The shaded region denotes chemical accuracy.

ties are in good agreement when the same active space is used but the DMC-CISDTQ/AE calculations use a smaller active space for second-row atoms. DMC-CISDTQ/AE ionisation potentials show much larger deviations from experimental reference values and the MAD is 80 ± 6 meV. Surprisingly the DMC-CISDTQ/AE electron affinities agree closely with CISD results, despite the CISDTQ wavefunctions recovering less correlation energy. The convergence of the CISDTQ wavefunction is slower for neutral atoms and anions and less correlation energy is recovered compared to cations. The errors in this slower convergence cancel out for electron affinities.

The most comprehensive set of multi-determinant values from the literature for both first- and second-row atoms are included in Table 5.1 and 5.2. The DMC-MC[7] results are only available for some atoms and the configurations were selected based on knowledge of degenerate orbitals. Similarly, the DMC-POEP[1] results use either a single or two-configuration wavefunction depending on the atom. Both methods use POEP orbitals.

A separate study used a CISD trial wavefunction from a HF reference but only calculated first-row ionisation potentials.[3] These results are compared in Figure 5.3. The errors for DMC-CISD/AE ionisation potentials calculated in this work are smaller for O but larger for B. The errors are roughly the same magnitude but have opposite sign for Ne and the overall error for both methods is 19meV. Although both methods used CI wavefunctions with single and double excitations (CISD) Seth et al. used the atomic multi-configuration Hartree-Fock (MCHF) package ATSP2K[41] and excitations were determined by principal quantum numbers and angular momentum numbers. The DMC-CISD/AE results presented here are consistent with previous work but orbitals were generated with a method that is transferrable to molecules.

Table 5.3: Mean absolute deviation (MAD) of DMC with different trial wavefunctions for ionisation potentials and electron affinities of first- and second-row atoms in meV. Uncertainty in last digit of DMC results is shown in parentheses.

| | IP | EA |
|---|---|---|
| DMC-B3LYP/AE | 77(6) | 34(6) |
| DMC-CISD/AE | 30(6) | 29(6) |
| DMC-CISDTQ/AE | 80(6) | 26(6) |
| DMC-B3LYP/BFD | 73(3) | 35(5) |
| DMC-CISD/BFD | 58(3) | 51(4) |
| DMC-CISDTQ/BFD | 55(3) | 44(4) |

The DMC-MC wavefunctions from Ref. 7 and the DMC-POEP wavefunctions from Ref 1 used a small set of configurations chosen based on *a priori* knowledge of degenerate orbitals. A more 'black-box' style approach was used in the present work where all possible single and double or single, double, triple and quadruple excitations were included in the CI calculation. When the DMC-CISD/AE and DMC-MC results are compared for the atoms Be through N and Al through P, DMC-CISD/AE performs better. The MAD for ionisation potentials is 21 ± 6 meV and just 13 ± 6 meV for electron affinities. As shown in Figure 5.2a, including more virtual orbitals in the active space recovers more correlation energy. The single and double excitations into higher virtual orbitals have a significant contribution to the quality of the nodal surface for DMC calculations. Furthermore, nodal correlation energy can be recovered for all species, including those that are predominantly single-reference. By using a more extensive wavefunction for all species the errors for DMC-CISD/AE are smaller than previous multi-determinant studies.

## 5.5 BFD pseudopotential

DMC methods scale as $Z^{5.5-6.5}$ with respect to atomic number $Z$ and the high-energy core electrons are routinely replaced by an effective potential to reduce the computational cost. The effect of these pseudopotentials on the correlation energy of neutral and charged first- and second-row atoms is investigated here. Deviations from experimental reference values are shown in Figure 5.1. The mean absolute deviations (MADs) for pseudopotential results are reported in Table 5.3.

The MAD for DMC-B3LYP/BFD ionisation potentials and electron affinities is equivalent to all-electron results. Figure 5.1 shows the errors are generally smaller for first-row atoms compared to all-electron results but this improvement is cancelled out by larger errors for second-row atoms. A previous study has shown core relaxation is important for first row atoms and freezing the doubly occupied $1s$ introduced substantial errors for CCSD(T) and FCIQMC first-row ionisation potentials.[6] Although the study only looked at first-row atoms it is likely removing the neon core for second-row atoms is having a similar, detrimental effect. The quality of the ionisation potential and electron affinity is improved for the degenerate species (Be, B, C and N) but worse for all other atoms. This suggests the errors introduced by removing the core electrons are negated by the errors associated with the degenerate orbitals.

The error of the DMC-B3LYP/BFD ionisation potential for Na is especially large compared to other atoms. The BFD pseudopotential replaces the $1s$, $2s$ and $2p$ electrons of second-row atoms with an angular-momentum dependent effective potential that acts on the valence electrons. This means the $Na^+$ cation is the pseudopotential core and the ionisation potential is the energy of the single-particle orbital. The Li ionisation potential is defined in a similar manner but the error associated with it is much smaller. The BFD pseudopotential is constructed to reproduce Dirac-Fock energies but this neglects correlation energy. The large error for the Na ionisation potential comes from the core correlation energy that is being ignored. For other atoms the error introduced by neglecting the correlation energy cancels out and reasonable values are obtained.

The DMC-CISD/BFD and the DMC-CIDSTQ/BFD results are included in Figure 5.1. Overall mean absolute deviations (MADs) are reported in Table 5.3. Electron correlation energy is shown in Figure 5.2b. Compared to all-electron results the BFD multi-determinant wavefunctions recover more correlation relative to the total energy of the atoms. For most systems the DMC-CISD/BFD and DMC-CISDTQ/BFD wavefunctions recover the same amount of correlation energy. The exception is the Li, B and Al anions where DMC-CISDTQ/BFD recovers more correlation energy than DMC-CISD/BFD. This is most likely a side effect of the truncation of determinants rather than a contribution from the additional excitations. It is not practical to include all determinants from a CISD (or CISDTQ) calculation in a DMC trial wavefunction and instead the number of determinants is truncated according to some threshold. An energy truncation scheme was used here[32] where the number of determinants was selected based on their contribution to the CI energy of the atom. Despite using the same truncation value the multi-determinant pseudopotential wavefunctions recovered significantly more correlation energy compared to the all-electron wavefunctions. The individual determinants have a bigger contribution to the final energy of the molecule and the BFD wavefunctions are more sensitive to the truncation.

The MAD for ionisation potentials decreases from 73 ±3 meV for single determinant DMC-B3LYP/BFD to 58 ± 3 meV for DMC-CISD/BFD (55 ± 3 for DMC-CISDTQ/DMC). The difference in the correlation energy recovered for the anions discussed above results in smaller errors for the DMC-CISDTQ/BFD electron affinities compared to the DMC-CISD/BFD results (44 ± 4 meV compared to 51 ± 4 meV) but in both cases the errors are larger than single-determinant DMC-B3LYP/BFD. Despite CISD (and CISDTQ) wavefunctions recovering a significant amount of nodal correlation energy the relative correlation energy for anions is worse compared to single-determinant results.

## 5.6 Summary

Overall, DMC performs well for the ionisation potentials and electron affinities of first- and second-row atoms. The quality of single-determinant DMC-B3LYP/AE electron affinities was better than ionisation potentials but including more determinants in the trial wavefunction improved the results. The DMC-CISD/AE results presented here are in good agreement with experiment and chemically accurate results have been obtained for the atoms Li through Ar. Including more virtual orbitals in the active space with just single and double excitations recovered a significant amount of nodal correlation energy. Including triple and quadruple excitations in the trial wavefunction had little effect on the nodal correlation energy recovered and the final energy differences. BFD pseudopotentials offer a good compromise between cost and accuracy and a significant amount of nodal correlation energy can be recovered by incorporating more determinants in the active space. In general, including more configurations in the trial wavefunction will recover more nodal correlation energy and lower the total energy of a system but for this to translate to accurate energy differences it needs to be balanced on all species.

## 5.7 References

[1] E. Buendía, F. J. Gálvez, P. Maldonado, A. Sarsa, *Chem. Phys. Lett.* **2014**, *615*, 21–25.

[2] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **2012**, *8*, 2181–2188.

[3] P. Seth, P. L. Ríos, R. J. Needs, *J. Chem. Phys.* **2011**, *134*, 084105.

[4] D. M. Cleland, G. H. Booth, A. Alavi, *J. Chem. Phys.* **2011**, *134*, 024112.

[5] Y. Nakatsuka, T. Nakajima, M. Nakata, K. Hirao, *J. Chem. Phys.* **2010**, *132*, 054102.

[6] G. H. Booth, A. Alavi, *J. Chem. Phys.* **2010**, *132*, 174104.

[7] P. Maldonado, A. Sarsa, E. Buenda, F. J. Glvez, *J. Chem. Phys.* **2010**, *133*, 064102.

[8] E. Buendía, F. J. Gálvez, P. Maldonado, A. Sarsa, *J. Chem. Phys.* **2009**, *131*, 044115.

[9] J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2008**, *128*, 174101.

[10] M. D. Brown, J. R. Trail, P. L. Ríos, R. J. Needs, *J. Chem. Phys.* **2007**, *126*, 224110.

[11] K. Hongo, Y. Kawazoe, H. Yasuhara, *Mater. Trans.* **2006**, *47*, 2612–2616.

[12] M. Casula, S. Sorella, *J. Chem. Phys.* **2003**, *119*, 6500–6511.

[13] F. J. Gálvez, E. Buenda, A. Sarsa, *Int. J. Quant. Chem.* **2002**, *87*, 270–274.

[14] J. W. Moskowitz, K. E. Schmidt, *J. Chem. Phys.* **1992**, *97*, 3382–3385.

[15] R. N. Barnett, P. J. Reynolds, W. A. Lester Jr., *J. Chem. Phys.* **1986**, *84*, 4992–4996.

[16] W. A. Al-Saidi, *J. Chem. Phys.* **2008**, *129*, 064316.

[17] E. Buendía, F. J. Gálvez, A. Sarsa, *J. Chem. Phys.* **2006**, *124*, 154101.

[18] M. Caffarel, J.-P. Daudey, J.-L. Heully, A. Ramárez-Solás, *J. Chem. Phys.* **2005**, *123*, 094102.

[19] M. C. Per, *CSIRO Quantum Monte Carlo software package*, **2017**.

[20] S. Dressler, W. Thiel, *Chem. Phys. Lett.* **1997**, *273*, 71–78.

[21] J. M. Martin, C. W. Bauschlicher, A. Ricca, *Comput. Phys. Comm.* **2001**, *133*, 189–201.

[22] B. N. Papas, H. F. Schaefer III, *THEOCHEM* **2006**, *768*, 175 – 181.

[23] S. E. Wheeler, K. Houk, *J. Chem. Theor. Comput.* **2010**, *6*, 395–404.

[24] D. Ceperley, *J. Stat. Phys.* **1986**, *43*, 815–826.

[25] A. Ma, N. D. Drummond, M. D. Towler, R. J. Needs, *Phys. Rev. E* **2005**, *71*, 066704.

[26] W. Foulkes, L. Mitás, R. J. Needs, G. Rajagopal, *Rev. Mod. Phys.* **2001**, *73*, 33–83.

[27] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[28] E. T. Swann, M. L. Coote, A. S. Barnard, M. C. Per, *Int. J. Quantum Chem.* **2017**, e25361.

[29] X. Zhou, F. Wang, *J. Comput. Chem.* **2017**, *38*, 798–806.

[30] J. R. Trail, R. J. Needs, *J. Chem. Phys.* **2013**, *139*, 014101.

[31] R. C. Clay III, M. A. Morales, *J. Chem. Phys.* **2015**, *142*, 234103.

[32] M. C. Per, D. M. Cleland, *J. Chem. Phys.* **2017**, *146*, 164101.

[33] J. Toulouse, R. Assaraf, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 244112.

[34] P.-O. Widmark, P.-A. Malmqvist, B. O. Roos, *Theor. Chim. Acta* **1990**, *77*, 291–306.

[35] M. C. Per, S. P. Russo, I. K. Snook, *J. Chem. Phys.* **2008**, *128*, 114106.

[36] M. Dolg, C. Filippi, *Private Communication* **2014**.

[37] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[38] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[39] M. Gordon, M. Schmidt in *Theory and Applications of Computational Chemistry: the first forty years*, C. Dykstra, G. Frenking, K. Kim, G. Scuseria (Eds.), Elsevier, Amsterdam, **2005**, pp. 1167–1189.

[40] M. C. Per, K. Walker, S. P. Russo, *J. Chem. Theory Comput.* **2012**, *8*, 2255–2259.

[41] C. F. Fischer, G. Tachiev, G. Gaigalas, M. R. Godefroid, *Comput. Phys. Comm.* **2007**, *176*, 559–579.

Chapter 6

# Difficult reactions

## 6.1 Introduction

The previous chapters have demonstrated the accuracy and reliability of DMC for a variety of systems. Single-determinant DMC was sufficient for chemically accurate results for most molecules and a multi-determinant wavefunction improved results for more challenging systems. The largest source of error in a DMC calculation comes from the fixed-node approximation but errors in the nodal surface usually cancel out for energy differences. The systems studied thus far have been limited to atoms and small or organic systems. Benchmarking results in the literature are also restricted to similar, simple electronic structures.[1,2] A more thorough investigation of the performance of DMC is presented in this chapter using a set of eighteen reactions that are particularly challenging for density functional theory (DFT).

DFT methods are a popular choice for low-cost electronic structure calculations. They have been broadly applied to a range of problems including transition metals,[3] molecular properties and spectroscopy for inorganic compounds,[4] complex chemical and biological systems[5] and hydrogen bonding in water complexes[6] to name just a few. Despite the successes in these area, there are severe limitations to the method. DFT suffers from self-interaction and delocalisation errors.[7–10] Poor results can be expected in these cases but it can also fail unexpectedly and catastrophically for simple systems like unsaturated hydrocarbons.[11] Most DFT methods fail to describe dispersion effects unless specifically parameterised for them.[12–14] As seen in Chapter 3 small, seemingly simple problems like hydrogen abstraction from methanol require high-level and costly methods for accurate results.[15–17]

Wavefunction theory (WFT) based methods, like the 'gold-standard' CCSD(T) do not suffer the same errors as DFT but become prohibitively expensive as system size increases. DMC is also unaffected by the delocalisation and self-interaction errors but scales as approximately $N^{3-4}$ with respect to system size, $N$. The Slater-Jastrow form of the trial wavefunction provides a good description of correlation effects and DMC is more suitable for dispersion problems.[6,18–22] Despite using Kohn-Sham orbitals in the trial wavefunction it has been shown to deliver chemically accurate results for the types of systems that DFT struggles with, such as the H-abstraction of methanol and Diels-Alder reactions as shown. The benchmarking of single- and multi-determinant DMC is

Table 6.1: Reference values (kcal/mol) for the difficult cases (DC18) test set. Reactions 1-9 are taken from Ref. 23, reactions 10-18 are taken from Ref. 14.

| | Reaction | Ref. Value |
|---|---|---|
| 1 | HCN . . . $BF_3 \rightarrow$ HCN + $BF_3$ | 5.7 |
| 2 | $C_6Cl_6$ + 6 HCl $\rightarrow$ 6 $Cl_2$ + $C_6H_6$ | 148.3 |
| 3 | $P_4 \rightarrow$ 4 P | 289.9 |
| 4 | $SF_6 \rightarrow$ S + 6 F | 477.5 |
| 5 | $PF_5 \rightarrow$ P + 5 F | 556.4 |
| 6 | $P_4O_{10} \rightarrow P_4$ + 5 $O_2$ | 719.7 |
| 7 | $C_6F_6 \rightarrow$ 6 C + 6 F | 1388.1 |
| 8 | $Si(OCH_3)_4 \rightarrow$ Si + 4 C + 4 O + 12 H | 2023.5 |
| 9 | urotropin $\rightarrow$ 6 C + 4 N + 12 H | 2151.1 |
| 10 | 2-pyridone $\rightarrow$ 2-hydroxypyridine | -1.0 |
| 11 | $(C_{20})_{cage} \rightarrow (C_{20})_{bowl}$ | -13.3 |
| 12 | hepta−1,2,3,5,6−hexaene $\rightarrow$ hepta−1,3,5−triyne | 14.3 |
| 13 | 2 tetramethyl−ethen $\rightarrow$ octamethylcylobutane | -19.2 |
| 14 | $CH_{12}$ isomerisation | -25.0[a] |
| 15 | carbo-[3]-oxacarbon isomerisation | -26.9 |
| 16 | $N_2CH_2$ + $C_2H_4 \rightarrow (CH_2)_3N_2$ | -38.1 |
| 17 | 4 Be $\rightarrow Be_4$ | -90.4[b] |
| 18 | 4 $S_2 \rightarrow S_8$ | -101.0 |

[a] Reference 24

[b] Reference 13

extended in this chapter using a test set of 18 difficult cases for DFT methods (DC18), shown in Table 6.1. Reactions 1 to 9 are taken from work by Truhlar et. al[23] (DC9T) and include non-hydrogen hypervalent compounds and more second-row and halogen atoms than previous test sets. Reactions 10 to 18 are taken from work by Grimme et. al.[14] (DC9G) and include larger organic molecules and isomerisation energies.

## 6.2 Method

Geometries and reference values were taken from previous work.[14,23] The reference reaction energies reported in Ref. 14 were taken from previous papers and were incorrect for reactions 14 and 17. The reference energies listed in in Table 6.1 were updated with values from Ref. 24 and Ref. 13 respectively.

DMC calculations were performed using target population sizes of 6400 walkers and an imaginary-time step of 0.01 a.u. Core electrons were replaced with the energy-consistent Burkatzki-Filippi-Dolg (BFD) potentials and associated associated triple-zeta (VTZ) basis sets,[25] with an improved basis and potential for H.[26] The size-consistent T-moves scheme was used for these non-local pseudopotentials.[27] The *eeN* term was omitted from the Jastrow factor.[15]

Single-determinant (SD) DMC calculations used B3LYP orbitals in the trial wavefunction. The CSFs for multi-determinant (MD) DMC were obtained from configuration

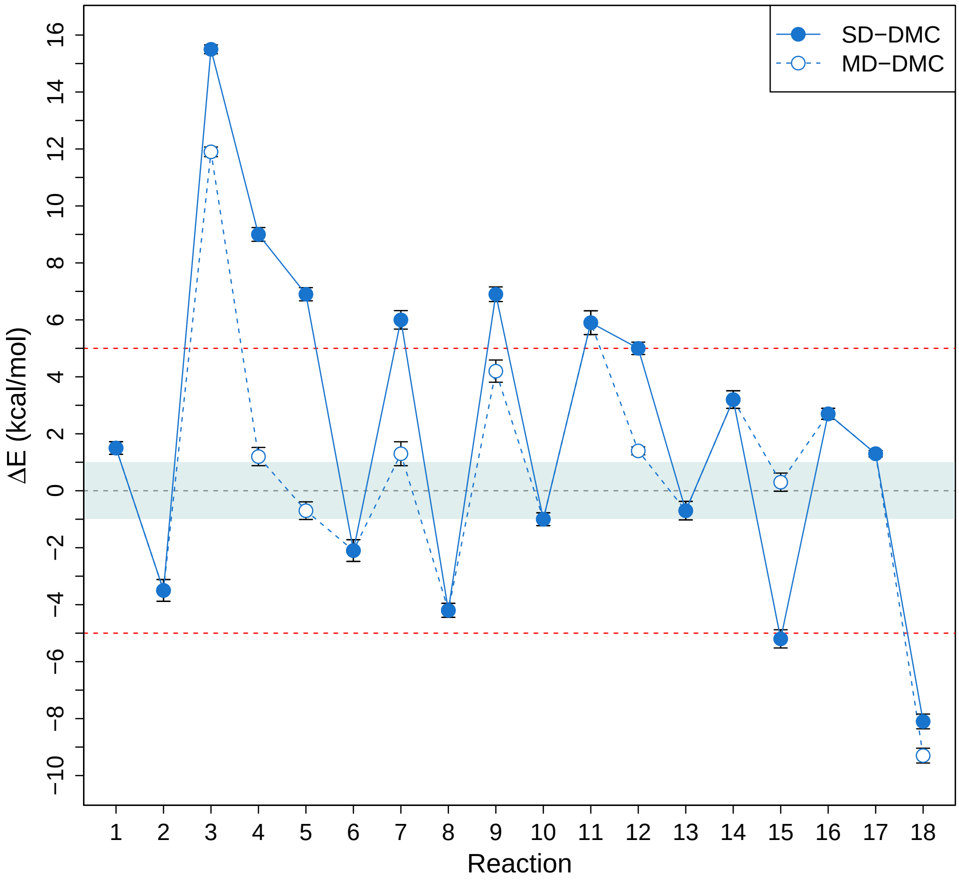


Figure 6.1: Deviations from reference values (ΔE) for the DC18 test set. The shaded region represents chemical accuracy, ± 1 kcal/mol. The red dashed line denotes errors of 5 kcal/mol. Reactions with errors greater than 5 kcal/mol were selected for a multi-determinant study. The lines between points are merely drawn to guide the eye.

interaction with single and double excitations (CISD) from a B3LYP reference wavefunction. The use of Kohn-Sham orbitals in CI wavefunctions have been shown to give better DMC nodal surfaces than Hartree-Fock orbitals.[28] CSFs were selected using the recently developed energy truncation method.[29] The variable parameters in the Jastrow factor and the CSF coefficients were simultaneously optimised by minimising the variational energy using an approach based on the linear method.[30] B3LYP and CI calculations were performed with GAMESS.[31,32] Quantum Monte Carlo calculations were performed using the CSIRO Quantum Monte Carlo software package.[33]

## 6.3 Single-determinant DMC

Single-determinant DMC deviations from reference values ($\Delta E$) are shown in Figure 6.1. Overall, the quality of results is much lower than previous test sets. Only four of the eighteen reactions come close to the 'chemical accuracy' standard of errors less than 1 kcal/mol. Reactions 1, 13 and 17 are all examples of systems with significant dispersion forces.[12–14] Most DFT methods fail to account for dispersion forces but DMC describes these effects with high accuracy[6, 18–22] and the small DMC errors for these reactions are unsurprising. Reaction 10 is the isomerisation of DNA base tautomers (2-pyridone → 2-hydroxypyridine). The aromatic structure of 2-hydroxypyridine is challenging for DFT methods and in most cases they predict the wrong energy ordering of isomers.[34] The reference isomerisation energy is -1.0 kcal/mol and DMC predicts no energy difference between the isomers.

Despite the good results for these four reactions the overall performance of single-determinant DMC for the DC18 test set is poor. Half of the reactions have errors greater than 5 kcal/mol and reactions 3, 4 and 18 have errors greater than 8 kcal/mol. The mean absolute deviation (MAD) is 4.9 kcal/mol for the entire set. As shown previously, DMC can be improved by incorporating more determinants in the trial wavefunction. Systems with errors greater than 5 kcal/mol were treated with multi-determinant methods with the exception of reaction 11.

Reaction 11, the relative energy of $C_{20}$ cage and bowl isomers, is a challenging problem for many methods. $C_{20}$ has three distinct low-lying isomers: a fullerene cage, a monocyclic ring and a bowl structure.[35] The relative energy of these isomers is extremely sensitive to the method used and most methods disagree on the magnitudes of the energies and even the relative ordering of isomers.[36–42] The reference value used here (13.3 kcal/mol) was calculated at an estimated CCSD(T)/CBS level[43] using MP2/TZV2d1f geometries.[39] Single-determinant DMC predicted a much larger energy difference of 19.2 ± 0.4 kcal/mol based on these MP2/TZV2d1f geometries. A recent study using a similar extrapolation scheme with PBE0/cc-pVTZ geometries found the relative energy between the bowl and cage isomers was approximately 8 kcal/mol.[42] Due to differences in geometries the DMC energy can't be compared with this new value. Part of the problem is thought to arise from partial multi-reference effects[44] and it is likely additional determinants could improve the energy, but given the system size and the unreliability of the reference data it was not included in the subsequent multi-determinant study.

## 6.4 Multi-determinant DMC

### 6.4.1 Active space selection

The choice of active space used for multi-determinant wavefunction expansions determines the accuracy of the final energy. Full configuration interaction incorporates all excitations in all levels. It is the most accurate method and will give the exact energy for the basis set limit but is prohibitively expensive for most systems. Instead, the active space and level of excitations are restricted but often there is no *a priori* way of knowing which orbitals or excitations are the most relevant. The active space needs to include the important bonding orbitals as well as sufficient virtual levels to capture the necessary excitations. In some cases this is obvious but most often it is selected intuitively or a guess is made at what the most relevant orbitals or excitations should be. For the ionisation potentials and electron affinities of first- and second-row atoms in Chapter 5, including triple and quadruple excitations had little effect on the final energy but higher energy virtual orbitals were extremely important. Selecting an appropriate active space can be complicated for reactions where number of electrons and orbitals needs to be balanced on both sides of a reaction equation. This becomes especially challenging for decomposition reactions where there is a large disparity in molecule sizes on either side of the equation. The active space needs to be sufficient on atoms to capture the important configurations but small enough for larger molecules to be computationally feasible. The DC18 set includes a wide range of molecules and reactions and the following guidelines were used to provide consistency. Core electrons and shells were replaced with a pseudopotential, equivalent to a helium core for first-row atoms and a neon core for second-row atoms.

- For first row atoms, all valence electrons and sufficient virtual levels to close the n=2 shell (i.e. $2s$ and $2p$ orbitals, 4 levels per atom)
- For second row atoms, all valence electrons and sufficient virtual levels to close the n=3 shell were included (i.e $3s$, $3p$ and $3d$ orbitals, 9 levels per atom)
- One level was included for each H atom in the molecule

The notation RAS($N_e, N_{\text{orb}}$) is used, denoting restricted active space (only single and double excitations were included) with $N_e$ electrons and $N_{\text{orb}}$ active orbitals. These methods produce too many determinants to be practical for DMC calculations and instead the expansion is truncated according to some threshold. In this work the number of determinants was selected using an energy-truncation scheme.[29] The error cancellation in multi-determinant expansions for reactions can be maximised by using different $E_{\text{trunc}}$ values, using the formula:

$$E^{P}_{\text{trunc}} = E^{R1}_{\text{trunc}} + E^{R2}_{\text{trunc}} \tag{6.1}$$

where $E^{R1}_{\text{trunc}}$ and $E^{R2}_{\text{trunc}}$ are the energy cut-offs of reactants and $E^{P}_{\text{trunc}}$ is the energy-cut off of the products.[29] The value of $E_{\text{trunc}}$ used was dependent on each reaction; $E_{\text{trunc}} = 0.01$ Ha or the smallest value possible to keep the number of CSFs below 1200 for the largest molecule in the reaction. A summary of the active space, $E_{\text{trunc}}$ and final number of CSFs used in each calculation is provided in Table 6.2.

### 6.4.2 Non-hydrogen hypervalent compounds

Non-hydrogen, hypervalent molecules are challenging for both DFT and WFT methods.[11,13,45] There are four decomposition energies of these types of molecules in DC18.

Table 6.2: Details of the final active spaces of $N_e$ electrons in $N_{\mathrm{orb}}$ active orbitals used for each molecule. Configuration space functions (CSFs) are symmetry-adapted linear combinations of the determinants (Dets).

| RXN | Mol | $E_{\mathrm{trunc}}$ (Ha) | $N_{\mathrm{e}}$ | $N_{\mathrm{orb}}$ | Dets | CSFs |
|---|---|---|---|---|---|---|
| 3 | $P_4$ | 0.055 | 20 | 36 | 4630 | 1131 |
| | P | 0.01375 | 5 | 9 | 13 | 6 |
| 4 | $SF_6$ | 0.05 | 28 | 47 | 4286 | 951 |
| | S | 0.05 | 4 | 8 | 1 | 1 |
| | F | 0.05 | 5 | 8 | 1 | 1 |
| 5 | $PF_5$ | 0.07 | 28 | 48 | 5128 | 1140 |
| | P | 0.01167 | 3 | 8 | 1 | 1 |
| | F | 0.01167 | 5 | 8 | 29 | 10 |
| 7 | $C_6F_6$ | 0.01 | 18 | 24 | 415 | 100 |
| | C | 0.01 | 2 | 3 | 1 | 1 |
| | F | 0.01 | 1 | 1 | 1 | 1 |
| 9 | urotropin | 0.04 | 56 | 52 | 2424 | 698 |
| | N | 0.04 | 5 | 4 | 1 | 1 |
| | C | 0.04 | 4 | 4 | 1 | 1 |
| 12 | hepta-1,2,3,5,6-hexaene | 0.01 | 32 | 32 | 4489 | 1030 |
| | hepta-1,3,5-triyne | 0.01 | 32 | 32 | 3583 | 874 |
| 15 | carbo-[3]-oxacarbon 1 | 0.03 | 30 | 36 | 3998 | 951 |
| | carbo-[3]-oxacarbon 2 | 0.03 | 30 | 36 | 4583 | 1045 |

Reaction 3, 4 and 5 are the decomposition energies of $P_4$, $SF_6$ and $PF_5$ respectively and all three have DMC errors greater than 6.5 kcal/mol. Reaction 17, the decomposition of a beryllium cluster, has an error of 1.3 ± 0.1 kcal/mol. Errors in the nodal surface can cancel out for energy differences provided there is an equivalent description of correlation for all species. These large errors in the decomposition energies of molecules containing second-row atoms point to a problem in the description of the molecule relative to the atoms. Reaction 6 is the decomposition of $P_4O_{10}$ into $P_4$ and $O_2$ but the error is 2.1 ± 0.4 kcal/mol, suggesting the errors for the molecules $P_4O_{10}$ and $P_4$ are cancelling out. Using more determinants can improve the description of the molecules in reactions 3, 4 and 5 relative to the atoms and reduce the error.

In reaction 3, the active space for $P_4$ was chosen to include all valence electrons and nine active orbitals for each atom (RAS(20, 36)). The error in the decomposition energy was reduced by approximately 3.5 kcal/mol to 11.9 ± 0.3 kcal/mol. This error is still significant but incorporating more orbitals into the active space could reduce it further. Including additional excitations had no effect on second-row ionisation potentials and electron affinities and is unlikely to decrease the error here.

Reactions 4 and 5 are similar to reaction 3 and single-determinant DMC also underbinds these molecules. Since they contain first- and second-row atoms the active space was initially chosen such that it included four levels for each first-row atom and nine levels for second-row atoms. The active space was RAS(48, 33) on $SF_6$ and RAS(40, 29) on $PF_5$. Although the active space included a large number of orbitals and electrons on both molecules there was no difference in the reaction energy compared to the single-determinant result. Previous QMC calculations have shown that multi-determinant DMC results are generally converged with smaller active spaces for molecules containing first-row atoms. A study looking at model retinal chromophores found no difference between the CAS(6,6) and CAS(6,12) space for $C_5NH_8$ and good results were seen with the smaller space.[46] Another study looking at models of the green fluorescent chromophore showed DMC results were converged with CAS(2,2) compared to CAS(12,11), even though the CASPT2 calculations used to generate the orbitals were not.[47] For second-row atoms larger active spaces with higher virtual orbitals are necessary. In a study of the G2 test set a modest active space incorporating valence electrons had a much greater effect on molecules with second row atoms compared to first row. In a small test with $PH_2$, $PH_3$ and $P_2$, a larger active space with more virtual orbitals reduced the error from 2.3 kcal/mol (s and p orbitals only) to 1.6 kcal/mol.[48]

Since the initial active space had no effect, the number of levels was increased to nine for all atoms. After freezing the doubly-occupied *s* orbitals on $PF_5$ the final active space was RAS(28, 48) and the magnitude of the error in reaction 5 was reduced by 6 kcal/mol. Using nine levels on all atoms (RAS(34,63)) resulted in too many CSFs for the DMC calculation of $SF_6$. The natural orbital occupation numbers (NOONs) from the CISD-B3LYP RAS(34,63) calculation were used to reduce the active space. All doubly-occupied orbitals with $\lambda_i > 1.991$ were frozen and any virtual level with $\lambda_i < 0.001$ was omitted. This reduced the active space for $SF_6$ to RAS(28, 47) but the atom active spaces were left unchanged. Using this modified active space lowered the error of reaction 4 to just 1.2 ± 0.3 kcal/mol. The active spaces used in the final DMC calculations of $SF_6$ and $PF_5$ included substantially more virtual levels than the initial active spaces but the number of electrons had to be reduced for the calculation to be computationally feasible. These results suggest that excitations into higher virtual orbitals are more important than including more electrons in the active space for DMC calculations of molecules containing second-row atoms.

### 6.4.3 Atomisation energies of large molecules

DFT errors are known to increase with system size and atomisation energies of large organic molecules like $C_6F_6$ (reaction 7) and urotropin ($C_6H_{12}N_4$, reaction 9) are particularly challenging.[11,24,49–51] DFT also has unexpectedly large errors for 'simple' saturated systems like urotropin and unsaturated compounds of the same size are described better.[11] The performance of DFT methods for these systems can be sporadic and unpredictable; most methods overbind $C_6F_6$ but B3LYP underbinds it by 2.9 kcal/mol.[49]

Of the ten decomposition energies in the DC18 test set, nine of them have DMC errors greater than 5 kcal/mol. Single-determinant DMC underestimated the binding energies of reactions 7 and 9 by at least 6.0 kcal/mol. Reaction 8, the decomposition of $Si(OCH_3)_4$, also has a large DMC error (4.2 ± 0.2 kcal/mol) but it was below the 5 kcal/mol threshold and was not studied further. The large errors for these systems can be attributed to a poor description of the nodal surface on the large molecules relative to the atoms. Similarly to the decomposition energies of second-row-containing molecules, including more determinants should improve the nodal surface of the molecules.

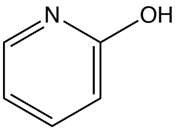
Using an active space with four levels per atom for $C_6F_6$ in reaction 7 was too computationally demanding and all doubly-occupied orbitals on atoms were frozen (RAS(18,14)). Using this active space in the DMC trial wavefunction reduced the error of reaction 7 to 1.3 ± 0.5 kcal/mol. A larger active space for urotropin in reaction 9 (RAS(56, 52), four levels for each C and N atom and 1 level for each H atom) still underestimated the decomposition energy by 4.2 ± 0.4 kcal/mol.

In both cases, a multi-determinant trial wavefunction reduced the DMC errors but the results are dependent on the molecule and active space used. For $C_6F_6$ a small active space with just 14 orbitals was sufficient. Urotropin has the same number of heavy atoms but a CISD trial wavefunction with an active space with 52 orbitals still resulted in a DMC error greater than 4 kcal/mol. The active space on $C_6F_6$ was much smaller than those used on the decomposition energies of hypervalent compounds, highlighting the unpredictable behaviour of these multi-determinant wavefunctions in DMC.

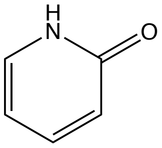
### 6.4.4 Isomerisation energies of organic molecules

Accurately describing the electron distribution in delocalised systems is a challenging problem for electronic structure methods. HF methods do not take electronic correlation into account and underestimate electron delocalisation but DFT methods have been shown to overestimate it.[52] Although DFT methods can accurately predict the structure of delocalised systems like cumulenes and carbomers they fail to predict the relative stabilities of isomers. They tend to overstabilise delocalised cumulenes relative to poly-enes and underestimate the energy difference.[53,54] DC18 contains four isomerisation energies of organic molecules, shown in Figure 6.2. QMC methods have been shown to recover a significant amount of correlation energy yet single-determinant DMC underestimated the energy differences for reactions 12 and 15 by approximately 5 kcal/mol. Reactions 10 and 14 are also isomerisation reactions but they have much smaller DMC errors.

DMC recovers a significant amount of correlation energy but the accuracy of energy differences is determined by error cancellation. When the errors in the nodal surface are equivalent the reaction energy will be good, regardless of how accurate the individual results are. For reactions 10 and 14 there is a good cancellation of errors and the reaction error is small. For reactions 12 and 15 there is an unbalanced description of the molecules and additional determinants are required. Choosing an active space for isomerisation energies is considerably easier than atomisation energies where the number of products and reactants can differ significantly but to maintain consistency with the previous calculations the same guidelines were used. For $C_7H_4$ (**12a** and **12b** in Figure 6.2) this gave an active space of 20 electrons in 32 orbitals for each isomer and reduced the error to 1.4 ± 0.3 kcal/mol. For the $C_9O_3$ isomers (**15a** and **15b** in Figure 6.2) the doubly-occupied s orbitals were frozen on each atom. The final active space (RAS(30, 36)) reduced the error



**10a**



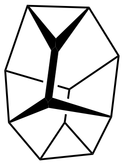
**10b**

$H_2C{=}C{=}C{=}C{=}C{=}C{=}CH_2$

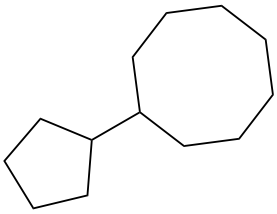
**12a**

$H_3C{-}C{\equiv}C{-}C{\equiv}C{-}C{\equiv}CH$

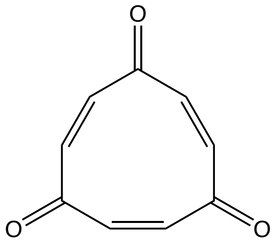
**12b**



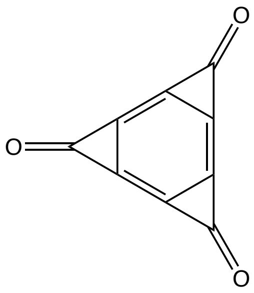
**14a**



**14b**



**15a**



**15b**

Figure 6.2: The structures of the isomers in reactions 10, 12, 14 and 15.

to 0.3 ± 0.3 kcal/mol.

The improvement made to the nodal surface by these additional determinants can be quantified by the nodal correlation energy, defined as:

$$\Delta E_{\text{corr}} = E_{\text{DMC-CI}} - E_{\text{DMC-B3LYP}}$$

where $E_{\text{DMC-CI}}$ is the multi-determinant energy and $E_{\text{DMC-B3LYP}}$ is the single-determinant energy. The nodal correlation energy can be used to establish which system was the most improved by a multi-determinant wavefunction. **15a** recovered approximately 12.5 kcal/mol correlation energy while **15b** recovered 5.4 kcal/mol. **12a** recovered 8.8 kcal/mol and **12b** recovered 5.2 kcal/mol. Although all four molecules recovered a significant amount of correlation energy in both reactions there was one isomer that recovered more. There is no distinction between the molecules that makes it clear as to when a multi-determinant wavefunction should be used. It can not always be attributed to the multi-reference character of the molecule. For example, the T1 diagnostic is a common metric used to assess the suitability of single-reference methods[55] and values over 0.02 suggest the need for a multi-reference calculation. The T1 diagnostic for one of the carbomer isomers in reaction 15 was 0.018 in the original work. This is close to the threshold of 0.02 but a rough CASSCF analysis showed no significant multi-reference character.[54] In this work using more determinants in the DMC wavefunction reduced the error by 5 kcal/mol.

### 6.4.5 Binding energy of sulphur

Reaction 18 is the binding of four sulfur dimers ($S_2$) to form a sulfur ring, $S_8$.[56] Multiply bonded sulphur compounds are known to be challenging for DFT methods[11] and DMC-B3LYP underestimated the binding energy by 8.0 kcal/mol. Using an active space of 9 levels on each atom was not feasible for a molecule this size. As shown for reactions 3, 4 and 5, higher virtual orbitals are especially important for multi-determinant DMC calculations with second-row atoms and an active space of 4 levels per atom would be inadequate. The largest active space considered for $S_8$ used an equivalent number of electrons and orbitals after freezing the s orbital on each atom (RAS(32, 32)). Energy differences rely on a balanced description of electron correlation on products and reactants and a suitable active space needs to be used on all species. Unfortunately this active space was not suitable and the error increased by 1.5 kcal/mol. Based on results from previous molecules in this set, more virtual levels need to be included in the active space but this as not possible at this stage. A CISDTQ wavefunction that includes more excitations is unlikely to improve the results and would be prohibitively expensive.

## 6.5 Comparison to other methods

Mean absolute deviations (MADs) for DMC and some popular WFT and DFT methods are reported in Figure 6.3. The MAD for single-determinant DMC for the two test sets

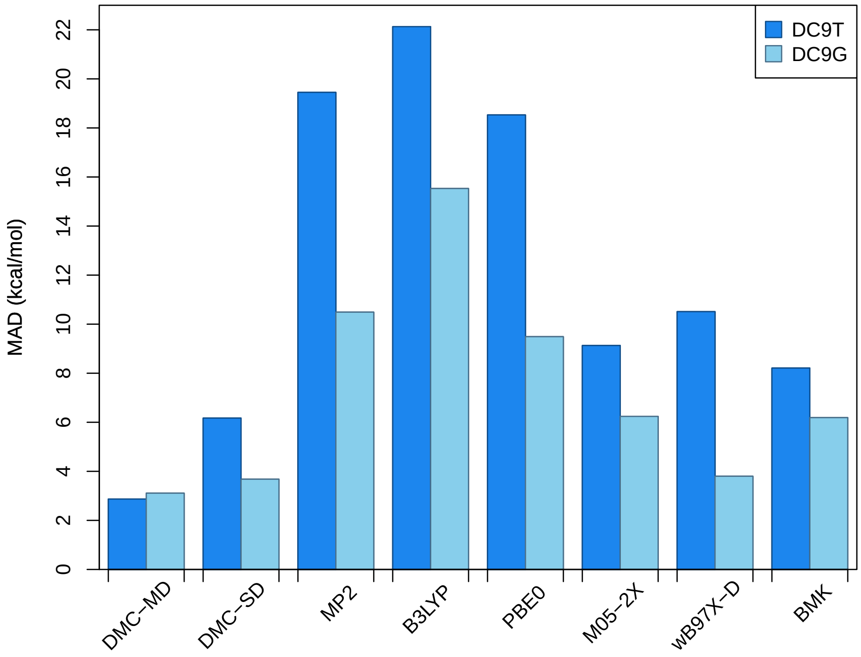


Figure 6.3: Mean absolute deviation (kcal/mol) for the DC9T and DC9G test sets. Methods shown include single-determinant DMC (DMC-SD), multi-determinant DMC (DMC-MD), *ab initio* wavefunction methods and DFT methods. Values for DC9T are taken from Ref. 23. Values for DC9G are taken from Ref. 14.

combined is 4.9 ± 4 kcal/mol, but the error for DC9G is 3.7 ± 4 kcal/mol compared to 6.2 ± 4 kcal/mol for DC9T. Despite these large errors, single-determinant DMC performs better than the other methods. DC18 is made up of reactions specifically chosen to challenge DFT methods and the large errors for DFT methods are unsurprising. Although DMC is not affected by the same problems as DFT methods, the errors for the two test sets follow the same trends. Reactions 1 to 9 (DC9T) are predominantly hypervalent compounds with second row atoms and these high-energy reactions are more challenging for all methods compared to reactions 10 to 18 (DC9G). Unlike the DFT and WFT methods shown, DMC can be improved by including more determinants in the trial wavefunction. Using a multi-determinant wavefunction for reactions 3, 4, 5, 7, 9, 12 and 15 reduced the errors associated with them, and the overall MAD is reduced to 3.1 ± 4 kcal/mol. This also reduced the disparity between the two subsets, the MADs for DC9T and DC9G were reduced to 3.4 ± 4 and 2.9 ± 4 kcal/mol respectively.

## 6.6 Summary

DMC is not affected by the same errors as DFT methods but the quality of reaction energies in the DC18 test set was low. Reactions 1 to 9 are more challenging for DFT methods and also had larger errors for DMC compared to reactions 10 to 18. Despite this, single-determinant DMC still had smaller errors for the DC18 test set compared to other methods. A unique advantage of DMC methods over DFT methods is they can be systematically improved by including more determinants in the wavefunction. Unfortunately it is not always clear when a multi-determinant wavefunction should be used or what the virtual orbitals the active space should include. In general, decomposition energies of large molecules require a multi-determinant wavefunction to better describe the nodal surface of the molecule relative to the atoms. Higher virtual orbitals are necessary for molecules with second-row atoms and including more determinants will only improve the final energy if sufficient virtual orbitals are included. Unfortunately including more determinants doesn't always work, as was seen for reaction 18 ($S_8$) and it may only remove some of the error, like reactions 3 and 9. For isomerisation reactions, it is not apparent which systems might need more determinants but a smaller active space is sufficient. Despite these drawbacks DMC performs consistently well for the range of reactions compared to other methods.

## 6.7 References

[1] A. Lüchow, *WIREs Comput. Mol. Sci.* **2011**, *1*, 388–402.

[2] B. Austin, D. Y. Zubarev, W. A. Lester, *Chem. Rev.* **2012**, *112*, 263–88.

[3] C. J. Cramer, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757–10816.

[4] F. Neese, *Coord. Chem. Rev.* **2009**, *253*, 526 – 563.

[5] H. Hu, W. Yang, *Annu. Rev. Phys. Chem.* **2008**, *59*, 573–601.

[6] B. Santra, A. Michaelides, M. Fuchs, A. Tkatchenko, C. Filippi, M. Scheffler, *J. Chem. Phys.* **2008**, *129*, 194111.

[7] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289–320.

[8] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Science* **2008**, *321*, 792–794.

[9] E. R. Johnson, P. Mori-Sánchez, A. J. Cohen, W. Yang, *J. Chem. Phys.* **2008**, *129*, 204112.

[10] V. Guner, K. S. Khuong, A. G. Leach, P. S. Lee, M. D. Bartberger, K. N. Houk, *J. Phys. Chem. A* **2003**, *107*, 11445–11459.

[11] S. Grimme, *J. Phys. Chem. A* **2005**, *109*, 3067–3077.

[12] J. A. Phillips, C. J. Cramer, *J. Chem. Theory Comput.* **2005**, *1*, 827–833.

[13] J. S. Lee, *J. Phys. Chem. A* **2005**, *109*, 11927–11932.

[14] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2010**, *6*, 107–126.

[15] E. T. Swann, M. L. Coote, A. S. Barnard, M. C. Per, *Int. J. Quantum Chem.* **2017**, e25361.

[16] E. Carvalho, A. N. Barauna, F. B. Machado, O. Roberto-Neto, *Chem. Phys. Lett.* **2008**, *463*, 33–37.

[17] J. Pu, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 773–778.

[18] M. Dubecký, P. Jurečka, R. Derian, P. Hobza, M. Otyepka, L. Mitáš, *J. Chem. Theory Comput.* **2013**, *9*, 4287–4292.

[19] M. Dubecký, R. Derian, P. Jurečka, L. Mitáš, P. Hobza, M. Otyepka, *Phys. Chem. Chem. Phys.* **2014**, *16*, 20915.

[20] M. C. Per, S. P. Russo, I. K. Snook, *J. Chem. Phys.* **2009**, *130*, 134103.

[21] J. Řezáč, P. Jurečka, K. E. Riley, J. Černỳ, H. Valdes, K. Pluháčková, K. Berka, T. Řezáč, M. Pitoňák, J. Vondrášek, P. Hobza, *Coll. Czechoslovak Chem. Comm.* **2008**, *73*, 1261–1270.

[22] S. Sorella, M. Casula, D. Rocca, *J. Chem. Phys.* **2007**, *127*, 014105.

[23] Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[24] P. R. Schreiner, A. A. Fokin, R. A. Pascal, A. de Meijere, *Org. Lett.* **2006**, *8*, 3635–3638.

[25] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[26] M. Dolg, C. Filippi, *Private Communication* **2014**.

[27] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[28] R. C. Clay III, M. A. Morales, *J. Chem. Phys.* **2015**, *142*, 234103.

[29] M. C. Per, D. M. Cleland, *J. Chem. Phys.* **2017**, *146*, 164101.

[30] J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 084102.

[31] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comp. Chem.* **1993**, *14*, 1347–1363.

[32] M. Gordon, M. Schmidt in *Theory and Applications of Computational Chemistry: the first forty years*, C. Dykstra, G. Frenking, K. Kim, G. Scuseria (Eds.), Elsevier, Amsterdam, **2005**, pp. 1167–1189.

[33] M. C. Per, *CSIRO Quantum Monte Carlo software package*, **2017**.

[34] M. Piacenza, S. Grimme, *J. Comput. Chem.* **2004**, *25*, 83–98.

[35] H. Prinzbach, A. Weiler, P. Landenberger, F. Wahl, J. Wörth, L. T. Scott, M. Gelmont, D. Olevano, B. V. Issendorff, *Nature* **2000**, *407*, 60–63.

[36] K. Raghavachari, D. L. Strout, G. K. Odom, G. E. Scuseria, J. A. Pople, B. G. Johnson, P. M. W. Gill, *Chem. Phys. Lett.* **1993**, *214*, 357–361.

[37] E. J. Bylaska, P. R. Taylor, R. Kawai, J. H. Weare, *J. Phys. Chem.* **1996**, *100*, 6966–6972.

[38] S. Sokolova, A. Lüchow, J. B. Anderson, *Chem. Phys. Lett.* **2000**, *323*, 229–233.

[39] S. Grimme, C. Mück-Lichtenfeld, *ChemPhysChem* **2002**, *3*, 207–209.

[40] W. An, Y. Gao, S. Bulusu, X. C. Zeng, *J. Chem. Phys.* **2005**, *122*, 204109.

[41] Y. Jin, A. Perera, V. F. Lotrich, R. J. Bartlett, *Chem. Phys. Lett.* **2015**, *629*, 76–80.

[42] D. Manna, J. M. Martin, *J. Phys. Chem. A* **2015**, *120*, 153–160.

[43] P. Jurečka, P. Hobza, *Chem. Phys. Lett.* **2002**, *365*, 89–94.

[44] M. Korth, W. Thiel, *J. Chem. Theory Comput.* **2011**, *7*, 2929–2936.

[45] L. A. Curtiss, K. Raghavachari, P. C. Redfern, J. A. Pople, *J. Chem. Phys.* **2000**, *112*, 7374–7383.

[46] O. Valsson, C. Filippi, *J. Chem. Theory Comput.* **2010**, *6*, 1275–1292.

[47] C. Filippi, F. Buda, L. Guidoni, A. Sinicropi, *J. Chem. Theory Comput.* **2012**, *8*, 112–124.

[48] F. R. Petruzielo, J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2012**, *136*, 124116.

[49] L. A. Curtiss, P. C. Redfern, K. Raghavachari, *J. Chem. Phys.* **2005**, *123*, 124107.

[50] S. Grimme, C. Diedrich, M. Korth, *Angew. Chemie Int. Ed.* **2006**, *45*, 625–629.

[51] S. Grimme, *Angew. Chemie Int. Ed.* **2006**, *45*, 4460–4464.

[52] A. E. Pomerantz, J. H. Han, C. B. Musgrave, *J. Phys. Chem. A* **2004**, *108*, 4030–4035.

[53] H. L. Woodcock, H. F. Schaefer, P. R. Schreiner, *J. Phys. Chem. A* **2002**, *106*, 11923–11931.

[54] C. Lepetit, H. Chermette, M. Gicquel, J. L. Heully, R. Chauvin, *J. Phys. Chem. A* **2007**, *111*, 136–149.

[55] T. J. Lee, P. R. Taylor, *Int. J. Quantum Chem. Quantum Chem. Symp. 23* **1989**, 199–207.

[56] S. Grimme, *J. Chem. Phys.* **2006**, *124*, 034108.

Chapter 7

# The isomerization of diazene ($N_2H_2$)

## 7.1 Introduction

Azo compounds, defined by −N=N− functional groups, have a broad range of applications. They are powerful and selective reducing agents and sources of free radicals.[1] They can undergo a reversible *cis/trans* photochemical or thermal transformation about the the -N=N- double bond, potentially acting as molecular switches or optical data storage systems.[2–4] Diazene ($N_2H_2$, also known as diimide) is the simplest azo compound. It has been used for stereospecific reduction of olefins[5] and as a ligand for transition metal complexes.[6]

The isomerisation of diazene has been extensively studied.[4,7–25] There are three isomers, *cis*, *trans* and *iso*-$N_2H_2$ and three isomerisation pathways; rotation about the -N=N- double bond, in-plane hydrogen inversion and an N-H bond cleavage/recombination. The transition state for the inversion pathway is predominantly single-reference but the rotational pathway is strongly correlated. The last pathway is often ignored due to the instability of the diazenyl radical. Rotation about a double bond involves breaking the $\pi$-bond and passing through a diradical transition state before reforming the $\pi$-bond. The configurational degeneracy can be seen in the potential energy curve of the rigid rotation of the *trans* structure about the double bond. Single-reference methods like CCSD overestimate the energy and a cusp-like feature is seen at a dihedral angle of $90^o$ if the energy is plotted as a function of dihedral angle.[11] Multi-reference methods give smoother curves and better estimates of the energy.[8,10,18] The complicated, multi-reference nature of the rotational transition state means it is often omitted from studies[4,21,22,24] but when both barrier heights are compared the results are inconsistent. In some cases the rotational barrier is much higher in energy than the inversion barrier[20,25] but other studies show the barrier heights are similar.[12,23]

Accurate relative energies need equivalent descriptions of correlation on all species but this is challenging when one structure has significant degeneracies. Single-reference methods give a poor description of the non-dynamical correlation on the rotational transition state. Multi-reference methods like multiconfigurational space self-consistent field (MC-

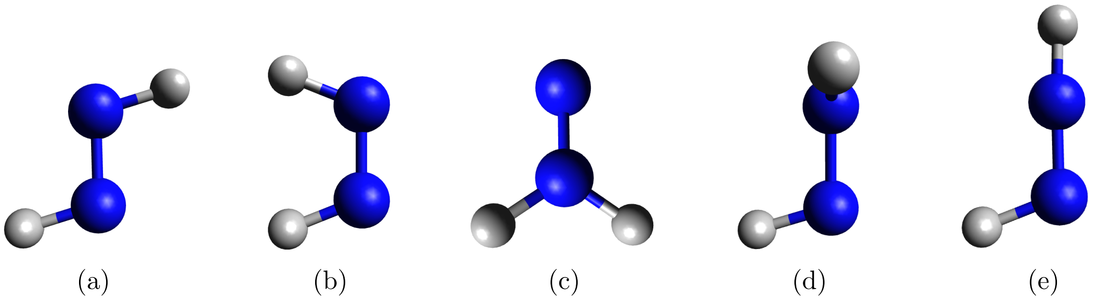


Figure 7.1: The diazene isomers and transition states a) *trans*-$N_2H_2$, b) *cis*-$N_2H_2$, c) *iso*-$N_2H_2$, d) TS1, out-of-plane rotational transition state e) TS2, in-plane hydrogen inversion transition state

SCF) or multi-reference configuration interaction (MRCI) methods, designed specifically for these types of degenerate systems, can fail to give a good description of the dynamic correlation on closed shell species. QMC methods like DMC are well suited to this type of problem. DMC with a trial wavefunction built from single-reference orbitals (e.g. HF, DFT) can fail for 'near-degeneracy' systems but including more determinants can account for the non-dynamic correlation. DMC has also previously been used to study the singlet and triplet excited states of *trans*-azobenzene, another popular azo-compound. Using a MCSCF trial wavefunction gave good agreement with experimental values.[26,27]

The major challenge of this system is finding a balanced description of the static and dynamic correlation on all species. Chapter 5 showed DMC with a multi-determinant trial wavefunction improved the description of correlation on challenging, degenerate species like Be, B, C and N. Including triple and quadruple excitations in the trial wavefunction had little effect on the final energy but the accuracy of the energy differences depended on the active space. The correlation energy of second-row anions and neutral atoms was poorly described relative to cations for smaller active spaces and the final energy differences were worse than single-determinant results. Chapter 6 showed that multi-determinant DMC calculations were no better than single-determinant DMC calculations with small active spaces for $PF_5$ and $SF_6$. The choice of multi-determinant method and active space used to generate the starting orbitals for a DMC calculation will ultimately govern the accuracy of the final energy. For single-determinant DMC it has been shown that Kohn-Sham orbitals give better trial wavefunctions than Hartree-Fock orbitals[28,29] but multi-determinant expansions are more complicated. In this chapter the effect that different multi-determinant methods and the choice of active space within these methods have on DMC total and relative energies is investigated. Three multi-determinant methods are compared; complete active space self-consistent field (MCSCF), configuration interaction with single and double excitations (CISD) and configuration interaction with single, double, triple and quadruple excitations (CISDTQ).

## 7.2 Methodology

The structures of the three isomers and two transition states are shown in Figure 7.1. Molecular geometries were taken from previous work where they were calculated using the 2-RDM method and an aug-cc-pVTZ basis set.[7] Two single-determinant methods (HF and B3LYP) are compared against three multi-determinant methods. Multi-configuration self consistent field (MCSCF) wave functions in complete active spaces (CAS) were generated using all possible excitations in a given active space. Configuration interaction (CI) calculations were truncated to include single and double excitations (CISD) or single, double, triple and quadruple excitations (CISDTQ) from a single-determinant B3LYP reference wavefunction. The use of Kohn-Sham orbitals in CI wavefunctions has been shown to give better DMC nodal surfaces than Hartree-Fock orbitals.[29] Full details of the active spaces are outlined in the results section below. The notation $(N_\mathrm{e}, N_\mathrm{orb})$ is used, where $N_\mathrm{e}$ is the number of active electrons and $N_\mathrm{orb}$ is the number of orbitals. B3LYP, CI and MCSCF calculations were performed with GAMESS.[30,31] The number of configuration state functions (CSFs) produced by multi-determinant methods is usually too large for DMC methods and they are commonly truncated according to some threshold. A weight-based truncation scheme was used here where the number of CSFs were selected such that the sum of the coefficients was 99.5% of the original expansion.

Fixed-node diffusion Monte Carlo calculations were performed using the CMQMC code[32] with a target population size of 6400 walkers and an imaginary time step size of 0.01 a.u. The Jastrow factor used here is the sum of electron-electron (*ee*) and electron-nucleus (*eN*) terms.[33] Burkatzki-Filippi-Dolg (BFD) pseudopotentials with the associated triple-zeta (VTZ) basis sets were used[34] with an improved H-atom potential.[35] DMC calculations involving nonlocal pseudopotentials used size-consistent T-moves with a symmetric branching term.[36] The variable parameters in the Jastrow factor and the CSF coefficients were optimised by minimising the variational energy using an approach based on the linear method.[37]

## 7.3 Single-determinant DMC

Energies relative to *trans*-$N_2H_2$ are reported in Table 7.1 for DMC using single-determinant methods (HF and B3LYP) as well as relative energies from the trial wavefunctions. Multi-reference configuration interaction with Davidson[38] and Pople[39] size consistency corrections and an aug-cc-pVQZ basis set (MRCI+Q/AVQZ) reference values are taken from Ref. 12.

There is good agreement between the reference values and the single-determinant DMC-HF and DMC-B3LYP for the predominantly single-reference species *cis*-$N_2H_2$, *iso*-$N_2H_2$ and the inversion transition state, TS2. Both DMC-HF and DMC-B3LYP overestimate the relative energy of the rotational transition state, TS1, by more then 20 kcal/mol. This is unsurprising given the strongly correlated nature of this structure. The DMC-B3LYP relative energies are lower than DMC-HF values for all structures and over-

Table 7.1: Relative energies with respect to the *trans* isomer (kcal/mol) calculated from single-determinant methods with BFD pseudopotentials. DMC statistical errors on the last digit are shown in parentheses. MRCI+Q/AVQZ reference values are taken from Ref. 12.

| Method | *cis*-$N_2H_2$ | *iso*-$N_2H_2$ | TS1 | TS2 |
|---|---|---|---|---|
| MRCI+Q/AVQZ | 5.05 | 24.04 | 54.96 | 51.07 |
| HF | 6.22 | 17.92 | 87.56 | 55.10 |
| B3LYP | 5.41 | 20.52 | 74.97 | 48.98 |
| DMC-HF | 5.8(1) | 25.8(1) | 79.1(1) | 54.0(1) |
| DMC-B3LYP | 5.6(1) | 24.2(1) | 77.5(1) | 53.7(1) |
| CCSD(T)/EBSL [a] | 5.38 | | 62.64 | 52.3 |
| 2-RDM [a] | 5.44 | | 50.73 | 53.84 |

[a] Reference 7

all DMC-B3LYP has a mean absolute deviation (MAD) of 6.5 kcal/mol compared to 7.4 kcal/mol for DMC-HF. This is consistent with previous work that has shown the use of a correlated method like B3LYP gives better nodal surfaces compared to Hartree-Fock.[28, 29] Single-determinant DMC won't properly describe the non-dynamical correlation that is important in multi-reference systems like the rotational transition state, TS1, but including more determinants in the trial wavefunction can recover this correlation energy.

## 7.4 Multi-determinant DMC

### 7.4.1 Total energies

The number of determinants, configuration state functions (CSFs) and total DMC energy for different multi-determinant trial wavefunctions are reported for each isomer Table 7.2. Figure 7.2 and Figure 7.3 show the total trial wavefunction and DMC energy as a function of the number of orbitals, $N_{\text{orb}}$, in the active space $(N_{\text{e}}, N_{\text{orb}})$.

Including more orbitals in the active space lowered the total energy of the trial wavefunction for all species. MCSCF energies are lower than both CISD and CISDTQ energies when the same active space is used. The DMC total energies are much less sensitive to the active space or method compared to the trial wavefunction energies and Figure 7.3 shows a closer view of the total DMC energies as a function of $N_{\text{orb}}$. DMC-MCSCF energies are still lower than DMC-CISD and DMC-CISDTQ energies for the same active space but the difference is much smaller than the trial wavefunctions. In general, increasing the active space for MCSCF trial wavefunctions decreases the total DMC energy but this trend is not consistent. For example, increasing the active space from MCSCF(4,3) to MCSCF(6,6) decreases the DMC energy of *cis*-$N_2H_2$ by 10 mHa, but increases the energy of *iso*-$N_2H_2$ by 2 mHa. The DMC total energy of the strongly correlated rotational transition state shows little change when the active space is increased from MCSCF(4,3) to MCSCF(8,8) but drops by 12 mHa when the full valence space, MCSCF(12, 10), is used.

MCSCF calculations produced substantially more CSFs than both CI methods for

Table 7.2: Total DMC energies, E (Ha), number of determinants (Det) and number of configuration state functions (CSFs) for the isomers and transition states of diazene using trial wavefunctions with different multi-determinant methods and active spaces ($N_e, N_{orb}$). DMC statistical uncertainty on the last digit is shown in parentheses.

(a) *trans*-$N_2H_2$

| Method | ($N_e, N_{orb}$) | Det | CSF | Total CSF | E (Ha) |
|---:|---:|---:|---:|---:|---:|
| CISD | (12,10) | 27 | 10 | 325 | -21.0453(4) |
| CISD | (12,12) | 124 | 41 | 703 | -21.0436(3) |
| CISD | (12,14) | 277 | 83 | 1225 | -21.0460(3) |
| CISD | (12,18) | 924 | 254 | 2701 | -21.0469(3) |
| CISDTQ | (12,10) | 26 | 11 | 5495 | -21.0451(1) |
| CISDTQ | (12,18) | 675 | 173 | 459971 | -21.0477(1) |
| CISDTQ | (12,20) | 1020 | 264 | 854050 | -21.0486(1) |
| MCSCF | (2,2) | 4 | 3 | 3 | -21.0348(2) |
| MCSCF | (4,3) | 1 | 1 | 6 | -21.0352(2) |
| MCSCF | (6,6) | 18 | 7 | 175 | -21.0425(2) |
| MCSCF | (8,8) | 74 | 23 | 1764 | -21.0467(2) |
| MCSCF | (12,10) | 161 | 57 | 13860 | -21.0480(2) |

(b) *cis*-$N_2H_2$

| Method | ($N_e, N_{orb}$) | Det | CSF | Total CSF | E (Ha) |
|---:|---:|---:|---:|---:|---:|
| CISD | (12,10) | 20 | 8 | 325 | -21.0367(1) |
| CISD | (12,12) | 139 | 40 | 703 | -21.0364(3) |
| CISD | (12,14) | 323 | 89 | 1225 | -21.0376(3) |
| CISD | (12,18) | 888 | 230 | 2701 | -21.0372(3) |
| CISDTQ | (12,10) | 9 | 6 | 5495 | -21.0361(1) |
| CISDTQ | (12,18) | 748 | 190 | 459971 | -21.0386(1) |
| CISDTQ | (12,20) | 1119 | 278 | 854050 | -21.0404(1) |
| MCSCF | (2,2) | 4 | 3 | 3 | -21.0260(2) |
| MCSCF | (4,3) | 1 | 1 | 6 | -21.0257(3) |
| MCSCF | (6,6) | 24 | 8 | 175 | -21.0359(2) |
| MCSCF | (8,8) | 59 | 18 | 1764 | -21.0381(2) |
| MCSCF | (12,10) | 138 | 42 | 13860 | -21.0376(2) |

(c) *iso*-$N_2H_2$

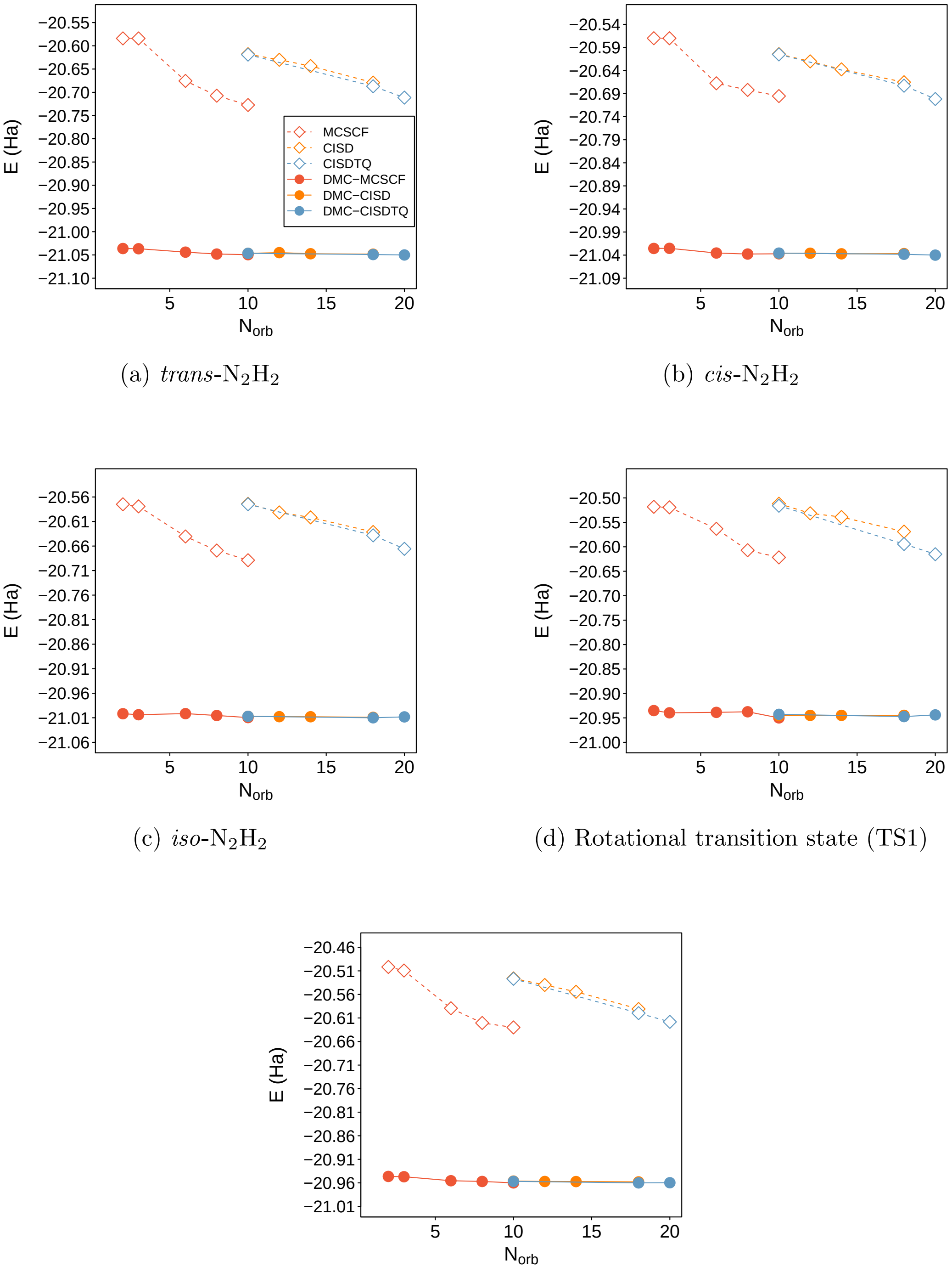
| Method | $(N_\mathrm{e}, N_\mathrm{orb})$ | Det | CSF | Total CSF | E (Ha) |
|---|---|---|---|---|---|
| CISD | (12,10) | 45 | 16 | 325 | -21.0055(4) |
| CISD | (12,12) | 148 | 45 | 703 | -21.0061(4) |
| CISD | (12,14) | 232 | 69 | 1225 | -21.0062(3) |
| CISD | (12,18) | 802 | 209 | 2701 | -21.0074(3) |
| CISDTQ | (12,10) | 22 | 10 | 5495 | -21.0054(1) |
| CISDTQ | (12,18) | 619 | 163 | 459971 | -21.0084(1) |
| CISDTQ | (12,20) | 1069 | 273 | 854050 | -21.0065(1) |
| MCSCF | (2,2) | 4 | 3 | 3 | -21.0000(2) |
| MCSCF | (4,3) | 3 | 3 | 6 | -21.0021(2) |
| MCSCF | (6,6) | 38 | 14 | 175 | -20.9998(2) |
| MCSCF | (8,8) | 138 | 42 | 1764 | -21.0037(2) |
| MCSCF | (12,10) | 200 | 63 | 13860 | -21.0079(2) |

(d) Rotational transition state (TS1)

| Method | $(N_\mathrm{e}, N_\mathrm{orb})$ | Det | CSF | Total CSF | E (Ha) |
|---|---|---|---|---|---|
| CISD | (12,10) | 38 | 14 | 325 | -20.9465(1) |
| CISD | (12,12) | 253 | 77 | 703 | -20.9459(3) |
| CISD | (12,14) | 432 | 126 | 1225 | -20.9459(4) |
| CISD | (12,18) | 1505 | 405 | 2701 | -20.9458(3) |
| CISDTQ | (12,10) | 39 | 13 | 5495 | -20.9438(1) |
| CISDTQ | (12,18) | 2447 | 600 | 459971 | -20.9484(1) |
| CISDTQ | (12,20) | 3576 | 858 | 854050 | -20.9449(1) |
| MCSCF | (2,2) | 4 | 3 | 3 | -20.9361(2) |
| MCSCF | (4,3) | 4 | 3 | 6 | -20.9407(2) |
| MCSCF | (6,6) | 45 | 14 | 175 | -20.9398(2) |
| MCSCF | (8,8) | 173 | 59 | 1764 | -20.9386(2) |
| MCSCF | (12,10) | 474 | 155 | 13860 | -20.9511(2) |

(e) Inversion transition state (TS2)

| Method | $(N_\mathrm{e}, N_\mathrm{orb})$ | Det | CSF | Total CSF | E (Ha) |
|---|---|---|---|---|---|
| CISD | (12,10) | 43 | 14 | 325 | -20.9587(3) |
| CISD | (12,12) | 220 | 66 | 703 | -20.9595(3) |
| CISD | (12,14) | 555 | 155 | 1225 | -20.9597(3) |
| CISD | (12,18) | 1718 | 461 | 2701 | -20.9604(3) |
| CISDTQ | (12,10) | 33 | 11 | 5495 | -20.9593(1) |
| CISDTQ | (12,18) | 1334 | 343 | 459971 | -20.9625(1) |
| CISDTQ | (12,20) | 1957 | 493 | 854050 | -20.9622(1) |
| MCSCF | (2,2) | 4 | 3 | 3 | -20.9488(2) |
| MCSCF | (4,3) | 2 | 2 | 6 | -20.9495(2) |
| MCSCF | (6,6) | 49 | 15 | 175 | -20.9579(2) |
| MCSCF | (8,8) | 121 | 41 | 1764 | -20.9595(2) |
| MCSCF | (12,10) | 309 | 96 | 13860 | -20.9625(2) |



(a) *trans*-$N_2H_2$

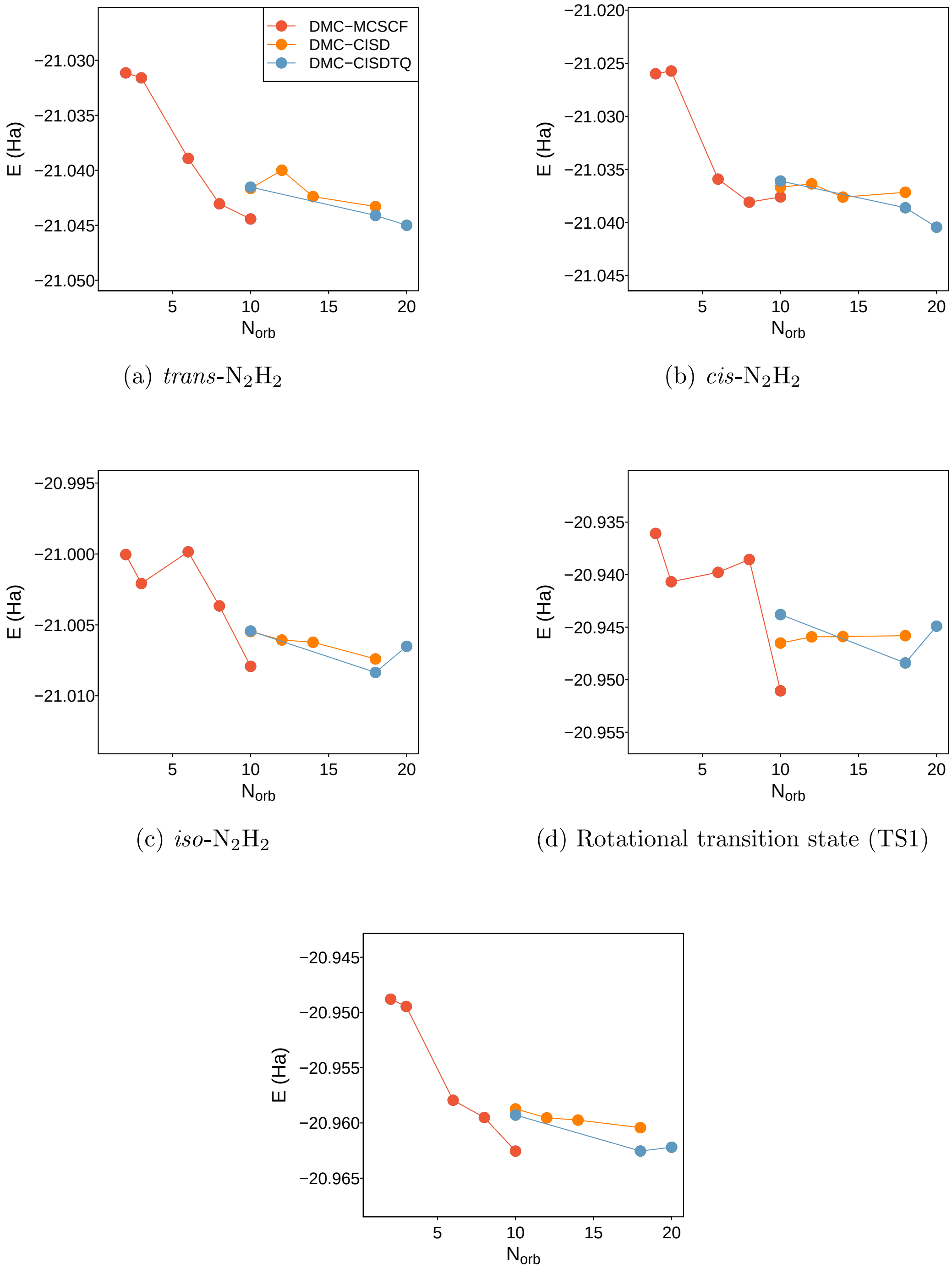
(b) *cis*-$N_2H_2$

(c) *iso*-$N_2H_2$

(d) Rotational transition state (TS1)

(e) Inversion transition state (TS2)

Figure 7.2: Total energies, E (Ha), of the diazene isomers and transition states as a function of the number of orbitals, $N_{\rm orb}$, in the active space $(N_{\rm e}, N_{\rm orb})$.



(a) *trans*-$N_2H_2$

(b) *cis*-$N_2H_2$

(c) *iso*-$N_2H_2$

(d) Rotational transition state (TS1)

(e) Inversion transition state (TS2)

Figure 7.3: Total DMC energies, E (Ha), of the diazene isomers and transition states as a function of the number of orbitals, $N_{\text{orb}}$, in the active space $(N_{\text{e}}, N_{\text{orb}})$.

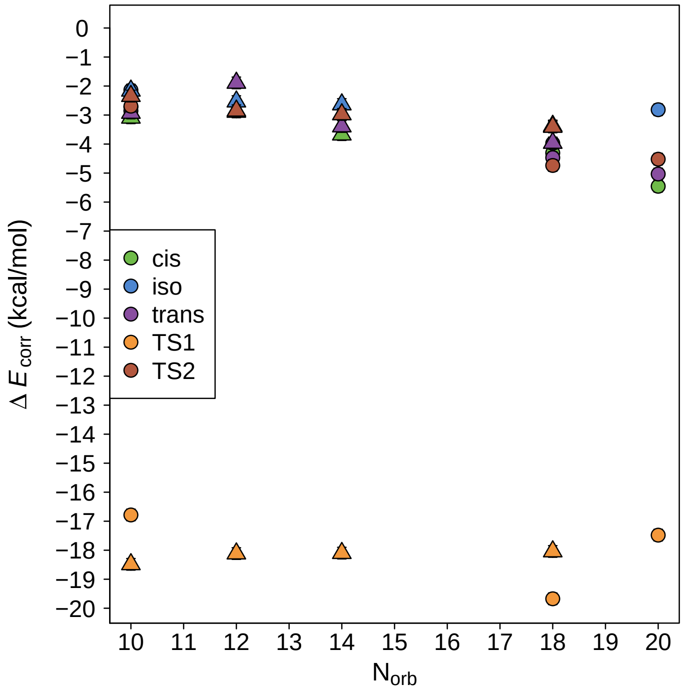


Figure 7.4: Nodal correlation energy (kcal/mol) for *cis, iso*, TS1 and TS2 (kcal/mol). CISD results are represented by triangles (△). CISDTQ results are represented by circles (○).

the same active space and more CSFs are included in DMC-MCSCF(12,10) calculations. Although CISDTQ produces more CSFs than CISD (5495 compared to 325) for the full-valence active space the number of CSFs used in the DMC calculation is virtually the same after the weight-based truncation scheme is applied. The number of CSFs was reduced such that the sum of the coefficients of the expansion used in DMC calculations was 99.5% of the original expansion and this suggests the determinants associated with the triple and quadruple excitations have very small weights for this system. For most species the DMC-CISDTQ energy is the same or slightly lower than DMC-CISD for the same active space. For the rotational transition state, TS1, the DMC-CISD(12, 10) energy is slightly lower than the DMC-CISDTQ(12,10) energy but the difference is only 2.7 mHa.

Correlation energy is usually defined as the difference between the exact energy and the Hartree-Fock value. Single-determinant DMC recovers a significant portion of this correlation energy and the remainder can be obtained by including more determinants to improve the nodal surface. The effect of these additional configurations on the total energy is described by the nodal correlation energy, defined as the difference between the single-determinant and multi-determinant DMC energies:

$$\Delta E_{\text{corr}} = E_{\text{DMC-CI}} - E_{\text{DMC-B3LYP}} \tag{7.1}$$

Table 7.3: DMC energies relative to *trans*-$N_2H_2$ (kcal/mol). DMC statistical uncertainty on the last digit is shown in parentheses. Reference MRCI+Q/AVQZ values are taken from Ref. 12.

| Method | $(N_e, N_{orb})$ | *cis*-$N_2H_2$ | *iso*-$N_2H_2$ | TS1 | TS2 |
|---|---|---|---|---|---|
| Ref. | | 5.05 | 24.04 | 54.96 | 51.07 |
| HF | | 5.79(12) | 25.84(12) | 79.14(13) | 53.98(12) |
| B3LYP | | 5.55(13) | 24.20(12) | 77.52(12) | 53.71(12) |
| CISD | (12,10) | 5.38(24) | 24.97(32) | 61.97(24) | 54.30(31) |
| CISD | (12,12) | 4.55(29) | 23.56(31) | 61.30(30) | 52.75(30) |
| CISD | (12,14) | 5.27(28) | 24.96(29) | 62.82(30) | 54.13(28) |
| CISD | (12,18) | 6.12(28) | 24.79(28) | 63.44(29) | 54.27(29) |
| CISDTQ | (12,10) | 5.68(12) | 24.92(11) | 63.59(12) | 53.88(11) |
| CISDTQ | (12,18) | 5.71(10) | 24.70(10) | 62.32(10) | 53.45(11) |
| CISDTQ | (12,20) | 5.13(10) | 26.42(10) | 65.08(10) | 54.23(10) |
| MCSCF | (2,2) | 5.67(12) | 21.85(11) | 62.13(12) | 53.95(12) |
| MCSCF | (4,3) | 5.84(13) | 20.97(12) | 59.52(13) | 53.77(12) |
| MCSCF | (6,6) | 4.10(12) | 26.78(11) | 64.50(12) | 53.14(11) |
| MCSCF | (8,8) | 5.41(11) | 26.95(11) | 67.72(11) | 54.62(11) |
| MCSCF | (12,10) | 6.60(11) | 25.43(10) | 61.07(10) | 53.71(10) |

Figure 7.4 shows the DMC-CISD and DMC-CISDTQ nodal correlation energy as a function of $N_{orb}$. The additional excitations in CISDTQ trial wavefunctions have little effect on DMC energies compared to CISD wavefunctions, consistent with results from Chapter 5. The amount of correlation energy recovered by the strongly-correlated transition state, TS1, is independent of the active space and all methods recover approximately 17-19 kcal/mol of nodal correlation energy for this species. Although the other species are predominantly single-reference, a significant amount of correlation energy (2-4 kcal/mol) is recovered by the multi-determinant expansions. Unlike TS1, the nodal correlation energy of the single-reference isomers shows a weak dependence on the number of orbitals in the active space and including more orbitals recovers slightly more correlation energy.

The total energy of all species was lowered by including additional determinants in the trial wavefunction and a significant amount of correlation energy was recovered for the multi-reference transition state. The different multi-reference characters of the isomers and transition states is the major challenge of this system and the relative energies with respect to the *trans* isomer will only be improved if there's a better description of the electron correlation on all species.

### 7.4.2 Relative energies

Relative energies with respect to the *trans* isomer are reported in Table 7.3. Relative energy errors are shown in Figure 7.5. The mean absolute deviation (MAD) of each method is shown in Figure 7.6.

For *cis*-$N_2H_2$ there is good agreement with reference values for all DMC methods and active spaces. This is unsurprising given the single-reference nature of this isomer. The *iso* isomer has a small error for most DMC-CI calculations but MCSCF trial wavefunctions give a bad description of the electron correlation of this single-reference isomer for DMC methods. DMC-MCSCF calculations with smaller active spaces (MCSCF(2, 2) and MCSCF(4, 3)) underestimate the relative energy with respect to *trans*-$N_2H_2$ by approximately 2 kcal/mol but DMC-MCSCF calculations with larger active spaces (MCSCF(6,6) and MCSCF(8,8)) overestimate the relative energy by approximately 2 kcal/mol. The inversion transition state, TS2, is also predominantly single-reference but the relative energy is over-estimated by approximately 2 kcal/mol for both single and multi-determinant DMC methods. Previous work with high-level single-reference methods has shown the CCSD(T)/EBSL barrier height is 52.30 kcal/mol and completely-renormalised coupled cluster (CR-CC(T)/EBSL) gives a barrier height of 53.84 kcal/mol.[7] Both of these values are in good agreement with both the single- and multi-determinant methods here, with barriers between 53-54 kcal/mol. All methods consistently overestimate the relative energy of TS1 by at least 5 kcal/mol. Multireference calculations from previous work agree with the MRCI+Q/AVQZ reference values used here. CCSD(T)/EBSL and 2-RDM results are included in Table 7.1 for comparison. 2-RDM is a parameterised multireference method and gives relative energy of 50.73 kcal/mol,[7] in better agreement (and lower than) the reference MRCI+Q/AVQZ values used here. In contrast, the multi-determinant DMC results are in better agreement with CCSD(T)/EBSL results from previous work[7] (see Table 7.1), where the relative energy was 62.64 kcal/mol. Using a multi-determinant trial wavefunction reduced the DMC error compared to single-determinant DMC results but including additional excitations or orbitals in the active space did not reduce the error further.

Although DMC-MCSCF total energies were lower than DMC-CISD or DMC-CISDTQ for the same active space all three methods give similar relative energies with respect to *trans*-$N_2H_2$ for the full valence active space (12, 10). DMC-CISD and DMC-CISDTQ relative energies are consistent across all active spaces considered. Relative energies of *iso*-$N_2H_2$ and TS1 vary unpredictably with the size of the active space in MCSCF trial wavefunctions. Overall DMC-CISD(12,12) has the smallest MAD but there is not one method that performs best for all four structures. DMC with MCSCF(4,3) determinants has the smallest error for TS1, the most challenging structure, but the largest error for single-reference *iso*-$N_2H_2$.

## 7.5 Future work

Despite an extensive study using three different multi-determinant trial wavefunctions and a range of active spaces the DMC error on the relative energy of rotational transition state could not be reduced below 5 kcal/mol. The fixed-node approximation means DMC can be severely limited by the quality of the nodal surface and it is not always possibly to recover all the correlation energy. The convergence of the DMC energy with respect to the method and active space used to generate the determinants and also the number of

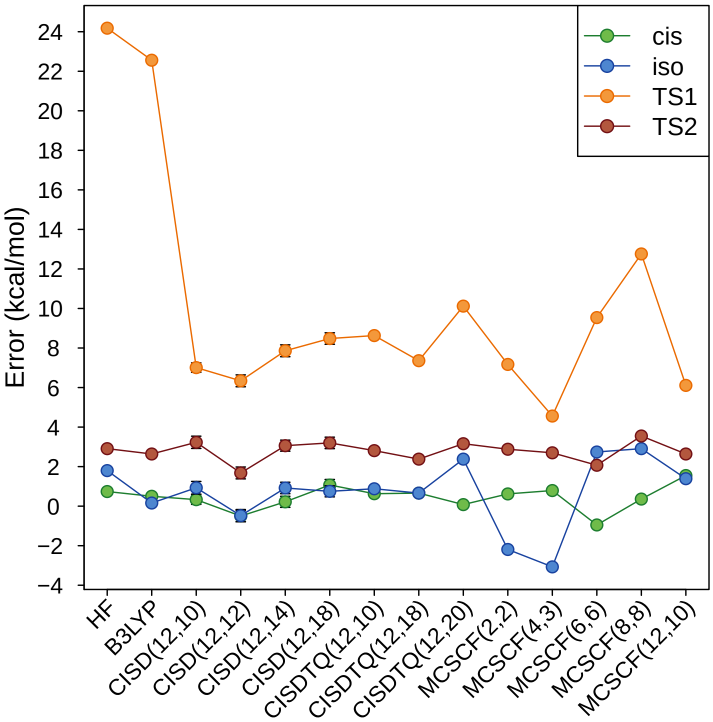


Figure 7.5: The errors (kcal/mol) of DMC relative energies with respect to *trans*-$N_2H_2$ for different trial wavefunctions. Errors are calculated with respect to reference MRCI+Q/AVQZ energies are taken from Ref. 12

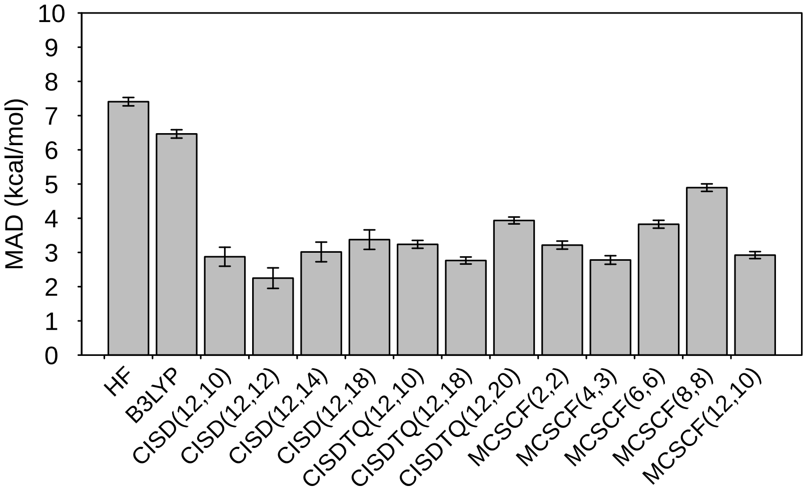


Figure 7.6: Mean absolute deviations (MAD) of the DMC relative energies for diazene isomers and transition states using different trial wavefunctions. Errors are calculated with respect to reference MRCI+Q/AVQZ energies are taken from Ref. 12

determinants used in the calculation is unique for each molecule. In this work no trial wavefunction gave a suitable nodal surface to adequately describe the static correlation of the rotational transition state relative to the *trans* isomer.

The isomerisation of diazene is a challenging problem but there are several avenues that could be explored further to reduce this error. The full-valence active space is a natural starting point for these types of studies. The orbitals that contribute to bonding need to be considered and generally these are found in the full-valence active space. For certain systems more orbitals need to be included, but it's not always apparent when this might be useful. It has been shown in several studies that for certain systems the orbitals active space must be extended beyond the full-valence space. Using a larger active space with just single and double excitations (SOCI) recovered more correlation energy for first-row atoms than just full-valence active space,[40] suggesting that excitations into higher virtual states can contribute significantly to the reduction of the fixed-node error. Another study looking at the binding energy of the beryllium dimer shown that the full valence active space was sufficient for the atom but was a poor choice for the dimer. Expanding the active space to include the $3s$ and $3p$ orbitals recovered more correlation energy but an even better result was obtained when the energy was extrapolated to the full CI limit.[41] Another study looking at the excitation energies of the green fluorescent protein noted that larger active spaces up to CAS(10,10) resulted in more accurate excitation energies even though the anionic chromophore does not have a strong multiconfigurational character.[42]

There is no 'one-size-fits-all' solution for this problem though. Previous studies have shown that trial wavefunctions with smaller active spaces can be sufficient to reduce the

DMC error of some systems. In a study looking at the photoisomerization of model retinal chromophores[43] it was shown that a CAS(6,6) expansion was sufficient. Increasing the number of virtual orbitals to 12 and 18 had no effect on the final excitation energy. For acrolein geometry optimisations an active space as small as CAS(2,2) was sufficient.[44]

In this work the size of the active space used in the trial wavefunction had little effect on the relative energies but larger active spaces improved the nodal surface and decreased the total DMC energy. Increasing the active space could lower the MCSCF energy even more. In Figure 7.3 the TS1 DMC-MCSCF energy appeared to have converged for both the (6,6) and (8,8) active spaces but there was a significant lowering of energy when the active space was expanded to MCSCF(12,10). In Figure 7.4 the amount of correlation energy recovered for DMC-CI calculations of single-reference species decreases slightly as the active space increases. This problem warrants an investigation using larger active spaces, but the accuracy of the result will still depend on the amount of correlation recovered by each species.

Optimising the orbitals in VMC would remove some of the ambiguity in the calculations by removing the starting orbitals as a variable. The choice of active space and excitation used would still affect the final results. It is the convergence of the total energy with respect to the size of the active space and number of determinants that ultimately governs the accuracy of the relative energies. This convergence is unique to each species but more correlation energy could be recovered by extrapolating to the full CSF limit. The number of CSFs used in the trial wavefunction is normally truncated according to some threshold. In this work the CSFs were truncated such that the sum of the coefficients was 99.5% of the original expansion. Some studies,[29, 40, 45] extrapolate the results to the full CSF limit but just a single point was used here. For atomisation energies of the G1 set extrapolating to the full CSF limit reduced the mean absolute deviation (MAD) by approximately 1 kcal/mol relative to the single-determinant results.[40] Although the error here is greater than 5 kcal/mol for all methods extrapolating to the full CSF limit could recover additional correlation energy for some species and reduce the DMC relative energy errors. .

The accuracy of fixed-node DMC is limited by the quality of the nodal surface but there are other methods available to recover the correlation energy. In addition to the multi-determinant expansion (static correlation) and Jastrow factor (dynamic correlation), a backflow transformation can be included. This backflow transformation allows further variations in the nodal surface, allowing for an improvement. In a study on the first-row atoms and ions it was shown that using a modest multi-determinant expansion with a backflow transformation could recover over 99% of the correlation energy at the DMC level.[46] Another possibility is auxiliary-field QMC (AFQMC). It uses the same imaginary-time propagation as DMC but stochastically samples from the determinant space. The chromium dimer is another example where DMC was unable to recover all of the correlation energy, regardless of the active space or trial wavefunction used.[47] The AFQMC potential energy curve (PEC) for the chromium dimer showed much better agreement with experiment.[48]

## 7.6 Summary

The isomerisation of diazene is a challenging problem and the solution was not as simple as just including more determinants, unlike the previous difficult cases in Chapter 6. DMC methods recover a significant amount of correlation energy but energy differences will only be accurate when the description of this correlation energy is equivalent on all species. The rotational transition state of the isomerisation of diazene is strongly multi-reference in nature and finding a good description of the static correlation of this system relative to the other isomers is problematic. In this work three different multi-determinant methods with different active spaces could not produce trial wavefunctions for DMC that reduced the error below 5 kcal/mol. Including more orbitals or excitations in the active space did not reduce this error further. Although MCSCF trial wavefunctions recovered more correlation energy and had lower total energies the relative energies of all methods was virtually the same. Using cheaper multi-determinant methods like CISD to generate trial wavefunctions is sufficient for these types of system. Further work is needed to develop and understand these methods. It is not clear why certain active spaces give smaller errors for MCSCF trial wavefunctions or why CISD or CISDTQ trial wavefunctions showed no change with active space. It is important to know the limitations of a given multi-determinant description of a trial wavefunction, to not only know when it might not work but to know what can be done to further improve it.

## 7.7 References

[1] C. E. Miller, *J. Chem. Edu.* **1965**, *42*, 254.

[2] H. Fliegl, A. Köhn, C. Hättig, R. Ahlrichs, *J. Am. Chem. Soc.* **2003**, *125*, 9821–9827.

[3] W. A. Sokalski, R. W. Góra, W. Bartkowiak, P. Kobyliński, J. Sworakowski, A. Chyla, J. Leszczyński, *J. Chem. Phys.* **2001**, *114*, 5504–5508.

[4] M. L. McKee, *J. Phys. Chem.* **1993**, *97*, 13608–13614.

[5] S. Vidyarthi, C. Willis, R. Back, R. McKitrick, *J. Am. Chem. Soc.* **1974**, *96*, 7647–7650.

[6] M. Veith, *Ang. Chem. Int. Ed.* **1976**, *15*, 387–388.

[7] A. M. Sand, C. A. Schwerdtfeger, D. A. Mazziotti, *J. Chem. Phys.* **2012**, *136*, 034112.

[8] M. Musiał, Ł. Lupa, K. Szopa, S. a. Kucharski, *Struct. Chem.* **2012**, *23*, 1377–1382.

[9] J. Jana, *Reports Theor. Chem.* **2012**, *1*, 1–10.

[10] U. S. Mahapatra, S. Chattopadhyay, *J. Chem. Phys.* **2011**, *134*, 044113.

[11] R. K. Chaudhuri, K. F. Freed, S. Chattopadhyay, U. Sinha Mahapatra, *J. Chem. Phys.* **2008**, *128*, 144304.

[12] M. Biczysko, L. A. Poveda, A. J. C. Varandas, *Chem. Phys. Lett.* **2006**, *424*, 46–53.

[13] X. Pu, N.-B. Wong, G. Zhou, J. Gu, A. Tian, *Chem. Phys. Lett.* **2005**, *408*, 101–106.

[14] D. Hwang, A. Mebel, *J. Phys. Chem. A* **2003**, *107*, 2865–2874.

[15] P. K. Chattarj, P. Perez, J. Zevallos, A. Toro-Labbe, *J. Mol. Struct.* **2001**, *580*, 171–182.

[16] V. Stepanic, G. Baranovic, *Chem. Phys.* **2000**, *254*, 151–168.

[17] J. M. L. Martin, P. R. Taylor, *Mol. Phys.* **1999**, *96*, 681–692.

[18] P. Mach, J. Masik, J. Urban, I. Hubac, *Mol. Phys.* **1998**, *94*, 173–179.

[19] B. Jursic, *Chem. Phys. Lett.* **1996**, *4*, 13–17.

[20] C. Angeli, R. Cimiraglia, H.-J. Hofmann, *Chem. Phys. Lett.* **1996**, *259*, 276–282.

[21] J. Andzelm, C. Sosa, R. A. Eades, *J. Phys. Chem.* **1993**, *97*, 4664–4669.

[22] B. J. Smith, *J. Phys. Chem.* **1993**, *97*, 10513–10514.

[23] H. J. A. Jensen, P. Joergensen, T. Helgaker, *J. Am. Chem. Soc.* **1987**, *109*, 2895–2901.

[24] C. A. Parsons, C. E. Dykstra, *J. Chem. Phys.* **1979**, *71*, 3025.

[25] N. W. Winter, *J. Chem. Phys.* **1975**, *62*, 1269.

[26] M. Dubecký, R. Derian, L. Mitáš, I. Štich, *J. Chem. Phys.* **2010**, *133*, 244301.

[27] M. Dubecký, R. Derian, L. Horváthová, M. Allan, I. Štich, *Phys. Chem. Chem. Phys.* **2011**, *13*, 20939–45.

[28] M. C. Per, K. A. Walker, S. P. Russo, *J. Chem. Theory Comput.* **2012**, *8*, 2255–2259.

[29] R. C. Clay III, M. A. Morales, *J. Chem. Phys.* **2015**, *142*, 234103.

[30] M. W. Schmidt, K. K. Baldridge, J. A. Boatz, S. T. Elbert, M. S. Gordon, J. H. Jensen, S. Koseki, N. Matsunaga, K. A. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery, *J. Comput. Chem.* **1993**, *14*, 1347–1363.

[31] M. Gordon, M. Schmidt in *Theory and Applications of Computational Chemistry: the first forty years*, C. Dykstra, G. Frenking, K. Kim, G. Scuseria (Eds.), Elsevier, Amsterdam, **2005**, pp. 1167–1189.

[32] M. Per, *CSIRO Quantum Monte Carlo software package*, **2017**.

[33] E. T. Swann, M. L. Coote, A. S. Barnard, M. C. Per, *Int. J. Quantum Chem.* **2017**, *117*, 1–7.

[34] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[35] M. Dolg, C. Filippi, *Private Communication* **2014**.

[36] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[37] J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 084102.

[38] S. R. Langhoff, E. R. Davidson, *Int. J. Quant. Chem.* **1974**, *8*, 61–72.

[39] J. A. Pople, R. Seeger, R. Krishnan, *Int. J. Quant. Chem.* **1977**, *12*, 149–163.

[40] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **2012**, *8*, 2181–2188.

[41] M. J. Deible, M. Kessler, K. E. Gasperich, K. D. Jordan, *J. Chem. Phys.* **2015**, *143*, 084116.

[42] C. Filippi, M. Zaccheddu, F. Buda, *J. Chem. Theory Comput.* **2009**, *5*, 2074–2087.

[43] O. Valsson, C. Filippi, *J. Chem. Theory Comput.* **2010**, *6*, 1275–1292.

[44] R. Guareschi, C. Filippi, *J. Chem. Theory Comput.* **2013**, *9*, 5513–5525.

[45] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, R. G. Hennig, *Phys. Rev. Lett.* **2007**, *98*, 110201.

[46] P. Seth, P. L. Ríos, R. J. Needs, *J. Chem. Phys.* **2011**, *134*, 084105.

[47] K. Hongo, R. Y. O. Maezono, *Int. J. Quantum Chem.* **2011**, *112*, 1243–1255.

[48] W. Purwanto, S. Zhang, H. Krakauer, *J. Chem. Phys.* **2015**, *142*, 064302.

Chapter 8

# Bias-free chemical test sets

## 8.1 Introduction

Benchmarking studies evaluate the accuracy of computational procedures. It is a well-established practice in computational chemistry and provides invaluable information for users of quantum chemical methods. Systematic benchmarking demonstrates the robustness of a method, identifies systems where it might fail and validates the implementation of algorithms within a new program. This thesis has made thorough use of these benchmarking protocols and published test sets to provide a well-rounded description of diffusion Monte Carlo (DMC) performance for a range of problems, significantly expanding the benchmarking already in the literature (for a thorough review of QMC applications see Refs 1–4).

For benchmarking to be useful it must be systematic and repeatable. This has been achieved by benchmarking against test sets of molecules and there are now hundreds, if not thousands, of test sets available in the literature with validated reference data. The first formalised test sets were introduced by Pople et al. while developing their G(n) series of composite methods. These initial test sets were built using small molecules with well-established gas-phase experimental values. They were limited to ionisation potentials, electron affinities and atomisation energies which were calculated from heats of formation. Over time, more molecules have been added to create more diverse test sets with larger systems and more properties.[5–11] Truhlar et al. advanced this type of standardised benchmarking by introducing test sets covering a wider range of physiochemical properties. This included test sets for chemical energies (CE345), physical energies (PE39), chemical structural properties (CS20) and physical structural properties (PS47).[12,13] They also introduced test sets like Database/3,[14] and its successor Database/4,[15] as sets with good chemical diversity without being as large as G2/97[8,9] or G3/99.[10] These early test sets were built using molecules explicitly selected for the accuracy of their reference data

High quality experimental data is largely confined to values like heats of formation for relatively small molecules. The properties that can be derived from these experimental values is limited and the first test sets were severely restricted in terms of electronic complexity and chemical properties. It is important to benchmark these properties but they are not a reliable indicator of a methods performance for other properties. The use of atomisation energies in benchmarking can lead to a strong bias towards an accurate description of the free atoms relative to molecules.[16,17] For example, the popular PBE[18] functional performs poorly for the W4-08 test set[19] of atomization energies with a mean absolute deviation (MAD) of 13.0 kcal/mol but accurately predicts the isomerisation energies for the molecules in the ISO34 test set[20] (MAD of 1.8 kcal/mol).[21]

The development of high-level quantum chemical methods like CCSD(T) means test sets can be built for any conceivable property or system and a computational chemist now has access to a myriad of test sets. Often there are many test sets for one problem and it is left to the chemist's discretion to choose which one to use. The construction of these test sets has been guided by 'chemical intuition' and they are biased by a limited understanding of chemical space and an aversion to overly-complicated systems. The poor construction of these test sets means the relative and absolute performance of a method is extremely dependent on the test sets used. For example, if a chemist was studying non-covalent interactions they could test a method with a large test set like JSCH-2005,[22] made up of 143 non-covalent complexes including DNA base pairs, amino acid pairs and other model complexes. If they wanted a smaller test set they could look at S22,[22,23] a subset of 22 complexes from JSCH-2005, or A24,[24] a test set of 24 noncovalent complexes specifically chosen to cover a wide range of interactions, including hydrogen bonding, $\pi - \pi$ stacking and mixed electrostatics-dispersion and dispersion-dominated interactions. If they wanted to study larger complexes they could use the L7[25] test set of seven large complexes of 48 to 112 atoms. If they were interested in molecule-specific dispersion interactions they could use X40,[26] a test set of non-covalent interactions for 40 halogenated molecules, or a water clusters test set featuring clusters of 2 to 10 water molecules.[27] The outcomes of their benchmarking study would be determined by the test sets they used. The popular B3LYP method has a root mean squared error (RMSE) of 1.08 kcal/mol for the A24 test set but 7.00 kcal/mol for the JSCH test set.[28] The relative performance of methods can also change; M062X-D3 and B97-D perform worse than SAPT2 for the A24 set but better for the JSCH test set.[28]

Newer test sets like GMTKN30,[21] a database of test sets for general main group thermochemistry, kinetics and non-covalent interactions, are intentionally built with redundancies so that omitting one or two test sets will not affect the relative performance of methods. A clear trend in the construction of these test sets is the emphasis on size for diversity and robustness. With each subsequent update of a test set more data points have been added (i.e. Gn series[7–11] , S22 to S22x5[29] or S66[30]). These test sets are powerful tools but benchmarking has become a cumbersome task requiring thousands of calculations, GMTKN30 is made up of 30 test sets to cover a wide cross section of chemical space and features 1218 single point calculations.

These test sets are constructed to model real life problems but are strongly influenced by chemical intuition. Chemists select reactions they believe will be the most diverse or most representative of the problem they're studying. Grimme et al. proposed a novel approach in which the test set was generated 'mindlessly'.[31] Rather than letting chemical intuition guide the selection of molecules for their test set they created a generator requiring explicit specification of any constraints prior to the systems being generated. As an example they used two sets of constraints to create two test sets. MB08-931 was generated using elemental probabilities of atoms such that the probability of atoms Na to Cl is one third the probability of atoms Li to F which in turn is one third the probability of hydrogen. The other set, MB08-ORG used organic molecule elemental probability distributions. Decomposition energies were used in an attempt to represent real world applications with their artificial molecules. By stepping away from the standard approach of looking at stable and realistic molecules the authors hoped to find a 'robust' method capable of handling any type of system. These two test sets were combined to form MB08-165. While this test set removes biases and gives a better approximation of chemical space it still requires 180 single-point calculations.

There is a need in the field to create smaller, more bespoke test sets. More time could be spent developing new methods if only a handful of calculations were needed for a robust benchmarking. This is not a novel concept, Truhlar et al. created smaller test sets by selecting a subset of an existing test set such that the error measure of the subset showed the smallest deviation from the entire test set for a given cost.[14] Representative subsets were found for four of their test sets; Database/3 was reduced to two subsets of 6 atomisation energies and 6 barrier heights,[32] NHTBH38/04 was reduced to three subsets of 6 reactions for heavy atom transfer (HATBH6), nucleophilic substitution (NSBH6) and unimolecular association (UABH76).[33] Two test sets for metal-ligand bond energies and transition metal atomisation energies were reduced to just four entries (MLBE4/05[34] and TMAE4/05[35]). This method reduces the number of calculations required for benchmarking new methods but still requires an enormous number of calculations with existing methods. For example, when looking for a smaller test set of Database/3,[14] a test set of 109 atomisation energies and 44 barrier heights,[32] they used three error metrics (mean signed error (MSE), mean unsigned error (MUE) and root-mean-square deviation (RMSD)) from 80 different methods. This meant over 12000 calculations were required to reduce it down to 6 atomisation energies and 6 barrier heights.[14] It is also limited by the diversity of the existing subset, relying on the assumption that the initial test set provided an adequate representation of chemical space. An alternative solution is to step away from the existing test sets and remove the human element by using tools like multivariate statistics instead.

### 8.1.1 Multivariate statistical techniques

Multivariate statistical techniques are powerful analytical tools that can be used to study intrinsic patterns in highly complex data sets. Multiple variables are analysed simultaneously to reveal correlated patterns and structural relationships. This information can then be used to reduce dimensionality with minimal loss of information. Unsupervised pattern recognition techniques like $k$-means clustering and principal components analysis (PCA) are routinely used in a variety of fields, from material[36] and earth[37] sciences, to pharmaceuticals.[38] More recently, archetypal analysis has been used in conjunction with PCA and $k$-means clustering to characterise datasets of diamond nanoparticles and graphene nanoflakes.[39] It has also been used to summarise a small test set of corrosion inhibitors[40] A brief summary of these methods is provided below.

### Principal component analysis

Principal component analysis (PCA) is a data manipulation method that converts a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (principle components) using an orthogonal transformation.[41] The axes of this new coordinate system are oriented to account for maximum variation in the data set. It takes an $n \times p$ data matrix $\mathbf{X}$ (made up of $n$ observations of $p$ variables) and uses an orthogonal linear matrix transformation to express the original data as a linear combination of scores and loadings, described by:

$$X = t^1 p_1' + t^2 p_2' + ... + t^A p_A' + E = TP' + E \tag{8.1}$$

Where X is the original data matrix, $A$ is the total number of extracted principal components ($A \leq p$) and $E$ is the residual matrix. The new latent variables, $t$ scores, show how the objects relate to each other, while the $p$ loadings are the weights of each original variable and show their importance in seen in the scores.

### *K*-means clustering

$K$-means clustering is a vector quantization method that groups data points together to form clusters. Each cluster is characterised by a representative structure known as a prototype. Given a set of $n$ observations $(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)$ where each observation is a $d$-dimensional real vector, $k$-means clustering aims to partition the $n$ observations into $k(\leq n)$ sets $\mathbf{S} = \{S_1, S_2, ...S_k\}$ so as to minimise an objective function $J$, usually the squared distance of each point to its closest centroid $c_i$:

$$J = \sum_{i=1}^{k} \sum_{x \in S_i} ||\mathbf{x} - c_i||^2 \tag{8.2}$$

The prototypes are the means of the clusters. This objective function is used as a measure of the quality of the clustering; a lower score indicates better clustering and the prototypes are more representative of the data. $K$-means clustering requires the user to specify the desired number of clusters *a priori*. An arbitrary choice of too few clusters can miss important information but too many clusters can be redundant. The explained variance as a function of number of clusters can be used to choose an appropriate number of clusters.

Clustering is a useful tool to summarise the data. It means chemical space, as defined by the set of descriptors used, can be represented by a subset of prototypes. Instead of benchmarking with hundreds or thousands of calculations the test set is reduced to tens of molecules. Unfortunately it is restricted to data for which there is a notion of centre, and it cannot handle non-globular clusters, or clusters of different sizes and densities, also has trouble with outliers. It also has limited flexibility, each data point is restricted to one cluster, with no intermediates allowed between clusters.[42] Careful selection of descriptors can avoid these problems and ensure it is effective.

### Archetypal analysis

Archetypal analysis[43] (AA) is a relatively new statistical procedure that can reduce the benchmarking space down to just a handful of key structures.[39] AA is a matrix factorisation method that seeks to represent each individual in a multivariate data set as a linear combination of pure types. It is a similar technique to PCA but where in PCA the coefficients can be negative and their sum is not restricted to one in AA the archetypes form a convex hull of the dataset. The archetypes represent the 'pure types' in the data and mean the results are more easily interpretable.

The effectiveness of AA was first demonstrated with datasets of the shape of human heads and air pollution.[43] It has since been applied to a range of fields, identifying representative genotypes within the human population[44] and analysing subtle biological variance in global gene expression[45] and variation in phenotypes.[46] It has also been used for signal enhancement and feature extraction of IR image sequences[47] as well as extracting features from different high-dimensional datasets[42] and even to identify work preferences for software engineers.[48] The use of these extreme points for benchmarking has also been demonstrated.[49] More recently it was applied to datasets of diamond nanoparticles[39] and a dataset of corrosion inhibitors.[40]

The methodology is as follows. A multivariate dataset of $n$ observations of $m$ attributes is represented by an $n \times m$ matrix, $X$. Archetypal analysis seeks to find a $k \times m$ matrix $Z$ such that each data point can be represented as a mixture of the $k$ archetypes. This is achieved by minimising the residual sum of squares:

$$RSS = \sum_{i=1}^{n} \|X_i - \sum_{j=1}^{k} \alpha_{ij} Z_i\|^2 = \sum_{i=1}^{n} \|X_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{i=1}^{n} \beta_{ij} X_i\|^2 \tag{8.3}$$

It is subject to two conditions:

1. $\sum_{j=1}^{k} \alpha_{ij} = 1$, with $\alpha_{ij} \geq 0$ and $i = 1, ..., n$
2. $\sum_{i=1}^{n} \beta_{ij} = 1$, with $\beta_{ij} \geq 0$ and $i = 1, ..., k$

The first constraint requires the data to be *best approximated by convex combinations of the archetypes* whilst the second constraint implies that the archetypes are convex combinations of the data points. An **R** package, *archetypes* allows for easy implementation.[50] Whilst there are some similarities in PCA and AA there are a few key differences worth making note of. The number of archetypes must be selected initially but in PCA the decision can be made *a posteriori*. Also, the archetypes are pure data that will not necessarily be represented in the dataset i.e archetypes can be selected from the data even when they're not in the original dataset. In contrast, the principal components are the directions of a new coordinate system. Thus the number of principal axes and archetypes needed to represent the same features of a dataset will not necessarily be the same. Prototypes selected with $k$-means clustering are the most representative structures and archetypes can be thought of as the outliers. Combining the two sets together creates a robust and bias-free test set.

### 8.1.2 Descriptors

Multivariate statistics require numerical inputs but traditional molecular structure representations like Cartesian coordinates are not amenable to these methods. Descriptors need to be permutationally invariant but a permutation-invariant version of Cartesian coordinates scales combinatorially.[51, 52] Fortunately, the advancement of quantitative structure-activity relationship (QSAR) models for medicinal chemistry has led to the development of a wide range of descriptors.[53]

The simplest of these descriptors are one-dimensional and based on physical properties of molecules, such as molecular weight, shape-factors, estimated logP, surface area, dipole moment, HOMO-LUMO gap etc. Geometric descriptors include moments of inertia, shadow indices, molecular volume, molecular surface area and gravitation indices. Electrostatic descriptors include minimum and maximum partial charges or molecule or particular types of atoms, polarity parameter.[54] These whole-molecule, low-dimension descriptors convey very little information about the substructure of the molecule. Descriptors that represent the one-, two- or three-dimensional structure of a molecule are preferable.

The one-dimensional structure of a molecule is the molecular formula. Molecular fingerprints deconstruct a molecule into a bit string with a simple 'yes-no' check for the presence of a predefined set of functional groups. They represent extremely high-dimensional chemistry-space, varying between 150 to 200 bits for MDL applications, up to a few thousand for Tripos and Daylight applications and even up to millions for pharmacore fingerprints, depending on the functional groups of interest.[55] Molecular quantum numbers are similar to fingerprints but include counts rather than bits for simple structural features such as atom, bond and ring types.[56] The combination of one-dimensional descriptors like physical properties with fingerprints has been used for diverse subset selection.[57] Topological descriptors consider the two-dimensional structure of the molecule and reflect features like size, shape, symmetry, branching, connectivity and cyclicity.[58–61] Fingerprints and topological descriptors don't contain any information on the stereochemistry of the molecule and topological descriptors emphasise structure and connectivity over the charge or type of atom. Information is sparse for small molecules with few functional groups and little to no connectivity.

Structural descriptors represent the three-dimensional structure of the molecule and are invariant to rotations, translations and permutations of equivalent atoms. Versions of these descriptors have been developed based on graph-theory procedures[52] and electronic structure methods[62] and have been used to gauge dis(similarity) of crystalline, disordered and molecular compounds.[51] Other types of descriptors have been developed to predict properties without expensive electronic structure calculations and include Coulomb matrices,[63,64] bags of bonds,[65] 'symmetry functions'[66] descriptors. Other approaches, like the smooth overlap of atomic positions (SOAP),[67] start with descriptors designed to represent local atomic environments but then combine them for a global measure of structure similarity.

The Coulomb matrix $M$[63] is a matrix representation of the three-dimensional structure of a molecule. It was introduced to correlate chemical structure to accuracy of approximate quantum chemical methods and has been used for machine learning models.[63,68,69] The entries of the matrix are given by:

$$M_{IJ} = \begin{cases} 0.5 Z_I^{2.4} \text{ for } I = J \\ \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \text{ for } I \neq J \end{cases} \tag{8.4}$$

where $\mathbf{R_I}$ and $Z_I$ are the Cartesian coordinates and nuclear charge of atom $I$ and $J$ respectively. Diagonal elements encode a polynomial fit of atomic energies to nuclear charge and off-diagonal elements of the matrix correspond to the Coulomb repulsion between atoms $I$ and $J$. The total energy of a molecule is invariant under rotation, translation and symmetry operations such as mirror reflections and this is reflected in the Coulomb matrix. Different numbers of atoms in molecules result in different dimensionalities of the Coulomb matrices. This is easily overcome by introducing 'dummy' atoms with zero nuclear charge, padding out the matrix with zeroes so all matrices have size $d \times d$ where $d$ is the maximal number of atoms per molecule. Depending on the range of molecular size within the dataset this can lead to very sparse matrices. These types of numerical descriptors are based on the structure of a molecule and don't require any electronic structure calculations aside from an initial geometry optimisation. If we can represent a molecules position in chemical space as a set of numeric descriptors we can use unbiased methods to sample from this chemical space and come up with better test sets.

Machine learning (ML) systematically identifies similarities among data to make quantitative predictions.[70] In this chapter, the performance of three classes of descriptors (one, two and three- dimensional) is investigated. The functional relationship between the descriptors and electronic energy of a system using ML techniques. New, unbiased test sets are developed and then used to assess the performance of DFT and QMC methods. DFT is a popular alternative to traditional *ab initio* wavefunction theory[71] and offers a good compromise between accuracy and cost. The biggest limitation for DFT is the unknown form of the exchange-correlation functional and performance for DFT methods can be sporadic, depending on the approximation to this functional. Often the methods are parameterised for a given set of problems and will perform well for systems similar to the training set, but can fail catastrophically for other systems.[72] QMC methods are a stochastic alternative to *ab initio* and DFT methods and have been shown to be highly accurate for energetic[73,74] and structural properties.[75]

## 8.2 Computational methods

### 8.2.1 Data set

The data set was built using geometries and energies taken from the NIST Computational Chemistry Comparison and Benchmark Database (CCCBDB).[76] The CCCBDB is an online database of experimental and computed thermochemical data for a selected set of 1709 gas-phase atoms molecule, ranging in size from 1 to 26 atoms. It encompasses a vast array of properties including enthalpies of formation, entropies, heat corrections, geometries and atomic charges. It is an extensive collection of molecules and energies but is not comprehensive for all methods and basis sets. Due to the nature of the data in the CCCBDB not all molecules have energies for all methods. A compromise was made between using a highly accurate reference method and using as many data points as possible to sample a broad section of chemical space. Molecules with $\omega$B97XD/6-31G* geometries and energies were used for the data set and G4 energies were used for reference values. The final data set had 1499 molecules.

### 8.2.2 Descriptors

MACCS 166 keys[77] fingerprints and 19 connectivity-based topological descriptors (outlined in Table 8.1) were used, both generated using RDKIT,[78] an open source toolkit for cheminformatics. The ordering of atoms in the Coulomb matrix is undefined and it can be challenging to find equivalent representations of molecules for reactions. Three methods have been proposed to overcome this.[69] The first is the eigenspectrum representation (EVEC), where each matrix **M** is represented by a vector of sorted eigenvalues. This drastic dimensionality reduction (reducing a $d \times d$ matrix to a vector of length $d$) can result in loss of information and introduce noise.[79] The second is the sorted Coulomb matrix (SCMAT), where rows (and columns) of **M** are ordered by their norm, such that $||M_i|| \geq ||M_{i+1}||$. This method is sensitive to slight variations in atomic coordinates that change the magnitude of $M_i$. The third solution is a set of Coulomb matrices where the ordering of atoms has been varied for each matrix (RCMAT).[69] All three versions were used here. Correlated columns were removed for each set of descriptors so that only the column with the highest variance was retained for each inter-correlated pair ($R^2 > 0.9$). This left a total of 149 fingerprint descriptors, 13 topological descriptors and 435 Coulomb matrix descriptors.

### 8.2.3 Machine learning

The magnitude of the error, $\Delta E$, defined as the difference between G4 and $\omega$B97X-D3/6-31G* energies:

$$\Delta E = E_{\mathrm{G4}} - E_{\omega\mathrm{B97X\text{-}D3/6\text{-}31G^*}} \tag{8.5}$$

was correlated to the descriptors using decision tree (DT),[80] and random forest (RF)[81] ML techniques. Decision trees are a binary, rule-based modelling technique that typically

Table 8.1: Two-dimensional topological descriptors

| Descriptor | Description |
|---|---|
| BalabanJ[a] | An index for a hydrogen-suppressed graph of $n$ nodes and $m$ edges. It ignores H atoms and does not include charge or atomic number. |
| BertzCT[b] | Measure of molecules 'complexity', sum of two terms representing complexity of bonding and complexity of distribution of heteroatoms. |
| IPC[c] | Information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen suppressed graph of a molecule |
| Hall-Kier $\alpha$[d] | Parameter derived from the ratio of the *covalent radius* $R_i$ of the $i$th atom relative to the $sp^3$ carbon. Non-zero contributions to $\alpha$ are given by heteroatoms or carbon atoms with a valence state different from $sp^3$ |
| $\kappa$ 1-3[d] | Takes into account the different shape contribution of heteroatoms and hybridisation states |
| $\chi$ 0-4v, 0-4n[d] | Molecular connectivity; characterises structural attributes of molecule |

[a]Ref. 60, [b]Ref. 59, [c]Ref. 58, [d]Ref. 61

uses an attribute selection search to construct binary rules of different combinations of attributes. A decision tree model approximates the dependent variable as rudimentary decision rules based on the values of a number of attributes. The number and specific types of attributes can vary to suit the needs of the task. Despite their simplicity, decision trees have been shown to perform fairly well with the added value of ease interpretation given the number of rules are not very large.[80] Random forests (RF) are an ensemble learning method that train a multitude of decision trees and output the mode of the classes in classification or mean prediction in regression of the individual trees. RF improve over decision trees with respect to overfitting to the training set. The algorithm for inducing a random forest combines 'bagging' and the random selection of features in order to construct a collection of decision trees with controlled variance.[81]

These RF models were used to correlate the energy difference to the structural fingerprints. The training set was generated from 50% of the data set sampled around the prototype and archetype structures, whilst the remaining 50% of the data was used as a test set. The training set was used to calibrated the RF models whilst the test set was used to test the prediction ability of the model. The quality of the fit can be described by an $R^2$ value:

$$R^2 = \frac{\sum_{i=1}^{N}(Y_i - P_i)^2}{\sum_{i=1}^{N}(Y_i - \bar{P})^2} \tag{8.6}$$

Where $N$ is the number of molecules, $Y_i$ and $P_i$ are the machine learning predicted and actual energy difference for molecule $i$, respectively. The average energy different over all structures is given by $\bar{P}$. When computed on the training set, $R^2$ measures how well the model fits the simulated data. To check for the possibility of overfitting, a technique known as internal three-fold-out (TFO) cross-validation was applied. The training set was divided into three subsets. One was removed while the other two were used to fit the regression model. The resulting model was then compared against the simulation data for the left-out subset. This process was repeated until all the subsets have been validated against each other.

### 8.2.4 Atomisation energies

The new test sets were used to assess the performance of a variety of methods. Estimated CCSD(T)/CBS reference atomisation energies using the same $\omega$B97XD/6-31G* geometries taken the CCCBDB[76] were calculated using:

$$E_{\mathrm{CBS}}^{\mathrm{CCSD(T)}} = E_{\mathrm{CBS}}^{\mathrm{MP2}} + (E^{\mathrm{CCSD(T)}} - E^{\mathrm{MP2}})|_{\text{small basis set}} \tag{8.7}$$

This difference does not depend significantly on the basis set and is a good approximation to CCSD(T)/CBS.[82] $E_{\mathrm{CBS}}^{\mathrm{MP2}}$ was determined using the extrapolation scheme of Helgaker and coworkers[83] and cc-pVTZ and cc-pVQZ basis sets.[84] The small basis set difference was calculated at the cc-pVDZ level.[84] For larger molecules (**48**, **167**, **264**, **768**, **769**, **773**, **1433**, **1477** and **1482**) MP2 calculations are replaced with the approximate resolution of the identity MP2 (RI-MP2) method.[85–87] Absolute and relative RI-MP2 energies agree well with MP2 values.[88]

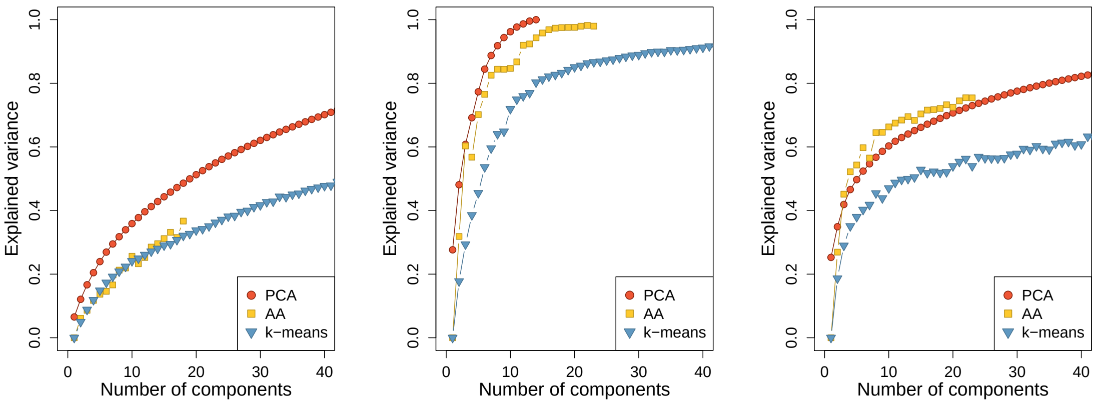
The accuracy of density functional theory (DFT) was investigated using 10 different exchange-correlation functionals. The types of functionals evaluated include local density approximation (SVWN[89,90]), generalised gradient approximation (GGA) functional (PBE[18]), meta-GGAs (TPSS,[91] M06L[92]), hybrid GGAs (B3LYP,[93,94] $\omega$B97XD,[95] PBE0[96,97]), hybrid meta-GGAs (B1B95,[98] MPW1B95[99]) and double-hybrid GGAs (B2P-LYP[100] and mPW2PLYP[101]). DFT calculations use the 6-311+G(3df,2p) basis set. The performance of MP2 and the G4(MP2)-6X composite method[102] was also compared. G4(MP2) is widely accepted as a highly accurate method close to chemical accuracy[72] and G4(MP2)-6X[102] is an approximation to G4(MP2) with improved speed for larger molecules. All DFT calculations including the starting orbitals for DMC were performed using Gaussian.[103] CCSD(T) and MP2 calculations were performed using Molpro.[104]

Diffusion Monte Carlo (DMC) calculations were performed using the CMQMC code.[105] Slater-Jastrow trial wavefunctions were used:

$$\Psi_T = e^J D^{\uparrow} D^{\downarrow} \tag{8.8}$$

where $D^{\uparrow}D^{\downarrow}$ are Slater determinants constructed from single-particle orbitals, taken from B3LYP calculations. $J$ is a two-term Jastrow factor containing explicit electron correlation terms.[106] The free parameters in the Jastrow factor were optimised by minimising the total energy at the variational Monte Carlo (VMC) level. Pseudopotentials are routinely used in DMC calculations to replace the chemically-inert core electrons and reduce local energy fluctuations, speeding up the calculations. Energy consistent Burkatzki-Filippi-Dolg (BFD) pseudopotentials with the associated triple-zeta basis set and improved H-atom potential[107, 108] were used here with imaginary timestep sizes $\tau = 0.01$. Nonlocal pseudopotentials were treated beyond the locality approximation in DMC using the size-consistent T-moves approach.[109]

## 8.3 Results and discussion



(a) (b) (c)

Figure 8.1: Explained variance of the entire data set as a function of number of components for principal component analysis (PCA), archetypal analysis (AA) and $k$-means clustering of a) fingerprint, b) topological and c) Coulomb matrix descriptors.

### 8.3.1 Principal component analysis

Principal component analysis (PCA) was used to explore the variance within the descriptors. Figure 8.1 shows the explained variance as a function of number of components for the three classes of descriptors. The fingerprints (FP) descriptors are essentially independent from each other; each variable only accounts for the presence or absence of a particular functional group. Not surprisingly, these descriptors have the lowest explained variance with the first, second and third principal components (PCs) accounting for only 6.6%, 5.6% and 4.6% of the variance respectively, for a combined total of only 16.7%. In comparison the PCs for the topological (TOPO) and Coulomb matrix (CMAT) descriptors recover a much higher proportion of the variance. The first, second and third PCs for TOPO explain 28.8%, 18.9% and 13.5% of the variance respectively (combined total of 61.2%). In the case of CMAT descriptors the first, second and third PCs explain 25.3%, 9.7% and 7.0% of the variance respectively for a combined total of 42.0%. The CCCBDB is a diverse data set and the fingerprints of the molecules are especially unique fingerprints. PCA is an orthogonal transformation and will not recover much variance for uncorrelated variables. In contrast, the TOPO and CMAT descriptors are more easily generalised by the PCs. There are fewer TOPO descriptors so the PC's will account for more variance. Despite having more variables (435 compared to 149) the PCs for CMAT recover more variance than for FP, suggesting the variables are more correlated.

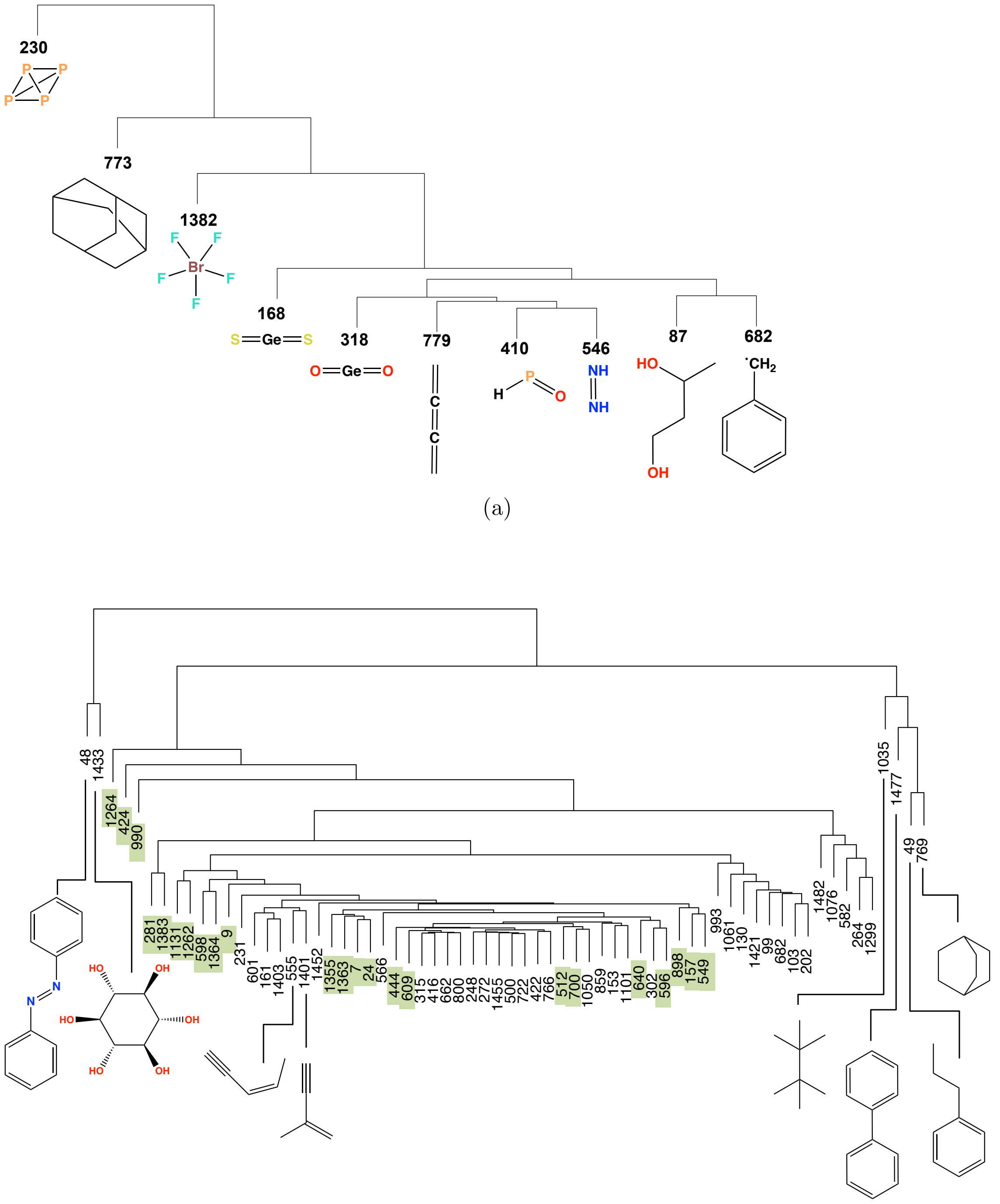
### 8.3.2 *K*-means clustering

To characterise the diversity of the data set, $k$-means clustering was used to identify groups (clusters) of molecules that share structural or functional similarity based on the three types of descriptors. This process has previously been used for a set of nanoparticles[39] and a test set of corrosion inhibitors.[40] For each cluster the molecule with the shortest Euclidean distance to the cluster centroid was selected as the cluster prototype. These were used for further analysis.

In $k$-means clustering the number of clusters is set *a priori*. A choice of too few clusters can miss important information whereas too many clusters becomes redundant. The optimum number of clusters was selected by analysing the amount of explained variance as a function of number of clusters (see Figure 8.1). To keep the number of clusters to a minimum clusters were selected to explain 70% of the explained variance (including more clusters would increase the explained variance, as required). This gave 10 TOPO prototypes and 66 CMAT prototypes. The TOPO descriptors describe a chemical space that is more easily summarised by a select few prototypes whereas the CMAT space is quite diverse and spread out. Clustering the fingerprint descriptors recovers very little information, a consequence of the diversity of the data set. Clustering recovers less variance for all three classes of descriptors when compared to PCA.

Hierarchical clustering was used to further analyse these prototype structures, measuring the similarity within the descriptors at different levels. The resulting dendrograms are shown in Figure 8.2a for TOPO descriptors and Figure 8.2b for CMAT descriptors. The TOPO prototypes are quite diverse and are depicted by nine hierarchical branches for the 10 prototypes. A broad range of functional groups is captured in this small set. Topological descriptors tend to emphasise physical structure or connectivity information over charge and this is seen in the grouping of the branches. Linear molecules ($GeO_2$ (**318**), $CH_2CCCH_2$ (**779**), $GeS_2$ (**168**)) and cage-like molecules (the $P_4$ cluster (**230**) and adamantane (**773**)) belong to the same branches. From a chemical perspective the separation is not intuitive. For example $GeO_2$ (**318**) is more similar to 1,2,3-butatriene (**779**) than $GeS_2$ (**168**). The species selected as prototypes include molecules that are challenging from a theoretical perspective (i.e. $P_4$[10,17] and diazene $N_2H_2$[110–129]). These statistically relevant molecules are unlikely to be those intuitively selected by chemists but are necessary to provide cases that challenge the accuracy and suitability of quantum chemical methods.

CMAT descriptors required 66 prototypes to account for 70% of the explained variance and the resulting dendrogram is crowded. By virtue of having more points there is a broader range of atoms and functional groups represented in the CMAT prototypes than



(a)

(b)

Figure 8.2: Hierarchical clustering dendrograms of the cluster prototypes of (a) TOPO and (b) CMAT descriptors. Molecules with at least one halogen atom (F, Cl or Br) are highlighted in green.

the TOPO prototypes and molecules range in size from two to twenty-six atoms. CMAT descriptors are calculated using nuclear charge and the three-dimensional coordinates of the molecule, and this is reflected in the dendrogram. Highlighted in green are all molecules with at least one halogen atom (F, Cl or Br) and in general they share the same branches. Organic molecules are also grouped together; for example, 3-penten-1-yne (**555**) and 1-buten-3-yne (**1401**) both have a carbon double and triple bond and share the same branch. Despite this, not all grouping is intuitive. For example the upper-left branch has *myo*-inositol (**1433**) with 6 OH groups is next to azobenzene (**48**) with two benzene rings and a nitrogen double bond. The upper-right branch shows four molecules with very different structures, despite containing only C and H atoms.

### 8.3.3 Archetypal analysis

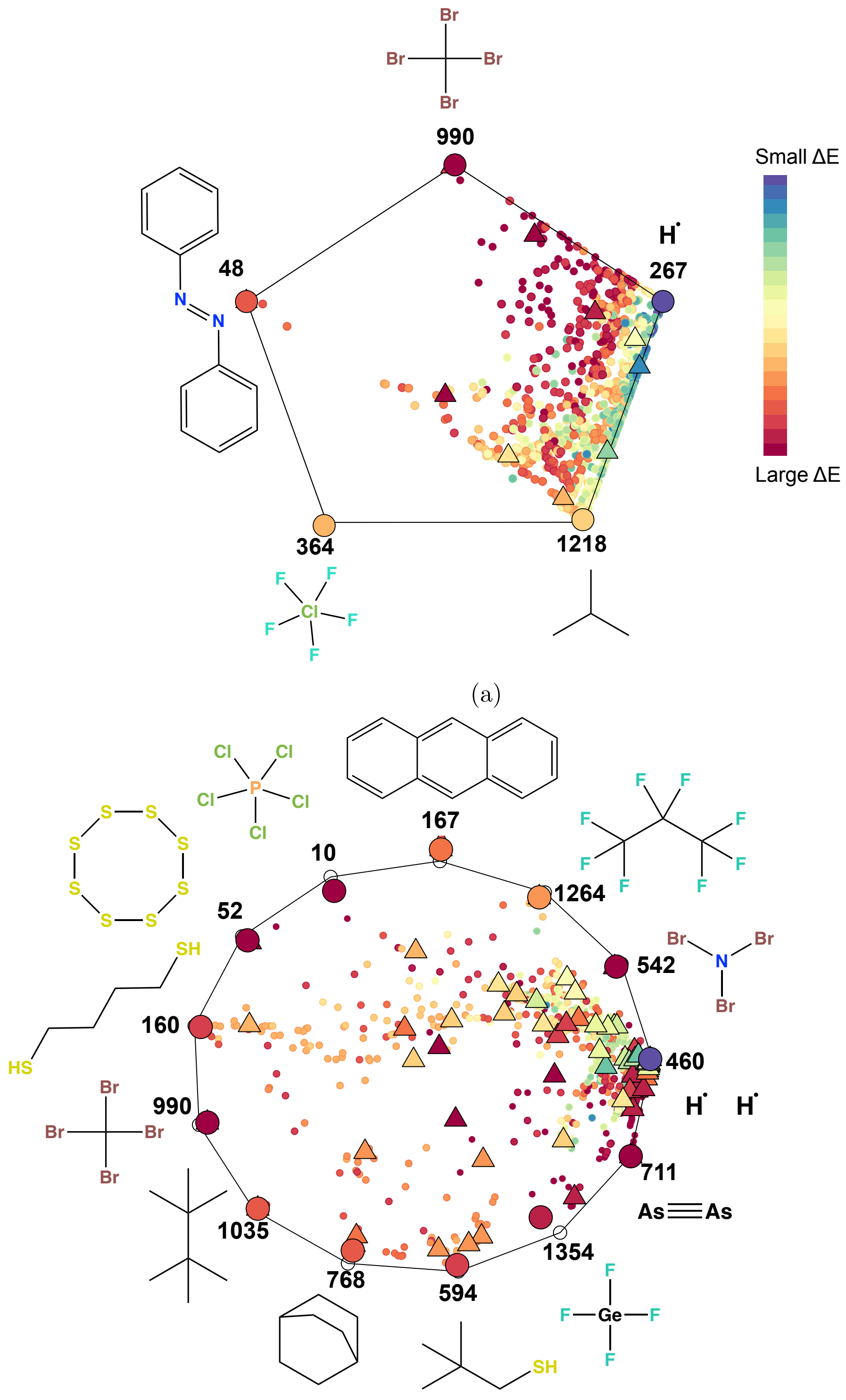
As stated above, $k$-means clustering only identifies structures that characterise the main groups of molecules that share similar properties. It fails to identify structures that represent unique combinations of features and lie on the complex hull of the data set. Archetypal analysis is needed to find the outliers. Archetypes were identified as molecules with the shortest Euclidean distance to each 'pure type'. The explained variance as a function of number of archetypes is shown in Figure 8.1. To keep the analysis of archetypes consistent with the prototypes, archetypes were selected to explain 70% of the variance, resulting in 5 TOPO archetypes and 13 CMAT archetypes. Fingerprint descriptors are poorly summarised by archetypes since they represent a high-dimensional chemical space that cannot be encapsulated by a convex hull.

Since each molecule can be described as a linear combination of archetypes, the data can be plotted on a two-dimensional simplex plots, as shown in Figure 8.3. The edges are defined by the archetypes and all other molecules are scattered at relative positions given by the contributions of each archetype towards each molecule. Points are coloured according to the magnitude of $\Delta E$, defined as the difference between G4 and $\omega$B97X-D3/6-31G$^*$ energies. Smaller errors are shown in indigo and larger errors in red.

The TOPO and CMAT archetype molecules cover a broad range of molecule sizes; the largest molecule for both sets of descriptors has fourteen heavy atoms (azobenzene (**48**) for TOPO and anthracene (**167**) for CMAT). Both descriptors include a small, hydrogen-containing archetype (the hydrogen radical (**267**) for TOPO and 2H (**460**) for CMAT) and for both descriptors this is the most important archetype. Using descriptors based entirely on the structural information of the molecules has found archetypes that distribute molecules according to their error, or rather, how challenging they are for computational methods. In general, the archetype molecules have relatively large errors, supporting the idea that archetypes are more unique and consequently more challenging molecules.

### 8.3.4 New test sets

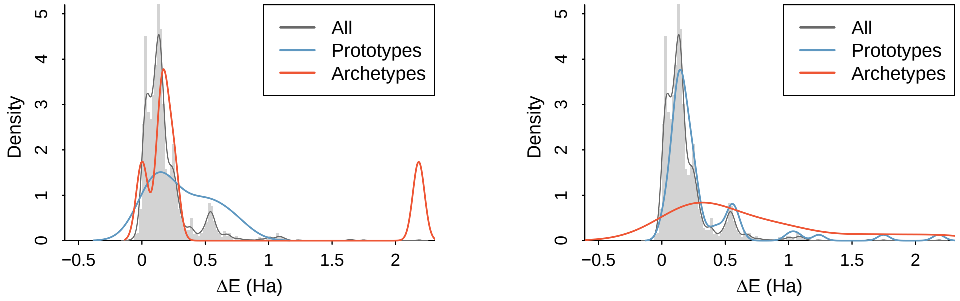
CMAT and TOPO descriptors emphasise different attributes of a molecule; topological descriptors highlight connectivity but Coulomb matrices emphasize charge. This can be very



(a)

(b)

Figure 8.3: Simplex plots of the (a) TOPO and (b) CMAT descriptors. Archetypes (large circles) are located toward the edges of the regular polygons and molecules are scattered as projections of the archetypes in the simplex. Prototypes are shown as triangles. Points are coloured according to the magnitude of ΔE.



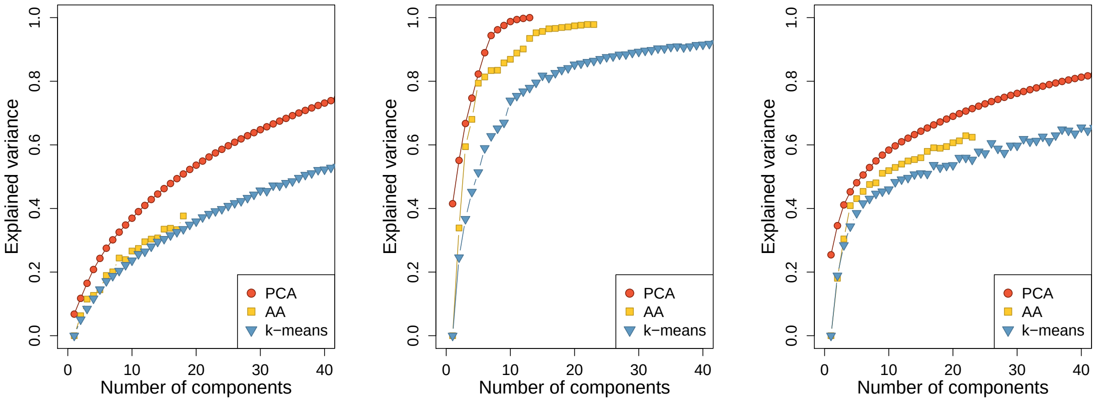
(a)

(b)

Figure 8.4: Histogram of the errors for the entire data set and new test sets based on (a) topological and (b) Coulomb matrix descriptors. The error (ΔE) is defined as the difference between the $\omega$B97XD/6-31G* and G4 energies, measured in Hartrees (Ha).

useful when deciding what types of descriptors to use when seeking a new test set, depending on future research plans. It is not unexpected that there is little overlap between their prototypes and archetypes, only $CBr_4$ (**990**) shows up as an archetype for both, the benzyl radical (**682**) is a prototype for both and azobenzene (**48**) is a TOPO archetype but a CMAT prototype. The broad distribution of the molecules in CMAT space is evidenced by the overlap of archetypes and prototypes. Azobenzene (**990**), 2,2,3,3-tetramethylbutane (**1035**) and octafluoropropane (**1264**) and are both CMAT prototypes and archetypes and it can be seen in the simplex plot that there are prototype structures lying on the archetype convex hull.

Once identified, the archetypes and prototypes can be combined to form a reliable, statistically significant and diverse test set for quantum chemical methods. The distribution of errors was compared, to assess the performance of CMAT and TOPO archetypes and prototypes in representing the database as a whole. Results are shown in Figure 8.4. The error distribution of the entire data set is shown in grey. The TOPO prototypes (light red) are more widely distributed than the CMAT prototypes (light blue) and the broad distribution of the CMAT prototypes more closely resembles that of the whole data set. In contrast, the TOPO archetypes (dark red) encompass molecules with both small and large errors. The CMAT archetypes are more widely distributed than the archetypes, but a skew towards larger errors is seen. For both descriptors the archetypes have larger errors than the prototypes. This again shows the prototypes are more representative structures whereas the archetypes (especially in the case of TOPO archetypes) capture molecules at the extreme ends of the distribution. These archetypes and prototypes were found using only structural information, yet the molecules identified as archetypes and prototypes reflect the energies of the whole data set. As seen in the dendrogram and simplex plots, these structures aren't the ones that would be intuitively selected, and might even be



(a) (b) (c)

Figure 8.5: Explained variance of the organic subset, as a function of number of components for principal component analysis (PCA), archetypal analysis (AA) and $k$-means clustering of a) fingerprint, b) topological and c) Coulomb matrix descriptors.

selected against for being problematic systems (i.e. $P_4$ and diazene $N_2H_2$). By removing human biases a diverse and statistically significant subset of molecules has been identified. These new test sets are labelled CMolsC-1[130] (CMAT test set) and CMolsT-1[131] (TOPO test set) and are available online.

### 8.3.5 Organic molecules

The CCCBDB is an especially diverse data set but most applications using test sets focus on a specific property or functional group. To demonstrate the suitability of this method for all properties, a smaller test set containing all molecules with at least one C and one H atom and no transition metals was built, to mimic an organic test set that would normally be used by chemists. This resulted in 700 molecules. Given the number of molecules and the number of CMAT descriptors, the CMAT descriptors were limited to only include ones that had values for more than 25% of the data set. This resulted in 194 variables, compared to 435 which were used for the entire data set. Once correlated columns were removed there were 13 TOPO descriptors and 143 fingerprint descriptors.

The same PCA, $k$-means clustering and AA as described above was performed and results are shown in Figure 8.5. Archetypes and prototypes were selected for 70% explained variance again, resulting in 5 TOPO archetypes, 11 TOPO prototypes and 62 CMAT prototypes. The CMAT archetypes did not recover 70% of the variance and were omitted from further study. The dendrograms for the CMAT and TOPO prototypes are shown in Figure 8.6. The TOPO prototypes for the organic subset are quite diverse, with eight distinct branches for the eleven molecules. They cover a broad range of functional groups with an emphasis on functional groups containing oxygen. For the CMAT prototypes there is a similar grouping of halogen-containing molecules, as was seen for the larger data set.

The simplex plot for the TOPO archetypes is shown in Figure 8.7. The TOPO archetypes again include the smallest molecule in the set (in this case the CH radical (**832**)) and one of the largest (adamantane (**773**)). Molecules are more evenly distributed in the simplex plot compared to Figure 8.3a but again molecules are distributed according to $\Delta$E, despite no energetic information being included in the descriptors. The CMAT descriptor was not suitable for this smaller data set as there was too much variance in the data set in CMAT space to be encapsulated by the archetypes. Despite this, the distribution of molecules within the simplex plot for the TOPO descriptors shows it is suitable for this smaller set.

The histogram of the errors of the archetypes and prototypes for the organic subset is shown in Figure 8.8. The distribution of archetypes and prototypes is very similar to those found for the whole test set. The TOPO archetypes capture molecules at the extremes of the distribution again, whereas the prototypes for both descriptors better represent the entire subset. This reiterates the suitability of prototypes and archetypes for forming a reliable and statistically significant data set. In cases where the CMAT descriptor is too diverse, the TOPO descriptor is still appropriate. These new test sets are labelled CMolsC-org[132] (CMAT organic test set) and CMolsT-org[133] (TOPO organic test set) and are available online. The use of this method has been demonstrated for organic molecules but the same approach could be used to find test sets of more exotic species.

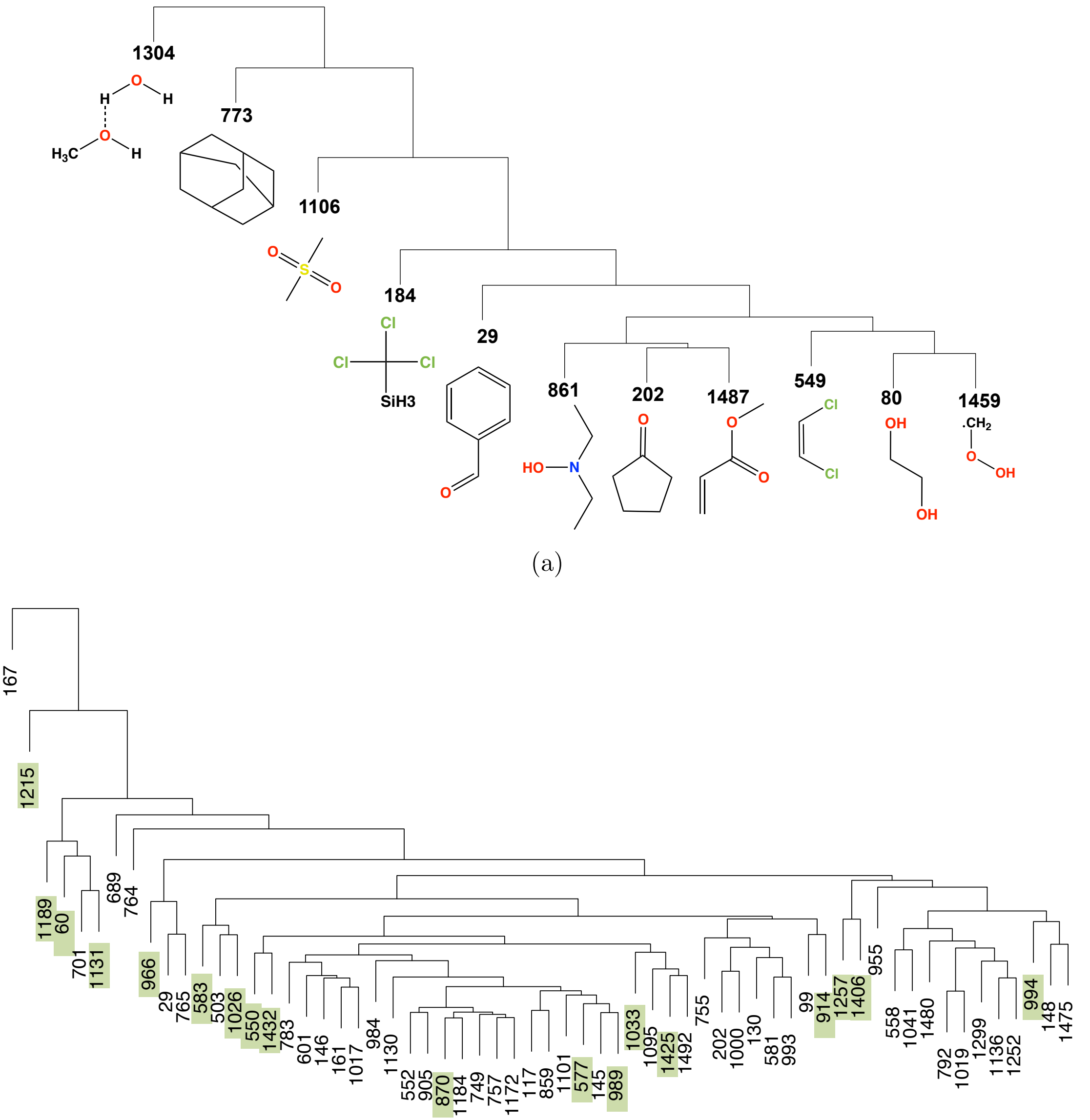
### 8.3.6 Machine learning

Regression machine learning (ML) models were calibrated to predict the difference between the $\omega$B97XD/6-31G$^*$ and G4 energies ($\Delta$E) using the TOPO and CMAT descriptors. Random forest (RF) models trained with 50% of the data set sampled around archetype and prototype structures are shown in Table 8.2. The optimal RF models exhibit cross-validation correlation coefficient ($R^2_{CV}$) of ~0.97 and ~0.74 for TOPO and CMAT descriptors, respectively.

Scatter plots shown in Figure 8.9 illustrate the better transferability capacity of the CMAT ML model, where the accuracy of the predictions of $\Delta$E values of the test set structures is >0.96, whilst the TOPO correlation is 0.2 units lower. The tridimensional information of the CMAT descriptors is more suitable for describing the energy difference between the two methods, which suggest that the set of archetype and prototype structures selected by this descriptor type are a better representation of the data set in the particular context of $\omega$B97XD/6-31G$^*$ and G4 energy gaps.

### 8.3.7 Atomisation energies

Total atomic energies are readily computed but in practice they are rarely reported on. Energy differences are of more chemical interest but accurate values rely on good cancellation of errors. Electron correlation energy makes up a small part of the total energy but becomes significant for energy differences. Atomisation test sets like the G(n) series[5,7–11]



(a)

(b)

Figure 8.6: Hierarchical clustering dendrograms of the cluster prototypes of (a) TOPO and (b) CMAT descriptors for the organic subset. Molecules with at least one halogen atom (F, Cl or Br) are highlighted in green.

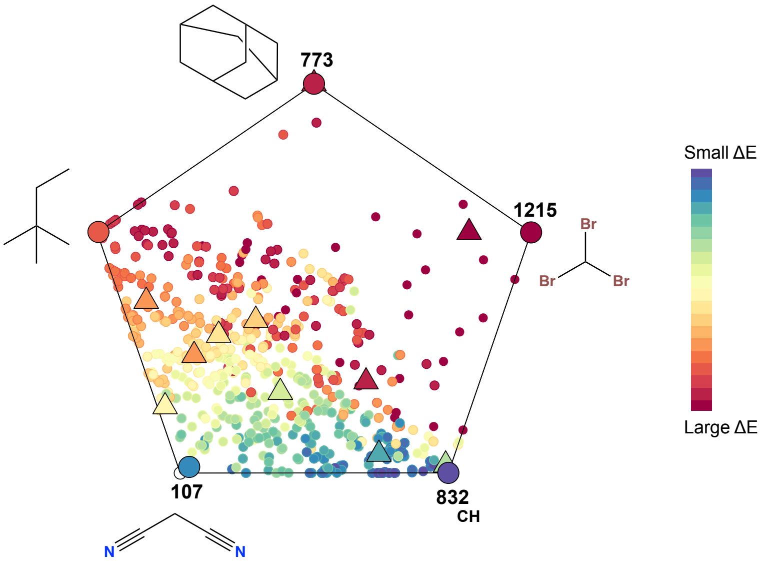


Figure 8.7: Simplex plot of the TOPO archetypes for the organic subset. Archetypes are located toward the edges of the regular polygons and the molecules are scattered as projections of the archetypes in the simplex. The large circles represent the archetypes and the triangles represent the prototypes. Points are coloured according to the magnitude of ΔE.

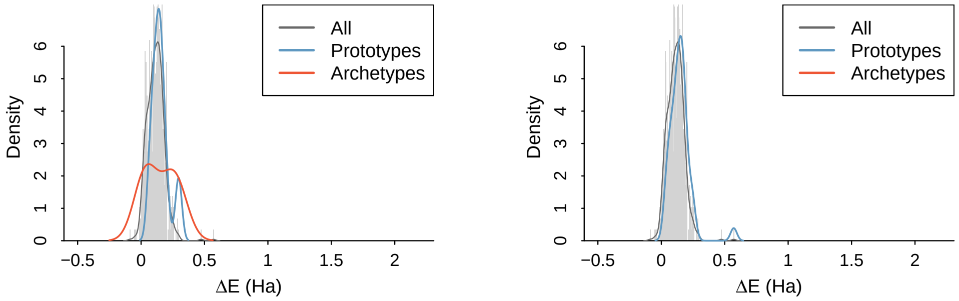
Table 8.2: Details of the optimum regression models of ΔE (defined as the difference between the $\omega$B97XD/6-31G$^*$ and G4 energies). $R_{\mathrm{CV}}$ is the cross-validation coefficient, $STD_{\mathrm{CV}}$ and $STD_{\mathrm{Test}}$ are the standard deviations of the cross-validation set and entire set, respectively.

| Model | Descriptor | Parameters$^{\mathrm{a}}$ | $R_{\mathrm{CV}}$ | $STD_{\mathrm{CV}}$ (Ha) | $R_{\mathrm{Test}}$ | $STD_{\mathrm{Test}}$ (Ha) |
|---|---|---|---|---|---|---|
| RF | CMAT | $n = 25,\ s = 2$ | 0.975 | 0.032 | 0.964 | 0.041 |
| | TOPO | $n = 25,\ s = 2$ | 0.739 | 0.105 | 0.733 | 0.104 |

$n$ is the number of estimator trees and $s$ is the minimum sample split in the RF model.

are predominantly made up of small, stable molecules with well-defined experimental values. These new CMolsC-1 and CMolsT-1 test sets provide a more robust test of a methods performance. Here they are used to test the performance of a selection of DFT methods as well as MP2, G4(MP2)-6X and DMC. Results are reported in Table 8.3.

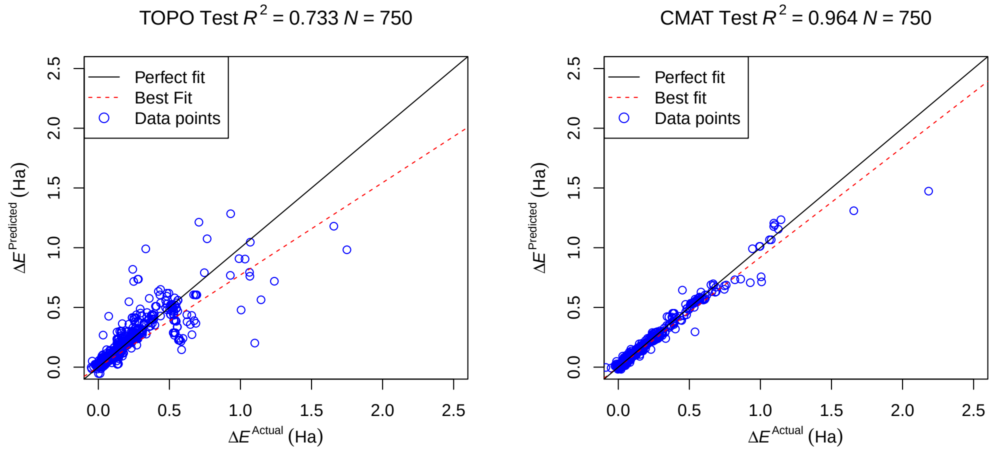
For the majority of methods the most challenging set of molecules is the CMolsC-1 test set. The archetypes are generally more challenging structures than prototypes, with larger MAD values for CMolsC-1A compared to CMolsC-1P. Results are not as consistent for CMolsT-1, functionals like M06L and B1B95 have larger errors for prototypes compared to archetypes but for $\omega$B97XD and the double-hybrid functionals B2PLYP and mPW2PLYP the archetypes are more challenging. Topological descriptors align more with our understanding of chemical intuition compared to the Coulomb matrix descriptors, and the CMolsT-1 test set highlights how parameterisation and the treatment of the exchange-correlation functional affect DFT performance. In contrast, DMC is not a parameterised method and the errors are consistent for archetypes and prototypes from both sets.



(a)

(b)

Figure 8.8: Histogram of the errors for the organic sub set and new test sets based on a) topological and b) Coulomb matrix descriptors. The error (ΔE) is defined as the difference between the ωB97XD/6-31G* and G4 energies, measured in Hartrees (Ha).



(a)

(b)

Figure 8.9: Scatter plot of ΔE predictions for the test set consisting of 50% of the data set using TOPO (a) and CMAT (b) descriptors. $\Delta E^{\text{Actual}}$ refers to the values calculated by energy differences between the two methods and $\Delta E^{\text{Predicted}}$ corresponds to the RF predictions.

Table 8.3: Mean absolute deviation (MAD) of DMC results from est. CCSD(T)/CBS reference values (kcal/mol) for atomisation energies of the molecules in the CMolsC-1 and CMolsT-1 test sets.

| | CMolsC-1A | CMolsC-1P | CMolsT-1A | CMolsT-1P | Overall |
|---|---|---|---|---|---|
| SVWN | 196.4 | 150.8 | 191.4 | 156.2 | 159.8 |
| PBE | 25.1 | 23.9 | 36.4 | 26.2 | 25.0 |
| TPSS | 4.6 | 6.3 | 6.4 | 9.0 | 6.3 |
| M06L | 10.9 | 8.3 | 7.4 | 11.0 | 8.9 |
| B3LYP | 22.1 | 11.1 | 12.7 | 12.4 | 12.8 |
| PBE0 | 8.8 | 8.5 | 9.9 | 9.9 | 8.8 |
| $\omega$B97XD | 8.2 | 5.6 | 7.4 | 5.4 | 6.0 |
| B1B95 | 9.0 | 6.6 | 5.7 | 9.7 | 7.2 |
| B2PLYP | 19.7 | 11.3 | 13.0 | 11.0 | 12.5 |
| mPW2PLYP | 18.0 | 10.5 | 13.0 | 10.9 | 11.7 |
| G4 | 4.0 | 2.9 | 4.0 | 4.1 | 3.2 |
| MP2 | 21.0 | 12.5 | 9.5 | 12.9 | 13.6 |
| DMC | 10.0 | 6.3 | 11.1 | 8.4 | 7.3 |

Density functionals can be grouped together according to their treatment of the exchange-correlation term in what is known as a 'Jacob's Ladder'.[134] There is significant variation within the different classes of methods. LDA functionals such as SVWN form the lowest rung of the Jacob's ladder and have the largest errors with an overall MAD of 159.8 kcal/mol. The double-hybrid functionals B2PLYP and mPW2PLYP lie at the top of the Jacob's ladder, but have comparable performance with lower level methods like B3LYP. They have been shown to be more basis-set dependent than hybrid functionals[135] but these two test sets highlight some potential shortcomings compared to other methods. The best DFT method is $\omega$B97XD, with an overall error of 6.0 kcal/mol. $\omega$B97XD is a long-range corrected functional and includes 100% Hartee-Fock exchange for long-range electron-electron interactions. It is known to perform well for challenging systems like the DC9 test set[135] and performs better than other functionals for non-covalent systems.[95] G4(MP2)-6X has the lowest MAD of all methods studied here. It is a composite method that uses an additivity scheme based on *ab initio* wavefunction calculations but has been parameterised on a large test set. DMC has an overall MAD of 7.3 kcal/mol and performs better than the double-hybrid functionals but worse than $\omega$B97XD and TPSS. These two new test sets include more exotic molecules with second and third-row atoms or hypervalent bonding that are generally more challenging for DMC methods. Unlike DFT methods, DMC is systematically improvable by including more determinants. In a previous study of G2 atomisation energies using a multi-determinant trial wavefunction reduced the error from 2.1 kcal/mol to 1.2 kcal/mol.[74] The same approach could reduce the error here.

## 8.4 Summary

The development of new quantum chemical methods requires extensive testing to demonstrate robustness and to identify potential weaknesses and shortcomings. Numerous databases are available to test and standardise this process. However, selecting robust structural test sets for specific problems is vulnerable to human bias and intuition.

The archetypes and prototypes found here constitute diverse subsets of molecular structures that can serve as reference to build robust calibration and test sets. Fingerprint descriptors, used in material science, are insufficient to describe chemical space, but both Coulomb matrix and topological descriptors were useful to identify the ideal subset of structures. Archetypal analysis found the molecules with the largest errors with reference to G4 methods without calculating the energy whilst the prototypes are a good representation of the data set as a whole. When tested on a smaller set of organic molecules, the archetypes and prototypes exhibit similar results. By calculating atomisation energies for the CMolsC-1 and CMolsT-1 test set it was shown DFT performance depends on the parameterisation and treatment of the exchange-correlation functional but DMC is a consistent method for challenging problems. These conclusions are difficult to draw from other test sets as they are biased by the same chemical intuition used to build the training sets.

Prototypes and archetypes can produce test sets without need of assumptions, prior knowledge or a completed body of work, as shown with both the general case and the specific case of organic molecules. These methods can be used to create small, robust test sets that avoid the burden on computational resources that is presented by the existing biased test sets. The test sets presented here are useful for main-group and organic chemistry but these methods can be extended to create test sets for virtually any property or type of system. Test sets for more exotic molecules or metallic systems could be constructed in the same manner, using descriptors that favour these exotic features. High-cost electronic structure calculations are not required to identify potentially large errors or challenging structures where popular methods might fail; just a sound statistical analysis of the chemical diversity of the structures is required.

## 8.5 References

[1] T. Pang, *Am. J. Phys.* **2014**, *82*, 980–988.

[2] A. Lüchow, *Wires Comput. Mol. Sci.* **2011**, *1*, 388–402.

[3] R. J. Needs, M. D. Towler, N. D. Drummond, P. López Ríos, *J. Phys. Condens. Matter* **2010**, *22*, 023201.

[4] B. Austin, D. Y. Zubarev, W. A. Lester, *Chem. Rev.* **2012**, *112*, 263–88.

[5] L. A. Curtiss, C. Jones, G. W. Trucks, K. Raghavachari, J. A. Pople, *J. Chem. Phys.* **1990**, *93*, 2537.

[6] J. A. Pople, M. Head-Gordon, D. J. Fox, K. Raghavachari, L. A. Curtiss, *J. Chem. Phys.* **1989**, *90*, 5622.

[7] L. A. Curtiss, K. Raghavachari, *J. Chem. Phys.* **1991**, *94*, 7221–7230.

[8] L. A. Curtiss, K. Raghavachari, P. C. Redfern, J. A. Pople, *J. Chem. Phys.* **1997**, *106*, 1063–1079.

[9] L. A. Curtiss, P. C. Redfern, K. Raghavachari, J. A. Pople, *J. Chem. Phys.* **1998**, *109*, 42–55.

[10] L. A. Curtiss, K. Raghavachari, P. C. Redfern, J. A. Pople, *J. Chem. Phys.* **2000**, *112*, 7374–7383.

[11] L. A. Curtiss, P. C. Redfern, K. Raghavachari, *J. Chem. Phys.* **2005**, *123*, 124107.

[12] Y. Zhao, D. G. Truhlar, *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[13] R. Peverati, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2012**, *14*, 13171.

[14] B. J. Lynch, D. G. Truhlar, *J. Phys. Chem. A* **2003**, *107*, 3898–3906.

[15] B. J. Lynch, Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 1643–1649.

[16] G. I. Csonka, A. Ruzsinszky, J. Tao, J. P. Perdew, *Int. J. Quantum Chem.* **2005**, *101*, 505–511.

[17] S. Grimme, *J. Phys. Chem. A* **2005**, *109*, 3067–3077.

[18] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[19] A. Karton, B. Ruscic, J. M. L. Martin, *J. Mol. Struct. (Theochem)* **2007**, *811*, 345–353.

[20] S. Grimme, M. Steinmetz, M. Korth, *J. Org. Chem.* **2007**, *72*, 2118–2126.

[21] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

[22] P. Jurečka, J. Sponer, J. Cerný, P. Hobza, *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

[23] T. Takatani, E. G. Hohenstein, M. Malagoli, M. S. Marshall, C. D. Sherrill, *J. Chem. Phys.* **2010**, *132*, 144104.

[24] J. Řezáč, P. Hobza, *J. Chem. Theory Comput.* **2013**, *9*, 2151–2155.

[25] R. Sedlak, T. Janowski, M. Pitoňák, J. Řezáč, P. Pulay, P. Hobza, *J. Chem. Theory Comput.* **2013**, *9*, 3364–3374.

[26] J. Řezáč, K. E. Riley, P. Hobza, *J. Chem. Theory Comput.* **2012**, *8*, 4285–4292.

[27] B. Temelso, K. A. Archer, G. C. Shields, *J. Phys. Chem. A* **2011**, *115*, 12034–12046.

[28] A. Li, H. S. Muddana, M. K. Gilson, *J. Chem. Theory. Comput.* **2014**, *10*, 1563–1575.

[29] L. Gráfová, M. Pitonák, J. Řezáč, P. Hobza, *J. Chem. Theory Comput.* **2010**, *6*, 2365–2376.

[30] J. Řezáč, K. E. Riley, P. Hobza, *J. Chem. Theory Comput.* **2011**, *7*, 2427–2438.

[31] M. Korth, S. Grimme, *J. Chem. Theory Comput.* **2009**, *5*, 993–1003.

[32] B. J. Lynch, Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2003**, *107*, 1384–1388.

[33] J. Zheng, Y. Zhao, D. G. Truhlar, *J. Chem. Theory Comput.* **2007**, *3*, 569–582.

[34] N. E. Schultz, Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 11127–11143.

[35] N. E. Schultz, Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 4388–4403.

[36] M. Fernandez, N. R. Trefiak, T. K. Woo, *J. Phys. Chem. C* **2013**, *117*, 14095–14105.

[37] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, H. Lam, *ACS Comb. Sci.* **2011**, *13*, 579–633.

[38] R. F. Murphy, *Nat. Chem. Biol.* **2011**, *7*, 327–330.

[39] M. Fernandez, A. S. Barnard, *ACS Nano* **2015**, *9*, 11980–11992.

[40] M. Fernandez, M. Breedon, I. S. Cole, A. S. Barnard, *Chemosphere* **2016**, *160*, 80–88.

[41] J. E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons, Inc., **1991**.

[42] M. Mørup, L. K. Hansen, *Neurocomputing* **2012**, *80*, 54–63.

[43] A. Cutler, L. Breiman, *Technometrics* **1994**, *15*, 661–675.

[44] P. Huggins, L. Pachter, B. Sturmfels, *Bull. Math. Biol.* **2007**, *69*, 2723–2735.

[45] J. C. Thøgersen, M. Mørup, S. Damkiæ r, S. Molin, L. Jelsbak, *BMC Bioinformatics* **2013**, *14*, 279.

[46] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, U. Alon, *Science* **2012**, *336*, 1157.

[47] S. Marinetti, L. Finesso, E. Marsilio, *Infrared Phys. Technol.* **2007**, *49*, 272–276.

[48] M. V. Kosti, R. Feldt, L. Angelis, *Empir. Softw. Eng.* **2016**, *21*, 1509–1532.

[49] G. C. Porzio, G. Ragozini, D. Vistocco, *Appl. Stoch. Model. Bus. Ind.* **2008**, *24*, 419–437.

[50] M. J. A. Eugster, F. Leisch, *J. Stat. Softw.* **2009**, *30*, 1–23.

[51] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys* **2016**, *18*, 13754–13769.

[52] A. Sadeghi, S. A. Ghasemi, B. Schaefer, S. Mohr, M. A. Lill, S. Goedecker, *J. Chem. Phys.* **2013**, *139*, 184118.

[53] A. Cherkasov, E. N. Muratov, D. Fourches, A. Varnek, I. I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y. C. Martin, R. Todeschini, V. Consonni, V. E. Kuz'Min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, *J. Med. Chem.* **2014**, *57*, 4977–5010.

[54] A. R. Katritzky, V. S. Lobanov, M. Karelson, *CODESSA 2.0 (Comprehensive Descriptors for Structural and Statistical Analysis*, University of Florida, USA, **1996**.

[55] E. J. Gardiner, V. J. Gillet, M. Haranczyk, J. Hert, J. D. Holliday, N. Malim, Y. Patel, P. Willett, *Stat. Anal. Data Mining* **2009**, *2*, 103–114.

[56] K. T. Nguyen, L. C. Blum, R. V. Deursen, J.-l. Reymond, *ChemMedChem* **2009**, 1803–1805.

[57] R. S. Pearlman, K. M. Smith, *Discovery* **1998**, 339–353.

[58] D. Bonchev, N. Trinajstic, *J. Chem. Phys.* **1977**, *67*, 4517.

[59] S. H. Bertz, *J. Am. Chem. Soc.* **1981**, *083*, 3599–3601.

[60] A. T. Balaban, *Chem. Phys. Lett.* **1982**, *89*, 399–404.

[61] L. H. Hall, L. B. Kier, *Rev. Comput. Chem.* **1991**, *2*, 367–422.

[62] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer New York, New York, NY, **2009**.

[63] M. Rupp, A. Tkatchenko, K.-R. Muller, O. A. Von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 58301.

[64] A. P. Bartók, M. J. Gillan, F. R. Manby, G. Csányi, *Phys. Rev. B* **2013**, *88*, 054104.

[65] O. A. Von Lilienfeld, *Int. J. Quantum Chem.* **2013**, *113*, 1676–1689.

[66] F. Pietrucci, W. Andreoni, *Phys. Rev. Lett.* **2011**, *107*, 085504.

[67] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.

[68] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, *New J. Phys.* **2013**, *15*, 095003.

[69] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, K. R. Muller, *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

[70] A. Varnek, I. Baskin, *J. Chem. Inf. Model.* **2012**, *52*, 1413–1437.

[71] W. Kohn, A. D. Becke, R. G. Parr, *J. Phys. Chem.* **1996**, *100*, 12974–12980.

[72] A. J. Cohen, P. Mori-Sánchez, W. Yang, *Chem. Rev.* **2012**, *112*, 289–320.

[73] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **2012**, *8*, 2181–2188.

[74] F. R. Petruzielo, J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2012**, *136*, 124116.

[75] D. M. Cleland, M. C. Per, *J. Chem. Phys.* **2016**, *144*, 124108.

[76] *NIST Computational Chemistry Comparison and Benchmark Database NIST Standard Reference Database Number 10*, `http://cccbdb.nist.gov/`, Release 18, October 2016.

[77] J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

[78] *RDKit : Open source Cheminformatics*, **2016**. `http://www.rdkit.org`.

[79] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *109*, 059802.

[80] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, **1993**.

[81] L. Breiman, *Machine Learning* **2001**, *45*, 5–32.

[82] P. Jurečka, P. Hobza, *Chem. Phys. Lett.* **2002**, *365*, 89–94.

[83] A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen, A. K. Wilson, *Chem. Phys. Lett.* **1998**, *286*, 243–252.

[84] T. H. Dunning, *J. Chem. Phys.* **1989**, *90*, 1007.

[85] M. Feyereisen, G. Fitzgerald, A. Komornicki, *Chem. Phys. Lett.* **1993**, *208*, 359–363.

[86] O. Vahtras, J. Almlöf, M. Feyereisen, *Chem. Phys. Lett.* **1993**, *213*, 514–518.

[87] D. E. Bernholdt, R. J. Harrison, *Chem. Phys. Lett.* **1996**, *250*, 477–484.

[88] P. Jurečka, P. Nachtigall, P. Hobza, *Phys. Chem. Chem. Phys.* **2001**, *3*, 4578–4582.

[89] J. C. Slater, *Phys. Rev.* **1951**, *81*, 385.

[90] S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* **1980**, *58*, 1200–1211.

[91] J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, *Phys. Rev. Lett.* **2003**, *91*, 146401.

[92] Y. Zhao, D. G. Truhlar, *J. Chem. Phys.* **2006**, *125*.

[93] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648–5652.

[94] P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *J. Phys. Chem.* **1994**, *98*, 11623–11627.

[95] J.-D. Chai, M. Head-Gordon, *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615–6620.

[96] M. Ernzerhof, G. E. Scuseria, *J. Chem. Phys.* **1999**, *110*, 5029–5036.

[97] C. Adamo, V. Barone, *J. Chem. Phys.* **1999**, *110*, 6158–6170.

[98] A. D. Becke, *J. Chem. Phys.* **1996**, *104*, 1040–1046.

[99] Y. Zhao, B. J. Lynch, D. G. Truhlar, *J. Phys. Chem. A* **2004**, *108*, 2715–2719.

[100] S. Grimme, *J. Chem. Phys.* **2006**, *124*, 034108.

[101] T. Schwabe, S. Grimme, *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398–4401.

[102] B. Chan, J. Deng, L. Radom, *J. Chem. Theory Comput.* **2011**, *7*, 112–120.

[103] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, . Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, *Gaussian-09 Revision E.01*, Gaussian Inc. Wallingford CT 2009.

[104] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, *WIREs Comput. Mol. Sci.* **2012**, *2*, 242–253.

[105] M. C. Per, *CSIRO Quantum Monte Carlo software package*, **2016**.

[106] E. T. Swann, M. L. Coote, A. S. Barnard, M. C. Per, *Int. J. Quantum Chem.* **2017**, *117*, 1–7.

[107] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[108] M. Dolg, C. Filippi, *private communication* **2014**.

[109] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[110] N. W. Winter, *J. Chem. Phys.* **1975**, *62*, 1269.

[111] C. A. Parsons, C. E. Dykstra, *J. Chem. Phys.* **1979**, *71*, 3025.

[112] H. J. A. Jensen, P. Joergensen, T. Helgaker, *J. Am. Chem. Soc.* **1987**, *109*, 2895–2901.

[113] J. Andzelm, C. Sosa, R. A. Eades, *J. Phys. Chem.* **1993**, *97*, 4664–4669.

[114] M. L. McKee, *J. Phys. Chem.* **1993**, *97*, 13608–13614.

[115] B. J. Smith, *J. Phys. Chem.* **1993**, *97*, 10513–10514.

[116] C. Angeli, R. Cimiraglia, H.-J. Hofmann, *Chem. Phys. Lett.* **1996**, *259*, 276–282.

[117] B. Jursic, *Chem. Phys. Lett.* **1996**, *4*, 13–17.

[118] P. Mach, J. Masik, J. Urban, I. Hubac, *Mol. Phys.* **1998**, *94*, 173–179.

[119] J. M. L. Martin, P. R. Taylor, *Mol. Phys.* **1999**, *96*, 681–692.

[120] V. Stepanic, G. Baranovic, *Chem. Phys.* **2000**, *254*, 151–168.

[121] P. K. Chattarj, P. Perez, J. Zevallos, A. Toro-Labbe, *J. Mol. Struct.* **2001**, *580*, 171–182.

[122] D. Hwang, A. Mebel, *J. Phys. Chem. A* **2003**, *107*, 2865–2874.

[123] X. Pu, N.-B. Wong, G. Zhou, J. Gu, A. Tian, *Chem. Phys. Lett.* **2005**, *408*, 101–106.

[124] M. Biczysko, L. Poveda, A. Varandas, *Chem. Phys. Lett.* **2006**, *424*, 46–53.

[125] R. K. Chaudhuri, K. F. Freed, S. Chattopadhyay, U. Sinha Mahapatra, *J. Chem. Phys.* **2008**, *128*, 144304.

[126] U. S. Mahapatra, S. Chattopadhyay, *J. Chem. Phys.* **2011**, *134*, 044113.

[127] J. Jana, *Reports Theor. Chem.* **2012**, *1*, 1–10.

[128] M. Musiał, Ł. Lupa, K. Szopa, S. A. Kucharski, *Struct. Chem.* **2012**, *23*, 1377–1382.

[129] A. M. Sand, C. a. Schwerdtfeger, D. A. Mazziotti, *J. Chem. Phys.* **2012**, *136*, 034112.

[130] E. Swann, M. Fernandez, A. Barnard, M. Coote, *CMolsC-1 Quantum Chemical Test Set. v1*, CSIRO Data Collection, **2017**. `10.4225/08/58bcf1565950a`.

[131] E. Swann, M. Fernandez, A. Barnard, M. Coote, *CMolsT-1 Quantum Chemical Test Set. v1*, CSIRO Data Collection, **2017**. `10.4225/08/58bcf21ca85b6`.

[132] E. Swann, M. Fernandez, A. Barnard, M. Coote, *CMolsC-org Quantum Chemical Test Set. v1*, CSIRO Data Collection, **2017**. `10.4225/08/58bcf3005e549`.

[133] E. Swann, M. Fernandez, A. Barnard, M. Coote, *CMolsT-org Quantum Chemical Test Set. v1*, CSIRO Data Collection, **2017**. `10.4225/08/58bcf2cf53bbe`.

[134] J. P. Perdew, K. Schmidt, *AIP Conf. Proc.* **2001**, *577*, 1–20.

[135] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2011**, *7*, 291–309.

Chapter 9

# Conclusion

Electronic structure theory is a powerful tool but more scalable methods are needed to take full advantage of the new wave of parallel computing. The stochastic nature of quantum Monte Carlo (QMC) means it is well suited for this style of computing. It is still a relatively new method and thorough benchmarking is needed before it is readily adopted by the quantum chemistry community. There also needs to be more transparency on the algorithmic choices made in its implementation. There are a number of parameters that can be adjusted that have a significant effect on the speed of calculations but not necessarily the accuracy. Using methanol as a test-case, a combination of parameters was found that reduced the speed of calculations without affecting the accuracy. This method was then applied to a set of reaction barrier heights to demonstrate the suitability of DMC for a wide range of chemically relevant systems. DMC performed consistently well for a variety of reaction barrier heights and had an average error of 0.9 kcal/mol across three diverse databases.

A unique advantage of QMC methods over traditional *ab initio* and DFT methods is the flexibility in the choice of trial wavefunction. The accuracy of DMC for the aforementioned barrier heights is in large part due to the Slater-Jastrow wavefunction explicitly accounting for the dynamic and non-dynamic electron correlation in the system. The trial wavefunction can also be systematically improved by including more determinants. This was demonstrated using the ionisation potentials and electron affinities of first- and second-row systems. These systems are extremely sensitive to the treatment of electron correlation and accurate values will only be obtained when there is a balanced description of correlation on the charged and neutral species. A significant portion of the correlation energy was recovered when using a multi-determinant expansion with just single and double excitations and higher-order excitations had little effect. Less correlation energy was recovered using a multi-determinant expansion with pseudopotentials but results were still good.

The advantage of a systematically improvable wavefunction was demonstrated with a set of eighteen challenging reactions. DMC methods had comparable performance with the best DFT methods for a range of these challenging systems. For reactions with particularly large errors a multi-determinant expansion using a CISD wavefunction improved the results. The results were dependent on the active space used and more virtual orbitals were necessary for reactions with second-row atoms. Not all systems could be improved. For large systems of second-row atoms, like $S_8$, a larger active space was too computationally demanding. For particularly challenging systems like diazene with a strong multireference character a multi-determinant wavefunction had little effect on the final energy.

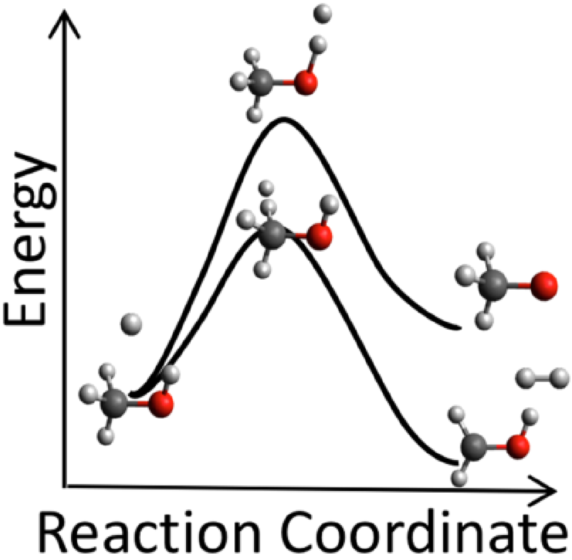
Extensive use was made of existing benchmarking protocols but the current methodology is tedious and rife with problems. We want methods that are accurate for all domains of chemical space but the current test sets are not testing this. Properties like atomisation energies don't commute with properties like barrier heights. Ideally the test sets we use to validate our methods should capture a representative subset of chemical space but current test sets are biased by our perceived chemical intuition and an human tendency for easy systems. Good performance for one test set does not guarantee good performance for another. Test sets like DC18 can be built using problematic reactions but there is no *a priori* way of knowing if a system will be challenging or not.

With better test sets we can spend less time verifying and justifying our methods and more time developing new ones. To this end a method was developed that used multivariate statistics to objectively select a representative subset. These methods were tested on a set of 1500 molecules using numerical descriptors based entirely on the structure of these molecules. No electronic-structure calculations were necessary. Using these methods identified a subset of molecules was identified that overlapped with the DC18 test set, but with no information known about them before hand. These test sets have the potential to improve the parameterisation of DFT methods. Empirical parameterisation is used in some methods to better describe the exchange-correlation functional but often this leads to skewed performance in favour of systems similar to the training set. These unbiased test sets could be used for a more robust parameterisation resulting in more universal functionals.

# Appendix 1

## A.1 Paper 1

**Efficient protocol for quantum Monte Carlo calculations of hydrogen abstraction barriers : Application to methanol**



Swann, E.T., Coote, M.L., Barnard, A.S., Per, M.C., 2017, **117**, *Int. J. Quantum Chem*, 1-7.

Received: 6 November 2016 | Revised: 22 December 2016 | Accepted: 13 January 2017

DOI 10.1002/qua.25361

FULL PAPER

WILEY International Journal of QUANTUM CHEMISTRY

# Efficient protocol for quantum Monte Carlo calculations of hydrogen abstraction barriers: Application to methanol

Ellen T. Swann[1,2] | Michelle L. Coote[2] | Amanda S. Barnard[1] | Manolo C. Per[1]

[1]Data61 CSIRO, Molecular & Materials Modelling, Door 34 Goods Shed, Village Street, Docklands, Victoria 3008, Australia

[2]ARC Centre of Excellence for Electromaterials Science, Research School of Chemistry, Australian National University, Canberra, Australian Capital Territory 2601, Australia

**Correspondence**
Dr. Manolo Per, Data61 CSIRO, Molecular & Materials Modelling, Door 34, Goods Shed, Village Street, Docklands, Melbourne, Victoria, Australia.
Email: manolo.per@data61.csiro.au

**Funding information**
This work was supported by the CSIRO Advanced Materials Future Science Platform. Computational resources were supplied by the Australian National Computing Infrastructure (NCI) national facility under the Partner Allocation Scheme grant dr4. ETS acknowledges support from the Office of the Chief Executive (OCE), CSIRO.

## Abstract

Accurate calculation of hydrogen abstraction reaction barriers is a challenging problem, often requiring high level quantum chemistry methods that scale poorly with system size. Quantum Monte Carlo (QMC) methods provide an alternative approach that exhibit much better scaling, but these methods are still computationally expensive. We describe approaches that can significantly reduce the cost of QMC calculations of barrier heights, using the hydrogen abstraction of methanol by a hydrogen atom as an illustrative example. By analysing the combined influence of trial wavefunctions and pseudopotential quadrature settings on the barrier heights, variance, and time-step errors, we devise a simple protocol that minimizes the cost of the QMC calculations while retaining accuracy comparable to large-basis coupled cluster theory. We demonstrate that this protocol is transferable to other hydrogen abstraction reactions.

KEYWORDS

methanol, quantum Monte Carlo, reaction barrier

## 1 | INTRODUCTION

Hydrogen abstraction reactions play important roles in many branches of organic chemistry, from the combustion of hydrocarbons[1] to damage in DNA as an indirect consequence of exposure to ionising radiation.[2] Accurate calculations of reaction barrier heights are required to build kinetic models for these processes; however, chemically reliable results often require a high-level treatment of electron correlation effects. This is illustrated by the H abstraction of methanol by an H atom, which has two main reaction channels,

$$CH_3OH + H \rightarrow CH_2OH + H_2 \qquad (1)$$

$$\rightarrow CH_3O + H_2 \qquad (2)$$

with $CH_2OH$ as the dominant product. These reactions are known to be important in the combustion of methanol under fuel-rich conditions,[3] and have been studied using a wide range of theoretical methods.[4–9] Despite the apparent simplicity of this system, studies have shown that accurate calculations of the barrier heights require methods that scale as a large power of the system size,[6] and that there are large discrepancies between methods. The performance of both MP2 theory and the popular B3LYP density functional[10] are particularly poor for this system.

Real-space quantum Monte Carlo (QMC) methods offer a stochastic alternative to traditional high-level electronic structure methods, and have been shown to be highly accurate for energetic[11,12] and structural properties.[13] The main advantages of QMC methods over more widely used alternatives such as coupled cluster theory are their low scaling with system size ($N^{3-4}$), and their immense parallelisability.[14] Despite these advantages there have been relatively few QMC calculations of H abstraction barrier heights. The earliest example, of the reaction $H_2 + H \rightarrow H + H_2$,[15,16] has very recently been revised to even higher accuracy.[17] Other examples include the reaction $OH + H_2 \rightarrow H_2O + H$,[18] and a study by Kollias et al. of the H abstraction of methanol by a Cl atom, which showed agreement with MP2 calculations.[19] More recent examples include the H abstraction by styrene of the H-terminated Si(001) surface,[20] and calculations of the barrier heights of three H-transfer reactions involving small molecules.[21]

To achieve resolutions of chemical accuracy in the barrier heights, statistical uncertainties in the stochastic QMC energies need to be on the order of fractions of a kcal/mol. Even though QMC scales well, this need for small uncertainties makes the calculations computationally expensive. As with other electronic-structure theories, efficient use of QMC methods requires a number of methodological choices to be made, including the choice of trial wavefunction and treatment of nonlocal pseudopotentials. Wavefunction choice is often discussed in reports of QMC calculations, but the effects of the parameters governing the treatment of pseudopotentials, including quadrature grids and cutoffs, are rarely mentioned.

In this work, we investigate the impact of these choices, and their mutual interactions, by performing a detailed study of the barrier heights of H abstraction in methanol by an H atom. We show that significant cost savings can be achieved, while still obtaining accurate barrier heights. In addition, we demonstrate that the approach is transferable, by evaluating the barrier heights of four unrelated H abstraction reactions. This enables a more black-box approach to be taken in similar QMC calculations.

## 2 | METHODS

The forward (F) and reverse (R) barrier heights of the reactions shown in Equations 1 and 2 are defined as the total energy differences

$$V_{1F}=E(\mathrm{TS1})-E(\mathrm{CH_3OH})-E(\mathrm{H}) \tag{3}$$

$$V_{1R}=E(\mathrm{TS1})-E(\mathrm{CH_2OH})-E(\mathrm{H_2}) \tag{4}$$

$$V_{2F}=E(\mathrm{TS2})-E(\mathrm{CH_3OH})-E(\mathrm{H}) \tag{5}$$

$$V_{2R}=E(\mathrm{TS2})-E(\mathrm{CH_3O})-E(\mathrm{H_2}) \tag{6}$$

where TS1 and TS2 are the transition-state structures for the reactions. The molecular geometries were obtained from B3LYP calculations with the Roos augmented triple-zeta (ATZ) basis set,[22] using Gaussian09.[23]

Practical QMC calculations require user-defined trial wavefunctions. The complexity of these wavefunctions strongly influence the computational cost of the calculations. Complicated wavefunctions are more expensive to optimize and evaluate at each Monte Carlo step, but more accurate wavefunctions lower the variance of the energy and therefore require fewer Monte Carlo steps to obtain a given statistical accuracy. In addition, the nodal surface of the wavefunction (the hypersurface on which it equals zero, and across which it changes sign) determines the systematic errors in fixed-node Diffusion Monte Carlo (DMC) calculations. The trial wavefunctions employed here have the Slater–Jastrow form,

$$\Psi_T=e^J D^{\uparrow} D^{\downarrow} \tag{7}$$

where the $D^{\uparrow,\downarrow}$ are Slater determinants constructed from single-particle orbitals, and $J$ is a Jastrow factor containing explicit electron correlation terms.

The Jastrow factor we use is a sum of electron-electron (*ee*), electron-nucleus (*eN*), and electron-electron-nucleus (*eeN*) terms,

$$J=\sum_{i>j}\sum_{A}\left[J_{ee}(r_{ij})+J_{eN}(r_{iA})+J_{eeN}(r_{iA},r_{jA},r_{ij})\right] \tag{8}$$

where $i$, $j$ label electrons, and $A$ labels nuclei. These terms were constructed as compactly supported natural polynomial expansions in the electron–electron and electron–nucleus distances,

$$J_{ee}(r_{ij})=f(r_{ij};L^{ee})\sum_{l=0}^{N_{ee}}\alpha_l\, r_{ij}^l \tag{9}$$

$$J_{eN}(r_{iA})=f(r_{iA};L^{eN})\sum_{l=0}^{N_{eN}}\beta_{l;A}\, r_{iA}^l \tag{10}$$

$$J_{eeN}(r_{iA},r_{jA},r_{ij})=f(r_{iA};L^{eeN})f(r_{jA};L^{eeN})\sum_{l,m,n=0}^{N_{eeN}}\gamma_{lmn;A}\, r_{iA}^l r_{jA}^m r_{ij}^n \tag{11}$$

where $L$ is the cutoff range, and $\{\alpha,\beta,\gamma\}$ are optimizable parameters. The cutoff function $f(r;L)$ is a $C^2$-smooth Wendland function[24] which goes to zero at $L$,

$$f(r;L)=\begin{cases}\left(1-\frac{r}{L}\right)^4\left(1+4\frac{r}{L}\right) & 0\le r\le L\\ 0 & r>L\end{cases} \tag{12}$$

All the calculations presented here use fixed ranges of $L=5$ Bohr. The relevant symmetries, electron–electron cusp condition, and the electron–nucleus no-cusp conditions were enforced by constraining the optimisable parameters in the Jastrow factor. We used the method described in the appendix of Ref. [25] for the more complicated *eeN* term. The free parameters in the Jastrow factor were optimized by minimising the total energy at the variational Monte Carlo (VMC) level, using the linear method of Toulouse and Umrigar.[26]

The orbitals used in the Slater determinants were taken from B3LYP calculations. Although the orbitals themselves contain no description of electron correlation, it has been shown that using orbitals from a correlated method such as B3LYP results in better QMC energies than using Hartree–Fock orbitals.[27] For all-electron calculations using the full electron-ion Coulomb potential, the orbitals were expanded in the Gaussian-type Roos ATZ basis set,[22] and cusp-corrected using a standard approach.[28]

In addition to all-electron calculations, we also used nonlocal pseudopotentials to represent the ionic cores. The use of pseudopotentials can greatly reduce the cost of QMC calculations, as the removal of the chemically inert core electrons reduces the fluctuations in the local energy. Evaluation of the local energy requires the nonlocal potential to be projected onto the trial wavefunction. Following Mitas et al.,[29] for each ion the contribution to this projection from an electron labeled $i$ can be written as a sum over angular momenta $l$,

$$\left(\frac{\hat{V}_{\mathrm{nl}}\Psi_T}{\Psi_T}\right)_i=\sum_l\frac{(2l+1)}{4\pi}v_l(r_i)\int_{4\pi}P_l[\cos\theta_i']\frac{\Psi_T(\ldots,\mathbf{r}_i',\ldots)}{\Psi_T(\ldots,\mathbf{r}_i,\ldots)}d\Omega_i' \tag{13}$$

where $v_l$ is the angular-momentum dependent radial potential, and the integral is over the surface of a sphere of radius $r_i$ centred on the ion. In practice the integral is evaluated using a deterministic approach,

$$\int_{4\pi}f(\mathbf{r}_i')d\Omega_i'\approx\sum_k^{N_Q}w_k f(\mathbf{r}_k) \tag{14}$$

where the $N_Q$ weights $w_k$ and points $\mathbf{r}_k$ are chosen according to a Gaussian quadrature rule, with values taken from Ref. [29]. This projection must be evaluated for each electron within range of each ion, at each step of the DMC calculation, so the number of quadrature points and the range of the radial potentials can have a large impact on the cost of the QMC calculation. We compared the effect of using two different pseudopotentials, both of which were explicitly constructed for use in QMC calculations, but using different methods. Both are Hartree–Fock pseudopotentials including scalar relativistic effects, but the Trail-Needs (TN)[30] potentials are shape-consistent, whereas the Burkatzki–Filippi–Dolg (BFD)[31] potentials are energy-consistent. Our BFD calculations used the associated valence triple-zeta (VTZ) basis sets, and an improved H-atom potential.[32] Calculations with the TN potentials used the aug-cc-pVTZ-CDF basis set from Ref. [33].

Nonlocal pseudopotentials were treated beyond the locality approximation in DMC using the size-consistent T-moves approach.[34] Imaginary time-step sizes of $\tau=(0.04, 0.02, 0.01, 0.005)$ a.u. were used for pseudopotential calculations. Smaller time-step sizes are required for all-electron calculations, and in this case values of $\tau=(0.02, 0.01, 0.005, 0.001)$ a.u. were used. All DMC energies were extrapolated to $\tau=0$ using quadratic fits. Target population sizes of 8000 walkers were used in all DMC calculations. All the QMC calculations were performed using the CSIRO Quantum Monte Carlo code.[35]

Forward and reverse barrier heights are not directly available from experiment. The highest level theoretical results available in the literature use coupled cluster methods, so to compare with these we also performed coupled cluster calculations using the CCSD(T) approach. These were performed using Molpro[36] with Dunning's aug-cc-pVQZ basis set[37] and an unrestricted Hartree–Fock reference state.

Finally, we evaluated the accuracy of density functional theory using 12 different exchange-correlation functionals. The types of functionals chosen were the local density approximation (LDA[38]), generalized gradient approximation (GGA) functionals (BLYP,[39,40] PBE,[41] B97D3[42]), meta-GGAs (TPSS,[43] M06L[44]), hybrid GGAs (B3LYP,[10] PBE0[45]), hybrid meta-GGAs (B1B95,[46] MPW1B95[47]), and double hybrids (B2PLYP,[48] mPW2PLYP[49]).

## 3 | RESULTS AND DISCUSSION

Barrier heights calculated using density functional, CCSD(T), and QMC methods are compared against results from the literature in Table 1.

Our CCSD(T) results, calculated at the B3LYP geometries, agree closely with previous calculations. When compared against the results of Carvalho et al.,[8] who used CCSD(T)/cc-pVTZ geometries, the largest deviation we observe is only 0.3 kcal/mol. This demonstrates the accuracy of the B3LYP geometries, despite the inability of that level of theory to predict accurate barrier heights.

None of the exchange-correlation functionals we used are able to recover all the barrier heights to within chemical accuracy of the CCSD(T) reference values. The most accurate functionals are the double-hybrids, which perform well for the forward barriers. However, the reverse barrier $V_{2R}$ deviates from the reference CCSD(T) value by over 5 kcal/mol. There is significant variation of barrier heights, even within the same class of functionals. For example, the forward barrier heights obtained using the meta-GGA functionals TPSS and M06L differ by around 7 kcal/mol. The extreme variability in the accuracy of the different density functionals emphasizes the need for extensive benchmarking using higher level methods.

**TABLE 1** Reaction barrier heights for the H abstraction of methanol by an H atom (in kcal/mol) using different methods

| Method | $V_{1F}$ | $V_{1R}$ | $V_{2F}$ | $V_{2R}$ |
|---|---|---|---|---|
| LDA | −3.5 | −0.1 | 1.8 | −7.7 |
| BLYP | 1.1 | 12.0 | 3.3 | 6.5 |
| PBE | 2.0 | 8.3 | 5.7 | 3.2 |
| B97D3 | 0.8 | 12.5 | 3.9 | 7.3 |
| TPSS | −0.9 | 11.4 | 1.1 | 7.3 |
| M06L | 7.0 | 10.1 | 7.8 | 8.2 |
| B3LYP | 3.6 | 13.3 | 6.9 | 9.2 |
| B3LYP/6-31+G(d,p) [6] | 3.2 | 12.9 | - | - |
| PBE0 | 5.5 | 11.0 | 10.4 | 8.0 |
| B1B95 | 6.8 | 13.5 | 11.3 | 10.5 |
| B1B95/MG3S [6] | 7.0 | 13.5 | - | - |
| mPW1B95 | 6.9 | 12.8 | 11.5 | 10.0 |
| mPW1B95/MG3S [6] | 7.1 | 12.9 | - | - |
| B2PLYP | 10.0 | 17.6 | 14.0 | 17.2 |
| mPW2PLYP | 9.8 | 17.0 | 13.9 | 16.7 |
| MP2/6-31+G(d,p) [6] | 16.8 | 18.0 | - | - |
| CCSD(T)/aug-cc-pVQZ [6] | 9.6 | 15.6 | - | - |
| CCSD(T)/cc-pVQZ [8] | 9.8 | 15.8 | 15.1 | 12.0 |
| CCSD(T)/aug-cc-pVQZ | 9.5 | 15.5 | 15.1 | 11.7 |
| DMC (All-electron) | 9.9 (1) | 15.5 (1) | 16.3 (1) | 12.1 (1) |
| DMC (BFD) | 9.8 (2) | 15.0 (2) | 15.9 (2) | 12.1 (2) |
| DMC (TN) | 9.9 (2) | 15.6 (2) | 15.7 (2) | 12.2 (2) |

Statistical uncertainties in the last digit of the DMC results are shown in parentheses. All DFT calculations used the Roos-ATZ basis, unless stated otherwise.

All-electron QMC calculations were performed using the complete Jastrow factor shown in Equation 8, including the *eeN* terms. The barrier heights obtained using this method agree closely with the CCSD(T) results, for all but the $V_{2F}$ barrier, where there is a difference of just over 1 kcal/mol. This disagreement is potentially due to the presence of nondynamical correlation effects. The $T_1$ diagnostic[50] is a widely used indicator of nondynamical effects in coupled-cluster calculations. Typically, $T_1$ values of 0.02 or greater are taken as an indication that a single determinant reference state is insufficient, though some researchers suggest this value should be higher for open-shell systems.[51] Nearly all of the structures have small $T_1$ values, the exceptions being $CH_3O$ (0.021) and TS2 (0.032).

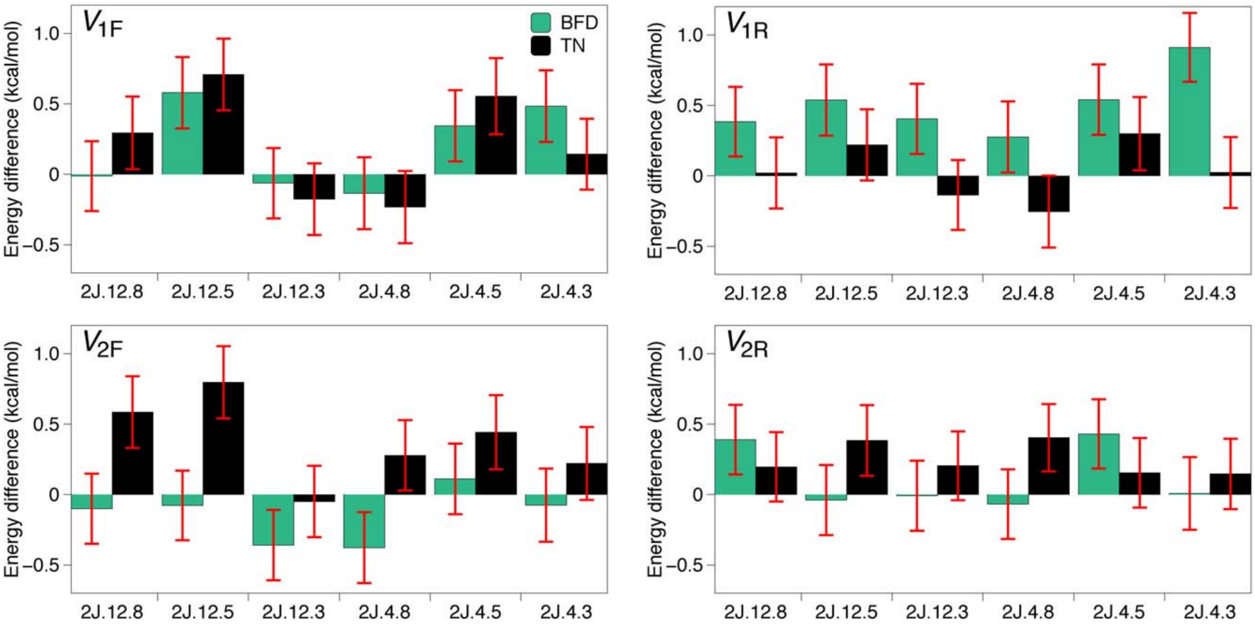


**FIGURE 1** Deviations of DMC barrier heights from the 3J.12.8 reference values, for different Jastrow factor and pseudopotential settings. The settings use the notation *x*.*y*.*z*, where *x* denotes the size of the Jastrow factor, *y* is the number of points used in the quadrature grid for evaluating nonlocal pseudopotentials, and *z* is a measure of the cutoff applied to the radial parts of the pseudopotentials. For more details see section 3

Reference pseudopotential QMC results were obtained using both BFD and TN pseudopotentials, with the same sized Jastrow factor used in the all-electron calculations. The barrier heights are all within two standard deviations of the all-electron results. These pseudopotential calculations are actually more expensive per Monte Carlo step than the all-electron approach (1.54x, for the DMC calculation of $CH_3OH$), due to the need to repeatedly evaluate the integral in Equation 13. The cost benefits of the pseudopotential approach come from the ability to use larger imaginary time-step sizes, and the reduced variance of the energy. As an example, for a fixed time-step size of $\tau$=0.005 a.u., the BFD calculation of $CH_3OH$ is over 20% faster than the all-electron calculation, to obtain the same statistical accuracy. This reduction in cost will be significantly larger for systems containing heavier elements.

## 3.1 | Approximations

To further reduce the computational cost of the QMC calculations, we investigated the effect of reducing both the complexity of the Jastrow factor and the treatment of the pseudopotentials. Dubecký et al.[52] have shown that for noncovalent interactions a two-body Jastrow factor is sufficient, and we consider the same modification here, removing the most expensive *eeN* terms. For the pseudopotentials, we reduced the number of quadrature points in the evaluation of Equation 13, and also reduced the range of both the local and nonlocal radial potentials. The notation used to define these settings is *x*.*y*.*z*. Here *x* denotes the size of the Jastrow factor, which is either 2J (indicating use of *ee* and *eN* terms) or 3J (indicating use of *ee*, *eN*, and *eeN* terms). The number of points used in the quadrature grid ($N_Q$ in Equation 14) is given by *y*. We used $y = 4$ tetrahedrally distributed points, and $y = 12$ icosahedrally distributed points, with the locations of the points on the unit sphere taken from Ref. [29]. Finally, *z* is a parameter which determines the ranges of the radial parts of both the local and nonlocal pseudopotentials. The ranges *r* of the local potentials are chosen such that *r* is the point furthest from the nucleus which deviates by more than $10^{-z}$ from the bare Coulomb potential. Similarly, the range of the radial part of a nonlocal potential is defined as the point furthest from the nucleus which deviates by more than $10^{-z}$ from zero. The potentials are set to zero outside these ranges. This method of defining the ranges leads to different values for each element and for both types of pseudopotential used. For the elements considered here, the BFD potentials are shorter ranged than the TN potentials for each of our choices of $z = 8$, 5, 3. Using this notation, the settings used for the reference pseudopotential calculations in Table 1 are 3J.12.8.

## 3.2 | Accuracy

As shown in Figure 1, the different settings used for the Jastrow factor and pseudopotentials have very little effect on the predicted barrier heights. The majority of settings result in no statistically significant change, and the largest changes are less than 1 kcal/mol. Overall the deviations from the reference values are statistically equivalent for both types of pseudopotential considered.

## 3.3 | Cost

Despite the relatively insignificant changes in the barrier heights, the reduced settings can have a very strong effect on the computational cost of the QMC calculations. The timings for a complete DMC

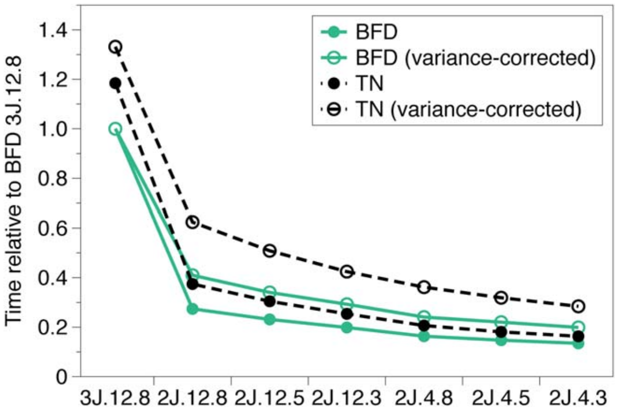


**FIGURE 2** DMC Timings for $CH_3OH$ relative to BFD.3J.12.8 settings. Solid points indicate relative times for a fixed number of Monte Carlo steps. Open points indicate relative times to achieve a fixed statistical uncertainty

calculation of $CH_3OH$, relative to the reference settings using the BFD pseudopotentials, are shown in Figure 2.

The largest time saving comes from eliminating the *eeN* terms in the Jastrow factor, which makes the DMC calculation of $CH_3OH$ 3x faster per Monte Carlo step. This simplified Jastrow factor increases the variance of the local energy (see section 3.4), but even when this effect is taken into account, we still obtain a speedup of around 2.5x for the time to achieve a fixed statistical accuracy in the total energy.

The next most important speedup comes from reducing the number of points in the quadrature grids. Evaluating the projection of the nonlocal operator onto the trial wavefunction requires multiple evaluations of the wavefunction ratio with the position of one electron moved. Even when using efficient methods for calculating this ratio, reducing the number of quadrature points from 12 to 4 results in a speedup of 1.7x when using the simpler 2J Jastrow factor. Reducing the ranges of the local and nonlocal parts of the pseudopotentials also reduces the cost of the calculations, but the improvement obtained is much smaller than when simplifying the Jastrow factor or reducing the number of quadrature points. When combined, all three measures provide a speedup greater than 5x, with no reduction in the quality of the barrier heights.

Our calculations using the TN pseudopotentials were always more expensive than when using the BFD potentials, as shown in Figure 2. There are a number of reasons for this, the most important being simply the size of the one-electron basis set used to construct the B3LYP orbitals in the trial wavefunction. When expressed in spherical harmonic (as opposed to Cartesian) Gaussians, the CDF basis set used with the TN potentials has 35% more primitives than the BFD basis set. There is also some contribution from the fact that the TN potentials contain a nonlocal term for H. The TN potentials contain *s*, *p*, and *d* channels for each element used in this work. The BFD potentials contain only *s* and *p* channels for C and O species, and only a local component for H. Finally, in the Gaussian representation of the pseudopotentials we used, the TN potentials contain many more terms than the BFD potentials. This has a very small effect on the cost, but it could be eliminated by representing both potentials on a radial grid.

## 3.4 | Variance

Reducing the complexity of the trial wavefunction by removing the *eeN* terms from the Jastrow factor leads to an increase in the variance of the local energy, shown in Figure 3. The results shown here are for energy-optimized trial wavefunctions. It should be possible to obtain lower variances by explicitly minimising the variance of the local energy rather than the total energy, though the gains are likely to be small.

Reducing the number of points in the quadrature grid has no effect on the variance when using the BFD pseudopotentials, but results in a 5% increase in the variance when using the TN potentials. This is likely due to the different angular momenta used in the construction of the different potentials. As aforementioned, the TN pseudopotentials use higher angular momentum terms than the BFD potentials, and so one would expect them to require a higher order quadrature rule. However, the effect is small, and does not translate in any statistically significant way to the quality of the energy barriers as shown in Figure 1. Reducing the ranges of the local and nonlocal radial potentials has no noticeable effect on the variance for either type of pseudopotential.

## 3.5 | Time-step error

The changes to the Jastrow factor and treatment of pseudopotentials also have an effect on the time-step error in DMC, as shown in Figure 4. We used a nonsymmetric branching factor in our DMC calculations with T-moves, which results in large time-step ($\tau$) errors, but with a predominantly linear behavior that is easily extrapolated to $\tau = 0$. Using a symmetric branching factor does result in a smaller error for a given value of $\tau$, but in practice we have frequently observed that the increased curvature means that reliable extrapolation to $\tau = 0$ still requires relatively small values of $\tau$. Using symmetric branching with T-moves is slightly more expensive than nonsymmetric branching, as it requires a second evaluation of the local energy at each DMC step if a T-move is accepted. Our current approach is to use nonsymmetric branching if performing a full extrapolation to $\tau = 0$, and to use symmetric branching if a single small value of $\tau$ is used.

As with the variance, the largest effect on the time-step error is the quality of the trial wavefunction. Using the larger 3J Jastrow factor

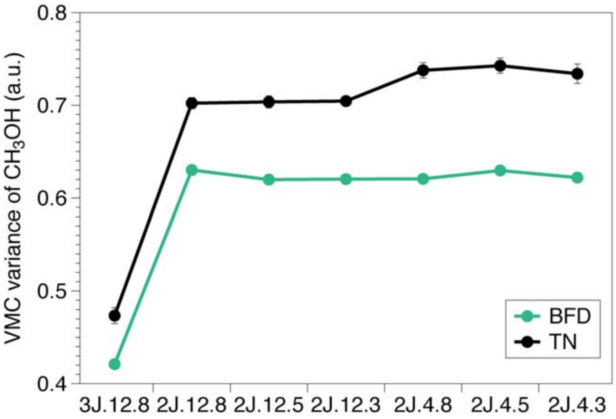


**FIGURE 3** VMC variance of the local energy for $CH_3OH$ using different settings for the Jastrow factor and pseudopotentials

**TABLE 2** Deviation of forward ($\Delta V_F$) and reverse ($\Delta V_R$) reaction barrier heights (in kcal/mol) from reference values, using different DMC protocols

| | BFD.3J.12.8 | | BFD.2J.4.5 | |
|---|---|---|---|---|
| Reaction | $\Delta V_F$ | $\Delta V_R$ | $\Delta V_F$ | $\Delta V_R$ |
| $H+HCl \rightarrow H_2+Cl$ | 0.052 (94) | −0.633 (92) | 0.02 (12) | −0.89 (12) |
| $OH+C_2H_6 \rightarrow C_2H_5+H_2O$ | 0.59 (17) | 0.48 (18) | 1.14 (22) | 0.75 (22) |
| $H+H_2S \rightarrow H_2+HS$ | 0.67 (11) | 0.77 (11) | 0.28 (12) | 0.42 (12) |
| $NH_2+CH_3 \rightarrow CH_4+NH$ | 0.63 (13) | 0.53 (13) | 0.48 (17) | 0.14 (18) |

Statistical uncertainties in the last digit of the DMC results are shown in parentheses.

results in smaller time-step errors than when using the 2J form. The quality of the Jastrow factor also has a small effect on the final $\tau = 0$ DMC energy, which comes from the projection of the nonlocal pseudopotential onto the trial wavefunction (Equation 13).

The majority of the quadrature grid and cutoff settings result in time-step errors that are mutually indistinguishable. Using the simpler 2J Jastrow factor, the use of a short range in the pseudopotentials has a larger effect than the number of quadrature points. Using the shortest range (corresponding to $z = 3$) produces noticeably higher energies, regardless of the number of quadrature points used. As the difference in cost between using ranges corresponding to $z = 3$ and $z = 5$ is so small, it is safer to use the larger value, which has no visible effect on the time-step error.

### 3.6 | Transferability

The results so far indicate that BFD.2J.4.5 is a reliable low-cost DMC protocol for calculating the barrier heights of the reactions (1) and (2). To assess its transferability to other systems, we compared it to the more expensive BFD.3J.12.8 protocol for four additional H abstraction reactions. These reactions were taken from the HTBH38/04 database[53] and include two reactions with the heavier elements S and Cl.

The results of these calculations are shown in Table 2, which lists the barrier heights relative to reference values obtained from a combination of theoretical and experimental results.[54] Both DMC protocols obtain barrier heights within chemical accuracy (within statistical uncertainties) of the reference values. They also agree with each other to within a fraction of a kcal/mol, showing that the reduced BFD.2J.4.5 protocol is transferable to other reaction barriers. The largest difference is in the forward barrier height for the reaction $OH + C_2H_6 \rightarrow C_2H_5 + H_2O$, though at 0.54 kcal/mol, this is still less than 2 statistical error bars.

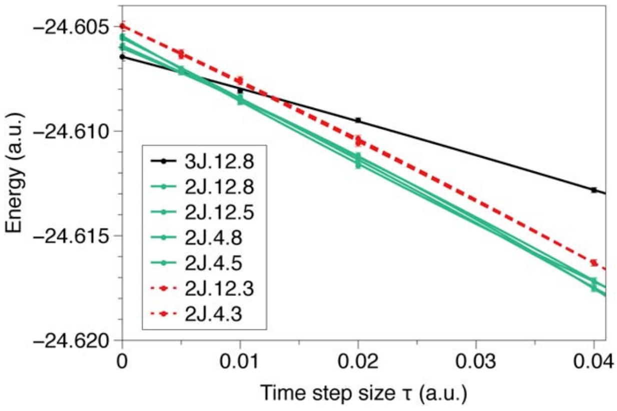


**FIGURE 4** DMC energies as a function of imaginary time-step for the TS2 geometry, using the BFD pseudopotentials

## 4 | CONCLUSIONS

We have used real-space QMC methods to calculate the reaction barrier heights of the two main channels for H abstraction of methanol by an H atom, a problem that requires a high-level treatment of electron correlation effects. The combination of B3LYP geometries and QMC energies predicts barrier heights that agree with CCSD(T) values to within chemical accuracy.

The cost of the QMC calculations can be minimized by simplifying the trial wavefunction and the treatment of nonlocal pseudopotentials. The largest cost saving can be achieved by using a simple Jastrow factor that includes only two-body correlation effects. By combining this simplified trial wavefunction with a sparse quadrature grid in the projection of the nonlocal pseudopotential, and applying cutoffs to the ranges of these potentials, we were able to reduce the cost of DMC calculations by a factor of 5x over our reference calculations, with no loss in accuracy.

In the notation defined in section 3, our recommended protocol is 2J.4.5, using the BFD pseudopotentials. We have shown that this approach is transferable to similar problems, including reactions involving the heavier elements S and Cl. However, a caveat is that one should be careful with the choice of integration grids for systems containing much heavier elements. These cases are likely to be more sensitive to the number of quadrature points due to the importance of larger angular momenta in the pseudopotentials.

Together with these cost-reducing measures, the accuracy, favorable scaling, and low memory requirements of QMC methods indicate this is a practical route to tackle H abstraction in much larger systems.

## ACKNOWLEDGMENTS

This work was supported by the CSIRO Advanced Materials Future Science Platform. Computational resources were supplied by the Australian National Computing Infrastructure (NCI) national facility under the Partner Allocation Scheme grant dr4. ETS acknowledges support from the Office of the Chief Executive (OCE), CSIRO. We thank Claudia Filippi for sharing an unpublished hydrogen pseudopotential and basis set.

## REFERENCES

[1] S. M. Sarathy, P. Oßwald, N. Hansen, K. Kohse-Höinghaus, *Prog. Energy Combust. Sci.* **2014**, *44*, 40.

[2] L. Pardo, R. Osman, J. Banfelder, A. P. Mazurek, H. Weinstein, *Free Radic. Res. Commun.* **1991**, *13*, 461.

[3] H. H. Grotheer, S. Kelm, H. S. T. Driver, R. J. Hutcheon, R. D. Lockett, G. N. Robertson, *Ber Bunsenges. Phys. Chem.* **1992**, *96*, 1360.

[4] J. T. Jodkowski, M. T. Rayez, J. C. Rayez, T. Bérces, S. Dóbé, *J. Phys. Chem. A* **1999**, *103*, 3750.

[5] Y. Y. Chuang, M. L. Radhakrishnan, P. L. Fast, C. J. Cramer, D. G. Truhlar, *J. Phys. Chem. A* **1999**, *103*, 4893.

[6] J. Pu, D. G. Truhlar, *J. Phys. Chem. A* **2005**, *109*, 773.

[7] E. F. V. Carvalho, A. N. Barauna, F. B. C. Machado, O. Roberto-Neto, *Int. J. Quantum Chem.* **2008**, *108*, 2476.

[8] E. F. V. Carvalho, A. N. Barauna, F. B. C. Machado, O. Roberto-Neto, *Chem. Phys. Lett.* **2008**, *463*, 33.

[9] R. Meana-Pañeda, D. G. Truhlar, A. Fernández-Ramos, *J. Chem. Phys.* **2011**, *134*, 094302.

[10] A. D. Becke, *J. Chem. Phys.* **1993**, *98*, 5648.

[11] M. A. Morales, J. McMinis, B. K. Clark, J. Kim, G. E. Scuseria, *J. Chem. Theory Comput.* **2012**, *8*, 2181.

[12] F. R. Petruzielo, J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2012**, *136*, 124116.

[13] D. M. Cleland, M. C. Per, *J. Chem. Phys.* **2016**, *144*, 124108.

[14] J. Kim, K. P. Esler, J. McMinis, D. M. Ceperley, in *Proceedings of the 2010 Scientific Discovery through Advanced Computing (SciDAC) Conference*, Chattanooga, Tennessee, July 11–15, **2010**. Oak Ridge National Laboratory. http://computing.ornl.gov/workshops/scidac2010/.

[15] F. Mentch, J. B. Anderson, *J. Chem. Phys.* **1984**, *80*, 2675.

[16] D. L. Diedrich, J. B. Anderson, *Science* **1992**, *258*, 786.

[17] J. B. Anderson, *J. Chem. Phys.* **2016**, *144*, 166101.

[18] J. C. Grossman, L. Mitas, *Phys. Rev. Lett.* **1997**, *79*, 4353.

[19] A. C. Kollias, O. Couronne, W. A. Lester, *J. Chem. Phys.* **2004**, *121*, 1357.

[20] Y. Kanai, N. Takeuchi, *J. Chem. Phys.* **2009**, *131*, 214708.

[21] F. Fracchia, C. Filippi, C. Amovilli, *J. Chem. Theory Comput.* **2013**, *9*, 3453.

[22] P. O. Widmark, P. A. Malmqvist, B. O. Roos, *Theor. Chim. Acta* **1990**, *77*, 291.

[23] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-;Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, T. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, D. J. Fox, *Gaussian 09*, Gaussian Inc. **2009**. http://gaussian.com/g09citation/.

[24] H. Wendland, *Adv. Comput. Math.* **1995**, *4*, 389.

[25] N. D. Drummond, M. D. Towler, R. J. Needs, *Phys. Rev. B* **2004**, *70*, 235119.

[26] J. Toulouse, C. J. Umrigar, *J. Chem. Phys.* **2007**, *126*, 084102.

[27] M. C. Per, K. A. Walker, S. P. Russo, *J. Chem. Theory Comput.* **2012**, *8*, 2255.

[28] M. C. Per, S. P. Russo, I. K. Snook, *J. Chem. Phys.* **2008**, *128*, 114106.

[29] L. Mitas, E. L. Shirley, D. M. Ceperley, *J. Chem. Phys.* **1991**, *95*, 3467.

[30] J. R. Trail, R. J. Needs, *J. Chem. Phys.* **2005**, *122*, 174109.

[31] M. Burkatzki, C. Filippi, M. Dolg, *J. Chem. Phys.* **2007**, *126*, 234105.

[32] M. Dolg, C. Filippi, private communication (**2014**).

[33] J. Xu, M. J. Deible, K. A. Peterson, K. D. Jordan, *J. Chem. Theory Comput.* **2013**, *9*, 2170.

[34] M. Casula, S. Moroni, S. Sorella, C. Filippi, *J. Chem. Phys.* **2010**, *132*, 154113.

[35] M. C. Per, *CSIRO Quantum Monte Carlo Software Package* **2016**.

[36] H. J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *2*, 242.

[37] T. H. Dunning, *J. Chem. Phys.* **1989**, *90*, 1007.

[38] S. H. Vosko, L. Wilk, M. Nusair, *Can. J. Phys.* **1980**, *58*, 1200.

[39] A. D. Becke, *Phys. Rev. A* **1988**, *38*, 3098.

[40] C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **1988**, *37*, 785.

[41] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865.

[42] S. Grimme, S. Ehrlich, L. Goerigk, *J. Comput. Chem.* **2011**, *32*, 1456.

[43] J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, *Phys. Rev. Lett.* **2003**, *91*, 146401.

[44] Y. Zhao, D. G. Truhlar, *J. Chem. Phys.* **2006**, *125*, 194101.

[45] C. Adamo, V. Barone, *J. Chem. Phys.* **1999**, *110*, 6158.

[46] A. D. Becke, *J. Chem. Phys.* **1996**, *104*, 1040.

[47] Y. Zhao, D. G. Truhlar, *J. Phys. Chem. A* **2004**, *108*, 6908.

[48] S. Grimme, *J. Chem. Phys.* **2006**, *124*, 034108.

[49] T. Schwabe, S. Grimme, *Phys. Chem. Chem. Phys.* **2006**, *8*, 4398.

[50] T. J. Lee, P. R. Taylor, *Int. J. Quantum Chem.* **1989**, *36*, 199.

[51] J. C. Rienstra-Kiracofe, W. D. Allen, H. F. Schaefer, *J. Phys. Chem. A* **2000**, *104*, 9823.

[52] M. Dubecký, R. Derian, P. Jurečka, L. Mitas, P. Hobza, M. Otyepka, *Phys. Chem. Chem. Phys.* **2014**, *16*, 20915.

[53] Y. Zhao, B. J. Lynch, D. G. Truhlar, *Phys. Chem. Chem. Phys.* **2005**, *7*, 43.

[54] B. J. Lynch, D. G. Truhlar, *J. Phys. Chem. A* **2002**, *106*, 842.

**How to cite this article**: Swann ET, Coote ML, Barnard AS, Per MC. Efficient protocol for quantum Monte Carlo calculations of hydrogen abstraction barriers: Application to methanol. *Int J Quantum Chem*. 2017;00:e25361. https://doi.org/10.1002/qua.25361