THE UNIVERSITY OF CHICAGO

DEVELOPMENTS AND APPLICATIONS OF AUTOMATED MULTICONFIGURATIONAL QUANTUM CHEMISTRY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF CHEMISTRY

BY

DANIEL SANBORN KING

CHICAGO, ILLINOIS

JUNE 2024

"You can use your brain to solve your problems"

– Mom

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I will start by thanking those who have most directly assisted in producing the following chapters, most centrally Laura Gagliardi, who has been a wonderful advisor to me during my PhD and has opened up a truly immense number of opportunities for me throughout my work: thank you for believing in my capacity to do science from the very beginning. Additionally I want to thank Matt Hermes, our staff scientist, who has been extremely impactful on my scientific thinking and ethics during my PhD. Matt is one of the most committed, smart, and engaged people I have ever met, and is centrally responsible for my excitement about the rigor and depth of computational science; I only hope that I have been able to share some of his enthusiasm with others.

I will next thank the fantastic collaborators I have had during my PhD, most principally Professor Don Truhlar, with which I have co-authored three papers throughout my PhD all included in this thesis. I would also like to thank Professor Max Delferro and his group, who supported me greatly in my SCGSR collaboration with Argonne National Laboratory. I want to thank Professor Shannon Stahl and his group for supporting me in a comprehensive computational study of fascinating copper electrocatalysis. I also want to thank Professor Brett Savoie for a fantastic collaboration in applying my automated multiconfigurational methods to organic transition states, which makes up the final chapter of this thesis. Finally, I want to thank Dr. Carlo Gaggioli and Dr. WooSeok Jeong for being excellent mentors to me during the first years of my PhD, and want to thank Jacob Wardzala, Matthew Hennefarth, Noah Dohrmann for working with me on different projects. My PhD would not have been nearly as interesting without you all.

Next, I want to thank the many professors who were vital in driving me towards chemistry and physics. Firstly, I would like to thank Dr. Dipannita Kalyani for being central in recognizing my well-meaning lack of experimental skill in lab and guiding me towards chemical theory, this truly had a large impact on my career and I have always been very

grateful for the advice. I would also like to thank my other fantastic chemistry professors Rob Hanson and Elodie Marlier for sharing their love of chemistry with me. Additionally, I would like to thank my fantastic physics professors Prabal Adhikari and Amy Kolan, whose challenging but inspiring courses were central in giving me the confidence to pursue a scientific career. Really I would like to thank the entire physics department at St. Olaf for being a first-rate pedagogical institution and doing a fantastic job.

I would also like to thank the National Science Foundation for sponsoring the national research experiences for undergraduates (REU), which allowed me to pursue summer research opportunities at the University of Kansas with Professor James Blakemore and at the University of California Los Angeles with Professor Ken Houk. These research experiences were really invaluable in motivating me to pursue a research career, and I thank both James and Ken for being wonderful mentors, for writing me recommendation letters, and everyone at the REU programs for organizing such incredible outreach opportunities. I do not know how I would have begun my journey into academia without these opportunities.

Finally, I would like to thank my amazing high elementary and high school science teachers Janet Gray-McKennis, Zeus Preckwinkle, Walt Kinderman, and Matt Silvia for fostering my love of science and encouraging me to pursue it outside of class. And of course, no acknowledgement in physics would be complete without thanking the amazing mathematics teachers I have had over the years, including Michael Caines, Mr. Espinoza, Dr. Karafiol, and Professor Matthew Wright. I would like to thank Matthew Wright in particular for introducing me to the YouTube channel 3Blue1Brown, which was invaluable in advancing my comprehesion of linear algebra and calculus. On that note I would like to acknowledge the countless other online and physical resources that have inspired me to pursue and taught me math and science over the years, as they have served as solid foundations for all my work.

Last but not least, I would like to thank my friends and family for helping me become who I am today. Of course, I most centrally want to thank my father and mother Ben and

Laura for bringing me into this world and helping me along every step of the way. I also want to thank my brother Christopher, my sister Annalucia, my best friend Henry, my amazing partner Bailey, and everyone else who has been by my side: I could not have done it without you all.

ABSTRACT

The efficient treatment of strong correlation in molecules is the foremost challenge in the accurate simulation of chemical systems at the quantum level. One of the most efficient methods for treating strong correlation is the "complete active space self-consistent field" (CASSCF) method, which places priors on the Hilbert space of electron configurations in which to carry out the calculation through use of an "active space" of electrons an orbitals in which to diagonalize the Hamiltonian. The problem of efficiently and automatically determining this active space, known as the "active space selection problem", is the focus of this thesis. Chapters 2, 3, and 4 principally concern the development of automated methods for efficiently determining the active space, while chapters 5, 6, and 7 concern the applications of these approaches.

CHAPTER 1

INTRODUCTION

1.1 Electron Correlation

The treatment of electron correlation is the foremost goal of electronic structure theory, and provides a foundation for all other areas of computational chemistry. While large, comprehensive simulations of materials often take the spotlight in industrial applications,[11] all of these methods ultimately rely on the understanding of electron behavior in various environments. This thesis is concerned with advancing and facilitating the use of some of the most accurate methods known for treating electron correlation in the hopes of ultimately bringing higher accuracy and understanding to the entire field of molecular simulation.

In this introduction, we briefly introduce the problem of electron correlation from a historical viewpoint, noting the central advances of (a) the treatment of exchange correlation by use of a Slater determinant[12] in 1929 and the development of Hartree-Fock theory in 1935,[13] and (b) the treatment of local, dynamic correlation by Kohn-Sham density functional theory beginning in 1965.[14] This work concerns the efficient use of methods to treat the remaining, non-local, multiconfigurational correlation, most prinicipally by the complete active space self-consistent field (CASSCF) method, which is held back by the problem of active space selection. The contributions of the following chapters to this problem are discussed, and the applications of these contributions found in later chapters provides inspiration for future work.

1.2 Exchange and Local Correlation

There are many ways to discuss the problem of electron correlation, and no rigorous separation of terms exists.[15] However, the problem can often be usefully separated into three different central contributions: (i) correlation that arises from symmetry (most profoundly

fermionic exchange symmetry), (ii) correlation that arises from the "dynamic", "short-range", or "local" avoidance of electrons from each other in real space, exemplified by the Coulomb cusp, and (iii) the remaining "nonlocal", "static", "multiconfigurational", or "multireference" correlation not captured by the previous categories. Learning to treat these different categories of correlation have coincided with fundamental advances in and uses of electronic structure theory.

Correlation arising from symmetry is generally the largest and most important term among these three categories, and arises centrally from the antisymmetric exchange symmetry of the electrons, causing the wave function to go to zero when electrons of the same spin occupy the same point in space. Efficiently treating the correlation that arises from this symmetry was overcome in the early days of quantum mechanics by Slater,[12] who proposed that the fundamental unit of computation should be determinants of single-particle functions:

$$\psi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_N(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_N(\mathbf{x}_N) \end{vmatrix}, \tag{1.1}$$

,

This wave function ansatz evidently captures the antisymmetric exchange symmetry of electronic wave functions, as particle exchange results in exchanging rows of the determinant, causing a change in sign. This ansatz was then adapted by Hartree, who variationally optimized this form with his father in the first application of the Hartree-Fock method in 1935, for the Beryllium atom.[13] The correlation arising from antisymmetry separates naturally out of the Hartree-Fock energy expression:

$$E_{HF} = V_{nn} + \sum_{pq} h_{pq} D_{pq} + \frac{1}{2} \sum_{pqrs} g_{pqrs} D_{pq} D_{rs} - \frac{1}{2} \sum_{pqrs} g_{pqrs} D_{ps} D_{rq} \tag{1.2}$$

where the last term, $K = -\frac{1}{2} \sum_{pqrs} g_{pqrs} D_{ps}$, a function of the one-body reduced density matrix (1-RDM) D_{pq} and two-electron integrals g_{pqrs}, represents the so-called "exchange" energy arising from this symmetry. One can immediately see from this expression that exchange is a significant and stabilizing effect (generally of similar magnitude to the Coulomb term $J = \frac{1}{2} \sum_{pqrs} g_{pqrs} D_{pq} D_{rs}$). While this term is evidently a "correlation", in the sense that it results in a lowering of the energy by means of electrons avoiding each other in space, it is so fundamental to the field that it is generally referred to separately as simply the "exchange" energy, with correlation reserved for the other categories. Treatment of this term accurately by Hartree-Fock inarguably gave rise to the first revolution in quantum chemistry, and nearly all methods begin from the starting point of the Hartree-Fock picture to this day.

It is probably fair to say that the second most important development in the field of electronic structure theory since Hartree-Fock has been the development of Kohn-Sham density functional theory (KS-DFT), beginning in 1965.[14] Starting from the famed Hohenberg-Kohn theorems[16] which reformulate quantum mechanics in terms of the electron density, Kohn-Sham theory aims to capture electron correlation as a functional of the electron density. However, despite this radical departure from standard quantum theory, the Kohn-Sham method actually takes on a very similar energy expression to Hartree-Fock (equation 1.2):

$$E_{\text{KS-DFT}} = V_{nn} + \sum_{pq} h_{pq} D_{pq} + \frac{1}{2} \sum_{pqrs} g_{pqrs} D_{pq} D_{rs} + \int_{\mathbf{r}} \epsilon_{\text{xc}}(\mathbf{r}) \rho(\mathbf{r}) \tag{1.3}$$

This similarity arises from the fact that KS-DFT also employs a determinant ansatz identical to equation 1.1 to model its density, most critically allowing it to utilize the same one-electron term $\sum_{pq} h_{pq} D_{pq}$ as HF for which there is not a good functional form purely in terms of the electron density. The most critical difference is that KS-DFT moves the treatment of

all correlation into an "exchange and correlation" (xc) functional (including the "exchange" correlation) of the electron density, in practice computed by integrating a *local* exchange-correlation energy, $\epsilon_{\mathrm{xc}}(\mathbf{r})$, on a grid over the entire molecule. Thus, the central advance of KS-DFT is not really a reformulation of quantum mechanics in terms of the density, but the finding that a good part of the correlation energy can be treated *locally*, without reference to the rest of the wave function.

This effective "near-sightedness" principle of KS-DFT[17] is physically sound, and arises from the concept of the "exchange-correlation hole" $n_{\mathrm{xc}}(\mathbf{r}, \mathbf{r'})$, which represents the function of the missing electron from the rest of the surroundings as a function of $\mathbf{r'}$ when an electron is known to be at $\mathbf{r}$:

$$n_{\mathrm{xc}}(\mathbf{r}, \mathbf{r'}) = \frac{P_2(\mathbf{r}, \mathbf{r'})}{\rho(\mathbf{r})} - \rho(\mathbf{r'}) \tag{1.4}$$

in which $P_2(\mathbf{r}, \mathbf{r'})$ is the pair density, equal to the diagonal elements of the reduced two-body density matrix (2-RDM, $P_2(\mathbf{r}, \mathbf{r'}) = d(\mathbf{r}, \mathbf{r}, \mathbf{r'}, \mathbf{r'})$). The exact exchange-correlation functional can then be reformulated in terms of this local function from standard Coulombic physics:

$$\epsilon_{\mathrm{xc}}(\mathbf{r}) = \frac{1}{2} \int_{\mathbf{r'}} \frac{n_{\mathrm{xc}}(\mathbf{r}, \mathbf{r'})}{|\mathbf{r} - \mathbf{r'}|} \tag{1.5}$$

This idea has been the cornerstone of the modern successful functionals such as PBE[18] and B3LYP,[19] which take ideas most centrally from studies of the homogeneous electron gas to derive local approximations of equation 1.5.[20] The efficient treatment of the local correlation well-treated by these approximations has arguably ushered in the second revolution of quantum chemistry, in which KS-DFT has become the standard approach for most applications.

1.3 Strong Correlation

The field has now solidly entered the phase of overcoming the third, most difficult category of correlation, which is that of the remaining (strong) *nonlocal* correlation not captured well by the correlation-hole picture assumed by KS-DFT functionals. Unfortunately, this challenge appears to require moving outside of the DFT framework: despite two decades of effort since the since the development of popular functionals such as PBE and B3LYP, little fundamental progress has been made in the development of new functionals, with the most notable developments being better-parameterized functionals such as M06[21] and the ωB97 series,[22] and the inclusion of force-field-like dispersion corrections (e.g. D3 corrections[23]).[24] Additionally, when success has come, it has often come only through reference back to the wave function picture, such as with the addition of exact exchange in hybrid functionals[19] or energies from perturbation theory in double-hybrid functionals.[25]

The fundamental problem arises from the fact that – to the best of my knowledge – there is no good way of engaging the concept of strong correlation from the density functional theory picture. It is inherently a wave function concept. Strong correlation arises when multiple electronic configurations contribute to the qualitative character of the quantum state, resulting in non-local and inherently quantum mechanical effects. The conditions in which this occurs are probably best viewed in the context of perturbation theory, in which the 2nd-order correction takes the rough form of

$$E_{(2)} = \sum_{i \neq j} \frac{V_{ij}^2}{E_j - E_i} \tag{1.6}$$

in which V_{ij} are the couplings between states i and j, and E_i and E_j their energies under the zeroth-order Hamiltonian. Strong correlation arises precisely when this term grows large: when (a) there is near-degeneracy between different electronic configurations, and (b) there is large coupling between these configurations (i.e., local interaction and not symmetry-

forbidden). Systems with these conditions are said to be "multiconfiguurational", and are generally ill-treated by KS-DFT.[26]

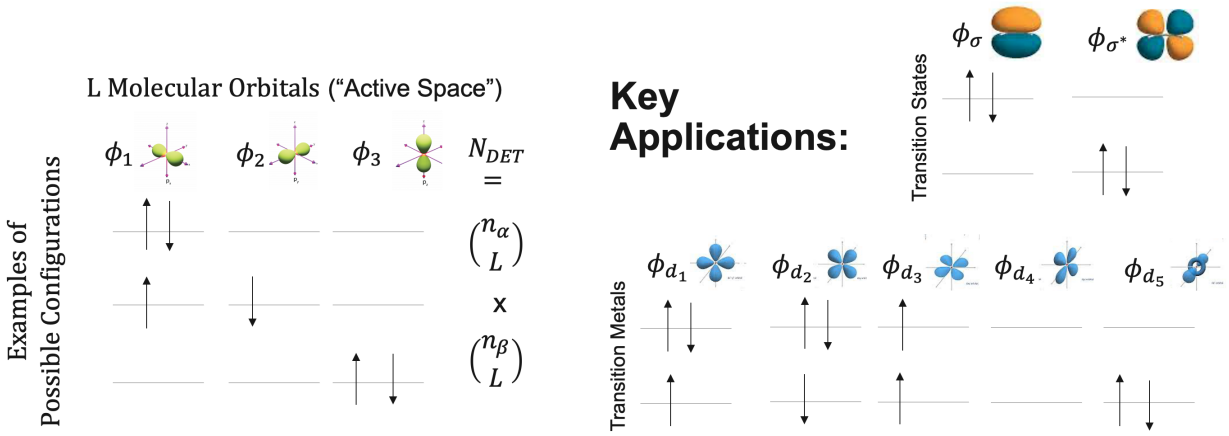

Figure 1.1: Left: Illustrative sketch of the active space concept, in which the number of determinants, N_{DET}, scales combinatorially with then number of electrons and orbitals chosen. Right: key applications for muliconfigurational methods with near-degenerate orbitals, (a) transition states, in which the bonding and antibonding configurations are near-degenerate, and (b) transition metals, in which different occupations of the d orbitals are degenerate.

However, despite the inherently quantum mechanical nature of strong correlation, most exemplary multiconfigurational systems are fairly intuitive to most chemists (Figure 1.1). Transition states are perhaps the marquee examples of commonly occurring multiconfigurational systems, in which the bonding and antibonding electronic configurations are very close in energy, resulting in a wave function that is a linear combination of both these possibilities. Another archetypal example is that of transition metals, in which the energies of the d-electrons are often weakly split, and thus multiple configurations of the electrons in the d-orbitals are often needed to accurately describe the wave function. Given the relevance of both these cases to modern chemistry, one can clearly see the importance of accurately treating the problem of strong correlation.

1.4 Treating Strong Correlation

Given the description of the problem, the solution to strong correlation in the wave function picture is self-evident: simply include all the relevant configurations in the Hilbert space and diagonalize. Unfortunately, identifying and including the relevant configurations *a priori* can be quite challenging, due to the exponential scaling of the number determinants with number of electrons and orbitals one considers (Figure 1.1). Including all electrons and orbitals in the active space is referred to as "full configuration interaction" (FCI), which can be carried out for only the smallest of systems (to about 20 electrons in 20 orbitals).[26]

Thankfully, research over the past 30 years has shown that the great majority of the possible electron configurations do not contribute to any one wave function – the configurations relevant for any one quantum state can generally be captured through a judicial selection of the Hilbert space. This observation has lead to two broad branches of development for efficiently identifying the relevant configurations in the Hilbert space: the "posteriors" approach, which aims to identify the Hilbert space through stochastic algorithms (e.g., Monte Carlo configuration interaction[27]) or entanglement characteristics (e.g., density matrix renormalization group (DMRG)[28]), and the "priors" approach, which aims to identify a subset of the Hilbert space prior to carrying out any calculation. The archetypal method of the latter category is the "complete active space self consistent field" (CASSCF) method, which takes on the wave function ansatz[29]

$$|\psi_{\mathrm{CAS}}\rangle = \sum_{n_1 n_2 ... n_n} c_{n_1 n_2 ... n_n} \, |22...n_1 n_2 ... n_n 00...\rangle \tag{1.7}$$

in which n_i are the occupation numbers (0, $\uparrow$, $\downarrow$, or 2) of an "active space" of orbitals and electrons treated multiconfigurationally (as in Figure 1.1).

This approach has proved to be a very efficent way of placing priors on the Hilbert space for multiconfigurational calculations and has seen much application in the literature.[26,30]

Although it may seem at first glance that the "priors" and "posteriors" methods are in conflict with one another, in reality the posteriors methods simply serve as more efficient solvers for the Hilbert spaces chosen through the priors methods. For example, DMRG allows active spaces to be treated with qualitative acccuracy up to about 100 orbitals in size,[31] rather than the roughly 20-orbital limit imposed by CASSCF alone.[26] An efficient combination of the prior and posterior approaches is the most effective for treating large systems with multiconfigurational approaches.

An additional point of concern is capturing the remaining dynamic correlation outside of the chosen subspace treated multiconfigurationally. This has historically been addressed by means of multireference perturbation theories (e.g., CASPT2[32,33] or NEVPT2[34,35]) or expansions (e.g. MRCI[36,37] or MRCC[38–40]). However, these approaches all require fairly dramatic amounts of additional computation on top of the multiconfigurational wave functions and limit the size of the active spaces that can be treated with these methods.[26] A promising new approach appears to be multiconfiguration pair-density functional theory,[41] in which the correlation energy is treated in real-space similar to KS-DFT:

$$E_{\text{MC-PDFT}} = V_{nn} + \sum_{pq} h_{pq} D_{pq} + \frac{1}{2} \sum_{pqrs} g_{pqrs} D_{pq} D_{rs} + \int_{\mathbf{r}} \epsilon_{ot}(\mathbf{r}) \rho(\mathbf{r}) \tag{1.8}$$

with the only key differences being (a) the use of density matrices D_{pq} and density ρ from a multiconfigurational wave function, and (b) the use of an "on-top" exchange-correlation functional ϵ_{ot} in place of ϵ_{xc} which takes as an argument the on-top density $\Pi(\mathbf{r})$, equal to the doubly diagonal elements of the two-body reduced density matrix (2-RDM, $\Pi(\mathbf{r}) = P_2(\mathbf{r}, \mathbf{r}) = d(\mathbf{r}, \mathbf{r}, \mathbf{r}, \mathbf{r})$).

Thus, the thesis of MC-PDFT is that the remaining dynamic correlation untreated by the active space can be captured *locally*, similar to KS-DFT, with the important muticonfigurational effects captured in the shape of the density $\rho(\mathbf{r})$ and one-body density matrix D_{pq}. One may argue that this is somewhat similar to how exchange is already treated in

KS-DFT, as the "exact" exchange term arising from symmetry is thrown out in the energy expression (equation 1.3) in place of a local functional, despite its implicit inclusion of exchange in the determinant form of the ansatz (equation 1.1). The importance of the 2-RDM and its small contribution to the correlation functional in MC-PDFT is currently under debate, with the multiconfigurational density coherence functional theory (MC-DFT)[42] being able to achieve good results in test cases without utilization of the 2-RDM. Calculating the energies of CASSCF wave functions using MC-PDFT has proved to be a very successful approach in many cases,[43] and has generally reproduced the results of the perturbation theories at significantly less expense.[6]

Nevertheless, despite the several successful applications of CASSCF in the literature and the increasing facility of obtaining quantitative energies with methods such as MC-PDFT, it has avoided large-scale application anywhere close to the use of popular methods such as KS-DFT. This is principally due to the problem of *active space selection* in utilizing CASSCF: how does one efficiently select the space of orbitals and electrons to treat at a high level for any given problem? While progress has been made on a case-by-case basis through selecting active spaces by-hand with trial and error,[30] carrying out CASSCF calculations in an automated and consistent fashion remains an open challenge. Overcoming this issue has been a highly active area of research over the past decade (as will be discussed thoroughly in the many chapters below), and concerns the principal contributions of this thesis.

1.5 Methods for Active Space Selection

The key contribution to the field of active space selection made in this thesis is the development of the "approximate pair coefficient" (APC) approach for estimating the importance of orbials for the active space prior to calculation.[5] This scheme employs Hartree-Fock matrix elements (nearly always calculated prior to initializing a multiconfigurational calculation) to estimate the interactions between different of inactive orbitals i and virtual orbitals a:

$$C_{ia} = \frac{0.5K_{aa}}{F_{aa} - F_{ii} + \sqrt{(0.5K_{aa})^2 + (F_{aa} - F_{ii})^2}} \tag{1.9}$$

in which F_{ii}, K_{ii}, F_{aa}, and K_{aa} are the diagonal Fock and exchange matrix elements of the orbitals, respectively. These "approximate pair coefficients" C_{ia} are then used to estimate the multiconfigurational character of different orbitals via their estimated Von-Neummann entropies over the probability of these orbitals being occupied or unoccupied (with intermediate normalization):

$$S_i = -\frac{1}{1 + \sum_a C_{ia}^2} \ln \frac{1}{1 + \sum_a C_{ia}^2} - \frac{\sum_a C_{ia}^2}{1 + \sum_a C_{ia}^2} \ln \frac{\sum_a C_{ia}^2}{1 + \sum_a C_{ia}^2} \tag{1.10}$$

$$S_a = -\frac{1}{1 + \sum_i C_{ia}^2} \ln \frac{1}{1 + \sum_i C_{ia}^2} - \frac{\sum_i C_{ia}^2}{1 + \sum_i C_{ia}^2} \ln \frac{\sum_i C_{ia}^2}{1 + \sum_i C_{ia}^2} \tag{1.11}$$

This schema is derived from a simple (2,2) FCI model calculation (e.g. minimal basis H_2), in which case the relationship is exact. This method was first published in the JCTC article "A Ranked-Orbital Approach to Selecting Active Spaces for High-Throughput Multireference Computation", which is reproduced here as chapter 2 of this thesis. This initial work tested the APC method in selecting active spaces from different sets of orbitals on a small subset of 20 test systems, for which it demonstrated good results.

The first large-scale test of the APC method took place in 2022 with the publication of "Large-Scale Benchmarking of Multireference Vertical-excitation Calculations via Automated Active-Space Selection" in JCTC,[6] which tested APC on the large and diverse QUESTDB database of vertical excitation energies gathered by Jacquemin, Loos, and coworkers.[44,45] Here, APC was shown to select good active spaces for about 80% of vertical excitations, depending on the active space and basis set size. This accuracy allowed for the first benchmarking of post-CASSCF methods such as NEVPT2 and MC-PDFT using automated and consistent active spaces, showing them to have errors of roughly 0.2 eV. These wave functions

(with full orbitals and CI vectors in the PySCF format) are freely available on Zenodo[46] and have already served as an excellent resource for the benchmarking of new post-CASSCF methods such as linearized pair-density functional theory.[47] This work is reproduced here in chapter 3.

However, despite the overall success in APC in selecting good active spaces for these excitations, a quantitative benchmarking was only possible by eliminating the 20% of poor active spaces by careful thresholding of the CASSCF error. This final 20% of poor active spaces was addressed in the 2023 publication "Variational Active Space Selection with Multiconfiguration Pair-Density Functional Theory", which developed a variational scheme to select between different active spaces based on energies from MC-PDFT. The variational selection in this case was achieved by constructing different sets of virtual orbitals through diagonalization of the operator $F - \lambda K$ with $\lambda \in [0, 2]$ to target between valence and Rydberg orbitals. This scheme in combination with DMRG solvers used to select larger active spaces (in this case without orbital optimization) was able to select good active spaces for 100% of the QUESTDB excitations and reproduced the good results found in the previous study.[6] This work is reproduced here in chapter 4.

1.6 Applications of Active Space Selection

The second half of this thesis, chapters 5, 6, and 7, concern different facets of applications for the automated multiconfigurational approaches developed in chapters 2, 3, and 4. The first of these chapters, chapter 5, titled "Machine-Learned Energy Functionals for Multiconfigurational Wave Functions" concerns the use of automated multiconfigurational data in machine learning applications, and reproduces work published in JPCL.[48] This work employed vertical excitation calculations on carbenes automated with APC to train a new MC-PDFT functional as a direct functional of the on-top density and pair-density.[5] The use of data from automated multiconfigurational calculations to train novel functionals and force fields

remains a key area for future work, and may well be the most impactful long-term use of these methods.

The second of these chapters, chapter 6, titled "Divergent Bimetallic Mechanisms in Copper(II)-Mediated C–C, N–N, and O–O Oxidative Coupling Reactions", reproduces work recently published in JACS. This work presents a detailed computational study of the Cu(II) imines in coupling together different substrates (imine, NH_2, OH, CN, and CCH), and was a collaboration with the group of Shannon Stahl at the University of Wisconsin. This work employed a set of automated multiconfigurational calclulations to characterize key transition states, which helped to distinguish between the biradical coupling mechanism found for imine, NH_2, OH, and a "diamond core" bimetallic coupling found for CN and CCH.

Finally, the last of these chapters, chapter 7, titled "Organic Reactivity Made Easy and Accurate with Automated Multireference Calculations", reproduces work recently published in *ACS Central Science*. This work presents a fruitful collaboration with the group of Brett Savoie at Purdue in which we undertook the automated characterization of hundreds of algorithmically generated transition states for organic reactions. We showed that a majority of these transition states possess significant multireference character, and that our automated APC scheme was able to reproduce the results of KS-DFT in cases of weak correlation while providing improvements over both KS-DFT and coupled cluster methods in cases of strong correlation. I believe this study was a huge step forward for the field of treating reactivity with multiconfigurational approaches, as we showed that a combination of automated active space selection combined with MC-PDFT is able to significantly overcome the troublesome "active space inconsistency error" in performing these types of calculations, to the point of being predictive in most cases.

1.7 Conclusions and Outlook

This introductory chapter has noted two "revolutions" in the treatment of correlation within electronic structure theory: the first revolution, in which exchange correlation was efficiently captured by the Hartree-Fock method, and the second revolution, in which correlation was efficiently captured locally by use of an exchange-correlation functional. It is my belief that we are now on the midst of a third revolution in electronic structure theory, in which we are learning to efficiently capture the remaining static, non-local correlation. Methods such as DMRG have proved efficient solvers for handling large and complex active spaces, and MC-PDFT has proved a robust method for capturing the remaining correlation outside of the active space. The work of active space selection presents what is perhaps the final barrier to seeing this revolution take place at large scale, at least for several important applications. With the remarkable success of even imperfect automated schemes such as shown in chapter 7 in providing accurate and predictive descriptions of chemical systems, we may be closer to this revolution taking place than we think.

However, much work remains to be done. While the APC method has proved remarkably successful for organic systems, transition metal complexes have proved to be more difficult to model and less well-treated by APC. This likely follows naturally from the fact that the minimal (2,2) orbital picture assumed by the derivation of the APC equation (1.9) breaks down when moving to the transition metal picture in which there are more than two localized near-degenerate orbitals. Different or new schemes will likely be necessary to treat these cases; for example, the work presented here on Cu N–N coupling relied on the atomic valence active space (AVAS) concept developed by Sayfutyarova and co-workers.[49] Perhaps combinations of this method with APC can play a role in extending selection to larger complexes. However, it is not clear how to treat excitations of transition metal complexes within the AVAS picture. My hope is that future developments may take inspiration from the variational concepts developed in chapter 4.

1.8 Other Works

I will conclude this introductory section by quickly touching on a few works accomplished during my PhD but not in the scope of active space selection and thus not included in this thesis. The largest project excluded from this thesis is a collaboration with experimentalists at Argonne National Laboratory, in which we used high-throughput experimentation to enhance the catalytic yield of metals deposited onto metal-organic frameworks. This was a quite fruitful collaboration for which I was able to receive SCGSR funding to work closely with experimental collaborators in integrating theory, and I was very thankful to be able to work in this direction due to their support. This work was recently published in *ACS Central Science.*[50]

Another project I would like to mention is a publication in *J. Phys. Chem. C.* titled the "Challenge of Small Energy Differences in Metal–Organic Framework Reactivity",[51] on which I was able to work with the talented undergraduate (now graduated) student Noah Dohrmann on the sensitivity of density functionals for describing small trends in metal organic framework reactivity. I believe this work complements the chapters included in this dissertation in showing the need for more accurate and sensitive electronic structure theories. A complete list of unincluded publications completed during my PhD can be found below.

I hope this chapter has served well as an introduction and motivation to the following chapters, and I thank everyone who has read this far for their interest in my work.

1.9 Appendix: List of Publications

Included as chapters in this thesis:

- **King, D.S.**; Gagliardi, L. A Ranked-Orbital Approach to Select Active Spaces for High-Throughput Multireference Computation. *J. Chem. Theory Comput.* **2021**, *17*, 2817-2831.

- **King, D.S.**; Hermes, M.R.; Truhlar, D.G.; Gagliardi, L. Large-Scale Benchmarking of Multireference Vertical-Excitation Calculations via Automated Active-Space Selection. *J. Chem. Theory Comput.* **2022**, *18*, 6065-6076.

- **King, D.S.**; Truhlar, D.G.; Gagliardi, L. Variational Active Space Selection with Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2023** *19* (22), 8118-8128

- **King,D.S.**; Truhlar, D.G.; Gagliardi, L. Machine-Learned Energy Functionals for Multiconfigurational Wave Functions. *J. Phys. Chem. Lett.* **2021**, *12*, 7761-7767.

- **King, D.S.**; Wang, F.; Gerken, J.B.; Gaggioli, C.A.; Guzei, I. A.; Jung, Y.K.; Stahl, S.; Gagliardi, L. Divergent Bimetallic Mechanisms in Copper(II)-Mediated C–C, N–N, and O–O Oxidative Coupling Reactions. *J. Am. Chem. Soc.* **2024**, *146* (5), 3521–3530

- Wardzala, J.W.†; **King, D.S.**†; Ogunfowora, L.; Savoie, B.; Gagliardi, L. Organic Reactivity Made Easy and Accurate with Atomated Multireference Calculations. *ACS. Cent. Sci.* **2024**

Publications not included as chapters:

- McCullough, K.E.†; **King,D.S.**†, Chheda, S.P.; Ferrandon, M.S.; Goetjen, T.A.; Syed, Z.H.; Graham, T.R.; Washton, N.M.; Farha, O.K.; Gagliardi, L.; Delferro, M. High-Throughput Experimentation, Theoretical Modeling, and Human Intuition: Lessons Learned in Metal-Organic-Framework-Supported Catalyst Design. *ACS Cent. Sci.* **2023**, *9*, 266-276.

- Dohrmann, N.†; **King, D.S.**†; Gaggioli, C.A.; Gagliardi, L. Challenge of Small Energy Differences in Metal–Organic Framework Reactivity *J. Phys. Chem. C* **2023**, *27*, 16891–16900

- Hennefarth, M.R.; **King, D.S**; Gagliardi, L. Linearized Pair-Density Functional Theory for Vertical Excitation Energies. *J Chem. Theory Comput.* **2023**, *19* (22), 7983-7988

- Zhou, C.; Hermes, M.; Wu, D.; Bao, J.J.; Pandharkar, R.; **King, D.S.**; Zhang, D.; Scott, T.; Lykhin, A.; Gagliardi, L.; Truhlar, D.G. Electronic Structure of Strongly Correlated Systems: Recent Developments in Multiconfiguration Pair-Density Functional Theory and Multiconfiguration Nonclassical-Energy Functional Theory. *Chem. Sci.* **2022**, *13*, 7685-7706.

- Jeong, W.; Stoneburner, S.J.; **King, D.S.**; Li, R.; Walker, A.; Lindh, R.; Gagliardi, L. Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation. *J. Chem. Theory Comput.* **2020**, *16*, 2389-2399.

CHAPTER 2

A RANKED-ORBITAL APPROACH TO SELECTING ACTIVE SPACES FOR HIGH-THROUGHPUT MULTIREFERENCE COMPUTATION

This chapter is reprinted with permissions from *J. Chem. Theory Comput.* **2021**, *17*, 5, 2817–2831

2.1 Abstract

The past decade has seen a great increase in the application of high-throughput computation to a variety of important problems in chemistry. However, one area which has been resistant to the high-throughput approach is multireference wave function methods, in large part due to the technicalities of setting up these calculations and in particular the not always intuitive challenge of active space selection. As we look towards a future of applying high-throughput computation to all areas of chemistry, it is important to prepare these methods for large-scale automation. Here, we propose a ranked-orbital approach to selecting active spaces with the goal of standardizing multireference methods for high-throughput computation. This method allows for the meaningful comparison of different active space selection schemes and orbital localizations, and we demonstrate the utility of this approach across 1120 multireference calculations for the excitation energies of small molecules; results reveal that it is helpful to distinguish the method used to generate orbitals from the method of ranking orbitals in terms of importance for the active space. Additionally, we propose our own active space selection scheme that estimates the importance of an orbital for the active space through a pair-interaction framework from orbital energies and features of the Hartree-Fock exchange matrix. We call this new scheme the "Approximate Pair Coefficient" (APC) method and it performs quite well for the test systems presented.

2.2 Introduction

In the past decade, the explosion of computational resources has led to the ability of many researchers to carry out high-throughput computational screenings of molecules and materials for several important applications in electrocatalysis,[52] gas storage,[53] and photochemistry.[54–56] Currently, these approaches rely overwhelmingly on some combination of molecular mechanics, semiempirical theories, density functional theory (DFT), and single-reference wave function theories (e.g. CCSD(T)) to calculate properties of interest.[52–56] However, one area where the high-throughput approach is poised to play a large role is in the development of new transition metal catalysts,[57,58] and these complexes are often strongly correlated and thus poorly described by single-reference methods such as DFT.[26,59–63] Furthermore, DFT frequently suffers from an inability to describe open-shell systems without resorting to broken-symmetry solutions,[61] and this becomes particularly severe when multiple low-lying spin states are important to consider (e.g. in application to spin-crossover complexes[64,65]). This feature also makes DFT difficult to use for the description of electronic excited states, and particularly at geometries far from equilibrium where these structures exhibit even stronger correlation.[63]

For these reasons, we expect that the expansion of reliable high-throughput computation to these problems will require the use of multireference approaches.[26] In addition to adding value in these cases, high-throughput multireference calculations have the potential to provide high-quality benchmarks and training data for new density functional and machine-learned approximations. Looking towards this future, recent research has gone into identifying chemical systems where multireference approaches would provide added value over DFT,[66] and here we have the goal of standardizing these calculations to run in an automated and robust fashion.

To achieve this, we turn our focus towards a unique issue that stands in the way of the most widely used multireference methods, which is the problem of active space selection.

In active space multireference computation the user must limit the size of the calculation uniquely for each system by selecting an "active space" of orbitals in which to expand the wave function configurationally, which is an approximation known as the "complete active space" (CAS) ansatz:

$$|\psi_{CAS}\rangle = \sum_{n_1 n_2 ... n_n} c_{n_1 n_2 ... n_n} |22...n_1 n_2 ... n_n 00...\rangle \tag{2.1}$$

In the above, n_i are the varying occupations $(0, \uparrow, \downarrow, 2)$ of the active space orbitals, and $c_{n_1 n_2 ... n_n}$ are the coefficients of each determinant $|22...n_1 n_2 ... n_n 00...\rangle$. Orbitals not in the active space have either constant 2 (inactive) or constant 0 (virtual or secondary) occupation.[26] The number of alpha and beta electrons in the active space is conserved in all determinants in the expansion (equation 2.1), and this number of electrons is generally set by the number of electrons in the occupied orbitals selected. The size of the chosen active space is commonly expressed as a number of electrons in a number of orbitals (N_{elec},N_{orbs}). For a wave function of maximum spin component along the laboratory axis ($S = S_z$), the effective degrees of freedom in equation 2.1 can be expressed through the number of "configuration state functions" (CSFs) as[67]

$$N_{CSF} = \binom{L}{\alpha}\binom{L}{\beta} - \binom{L}{\alpha+1}\binom{L}{\beta-1} \tag{2.2}$$

where L is the number of orbitals and α and β the number of alpha and beta electrons in the active space, respectively.

Today, there are many approaches for optimizing the CAS ansatz. Obtaining the coefficients in equation 2.1 through exact diagonalization is known as CASCI, while optimizing the orbitals and the coefficients simultaneously is known as CASSCF.[68] Currently, the maximum active space that can be computed with these methods is about (20,20).[26] To expand beyond this limit, several methods exist for approximating the coefficients in equation 2.1,

such as the density matrix renormalization group method (DMRG),[69–71] full configuration interaction quantum Monte Carlo (FCIQMC),[27,72] and neural networks.[73] These approximate methods can handle up to hundreds of orbitals in the active space.[31] Often, the active orbitals are variationally optimized in tandem with the coefficients in equation 2.1, which spawns methods with the self-consistent field (SCF) suffix (e.g. CASSCF and DMRGSCF). Additionally, results from CAS-type wave functions are often enhanced through the addition of dynamic correlation through multireference perturbation theory via CASPT2[74] or n-electron valence perturbation theory (NEVPT2).[35] The addition of dynamical correlation through these methods limits calculations to only about 14 orbitals in the active space.[26]

Regardless of the method used to optimize or improve the CAS wave function, an active space must be selected. Even if the orbitals are variationally optimized as in CASSCF, the initial guess can greatly influence the quality of the result obtained due to the variety of local minima on the optimization surface.[30] If one selects an active space that is too large, the calculation becomes exponentially more expensive and potentially unaffordable, while if one selects an active space that is too small or that does not include the important orbitals, the wave function can be qualitatively wrong. The past five years have seen a large amount of research on the topic of automatically selecting the active space.[1,31,49,75–84]

A new approach for selecting active spaces that has gathered a lot of attention in recent years goes by the name of AutoCAS,[31,76] and is centered around the idea of choosing orbitals that vary in occupation $(0, \uparrow, \downarrow, 2)$ within a low-cost or even partially converged DMRG calculation. This variance is measured through the single-orbital entropy, given for an orbital i as[85]

$$S_i = -\sum_{j=\{0,\uparrow,\downarrow,2\}} \rho^i_{jj} \ln \rho^i_{jj} \tag{2.3}$$

where ρ^i is the one-orbital reduced density matrix for orbital i, the configurational analogue of the one-particle reduced density matrix obtained by tracing over all other configurational

degrees of freedom:[85]

$$\rho^i = \sum_{\{\mathbf{n}\neq n^i\}} \langle \mathbf{n}|\psi\rangle \langle \psi|\mathbf{n}\rangle = \sum_{kj} c_k c_j^* \left(\sum_{\{\mathbf{n}\neq n^i\}} \langle \mathbf{n}|\mathbf{n}_k\rangle \langle \mathbf{n}_j|\mathbf{n}\rangle \right) \left|n_k^i\right\rangle \left\langle n_j^i\right| \tag{2.4}$$

where $\mathbf{n} \neq n^i$ are all possible occupations of the other orbitals (excluding orbital i). Note that the orbital entropies are state, localization, and orbital-dependent.

Stein and Reiher have proposed two schemes for selecting the orbitals from the orbital entropies above: one scheme based off a global threshold and a more complicated flowchart scheme based off of identifying plateaus in threshold diagrams.[76] In this work we investigate the performance of the former scheme due to the difficulty of using the latter outside of the software package they have released to run calculations in OpenMolcas and the similarity of its results to the latter in several reported cases.[31,67,76] To make things clearer, we refer to this global threshold scheme as "EntropyCAS" and to the more complicated scheme as "AutoCAS" to avoid confusion. The EntropyCAS procedure suggested by Stein and Reiher is to choose all orbitals with orbital entropy $S_i > 0.1 S_{max}$, where S_{max} is the entropy of the highest-entropy orbital in the ground state.[31,76] When multiple states are considered, Stein and Reiher suggest selecting the union of all orbitals with $S > 0.1 S_{max}$ in their respective states,[86] and we refer here to this extended and more expensive scheme as EntropyCAS+.

Here, we investigate the performance of the EntropyCAS and EntropyCAS+ procedures for the problem of computing ground-state to first-excited-state singlet ($S_0 \rightarrow S_1$) and doublet ($D_0 \rightarrow D_1$) excitation energies for twenty small molecules using state-averaged (SA) CASSCF/ NEVPT2 calculations, as was recently investigated by Bao and Truhlar.[1] We find that while the EntropyCAS procedure excels at detecting good orbitals for the active space, the threshold scheme proposed by Stein and Reiher is too unwieldy for high-throughput computation.

To remedy this, a modified ranked-orbital procedure is proposed which provides consistently-

sized active spaces and allows us to compare the quality of both active space orbitals and orbital selection schemes for this problem. This ranked-orbital procedure is extended to two other threshold selection schemes, high-spin UNO[1,87] and AVAS,[49] with similar results. To demonstrate the robustness of this approach for high-throughput computation, we carry out 1120 SA-CASSCF/NEVPT2 calculations which serves to highlight trends in the application of the CASSCF/NEVPT2 method.

Finally, with the goal of accelerating high-throughput computation with the ranked-orbital EntropyCAS procedure, we attempt to approximate the orbital entropy from a pair-coefficient framework by the readily available molecular orbital energies and elements of the exchange matrix from Hartree-Fock. This new approximation is called the "Approximate Pair Coefficient" (APC) method, and performs about equivalently to the modified EntropyCAS procedure for these simple systems. Taking inspiration from molecular-orbital based machine learning (MOB-ML),[88] we attempt to improve this approximation through a machine learning scheme using more information from the Hartree-Fock matrices. While we find that this improvement has little effect on the performance of the active spaces for these simple systems, we hope that the work here inspires future efforts in approximating the orbital entropies for more complex cases.

2.3 Methods

Excitation Energies. Geometries and reference values for excitation energies were taken from the previous work of Bao and coworkers.[1,77] Here, we select a subset of these reference values consisting of $S_0 \rightarrow S_1$ or $D_0 \rightarrow D_1$ excitation energies of 12 singlet and 8 doublet DFT-optimized structures, with singlet reference values taken from experiment and doublet reference values obtained from high-quality multireference configuration interaction calculations with the Davidson correction (MRCI+Q). Although these systems are small, the majority of their first excited states are thoroughly multiconfigurational (supporting infor-

mation). After the active space is selected by various schemes, final SA-CASSCF/NEVPT2 calculations with the aug-cc-pVTZ basis[89] were performed using PySCF,[90] with state averaging done over the five lowest-energy states with the same spin as the ground state. The maximum number of macro cycles in the CASSCF optimization procedure was set to 200, and the CASSCF orbitals were taken regardless of convergence after this limit was reached (sloppy convergence, see supporting information).

EntropyCAS/EntropyCAS+. Orbital entropies for orbitals generated in PySCF were calculated by interfacing with QCMaquis[71] via the FCIDUMP[91] file interface. DMRG calculations were initialized using the CIDEAS initial guess,[92] and information from this initial calculation was used to employ an optimized Fiedler ordering[93] of the DMRG orbitals for a larger calculation with a bond order of $M = 450$. Then, to ensure convergence of the orbital entropies, information from this $M = 450$ calculation was used to initialize a larger $M = 500$ calculation with an updated Fiedler ordering and if necessary this process was repeated increasing M by 50 until all orbital entropies were converged to within 0.01 units. Entropies for the first excited state were calculated in tandem by enforcing orthogonality to the ground state at each step (guess, $M = 450$, $M = 500$...). This process was continued until convergence was met in both the ground and first excited state entropies.

Orbitals for this procedure were generated from ROHF solutions and several localization schemes (canonical (HF), Boys,[94] Pipek-Mezey with Löwdin charges (PM),[95,96] and Edmiston-Ruedenberg[97] (ER)) implemented in PySCF. Orbitals were localized in a split-localized procedure where the doubly-occupied orbitals were localized in a space of all doubly occupied orbitals and the virtual orbitals were localized in a space of 40 virtual orbitals, selected by two different schemes (supporting information); any singly-occupied orbitals remained unchanged.

High-Spin Unrestricted Natural Orbitals (UNO(HS)). Selecting UHF natural orbitals (UNOs) for the active space based on their occupation number is one of the oldest schemes

for selecting active spaces,[87] but as recently noted is still capable of selecting good active spaces for many difficult systems.[83] However, because all systems here are at equilibrium geometry and weakly correlated, the standard UNO-CAS procedure is not viable as an unrestricted UHF solution does not exist separately from the RHF solution at many of these geometries. To amend this, we take inspiration from the work of Bao and Truhlar[1] who used high-spin UHF natural orbitals to construct active spaces for these systems. For singlet systems, we compute the UHF wave function with $S_z = 2$ and for doublet systems we compute the UHF wave function with $S_z = 5/2$. The natural orbitals and occupation numbers are then obtained by solving the relevant eigenvalue problem,[87]

$$S^{1/2}(D_\alpha + D_\beta)S^{1/2}(S^{1/2}C) = \sigma(S^{1/2}C) \tag{2.5}$$

where S is the atomic orbital overlap matrix, C is the molecular orbital coefficient matrix of the UNOs to be obtained, σ is a diagonal matrix containing the occupation numbers, and D_α and D_β are the alpha and beta density matrices in the atomic orbital basis.

AVAS. The atomic valence active space (AVAS) method was published by Sayfutyarova and coworkers in 2017 and is based on the insight that active spaces are generally selected by thinking about atomic orbitals and not molecular ones.[49] Once a single-determinant wave function is acquired, the user selects a set of A atomic orbitals from a minimal basis, and then the doubly occupied and virtual orbitals are localized separately to form a basis of at most $2A$ molecular orbitals that completely embed the user-selected atomic orbitals. A question remains for singly-occupied orbitals, and the authors suggest a few ways of dealing with these. Here we calculate the wave function of all doublet structures using ROHF determinants, and use the approach suggested by Sayfutyarova and coworkers of carrying over all singly-occupied molecular orbitals into the active space without localization.

As described, AVAS is not strictly a fully automatic scheme as the user must select the atomic orbitals to embed by hand. However, the singular values from the singular value

decomposition used to embed the orbitals are suggested by the authors as a way to qualify orbitals for the active space. To construct a fully automatic scheme we ask AVAS to embed all the valence orbitals in a minimal basis for the system and use the provided singular values as qualifiers in the ranked-orbital procedure described below. The AVAS orbitals and singular values were obtained via its implementation within PySCF.

Figures. Most figures were generated with Seaborn,[98] which calculates 95% confidence intervals by bootstrapping the mean value over 1000 random samplings. Orbital isosurfaces enclosing 80% of orbital electron density were generated using IboView.[99]

2.4 Results and Discussion

2.4.1 The Limitations of Threshold Schemes

As a simple first test of the viability of EntropyCAS and EntropyCAS+ as high-throughput active space selection schemes, we chose to select active spaces for 20 excitation energies of small systems investigated previously by Bao and Truhlar.[1] To calculate the excited states of ethylene, Stein and Reiher calculated entropies for and selected the active space from 12 valence and 12 Rydberg orbitals generated from a Hartree-Fock calculation in the large ANO-RCC[100–102] basis (8s8p4d3f2g for carbon and 6s4p3d1f for hydrogen).[86] Here, we calculate entropies for and select orbitals from the lowest 30 orbitals in energy for all twenty systems investigated. To investigate the effect of orbital localization in the EntropyCAS and EntropyCAS+ methods we used both canonical (HF) and Boys-localized orbitals (EntropyCAS/HF, EntropyCAS+/HF, EntropyCAS/Boys and EntropyCAS+/Boys).

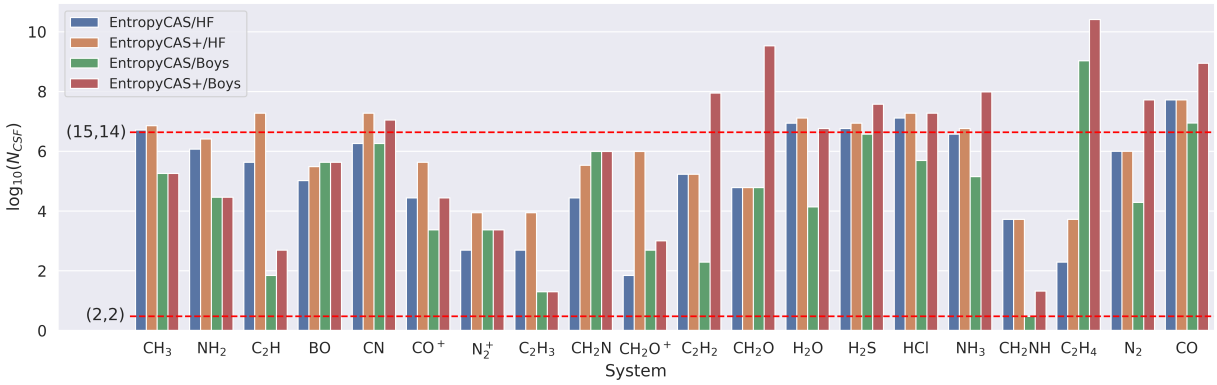

Figure 2.1: Comparison of the size of the active spaces selected by EntropyCAS/HF, EntropyCAS+/HF, EntropyCAS/Boys, and EntropyCAS+/Boys for the twenty different systems investigated, as plotted by the base-10 logarithm of the number of configurations in the selected active space ($\log_{10} N_{CSF}$). No method selects spaces for all systems under the affordable CASSCF/NEVPT2 limit of (15,14) (top horizontal dotted line).

The active space selections of EntropyCAS and EntropyCAS+ using the $0.1S_{max}$ threshold suggested by Stein and Reiher are shown in figure 2.1, plotted against the $\log_{10} N_{CSF}$ of configurations in the selected active space, with N_{CSF} as calculated by equation 3.4. Dotted red lines indicate the $\log_{10} N_{CSF}$ for a minimum active space of (2,2) and the maximum affordable active space using CASSCF/NEVPT2 of (15,14).[26] We find that for many systems the active spaces selected by the EntropyCAS and EntropyCAS+ procedures are larger in size than the affordable (15,14) limit for CASSCF/NEVPT2 and that no method selects an active space below this limit for all systems. Furthermore, because the orbital entropies are localization dependent,[103] the size of the selected active space varies heavily by system, orbital localization, and selection method (EntropyCAS vs. EntropyCAS+). We use these results to highlight what we believe to be limitations of threshold schemes:

- Threshold schemes have no regard for the affordability of the selected active space for the CAS method being employed. This makes them difficult to adapt for the large variety of CAS methods (exact diagonalization, DMRG, FCIQMC) that vary heavily in their preferred active space size.

- Threshold schemes are very hard to compare with one another. For example, if one approach (e.g. EntropyCAS+) selects an active space with many orders of magnitude more configurations than another scheme (e.g. EntropyCAS), it makes it very difficult to meaningfully compare these results.

- Threshold schemes make it difficult to compare orbitals with one another, as in cases where localizing the orbitals results in an active space with orders of magnitude less configurations (e.g. C_2H) or more configurations (e.g. C_2H_4).

- The variability of active space sizes selected by threshold schemes makes it hard to automate. Spaces with orders of magnitude more configurations will require drastically different amounts of computational resources than smaller ones.

While one may hope to remedy the issues above through more sophisticated selection schemes such as AutoCAS or by reiterating the selection scheme on converged CASSCF orbitals, this in principle does not avoid any of the criticisms above. Although more sophisticated schemes such as AutoCAS would likely choose much smaller spaces, any scheme which determines the size of the active space from developer-set parameters retains its agnostic nature towards the CAS solver the user would like to use and is thus prone to choose a space that is too small or too large (indeed, the scheme of Stein and Reiher was developed in the context of DMRG).[31,76] Similarly, due to the variety of CASSCF minima on the orbital optimization surface, there is no reason to think that a reiterative scheme would completely eliminate all variance in the final active space size with respect to the starting basis employed.

Despite these issues, on physical grounds the orbital entropies used by the EntropyCAS and EntropyCAS+ procedures stand on good terms; it is simply the procedural act of selecting the active spaces via a threshold scheme that results in these unfavorable qualities. These problems also apply to threshold occupation number schemes such as UNO-CAS.[87]

Here, we modify these threshold schemes in order to select consistent, flexible, and affordable active spaces for high-throughput computation.

2.4.2 The Ranked-Orbital Approach to Selecting Active Spaces

One will note that all the problems with threshold schemes mentioned above can be resolved by simply requiring threshold schemes to select active spaces of a consistent size. How to go about doing this in a flexible way, however (as opposed to, for example, simply limiting the number of orbitals in the active space) is an open question. Here, we propose a ranked-orbital approach to selecting active spaces that can easily be adapted to any threshold scheme:

1. The user specifies a maximum CAS space of $\max(N_{elec}, N_{orbs})$ and this space is converted to a maximum N_{CSF}^{MAX} via equation 3.4, with $S = S_z = 0$ for an even number of electrons and $S = S_z = 1/2$ for an odd number of electrons.

2. The selection scheme *ranks* all candidate orbitals in order of importance

3. The lowest-importance orbital is repeatedly dropped from the active space until $N_{CSF} \leq N_{CSF}^{MAX}$

4. If an orbital is dropped that results in an unreasonable active space (with reasonability here defined as having at least one occupied orbital and two unoccupied orbitals in the active space, to ensure stability of the CASSCF solver), the next lowest orbital is dropped instead.

We note that all that is strictly required by the above algorithm is the maximum number of CSFs, N_{CSF}^{MAX}. However, as computational chemists rarely discuss active spaces in this language, we have the user set N_{CSF}^{MAX} with reference to the size of a real active space: max(7,6) ($N_{CSF}^{MAX} = 490$), max(8,8) ($N_{CSF}^{MAX} = 1764$), max(10,10) ($N_{CSF}^{MAX} = 19404$), and max(12,12) ($N_{CSF}^{MAX} = 226512$), etc. Here we convert the following threshold schemes in this way:

- High-spin UNO-CAS (UNO(HS)), with natural orbitals ranked by the absolute deviation of their occupation number from 0 or 2.

- EntropyCAS, with arbitrary candidate orbitals ranked by their orbital entropy from DMRG

- EntropyCAS+, with arbitrary candidate orbitals ranked by their average, max-normalized orbital entropy from DMRG in all relevant states (here the ground and first excited state):

$$S_i = \frac{1}{N} \sum_{n}^{N} \frac{S_{ni}}{S_n^{max}} \tag{2.6}$$

- AVAS, with SVD orbitals ranked by their singular values from embedding all valence orbitals in a minimal basis

To illustrate the robustness of this approach for high-throughput computation, we now evaluate the performance of the above schemes in the ranked-orbital procedure by having them choose active spaces for the 20 small-molecule excitation energies mentioned previously with maximum active space sizes of max(7,6) and max(10,10). At the max(7,6) level, good results should be obtainable with a mean error of about 0.17 eV, as achieved by Bao and Truhlar with spaces of about this size using jun-cc-pvTZ and CASPT2.[1] The max(10,10) level is chosen to demonstrate that the modified schemes are able to select active spaces with roughly two orders of magnitude more configurations that correspondingly improve the results obtained.

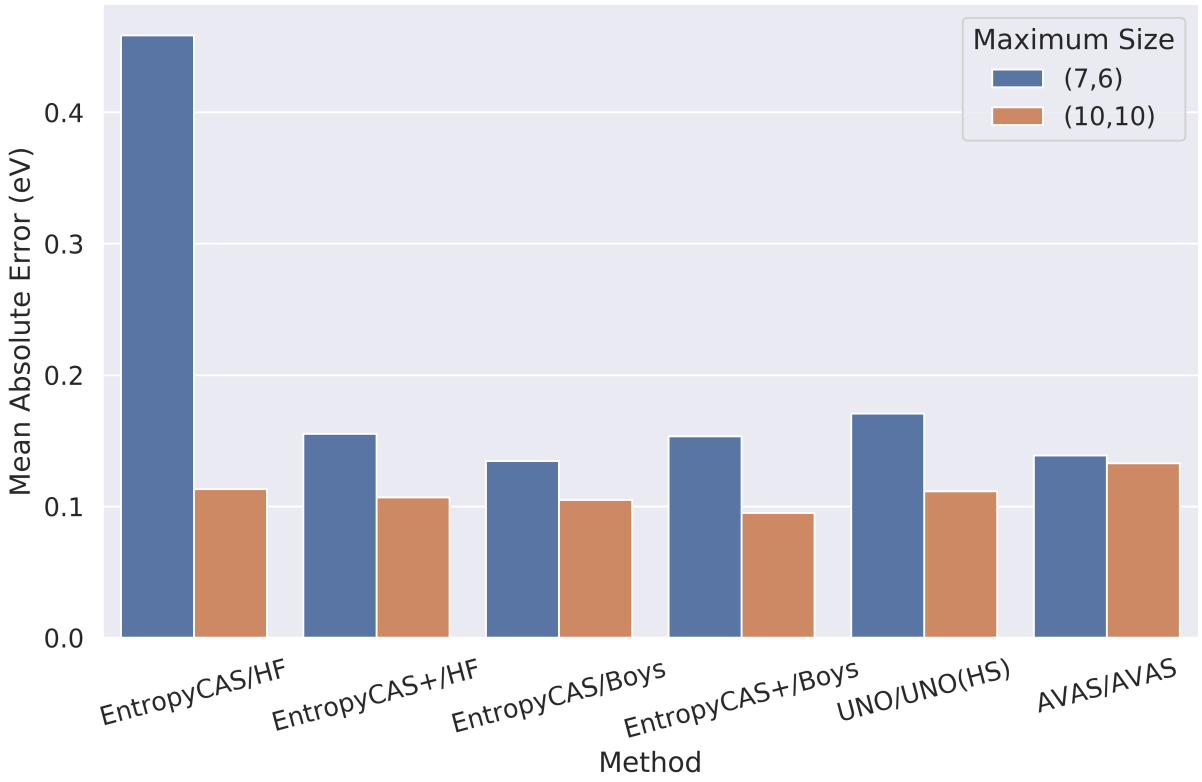

Figure 2.2: Performance of six different threshold schemes that have been modified by the ranked-orbital procedure at maximum active space sizes of (7,6) and (10,10). The ranked-orbital scheme allows for a meaningful comparison between active space selection schemes, orbital localization schemes, and active space sizes.

The results of six different threshold schemes that have been modified by the ranked-orbital procedure over the 20 excitation energies in the test set are plotted in figure 2.2. Because the selected active spaces are limited to a consistent size, the scheme allows for a meaningful comparison between the quality of different methods. For example, the best method at max(7,6) is EntropyCAS/Boys with a mean error of 0.13eV while the best method at max(10,10) is EntropyCAS+/Boys with a mean error of 0.11eV. EntropyCAS/HF performs quite poorly at the max(7,6) level, mostly due to a poor selection for CH_2O (supporting information).

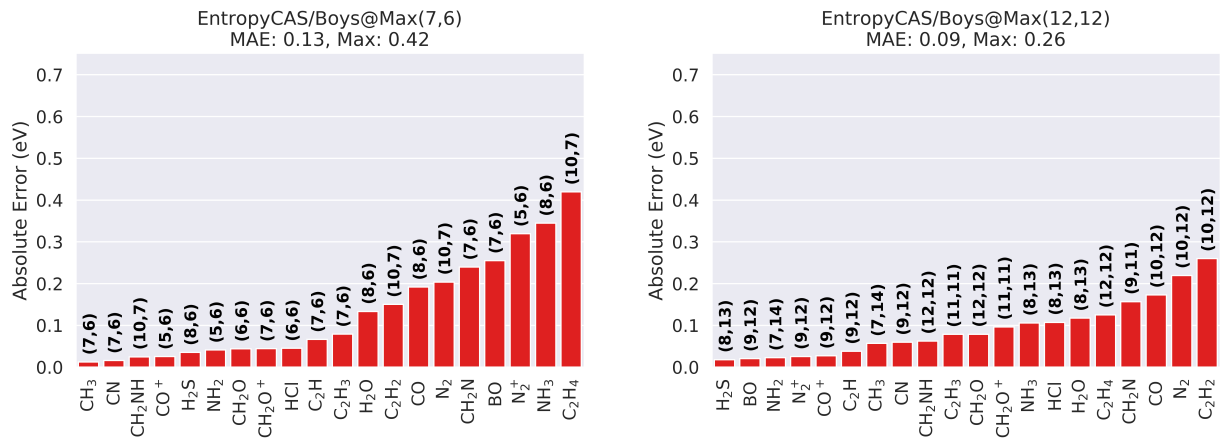

Figure 2.3: Performance of the modified ranked-orbital EntropyCAS scheme at max(7,6) and max(12,12). The ranked-orbital procedure allows for flexibility in the chosen active space while being able to select calculations of a consistent and managable size, by fixing the maximum number of CSFs.

To illustrate the flexibility of the ranked-orbital procedure in selecting active spaces of a consistent size while maintaining flexibility, the performance and selection of the ranked-orbital EntropyCAS procedure using Boys orbitals at max(7,6) and max(12,12) is shown in figure 2.3. It can be seen that limiting only the number of CSFs (and not the number of orbitals or electrons) allows for different numbers of orbitals and electrons to be selected for each system. For example, active spaces from (10,7) to (5,6) are selected at the max(7,6) level and from (9,12) to (7,14) at the max(12,12) level.

Benefits and Drawbacks. The main benefits of the ranked-orbital scheme are that it resolves all of the rather critical limitations of threshold schemes for high-throughput computation and makes it easier to compare different approaches for selecting active spaces for a given problem. The main drawback of this approach, however, is that the user must select the maximum active space size (or equivalently, the maximum number of CSFs). While this concession makes the schemes in some sense less "automatic", we believe this trade-off to be inevitable and worthwhile given the large variety of active spaces demanded by the large variety of CAS solvers (exact diagonalization, DMRG, FCIQMC), and in line with other methods

that require users to select the size of their approximation such as CISD/CISDT/CISDTQ. Although one may wish to further automate the problem of choosing the maximum number of CSFs (specifically to avoid drastically inadequate calculations for a given property/system), this choice is inherently dependent on the property of interest and the CAS solver used; while for specific applications this further automation may prove quite fruitful, the problem of flagging any calculation as entirely inadequate is quite difficult and we suspect the frontier of adequacy will evolve considerably as computational power increases and new methods are developed.

2.4.3 A High-Throughput Examination: 1120 SA-CASSCF/NEVPT2 Calculations

To further demonstrate the robustness of the ranked-orbital approach for high-throughput multireference computation and its utility for evaluating and comparing the effectiveness of different orbitals and methods, we calculated excitation energies for the 20 small systems using SA-CASSCF/NEVPT2 and choosing from the lowest 30 orbitals in energy with the EntropyCAS and EntropyCAS+ ranked-orbital approaches. Four different localization methods were used to generate the initial orbitals for selection: Boys, Pipek-Mezey, and Edmiston-Ruedenberg in addition to canonical orbitals. For each non-canonical method, the virtual orbitals were localized over two different subsets of virtual canonical orbitals, either the lowest 40 virtuals in energy or the lowest 20 and the highest 20 virtuals in energy, the intuition behind the second scheme being to include Rydberg character in the final localized orbitals. In addition, we investigated the effectiveness of the ranked-orbital approach at four maximum active space sizes: max(7,6), max(8,8), max(10,10), and max(12,12), which differ in their number of CSFs by roughly an order of magnitude each. In total, we calculated excitation energies for 20 systems over seven localization schemes (cannonical orbitals + two types of Boys, Pipek-Mezey, and Edmiston-Ruedenberg) at four different maximum active

space sizes (max(7,6), max(8,8), max(10,10), and max(12,12)) using two different selection methods (ranked-orbtial EntropyCAS and EntropyCAS+), for a total of 1120 calculations.

With the data from this study we hope to demonstrate the utility of the ranked-orbital procedure by answering the following questions concerning the calculation of the first excitation energies of small molecules:

- How does the maximum active space size and orbital localization affect the quality of the results?
- Does the consideration of excited-state entropies through EntropyCAS+ affect the quality of the results?
- How does the addition of dynamical correlation through NEVPT2 interact with the maximum size of the active space and the ranked-orbital procedure?

To answer these questions, we will show the mean absolute error of the calculated excitation energies with respect to the reference values calculated by Bao and Truhlar[1] over subsets of these calculations with given settings. For example, to analyze the effectiveness of EntropyCAS+ we compare the mean absolute error of calculations using EntropyCAS vs. calculations using EntropyCAS+. Mean values and confidence intervals are taken over the complete subset of the 1120 calculations with the specified settings– for example, EntropyCAS calculations will account for $1120/2 = 560$ calculations differing in system, maximum size, and localization. Similarly, EntropyCAS calculations initialized with HF orbitals will account for $560/7 = 80$ calculations differing by system and maximum size, and EntropyCAS calculations initialized with Boys orbitals will account for $560/7 * 2 = 160$ calculations, due to the two types of non-canonical virtual localization schemes. Exactly what calculations are being averaged over is specified in the captions of each figure below.

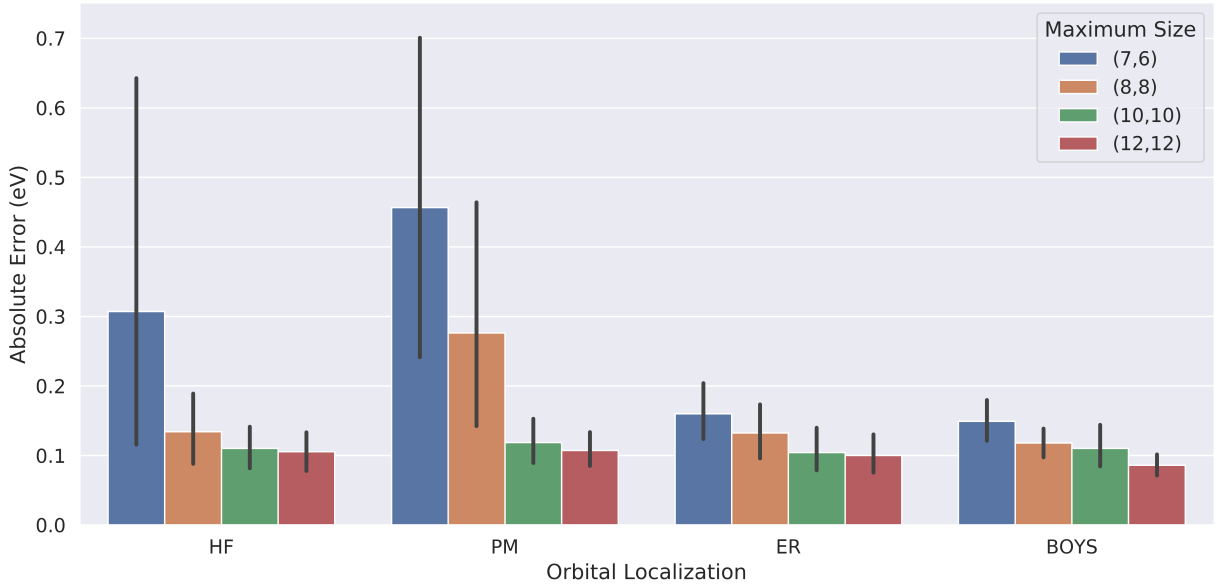

Figure 2.4: Performance of the ranked EntropyCAS/EntropyCAS+ procedures over different orbital localizations and maximum active space sizes, plotted by the error of their final CASSCF/NEVPT2 excitation energies from reference values; bootstrapped 95% confidence intervals are shown by vertical bars in black. The statistics of each bar are taken over 40 calculations for HF and over 80 calculations for localized schemes (due to the two types of virtual localization) that differ in system and selection method (EntropyCAS vs. EntropyCAS+). Regardless of orbital localization, a convergent decrease in the mean absolute error in the excitation energies is observed with increasing maximum active space size.

Active Space Size and Orbital Localization. The performances of the ranked-orbital EntropyCAS/EntropyCAS+ procedures are shown in figure 2.4. A convergent decrease in the mean absolute error is observed for all localization schemes with increasing active space size. Excepting the Pipek-Mezey scheme, which seems to have pathological behavior due to its implementation with Lödwin charges in a triple-zeta basis,[96] orbital localization greatly increases the quality of the results obtained with respect to the canonical orbitals from Hartree-Fock (HF). Boys-localized max(7,6) spaces generate results of roughly the same quality as HF max(8,8) spaces, the latter of which has roughly an order of magnitude more CSFs. We find Boys-localized orbitals to be the overall best in quality, with the best performance at both (7,6) and (12,12) spaces, and also note that IBO localization[104] is a promising

localization scheme for this application but was not explored here.

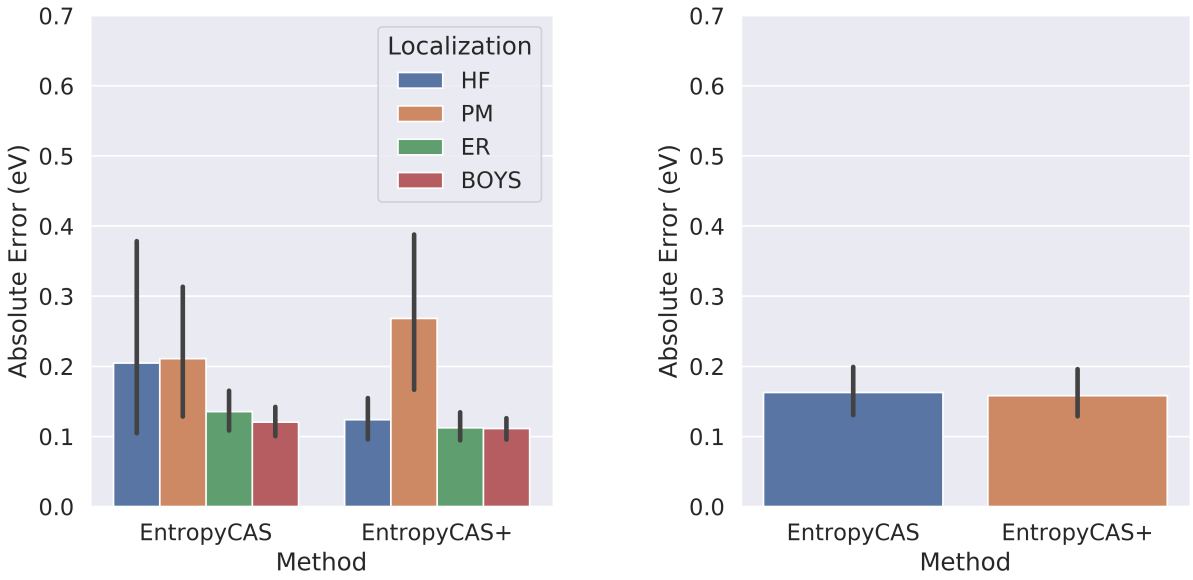

Figure 2.5: Left: Comparison of the EntropyCAS and EntropyCAS+ procedures by orbital localization. Statistics are taken over 80 calculations for HF and over 160 calculations for localized schemes (due to the two types of virtual localization) that differ in system and maximum active space size. Right: Comparison of the EntropyCAS and EntropyCAS+ procedures overall. Statistics are taken over 560 calculations in each bar that differ in system, maximum active space size, and localization. Confidence intervals at 95% are shown in black.

EntropyCAS vs. EntropyCAS+. Figure 2.5 shows the performance of the EntropyCAS and EntropyCAS+ procedures over the entire dataset partitioned by orbital localization and overall. Surprisingly, we find there to be no significant improvement when taking into account the excited-state orbital entropies (EntropyCAS+), except in the case of the HF orbitals, where errors are greatly reduced by almost 0.08 eV. One explanation for this is that the localization scheme is able to produce orbitals that are better for both the ground and excited states, and hence only using the orbital entropies from ground state performs well in these cases. From these data, we highly recommend the EntropyCAS+ procedure for $S_0 \rightarrow S_1$ and $D_0 \rightarrow D_1$ excitation energies when employing HF orbitals, but at least for the systems studied here, when employing localized orbitals, only considering the ground

state entropies is likely sufficient. Over the entire dataset we find no significant difference between the EntropyCAS and EntropyCAS+ procedures (although it should be noted that HF calculations make up only one out of every seven calculations). Interestingly, we find that the excitation energies tend to be overestimated (about 78% of calculations) significantly more than they are underestimated (about 22% of calculations), regardless of the selection method used (supporting information).

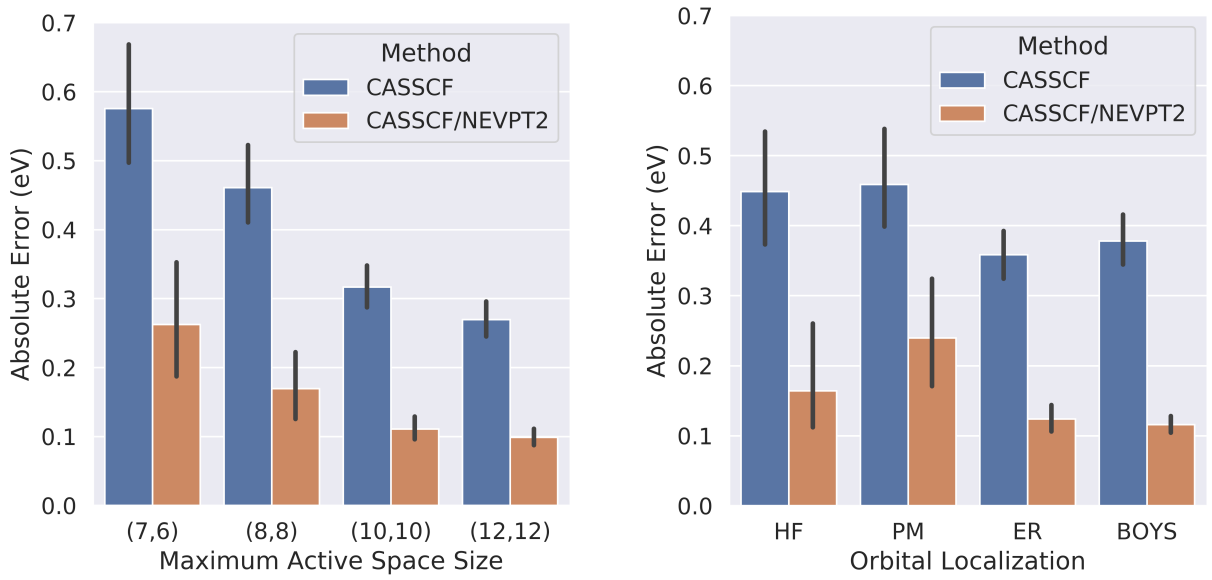

Figure 2.6: Left: Comparison of CASSCF vs. CASSCF/NEVPT2 by maximum active space size. The statistics of each bar are taken over 280 calculations that differ in system, localization, and selection method (EntropyCAS vs. EntropyCAS+). Right: Comparison of CASSCF vs. CASSCF/NEVPT2 by orbital localization. The statistics of each bar are taken over 160 calculations for HF and over 320 calculations for localized schemes (due to the two types of virtual localization) that differ in maximum active space size and and selection method (EntropyCAS vs. EntropyCAS+).

CASSCF vs. CASSCF/NEVPT2. Figure 2.6 shows the performance of CASSCF vs. CASSCF / NEVPT2 by maximum active space size and orbital localization. Expectantly, we see that the improvement with the addition of NEVPT2 decreases in magnitude with active space size, with excitation energies improved by an average of 0.31 eV at the max(7,6) level to only 0.17 eV at the max(12,12) level. Interestingly, we observe no significant difference in the

NEVPT2 improvement of the results by orbital localization, implying that the majority of the improvement in using Boys and Edmiston-Ruedenberg orbitals comes from improvements in the CASSCF wave function and not in their interaction with NEVPT2. Overall, we find, as expected, the performance of the NEVPT2 correction to be quite impressive for this problem, being able to consistently improve the CASSCF result up to errors of about 2 eV (supporting information).

2.4.4 Error Estimators for CASSCF/NEVPT2

For high-throughput screenings utilizing multireference calculations, it is desirable to develop estimators of the error of a given CASSCF/NEVPT2 result without the use of reference data. Recently, several such error estimators have been proposed by authors developing active space selection schemes:[49,80,83]

- Small singular values σ_i of the overlap matrix between the initial (selected) and final (optimized) active spaces in the CASSCF procedure,[49]

$$S_{change} = (C_{act}^{final})^{\dagger} S C_{act}^{initial} \tag{2.7}$$

$$= \sum_i \sigma_i u_i v_i^T \tag{2.8}$$

  where S is the atomic orbital overlap matrix, C_{act}^{final} are the molecular orbital coefficients of the final active space, $C_{act}^{initial}$ are the coefficients of the initial active space, and v_i and u_i are the singular vectors.

- Large differences in energy between CASSCF and CASCI,[83,84] $E_{CASSCF} - E_{CASCI}$ or $\Delta E_{CASCI}^{CASSCF}$

- Large numbers of iterations/macro cycles undertaken by the CASSCF optimization procedure, N_{iter}.[84]

- Large absolute differences between the CASSCF and NEVPT2 excitation energies,[1] $|\Delta E_{NEVPT2} - \Delta E_{CASSCF}|$ or $|\Delta\Delta E^{NEVPT2}_{CASSCF}|$

Through an analysis of the 1120 calculations above we hope to quantify the effectiveness of these different methods for estimating the error of a given CASSCF/NEVPT2 result and suggest good thresholds for utilizing these values.

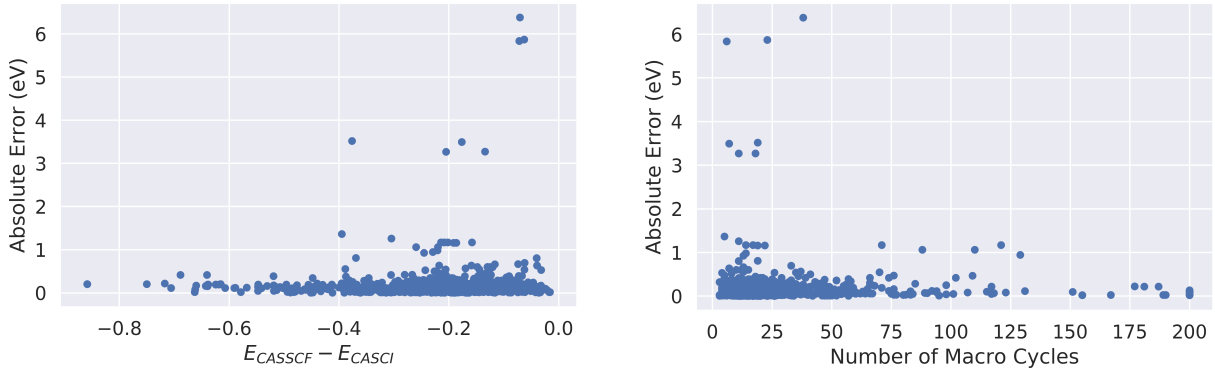

Figure 2.7: Left: Absolute errors of all 1120 calculated excitation energies with respect to the reference values of Bao and Truhlar,[1] plotted against their state-averaged $\Delta E^{CASSCF}_{CASCI}$. Right: Absolute errors of all 1120 calculated excitation energies with respect to the reference values of Bao and Truhlar,[1] plotted against the number of macro cycles in the CASSCF procedure, N_{iter}. Neither value has any significant correlation with the error of calculated excitation energies.

$\Delta E^{CASSCF}_{CASCI}$ and N_{iter}. Figure 2.7 shows the performance of the state-averaged $\Delta E^{CASSCF}_{CASCI}$ and N_{iter} as error estimators of the CASSCF/NEVPT2 results. We find that both of these values have no significant correlation with the absolute error of the calculated excitation energies. Interestingly, we find that calculations initialized with canonical (HF) orbitals change in the state-averaged energy only about half as much on average than those initialized by localized orbitals, and that $\Delta E^{CASSCF}_{CASCI}$ remains surprisingly consistent with maximum active space size (supporting information).

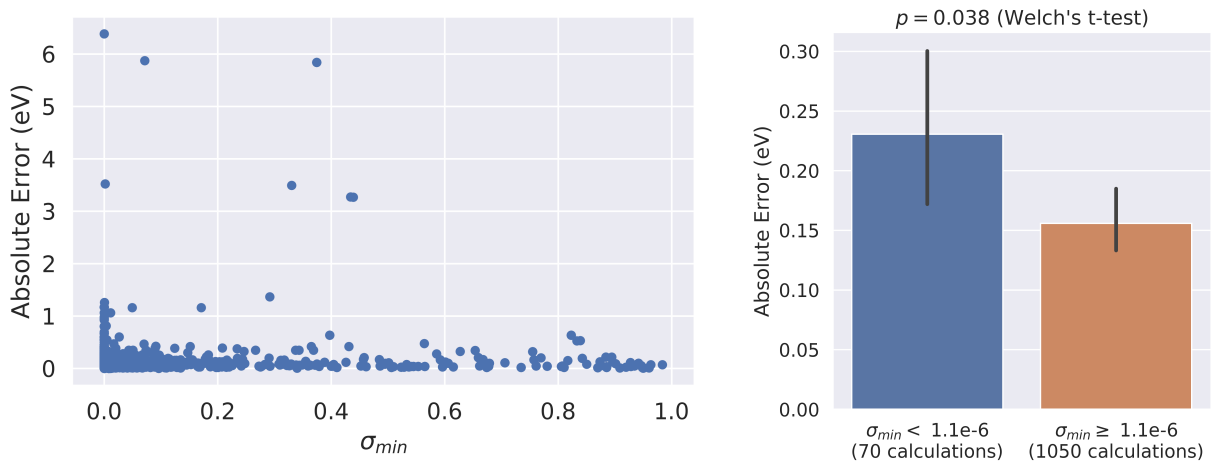

Figure 2.8: Left: Absolute errors of all 1120 calculated excitation energies with respect to the reference values of Bao and Truhlar,[1] plotted against the minimum singular value σ_{min} of their active space overlap matrix (equations 2.7 and 2.8). Right: Performance of the suggested threshold of 1.1e-6, which demonstrates a statistically significant difference between the two groups of calculations under Welch's t-test.[2]

Active Space Overlap. Small singular values of the active space overlap matrix (equations 2.7 and 2.8) indicate that an orbital was rotated out completely during the CASSCF optimization procedure, and has thus been proposed by Sayfutyarova and coworkers as a way to judge the quality of a given active space.[49] Figure 2.8 demonstrates the performance of the minimum singular value of the active space overlap matrix, σ_{min}, as an error estimator of the CASSCF/NEVPT2 results. We find that σ_{min} decreases with maximum active space size in correspondence with larger N_{iter}, implying that the starting orbitals are further away from local minima on the larger CASSCF optimization surfaces; this is likely indicative of additional orbitals capturing a vanishing amount of static correlation, as also supported by the diminishing returns in absolute error with increasing maximum active space size (figure 2.4). While there appears to be merit to using σ_{min} to judge the quality of a finalized active space, we find the difference in error to only be significant at extremely low values of σ_{min}. The right of figure 2.8 demonstrates the performance of our suggested threshold of 1.1e-6, which classifies a subset of 70 calculations (about 6%) that has a significantly higher mean

error by about 0.08 eV.

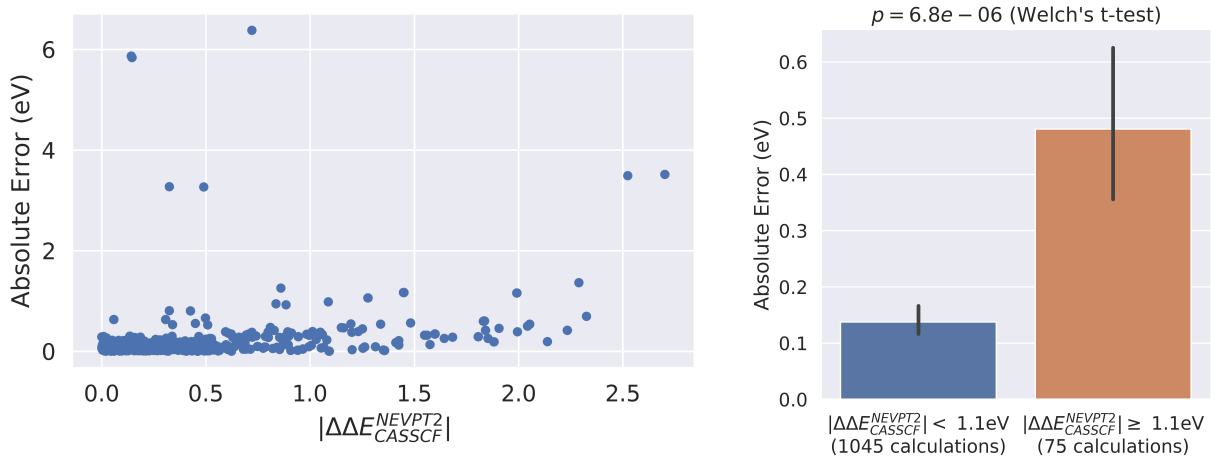

Figure 2.9: Left: Absolute errors of all 1120 calculated excitation energies with respect to the reference values of Bao and Truhlar,[1] plotted against their $|\Delta\Delta E^{NEVPT2}_{CASSCF}|$. Right: Performance of the 1.1 eV threshold suggested by Bao and Truhlar, which demonstrates a significant difference between the two groups of calculations under Welch's t-test.[2]

$|\Delta\Delta E^{NEVPT2}_{CASSCF}|$. Bao and Truhlar suggested classifying an excitation energy result as "reliable" if $|\Delta\Delta E^{NEVPT2}_{CASSCF}| \leq 1.1$ eV.[1] Figure 2.9 shows the performance of this test as an error estimator of the CASSCF/NEVPT2 results, and indeed we find the 1.1 eV threshold suggested by Bao and Truhlar to separate the calculations into significantly different groups, classifying a subset of 75 calculations (about 7%) that has a significantly higher mean error by about 0.34 eV. In the supporting information we suggest optimized thresholds for using $|\Delta\Delta E^{NEVPT2}_{CASSCF}|$ as a weak error classifier, as well as further analyses of all estimators with respect to orbital localization and active space size.

2.4.5 Approximations of the Orbital Entropy

To increase the viability of high-throughput multireference calculations, good active spaces should be able to be selected at low cost. While the EntropyCAS and EntropyCAS+ schemes can certainly select good active spaces in a physically motivated fashion, the computation

of the DMRG orbital entropies requires a fair amount of computation, with the limiting factor being high memory. In this section, we attempt to approximate the orbital entropy by analyzing the multiconfigurational character of a two-configuration system (e.g. minimal-basis H_2). The wave function for this system may be written in intermediate normalization as

$$|\psi\rangle = |20\rangle + c\,|02\rangle \tag{2.9}$$

The multiconfigurational character of this system is determined entirely by the pair coefficient c. The approach here is to model the entire wave function expansion as a set of doubly-occupied and virtual pairs, with each pair behaving like the two-configuration model system. In other words, each doubly occupied orbital interacts in a pairwise fashion with every virtual orbital, and every virtual orbital interacts in a pairwise fashion with each doubly occupied orbital. Given a set of pair coefficients for a single doubly occupied orbital i with virtual orbitals a, c_{ia} (with each c_{ia} as in equation 2.9), we can write the one-orbital reduced density matrix of the doubly occupied orbital, ρ^i, as roughly

$$\rho^i \approx \frac{1}{1+\sum_a c_{ia}^2}\left(|2\rangle\langle 2| + \sum_a c_{ia}^2\,|0\rangle\langle 0|\right) \tag{2.10}$$

where $\frac{1}{1+\sum_a c_{ia}^2}$ is a leading normalization factor. Similarly, we can write the one-orbital reduced density matrix for a virtual orbital a interacting in a pairwise fashion with doubly occupied orbitals i through pair coefficients c_{ia} as roughly

$$\rho^a \approx \frac{1}{1+\sum_i c_{ia}^2}\left(|0\rangle\langle 0| + \sum_i c_{ia}^2\,|2\rangle\langle 2|\right) \tag{2.11}$$

Then, the entropy of a doubly occupied orbital i is approximated via equation 2.3 as

$$S^i \approx -\frac{1}{1+\sum_a c_{ia}^2} \ln \frac{1}{1+\sum_a c_{ia}^2} - \frac{\sum_a c_{ia}^2}{1+\sum_a c_{ia}^2} \ln \frac{\sum_a c_{ia}^2}{1+\sum_a c_{ia}^2} \tag{2.12}$$

and for a virtual orbital a as

$$S^a \approx -\frac{1}{1+\sum_i c_{ia}^2} \ln \frac{1}{1+\sum_i c_{ia}^2} - \frac{\sum_i c_{ia}^2}{1+\sum_i c_{ia}^2} \ln \frac{\sum_i c_{ia}^2}{1+\sum_i c_{ia}^2} \tag{2.13}$$

Thus, if we can approximate the matrix of pair coefficients c_{ia} we can approximate the orbital entropies S^i and S^a. To approximate the pair coefficients, we turn back to our model system (equation 2.9), in which c is given exactly by the solution to the CI eigenvalue problem[105]

$$\begin{pmatrix} 0 & (12|12) \\ (12|12) & 2\Delta \end{pmatrix} = \begin{pmatrix} 1 \\ c \end{pmatrix} E_{corr}$$

where $(12|12)$ is the 2-electron exchange integral between orbitals 1 and 2, and Δ is half the difference in energy between $|20\rangle$ and $|02\rangle$. Solving this eigenvalue problem for c yields an analytical expression in terms of the exchange integrals and Δ,

$$c = -\frac{(12|12)}{\Delta + \sqrt{(12|12)^2 + \Delta^2}} \tag{2.14}$$

which brings the problem down to approximating the terms in this expansion for a given doubly occupied orbital i and virtual orbital a in a real system. Fairly easily we can make the approximation that $\Delta_{ia} \approx \epsilon_a - \epsilon_i$, where ϵ_i are the orbital energies (or for non-canonical orbitals, diagonal elements of the Fock matrix F_{ii}). However, the exchange integrals $(ia|ia)$ are quite costly to compute when extrapolating to larger systems as they require a molecular orbital integral transformation which scales as N^5. To approximate these integrals we examine two expressions for the orbital energy of the virtual molecular orbital in the model system, the first given by the diagonal elements of the diagonalized Fock matrix,

$$\epsilon_2 = h_{22} + J_{22} - 0.5K_{22} \tag{2.15}$$

and the second given by Koopman's theorem in terms of the molecular orbital integrals,

$$\epsilon_2 = (2|h|2) + 2(11|22) - (12|12) \tag{2.16}$$

Comparing these expressions, we match up the exchange terms and make the approximation that

$$(12|12) \approx 0.5K_{22} \tag{2.17}$$

which turns out to be exact in the case of minimal basis H_2. In the general case, for an arbitrary doubly occupied orbital i and virtual orbital a we make the approximation that

$$(ia|ia) \approx 0.5K_{aa} \tag{2.18}$$

With these approximations in hand, we approximate the final pair coefficient between a given doubly occupied orbital i and virtual orbital a as

$$c_{ia} = -\frac{0.5K_{aa}}{(\epsilon_a - \epsilon_i) + \sqrt{(0.5K_{aa})^2 + (\epsilon_a - \epsilon_i)^2}} \tag{2.19}$$

These coefficients are then gathered for each orbital and used in equations 3.2 and 3.3 to approximate the orbital entropies. We henceforth refer to this approximation as the "approximate pair coefficient" (APC) approximation. The scheme makes no attempt to approximate the entropies of or interactions with singly occupied orbitals, and instead assigns them the maximum entropy value across all virtual and doubly occupied orbitals.

As a reference scheme, we also explore not making the approximation in equation 2.18 and using the exact exchange integrals; we call this much more expensive approximation

"APCX". Finally, taking inspiration from the recent work of Welborn and co-workers,[88] we attempt to enhance this core approximation by using other elements of the HF Coulomb, exchange, kinetic and potential energy matrices with machine learning (supporting information). The model was trained on the entropies of the orbitals used in the 1120 calculations of the previous section (for a total of 20 systems * 7 localizations * 30 orbitals each = 4200 points), and a full description of this scheme is available in the supporting information; we refer to this scheme as "APCML".

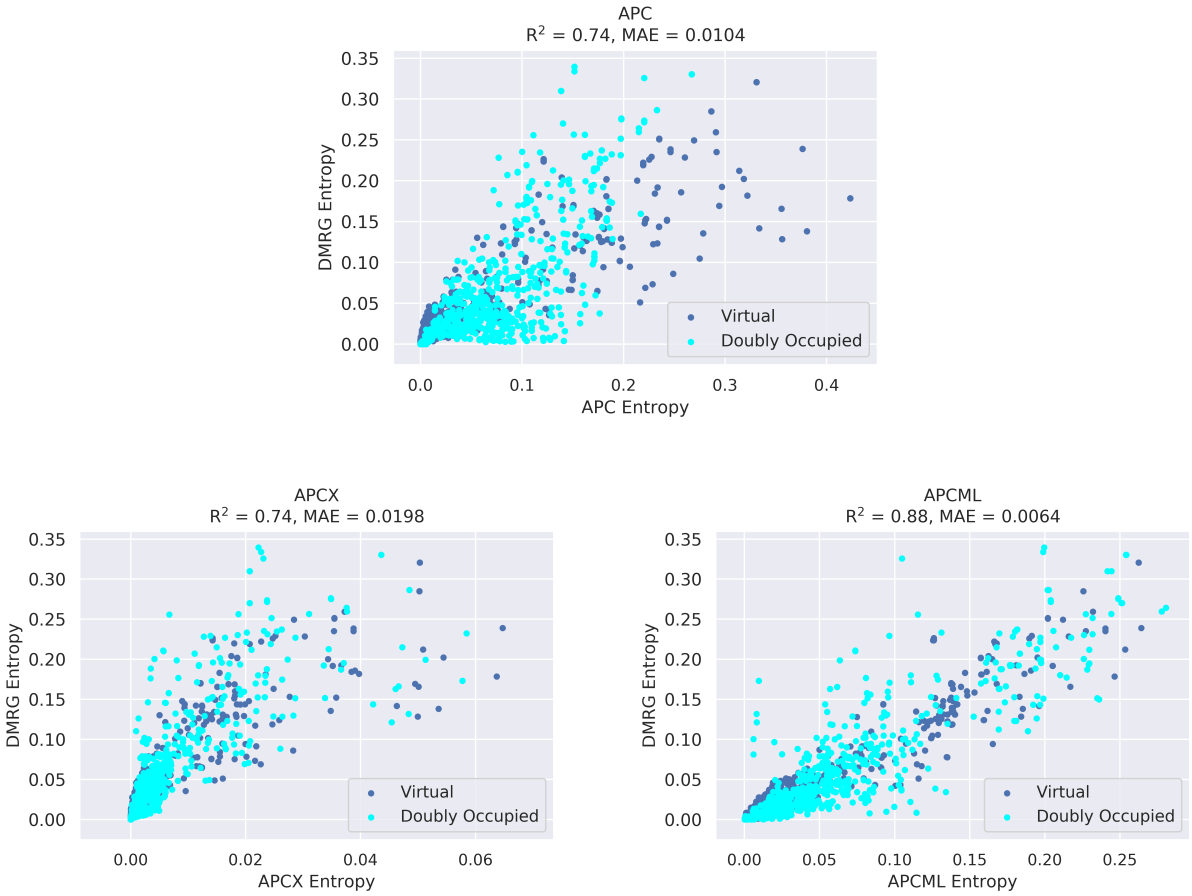

Figure 2.10: Top: APC entropies vs. DMRG entropies. Bottom: APCX and APCML entropies vs. DMRG entropies. The approximate pair coefficient (APC) approximation is a surprisingly accurate approximation of the orbital entropy for these simple systems.

Figure 2.10 demonstrates the surprisingly good performance of the APC entropies as a first-order approximation to the DMRG entropies for doubly occupied and virtual orbitals,

with a Pearson's R^2 value of 0.76 and a mean absolute error of 0.0104 over all orbitals (compared to a standard deviation of $\sigma = 0.046$). Errors tend to be higher for doubly occupied orbitals ($R^2 = 0.64$, MAE = 0.0240) and lower for virtual orbitals ($R^2 = 0.83$, MAE = 0.0064), with the main error being an overestimation of doubly occupied orbitals. However, the standard deviation of the doubly occupied orbitals is twice as large ($\sigma = 0.066$ vs. $\sigma = 0.033$). Additionally, although performance in all three schemes is significantly worse for higher-entropy orbitals (e.g. S > 0.05), this does not appear to affect the ranking precision; we find all APC schemes to rank the orbitals with about 88% precision compared to DMRG, and include a lesser number of important orbitals ($S > 0.05$) in their top 6 ranking in only 1.5-3% of cases (supporting information).

Surprisingly, we find that the APC approximation performs significantly better than APCX in approximating the magnitude of the DMRG entropies, indicating a fortunate cancellation of error. One explanation is that APC overcorrelates pairs of orbitals by approximating the integrals as diagonal elements (equation 2.18), which cancels out the correlation lost by only considering a pairwise framework. APCML performs slightly better with an R^2 of 0.88 and MAE of 0.0064. Both APCX and APCML continue to perform worse for doubly occupied orbitals and better for virtual orbitals.

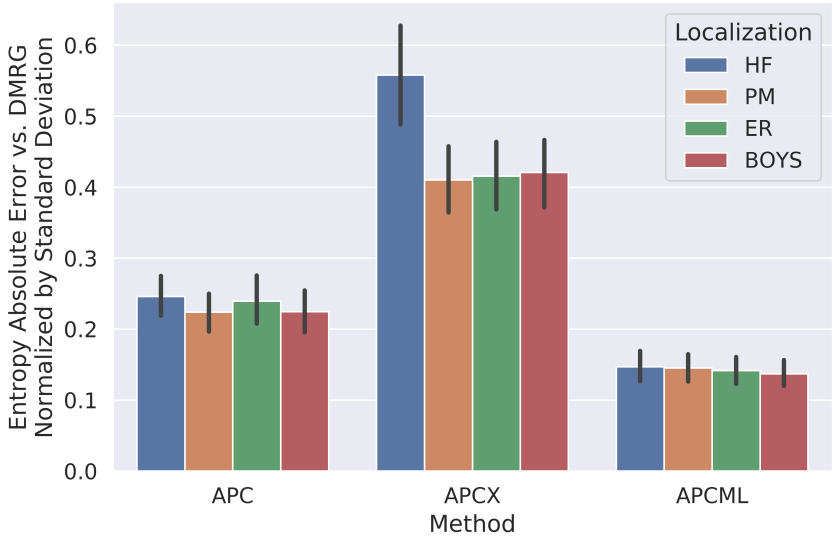

Figure 2.11: Error of different approximate methods for the DMRG entropy vs. the DMRG values, normalized by the standard deviation of entropy values in that orbital type (supporting information); bootstrapped 95% confidence intervals are shown by vertical bars in black. Statistics are taken over a subset of 20 calculations for HF and 40 calculations for localized schemes (due to the two different types of virtual localization) that differ by system. Surprisingly, there is no significant drop of in the performance of the APC schemes when applied to localized orbitals.

Since the APC approximation is centered on arguments considering HF canonical orbitals, one might think that it would perform worse for localized orbitals. Figure 2.11 shows that we find no significant difference in the performance of the APC approximation by orbital type, except in the case of APCX which surprisingly performs significantly worse for HF orbitals; this further implies a very fortunate cancellation of error in the APC approximation.

While the agreement with the DMRG orbital entropies is promising, a final evaluation of a scheme should rely on the quality of the active spaces it selects for a specific problem. To compare to DMRG values, the APC/APCX/APCML models analyzed interactions only between the same lowest 30 orbitals in energy as were analyzed in the DMRG calculation. To turn these into general schemes for systems of arbitrary size, we analyze the interactions between all doubly occupied orbitals and the lowest 23 virtual orbitals in energy (HF, Boys, AVAS) or the highest 23 virtual orbitals in occupation number (UNO(HS)).

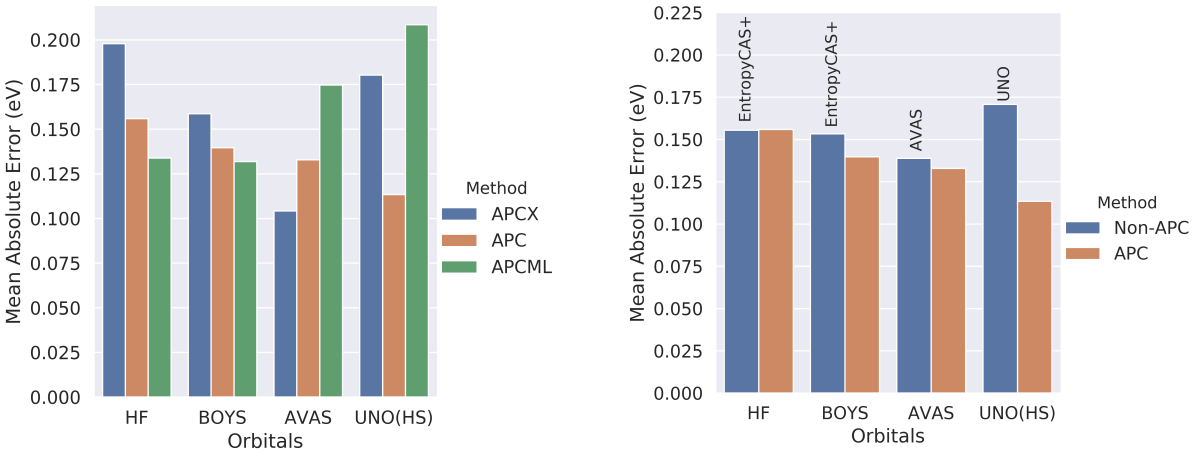

Figure 2.12: Left: Performance of APC/APCX/APCML selection on orbitals from different active space selection schemes at the max(7,6) level. Right: Performance of APC selection vs. non-APC selection at the max(7,6) level.

Figure 2.12 demonstrates the performance of the three APC schemes when choosing active spaces for different types of orbitals (HF, Boys, AVAS, and high-spin unrestricted natural orbitals (UNO(HS)), and the performance of the APC scheme in comparison to non-APC schemes. Surprisingly, we find the cheap and understandable APC scheme to perform the best overall, when compared to APCX and APCML. While APCML performs slightly better than APC for the HF and Boys-localized orbitals, APCML performs quite poorly for AVAS and UNO(HS) orbitals, and appears to be an example of overfitting and performing poorly on orbital types not included in the training data.

We wish to highlight the performance of the APC scheme with high-spin UNO orbitals, which is quite remarkable: the difference in performance between selecting the orbitals based on their UHF occupation number (UNO) and selecting them by the APC scheme is almost 0.06 eV, which is an excellent example of how orbital ranking can have a large impact on the quality of the results. Furthermore, the quality of the results obtained with the APC/UNO(HS) scheme are the best at the max(7,6) level, and even comparable with active space selections at the max(10,10) level; this would seem to imply that the UNO(HS) scheme is quite good for *producing* the orbitals for calculating excitation energies (as supported by

the work of Bao and Truhlar[1]), but rather poor at *ranking* them in terms of importance. This brings forth the possibility that approaches that mix orbital construction and active space selection could be ideal for certain types of problems.

As a final note, it appears that the concept of learning the orbital entropies has been investigated concurrently by Golub and coworkers,[80] who focused on learning the entropy for transition metal systems in much more difficult cases. We note that the approach here is much less expensive due to its featurization from solely the HF matrices and not from molecular orbital integrals, but their results are quite promising and we hope that the model employed here as well as the APC approximation helps to develop future work in this direction. We note that learning to rank algorithms[106] have a strong use case for this problem, but were not pursued here due to separate models for the doubly occupied and virtual orbitals. Additionally, the approach of using features of the HF exchange matrix to estimate energies has been explored in several papers,[88,107,108] and we hope that the APC framework developed here helps to gain insight into these models. Finally, despite its successes here, we note that the APC scheme is likely to perform worse in much larger systems and in cases where the HF determinant is a drastically poor approximation to the true wave function.

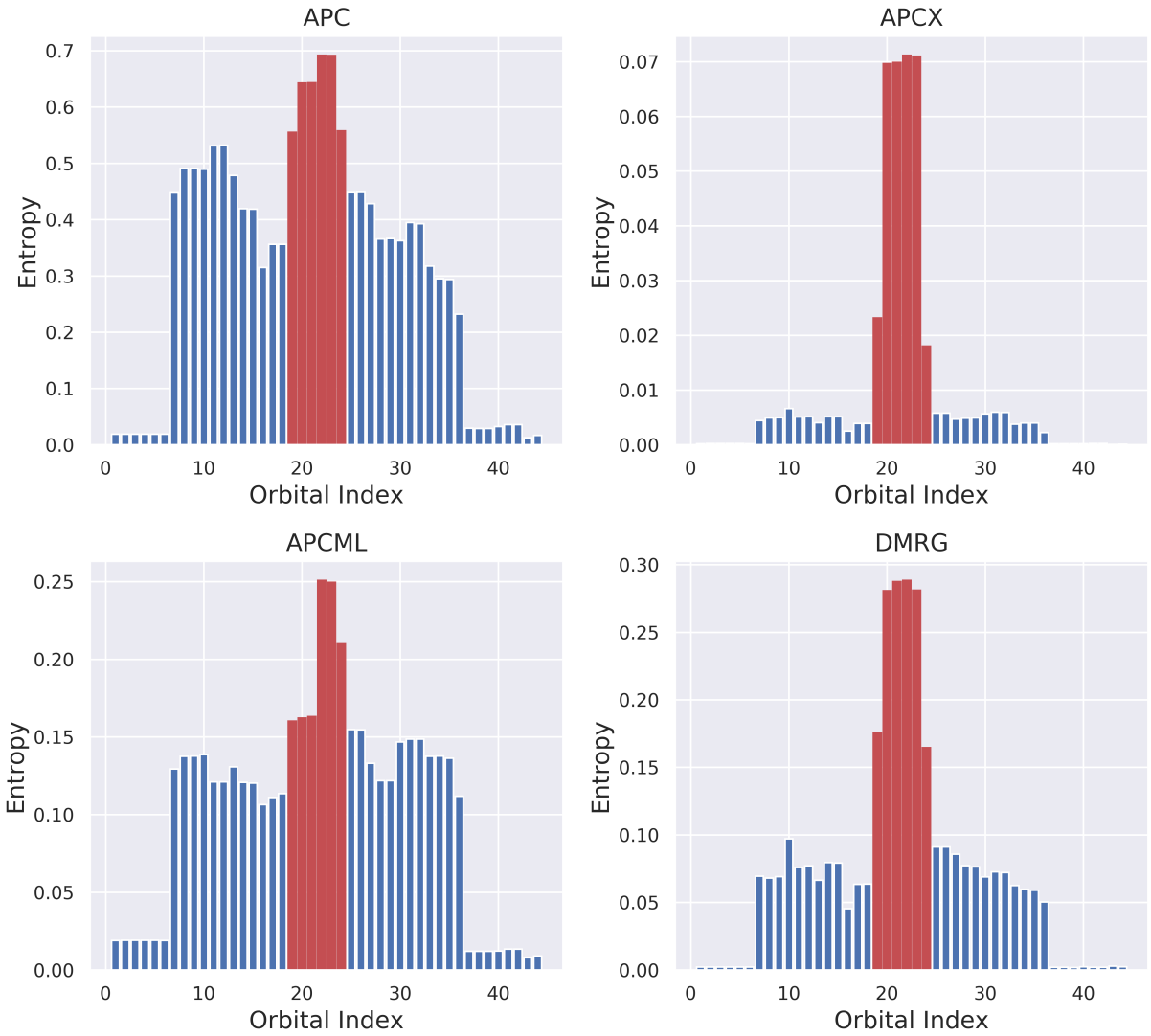

Figure 2.13: The APC, APCX, APCML, and DMRG entropies for all doubly occupied orbitals and the first 23 ground-state virtual UNO orbitals highest in occupation number for the benzene geometry of Bao and Truhlar,[1] with orbitals indexed by decreasing occupation number (the HOMO is orbital 21). All schemes select the chemically intuitive (6,6) active space of the π system at the max(7,6) level, in agreement with orbital entropies from DMRG.

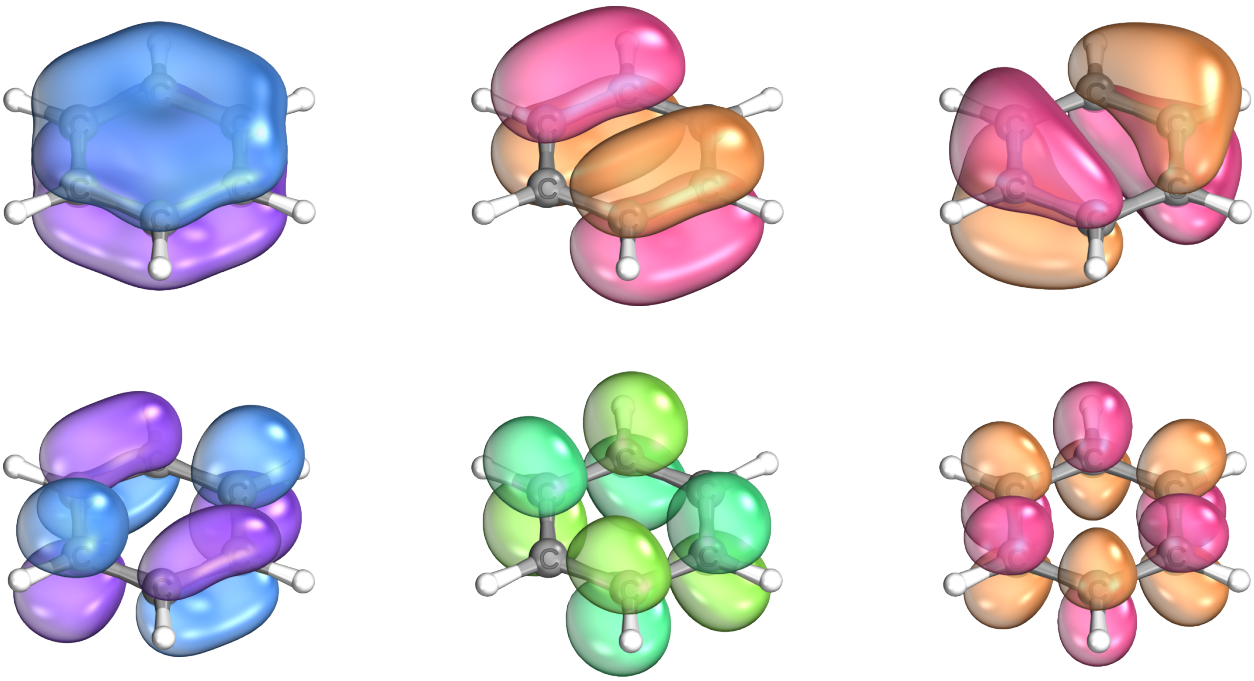

Figure 2.14: The six unrestricted natural orbitals of benzene selected by all APC schemes and EntropyCAS at the max(7,6) level. Top: Orbitals 19-21. Bottom: Orbitals 22-24.

To demonstrate the chemical utility of the APC schemes, we set out to test the APC predictions for the chemically intuitive case of benzene. Figure 2.13 shows the predicted entropies of the three different APC schemes for the ground-state UNO orbitals of the benzene geometry of Bao and Truhlar.[1] This is a case in which the UNO scheme is well-known to be able to select the chemically intuitive (6,6) space with its standard threshold of 0.02 (here, the UNO orbitals are well defined due to the existence of a non-RHF solution).[7,83] In figure 2.13, it is seen that APCX is able to identify the most important orbitals for the active space quite strongly, and while APC and APCML appear to significantly overcorrelate the lower doubly occupied and higher virtual orbitals compared to DMRG, all three schemes are able to rank the same six orbitals as the most important for the active space. Delightfully, in line with chemical intuition, all three APC schemes are able to choose the correct (6,6) space at the max(7,6) level.

2.4.6 Conclusions

In this work we have presented the ranked-orbital approach to selecting active spaces with the goal of standardizing active space multireference methods for high-throughput computation. Through an application of this approach to 1120 multireference calculations for the first excitation energies of small molecules, we showed how this method can be used to compare the quality of different orbitals and selection schemes in a meaningful fashion. Concerning selection with entropy-based procedures, we find that localized orbitals perform better than non-localized orbitals for the problem of calculating excitation energies, and that EntropyCAS is comparable to EntropyCAS+ in performance when localized orbitals are employed. Additionally, we analyzed the effectiveness of methods for estimating the error of CASSCF/NEVPT2 results, including active space overlap, N_{iter}, $\Delta E_{CASCI}^{CASSCF}$, and $|\Delta\Delta E_{CASSCF}^{NEVPT2}|$. Among these, we find $|\Delta\Delta E_{CASSCF}^{NEVPT2}|$ to be the most robust.

Next, inspired by the performance of entropy-ranked methods for this problem but discouraged by their computational cost, we attempted to estimate the entropy in a physically motivated fashion from orbital energies and features of the HF exchange matrix in a pair-interaction framework. We call this new scheme the "approximate pair coefficient" (APC) method, and it performs quite well for the test systems presented, being able to select good active spaces over many different types of orbitals. APC-selected high-spin UNO orbitals appears to be a very effective approach for calculating the first excitation energies of small molecules, indicating the utility of separating the steps of orbital generation and ranking the orbitals in terms of importance for the active space. Future work will likely focus on testing the APC scheme for more difficult cases and on the application of the ranked-orbital approach to high-throughput multireference computation for important problems in chemistry.

As a final note, we expect a main criticism of this work to be that fixing the maximum active space size in the ranked-orbital scheme allows the user to run calculations that necessarily result in an inadequate zeroth-order description of the wave function. Broadly speaking, we

respond to this criticism with two points: one, determining whether a calculation will give a qualitatively accurate and/or "useful" result is a very fuzzy question, and two, this criticism reflects a somewhat restrictive viewpoint of computational science in which the user must be "warned" that they might be doing their calculations wrong. When one performs a DFT calculation the user is generally not warned that “this functional was not parameterized for this application”, they simply do the calculation and check against some reference to see if it is good enough for what they want to investigate. We believe that computational scientists should not be closely monitored to keep them from doing calculations “the wrong way”– it is the responsibility of the scientist and not the computer to determine if their results are useful by the metrics they wish to consider. In this vein, we see the computational affordability, consistency, and comparability achieved by the ranked-orbital approach to be worthwhile tradeoffs for allowing the user select a maximum active space size that may be insufficient.

The authors thank the Inorganometallic Catalyst Design Center (ICDC) under DOE award DE-SC0012702. Additionally, the authors thank the Minnesota Supercomputing Institute (MSI) for access to computational resources and Andrew Walker for help investigating databases of molecular geometries.

CHAPTER 3

LARGE-SCALE BENCHMARKING OF MULTIREFERENCE VERTICAL-EXCITATION CALCULATIONS VIA AUTOMATED ACTIVE-SPACE SELECTION

This chapter is reprinted with permissions from *J. Chem. Theory Comput.* **2022**, *18*, 10, 6065-6076

3.1 Abstract

We have calculated state-averaged complete-active-space self-consistent-field (SA-CASSCF), multiconfiguration pair-density functional theory (MC-PDFT), hybrid MC-PDFT (HMC-PDFT), and n-electron valence state second-order perturbation theory (NEVPT2) excitation energies with the approximate pair-coefficient (APC) automated active-space selection scheme for the QUESTDB benchmark database of 542 vertical excitation energies. We eliminated poor active spaces (20-30% of calculations) by applying a threshold to the SA-CASSCF absolute error. With the remaining calculations, we find that NEVPT2 performance is significantly impacted by the size of the basis set the wave functions are converged in regardless of the quality of their description, which is a problem absent in MC-PDFT. Additionally, we find that HMC-PDFT is a significant improvement over MC-PDFT with the tPBE density functional, and that it performs about as well as NEVPT2 and second-order coupled cluster (CC2) on a set of 373 excitations in the QUESTDB database. We optimized the percentage of SA-CASSCF energy to include in HMC-PDFT when using the tPBE on-top functional, and we find the 25% value used in tPBE0 to be optimal. This work is by far the largest benchmarking of MC-PDFT and HMC-PDFT to date, and the data produced in this work is useful as a validation of HMC-PDFT and of the APC active-space selection scheme. We have made all the wave functions produced in this work (orbitals and CI vectors) available

to the public and encourage the community to utilize this data as a tool in the development of further multireference model chemistries.

3.2 Introduction

The accurate treatment of excited states is critical for understanding photochemical phenomena,[109–115] and it has been a long-standing goal of the electronic structure community.[116–127] Although treating excited states is difficult in general, it is particularly challenging when single-determinant methods such as Hartree-Fock or Kohn-Sham density functional theory provide a poor reference state for predicting excited states. This can occur either because the excited states vary greatly from the ground state (e.g., double excitations[128]) or because the ground state itself is not well-described (e.g., strongly correlated systems[129–132]). One can overcome these deficiencies by using multiple-determinant reference states, and the methods that take this approach are called multireference methods.

The most popular multireference method is the complete active space self-consistent field (CASSCF) method,[68] which expresses approximate wave functions in the space of all possible configurations of electrons in an "active space" of orbitals and electrons. These wave functions can then serve as references for perturbation theories such as MC-QDPT,[133,134] CASPT2,[135,136] and NEVPT2.[34,35] Alternatively, quantitative accuracy can be achieved by using a nonclassical-energy functional applied to the the converged wave function in multiconfiguration nonclassical functional theory (MC-NCFT).[41,48,137–139] The total energy is then a sum of the classical portion of the CASSCF energy and nonclassical energy from the functional.

The most common form of MC-NCFT utilizes nonclassical-energy functionals obtained by translating Kohn-Sham exchange-correlation functionals for use with multiconfigurational wave functions via the on-top pair density and is called multiconfiguration pair-density functional theory (MC-PDFT). The translated PBE functional (tPBE) has been used as the

functional in the majority of MC-PDFT calculations to date. The nonclassical energy from a density functional can be mixed with the nonclassical part of the CASSCF energy to form a "hybrid" nonclassical functional, for example, using a 0.75:0.25 mixture of tPBE and CASSCF nonclassical energies yields the tPBE0 functional.[140]

Difficulties encountered in all such post-CASSCF methods are making the active space large enough and well-balanced enough to converge to a qualitatively accurate description of the underlying wave function(s). The results can depend significantly on the size and nature of the active space and the initial orbital guess.[30] Moreover, in many occasions the orbitals will change character during their optimization. For these reasons, such calculations often require expert human guidance to carefully choose the active space size and composition and monitor them during the calculations.

Although CASSCF has been used since the 1980s,[68] the prospect of automated active space selection has only received significant attention within the last decade or so.[1,3,7,31,49,75–79,81–84,86,87,103,141–143] Recently, we published the ranked-orbital approach to select active spaces and the approximate pair coefficient (APC) approximation for low-cost estimates of the orbital entropies used in the ranking.[5] This automated scheme, inspired by the entropy-driven approach of Stein and Reiher,[76] allows for the flexible selection of active space size with a hierarchy of levels (max(8,8), max(10,10), max(12,12)...) reminiscent of the CI level sequence (CISD, CISDT, CISDTQ, ...).

Recently, Jacquemin and coworkers published the QUESTDB benchmark dataset of 542 vertical excitation energies on a diverse set of small and midsize main-group molecules, calculated via a variety of high-level wave function methods in the aug-cc-pVTZ[144,145] basis.[44,45,128,146–148] In the present paper we have undertaken the automated calculation of these excitation energies with SA-CASSCF, NEVPT2, and MC-PDFT using the APC-ranked-orbital active space selection scheme. To benchmark and analyze the performance of various multireference methods on this diverse set of excitations, we eliminate poor active

spaces (20-30% of calculations) by setting an error threshold on the SA-CASSCF excitation energy because that has previously been shown to be good way to judge the quality of the active space.[1]

By analyzing results across different active space and basis set size choices, we find different trends in the performance of MC-PDFT and NEVPT2 where the performance of NEVPT2 is overly dependent on the basis set in which the underlying wave function is converged. Additionally, we are able to produce the first large-scale and robust comparison of MC-PDFT to other single-reference methods such as CC2, and find the CASSCF mixing parameter of 0.25 used in tPBE0 to be optimal. We have made all the wave functions converged in this work available to the public via publication of the converged orbitals and CI vectors and encourage others to use these data in the development of further multireference model chemistries.

3.3 Methods

Data Overview. The data we have examined can be found in the QUESTDB dataset,[45] which consists of 542 vertical excitations of small and midsize main-group molecules (molecules with 1-10 non-hydrogenic atoms). Of these excitations, 491 are from singlet ground states and 51 are from doublet ground states. Every excitation in the QUESTDB dataset is specified by its spatial and spin symmetries, and benchmark values are reported as "theoretical best estimates" (TBEs) calculated with a variety of high-level methods with the aug-cc-pVTZ[144,145] basis. These TBEs have been used in this work to judge the errors of all computed excitation energies, even those obtained with a different basis set.

Active Space Selection. To obtain orbitals for the active space selection scheme, we started with a restricted Hartree-Fock singlet wave function for closed-shell molecules and a restricted open-shell Hartree-Fock doublet wave function for doublet molecules, as calculated using PySCF.[90] The molecular point group was reduced to the highest available symmetry

implemented for the PySCF SA-CASSCF solver: C_{2h}, C_{2v}, C_s, or D_{2h}. The APC-ranked-orbital active-space-selection scheme[5,48] starts with a set of candidate localized orbitals, ranks them by their approximated orbital entropies, and then eliminates orbitals starting from the lowest-entropy orbitals (those with the highest entropies are considered to be the most important) until the active space size reaches a predetermined maximum number of configuration state functions. We next describe the generation of candidate orbitals, then the ranking scheme, and finally the maximum-size criteria.

Following previous work,[48] up to 23 lowest-energy virtual orbitals of the Hartree-Fock calculation were selected, and orbitals within this subset were grouped by symmetry and Boys-localized[94] within each symmetry. Likewise, up to 23 highest-energy doubly occupied orbitals were also grouped by symmetry and Boys-localized within each symmetry. These two sets of localized orbitals (and the one singly occupied orbital, when present) were then considered as candidates for the active space. Next we describe how we ranked the localized orbitals.

In the originally published APC ranking scheme, given a set of doubly occupied candidate orbitals i and virtual orbitals a, one calculated the approximate-pair-coefficient (APC) matrix C_{ia} as

$$C_{ia} = \frac{-0.5K_{aa}}{F_{aa} - F_{ii} + \sqrt{(0.5K_{aa})^2 + (F_{aa} - F_{ii})^2}} \tag{3.1}$$

where K_{aa} is the diagonal virtual element of the exchange matrix and F_{aa} and F_{ii} are diagonal virtual and doubly occupied elements of the Fock matrix in the MO basis. The entropies of doubly occupied orbitals i are calculated as

$$S^i \approx -\frac{1}{1+\sum_a C_{ia}^2} \ln \frac{1}{1+\sum_a C_{ia}^2} - \frac{\sum_a C_{ia}^2}{1+\sum_a C_{ia}^2} \ln \frac{\sum_a C_{ia}^2}{1+\sum_a C_{ia}^2} \tag{3.2}$$

and those of virtual orbitals a as

$$S^a \approx -\frac{1}{1+\sum_i C_{ia}^2} \ln \frac{1}{1+\sum_i C_{ia}^2} - \frac{\sum_i C_{ia}^2}{1+\sum_i C_{ia}^2} \ln \frac{\sum_i C_{ia}^2}{1+\sum_i C_{ia}^2} \tag{3.3}$$

Finally, any singly occupied orbitals are assigned the maximum entropy value from the set of doubly occupied and virtual orbital entropies ($\{S_i\}$,$\{S_a\}$). Note that the removal of a virtual orbital from consideration affects all doubly occupied entropies and vice-versa. This method is inexpensive because it uses only easily calculated diagonal Hartree-Fock matrix elements (supporting information).

However, in the present work we have found that in larger molecules (with >350 aug-cc-pVTZ basis functions) the APC entropies tend to overestimate the interaction of some virtual orbitals with the doubly occupied orbitals, artificially inflating the entropies of all doubly occupied orbitals and causing the selection of highly imbalanced active spaces (spporting information). To overcome this issue, we propose an algorithmic extension of APC in which high-entropy virtual orbitals are removed from consideration when calculating entropies and then assigned the maximum entropy value (i.e. treated in the same way as singly occupied orbitals). The algorithm takes the following steps:

- Provide the sets of candidate doubly occupied/ singly occupied/virtual orbitals $(\{L_i\}, \{L_s\}, \{L_a\})$.
- Calculate entropies $(\{S_i\}, \{S_s\}, \{S_a\}) = \text{APC}(\{L_i\}, \{L_s\}, \{L_a\})$ and then remove the highest-entropy virtual orbital from L_a and put it in L_s. Repeat N times.
- Return $(\{S_i\}, \{S_s\}, \{S_a\})$.

The above algorithm has a single parameter N (APC-N) which is the number of times the highest-entropy virtual orbital is removed. In this work we have found a good value of N to be 2 (a scheme we refer to as APC-2), and we find that using APC-2 entropies results in more balanced active spaces and lower SA-CASSCF error than APC (supporting information).

The APC-2 entropies are then used to rank and select the orbitals for the active space by dropping the lowest-entropy orbitals until the active space size is lower than some maximum active space size. However, we note that the current scheme should be improved for the treatment of orbitals with degenerate entropies (supporting information).

In more detail, the active space size for a given set of N_{orb} active orbitals containing N_{elec} active electrons is calculated via the equation:[67]

$$N_{\mathrm{CSF}}(N_{elec}, N_{orb}) = \binom{N_{orb}}{N_\alpha}\binom{N_{orb}}{N_\beta} - \binom{N_{orb}}{N_\alpha + 1}\binom{N_{orb}}{N_\beta - 1} \tag{3.4}$$

where $N_\alpha + N_\beta = N_{elec}$ and $N_\alpha = N_\beta$ for even N_{elec} and $N_\alpha = N_\beta + 1$ for odd N_{elec}. The maximum active space size $N_{\mathrm{CSF}}^{\mathrm{Max}}$ is set via a specification of a maximum number of active electrons and orbitals (N_{elec}^{Max},N_{orb}^{Max}) whose size is calculated via equation 3.4; this maximum active space choice is notated as max(N_{elec}^{Max},N_{orb}^{Max}). In this work we calculate results at three choices of max(N_{elec}^{Max},N_{orb}^{Max}): max(8,8) ($N_{\mathrm{CSF}}^{\mathrm{Max}} = 1764$), max(10,10) ($N_{\mathrm{CSF}}^{\mathrm{Max}} = 19404$), and max(12,12) ($N_{\mathrm{CSF}}^{\mathrm{Max}} = 226512$). Following this specification, all orbitals are selected and then the lowest-entropy orbital is successively dropped until the size of the active space calculated via equation 3.4 is less than or equal to $N_{\mathrm{CSF}}^{\mathrm{Max}}$.

As an example, we guide the reader through choosing a max(4,4) active space ($N_{\mathrm{CSF}}^{\mathrm{Max}} = 20$) from a set of orbitals with occupancies and entropies $\{(n_j, S_j)\}$:

$$\{(2, 0.05), (2, 0.5), (2, 0.9), (1, 0.9), (0, 1.2), (0, 0.2), (0, 0.1)\} \tag{3.5}$$

In this case the active space is selected as:

- (2,0.05),(2,0.5),(2,0.9),(1,0.9),(0,1.2),(0,0.2),(0,0.1) | (7,7) $N_{CSF} = 784$
- (2,0.5),(2,0.9),(1,0.9),(0,1.2),(0,0.2),(0,0.1) | (5,6) $N_{CSF} = 210$
- (2,0.5),(2,0.9),(1,0.9),(0,1.2),(0,0.2) | (5,5) $N_{CSF} = 75$

- (2,0.5),(2,0.9),(1,0.9),(0,1.2) | (5,4) $N_{CSF} = 20$

with a resulting selected (5,4) active space.

Calculation of the Excitation Energies. Calculations of excited-state wave functions were carried out by state-averaged CASSCF, averaging over the ground state and the minimum necessary number of excited states of the symmetry specified by QUESTDB. For example, in a C_{2v} molecule (e.g., water) if the symmetry of the excited state under consideration is specified to be $^1A_2$ with no lower $^1A_2$ excitations present, then the state averaging was done evenly over the $^1A_1$ ground state and the $^1A_2$ excited state. For higher $^1A_2$ excitations, however, the state averaging included an additional $^1A_2$ state for each $^1A_2$ excitation lower in energy (again with weights for state averaging being the same for all states averaged). Standard convergence parameters were employed, and for a few poor active space choices (0.4% of cases) the calculations failed to converge.

Because the highest available point groups supported by the SA-CASSCF solver in PySCF have lower symmetry than those specified by QUESTDB for single atoms and diatomics, the point groups sometimes had to be reduced to the highest-symmetry subgroup. Additionally, the labeling of different irreps is sometimes a choice of axis convention, such as between B_1 and B_2 in C_{2v} or between $B_{1g}/B_{2g}/B_{3g}$ and $B_{1u}/B_{2u}/B_{3u}$ in D_{2h}; we have done our best to match the irrep we think was used in QUESTDB. Calculations were done with the highest M_S allowed by the spin symmetry (e.g., if an excited state has $S = 1$, then for 8 electrons in the active space the active space would have 5 α and 3 β electrons).

The tPBE and NEVPT2 energies of the converged SA-CASSCF states were then calculated using the implementations of these methods in PySCF. Our implementation of MC-PDFT within PySCF is currently available in the mrh repository.[149] Additionally, tPBE0 energies were calculated by averaging the SA-CASSCF and tPBE energies:[140]

$$E_{\mathrm{tPBE0}} = 0.25E_{\mathrm{SA-CASSCF}} + 0.75E_{\mathrm{tPBE}} \tag{3.6}$$

The only implementation of NEVPT2 currently in PySCF is strongly contracted NEVPT2 (SC-NEVPT2),[35] and our NEVPT2 calculations use this.

In order to maintain a consistent labeling, the excited state to be compared to the QUESTDB excitation energy was chosen to be the state highest in energy as judged by tPBE. Although this is not a fail-proof scheme in terms of isolating the "same" QUESTDB state of the specified symmetry due to root flipping, we have found it to be satisfactory for our work as the converged QUESTDB wave functions are unavailable and labels such as "$n \rightarrow \pi^*$" are ambiguous non-observables. However, because the present work shows that tPBE0 is more accurate than tPBE, we suggest ordering the states by tPBE0 in future work.

Method Timing. All converged CASSCF wave functions (orbitals and CI vectors) were saved to disk at the end of the calculation. Timings for tPBE and NEVPT2 calculations were achieved by loading in the converged CASSCF wave functions, computing the relevant quantity (the tPBE nonclassical energy or the NEVPT2 perturbative correction) and then saving the results. The amount of resources requested for each calculation was determined by an empirically derived formula dependent on the number of aug-cc-pVTZ basis functions in the underlying molecule (supporting information), and so timings between the tPBE and NEVPT2 implementations available in PySCF can be fairly compared (although we note that methodologies can always be further optimized).

Plotting. Figures were made in Python using matplotlib as enhanced by Pandas[150,151] and Seaborn.[98] Seaborn calculates 95% confidence intervals for the mean values reported in plots by bootstrapping the mean value over 1000 random samplings of the underlying data.[152–154]

3.4 Results

3.4.1 Eliminating Poor Active Spaces

We calculated excitation energies for all 542 vertical excitations listed in the QUESTDB database with six combinations of active space and basis set: four involving max(12,12) APC-2 active spaces with decreasing basis size (aug-cc-pVTZ[144,145], jun-cc-pVTZ,[155] cc-pVTZ,[156,157] cc-pVDZ[156,157]) and two involving jun-cc-pVTZ with decreasing active space size (max(10,10) and max(8,8)). We will refer to these combinations throughout the paper as Aug(12,12), Jun(12,12), TZ(12,12), DZ(12,12), Jun(10,10), and Jun(8,8).

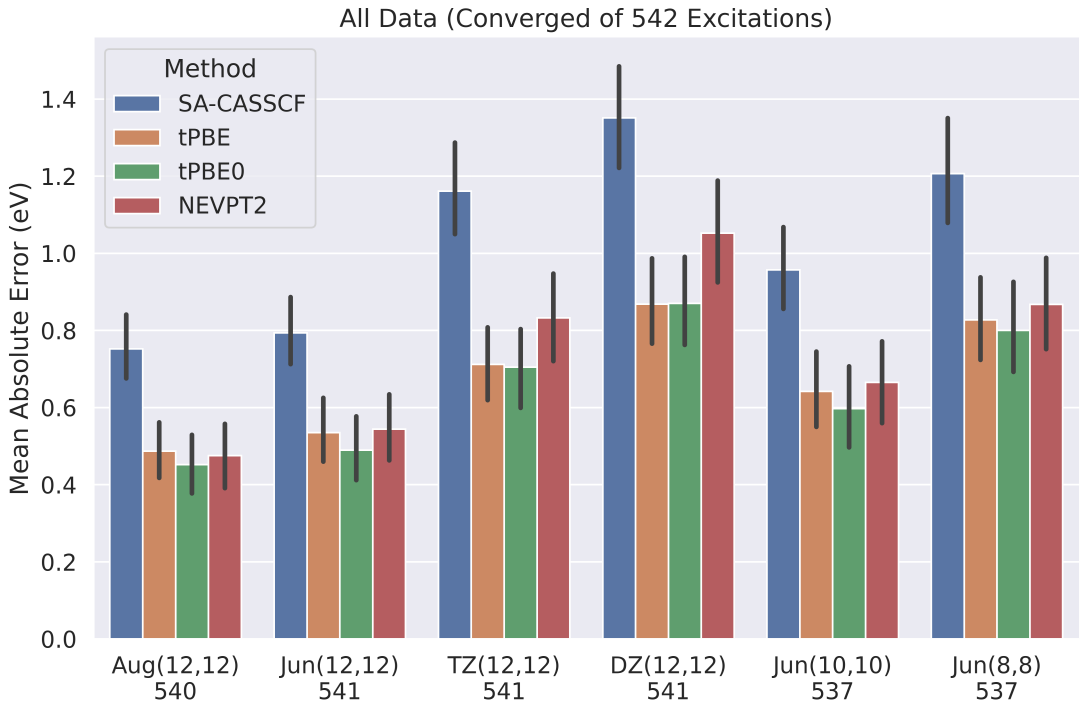

Figure 3.1: Comparison of the mean absolute errors of SA-CASSCF, tPBE, tPBE0, and NEVPT2 across different active space and basis set sizes for all converged calculations. The number of converged excitations with each combination of active space and basis is shown below each column, and 95% confidence intervals for each mean are shown in black.

Figure 3.1 shows the mean absolute errors of SA-CASSCF, tPBE, tPBE0, and NEVPT2 that we obtain for all wave functions converged at each combination of active space and basis

set tested in this work. Adding together the number of converged calculations at each active space and basis set shown at the bottom of Figure 1 yields 3237 calculations. As expected, we find that the SA-CASSCF error increases when we move from a larger to a smaller basis set with a given active space scheme or when we move from a larger active space to a smaller one with a given basis set. However, in order to reasonably evaluate the accuracies of these methods, we need to eliminate results whose error is driven mainly by poorly chosen active spaces. To analyze only cases with reasonable active spaces we set a threshold T on the SA-CASSCF error of 1.1 eV ($T_{\mathrm{SA-CASSCF}} = 1.1$ eV). That is, we consider that the APC scheme has produced a good active space if the error in the SA-CASSCF excitation energy is less than 1.1 eV.

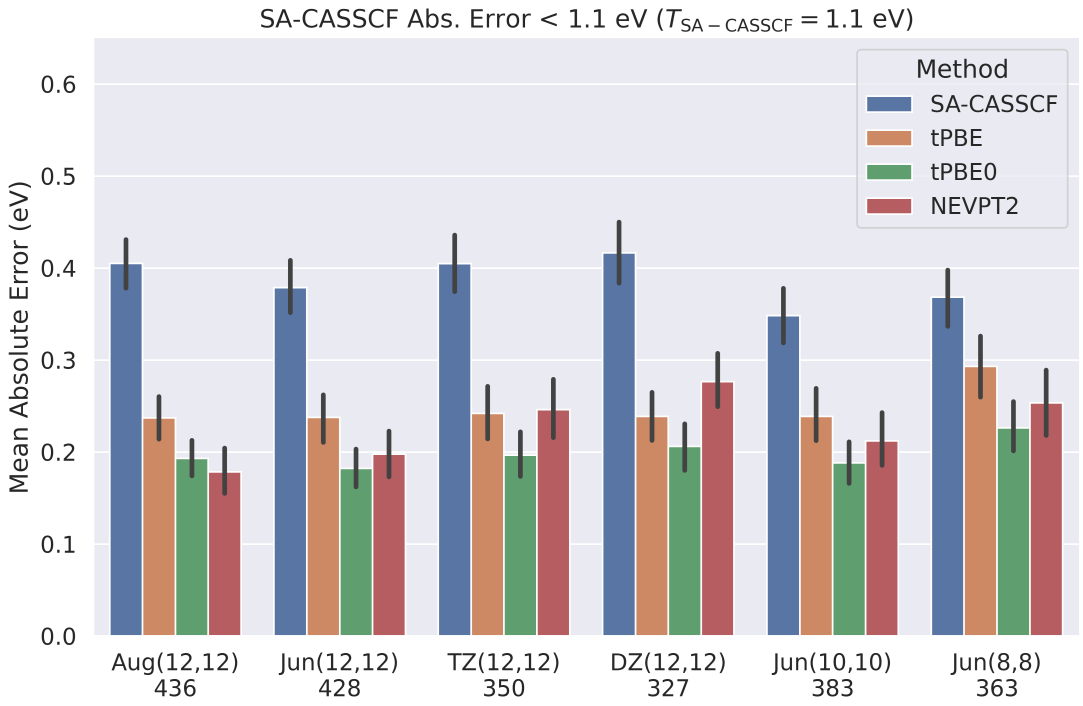

Figure 3.2: Comparison of the mean absolute errors of SA-CASSCF, tPBE, tPBE0, and NEVPT2 excitations across different active space and basis set sizes included by $T_{\mathrm{SA\text{-}CASSCF}} = 1.1$ eV. The number of excitations included in this analysis for each combination of active space and basis set is shown below each group of bars, and 95% confidence intervals for each mean are shown in black.

Figure 3.2 shows the performance of SA-CASSCF, tPBE, tPBE0, and NEVPT2 at different active space and basis set sizes after using the 1.1 eV SA-CASSCF error cutoff to eliminate poor active space choices. As expected, instead of observing an increasingly poor performance for SA-CASSCF excitations as active space and basis set size is decreased, we instead see a consistent error of roughly 0.39 ± 0.03 eV with an increasing amount of excitations excluded by $T_{\text{SA-CASSCF}} = 1.1$ eV. The number of excluded excitations roughly doubles from 19.2% at Aug(12,12) to 39.6 % at DZ(12,12) with decreasing basis size and to 32.4 % at Jun(8,8) with decreasing active space size. We note the very small increase of 8 excluded excitations upon moving from Jun(12,12) to Aug(12,12), highlighting the very efficient nature of the jun basis set.[155] Of course, with a better automatic active space selection scheme one would observe an increased amount of excitations included at each active space and basis set size, but the error will remain fairly consistent.

As we found for SA-CASSCF, we find that tPBE (0.25 ± 0.02 eV) and tPBE0 (0.20 ± 0.02 eV) maintain relatively consistent errors across different active spaces and basis set sizes when the 1.1 eV SA-CASSCF error threshold is applied. This is an intuitive result, as the accuracy of MC-PDFT is primarily contingent on the quality of the SA-CASSCF density and on-top density and on the quality of the on-top functional; if one eliminates the poor active spaces, then the functional (correlation) error may dominate, and this is approximately independent of the active space and basis set. In contrast, NEVPT2 shows quantitatively worse results as the basis set is decreased even as the wave function remains qualitatively well-described. This makes sense because the power of NEVPT2 to change the SA-CASSCF energy stems from its perturber states, which are less capable of describing dynamic correlation within a smaller basis because the smaller basis set cannot represent the virtual-orbital space as well.[32,33,158]

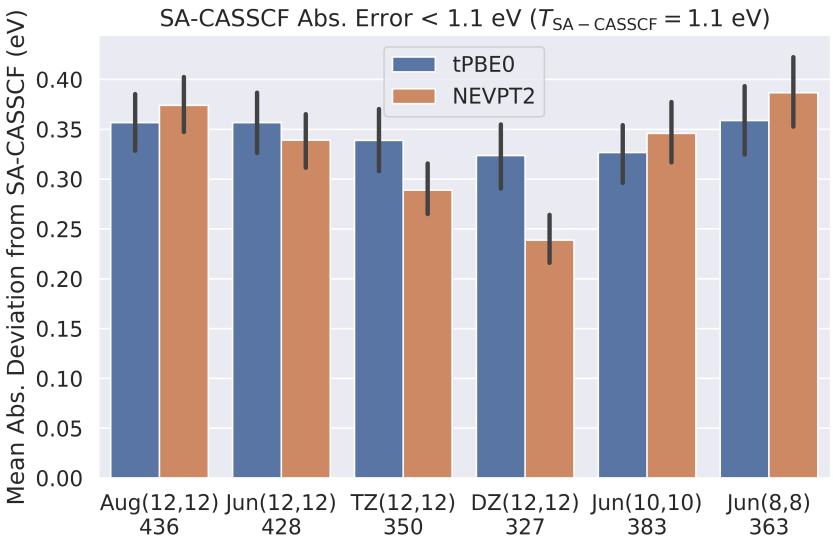

Figure 3.3: Mean absolute changes to the SA-CASSCF excitation energy made by tPBE0 and NEVPT2 across different active space and basis set calculations included by $T_{\text{SA-CASSCF}} = 1.1$ eV.

Figure 3.3 shows that the difference between the tPBE0 excitation energy and the SA-CASSCF excitation energy remains fairly consistent across active spaces and basis sets, but there is a significant drop in the NEVPT2 correction when moving from aug-cc-pVTZ to cc-pVDZ, resulting in increased NEVPT2 error. Figure 3.3 combined with Figure 3.2 shows clearly how the NEVPT2 results degrade in quality with decreasing size of the basis set, while the performance of tPBE0 remains consistent. As the basis set is decreased in size, the mean absolute change to the SA-CASSCF excitation energy decreases for NEVPT2 while remaining constant for tPBE. These results provide a plausible explanation of the discrepancy in mean absolute error found for SC-NEVPT2 between the study of Schapiro et. al.[159] (0.23 eV) and the more recent study of Sarkar et. al.[160] (0.15 eV). They imply that it is due to the fact that the Sarkar study used the aug-cc-pVTZ basis while the Schapiro study employed the cc-pVTZ basis. However, our results point to this being caused by a strictly poorer performance of NEVPT2 with the smaller basis set and not due to a poorer zeroth-order description of the underlying wave functions.

Further discussion of the error threshold is given in the Supporting Information, which shows the 1.1 eV SA-CASSCF error threshold to be optimal (albeit imperfect) for isolating subsets of automated wave function calculations that reproduce results curated by hand.[160] Additionally, we analyze alternative error thresholds on the NEVPT2 and tPBE0 error. However, an error criterion cannot be used when a benchmark excitation energy or experimental excitation energy is not available. Nevertheless, when an accurate value is not available, one can still use this criterion (although with somewhat less reliably) by comparing to one's best estimate rather than to an accurate value. Clearly, if one's best estimate is good, this will work as well as comparing to an accurate value.

Finally, one might wonder what one can do to fix the active space if a calculation goes poorly. Of course, increasing the size of the active space via $N_{\mathrm{CSF}}^{\mathrm{Max}}$ is a worthwhile option to explore if affordable, and it is clearly seen in Figure 3.2 how this significantly increases the success rate of the selection algorithm. However, following our previous work,[5] we also recommend experimenting with different orbital localization schemes for initializing the ranked-orbital selection as this can be a low-cost way to converge to a reasonable result.

3.4.2 Comparison to Single-Reference Methods

Data Overview. In the QUESTDB database,[45] excitations from many methods are only reported for the 491 excitations from closed-shell (S_0) molecules, and, due to double excitations and strongly mixed states, results from most methods are only available for about 460 of these excitations (supporting information). Our Aug(12,12) results comprise of 436 excitations included by $T_{\mathrm{SA-CASSCF}} = 1.1$ eV, 399 of which come from closed-shell molecules. Combining all methods and leaving out STEOM-CCSD, CCSDR(3), and CCSDT-3 for which there is significantly less available data (supporting information), there are a total of 373 excitations consistently available for comparison with SA-CASSCF, tPBE0, NEVPT2, and 12 other methods in the QUESTDB database. Unlike Jacquemin and coworkers, we have

not limited ourselves to comparisons based on "safe"[45,148] excitations, and this includes 29 excitations that would otherwise have been excluded (out of a total of 57 unsafe excitations in the total set of 542).

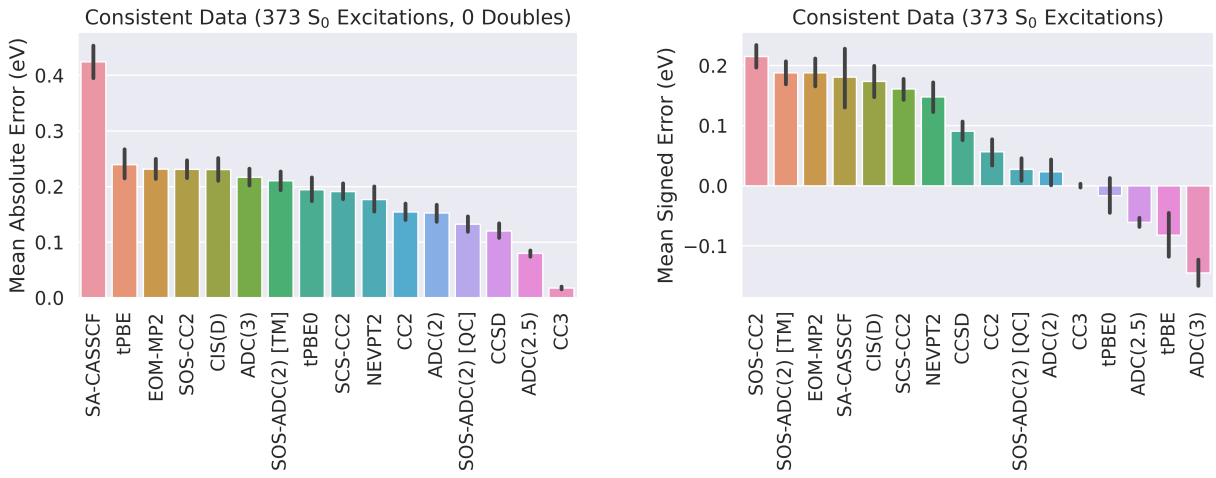

Figure 3.4: Comparison of the mean signed and unsigned errors of various methods on the 373 Aug(12,12) excitations included by $T_{\mathrm{SA-CASSCF}} = 1.1$ eV error threshold. The 95% confidence intervals are shown in black. Left: Mean absolute errors. Right: Mean signed errors.

Figure 3.4 shows the mean absolute and signed errors of SA-CASSCF, tPBE, tPBE0, and NEVPT2 in comparison with 12 other methods in the QUESTDB database on the set of 373 excitations. First considering the mean absolute errors, we find that both NEVPT2 (0.18 eV) and tPBE0 (0.19 eV) have accuracy on par with CC2 (0.15 eV), with tPBE lagging significantly behind (0.24 eV). However, we note that the errors we report here for tPBE, tPBE0, and NEVPT2 are likely slightly overestimated, as our CASSCF error threshold is imperfect and fails to eliminate all cases with poor active space choices (as discussed in the Supporting Information). Furthermore, this consistently available data set excludes all double excitations, for which the performance of the multireference methods is far superior (as discussed below).

Nevertheless, Trends in the signed errors are particularly interesting, with all but four methods (ADC(3), ADC(2.5), tPBE, and tPBE0) overestimating excitations; this implies

biased relative overstabilization of the ground state for most of the methods. One can clearly see how tPBE0 benefits from balancing the treatment of exchange and correlation, with SA-CASSCF overestimating excitations by 0.18 eV and tPBE underestimating by 0.08 eV such that tPBE0 has nearly zero mean signed error. We note that the same good balance seems to occur in ADC(2.5),[161] which averages ADC(3) and ADC(2).

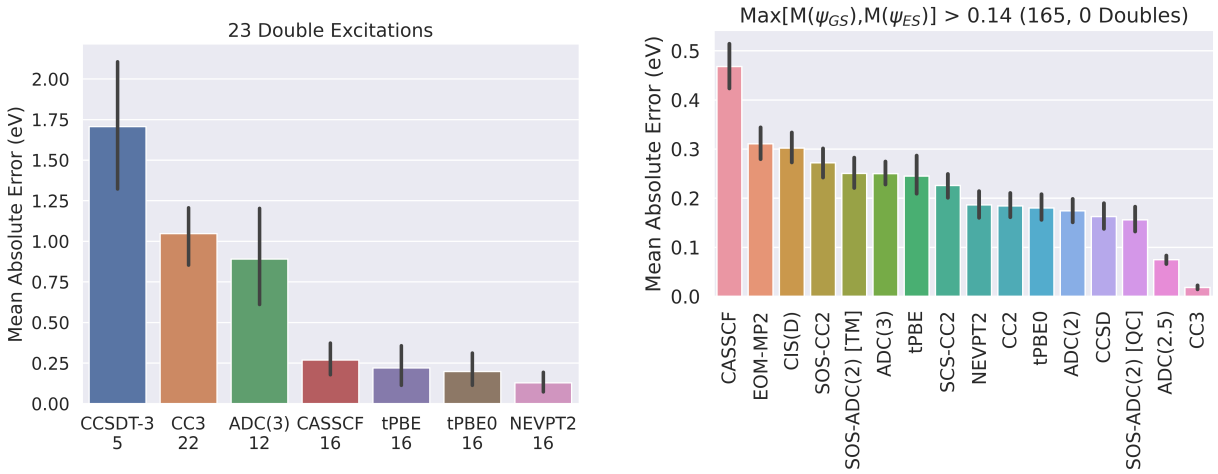

Figure 3.5: Left: Comparison of the mean absolute error of different methods on the entire subset of 23 double excitations in the QUESTDB dataset. The amount of excitations available for each method (with SA-CASSCF, tPBE, tPBE0, and NEVPT2 included via a 1.1 eV SA-CASSCF error threshold) is marked under each bar. Right: Comparison of the mean absolute errors of various methods on the 165 Aug(12,12) excitations included by $T_{\mathrm{SA-CASSCF}} = 1.1$ eV with high multireference character ($\mathrm{Max}[\mathrm{M}(\psi_{GS}),\mathrm{M}(\psi_{ES})] > 0.14$) and data available for every method shown, where M is the M diagnostic[3] of the corresponding wave function. 95% confidence intervals are shown in black.

The left of Figure 3.5 shows the mean absolute error of different methods on excitations classified as double excitations for all methods with any calculated double excitations in the QUESTDB database; our automated approach was able to converge results within the 1.1 eV SA-CASSCF error threshold for 16/23 (70%) of the double excitations that have TBEs available, which is only slightly lower than the overall un-dropped-out fraction of 436/542 (80%) in the Aug(12,12) calculations. In keeping with the usual recommendation to use multireference methods for this class of excitation, we find that multireference methods are

the only methods that perform consistently well on double excitations.

However, double excitations are not the only category of excitations which are a challenge for single-reference approaches. To quantify the multireference character of single excitations we have calculated the M diagnostic[3] of the ground and excited state Aug(12,12) wave functions included by the 1.1 eV SA-CASSCF error threshold and use the maximum of these values, M_{Max} = Max[M(ψ_{GS}),M(ψ_{ES})]. Doing so, we find that the lowest M_{Max} calculated for a double excitation is 0.14 (supporting information), and use this threshold as a classifier for identifying highly multireference single excitations; it happens to fall at slightly above the 50th percentile in the M_{Max} distribution (supporting information). The right of figure 3.5 shows a comparison of the mean absolute errors of SA-CASSCF, tPBE, tPBE0, and NEVPT2 to 12 single-reference methods on the 165 excitations with $M_{Max} > 0.14$ included by $T_{SA-CASSCF} = 1.1$ eV and data available for every method shown. Comparing to Figure 3.5, we find that the performance of nearly all single-reference methods deteriorates significantly by about 0.05-0.07 eV when we consider only this high-M_{Max} subset; this brings the performance of CCSD into line with tPBE0.

In summary, we find tPBE0 and NEVPT2 to perform competitively on single excitations when compared to single-reference methods (Figure 3.4 and Figure 3.5) and to be the only methods capable of reasonably describing double excitations (Figure 3.5). As such, we recommend tPBE0 and NEVPT2 as robust methods for calculating all classes of vertical excitations, although the active space selection scheme may sometimes fail.

3.4.3 Performance by Excitation Type

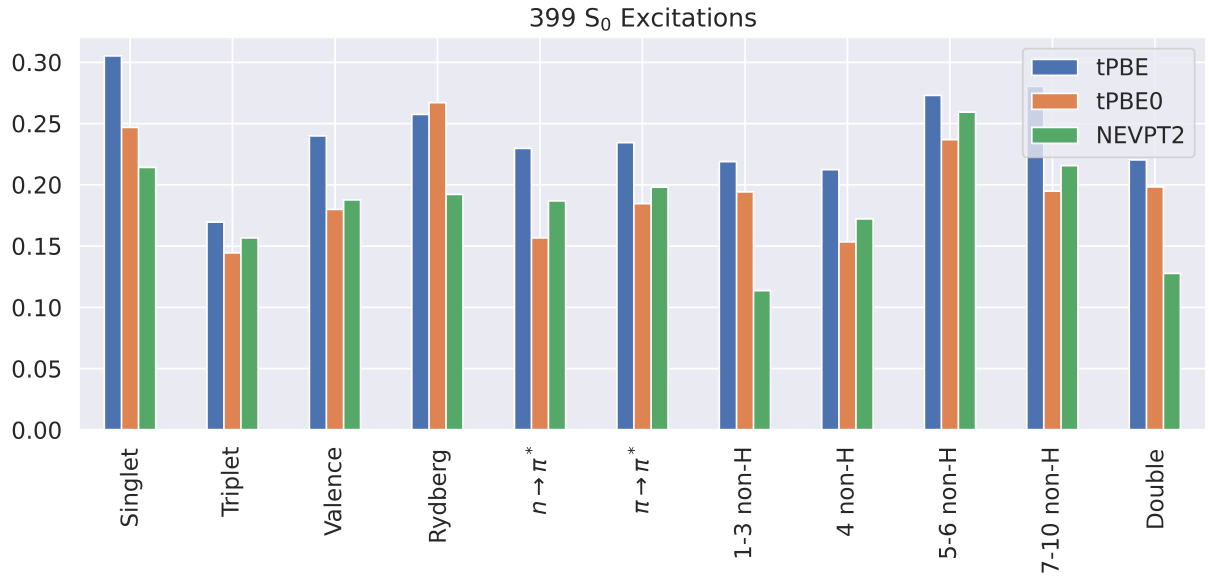

Figure 3.6: Mean absolute errors (in eV) of tPBE, tPBE0, and NEVPT2 Aug(12,12) calculations on various types of S_0 excitations included by the threshold $T_{\text{SA-CASSCF}} = 1.1$ eV.

Figure 3.6 shows the errors classified by excitation type. In line with the Sarkar study,[160] we find that NEVPT2 is more accurate for triplet excitations than singlet excitations, and tPBE and tPBE0 follow this same trend. The figure shows that, with the exception of Rydberg states, tPBE0 has better performance than tPBE for every excitation category, and therefore we recommend the use of tPBE0 rather then tPBE for calculating excitation energies of valence excitations. We also recommend tPBE0 for calculating a spectrum containing both valence and Rydberg excitations since the performance of the two methods is very similar (on average) for Rydberg states.

3.4.4 Method Timing : tPBE0 vs. NEVPT2

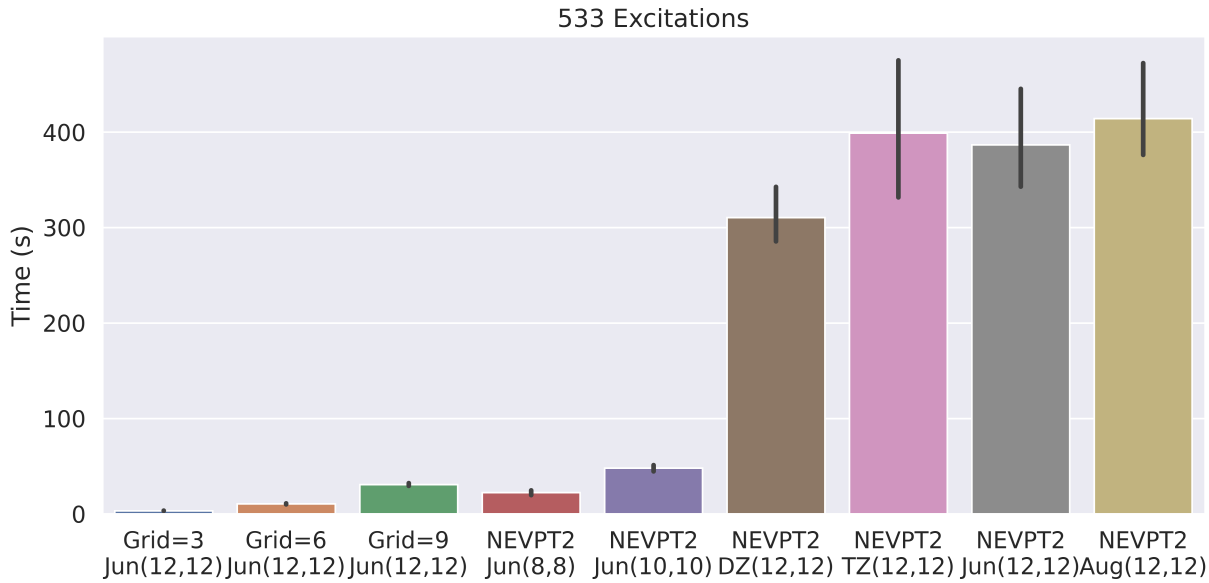

Figure 3.7: Comparison of the mean compute times for the post-SCF portion of tPBE calculations with various grid specifications and for the post-SCF portion of NEVPT2 calculations with various active spaces and basis set sizes on the set of 533 excitations that were converged with all active spaces and basis sets. The costs of the SA-CASSCF portions of the calculations were removed from these comparisons by caching the converged wave functions.

Figure 3.7 shows the average time consumed by the calculation of the NEVPT2 perturbative correction at different active space/basis set sizes and compares these timings to those for the calculation of the tPBE on-top energy by the methodology in section 3.3. We find that at the normal grid size (grids_level = 3 in PySCF), tPBE is on average 114× less expensive than NEVPT2 for the large max(12,12) active spaces. This is because – as is well known – the cost of NEVPT2 scales very poorly with the size of the active space, while the cost of tPBE0 remains independent of that. Furthermore, the memory required for NEVPT2 also increases with active space size. It is around the max(12,12) active space size that the compute time for the perturbative correction begins to exceed the compute time of the underlying SA-CASSCF step, while the compute time of tPBE remains low.[162] For smaller active spaces such as max(8,8), the cost of NEVPT2 is comparable to that of tPBE and

tPBE0.

Keeping the cost down is important in many applications. Figure 3.7 shows that the dependence of MC-PDFT compute times on grid size is a significant consideration; we observe a roughly 10× increase in cost from grids_level = 3 (3.4 s) to grids_level = 9 (30.8 s). Our studies find that the standard grids_level = 3 in PySCF is sufficient for excitations such as those we have calculated because we only see a significant change between the maximum and default grid size for a single excitation (supporting information). Therefore we recommend standard grid sizes for most applications involving state-averaged MC-PDFT.

3.4.5 Method Timing : tPBE0 vs. CC2 and CCSD

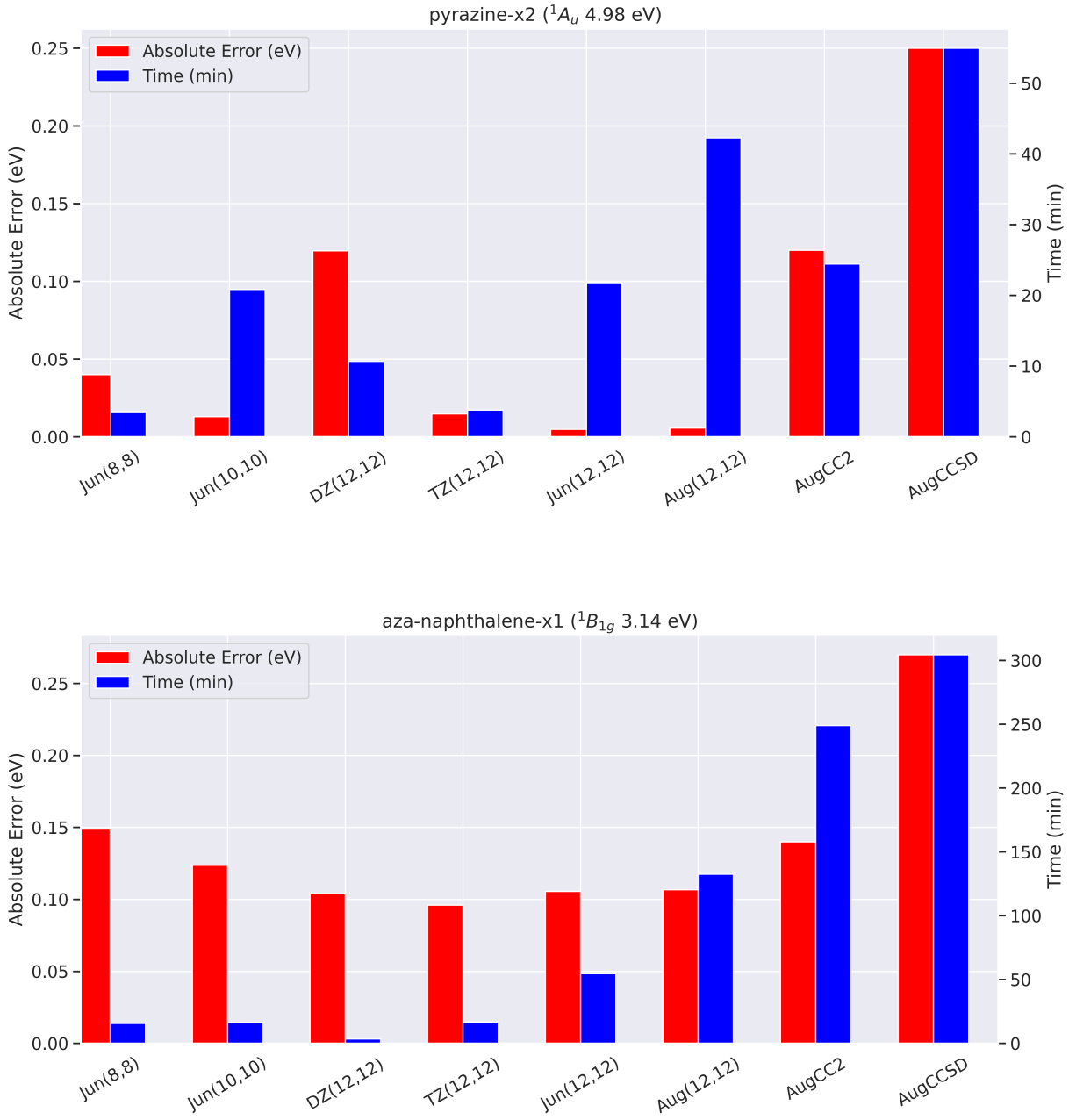

Figure 3.8: Comparison of timings and accuracy between tPBE0 at the six active space/basis set combinations explored in this work and CC2 and CCSD in the aug-cc-pVTZ basis. Timings for tPBE0 include the steps of RHF convergence, Boys orbital localization, active space selection, CASSCF optimization, and computation of the tPBE0 nonclassical energy. Timings for CC2 and CCSD were computed in the aug-cc-pVTZ basis using their implementation in Psi4[4] and were confirmed to reproduce the Jacquemin results.

In an effort to give greater context to the standing of tPBE0 as a method for calculating vertical excitations outside of cases where multireference methods are absolutely needed (such as double excitations), we compare the timings of complete CASSCF+tPBE0 calculations to those of CC2 and CCSD. Figure 3.8 shows the comparison of such timings for tPBE0 (including RHF convergence, Boys orbital localization, active space selection, CASSCF optimization, and computation of the tPBE0 nonclassical energy) and CC2 and CCSD as computed in Psi4[4] for two excitations in QUESTDB. All calculations were given the same amount of computational resources as outlined in the supporting information. We have chosen to show timings and accuracies for both a "medium-sized" excitation (pyrazine-x2, with 368 aug-cc-pVTZ basis functions) and a "large-sized" excitation (aza-naphthalene-x1, with 552 aug-cc-pVTZ basis functions). Additionally, Figure 3.8 shows timings and accuracies for tPBE0 and all 6 of the active space and basis set combinations explored in this work.

Focusing first on the aug-cc-pVTZ calculations, one can see that tPBE0 takes a comparable amount of time compared to CC2 and CCSD, both for pyrazine and aza-naphthalene. However, in both of these cases costs can be cut significantly while maintaining accuracy by decreasing active space and basis set size. As demonstrated by Figure 3.8, through a judicious choice of active space and basis set, tPBE0 has the potential to be much less expensive than comparative single-reference approaches while achieving similar accuracy or better. For aza-naphthalene-x1, tPBE0 is about 16× as fast as CC2 at Jun(8,8) and about 72× as fast at DZ(12,12). The speedup one can obtain tends to be greater when considering larger systems.

However, the idealized (albeit real) case shown for aza-naphthalene-x1 is far from general. Firstly, one can only reduce basis set and active space size so far before one's results become highly inaccurate with tPBE0, and the point at which this happens is highly excitaiton dependent and somewhat dependent on the active-space-selection scheme. Secondly, the timing behavior of CASSCF+tPBE0 is not always as well behaved: CASSCF optimization

relies on a highly nonconvex and nonlinear optimization process which may not conform expected timing trends. An example of this can be seen in the pyrazine-x2 timings in Figure 3.8, where tPBE0@DZ(12,12) takes significantly more time than tPBE0@TZ(12,12). Further taking into account differences between implementations, we present Figure 3.8 only to give readers a rough sense of timings for tPBE0 with respect to comparably accurate single-reference methods on different system sizes.

Additionally, we attempted to compute timings for CC3 for these two excitations: the pyrazine-x2 result was computed in 1765 min (29 hours) and aza-naphthalene-x1 was not able to finish within the 36 hour time limit allowed by the resources available for these calculations. Finally, we note that CCSD also includes an iterative step, but a study of the convergence issues in CCSD is beyond our scope.

3.4.6 Optimizing the Mixing Parameter in Hybrid tPBE

A major motivation of this work was to generate data for benchmarking and improving MC-PDFT. As a first use of our data to optimize MC-PDFT functionals, we have investigated the optimal mixing parameter λ for hybrid tPBE (htPBE, for which the energy is given by $\lambda E_{\mathrm{SA-CASSCF}} + (1-\lambda)E_{\mathrm{tPBE}}$) over the Aug(12,12) database. We have chosen this set of excitations because it is likely to have the smallest amount of poor active spaces erroneously included by the $T_{\mathrm{SA-CASSCF}} = 1.1$ eV error threshold. In other words, we expect this set of excitations to have the largest percentage of well-chosen active spaces.

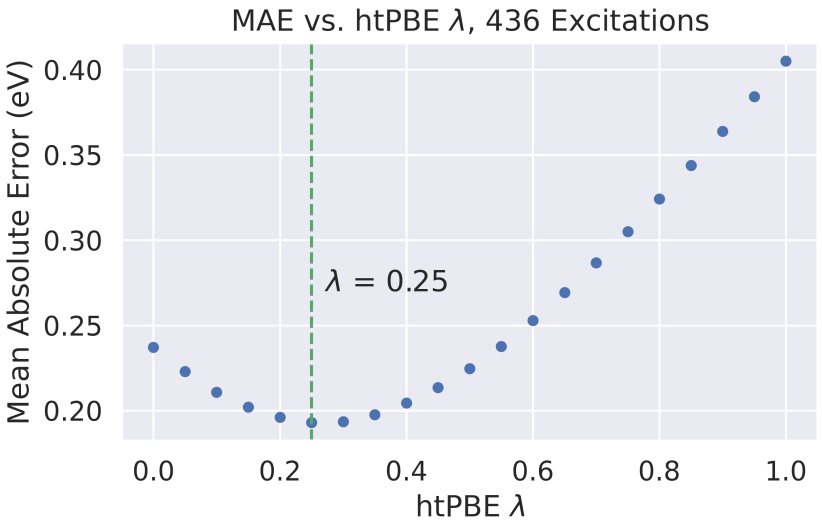

Figure 3.9: Mean absolute errors of different mixing parameters λ in energies computed by htPBE for the 436 Aug(12,12) excitations included with $T_{\mathrm{SA-CASSCF}} = 1.1$ eV. The optimal value of $\lambda = 0.25$ (the same as in tPBE0) is marked with a dashed green line.

Figure 3.9 shows the optimization of λ on the Aug(12,12) set of included excitations. Delightfully, we find that $\lambda = 0.25$ – the same parameter used in tPBE0 – is optimal for this set of excitations, in agreement with the much smaller study previously conducted on the EE27 database.[140] Therefore we recommend using tPBE0 for excitation energies in the general case and especially for excitations similar to those in the QUESTDB dataset. Optimizing the parameter over all active spaces and basis sets results in only a slightly shifted value of $\lambda = 0.3$, which appears to be offset mostly by the greater number of poor active spaces included in the Jun(8,8) excitation energies (supporting information); using the more robust tPBE0 error threshold (discussed in the SI) removes this discrepancy (supporting information). This suggests that a higher value of λ may be optimal for cases in which wave function error dominates.

3.5 Conclusion and Future Work

The work presented here is the largest application to date of automated multireference calculations on a broad range of molecules. The generation of 3237 multireference excitation energies has allowed us to gain insight into how to eliminate poorly chosen active spaces and has identified trends in the performance of MC-PDFT and NEVPT2. This work has been possible only through the careful work of Loos, Jacquemin, and coworkers in compiling the QUESTDB dataset[44,45,128,147,148,163] and the recent work of of Sarkar et. al.[160] which has enabled us to compare our automatically generated results to hand-selected active space calculations.

We see this initial publication as laying the groundwork for several future applications related to MC-PDFT and high-throughput multireference calculations including:

- Using the generated data to train and test novel functionals for MC-NCFT, representing a continuation of our initial work that used carbene singlet-triplet excitation energies to train machine-learned functionals.[48]

- Improving the active space selection scheme. Our finding that error thresholds can be used to determine the fraction of poor wave functions in the calculated excitation energies can be used as a measure to benchmark the effectiveness of different active space selection schemes.

- Determining if a selected active space is well-chosen without reference to the underlying benchmark values. For specific active spaces and basis sets there appears to be promise in looking at differences between different methods (supporting information), but a method that is generalizable across active spaces and basis sets has yet to be found.

Additionally, we expect that the wave functions converged in this work will be of interest for the development of different post-CASSCF methods such as multireference adiabatic

connection (AC)[164] and algebraic diagrammatic construction (ADC).[165] For this reason, we are making all 3237 converged wave functions freely available for public use. We hope that this data will be useful to the electronic structure community both for comparing to the results published here and for developing and testing their own methods.

In summary, we have carried out the largest benchmarking of SA-CASSCF and MC-PDFT to date. This was accomplished by means of an automatic active-space selection scheme and use of a SA-CASSCF error threshold to eliminate poor active-space choices. On a set of 373 aug-cc-pVTZ excitation energies, we find that tPBE0 and NEVPT2 perform with similar accuracy to CC2, while tPBE lags behind. However, the accuracy of NEVPT2 degrades with basis set size even as the quality of the underlying density and on-top pair density appear to remain the same. As expected, we find that tPBE0 is orders of magnitude less expensive than NEVPT2 for larger active spaces, and we recommend its use for the calculation of a broad range of excitation energies, including double excitations.

This work is supported by the National Science Foundation under grant CHE-2054723. Additionally, the authors thank the Research Computing Center (RCC) at the University of Chicago for access to computational resources, as well as Matthew Bousquet for helping to identify and categorize the QUESTDB dataset. Finally, the authors thank Ricardo Almada Monter for help computing timings for CC2, CCSD, and CC3.

We have made available all the data necessary to reproduce and explore our results (orbitals, CI vectors, and metadata for all 3237 converged SA-CASSCF calculations) via Zenodo: https://doi.org/10.5281/zenodo.6644169.

CHAPTER 4

VARIATIONAL ACTIVE SPACE SELECTION WITH MULTICONFIGURATION PAIR-DENSITY FUNCTIONAL THEORY

This chapter is reprinted with permissions from *J. Chem. Theory Comput.* **2023**, *19*, 22, 8118-8128

4.1 Abstract

The selection of an adequate set of active orbitals for modeling strongly correlated quantum states is difficult to automate because it is highly dependent on the states and molecule of interest. Although many approaches have shown some success, no single approach has worked well in all cases. In light of this, we present the "discrete variational selection" (DVS) approach to active space selection in which one generates multiple trial wave functions from a diverse set of systematically constructed active spaces and then selects between these wave functions variationally. We apply this DVS approach to 207 vertical excitations of small-to-medium-sized organic and inorganic molecules (with 3 to 18 atoms) in the QUESTDB database by (i) constructing various sets of active space orbitals through diagonalization of parameterized operators and (ii) choosing the result with the lowest average energy among the states of interest. This approach proves ineffective when variationally selecting between wave functions using the DMRG/CASSCF energy, but is able to provide good results when variationally selecting between wave functions using the energy of the tPBE functional from multiconfiguration pair-density functional theory (MC-PDFT). Applying this DVS-tPBE approach to selection among state-averaged density matrix renormalization group (SA-DMRG) wave functions, we obtain a mean unsigned error of only 0.17 eV using hybrid MC-PDFT. This result matches that of our previous benchmark without the need to filter out poor

active spaces, and with no further orbital optimization following active space selection of the SA-DMRG wave functions. Furthermore, we find that DVS-tPBE is able to robustly and effectively select between the new SA-DMRG wave functions and our previous SA-CASSCF results.

4.2 Introduction

The accurate treatment of excited electronic states of molecules is a long-standing and active area of research in computational chemistry.[43,116–127,166–170] It is especially difficult when a single-determinant ground state provides a poor reference for computing the excited states (e.g., double excitations[128,171] or strongly correlated systems[129–132,172]). A useful form of wave function for overcoming such difficulties is the complete active space configuration interaction (CASCI) trial function:

$$|\Psi_{\text{CASCI}}\rangle = |22...2\rangle \wedge \sum_{n_1 n_2..n_L} C_{n_1 n_2..n_L} |n_1 n_2 ... n_L\rangle \tag{4.1}$$

in which $|22...2\rangle$ is a single Slater determinant consisting of doubly occupied orbitals (called inactive orbitals in the CASCI context), the $C_{n_1 n_2..n_L}$ are coefficients, and the determinants $|n_1 n_2 ... n_L\rangle$ span the space of all possible configurations obtained by distributing a fixed number N_{elec} of active electrons among L active orbitals. Each determinant $|n_1 n_2 ... n_L\rangle$ is defined by its orbital occupation numbers $n_i \in \{0, \uparrow, \downarrow, 2\}$ of the active electrons in the active orbitals, and diagonalization of the Hamiltonian in this space (and in the mean field of $|22...2\rangle$) is known as CASCI. However, because the size of the space scales exponentially with the number of orbitals L, this approach is only feasible up to active space sizes of about 20 electrons in 20 orbitals.[26]

Many methods exist to approximate the solution for the coefficients $C_{n_1 n_2..n_L}$ in equation 4.1.[173] Among the most successful approaches in this regard is the density matrix renormal-

ization group configuration interaction (DMRG) method,[28,69,70,168,174–183] in which the coefficients of equation 4.1 are approximated by the matrix product[176]

$$C_{n_1 n_2 .. n_L} = \sum_{ij...(L-1)} A_i^{n_1} A_{ij}^{n_2} A_{jk}^{n_3} ... A_{(L-1)}^{n_L} \tag{4.2}$$

In equation 4.2, each possible occupation of each active orbital n_i is given by its own matrix or vector A^{n_i}. The maximum inner dimension of these matrices is called the bond order M and is the number of states retained during the renormalization step. As $M \to \infty$, results obtained with this method approach those obtained with full diagonalization (although useful results for well-chosen active spaces are generally obtainable with practical values of M). Using this approach, it is possible to describe active spaces with up to about 100 orbitals.[31]

The success of CAS-based methods relies heavily on the construction and selection of the orbitals defining the active space, because this selection affects both the convergence of self-consistent-field iterations and the quality of the energetic results. Variationally optimizing the active-space orbitals is known as CAS self-consistent field (CASSCF)[68] when used with a full-configuration-interaction solver or as DMRG-SCF[184,185] when used with a density-matrix-renormalization-group solver. To try to obtain a consistent treatment of multiple states, one may optimize the state-averaged (SA) energy with respect to the active orbitals, yielding SA-CASSCF[186] or state-averaged DMRG-SCF (SA-DMRG-SCF).[166,185,187] (If orbitals are predetermined rather than optimized, one may obtain state-averaged DMRG (SA-DMRG)). We emphasize two difficulties with conventional methods of optimizing orbitals: (i) the energetic optimization is prone to converging to local minima, and (ii) state-averaged variational optimization of orbitals is not necessarily optimal for computing energy differences between states, especially when states with different characters are considered;[188] the latter of these difficulties is made worse by the fact that the CASSCF orbitals are generally optimized without regard for post-CAS correlation generally included in the computation of

excitation energies. Furthermore, orbital optimization significantly increases the cost of the computation. Thus, although SCF generally helps improve the quality of the active space, it does not eliminate the need to develop good active space construction and selection schemes for excited-state calculations, and it comes at a computational cost.

Because of the above considerations, active-space construction and selection remains a vigorous area of research, and several approaches have been proposed to date.[1,3,5,6,30,31,49,75–79,81,87,103,141–143,170,189–202] The most commonly applied method method involves chemical intuition with trial and error.[30,190] However, this approach is unsystematic and difficult to apply in a high-throughput fashion. In recent years, there has been much interest in developing more systematic methods for fashioning active spaces.[1,3,5,6,31,49,75–79,81,87,103,141–143,170,189,191–201] A tool called AVAS, developed by Sayfutyarova and coworkers,[49] allows one to semiautomatize the active-space construction by using molecular orbitals that overlap optimally with a user-selected set of atomic orbitals. Other approaches involve some preliminary calculations, such as the natural orbital occupancies of a unrestricted Hartree-Fock (UHF) calculation,[7,83,87,189,191] entanglement information from a large DMRG calculation,[31,76,86,203] the quantitative accuracy of some physical observable such as the dipole moment,[201] machine learning predictive models,[81,143] and physically motivated equations based on information such as HF matrix elements.[5,6] An assumption of all these approaches is that the key physics necessary to construct and select the active space can be captured by a preliminary calculation (UHF, DMRG, etc.). However, as we will show, even when selecting large active spaces (e.g., with 40 orbitals) for small molecules, it is difficult to make even qualitatively accurate active spaces for any given excitation with a single method.

In recent work,[5] we employed one such automated approach, approximate pair coefficient (APC) active space selection, on the extensive QUESTDB database[148] of accurate vertical excitation energies for small-to-medium-sized organic systems. Through this, we were able to

carry out extensive benchmarking of post-SA-CASSCF methods such as n-electron valence perturbation theory (NEVPT2)[34,35] and multiconfiguration pair-density functional theory (MC-PDFT)[41] using the translated PBE (tPBE) functional with active spaces generated by the automated approach. However, to ensure accurate evaluation of the post-SA-CASSCF methods and distinguish errors arising from poor active spaces, we considered only the active spaces for which the SA-CASSCF result fell within 1.1 eV of the best estimate in the QUESTDB database.[5] This criterion was satisfied for 363–436 (68–82%) of the 532 excitations in the database, depending on the active-space size and basis set. Although the APC scheme for these cases proved to be competitive with active spaces selected by hand[137,159,204], it was observed that the remaining active spaces (18–32% of the excitations) exhibited very high errors, sometimes exceeding 5 eV. Consequently, the predictive utility of these active spaces for our purposes was only partially acceptable.

In the present study our objective is to develop a new framework for active space selection that is more broadly accurate for predicting vertical excitation energies. The key element of the new method is the premise that *no single active-space-selection scheme will be successful in all cases*. Therefore, we hypothesize that the important missing component of the current schemes is the lack of a way to effectively *choose* between active spaces generated by different methods or different parameters. To address this, we propose the "discrete variational selection" (DVS) approach to active space selection in which (i) one generates trial wave functions with a variety of active spaces constructed with different methods (e.g., any of those mentioned or cited above) or different parameters, and then (ii) one chooses between the generated active-space wave functions variationally.

In this work, we apply this DVS approach to the calculation of 207 vertical excitations of small-to-medium sized main-group molecules in the QUESTDB database.[148] These excitations provide a rich variety of different types of excitations (e.g., Rydberg and valence excitations of organic molecules, including both $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ excitations, and ex-

citations of inorganic molecules) and thus present a demanding challenge to the systematic prediction of their excitation energies. We find that the DVS scheme is unsuccessful when variationally selecting between results using the CASSCF/DMRG energy, but performs well when applied using the tPBE energy from MC-PDFT. Applying this DVS-tPBE approach to selection among systematically constructed wave functions with SA-DMRG, we are able to obtain a mean unsigned error of only 0.17 eV with hybrid MC-PDFT. This result reproduces that of our previous benchmarks of hybrid MC-PDFT[6] without the need to filter out poor active spaces, and with no furhter orbital optimization following the active space selection of the SA-DMRG wave functions. Furthermore, we find that DVS-tPBE is able to robustly select between the newly generated SA-DMRG wave functions and our previously generated SA-CASSCF results of our previous study.[6]

4.3 Theory and Methods

In this section, we provide an overview of MC-PDFT[41] and hybrid MC-PDFT,[140] and provide a detailed descripion of the approach used in this work to systematically construct the active spaces and SA-DMRG wave functions for DVS-tPBE. Finally we provide a detailed description of the QUESTDB data used to judge the performance of this method.

4.3.1 Multiconfiguration Pair-Density Functional Theory

The energy expression of multiconfiguration pair-density functional theory may be written as[41]

$$E_{\text{MC-PDFT}} = V_{nn} + \sum_{ij} h_{ij}\gamma_{ij} + \frac{1}{2}\sum_{ijkl} g_{ijkl}\gamma_{ij}\gamma_{kl} + E_{\text{NE}} \tag{4.3}$$

where V_{nn} is the nuclear repulsion, i, j, k, and l are orbital indices, h_{ij} is a one-electron integral, γ_{ij} is the one-electron density matrix, g_{ijkl} is a two-electron integral, and E_{NE} is the nonclassical-energy functional. In most of our work, E_{NE} is written as a function of the

electron density ρ and the on-top pair density Π and is called an on-top density functional. Recently, we have begun to explore different types of nonclassical-energy functionals derived from machine learning[48] or the density coherence,[139] and we refer the reader to a recent review.[43] However, all practical applications so far have employed a translated version of the PBE[18] Kohn-Sham functional which is an on-top functional denoted as tPBE.

The on-top functional may also be combined with the wave function exchange-correlation energy to form hybrid MC-PDFT,[140] for which the energy expression becomes

$$E_{\text{HMC-PDFT}} = XE_{\text{SA-CASSCF}} + (1 - X)E_{\text{MC-PDFT}} \tag{4.4}$$

where $E_{\text{SA-CASSCF}}$ is the SA-CASSCF energy computed by wave function theory, and X is a parameter. We have often found good results using tPBE with $X = 0.25$,[6] which is called tPBE0.

4.3.2 Systematically Constructed Active Spaces for DVS-tBPE

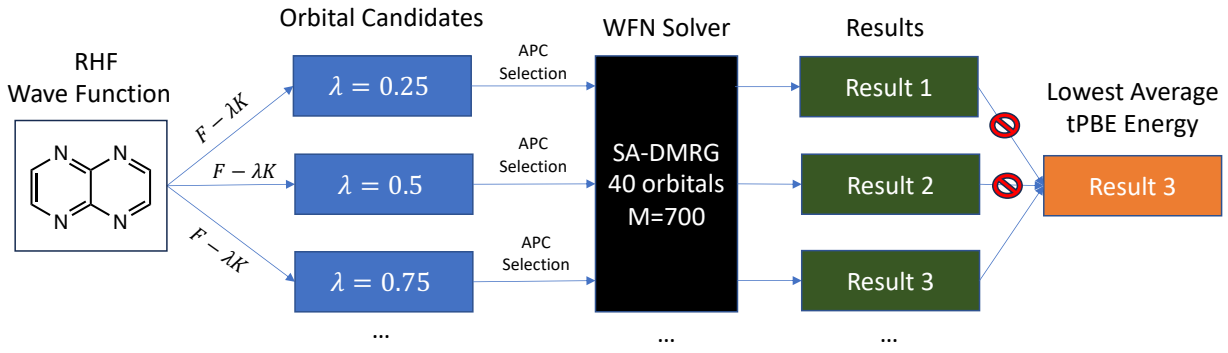

Figure 4.1: Schematic of the scheme used to systematically construct active spaces for DVS-tPBE. Starting from an RHF or ROHF wave function, different sets of orbitals are generated by diagonalizing $F - \lambda K$ in the space of virtual orbitals. Active spaces of 40 orbitals are then selected from these orbital candidates using APC selection.[5,6] Wave functions are then generated using these selected active spaces by SA-DMRG. The final step represents the DVS-tPBE approach in which the final result is chosen as the one with the lowest sum of the tPBE energies between the two states of interest.

The outline of the scheme used to systematically generate active spaces for DVS-tPBE is shown in Figure 4.1. The approach consists of (i) calculation of initial Hartree-Fock wave functions, (ii) virtual orbital construction via diagonalization of a parameterized operator, (iii) selection of active spaces with the approximate-pair-coefficient selection (APC) method,[5,6] and (iv) generation of SA-DMRG wave functions in these active spaces using Block2[205]. In the following we explain these components step-by-step.

Hartree-Fock Calculations. Hartree-Fock orbitals were generated for closed-shell singlet ground states by restricted Hartree-Fock (RHF) theory[206] and for doublet ground states by restricted open-shell Hartree-Fock (ROHF) theory[207] using the aug-cc-pVTZ basis[208,209] as was used for the theoretical best estimates listed in the QUESTDB database. (Note that good results for these excitations are likely achievable with the smaller jun-cc-pVTZ[155] basis as was observed in our previous benchmark.[6])

The definition of the Fock operator in ROHF theory is ambiguous, but the choice must be specified as it affects the basis of orbitals from which we select to form the SA-DMRG wave functions as well as the inputs into the APC theory[5,6] used to select the active space from these orbitals. For the present article, we employ Roothaan's effective Fock operator[207,210], which is the default choice in PySCF.[211,212]

Orbital Construction. Starting from the set of canonical orbitals obtained from the RHF or ROHF wave function, we index the doubly occupied orbitals with i and the virtual orbitals with a. We then generate multiple trial orbital sets for a calculation by diagonalizing the parameterized operator

$$G = F - \lambda K \tag{4.5}$$

in the space of the RHF or ROHF virtual orbitals, where F and K are the Fock and exchange matrices generated from the RHF or ROHF density matrix, and λ is a tunable parameter used to generate different sets of orbitals. Each of these trail orbital sets serves as a set of candidate orbitals from which the active space will be selected for separate multiconfigurational wave

function calculations; after their generation in this step they remain unchanged. The next step serves to select 40 active orbitals and a set of inactive orbitals from each of these initial sets of orbitals.

APC Active Space Selection. The approximate pair coefficient (APC) method is a method for estimating the one-orbital reduced density matrix entropies of candidate orbitals for the active space from HF matrix elements.[5] Using this approach, it is possible to efficiently estimate orbital importance for the active space (and thus rank the orbitals appropriately), as higher orbital entropy is a measure of higher multireference character. For a doubly occupied orbital i and virtual orbital a, the approximate pair coefficient between these two orbitals is defined by[5]

$$C_{ia} = \frac{0.5K_{aa}}{F_{aa} - F_{ii} + \sqrt{(0.5K_{aa})^2 + (F_{aa} - F_{ii})^2}} \tag{4.6}$$

where F and K are again the Fock operator and exchange operator generated from the HF density matrix. The entropies of doubly occupied orbitals and virtual orbitals are defined as

$$S_i = -\frac{1}{1+\sum_a C_{ia}^2} \ln \frac{1}{1+\sum_a C_{ia}^2} - \frac{\sum_a C_{ia}^2}{1+\sum_a C_{ia}^2} \ln \frac{\sum_a C_{ia}^2}{1+\sum_a C_{ia}^2} \tag{4.7}$$

and

$$S_a = -\frac{1}{1+\sum_i C_{ia}^2} \ln \frac{1}{1+\sum_i C_{ia}^2} - \frac{\sum_i C_{ia}^2}{1+\sum_i C_{ia}^2} \ln \frac{\sum_i C_{ia}^2}{1+\sum_i C_{ia}^2} \tag{4.8}$$

where the sums over i includes all HF doubly occupied orbitals, and the sums over a initially includes all virtual orbitals generated in the orbital construction step. We will eventually select high-entropy orbitals for the active space, but in our previous work[6] we have found the entropies calculated with the full sums to be overly biased towards doubly occupied orbitals, resulting in less-than-optimal active spaces. Therefore, we use a virtual-orbital removal step in which the C_{ia} involving the highest-entropy virtual orbital is removed from the sums in equations 4.7, and the entropies are recalculated. After N such virtual-orbital removal steps

are taken, the entropies of the removed virtual orbitals are set to the maximum entropy of the remaining orbitals plus some small value, decreasing in order of removal; we have found good results for small-to-medium-sized organic molecules with $N = 2$,[6] which is used in this work.

Having calculated orbital entropies for each trial set of doubly occupied orbitals i and virtual orbitals a (constructed in step (ii)), the 40 highest-entropy orbitals are selected as the active orbitals, and the other orbitals are dropped from the active space. Any dropped doubly occupied orbitals become inactive, whereas dropped virtual orbitals become secondary. As such, there are always 40 active orbitals in the active space of each subsequent SA-DMRG calculation, and the number of inactive orbitals is the number of doubly occupied orbitals dropped from the active space in the above selection stage. We note that although we do not exclude core orbitals from selection, they are highly biased against by the APC scheme (equation 4.6) and mostly harmless if added to the active space (generally when one has exhausted all other orbitals). The number of active electrons in each calculation is set to two times the number of doubly occupied orbitals remaining in the active set, and the number of inactive electrons is equal to two times the number of inactive orbitals.

Computation of SA-DMRG Wave Functions. Having selected a 40-orbital active space for each set of trial orbitals in step (iii) (one for each value of λ selected in step (ii)), density matrix renormalization group calculations were carried out without re-optimization of orbitals by using the state-averaged DMRG (SA-DMRG) in Block2[205] as integrated into PySCF.[211,212] The maximum bond dimension of these calculations (i.e., the maximum number of renormalized states) was fixed to $M = 700$. The choice of selecting 40-orbital active spaces with a bond dimension of $M = 700$ was made because it provided good accuracy while still remaining computationally affordable (here defined as being able to run on 24 Intel Cascade Lake cores with 96 GB of memory in less than a few hours for all systems). As in our previous study,[6] excited-state wave functions were calculated in a state-averaged fashion

averaging over the ground state and the required number of excited states (for example, to approximate the 2 $^1A_2$ state, we would include the ground state of symmetry $^1A_1$ and two states of symmetry $^1A_2$, as in our previous work[6]).

We then select among the multiple wave function results generated for each excitation via steps (i)-(iv) by variational selection with tPBE. In particular, we select the SA-DMRG wave function that yields the lowest sum of tPBE absolute energies between the ground state and the excited state of interest. Energies with tPBE were calculated using a version of PySCF that incorporates the MRH code[149] now available in PySCF-Forge.[213] Grid integration was carried out for evaluation of the on-top functional with fineness grids_level = 3, as judged to be sufficient in our previous benchmark study.[6]

4.3.3 Benchmarking Data

We investigate the approach described above on a subset of theoretical best estimates in the QUESTDB dataset[148] for vertical excitation energies of small-to-medium-sized main-group molecules. This set of excitations includes many of the most widely studied molecules of the quantum chemistry community (e.g., water, ethylene, and naphthalene) as well as a rich variety of different excited states (e.g., valence, Rydberg, $n \rightarrow \pi^*$, and $\pi \rightarrow \pi^*$ excitations in many organic molecules with as many as 18 atoms plus excitations in H_2S, HPO, HPS, HSiF, and HNO). Thus, it presents a difficult challenge for any active space selection scheme that hopes to be predictive in its calculations. We form a subset of these excitations by applying the following constraints:

- Excitations must be labeled as "safe" in the original QUESTDB dataset (considered by the authors of that work as chemically accurate or within 0.05 eV of the FCI limit for the given geometry and basis set).[148]
- The full symmetry of the molecule must be supported in the CAS module of PySCF; this limits us to molecules with symmetries C_s, C_{2v}, C_{2h}, and D_{2h}.

- The symmetry of the states must be unambiguously specified with regard to axis convention. This excludes excitations involving the irreps B_1 and B_2 in C_{2v} and B_{1g}, B_{2g}, B_{3g}, B_{1u}, B_{2u}, B_{3u} in D_{2h}.

These criteria exclude any possibility of the calculated excitations being inaccurate due to unavailable symmetry or mislabeled symmetry. After eliminating data according to these criteria, we are left with a set of 207 excitations for testing the present approach (199 excitations from singlet states and eight from doublet states).

The 2 ^1A_g State of Ethylene. Special attention is given to the theoretical best estimate listed in QUESTDB for the 2 ^1A_g state of ethylene , which is characterized by Loos and and coworkers as a valence (π,π) $\rightarrow$ (π*,π*) double excitation at roughly 12.15 eV, referencing a 2004 study by Barbatti et al.[128,214] However, in the comprehensive 2014 study on the excited states of ethylene carried out by Feller et. al.,[215] the 2^1A_g state of ethylene is clearly characterized by both experiment and theory as a single (π,3p) Rydberg excitation at about 8.45 eV. Although we have been able to converge to the double excitation in the ^1A_g irrep described by Loos et. al. with some active space selections, it is clear that our best estimates converge to the lower Rydberg excitation supported by the Feller et al.[215] study. Thus, we have changed the theoretical best estimate of this excitation in the QUESTDB database to the value of 8.45 eV reported by Feller et al.[215]

4.4 Results

30-Excitation Tests. We first show the robustness of the new active space selection approach by carrying out calculations for a set of 30 excitations for which APC selection in the aug-cc-pVTZ basis[6] had a SA-CASSCF tPBE0 error greater than 0.55 eV; these 30 excitations involve a set of 17 molecules. We generated nine active spaces for each excitation by the method explained in section 4.3.2, using λ equal to 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, and 2.0 in step (ii).

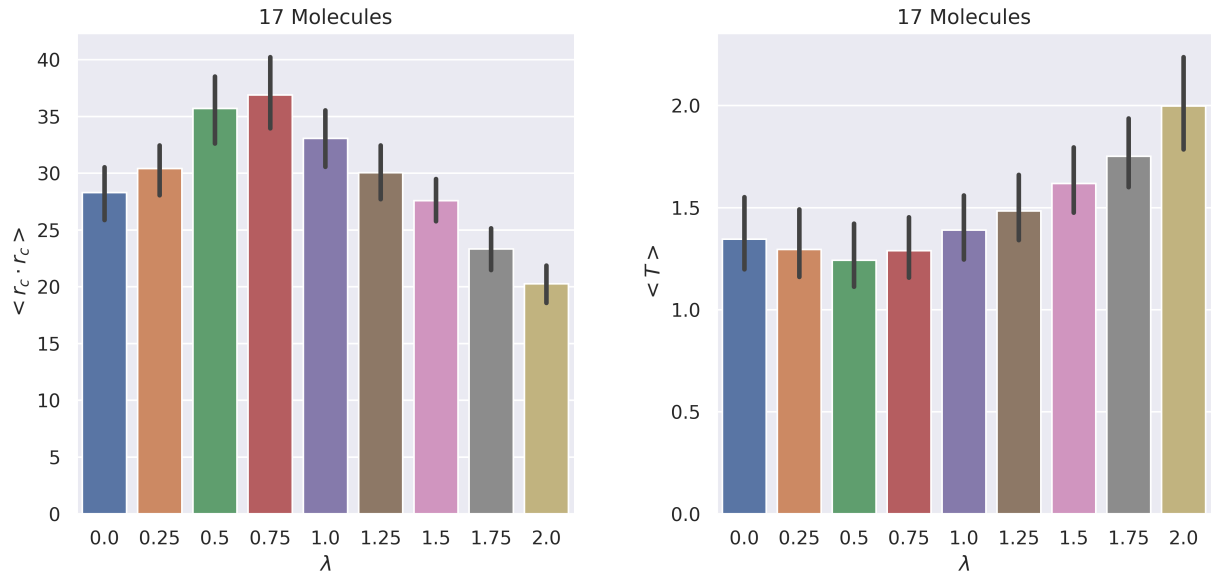

Figure 4.2: Left: Averaged squared distance from the centroid $< \mathbf{r}_c \cdot \mathbf{r}_c >$ over the selected active orbitals for the 17 unique molecules of the 30-excitation test set. Right: Averaged kinetic energy of the selected active orbitals for the 17 unique molecules of the 30-excitation test set.

Figure 4.2 shows trends in the orbital character of the selected active orbitals as λ is varied from 0 to 2 for the 17 molecules present in the 30-excitation test subset. The left side of Figure 4.2 shows the averaged squared distance from the centroid (with the centroid defined as the averaged coordinates of all nuclei in the molecule) over the selected active orbitals. The figure shows that different values of λ lead to significantly different averaged diffuse character of the selected orbitals, with the most diffuse character for $\lambda = 0.75$. The right side of Figure 4.2 shows the average kinetic energy of the selected orbitals, and illustrates the well-known quantum mechanical relation by which average kinetic energy in inversely related to average spatial extent. Thus, modifications of λ provide an effective means to explore active spaces targeting different kinds of states, e.g., Rydberg vs. valence excitations. This is demonstrated clearly in the calculation of the 2^1A_g state of ethylene, where $\lambda = 0.25$ selects an active space converging to the valence doubly excited ^1A_g of Loos and coworkers,[128] while $\lambda = 0.75$ converges to the lower-energy singly excited Rydberg state

(Supporting Information).[215]

We next examine the accuracy of excitation energies calculated from the selected active spaces by tPBE0. To prevent confusion, we stress that although the DVS-tPBE selection scheme employs the tPBE functional for variational selection, our calculations of excitation energies are based on tPBE0. These choices simply reflect that tPBE performs better in the selection scheme (as discussed below), whereas tPBE0 gives more accurate excitation energies (as shown in previous work[6] and discussed below.).

The left side of Figure 4.3 shows the absolute error of the tPBE0 calculations of the excitation energies with active spaces generated by the nine values of λ. As can be seen, no single value of λ yields accurate results for all 30 cases. For each value of λ, several excitations have an error greater than 1 eV. Although the mean absolute error is lowest for $\lambda = 1$ (0.56 eV), this is much larger than the mean absolute error of our previous benchmark results (0.19 eV) when we excluded poor active spaces. However, for all 30 excitations, the new scheme produces at least one value of λ that gives an absolute error less than 0.55 eV (the threshold for qualitative accuracy found in our previous benchmark).[6] This motivated the use of a variational scheme to find the best value of λ for each case.

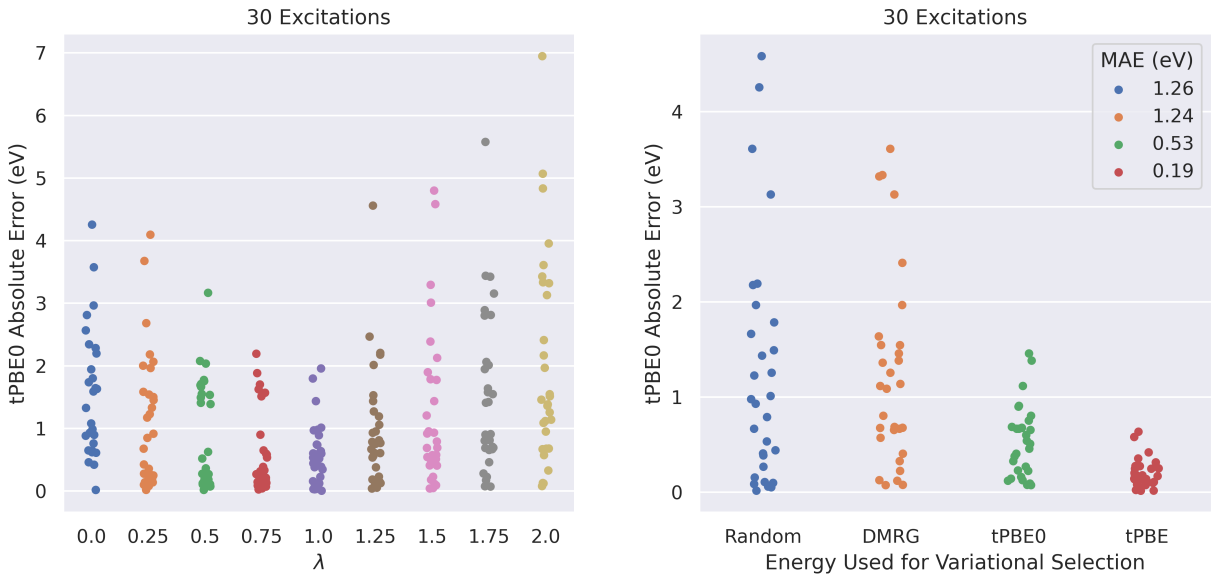

Figure 4.3: Left: tPBE0 absolute error of 30 difficult vertical excitations using active spaces selected with different values of λ. Right: Errors of these excitations with different values of λ selected variationally by different energies: random selection, variational selection with DMRG, tPBE0, and tPBE

As mentioned above, the criterion we use in DVS-tPBE is to choose the active space that gives the lowest sum of the tPBE absolute energy for the ground state and the excited state under investigation. To show the effectiveness of this approach, we compare this selection rule to three other schemes: variational selection using the summed DMRG energy, variational selection using the summed tPBE0 energy, and random selection. The right side of Figure 4.3 compares these approaches in choosing among the active spaces generated with different values of λ. The figure shows that tPBE distinguishes robustly between qualitatively accurate and inaccurate complete-active-space wave functions (i.e., there are no very large errors), while DMRG does little better than random chance, which does very poorly. The mean absolute error of the tPBE0 excitation energies is 1.24 eV with DMRG used for selection, as compared to 0.19 eV with tPBE used for selection. Furthermore, the maximum absolute error decreases from 3.61 eV with DMRG selection to 0.63 eV with tPBE selection. As a hybrid between tPBE and DMRG, selection with tPBE0 performs midway

between these two approaches.

We also examined other ways to try to select the best active space from the trial set, but none worked as well as the tPBE selection. For example, using as the sum of orbital entropies in the active space[31] or the sum of occupation number deviations from zero or 2 are unable to select well between the different values of λ (see the Supporting Information for details of these tests).

207-Excitation Tests. We next consider the performance of DVS-tPBE on the entire set of 207 excitations in the QUESTDB database that meet the selection criteria criteria of section 4.3.3. For this larger test, we used only four values of λ to generate active spaces: $\lambda = 0.25, 0.5, 0.75$, and 1.25. These values of λ were chosen based on their good performance on the 30-excitation tests (see the Supporting Information for more discussion of this point).

The left panel of Figure 4.4 shows the mean absolute errors achieved by DMRG, tPBE, and tPBE0 transition energy calculations with DVS-tPBE active-space selection for the full set of 207 excitations. These results are compared to our previous benchmark for the subset that excluded poor active spaces (those with with SA-CASSCF errors greater than 1.1 eV). As can be seen, errors for all three of these methods are as good as or exceed the performance of the previous benchmark. The comparison of mean unsigned errors is as follows:

- DMRG/CASSCF: 0.46 eV presently vs. 0.37 eV previously.
- tPBE: 0.20 eV presently vs. 0.21 eV previously.
- tPBE0: 0.17 eV presently vs. 0.18 eV previously.

We note that the performance using the wave function energy (DMRG/CASSCF) is slightly worse, as might have been expected due to the bias of the previous benchmark in excluding SA-CASSCF errors larger than 1.1 eV. However, we stress that here we achieved this comparable performance without excluding any cases, whereas previously the errors were only for the better active spaces. The results of our previous benchmark without excluding poor

active spaces are shown by the green bars in Figure 4.4; as can be seen inclusion of these active spaces significantly diminishes the performance of the method and returns a tPBE0 mean absolute error of 0.39 eV.

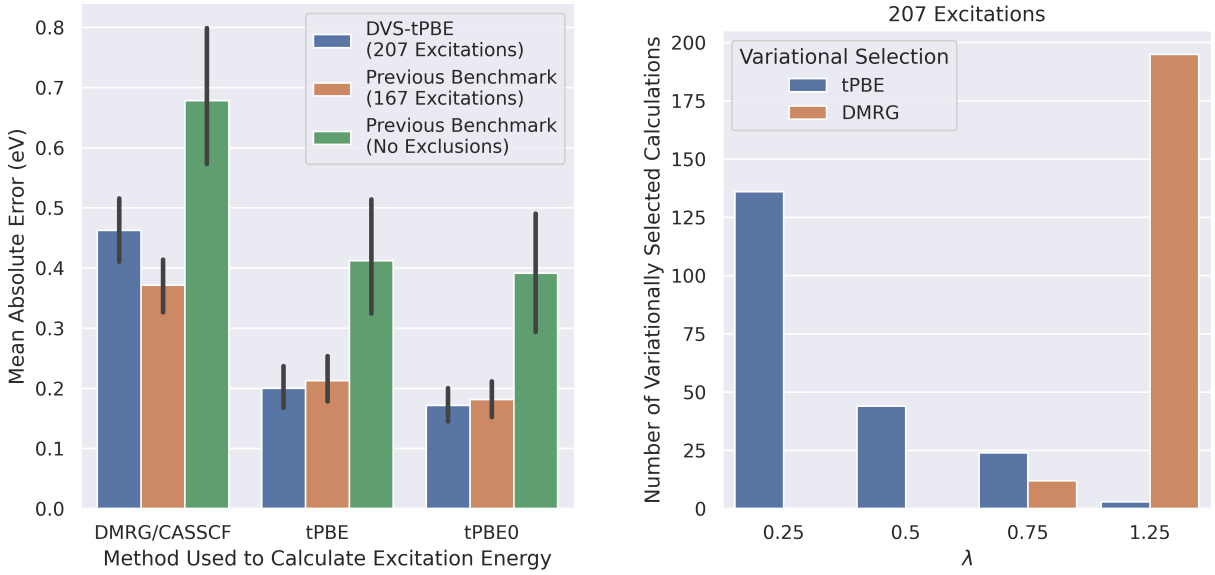

Figure 4.4: Left: Mean absolute errors achieved by DMRG/CASSCF, tPBE, and tPBE0 on the 207-excitation test set with active spaces selected by DVS-tPBE compared to the active spaces used in our previous benchmark, both before (no exclusions) and after (167 excitations) eliminating the poor active spaces. Right: Comparison of number of wave functions variationally selected with tPBE vs. number of wave functions variationally selected with DMRG at each value of λ.

The key to the success of tPBE selection as compared to DMRG selection seems to be that it chooses lower values of λ. The right panel of Figure 4.4 shows the frequency with which each value of λ was chosen in the tPBE selection compared to selection with DMRG. The figure shows that the frequency decreases quickly as a function of λ for tPBE selection. In contrast, this trend is reversed in the variational selection by DMRG, for which the selected values of λ are instead clustered heavily around $\lambda = 1.25$. The same trend toward preferring higher λ is also found in the tests on smaller 30-excitation dataset where we explored λ values as high as $\lambda = 2$. In that case we found that the selections by DMRG are clustered around $\lambda = 2$ (see Supporting Information).

The key to the success of tPBE selection as compared to DMRG selection seems to be that it chooses lower values of λ. The right panel of Figure 4.4 shows the frequency with which each value of λ was chosen in the tPBE selection compared to selection with DMRG. The figure shows that the frequency decreases quickly as a function of λ for tPBE selection. In contrast, this trend is reversed in the variational selection by DMRG, for which the selected values of λ are instead clustered heavily around $\lambda = 1.25$. The same trend toward preferring higher λ is also found in the tests on smaller 30-excitation dataset where we explored λ values as high as $\lambda = 2$. In that case we found that the selections by DMRG are clustered around $\lambda = 2$ (see Supporting Information).

We next evaluate the usefulness of variational selection with tPBE for the problem of comparing active spaces of vastly different sizes. To do this, we use the publicly available SA-CASSCF wave functions of our previous benchmark study,[6] but here not excluding any wave functions due to poor active spaces. We then use tPBE to variationally select among five active space results: the previous SA-CASSCF results (with active spaces of about 12 active orbitals) and the four new SA-DMRG results generated with the four values of λ (large active spaces with 40 active orbitals and $M = 700$). We label this broader selection scheme as DVS*-tPBE, and the left panel of Figure 4.5 shows the results of this scheme compared to simply using the previous SA-CASSCF wave function results (again, not excluding any active spaces due to poor selection). Although the performance of DVS*-tPBE is slightly reduced compared to DVS-tPBE (0.20 eV tPBE0 error vs. 0.17 eV), variational selection with tPBE is able to robustly discriminate against the outlier SA-CASSCF active spaces.

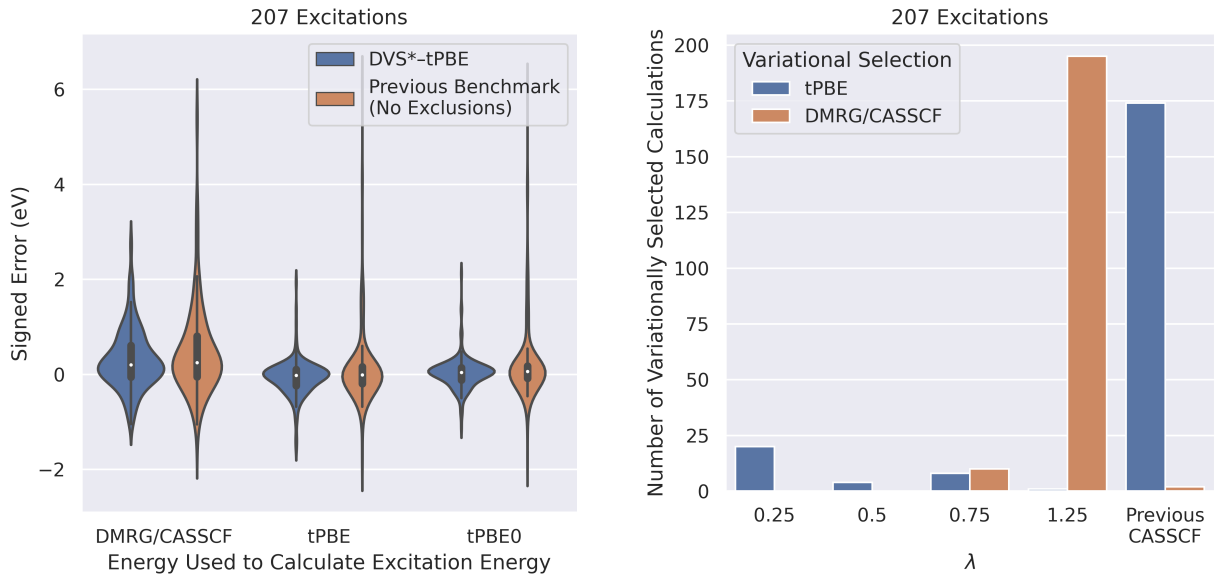

Figure 4.5: Left: Violin plots comparing the distribution of errors for three kinds of energy calculations (DMRG/CASSCF, tPBE, and tPBE0) on the 207-excitation test set when using tPBE to select among the previous pair of SA-CASSCF wave functions and the four pairs of SA-DMRG wave functions generated in this work to the the distribution of errors using just the SA-CASSCF wave functions on the same test set (not excluding poor active spaces). Right: Number of wave functions variationally selected from among the five trial wave functions by using tPBE to select or DMRG/CASSCF to select. Note that the selection among five trial pairs of wave functions is labeled in the plot as DVS*-tPBE, and the previously generated SA-CASSCF wave functions are labeled as "Previous Benchmark."

The right panel of Figure 4.5 again compares the distribution of wave functions variationally selected (among these five trial wave functions) by tPBE to those selected by DMRG/CASSCF. The figure shows that, although the active-space wave functions selected by DVS*-tPBE have significantly smaller mean absolute errors (0.20 vs. 0.39 eV), most of the wave functions variationally selected by tPBE come from the previous SA-CASSCF wave functions. Thus, variational selection with tPBE mainly improves the results by avoiding poor SA-CASSCF wave functions and replacing them with relatively good SA-DMRG wave functions. In contrast, variational selection with DMRG/CASSCF yields mainly wave functions generated with high values of λ and hardly any of the SA-CASSCF wave functions from our previous work.[6]

Although we have emphasized the excitation energies calculated by tPBE0, examination of the above results shows that tPBE excitation energies are – on average – only slightly worse. Another conclusion that can be drawn from the above comparisons is that tPBE and tPBE0 excitation energy calculations are not overly sensitive to the nature of the multiconfigurational wave functions. We obtain good results both with the selection among four active spaces for SA-DMRG calculations and with the selection among five trial active spaces, although in the latter case a DMRG active space is not usually the one chosen. Therefore, for the great majority of the excitations, we get good results with MC-PDFT and HMC-PDFT with quite different kinds of multiconfigurational wave functions.

Finally, we performed some tests to evaluate the sensitivity of DVS-tPBE to the number of active orbitals chosen. Keeping the bond dimension (700) and number of active spaces (4) fixed, we find that increasing the number of orbitals to 40 is the point at which our tPBE results start to replicate the accuracy of our previous study;[6] selecting 30 orbitals significantly decreases performance (Supporting Information). Thus, the success of the approach in avoiding the expensive step of orbital optimization is largely enabled by the large active spaces afforded by SA-DMRG.

4.5 Concluding Remarks

The goal of this work was to develop an automatized framework for selecting active spaces for calculating vertical excitation energies with useful predictive accuracy. Towards this goal, we have presented the discrete variational selection (DVS) approach to active space selection in which one generates multiple trial wave functions from a set of constructed active spaces and employs a variational selection scheme to choose the final result. To practically implement this approach for vertical excitation energies in the QUESTDB database, we have presented a scheme in which, for each excitation, (i) an RHF or ROHF wave function is calculated for the ground state, (ii) different sets of candidate orbitals are generated by diagonalization

of a parameterized operator, (iii) 40-orbital trial active spaces are chosen from these sets using APC selection, (iv) ground and excited-state SA-DMRG wave functions are calculated for each of these active spaces with a bond dimension of $M = 700$, and (v) the final result is chosen from among the resulting wave functions as the one that gives the lowest sum of absolute energies of the ground state and the excited state under consideration. We find that this approach performs poorly when using the DMRG/CASSCF absolute energies to select between wave functions, but robustly when using the absolute energy given by the translated tPBE functional of MC-PDFT (DVS-tPBE).

We have tested this method on 207 vertical excitations in the QUESTDB dataset (199 excitations from singlet states and eight from doublet states). When choosing between only four trial active spaces with no further orbital optimization, we are able to obtain equally as accurate tPBE0 results as in our previous benchmark[6] but now for all systems without the need to filter out poor active spaces. The success of this approach in avoiding the costly step of orbital optimization is largely enabled by the large active spaces afforded by SA-DMRG, and it is consistent with the recent perspective that "CASCI is not merely an approximation to CASSCF, in that it can be designed to have important qualitative advantages over CASSCF."[195] While the results in the article proper show that this approach is successful for systems in QUESTDB,[148] we show in Supporting Information that it can also have success in the transition metal system MnO_4^- with only minor modification (using larger N in the calculation of the APC entropies).[216]

Of course, application to different systems may require a greater number of orbitals and larger bond dimension, or a different approach entirely to constructing the candidate active space wave functions. Towards this end, we have shown that DVS-tPBE remains effective even when choosing between the large SA-DMRG active spaces of this work and the smaller SA-CASSCF active spaces of our previous benchmark.[6] That is, if we enlarge the trial set of active spaces to include both those from the SA-CASSCF calculations with small active

spaces and the new large active spaces (with 40 active orbitals), and we choose among them with variational selection by tPBE, we again obtain good results, even though we are now comparing quite different kinds of wave functions. These results show that DVS-tPBE is able to choose robustly between active spaces of vastly different sizes. This flexibility provides the basis for the further development of DVS-tPBE to applications of more metal-containing systems, extended organic systems, and adiabatic excitations.

In summary, we have proposed an approach for automatically selecting between active spaces for vertical excitations variationally through use of the tPBE energy from MC-PDFT. We have practically implemented this approach for the QUESTDB database through use of a parameterized operator to generate different active spaces and large SA-DMRG wave functions. Our results show that such an approach can potentially enable the application of CAS-based approaches in a high-throughput and predictive fashion. Although one cannot guarantee that any single active-space selection method will always work well, discrete variational selection with tPBE (DVS-tBPE) appears robust.

Converged density matrices of all 40-orbital DMRG calculations for the singlet and triplet QUESTDB excitations are available on Zenodo.[46] The code used for APC active space selection is now available in PySCF.

We thank Matthew Hennefarth and Matthew Hermes for helpful discussions. This work is supported by the National Science Foundation under grant CHE-2054723. We thank the Research Computing Center (RCC) at the University of Chicago for computational resources.

CHAPTER 5

MACHINE-LEARNED ENERGY FUNCTIONALS FOR MULTICONFIGURATIONAL WAVE FUNCTIONS

This chapter is reprinted with permissions from *J. Phys. Chem. Lett.* **2021**, *12*, 32, 7761–7767

5.1 Abstract

We introduce multiconfiguration data-driven functional methods (MC-DDFMs), a group of methods which aim to correct the total or classical energy of a qualitatively accurate multiconfigurational wave function using a machine-learned functional of some featurization of the wave function such as its density, on-top density, or both. On a dataset of carbene singlet-triplet energy splittings, we show that MC-DDFMs are able to achieve near-benchmark performance on systems not used for training with a robust degree of active-space independence. Beyond demonstrating that the density and on-top density hold the information necessary to correct the singlet-triplet energy splittings of multiconfigurational wave functions, this approach shows great promise for the development of functionals for MC-PDFT because corrections to the classical energy appear to be more transferable to types of molecules not included in the training data than corrections to total energies such as yielded by CASSCF or NEVPT2.

5.2 Results and Discussion

Although current Kohn-Sham density functional theory (KS-DFT) is highly accurate for many interesting chemical systems, it is well-known to be less accurate for strongly correlated systems than for systems well-described by a single Slater determinant.[59,61,62,217–220] This has motivated interest in combining density functionals with multiconfigurational wave

function methods[26,221–223] (e.g., CASSCF) that explicitly express the wave function as a superposition of electronic configurations. However, because multiconfigurational wave function methods are generally limited to a set of configurations that is too small to yield quantitatively accurate correlation energies, one must augment them by a post-MCSCF procedure in order to obtain quantitative accuracy. The most widely used of these methods include multireference perturbation theory (MRPT)[26,32,35,224] (e.g., CASPT2 and NEVPT2) and multireference configuration interaction (MRCI),[225,226] which are both very expensive.

As an alternative to MRPT and MRCI, we have proposed multiconfiguration pair-density functional theory (MC-PDFT)[41] and multiconfiguration density-coherence functional theory (MC-DCFT).[139] These methods share the feature that they compute an energy by combining wave function theory for the classical components (kinetic energy, electron-nuclear attraction, and classical electron-electron interactions) with a functional for the nonclassical components of the energy (exchange and correlation), and together they may be grouped as examples of multiconfigurational nonclassical functional theory (MC-NCFT). The general MC-NCFT energy expression is given by:

$$E_{\text{MC-NCFT}}[\psi^{\text{MC}}] = E_{\text{class}}^{\text{MC}} + E_{\text{nc}}[f[\psi^{\text{MC}}]] \tag{5.1}$$

where the classical energy $E_{\text{class}}^{\text{MC}}$ accounts for nucleus-nucleus repulsion, nucleus-electron attraction, classical electron-electron repulsion, and electron kinetic energy, and E_{nc}^{MC} is a nonclassical functional (NCF) dependent on a featurization f of the reference wave function ψ^{MC}, which may be the density, on-top density, density coherence, gradients of these quantities, or any other featurization of the wave function.

Inspired by both the success of these methods and recent work that has used neural networks to develop density functionals for KS-DFT,[227–259] we introduce a broader class of methods named "multiconfiguration data-driven functional methods" (MC-DDFMs) which aim to correct the classical or total energy E_{ref} of a multiconfigurational wave function

method through the use of a machine-learned functional E_{ML}:

$$E_{\text{MC-DDFM}}[\psi^{\mathrm{MC}}] = E_{\mathrm{ref}}^{\mathrm{MC}} + E_{\mathrm{ML}}[f[\psi^{\mathrm{MC}}]] \tag{5.2}$$

in which E_{ML} plays the generalized role of E_{nc}. In this work we introduce four new MC-DDFMs which use functionals of ρ^{MC} and Π^{MC} trained to correct four different reference energies E_{ref}:

1) Data driven functional '21 (DDF21), a MC-NCFT functional trained to correct the classical energy:

$$E_{\mathrm{DDF21}} = E_{\mathrm{class}}^{\mathrm{MC}} + E_{\mathrm{DDF21}}[\rho^{\mathrm{MC}}, \Pi^{\mathrm{MC}}] \tag{5.3}$$

2) ΔtPBE-21, a functional trained to correct the translated PBE (tPBE) energy of MC-PDFT:[41]

$$E_{\Delta\text{tPBE-21}} = E_{\mathrm{tPBE}}^{\mathrm{MC}} + E_{\Delta\text{tPBE-21}}[\rho^{\mathrm{MC}}, \Pi^{\mathrm{MC}}] \tag{5.4}$$

3) ΔCASSCF-21, a functional trained to correct the CASSCF energy:

$$E_{\Delta\text{CASSCF-21}} = E_{\mathrm{CASSCF}}^{\mathrm{MC}} + E_{\Delta\text{tPBE-21}}[\rho^{\mathrm{MC}}, \Pi^{\mathrm{MC}}] \tag{5.5}$$

4) ΔNEVPT2-21, a functional trained to correct the NEVPT2 energy:

$$E_{\Delta\text{NEVPT2-21}} = E_{\mathrm{NEVPT2}}^{\mathrm{MC}} + E_{\Delta\text{NEVPT2-21}}[\rho^{\mathrm{MC}}, \Pi^{\mathrm{MC}}] \tag{5.6}$$

Below, we present the development of these MC-DDFMs as well as three different tests of their generalization to molecules outside of the training set: (i) test data similar to training data; (ii) test data using other active spaces; and (iii) test data using aryl and biradical systems. These results are encouraging, and we believe that further progress in this direc-

tion – particularly towards designing new functionals for MC-NCFT – has the potential to systematically achieve low-cost quantitative accuracy for a variety of different wave function methods.

Training Geometries. We have taken our training geometries from the QMSpin database of Schwilk et al.,[260] which contains carbenes optimized in the singlet state using CASSCF(2,2)/cc-pVDZ-F12 as well as benchmark-quality vertical singlet-triplet splittings obtained using explicitly correlated multireference configuration interaction with single and double excitations and the Davidson quadruples correction (MRCISD-F12+Q).[261–264] In this work we have used a subset of these carbenes that contain only carbon and hydrogen atoms.

Network Architecture. We have taken an approach very similar to the recent work of Dick and Fernandez-Serra in their development of NeuralXC.[253] Atomic feature vectors for atoms I are obtained by projecting the density ρ^{MC} and on-top density Π^{MC} onto atom-centered basis functions ϕ_{nlm} via quadrature:

$$c^{I,\rho}_{nlm} = \int_{\mathbf{r}} \phi^{I}_{nlm}(\mathbf{r})\rho^{\mathrm{MC}}(\mathbf{r}) \qquad c^{I,\Pi}_{nlm} = \int_{\mathbf{r}} \phi^{I}_{nlm}(\mathbf{r})\Pi^{\mathrm{MC}}(\mathbf{r}) \tag{5.7}$$

and these features are then made rotationally invariant by the transformations:[253,259]

$$d^{I,\rho}_{nl} = \sum_{m} (c^{I,\rho}_{nlm})^2 \qquad d^{I,\Pi}_{nl} = \sum_{m} (c^{I,\rho}_{nlm})^2 \tag{5.8}$$

In this work we used the 108 optimized basis functions developed by Chen et. al. for featurization on each atom;[265] this results in a total of 36 rotationally invariant features for each atom I and density ζ (ρ or Π): 12 "s" features ($l = 0$, $d^{I,\zeta}_{1,0} \dots d^{I,\zeta}_{12,0}$), 12 "p" features ($l = 1$, $d^{I,\zeta}_{2,1} \dots d^{I,\zeta}_{13,1}$), and 12 "d" features ($l = 2$, $d^{I,\zeta}_{3,2} \dots d^{I,\zeta}_{14,2}$). We then input each atomic feature vector $v_I = \{d^{I,\rho}_{nl}, d^{I,\Pi}_{nl}\}$ into its respective element network, f_{λ_I} to obtain the total energy correction:

$$E = \sum_I f_{\lambda_I}(v_I) \tag{5.9}$$

as in the work of Behler and Parrinello.[266]

Networks were implemented and developed in PyTorch[267] from the starting point of NeuralXC available on GitHub.[268] Element networks consist of an input layer, n_{layers} fully connected hidden layers each with n_{nodes}, and a one-node output layer, with n_{layers} and n_{nodes} treated as hyperparameters. The GELU activation function[269] was used for all nodes. Although overfitting is a concern given the large amount of features used, scores on the test set are not overly large compared to the training set, differing at most by 0.03 eV (see Supporting Information).

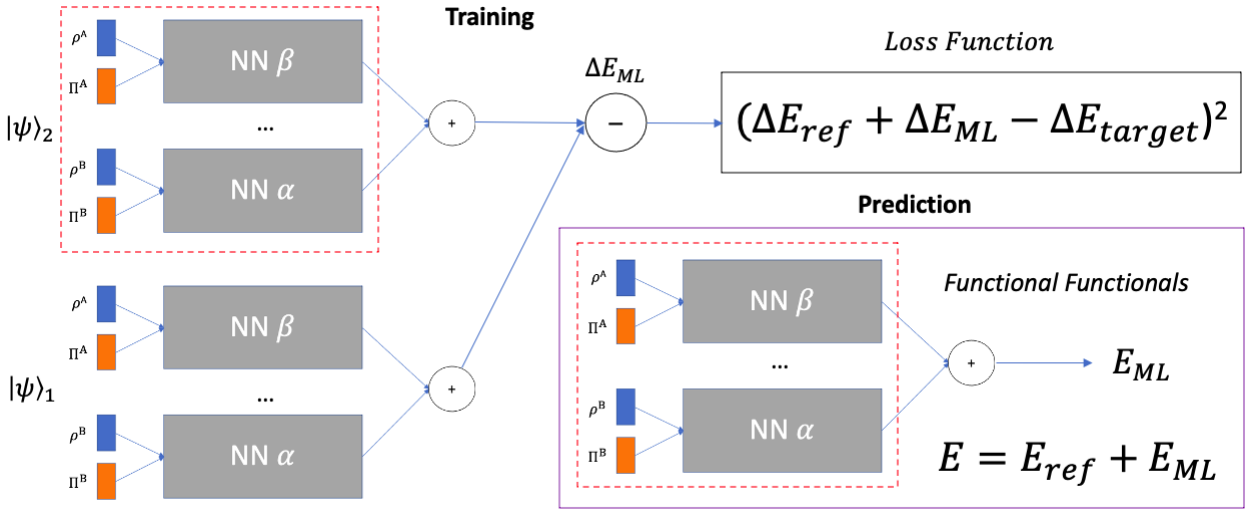

Figure 5.1: Network training scheme. Given a starting reference energy E_{ref} with output ΔE_{ref}, the element networks $\{\alpha, \beta, \ldots\}$ are regressed to minimize the mean squared deviation between corrected energy differences $\Delta E_{\text{ref}} + \Delta E_{\text{ML}}$ and the target energy difference ΔE_{target}.

Network Training. We focus our non-classical functionals on predicting benchmark-quality energy differences between two states $|\psi\rangle_1$ and $|\psi\rangle_2$, in particular the singlet and triplet state of a single geometry. Given a difference in energy between these states from a reference method, ΔE_{ref}, we train functionals to minimize the mean squared deviation

between the corrected energy difference, $\Delta E_{\text{ref}} + \Delta E_{\text{ML}}$, and a target energy difference, ΔE_{target} (in this work, singlet-triplet energy splittings from MRCISD-F12+Q); this training scheme is outlined in Figure 5.1. This centering of the loss function solely on relative energies stands in contrast to previous work in NeuralXC,[253] DeepKS,[258] OrbNet,[270] and KDFA,[259] but it has three advantages: (i) it allows benchmark results to be obtained from a variety of different sources (including experiment, which almost always yields relative energies); (ii) relative energies are the quantities of most interest to chemists, since bond energies, energies of reaction, and barrier heights are all relative energies; and (iii) theoretical data used for training is almost always more accurate for relative energies than for absolute energies.

For optimization of parameters and hyperparameters, the 360 carbenes were split into a training set of 287 carbenes, a validation set of 37 carbenes, and a test set of 36 carbenes. All features were normalized using a StandardScaler fit on the training set,[271] and networks were optimized to reduce the mean squared error loss over the entire training set in Pytorch using the Adam optimizer[272] with a learning rate of 0.01 for a maximum of 20001 steps. A PyTorch scheduler (torch.optim.lr_scheduler.ReduceLROnPlateau) was used to decrease the learning rate over time upon an observed plateau in the loss to a minimum learning rate of 1.1e-7, after which the training was stopped early. The hyperparameters considered were the weight decay of the Adam optimizer and the number of nodes and layers in the element networks, and these hyperparameters were optimized using Optuna[273] by minimizing loss on the validation set. The final hyperparameters of all networks and the ranges explored are given in the Supporting Information.

Wave Function Generation. State-averaged (2,2)-CASSCF wave functions, along with tPBE and NEVPT2 energies for the singlet and triplet states of each carbene, were obtained using PySCF,[90] as integrated with MC-PDFT capabilities using publicly available development code.[274] Atomic feature vector inputs (eq 5.7) were obtained via quadrature using the highest grid quality (grid_level=9). During development it was found that these input

features converge at significantly lower thresholds than the CASSCF energy, and therefore more stringent CASSCF optimization parameters were used in obtaining the singlet and triplet wave functions to insure consistency (mc.conv_tol = 1e-10, mc.conv_tol_grad = 1e-6, mc.ah_lindep = 1e-14, and mc.ah_conv_tol = 1e-12).

Active Space Selection. With the exception of benzene, all active spaces for CASSCF calculations were chosen automatically using the ranked-orbital approach.[5] The highest 23 doubly occupied orbitals and the lowest 23 virtual orbitals of an ROHF wave function were individually Boys-localized[94] and the approximate pair coefficient (APC) method[5] was employed on all doubly occupied orbitals and the localized virtual orbitals to approximate orbital entropies (the remaining virtual orbitals were not considered for the active space). These entropies were then used to rank the orbitals in terms of importance, and the final active space was selected by setting a maximum number of allowed CSFs in the wave function expansion (e.g., max(2,2), max(4,4), and max(6,7)) and dropping orbitals from the active space until the size of the active space satisfied the threshold. In the training data we selected all active spaces at the max(6,7) level.

Active Space Error. Although the ranked-orbital approach above is imperfect at ranking orbitals in importance for the active space, at the max(6,7) level our method failed to select active spaces with qualitatively accurate CASSCF excitation energies (<1 eV in absolute error) in only a small number of cases; these cases were rejected from the training, validation, and test sets. However, in addition to the calculations at the max(6,7) level that were used to train the functionals, we performed some tests with minimal active spaces generated at the max(2,2) level, which requires a perfect ranking of the orbitals; in these tests we experienced a much higher failure rate (33%), and therefore these tests were carried out on a test subset of only 24 carbenes (listed in the Supporting Information).

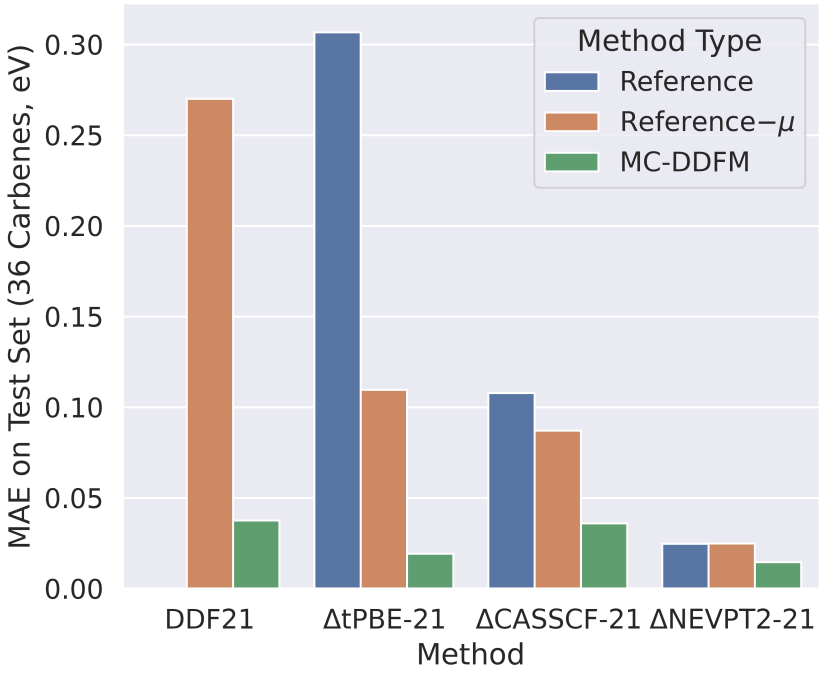

Figure 5.2: Mean absolute errors (MAEs) on MRCISD-F12+Q benchmark data for a test set of 36 carbenes excluded from the training data. For each MC-DDFM (DDF21, ΔtPBE-21, ΔCASSCF-21, and ΔNEVPT2-21, shown in green), we show the performance of its reference method (tPBE, CASSCF, and NEVPT2, shown in blue) as well as a one-parameter mean-corrected method (Reference-μ) shown in orange. The MAE of the CASSCF classical energy (1.1eV) is not shown due to scale.

Results. Figure 5.2 shows the performance of the four MC-DDFMs in comparison to their respective reference methods on the test set of 36 carbene singlet-triplet energy splittings. For comparison, we also show the performance of a simple one-parameter mean correction to the singlet-triplet energy splittings, in which ΔE_{ref} is corrected by its mean deviation from MRCISD-F12+Q on the training data. Encouragingly, all four functionals are able to greatly improve upon these one-parameter corrections, surpassing the mean absolute errors (MAEs) of their reference methods by factors of 29 (DDF21), 16 (ΔtPBE21), 3 (ΔCASSCF-21), and 2 (ΔNEVPT2-21). Additionally, although all functionals presented in the article proper depend on both the density and on-top density, additional results given in the Supporting Information show that we obtain similarly high accuracy using only density features or only

on-top density features.

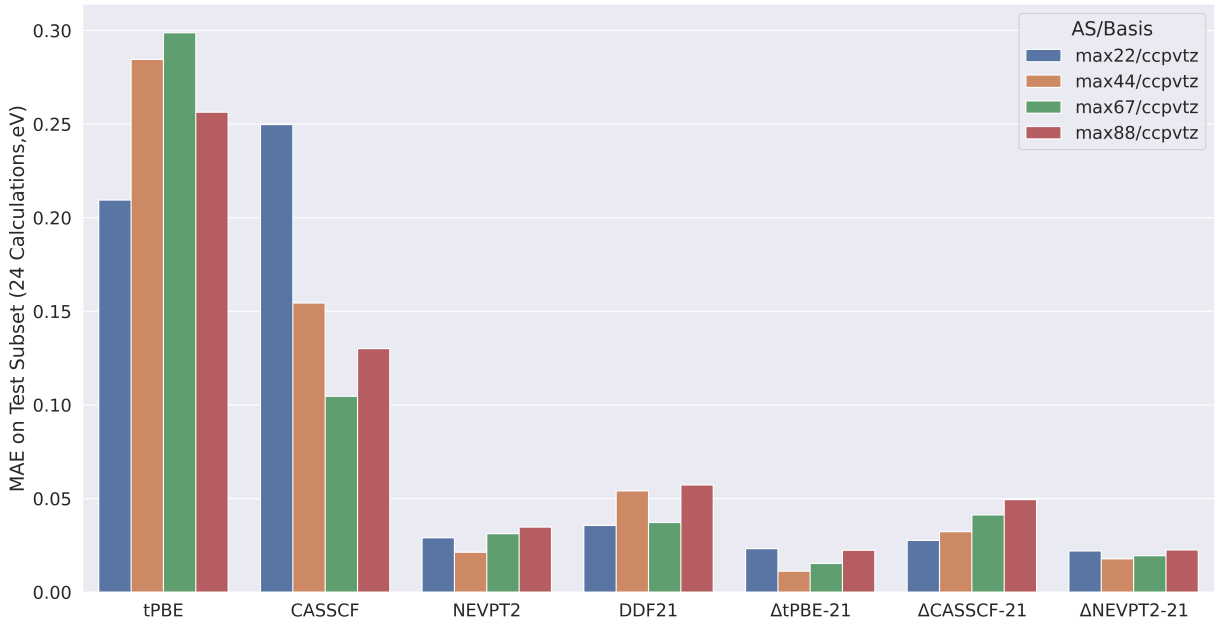

Figure 5.3: Mean absolute errors on MRCISD-F12+Q benchmark data from a test subset of 24 carbenes for which our automated scheme chose a reasonable (2,2) active space, tested with the cc-PVTZ basis at four different active space sizes: max(2,2), max(4,4), max(6,7), and max(8,8).

We tested the active space dependence of our data-driven functional methods on 24 carbenes with four different active space sizes whose number of configurations vary by four orders of magnitude: max(2,2), max(4,4), max(6,7), and max(8,8). Figure 3 shows that all MC-DDFMs maintain their near-benchmark accuracy across this wide range of active spaces, despite being trained on only max(6,7) active spaces. We note that this active space robustness is likely a result of the sole dependence of our loss function on relative energies rather than absolute ones. However, we find that one drawback of our approach is that the parameters do not seem to be easily transferable to other basis sets; when switching to either a cc-pVDZ or cc-pVQZ basis the errors of the MC-DDFMs tend to increase dramatically (Supporting Information).

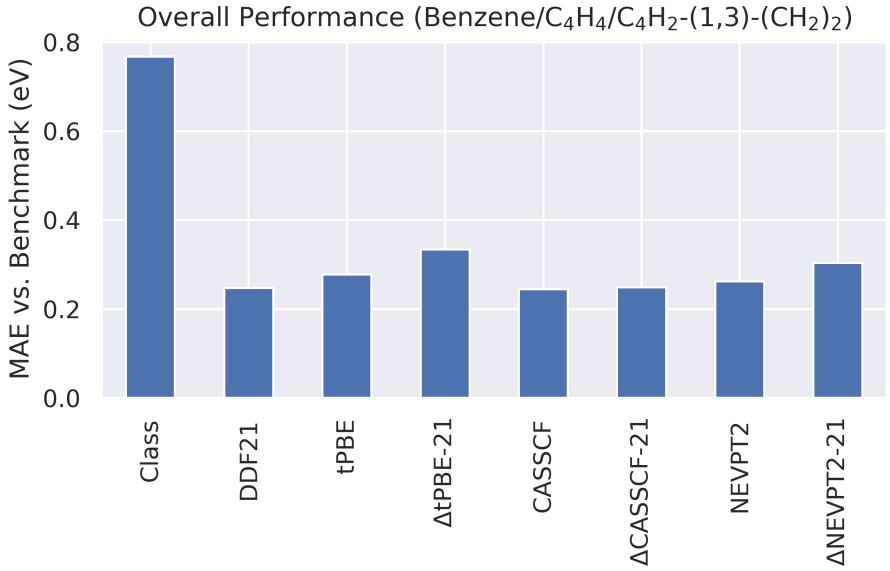

Figure 5.4: Mean absolute error of reference and data-driven functional methods on three difficult singlet triplet energy splittings, consisting of one aryl system (C_6H_6, using the standard minimal cc-pVTZ@UNO-(6,6) active space[7]) and two biradical systems (cyclobutadiene, C_4H_4, and 1,3-bis(methylene)-cyclobutadiene (C_4H_2-(1,3)-$(CH_2)_2$), using automatically selected max(10,10) active spaces).

As a final test of generalizability, we tested the MC-DDFMs on three difficult singlet–triplet energy splittings quite different than any data in the training set: benzene and two biradical systems; cyclobutadiene (C_4H_4) and 1,3-bis(methylene)cyclobutadiene (C_4H_2-(1,3)-$(CH_2)_2$) (Figure 5.4). These systems were taken from previous benchmark studies on translated functionals,[275,276] with benchmarks for benzene taken from experiment[277] and benchmarks for the biradicals from theoretical results.[278] While MC-DDFMs correcting total energies (ΔtPBE-21, ΔCASSCF-21, and ΔCASSCF-21) all performed worse on average than their respective reference methods, DDF21 maintains a large improvement upon the CASSCF classical energy, reducing its MAE from 0.77 eV to only 0.25 eV. This suggests that corrections to the classical energy – as is done in MC-NCFT – may be more transferable to types of molecules not included in the training data than corrections to "complete" methods such as CASSCF or NEVPT2. Similar generalizability in this regard is achieved

by MC-DDFMs trained solely on the density or on-top density (Supporting Information).

Related Work. This work builds on the development of machine-learned density functionals for KS-DFT,[227–259] in addition to machine-learned density or density matrix functionals developed to correct the energy of semiempirical methods (OrbNet)[270] or Hartree-Fock (MOB-ML).[88,107,279] In addition, machine learning has also been used to predict the density itself[280–284] and even the on-top pair density;[285] in principle these methods could be combined with the energy functionals we have presented here to predict the energy directly from a molecular geometry. Other related work is the application of machine learning or theory to predict multireference character, which could help to identify when multiconfigurational methods should be used.[286–288] In addition, several previous studies have attempted to predict singlet-triplet energy splittings directly from the molecular geometry, often achieving high accuracy with respect to their benchmark data.[65,289–292] Given these works, there is no reason to think that the prediction of singlet-triplet energy splittings could not be achieved by a much simpler method, but we emphasize that the central contribution of this work is not in predicting accurate singlet-triplet energy splittings at low cost but in demonstrating the potential for data-driven improvement in MC-NCFT.

Conclusions. We have presented a data-driven approach to the development of energy functionals for multiconfigurational wave functions utilizing neural networks parametrized in terms of the density and on-top density. Using a dataset of carbene singlet-triplet energy splittings taken from the QMSpin database,[260] we find that the new multiconfiguraional data-driven functional methods (MC-DDFMs) are able to achieve benchmark-quality accuracy on carbenes not included in the training set and improve markedly on approaches using translated MC-PDFT functionals even when extended to different active spaces. Beyond demonstrating that the density and on-top density hold the information necessary to correct the singlet-triplet energy splittings of multiconfigurational wave functions, this approach shows great promise for multiconfigurational nonclassical functional theory, because

corrections to the classical energy appear to be more transferable to types of molecules not included in the training data than corrections to total energies such as yielded by CASSCF or NEVPT2. It will be interesting to see if this good performance can be maintained when the functionals are parameterized using larger and more diverse sets of training data.

CHAPTER 6

DIVERGENT BIMETALLIC MECHANISMS IN COPPER(II)-MEDIATED C–C, N–N, AND O–O OXIDATIVE COUPLING REACTIONS

This chapter is reprinted with permissions from *J. Am. Chem. Soc.* **2024**, *146*, 5, 3521–3530

6.1 Abstract

Copper-catalyzed aerobic oxidative coupling of diaryl imines provides a route for conversion of ammonia to hydrazine. The present study uses experimental and density functional theory computational methods to investigate the mechanism of N–N bond formation, and the data support a mechanism involving bimolecular coupling of Cu-coordinated iminyl radicals. Computational analysis is extended to Cu^{II}-mediated C–C, N–N, and O–O coupling reactions involved in the formation of cyanogen (NC–CN) from HCN, 1,3-butadiyne from ethyne (i.e., Glaser coupling), hydrazine from ammonia, and hydrogen peroxide from water. The results reveal two different mechanistic pathways. Heteroatom ligands with an uncoordinated lone pair (iminyl, NH_2, OH) undergo charge transfer to Cu^{II}, generating ligand-centered radicals that undergo facile bimolecular radical-radical coupling. Ligands lacking a lone pair (CN and CCH) form bridged binuclear diamond-core structures that undergo C–C coupling. This mechanistic bifurcation is rationalized by analysis of spin densities in key intermediates and transition states, as well as multiconfigurational calculations. Radical-radical coupling is especially favorable for N–N coupling owing to energetically favorable charge transfer in the intermediate and thermodynamically favorable product formation.

6.2 Introduction

Copper-mediated oxidative coupling of organic molecules was discovered more than 150 years ago, when Glaser demonstrated the oxidative coupling of phenylacetylene (Figure 6.1A).[293–295] Cu-catalyzed reactions of this type continue to be the focus of extensive study,[296–299] and many are compatible with molecular oxygen as a stoichiometric oxidant. Despite the extensive history of these reactions, their mechanisms remain poorly understood. Complications arise from the likely participation of more than one copper species, the potential accessibility of three formal oxidation states (Cu^{I}, Cu^{II}, and Cu^{III}), and the possibility of both one- and two-electron redox steps in the mechanism.[300–302] The mechanism of the Glaser coupling has been the subject of extensive investigation,[295] and several studies[303–305] favor a binuclear reductive-elimination pathway involving a dimeric Cu^{II} intermediate with bridging acetylides (Figure 6.1B), similar to that originally proposed by Bohlmann.[306,307] Recent studies by some of us have been exploring Cu-catalyzed N–N coupling reactions that share many features in common with the Glaser coupling.[308–310] Both reaction classes feature Cu catalysts with imine ligands (i.e., pyridine or diaryl imine) and promote efficient oxidative homocoupling with O_2 as the terminal oxidant. The oxidative coupling of diaryl imines to generate an azine product was introduced by Hayashi and coworkers as a key step in the production of hydrazine (Figure 6.1C).[311,312] The catalytic rate law for this reaction features a second-order dependence on [Cu],[309] which led to a mechanistic proposal for an iminyl-bridged binuclear Cu intermediate that resembles the Glaser coupling mechanism (Figure 6.1D). The present study was initiated to probe this mechanistic pathway and the mechanistic relationship between Cu-catalyzed N–N and C–C coupling reactions.

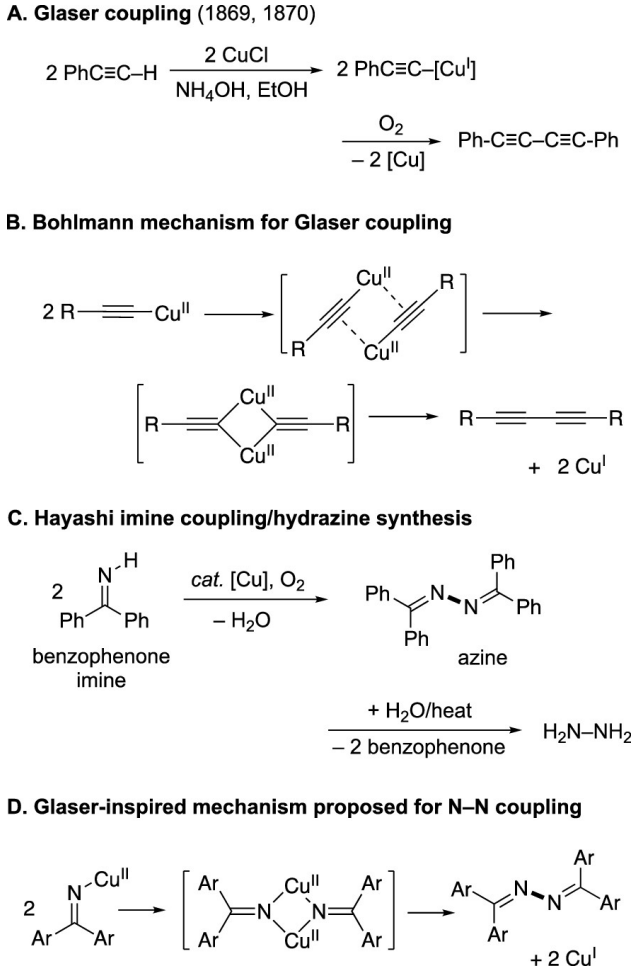

Figure 6.1: Prominent Cu-catalyzed aerobic oxidative coupling reactions and their proposed mechanisms: Glaser homocoupling of alkynes (A) and proposed binuclear mechanism (B); Hayashi homocoupling of diarylimines (C) and proposed binuclear mechanism (D).

The Glaser and Hayashi coupling reactions are prominent examples of a broader series of related Cu-catalyzed oxidative reactions (Figure 6.2). Cu-mediated reversible oxygen-oxygen bond cleavage/formation via diamond-core structures is well established in O_2 activation enzymes and model systems;[313–316] however, these processes involve an oxo-bridged dimer with a formal Cu^{III} redox state. The diamond-core intermediates in Figures 1B and 1D are formally assigned as Cu^{II} species. The latter proposal is consistent with observations that Cu^{II} can mediate these reactions in the absence of a secondary oxidant, and the observation that electrochemical N–N coupling operates at potentials close to the copper(II/I) potential.

Thus, despite geometric similarities, the oxidative homocoupling reactions considered here are electronically distinct (i.e., Cu(II) vs. Cu(III)) from Cu/O_2 reactions that proceed via a $Cu_2^{III}(\mu\text{-}O)_2$ intermediate.

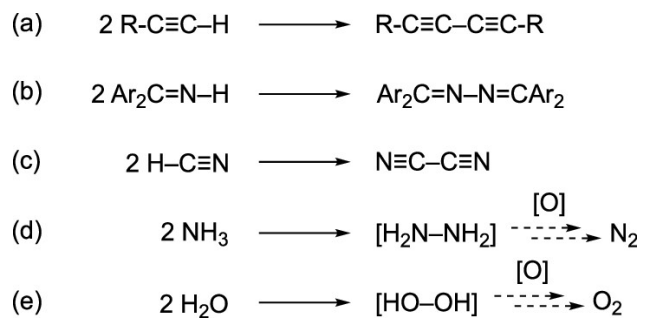

Figure 6.2: Complementary Cu-catalyzed oxidative homocoupling reactions of alkynes (a), diarylimines (b), hydrogen cyanide (c), ammonia (d), and water (e), each of which is considered in the present study.

Previous results obtained from Cu-catalyzed aerobic oxidative coupling of diaryl imines[309,310] are supplemented here by additional structural, electrochemical, and stoichiometric reactivity studies of diarylimine-Cu species. These data provide experimental benchmarks for density functional theory (DFT) computational analysis of the N–N coupling mechanism. The computational results show that the previously proposed N–N coupling pathway, featuring a binuclear diamond-core intermediate (cf. Figure 6.1D), is energetically prohibitive. This outcome contrasts with the facile C–C coupling via the analogous alkynyl-bridged binuclear intermediate (cf. Figure 6.1B). A lower-energy pathway for N–N coupling is identified that involves bimolecular radical-radical coupling between two iminyl-Cu complexes. Electronic-structural analysis of this pathway shows that deprotonation of a coordinated imine ligand triggers nitrogen-to-Cu^{II} charge transfer, forming a species that has significant Cu^{I}-iminyl radical character. These two mechanistic pathways – (1) binuclear reductive-elimination via a bridged diamond-core intermediate and (2) bimolecular radical-radical coupling – have been analyzed for each of the C–C, N–N, and O–O coupling reactions in Figure 6.2. The results reveal why C–C coupling favors binuclear reductive elimination, while heteroatom-

heteroatom coupling favors radical-coupling.

6.3 Results and Discussion

Experimental and computational benchmarks. Experimental studies employed a diaryl imine with two *p*-fluoro substituents, 4,4'-difluorobenzophenone imine (**1**), to facilitate product analysis by ^{19}F NMR spectroscopy. Stoichiometric reactions of **1** with $Cu(OTf)_2$ have been reported previously to form the N–N coupled azine product **2** at room temperature in *N,N*-dimethylformamide.[309] Similar reactivity was observed here in acetonitrile (MeCN) (Figure 6.3A), complicating efforts to characterize Cu^{II}/diaryl imine complexes. This reactivity could be slowed significantly at lower temperature, however. Addition of other solvents to this solution, including CH_2Cl_2, toluene, THF, and EtOAc, led to formation of a brown solid, with orange crystals also observed when adding CH_2Cl_2 and toluene. Crystals suitable for X-ray diffraction analysis were obtained by dissolving $Cu(OTf)_2$ and 4 equiv of **1** in MeCN, and layering over CH_2Cl_2 at –45 °C. The X-ray crystal structure revealed the formation of $[Cu(\mathbf{1})_4](OTf)_2$ (**3**) (Figure 6.3B and 6.3C). Consistent with the stoichiometric reactivity of **1** and $Cu(OTf)_2$ noted above, this complex is not stable when dissolved in MeCN at room temperature, but undergoes spontaneous reaction to generate azine **2**.

The X-ray crystal structure of **3** was used as a benchmark for DFT computational studies. DFT geometries, frequencies, and energies were calculated using unrestricted Kohn-Sham DFT using the M06 functional[21] with ultrafine grid quality in Gaussian16,[317] using the def2-TZVPP basis for transition metals and the def2-SVP basis for all other atoms.[318,319] The polarizable continuum model (PCM) was used to model the MeCN solvent (see Supporting Information for additional details). A survey of 14 structural variations of Cu^{II} complexes were evaluated by these methods. The structures differ in the number of imine and acetonitrile ligands (3–5 each, see the Supporting Information for details). The most stable structure has four imine and no acetonitrile ligands, closely resembling the structure

obtained from X-ray diffraction analysis of **3** (Figure 6.3C). Good agreement is observed between the 4-coordinate complex predicted by theory and the experimental crystal structure. For example, the experimental and computed Cu-N bond distance is 1.99 ± 0.01 Å in both cases.

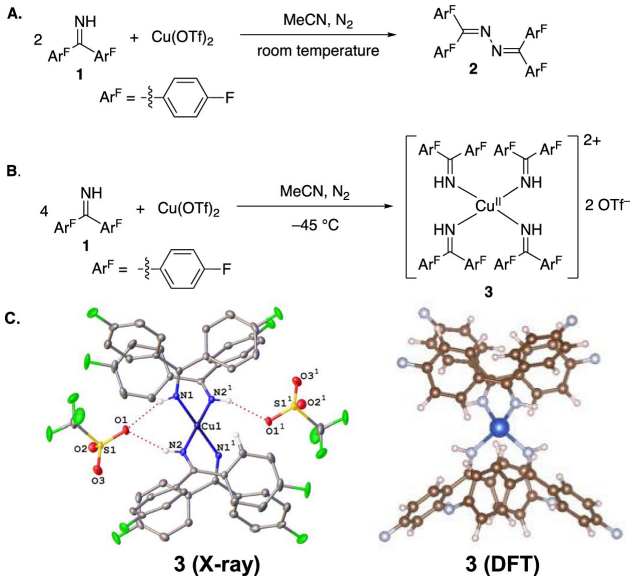

Figure 6.3: A) Cu^{II}-mediated imine coupling proceeds spontaneously at room temperature. B) Low-temperature preparation of the tetraimine complex **3**. C) Experimental and computational structure of $[Cu^{II}(\mathbf{1})_4]^{2+}$ (left and right, respectively). The symmetry code for the experimental structure is 1:1-X,1-Y,-Z. The experimental crystal structure includes two triflate counterions (non-imine hydrogens are not shown for clarity), while the DFT optimization of **3** was performed without the two triflate counterions.

Cyclic voltammetry analysis was performed with $[Cu^{I}(MeCN)_4]PF_6$ in the presence of a series of diaryl imines with different substituents in the 4,4'-positions: X = MeO, H, F, Cl, CF3 , and the $Cu^{II/I}$ redox process is evident as a quasi-reversible wave in acetonitrile (Figure 6.4A). These results were complemented by computational analysis of the reduction potential of the different $[Cu^{II}(imine)_4]^{2+}$ species by evaluating the free energy of electron transfer from ferrocene (Fc) to CuII (see the Supporting Information for details). The experimental and computed redox potentials show good agreement (Figure 4B), with the absolute reduction potential for $[Cu(\mathbf{1})_4](OTf)_2$ (3) differing by only 0.06 eV. These differences between

experimental and computational values are within the range expected from DFT calculations and likely incorporates experimental solvation effects not captured fully by the PCM method.

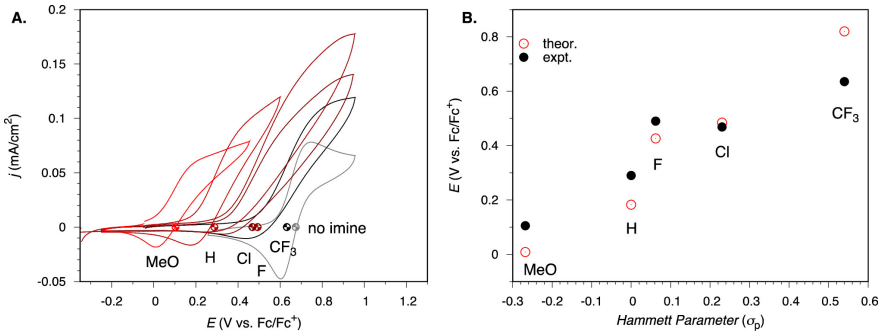

Figure 6.4: Comparison of experimental (A) and DFT computational (B) redox potentials for Cu^{II}/Cu^{I} vs. Fc/Fc^{+} as the para-substituents of the diaryl imine are changed.

Analysis of pathways for N–N bond formation. Previous kinetic studies of Cu-catalyzed oxidative homocoupling of 1 revealed a second-order kinetic dependence on [Cu][309] and led to the proposed Glaser-like coupling mechanism, involving binuclear N–N reductive elimination via a diamond-core intermediate (cf. Figure 6.1D). The energetics of this pathway were probed by DFT methods, starting from $[Cu(\mathbf{1})_4](OTf)_2$ (**3**) (Figure 6.5A). A three-coordinate Cu-iminyl complex **Int-1** can be generated from **3** by using one of the imine ligands as a Brønsted base to deprotonate another imine ligand. This step is calculated to be endergonic by 14.0 kcal/mol, but **Int-1** can undergo favorable dimerization to afford the bis-iminyl diamond-core intermediate (**dimer**, $\Delta G^{o} = 12.1$ kcal/mol with respect to **1**). Efforts were made to identify a transition state for N–N bond formation from this dimer by scanning the N–N distance; however, shorter N–N distances led to a monotonic increase in energy, reaching >40 kcal/mol with respect to **3** at a bond length 2.0 Å (Figure 6.5B). The origin of this behavior will be considered further below, but these results prompted us to consider an alternative pathway.

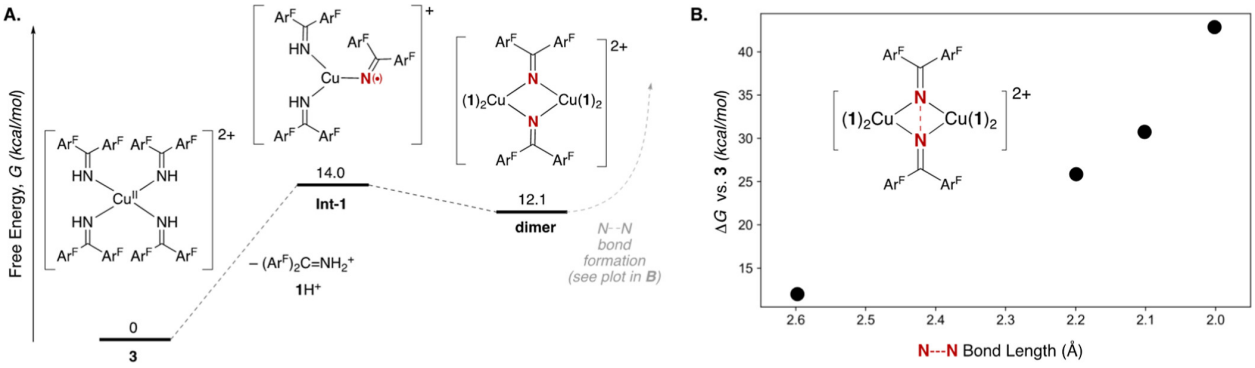

Figure 6.5: Free-energy diagram for a Glaser-like binuclear reductive elimination pathway for CuII-mediated N–N bond formation from imines. (A) DFT-based energy diagram for N–N bond formation, starting from $[Cu(\mathbf{1})_4](OTf)_2$ (**3**). (B) Free energy changes observed upon scanning the N–N bond in the **dimer** from 2.2 to 2.0 Å, while maintaining the angle between N–N and Cu–Cu. The data indicate that N–N bond formation from the diamond-core intermediate is unfavorable.

Deprotonation of the imine ligand in the formation of **Int-1** significantly changes the electronic structure of the Cu complex. Analysis of the Hirshfeld spin[8] distribution in **3** and **Int-1** reveals that the primary localization of spin shifts from predominantly Cu (0.66) in **3** to predominantly N (0.56) in **Int-1** (Figure 6.6), reflecting significant charge transfer from the non-coordinated lone pair of the anionic iminyl ligand into the d_{xy} orbital of Cu^{II}. Thus, **Int-1** is best described as a "Cu^{I}-(iminyl radical)" complex. This complex can dimerize to afford the diamond-core structure, but the **dimer** does not provide a pathway for N–N coupling, as shown in Figure 6.5. The N-centered radical character of **Int-1**, however, suggested the possibility of a radical-radical coupling pathway via direct N–N bond formation between two equivalents of **Int-1**. Transition state **TS-1** was identified at 26.0 kcal/mol with respect to **3**, providing an energetically accessible pathway to the azine-Cu^{I} dimer **Int-2**. This barrier is further reduced to 21.8 kcal/mol upon considering spin corrections (Supporting Information). The significant spin density on the reacting N atoms (0.59) in **TS-1** disappears when the N–N bond is formed in **Int-2**. Overall, this radical-radical coupling pathway readily rationalizes the experimental observations, including the bimolecular rate-law associated with N–N coupling and increased yields for the stoichiometric reaction when using stronger

Brønsted bases.[309] The electronic structural analysis shows that Cu^{II}-mediated oxidation of the imine substrate is triggered by deprotonation of the coordinated imine, prior to reaction with a second equivalent of Cu.

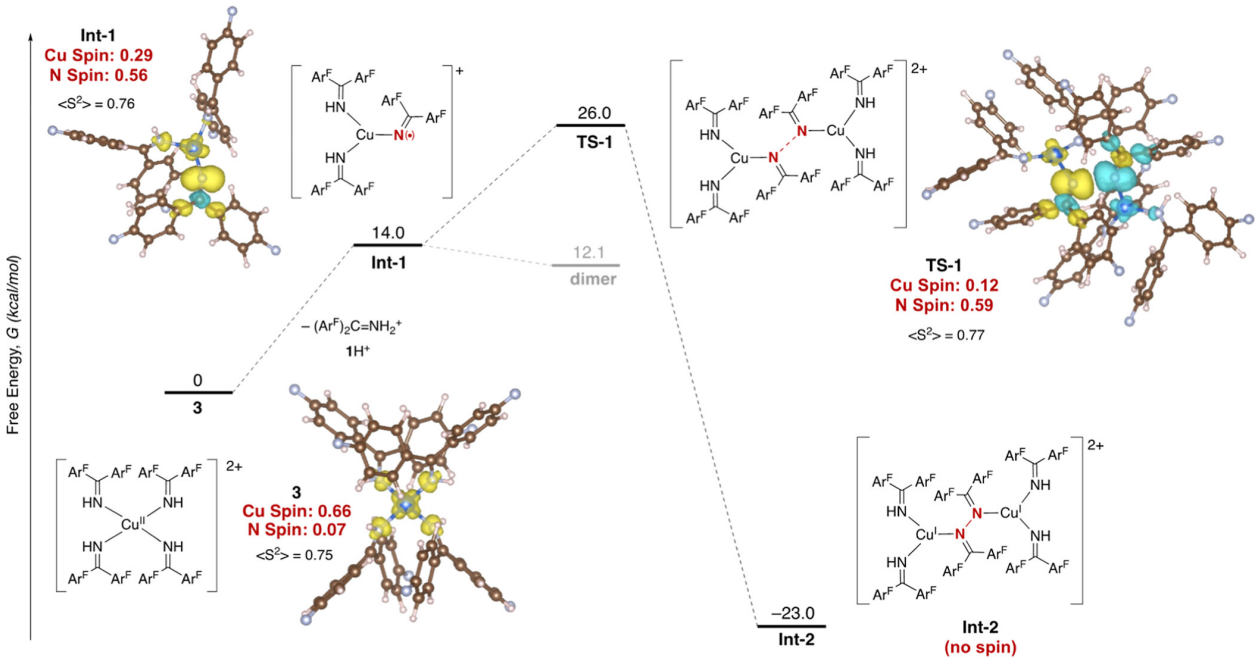

Figure 6.6: Free-energy diagram for a bimolecular radical-radical coupling pathway for Cu^{II}-mediated N–N bond formation from imines, and electronic structure of the key intermediates and transition state showing Hirshfeld spin distribution (yellow = up spin, blue = down spin).[8] Deprotonation of an imine ligand in **3** generates **Int-1**, which has significant Cu^{I}-(iminyl radical) character. Relevant bond lengths in **TS-1** (all in Å): Cu – N = 1.91, N – N = 2.38, N – Cu = 3.71.

Comparison of pathways for N–N, O–O, and C–C bond formation. Recognizing that the radical-radical coupling mechanism in Figure 6.6 could be involved in other Cu-catalyzed oxidative coupling reactions, we extended the DFT calculations to reactions with a series of anionic ligands derived from small molecules: hydroxide (-OH), amide ($-NH_2$), cyanide (-CN), and ethyne (-CCH). The ancillary diaryl imine ligands were replaced with dimethyl imines to facilitate the analysis. The viability of this substitution was confirmed by probing radical-radical homocoupling of $N{=}C(Ar^F)_2$ with the dimethyl imine ancillary ligands. The computed free-energy barrier for conversion of **Int-1** to **TS-1** is $\Delta G^{\dagger}$ = 12.8 kcal/mol (Sup-

porting Information), closely matching that obtained with full diaryl imine ancillary ligands ($\Delta G^{\dagger}$ = 12.0 kcal/mol, Figure 6.6).

The analysis of other substrates revealed a distinct bifurcation between the two reaction pathways for substrates with and without lone pairs (NH_2, OH versus CN, CCH). A radical-radical coupling transition state was identified for the coupling of NH_2 and OH, with free-energy barriers of 18.1 kcal/mol (NH_2) and 42.3 kcal/mol (OH) between structures analogous to **Int-1** and **TS-1** (Figure 7; the much higher barrier for HO-OH coupling reflects uphill thermodynamics associated with the formation of H_2O_2 from Cu^{II}). No transition state could be found for NH_2 or OH homocoupling via the binuclear reductive elimination pathway, similar to the results obtained for imine homocoupling, shown in Figure 6.5. While diamond-core dimer structures were identified, transition states could not be found. For example, a constrained scan of the HO—OH bond distance from a $Cu_2(OH)_2$ diamond-core geometry reaches energies higher than 100 kcal/mol as the O–O distance approaches that of H_2O_2 (see Supporting Information). Scans were evaluated with and without fixing the angle between the X–X and Cu–Cu bonds, but neither led to a transition state that formed the product (see Supporting Information for additional details).

With carbon-based ligands, CN and CCH, the opposite outcome was observed. No transition states were identified for the radical-radical coupling pathway. Instead, the monomeric Cu^{II}–CN and –CCH structures form dimeric diamond-core intermediates with low-energy pathways for C–C bond formation via binuclear reductive elimination transition states: $\Delta G^{\dagger}$ = 15.8 and 3.5 kcal/mol for CN and CCH, respectively (Figure 6.7). While these results closely resemble previous studies of Glaser homocoupling of alkynes,[303–305,307] we are not aware of previous studies of CN homocoupling.[320,321] The results show that the mechanistic bifurcation correlates with the presence (OH, NH_2, N=CAr_2) or absence (CN, CCH) of a substrate lone pair, with the former accessing a radical-radical coupling mechanism and the latter a binuclear reductive elimination pathway. Within the two mechanistic classes,

the calculated activation energies for these reactions correlated with the overall reaction energies ΔG^o, calculated by considering equilibrium Cu^I species formed in these reactions (Figure 6.7; see xyz in the Supporting Information for details). We have found the triplet solution to be higher in energy than the broken-symmetry solutions in all cases. However, spin correction lowers the computed barrier of the radical-radical couplings and raises that of the dimeric couplings, further demonstrating the bifurcation between these mechanisms (Supporting Information).

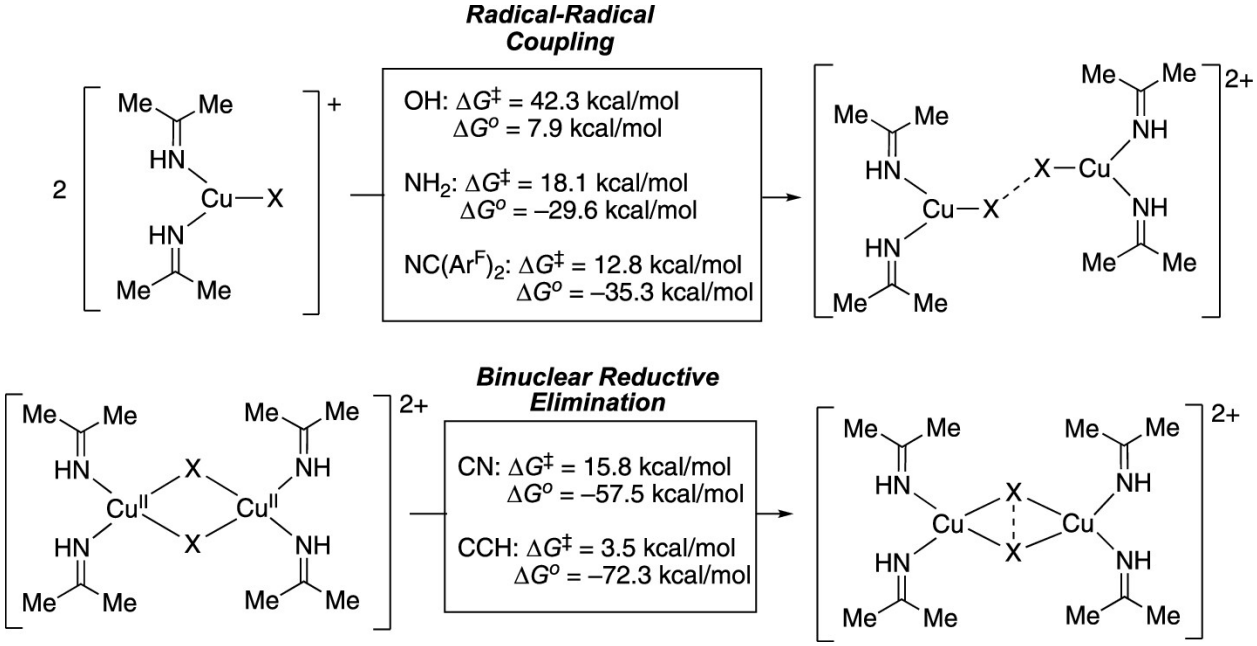

Figure 6.7: Activation free energies ($\Delta G^{\dagger}$) and reaction free energies (ΔG^o) for Cu-mediated oxidative homocoupling of OH, NH_2, $NC(Ar^F)_2$, CN and CCH ligands. Only one of the two mechanisms, (**1**) radical-radical coupling or (**2**) binuclear reductive elimination, was found to be accessible for each substrate, with the preferred pathway correlating with the presence (OH, NH_2, $N{=}C(Ar^F)_2$) or absence (CN, CCH) of a lone pair on the substrate. Reaction free energies are estimated with respect to the energy of the homocoupled X-groups (e.g., H_2O_2, N_2H_4) and the calculated Cu^I equilibrium species (see Supporting Information).

Electronic structural analysis of key intermediates and transition states. The origin of the mechanistic bifurcation evident among the different substrates in Figure 6.7 is illuminated by comparison of the different spin densities in Cu complexes involved in the two mechanisms (Figure 6.8). For the heteroatom substrates, X = OH, NH_2, $N{=}CAr^F{}_2$,

Hirshfeld spin densities were calculated for the **Int-1**-like monomers (Figure 6.8A, left) and the **TS-1**-like transition states (Figure 6.8A, right). For the carbon substrates, X = CN, CCH, the transition states for C–C bond formation optimized to a closed shell structure with no spin, so the monomeric **Int-1**-like structures (Figure 6.8B, left) were evaluated with the corresponding binuclear dimer structures (Figure 6.8B, right).

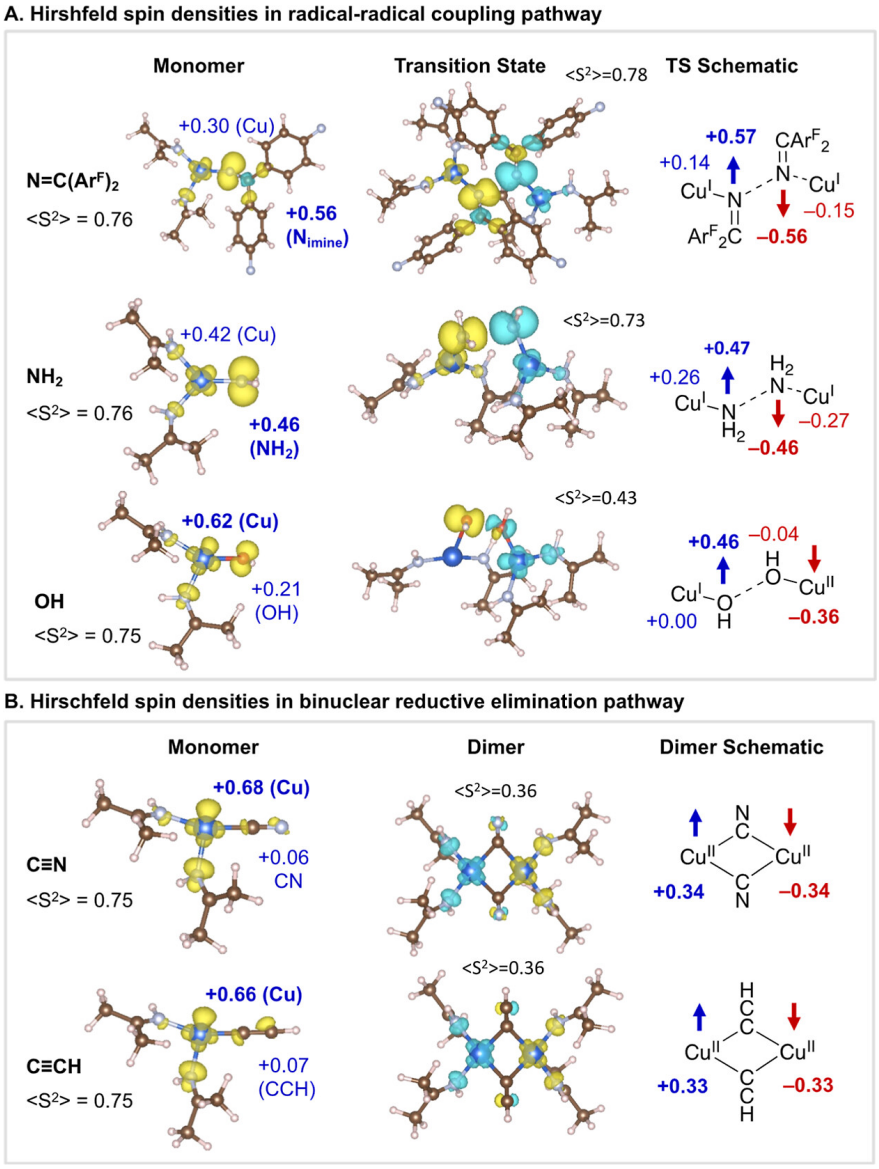

Figure 6.8: Calculated Hirshfeld spin densities of intermediates and transition states for the reactions of (A) heteroatom (OH, NH_2, N=CAr_2) and (B) carbon-based (CN, CCH) substrates. The monomeric $[(Me_2C{=}NH)_2Cu{-}X]^+$ species are shown on the left for each substrate, together with the transition-state structures for radical-radical coupling (A), and the dimeric intermediate for binuclear reductive elimination (B). The transition state for binuclear reductive elimination was calculated to be closed shell. The spin density maps are show with up spin in yellow and down spin in light blue, while the line drawings show up spin in dark blue and down spin in dark red. Relevant bond lengths in transition-state structures (all in Å): imine coupling: Cu-N = 1.97, N-N = 2.41, N-Cu (far) = 3.57; OH coupling: Cu-O = 2.32, O-O = 1.80, O-Cu = 3.27; CN coupling: Cu-C = 1.99, C-C = 1.79; CC coupling: Cu-C = 1.96, C-C = 2.02.

Several trends are evident in the calculated spin densities. Among the five monomeric structures (Figure 6.8, left), the nitrogen-based substrates, N=C(Ar^F)$_2$ and NH_2, have significant radical character, with spin densities of 0.56 and 0.46, respectively. At the other extreme, the carbon-based substrates, CN and CCH, have negligible radical character, with spin densities of only 0.06 and 0.07, respectively. OH represents an intermediate case, with the majority of the spin density located in a Cu d_{xy} orbital (0.62), while significant spin also resides on oxygen, in an orbital with sp character (0.21).

Distinctions are also evident in the transition states/intermediates associated with each of these substrates (Figure 6.8, right). The transition states for radical-radical coupling of the nitrogen-based substrates show symmetrical structures, with the spin retained with an equal distribution between the two nitrogen atoms undergoing bond formation. O–O bond formation favors a radical-radical coupling pathway, but the transition-state structure has an unsymmetrical distribution of spin. Relative to the monomeric Cu–O fragments, which have most of the spin localized on Cu, one of the Cu–O fragments in the transition state has increased spin density at oxygen (0.46, relative to 0.21 in the monomer), while the other has a decreased spin density at oxygen (0.04). This mismatch in spin localization correlates with a significantly higher barrier computed for O–O bond formation.

In the diamond-core intermediates leading to C–C bond formation, the spin density in the dimer is almost entirely located on Cu ($\pm$ 0.33–0.34) and the ancillary imine ligands (Figure 8B, right). The C–C bond-forming transition state for CN and CCH homocoupling has no spin (both optimize to closed shell determinants), and the forming C–C bonding orbital is heavily mixed with Cu-centered orbitals (see Supporting Information). These observations may be compared to the diamond-core [Cu_2[N=C(Ar^F)$_2$]$_2$ dimer intermediate, which does not undergo N–N coupling (see Figure 6.5). This structure also shows negligible spin density (0.01) at the bridging nitrogen atoms and significant spin on Cu (0.59) (see Supporting Information). An important distinction between the carbon- and heteroatom N/O-bridged

diamond-core structures is the C-atom uses a single lone pair to bridge the two Cu atoms, while the O- and N-atoms use two orthogonal lone pairs to bridge the two Cu atoms. This distinction has significant implications for the orbital interactions in the binuclear reductive elimination transition state, which supports C–C, but not N–N or O–O bond formation.

We have further analyzed the electronic structure of the diamond-core and biradical transition states through density matrix renormalization group (DMRG) calculations in Block2[205] through the PySCF[90] interface. Atomic valence active spaces[49] (AVAS) were chosen from a Hartree-Fock determinant targeting the relevant C and N 2s and 2p and Cu 3d orbitals for the active space. Figure 6.9 shows the resulting bonding and antibonding natural orbitals of these calculations: imine biradical coupling (left) shows an independent bonding/antibonding interaction, whereas the antibonding σ^* orbital of the forming C_4H_2 unit is hybridized with the Cu d orbitals. This difference in electronic structure clearly distinguishes between the two mechanisms, and these differences are reproduced in the multiconfigurational natural orbitals of the corresponding transition-state structures involved in OH, NH_2, and CCH coupling (Supporting Information).

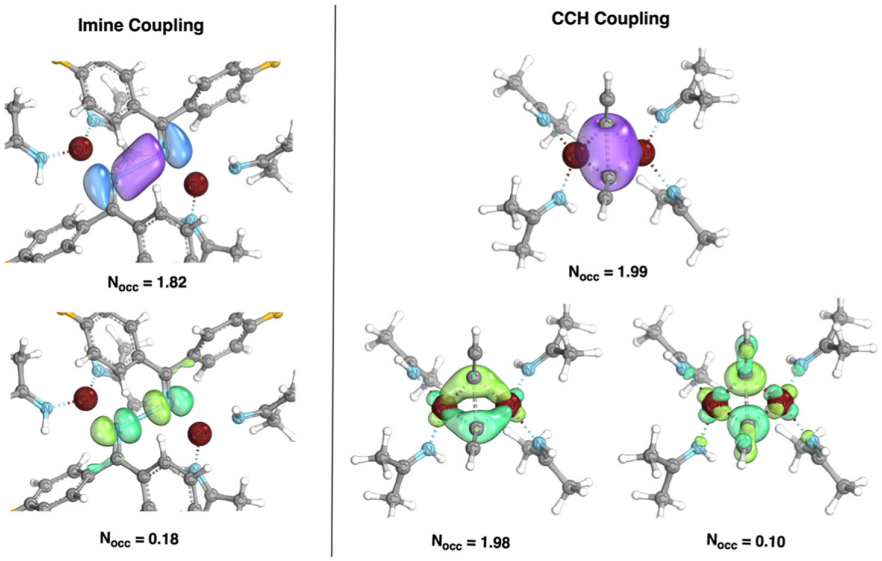

Figure 6.9: Natural orbitals from AVAS-DMRG calculations of the structures for the biradical imine coupling and diamond-core CCH coupling transition states. Left: σ and σ^* orbitals of the biradical imine coupling transition state, which are independent of the Cu d orbitals. Right: σ and σ^* orbitals of the CCH coupling transition state, in which the σ^* orbital is hybridized with the Cu d_{xy} orbitals.

Relationship to other literature reports of Cu-catalyzed oxidative homocoupling reactions. The data reported here has several important connections to previous data reported in the literature. The oxidative N–N coupling of diarylketimes to the corresponding azines has been achieved under electrochemical conditions.[310] This reaction may be promoted by a Brønsted base (di-*n*-butyl phosphate) or with a pyridine-ligated copper catalyst that resembles the system described here. The base-promoted reactions is proposed to proceed via proton-coupled electron transfer from neutral imines, generating free N-centered radicals that can undergo radical-radical coupling. The data outlined herein suggest that the Cu catalyst mediates N–N bond formation via a different radical-radical coupling pathway that features Cu-ligated N-centered radicals. Participation by the Cu center significantly lowers the energy requirement for formation of the N-centered radical. This lower barrier is manifested by the ca. 1 V lower overpotential required to promote electrochemical N–N bond formation, reflecting the different potentials needed to oxidize

Cu^{I} to Cu^{II} versus promoted electrochemical PCET from the substrate.[310] These observations further reflect that imine coordination to Cu^{II} facilitates substrate deprotonation (via Lewis-acid coordination-activation) and electron transfer (inner-sphere ligand-to-metal charge transfer upon deprotonation).

The involvement of N-centered radicals in stoichiometric and catalytic N–N coupling reactions has been directly characterized in recent studies by the groups of Warren[322,323] and Peters.[324] These efforts included the synthesis and structural, reactivity, and computational analysis of Cu^{II} and Ni^{III} complexes bearing well-defined ancillary ligands, β-diketiminates for Cu and a tetrapodal (SiP2S) ligand for Ni. Warren and coworkers further showed that a β-diketiminate-ligated Cu complex can serve as an electrocatalysts for ammonia oxidation to dinitrogen,[323] complementing the aforementioned aerobic and electrocatalytic coupling of diarylimines.[309,310] In both cases, the key N–N bond-forming step is proposed to involve bimolecular reaction of two N-centered radicals coordinated to Cu. This ligand radical character and facile homocoupling reactivity is not matched by hydroxide, as shown in Figures 6.7 and 6.8. These observations align with literature reports showing that oxygen-radical character in mononuclear Cu–OH complexes does not arise until the complexes are increased by one oxidation state relative to those presented here (i.e., formally Cu^{II} to Cu^{III}, although ligand-based redox negates this formal description[325]).[326–329] Such complexes are potential intermediates in electrocatalytic water oxidation mediated by homogeneous Cu-complexes.[330–332] Once the complexes reach this oxidation state, their electronic structure introduces the possibility of O–O bond formation via the now-classic Cu_2O_2-diamond-core structure, corresponding to the binuclear reductive elimination pathway that proved untenable at the lower oxidation state (cf. Figure 6.7).[313–316,333,334]

6.4 Conclusion

This analysis of the Cu^{II}-catalyzed oxidative homocoupling of diaryl imines leads to a mechanistic conclusion different from that proposed previously. Specifically, this reaction does not mimic Glaser-type alkyne coupling, but rather features a radical-radical coupling pathway. Coordination of the neutral imine to Cu forms stable Cu^{II} complexes in solution. Upon deprotonation, the imine ligand undergoes significant change in its electronic structure, generating a Cu-stabilized iminyl radical that is equipped to undergo facile N–N coupling. The ability to access this reactivity from the Cu^{II} oxidation state enables catalytic turnover to proceed with O_2 as the stoichiometric oxidant or under modest electrochemical potentials. More broadly, the results presented herein highlight two distinct pathways for homocoupling depending on whether the reacting groups are carbon (-CN, -CCH) or heteroatoms (-N=CAr_2, -NH_2, and -OH) ligands. This study is the first to show this distinction among these important historical and contemporary coupling reactions.

CHAPTER 7

ORGANIC REACTIVITY MADE EASY AND ACCURATE WITH AUTOMATED MULTIREFERENCE CALCULATIONS

This chapter is reprinted with permissions from *ACS Cent. Sci.* **2024** (*Accepted*)

7.1 Abstract

In organic reactivity studies, quantum chemical calculations play a pivotal role as the foundation of understanding and machine learning model development. While prevalent black-box methods like density functional theory (DFT) and coupled-cluster theory (e.g., CCSD(T)) have significantly advanced our understanding of chemical reactivity, they frequently fall short in describing multiconfigurational transition states and intermediates. Achieving a more accurate description necessitates the use of multireference methods. However, these methods have not been used at scale due to their often-faulty predictions without expert input. Here, we overcome this deficiency with automated multiconfigurational pair-density functional theory (MC-PDFT) calculations. We apply this method to 908 automatically generated organic reactions. We find 68% of these reactions present significant multiconfigurational character, in which the automated multiconfigurational approach often provides a more accurate and/or efficient description than DFT and CCSD(T). This work presents the first high-throughput application of automated multiconfigurational methods to reactivity, enabled by automated active space selection algorithms and the computation of electronic correlation with MC-PDFT on-top functionals. This approach can be used in a black-box fashion, avoiding significant active space inconsistency error in both single- and multireference cases and providing accurate multiconfigurational descriptions when needed.

7.2 Introduction

In the past 20 years, quantum chemistry has made great strides in describing chemical reactivity; widely-used methods such as density functional theory (DFT) and coupled-cluster methods (e.g., CCSD(T)) have become a rich source of data for the understanding of chemical reactions and the development of machine learning algorithms.[335,336] However, despite their black-box nature, these methods face limitations on systems poorly described by a single electronic configuration, i.e. multiconfigurational or strongly correlated systems.[337–341] A key example of these systems is familiar to most chemists: that of the transition state, in which the electronic character is often split between describing that of the reactant and the product. Given the ubiquitous nature of transition states in chemistry, it may then be a wonder how these approaches have proven successful in so many applications. The answer is that for many important cases these methods are simply able to overcome this difficulty despite the fundamental struggle with multiconfigurational character. Nevertheless, in automated applications of quantum chemistry such as reaction network exploration,[342] the poorer description of multiconfigurational species can rear its head in key places and significantly impact results.

As such, describing strong correlation in transition states has long been poised as a potential application for multiconfigurational approaches such as complete active space self-consistent field (CASSCF) theory.[26,343] This approach overcomes the difficulty of describing multiconfigurational systems by describing the state as a superposition of the possible electronic configurations in an "active space" of orbitals and electrons:[29]

$$|\Psi_{\text{CASSCF}}\rangle = \sum_{n_1 n_2 ... n_L} C_{n_1 n_2 ... n_L} \, |22...n_1 n_2 ... n_L 00...\rangle \tag{7.1}$$

in which $n_1 n_2 ... n_L$ enumerates the possible occupations of the L active orbitals. With a good choice of active space, all static correlation can be addressed with far fewer configu-

rations than FCI and comparable expense to DFT.[6] However, despite the many academic applications of these approaches in the literature,[344–350] the widespread adoption of these methods for reactivity has been hindered by the challenge of choosing a consistent and adequate active space along the reaction surface.[86,351] The CASSCF energy expression is given by

$$E_{\text{CASSCF}} = V_{\text{NN}} + \sum_{pq} h_{pq} D_{pq} + \sum_{pqrs} g_{pqrs} d_{pqrs} \tag{7.2}$$

where D_{pq} and d_{pqrs} are the CASSCF one- and two-body reduced density matrices. If the active space is chosen inconsistently between two geometries, one will obtain an unphysical "active space inconsistency error" (ASIE) resulting from the inconsistent treatment of correlation in the density matrices of equation 7.2. This error generally remains present even when addressing the remaining dynamic correlation perturbatively with methods such as CASPT2[352,353] or NEVPT2.[34,35]

The most common approach for reducing ASIE involves interpolating the active space orbitals between geometries, providing a continuous set of orbitals along the reaction coordinate.[351] However, this approach is quite cumbersome: active orbitals often rotate in and out of the active space randomly during this procedure, and the active space may change size along a reaction coordinate, such as when moving from a fairly uncorrelated reactant to a correlated transition state. Furthermore, this interpolation scheme dramatically increases the cost of the calculation relative to approaches such as DFT, as CASSCF calculations are necessary along several points between the reactant and product, whereas KS-DFT only requires calculations at the individual end points.

In this light, we note the broad success of KS-DFT in modeling reactivity, which models all densities via a single determinant and calculates energies via use of an exchange-correlation functional:

$$E_{\text{KS-DFT}} = V_{\text{NN}} + \sum_{pq} h_{pq} D_{pq} + \sum_{pqrs} g_{pqrs} D_{pq} D_{rs} + E_{\text{xc}}[\rho] \tag{7.3}$$

Despite the fact that the KS-DFT determinant inevitably describes the density matrices of reactants and transition states with different accuracy (i.e., the exact two-body density matrices d_{pqrs} differ more or less from the single-determinant $D_{pq}D_{rs}$), KS-DFT is able to obtain good results in reactivity through use of an exchange functional of the density $E_{xc}[\rho]$. This statement also applies to the success of "density corrected" DFT (DC-DFT),[354] in which the densities used in the KS-DFT energy expression (eq. 7.3) come from HF determinants (i.e., the functional has no input on the density, but only the energy calculation). This leads to the hypothesis that the ASIE found in CASSCF and NEVPT2 may come in large part from unequal contribution of the density cumulant between two geometries, $d_{pqrs} - D_{pq}D_{rs}$. A multiconfigurational approach that avoids use of the density cumulant by means of an exchange-correlation functional may inherit much of the equal-footing properties of KS-DFT and prove more robust against ASIE.

One such method that achieves this goal is called multiconfigurational pair-density functional theory (MC-PDFT).[41] This theory more-or-less shares an energy expression with KS-DFT:

$$E_{\text{MC-PDFT}} = V_{\text{NN}} + \sum_{pq} h_{pq} D_{pq} + \sum_{pqrs} g_{pqrs} D_{pq} D_{rs} + E_{\text{ot}}[\rho, \Pi] \tag{7.4}$$

with two key differences: (i) the exchange-correlation functional is replaced with an "on-top" functional E_{ot} which is a functional of both the density ρ and on-top density Π, and (ii) the density arguments D_{pq}, ρ, and Π come from a multiconfigurational (generally CASSCF) wave function. The on-top pair density, derived from the two-particle density matrix, describes the probability of finding two electrons at the same point in space. In practice, the on-top functional is a "translated" functional (most often translated PBE[18], tPBE) in which the

density and on-top density are used to manufacture effective spin densities for use in the KS-DFT energy expression (eq. 7.3). Thus, as MC-PDFT more-or-less shares eq. 7.3 with KS-DFT, MC-PDFT appears promising for attenuating part of the active space inconsistency error, especially when paired with automated methods for choosing the active space in a reliable and consistent fashion.[5,6,31,76,86,351,355] While MC-PDFT has been tested on a wide variety of systems and excitations,[6,141,347,356] it has yet to be tested in a high-throughput fashion for reactivity.

Here we provide the first such test by applying automated MC-PDFT to the calculation of 908 automatically generated organic reactions in the RGD1 database.[357] These data present a rich variety of organic reactivity and a challenging test for multiconfigurational approaches that is germane to reaction network exploration. Our results highlight the robustness of automated MC-PDFT in this domain compared to other perturbative multiconfigurational approaches such as NEVPT2[34,35] and outline the opportunity and challenges for applying multiconfigurational methods to high-throughput main-group reactivity. We find that combining the approximate pair coefficient active space selection scheme (APC) with MC-PDFT (referred to as APC-PDFT) generates robust results, with APC-PDFT reproducing DFT results for a set of single reference reactions. In addition, we show the deviation in relative energies from single reference are correlated to level of multiconfigurational character, with DFT and CCSD(T) becoming less reliable for strongly correlated systems (68% of reactions), and APC-PDFT providing better results in many of these cases.

7.2.1 Methods

The main barrier to automating multiconfigurational approaches is automatically selecting the active space in a robust fashion. Methods for automatically selecting active spaces continue to be an active research topic, and several approaches exist.[31,49,76,86,87,351,355,356,358] Here, we employ approximate pair coefficient (APC) selection,[5,6] in which candidate Hartree-

Fock orbitals are ranked for the active space by means of their approximate pair coefficient interaction with other orbitals. We note that APC is a ranked-orbital approach, where the user defines a maximum active space size. This method allows the practitioner to prevent the selection scheme from picking active spaces larger than are computationally feasible and it also allows for flexibility towards solvers with different practical size limitations (i.e. CAS vs. DMRG). The drawback is that the user has to define this maximum size manually which can result in an unnecessarily large active space. Given doubly occupied orbitals i and virtual orbitals a, approximate pair coefficients are calculated as

$$C_{ia} = \frac{0.5K_{aa}}{F_{aa} - F_{ii} + \sqrt{(0.5K_{aa})^2 + (F_{aa} - F_{ii})^2}} \tag{7.5}$$

where F_{ii}, F_{aa}, and K_{aa} are the respective diagonal elements of the Fock and exchange matrices. The entropies of doubly occupied orbitals i and virtual orbitals a are then calculated by summing over their approximated interactions (intermediate normalization):

$$S_i = -\frac{1}{1+\sum_a C_{ia}^2} \ln \frac{1}{1+\sum_a C_{ia}^2} - \frac{\sum_a C_{ia}^2}{1+\sum_a C_{ia}^2} \ln \frac{\sum_a C_{ia}^2}{1+\sum_a C_{ia}^2} \tag{7.6}$$

$$S_a = -\frac{1}{1+\sum_i C_{ia}^2} \ln \frac{1}{1+\sum_i C_{ia}^2} - \frac{\sum_i C_{ia}^2}{1+\sum_i C_{ia}^2} \ln \frac{\sum_i C_{ia}^2}{1+\sum_i C_{ia}^2} \tag{7.7}$$

Interactions with singly occupied orbitals are left uncalculated, and singly occupied orbitals are automatically given the highest possible entropy. As the pair coefficients are generated from Fock and exchange matrix elements which change adiabatically with the molecular geometry, the APC scheme aims to select moderately consistent (but not exactly consistent) active spaces across the reaction coordinate.

Finally, due to the observed biasing of APC entropies towards doubly occupied orbitals[5,6] a series of virtual orbital removal steps are employed N times in which the highest-entropy virtual orbital is removed from the sums in equations 7.6 and 7.7 and the entropies are

recalculated; these highest-entropy virtual orbitals are then assigned the highest entropy at the end of the calculation. For small-to-medium sized organic systems we have found good results with $N = 2$,[6] which we have used here. However, this parameter appears to have less impact due to the fixed active space size we employ here to enforce active space size consistency between different geometries (described below). Implementation of APC is now available in PySCF.[359,360]

Candidate HF orbitals are then ranked in importance by their orbital entropies, with this ranking used to choose an active space meeting some user-defined size requirement (e.g., a 12 electron in 12 orbital or (12,12) active space). Here, to select consistent active space sizes between geometries we employ a simple size requirement in which for an (A,B) active space, where A and B are the number of active electrons and orbitals respectively, the $A/2$ highest-entropy doubly occupied orbitals and the $B - A/2$ highest-entropy virtual orbitals are added to the active space; we refer to these active spaces as APC-(A,B). CASSCF calculations initialized from these active spaces in the cc-pVDZ basis[361,362] were then carried out in PySCF.[359,360] These CASSCF wave functions were then used for the calculation of MC-PDFT (tPBE) and NEVPT2 energies, also implemented in PySCF and PySCF-Forge.[213]

Multiconfigurational (or equivalently, multireference (MR)) character in the resulting wave functions is calculated via the M-diagnostic,[3] which measures multiconfigurational character as a function of the natural orbital occupancies:

$$M = \frac{1}{2}(2 - n_{\mathrm{HDOMO}} + n_{\mathrm{LUMO}} + \sum_{j_{\mathrm{SOMO}}} |n_j - 1|) \tag{7.8}$$

Here n_{HDOMO}, n_{LUMO}, and n_{SOMO} are the average occupations of the highest doubly occupied, lowest unoccupied, and any singly occupied orbitals in the active space. An M diagnostic less than 0.05 is considered minimally multiconfigurational, $0.05 < \mathrm{M} < 0.1$ moderately MR, and $\mathrm{M} > 0.1$ substantially MR.

7.2.2 Data

The reactions for this benchmark were taken from the Reaction Graph Depth 1 (RGD1) dataset for CHON-containing molecules.[357] In brief, these reactions were generated using generic graph-based reaction rules applied to neutral closed-shell reactants sampled from PubChem. Transition state, reactant, and product geometries for each reaction were optimized at the B3LYP-D3/TZVP level. Three subsets of RGD1 were used for this work. These are a random five percent (400) of the break two form one (B2F1) reactions, 400 break two form two (B2F2) reactions, and a "small molecule" dataset of 108 reactions in RGD1 with $<$ 5 non-hydrogen atoms. The B2F1 reactions, which break two bonds and form one bond as the reaction progresses from reactant to product, have an increased likelihood of showing MR character due to the uneven number of bonds formed and broken in the reaction, whereas the B2F2 reactions, which have two bonds broken and two bonds formed throughout the reaction, have closed-shell reactants and products (Supporting Information). To provide reference results for comparison to the automated multiconfigurational approach, CCSD(T) and B3LYP-D3 (with zero damping) results were recalculated in the cc-pVDZ basis in PySCF using the all-atom pre-associated reactants and products provided by RGD1.

7.3 Results

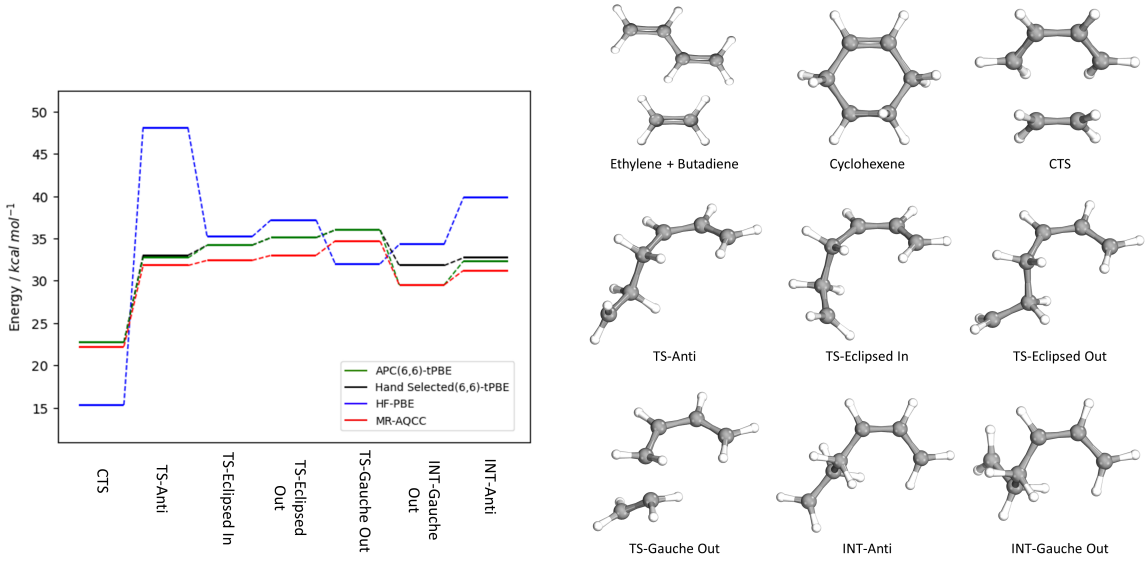

Figure 7.1: Electronic energies of each state in the concerted transition state (CTS) and biradical reaction pathways relative to the reactants. Four methods are shown: APC(6,6)-tPBE (green, this work), hand-selected (6,6)-tPBE (black),[9] HF-PBE (blue), and reference MR-AQCC results (red).[10] The structures of each transition state and intermediate are displayed on the right.

As a first test of our methodology, we explore the performance of APC-tPBE on the Diels-Alder reaction between butadiene and ethylene. This reaction presents a well-studied series of transition states and intermediates[9,10] that provide a clear challenge for automated multi-configurational approaches, as all states contain a significant amount of multiconfigurational character ($M > 0.1$). Figure 7.1 shows the tPBE results obtained with our automated APC(6,6) active spaces compared to previous literature results using hand-selected (6,6) active spaces,[9] as well as reference multireference averaged quadratic coupled cluster (MR-AQCC) calculations.[10] The study from Lischka *et al.* showed the MR-AQCC results to be in good agreement with experiment for the accepted reaction pathway.

As is seen, the automatically selected active spaces are able to reproduce the tPBE results

(in good agreement with the MR-AQCC results) of the hand-selected active spaces in all transition states, despite not directly enforcing any consistency between active spaces beyond the size. For reference, we show the single-reference limit of MC-PDFT in which the CASSCF wave function densities are replaced with HF densities (equivalent to so-called "density-corrected" PBE[354]); here we refer to this approach as HF-PBE. Unlike APC(6,6)-tPBE, HF-PBE dramatically overestimates the stability of the concerted transition state (CTS) while greatly underestimating the stability of the TS-Anti transition state and intermediate. Results with an APC(12,12) active space as well as KS-DFT and CCSD(T) are reported in the Supporting Information. The larger active space results are in good agreement with the APC(6,6) performance. Thus, our automated scheme successfully reproduces the important multiconfigurational results.

Given the success of our methodology in reproducing Diels-Alder results, we turn to the 908-reaction subset of RGD1 reactions for further testing. Our calculations show that this set of reactions shows a broad distribution of multiconfigurational character as measured by the M-diagnostic (Supporting Information), with 32% of reaction energies and 63% of activation energies demonstrating significant multiconfigurational character ($M > 0.1$), for a total of 68% of reactions exhibiting such character in at least one state overall. To account for the cases with the most multiconfigurational character, we have chosen large APC(12,12) active spaces for each state in these reactions. This active space size is significantly larger than necessary for most reactions in the dataset, resulting in inconsistent but unimportant orbitals between the reactants and products of some reactions. These orbital inconsistencies represent a second test of the robustness of MC-PDFT.

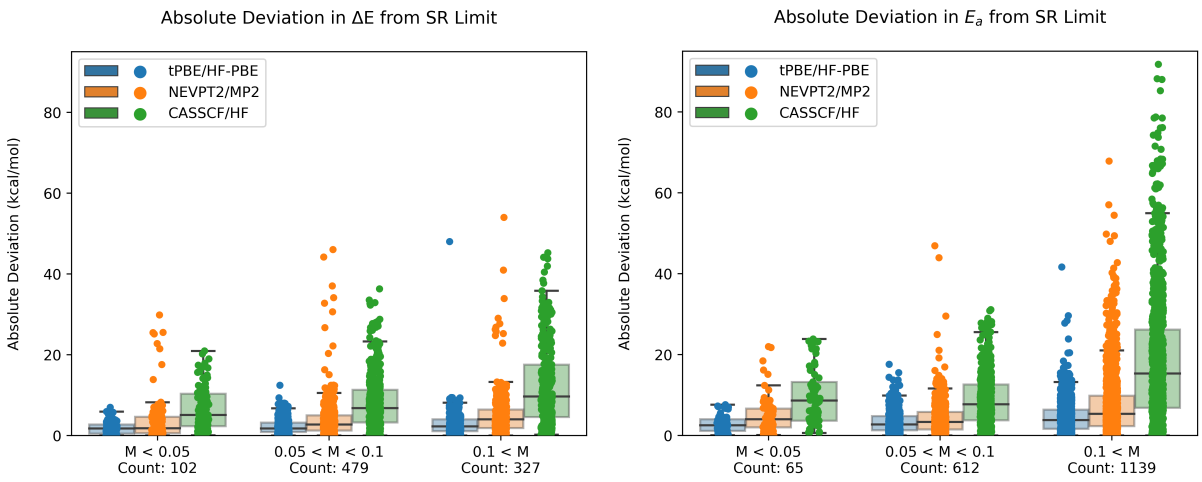

Figure 7.2: Whisker plots of deviations from single-reference limits (right: ΔE left: E_a) of APC-tPBE, APC-NEVPT2 and APC-CASSCF, stratified by the degree of multiconfigurational character as measured by the M diagnostic. The number of reactions in each M diagnostic category are displayed below each label. Mean absolute deviations (MAD) in systems with low multiconfigurational character ($M < 0.05$, in kcal/mol, ΔE /E_a): 1.8/2.8 (tPBE); 4.0/5.2 (NEVPT2); 6.8/9.8 (CASSCF). Mean absolute deviations (MAD) in systems with high multiconfigurational character ($M > 0.1$, in kcal/mol, ΔE /E_a): 3.1/4.6 (tPBE); 5.2/7.6 (NEVPT2); 12.3/19.0 (CASSCF).

Figure 7.2 shows the absolute deviation in the reaction energy, ΔE, and the activation energy, E_a (both forward and backward), for all examined reactions from the single reference limit (SRL) for CASSCF (SRL: HF), tPBE (SRL: HF-PBE), and NEVPT2 (SRL: MP2). This deviation is stratified by three degrees of multirference (MR) character (low ($M < 0.05$), moderate ($0.05 < M < 0.1$), and high ($0.1 < M$). As shown clearly, both the mean absolute deviation (MAD) from the SRL and overall spread of the data increases from the low M to the high M categories. In the cases with low multiconfigurational character, $M < 0.05$, tPBE successfully reproduces the single-reference limit with a mean deviation of $\pm$ 1.8 kcal/mol for ΔE and $\pm$ 2.8 kcal/mol for E_a, with an average between these two of $\pm$ 2.2 kcal/mol. In contrast, CASSCF and NEVPT2 reproduce these limits with a mean deviation of $\pm$ 7.9 kcal/mol and 4.4 kcal/mol repectively, with much larger maximum deviations

(as high as 20 kcal/mol). These results show that MC-PDFT is significantly more robust in the single-reference limit towards active space inconsistency error (ASIE) than competing multiconfigurational approaches, making it ideal for high-throughput application. Surprisingly, we find that this robustness carries over to the performance of hybrid PDFT as well, despite it being an admixture of CASSCF and tPBE; this point bears technical discussion and is discussed in the Supporting Information. A similar analysis, using the square of the coefficient of the leading configuration, C_0^2, as the multireference diagnostic can also be found in the Supporting Information.

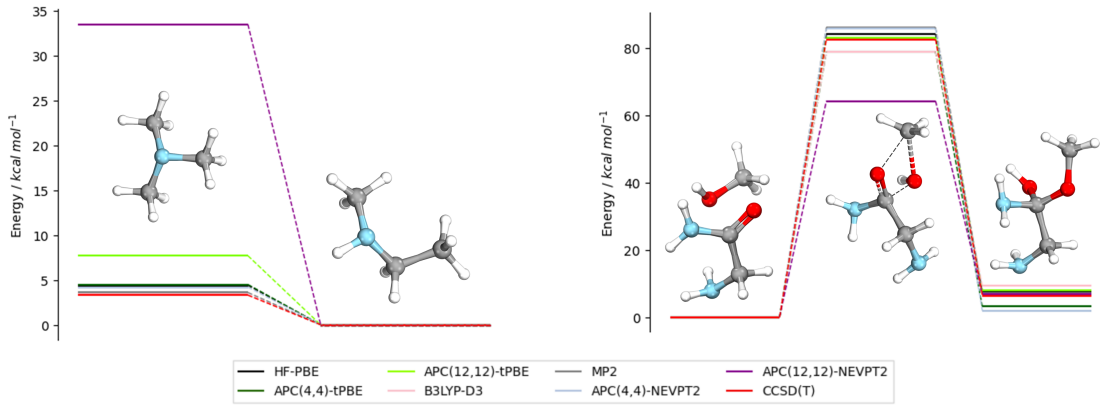

Figure 7.3: Reactions MR_3361_1 (rearrangement of trimethylamine) and MR_619998_2 (hemiacetal formation from methanol and glycinamide). Six methods are shown on each plot: APC(12,12)-tPBE (light green), APC(12,12)-NEVPT2 (purple), APC(4,4)-tPBE (dark green), APC(4,4)-NEVPT2 (silver), HF-PBE (black), MP2 (grey), B3LYP-D3 (pink), and reference CCSD(T) (red). Energies shown are calculated relative to the lowest energy state (right: reactants, left: products). Since the transition state of the trimethylamine rearrangement reaction is reasonably multireference (M=0.49), it is excluded here.

Two examples where tPBE shows improved reliability for a single-reference reaction are shown in Figure 7.3. The first is a trimethylamine rearrangement reaction, where the APC(12,12)-CASSCF wave functions for the reactant and product are mostly well-described by a single determinant, with M diagnostics below 0.03. Thus, the overall reaction energy is expected to be similar between each MR approach and its single reference parallel. As is seen, APC-tPBE successfully reproduces HF-PBE to within 3 kcal/mol, a result that is

similarly in-line with B3LYP-D3 and CCSD(T). Though this deviation is slightly larger than chemical accuracy, it presents a substantial improvement over APC(12,12)-NEVPT2, which shows a clear deviation from all other methods, overestimating the energy of the reactant by roughly 30 kcal/mol, despite using the same underlying APC-CASSCF wave functions as APC-tPBE. This drastic difference from the single-reference result is emblematic of ASIE, where orbital rotation between the product and reactant results in drastically unphysical results. Since the reaction is known to be single-reference, this ASIE can be eliminated through the selection of a smaller active space: APC(4,4)-NEVPT2 produces results in line with CCSD(T) and density functional approaches, and the APC(4,4)-tPBE results come closer in line with CCSD(T).

The second case presents the formation of a hemiacetal from methanol and glycinamide. Here, all three states exhibit an M of less than 0.05, indicating both the reaction and activation energies should be well described by a single-determinant wave function. Despite this, both the forward and reverse barriers are predicted to be 20 kcal/mol lower with APC(12,12)-NEVPT2 than MP2. By comparison, APC-tPBE agrees to within chemical accuracy (1 kcal/mol) with the single-reference limit of HF-PBE and CCSD(T). Once again, the smaller APC(4,4) active space largely remedies this unphysical error with NEVPT2, demonstrating the error to be due to ASIE. An in-depth evaluation of the active space dependence of tPBE and NEVPT2 for these two reactions, as well as CASSCF is included in the Supporting Information.

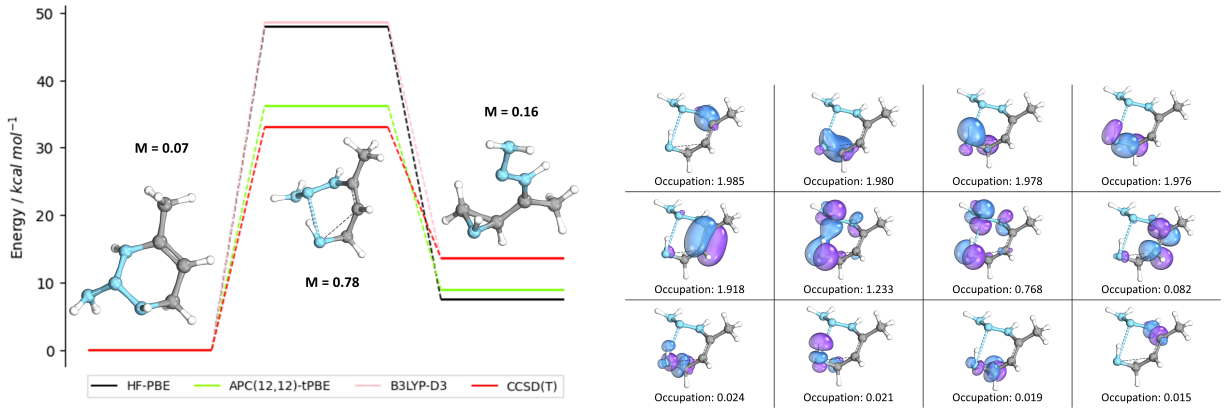

Figure 7.4: Reaction MR_186317_0 (ring-opening/ring-closing reaction of $N_4C_4H_{10}$). The APC(12,12)-tPBE (green), HF-PBE (black), B3LYP-D3 (pink), and CCSD(T) (red) energy diagrams are displayed on the left. The transition state active orbitals and their occupations are shown on the right.

We next show by example how multiconfigurational effects can result in important deviations from DFT and CCSD(T) in the RGD1 dataset. The first example is shown in Figure 7.4, which highlights the most common type of deviation from single reference in which the transition state exhibits the largest degree of multiconfigurational character (M = 0.767). The transition state orbitals of this ring-opening/ring-closing reaction show significant multiconfigurational character in both the bond breaking of the 6-membered ring and the C-C double bond rearranging to form the 3-membered ring. The concerted nature of this ring-opening reaction makes this a difficult case for single-reference approaches, much like the Diels-Alder reaction studied prior (Figure 7.1). As a result, B3LYP-D3 and HF-PBE overestimate the activation energy of the forward reaction by 12 kcal/mol relative to APC-tPBE. In this case, the multiconfigurational character is able to be captured by CCSD(T), which is largely in agreement the automated APC-tPBE results. The chosen orbitals and their occupations for the transition state are shown alongside the energy diagram.

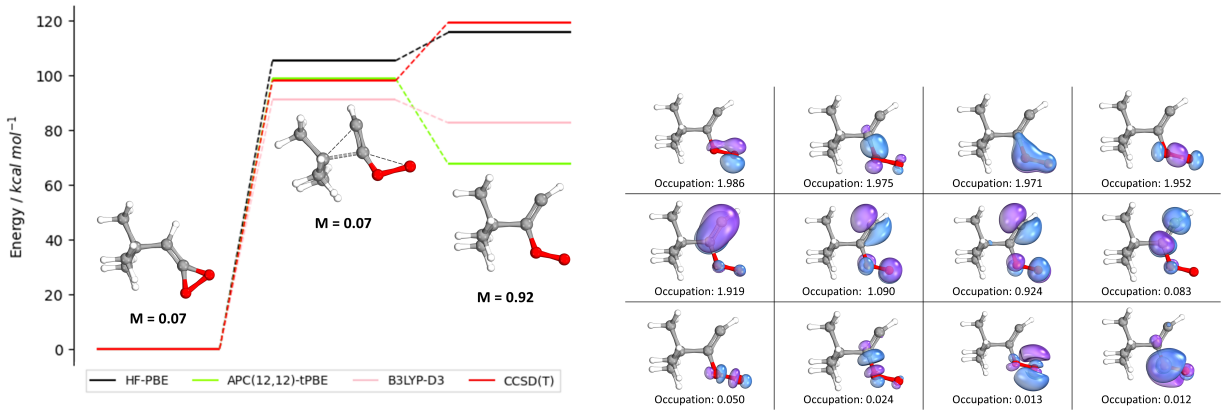

Figure 7.5: Reaction MR_673407_0 (ring opening of 3-membered heterocycle). The APC(12,12)-tPBE (green), HF-PBE (black), B3LYP-D3 (pink) and CCSD(T) (red) energy diagrams are displayed. The product active orbitals and their occupations are shown on the right.

Figure 7.5 presents a second case, in which the ring-opening of a 3-membered heterocycle forms an oxygen diradical with significant multiconfigurational character. As is seen, the HF determinant is completely incapable of describing this diradical product, predicting it to lie overestimating the energy of this product relative to the reactant by 60 kcal/mol – higher in energy than the transition state. Due to this terrible description given by HF, CCSD(T) also dramatically overestimates the energy of the biradical relative to the transition state. The unrestricted nature of B3LYP-D3 is able to account for the multiconfigurational character of the biradical somewhat, predicting a shallow barrier of 8.5 kcal/mol relative to the transition state. In contrast, APC-tPBE predicts a significantly more stable product, with a barrier of 31.3 kcal/mol relative to the transition state, and in much better agreement with the CCSD(T) reference values for the single-reference reactant and transition state. We believe these APC-tPBE results give a much more accurate description than either DFT or CCSD(T), and serve to highlight the necessity of multiconfigurational approaches for some reactions containing significant multiconfigurational character.

As a study of basis set dependence, we have investigated the behavior of B3LYP, APC-tPBE, and CCSD(T) in the larger cc-pVTZ basis for the case studies presented in Figures 3-5 (Supporting Information). Overall, we find the APC-tPBE to be remarkably consistent with respect to basis set size, with nearly all results in the cc-pVDZ basis set being well-reproduced in the larger cc-pVTZ basis and qualitatively similar correlating orbitals being chosen in all cases. However, a large discrepancy is found in the cc-pVTZ description of MR_673407_0, in which the APC(12,12)-tPBE reaction energy changes from 31.3 kcal/mol in the cc-pVDZ basis to 12.9 kcal/mol in the cc-pVTZ basis. We find that this discrepancy is due to an abnormally large ASIE in the cc-pVTZ basis, which can be eliminated by executing a CASCI in only the (4,4) active space of correlating orbitals (visually identical to those of the cc-pVDZ basis), which largely reproduces the results shown in Figure 7.5. This process of recomputing reaction energies using CASCI calculations in only the space of correlating orbitals is promising for further reducing ASIE in APC-tPBE and will be explored in future work.

7.4 Discussion/Conclusion

We have here presented the first large-scale automated multiconfigurational approach to the modeling of organic reactivity, which provides a compelling alternative to DFT and CCSD(T) for interrogating chemical space. These multiconfigurational methods have been held back from high-throughput application for decades due to the problem of active space inconsistency error (ASIE), which is here overcome through the increased robustness of the MC-PDFT method to ASIE and automated active space selection with the approximate pair coefficient (APC) approach. We have applied this automated APC-PDFT approach to the calculation of 908 main group reactions from the RGD1 database, which successfully reproduces the single-reference limit with ASIE of $\pm$2.2 kcal/mol (similar to deviations between different density functionals) while providing more accurate multiconfigurational descriptions

than DFT and CCSD(T) in many of the 68% of reactions containing multiconfigurational character. Taken at face value, these results make it possible to for the first time envision the high-throughput use of multiconfigurational methods in this domain, potentially increasing the accuracy of predictions at significantly lower cost (and possibly higher accuracy) than CCSD(T).

Of course, there are limitations. Firstly, there is no reason to expect good results if a sufficient active space is not chosen for all geometries. In the best case, one will reproduce HF-PBE, which may or may not be adequate.[354] In the worst case, describing only some multiconfigurational states with good active spaces may result in an imbalanced treatment and actively worse predictions. How can one be sure that this is not the case? The APC(12,12) active spaces chosen in this work seem to have been sufficient for this application, but further development will be needed for application to larger organic complexes and transition metal systems. Ultimately, different approaches need to be tested on a wide variety of systems and investigated on a case-by-case basis to be trusted.

Secondly, the active space dependence of MC-PDFT may be larger than is comfortable in some sensitive systems. For example, previous work on H_2 dissociation has shown that the predicted dissociation energy of MC-PDFT can vary by over 10 kcal/mol increasing the active space size from a minimal (2,2) to (2,28).[363] Nevertheless, this work has shown that cases such as this are more likely to be outliers than the norm; H_2 dissociation is a well-known failing of restricted HF and DFT, and thus the active space likely has an outsized impact on the performance of MC-PDFT in this case. The generally active-space-independent nature of APC-PDFT beyond a minimum size is further shown by recent studies calculating vertical excitation energies.[6]

Regardless of these remaining challenges, the throughput, automation, and robustness achieved here represent a milestone in applying multiconfigurational methods to main group reactivity and suggest further general-use implementations are possible. The next frontier

involves extending this approach to encompass full reaction networks and larger compounds, promising a more comprehensive understanding of complex chemical processes.

This work is supported by the National Science Foundation under grant CHE-2054723. The work performed by L. O. and B.M.S was made possible by the Office of Naval Research (ONR) through support provided by the Energetic Materials Program (MURI grant number: N00014-21-1-2476, Program Manager: Dr. Chad Stoltz). We thank the Research Computing Center (RCC) at the University of Chicago for computational resources. The authors also acknowledge Mitchell Haselow for participating in project planning discussions.

Converged CI vector, molecular orbital coefficients, and energies for all reactions can be found at: https://doi.org/10.5281/zenodo.10265717.

CHAPTER 8

CONCLUSION

This thesis has undertaken the study of models for determining the active space for multiconfigurational calculations, most principally for the complete active space self-consistent field (CASSCF) method. Despite the somewhat opaque nature of the problem, I hope to have shown in the preceding chapters that progress in this field is possible, and that applications of these methods are vital to advancing our understanding of quantum chemistry.

The first three chapters have concerned the development of methods for selecting the active space, most principally the approximate pair coefficient (APC) method, which was developed from a minimal (2,2) model of the active space to estimate the multiconfigurational character of orbitals from Hartree-Fock matrix elements prior to calculation. Chapter 2 discussed the failures of prior schemes to select the active space, mainly the necessity of further computations or the inflexibility of the active space size, and concluded with the development of the APC method, which proved to be generalizable to a large number of different types of orbitals. Chapter 3 showed the first large-scale application of the APC method, in which vertical excitation energies were calculated for the large and diverse QUESTDB dataset, allowing for the first automated large-scale benchmarking of multiconfigurational approaches. Finally, chapter 4 proposed a variational extension of the APC appraoch in which multiple active spaces are considered and chosen between depending on their energy calculated from multiconfigurational pair-density functional theory (MC-PDFT).

The final three chapters have concerned the applications of automated multiconfigurational methods. Chapter 5 discussed the development of what to the best of my knowledge is the first machine-learned functional trained on automated multiconfigurational data, developing a new series of functionals for use in MC-PDFT. Chapter 6 discussed a detailed computational study on the coupling of different substrates by Cu(II) imines, and employed automated multiconfigurational calculations to elucidate differences in the two mechanisms.

Finally, chapter 7 concerned the application of the APC method to a large dataset of algorithmically generated transition states, showing that it can provide more accurate results than standard approaches in many cases. I believe this last work is likely to be the most impactful on the field and hope to see more studies in this direction in the near-future.

However, despite the remarkable success of the existing schemes, much remains to be done. Future work will likely focus on developing better methods for the selection of active spaces for transition metal systems, and integrating active space selection schemes with non-single-point applications such as dynamics. I hope that this work can serve as good inspiration for future efforts.

REFERENCES

[1] Bao, J. J.; Truhlar, D. G. Automatic Active Space Selection for Calculating Electronic Excitation Energies Based on High-Spin Unrestricted Hartree–Fock Orbitals. *J. Chem. Theory Comput.* **2019**, *15*, 5308–5318.

[2] Welch, B. L. The Generalization Of 'Student's' Problem When Several Different Population Variances Are Involved. *Biometrika* **1947**, *34*, 28–35.

[3] Tishchenko, O.; Zheng, J.; Truhlar, D. G. Multireference Model Chemistries for Thermochemical Kinetics. *J. Chem. Theory Comput.* **2008**, *4*, 1208–1219.

[4] Parrish, R. M.; Burns, L. A.; Smith, D. G.; Simmonett, A. C.; DePrince III, A. E.; Hohenstein, E. G.; Bozkaya, U.; Sokolov, A. Y.; Di Remigio, R.; Richard, R. M.; others Psi4 1.1: An Open-Source Electronic Structure Program Emphasizing Automation, Advanced Libraries, and Interoperability. *J. Chem. Theory Comput.* **2017**, *13*, 3185.

[5] King, D. S.; Gagliardi, L. A Ranked-Orbital Approach to Select Active Spaces for High-Throughput Multireference Computation. *J. Chem. Theory Comput.* **2021**, *17*, 2817–2831.

[6] King, D. S.; Hermes, M. R.; Truhlar, D. G.; Gagliardi, L. Large-Scale Benchmarking of Multireference Vertical-Excitation Calculations via Automated Active-Space Selection. *J. Chem. Theory Comput.* **2022**, *18*, 6065–6076.

[7] Tóth, Z.; Pulay, P. Finding Symmetry Breaking Hartree-Fock Solutions: The Case of Triplet Instability. *J. Chem. Phys.* **2016**, *145*, 164102.

[8] Hirshfeld, F. L. Bonded-Atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta.* **1977**, *44*, 129–138.

[9] Mitchell, E. C.; Scott, T. R.; Bao, J. J.; Truhlar, D. G. Application of Multiconfiguration Pair-Density Functional Theory to the Diels–Alder Reaction. *J. Phys. Chem. A* **2022**, *126*, 8834–8843.

[10] Lischka, H.; Ventura, E.; Dallos, M. The Diels–Alder Reaction of Ethene and 1,3-Butadiene: An Extended Multireference Ab Initio Investigation. *ChemPhysChem* **2004**, *5*, 1365–1371.

[11] Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)evolution. *ACS Cent. Sci.* **2018**, *4*, 144–152.

[12] Slater, J. C. The Theory of Complex Spectra. *Phys. Rev.* **1929**, *34*, 1293.

[13] Hartree, D. R.; Hartree, W. Self-Consistent Field, With Exchange, for Beryllium. *Proc. R. Soc. Lond.* **1935**, *150*, 9–33.

[14] Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.

[15] Izsák, R.; Ivanov, A. V.; Blunt, N. S.; Holzmann, N.; Neese, F. Measuring Electron Correlation: The Impact of Symmetry and Orbital Transformations. *J. Chem. Theory Comput.* **2023**, *19*, 2703–2720.

[16] Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864.

[17] Prodan, E.; Kohn, W. Nearsightedness of Electronic Matter. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 11635–11638.

[18] Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

[19] Becke, A. D. Density-Functional Thermochemistry. I. The Effect of the Exchange-Only Gradient Correction. *J. Chem. Phys.* **1992**, *96*, 2155–2160.

[20] Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods*; Cambridge university press, 2020.

[21] Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Noncovalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-Class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

[22] Chai, J.-D.; Head-Gordon, M. Systematic Optimization of Long-Range Corrected Hybrid Density Functionals. *J. Chem. Phys.* **2008**, *128*.

[23] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*.

[24] Mardirossian, N.; Head-Gordon, M. Thirty Years of Density Functional Theory in Computational Chemistry: An Overview and Extensive Assessment of 200 Density Functionals. *Mol. Phys.* **2017**, *115*, 2315–2372.

[25] Goerigk, L.; Grimme, S. Double-Hybrid Density Functionals. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 576–600.

[26] Gaggioli, C. A.; Stoneburner, S. J.; Cramer, C. J.; Gagliardi, L. Beyond Density Functional Theory: The Multiconfigurational Approach To Model Heterogeneous Catalysis. *ACS Catal.* **2019**, *9*, 8481–8502.

[27] Li Manni, G.; Smart, S. D.; Alavi, A. Combining the Complete Active Space Self-Consistent Field Method and the Full Configuration Interaction Quantum Monte Carlo Within a Super-Ci Framework, With Application to Challenging Metal-Porphyrins. *J. Chem. Theory Comput.* **2016**, *12*, 1245–1258.

[28] Wouters, S.; Van Neck, D. The Density Matrix Renormalization Group for Ab Initio Quantum Chemistry. *Eur. Phys. J. D* **2014**, *68*, 272.

[29] Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. M. A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-Ci Approach. *Chem. Phys.* **1980**, *48*, 157–173.

[30] Veryazov, V.; Malmqvist, P. ; Roos, B. O. How to Select Active Space for Multiconfigurational Quantum Chemistry? *Int. J. Quantum Chem.* **2011**, *111*, 3329–3338.

[31] Stein, C. J.; Reiher, M. autoCAS: A Program for Fully Automated Multiconfigurational Calculations. *J. Comput. Chem.* **2019**, *40*, 2216–2226.

[32] Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-Order Perturbation Theory With a CASSCF Reference Function. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

[33] Andersson, K.; Malmqvist, P.; Roos, B. O. Second-order Perturbation Theory With a Complete Active Space Self-consistent Field Reference Function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

[34] Dyall, K. G. The Choice of a Zeroth-order Hamiltonian for Second-order Perturbation Theory With a Complete Active Space Self-consistent-field Reference Function. *J. Chem. Phys.* **1995**, *102*, 4909–4918.

[35] Angeli, C.; Cimiraglia, R.; Evangelisti, S.; Leininger, T.; Malrieu, J. P. Introduction of N-Electron Valence States for Multireference Perturbation Theory. *J. Chem. Phys.* **2001**, *114*, 10252–10264.

[36] Knowles, P. J.; Werner, H.-J. An Efficient Method for the Evaluation of Coupling Coefficients in Configuration Interaction Calculations. *Chem. Phys. Lett.* **1988**, *145*, 514–522.

[37] Lischka, H.; Shepard, R.; Brown, F. B.; Shavitt, I. New Implementation of the Graphical Unitary Group Approach for Multireference Direct Configuration Interaction Calculations. *Int. J. Quantum Chem.* **1981**, *20*, 91–100.

[38] Kaldor, U.; Roszak, S.; Hariharan, P. C.; Kaufman, J. J. Multireference Coupled Cluster and Multireference Configuration Interaction Studies of the Potential Surfaces for Deprotonation of NH+4. *J. Chem. Phys.* **1989**, *90*, 6395–6400.

[39] Oliphant, N.; Adamowicz, L. Multireference Coupled-cluster Method Using a Single-reference Formalism. *J. Chem. Phys.* **1991**, *94*, 1229–1235.

[40] Mahapatra, U. S.; Datta, B.; Mukherjee, D. A Size-Consistent State-Specific Multireference Coupled Cluster Theory: Formal Developments and Molecular Applications. *J. Chem. Phys.* **1999**, *110*, 6171–6188.

[41] Li Manni, G.; Carlson, R. K.; Luo, S.; Ma, D.; Olsen, J.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2014**, *10*, 3669–3680.

[42] Zhang, D.; Truhlar, D. G. An Accurate Density Coherence Functional for Hybrid Multiconfiguration Density Coherence Functional Theory. *J. Chem. Theory Comput.* **2023**, *19*, 6551–6556.

[43] Zhou, C.; Hermes, M.; Wu, D.; Bao, J. J.; Pandharkar, R.; King, D. R.; Zhang, D.; Scott, T.; Lykhin, A.; Gagliardi, L.; Truhlar, D. G. Electronic Structure of Strongly Correlated Systems: Recent Developments in Multiconfiguration Pair-Density Functional Theory and Multiconfiguration Nonclassical-Energy Functional Theory. *Chem. Sci.* **2022**, *13*, 7685–7706.

[44] Loos, P.-F.; Scemama, A.; Blondel, A.; Garniron, Y.; Caffarel, M.; Jacquemin, D. A Mountaineering Strategy to Excited States: Highly Accurate Reference Energies and Benchmarks. *J. Chem. Theory Comput.* **2018**, *14*, 4360–4379.

[45] Loos, P.-F.; Scemama, A.; Jacquemin, D. The Quest for Highly Accurate Excitation Energies: A Computational Perspective. *J. Phys. Chem. Lett.* **2020**, *11*, 2374–2383.

[46] King, D. S.; Truhlar, D. G.; Gagliardi, L. Variational Active Space Selection With Multiconfiguration Pair-Density Functional Theory. `https://doi.org/10.5281/zenodo.8157623`, 2023.

[47] Hennefarth, M. R.; King, D. S.; Gagliardi, L. Linearized Pair-Density Functional Theory for Vertical Excitation Energies. *J. Chem. Theory Comput.* **2023**, *19*, 7983–7988.

[48] King, D. S.; Truhlar, D. G.; Gagliardi, L. Machine-Learned Energy Functionals for Multiconfigurational Wave Functions. *J. Phys. Chem. Lett.* **2021**, *12*, 7761–7767.

[49] Sayfutyarova, E. R.; Sun, Q.; Chan, G. K.-L.; Knizia, G. Automated Construction of Molecular Active Spaces From Atomic Valence Orbitals. *J. Chem. Theory Comput.* **2017**, *13*, 4063–4078.

[50] McCullough, K. E.; King, D. S.; Chheda, S. P.; Ferrandon, M. S.; Goetjen, T. A.; Syed, Z. H.; Graham, T. R.; Washton, N. M.; Farha, O. K.; Gagliardi, L.; others High-Throughput Experimentation, Theoretical Modeling, and Human Intuition: Lessons Learned in Metal–Organic-Framework-Supported Catalyst Design. *ACS Cent. Sci.* **2023**, *9*, 266–276.

[51] Dohrmann, N.; King, D. S.; Gaggioli, C. A.; Gagliardi, L. Challenge of Small Energy Differences in Metal–Organic Framework Reactivity. *J. Phys. Chem. C* **2023**, *127*, 16891–16900.

[52] Tran, K.; Ulissi, Z. W. Active Learning Across Intermetallics to Guide Discovery of Electrocatalysts for CO 2 Reduction and H 2 Evolution. *Nat. Catal.* **2018**, *1*, 696–703.

[53] Thornton, A. W.; Simon, C. M.; Kim, J.; Kwon, O.; Deeg, K. S.; Konstas, K.; Pas, S. J.; Hill, M. R.; Winkler, D. A.; Haranczyk, M.; Smit, B. Materials Genome in Action: Identifying the Performance Limits of Physical Hydrogen Storage. *Chem. Mater.* **2017**, *29*, 2844–2854.

[54] Teunissen, J. L.; De Proft, F.; De Vleeschouwer, F. Tuning the HOMO–LUMO Energy Gap of Small Diamondoids Using Inverse Molecular Design. *J. Chem. Theory Comput.* **2017**, *13*, 1351–1365.

[55] Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *J. Phys. Chem. Lett.* **2013**, *4*, 1613–1623.

[56] Shu, Y.; Levine, B. G. Simulated Evolution of Fluorophores for Light Emitting Diodes. *J. Chem. Phys.* **2015**, *142*, 104104.

[57] Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354–2377.

[58] Vogiatzis, K. D.; Polynski, M. V.; Kirkland, J. K.; Townsend, J.; Hashemi, A.; Liu, C.; Pidko, E. A. Computational Approach to Molecular Catalysis by 3d Transition Metals: Challenges and Opportunities. *Chem. Rev.* **2019**, *119*, 2453–2523.

[59] Cramer, C. J.; Truhlar, D. G. Density Functional Theory for Transition Metals and Transition Metal Chemistry. *Phys. Chem. Chem. Phys* **2009**, *11*, 10757.

[60] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chem. Rev.* **2012**, *112*, 289–320.

[61] Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham Density Functional Theory Descending a Staircase. *J. Chem. Phys.* **2016**, *145*, 130901.

[62] Becke, A. D. Perspective: Fifty Years of Density-Functional Theory in Chemical Physics. *J. Chem. Phys.* **2014**, *140*, 18A301.

[63] Lischka, H.; Nachtigallová, D.; Aquino, A. J. A.; Szalay, P. G.; Plasser, F.; Machado, F. B. C.; Barbatti, M. Multireference Approaches for Excited States of Molecules. *Chem. Rev.* **2018**, *118*, 7293–7361.

[64] Ashley, D. C.; Jakubikova, E. Ironing Out the Photochemical and Spin-Crossover Behavior of Fe(II) Coordination Compounds With Computational Chemistry. *Coord. Chem. Rev.* **2017**, *337*, 97–111.

[65] Janet, J. P.; Chan, L.; Kulik, H. J. Accelerating Chemical Discovery With Machine Learning: Simulated Evolution of Spin Crossover Complexes With an Artificial Neural Network. *J. Phys. Chem. Lett.* **2018**, *9*, 1064–1071.

[66] Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J. Semi-Supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost. *J. Phys. Chem. Lett.* **2020**, *11*, 6640–6648.

[67] Aquilante, F. et al. Molcas 8: New Capabilities for Multiconfigurational Quantum Chemical Calculations Across the Periodic Table. *J. Comput. Chem.* **2016**, *37*, 506–541.

[68] Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. A Complete Active Space SCF Method (CASSCF) Using a Density Matrix Formulated Super-Ci Approach. *Chem. Phys.* **1980**, *48*, 157–173.

[69] White, S. R. Density Matrix Formulation for Quantum Renormalization Groups. *Phys. Rev. Lett.* **1992**, *69*, 2863–2866.

[70] White, S. R. Density-Matrix Algorithms for Quantum Renormalization Groups. *Phys. Rev. B* **1993**, *48*, 10345–10356.

[71] Keller, S.; Dolfi, M.; Troyer, M.; Reiher, M. An Efficient Matrix Product Operator Representation of the Quantum Chemical Hamiltonian. *J. Chem. Phys.* **2015**, *143*, 244118.

[72] Booth, G. H.; Thom, A. J. W.; Alavi, A. Fermion Monte Carlo Without Fixed Nodes: A Game of Life, Death, and Annihilation in Slater Determinant Space. *J. Chem. Phys.* **2009**, *131*, 054106.

[73] Yang, P.-J.; Sugiyama, M.; Tsuda, K.; Yanai, T. Artificial Neural Networks Applied as Molecular Wave Function Solvers. *J. Chem. Theory Comput.* **2020**,

[74] Roos, B. O.; Linse, P.; Siegbahn, P. E.; Blomberg, M. R. A Simple Method for the Evaluation of the Second-Order-Perturbation Energy From External Double-Excitations With a CASSCF Reference Wavefunction. *Chem. Phys.* **1982**, *66*, 197–207.

[75] Bao, J. L.; Sand, A.; Gagliardi, L.; Truhlar, D. G. Correlated-Participating-Orbitals Pair-Density Functional Method and Application to Multiplet Energy Splittings of Main-Group Divalent Radicals. *J. Chem. Theory Comput.* **2016**, *12*, 4274–4283.

[76] Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *J. Chem. Theory Comput.* **2016**, *12*, 1760–1771.

[77] Bao, J. J.; Dong, S. S.; Gagliardi, L.; Truhlar, D. G. Automatic Selection of an Active Space for Calculating Electronic Excitation Spectra by MS-CASPT2 or MC-PDFT. *J. Chem. Theory Comput.* **2018**, *14*, 2017–2025.

[78] Khedkar, A.; Roemelt, M. Active Space Selection Based on Natural Orbital Occupation Numbers From N-Electron Valence Perturbation Theory. *J. Chem. Theory Comput.* **2019**, *15*, 3522–3536.

[79] Sayfutyarova, E. R.; Hammes-Schiffer, S. Constructing Molecular -Orbital Active Spaces for Multireference Calculations of Conjugated Systems. *J Chem Theory Comput* **2019**, *15*, 1679–1689.

[80] Golub, P.; Antalik, A.; Veis, L.; Brabec, J. Machine Learning-Assisted Selection of Active Spaces for Strongly Correlated Transition Metal Systems. *J. Chem. Theory Comput.* **2021**, *17*, 6053–6072.

[81] Jeong, W.; Stoneburner, S. J.; King, D.; Li, R.; Walker, A.; Lindh, R.; Gagliardi, L. Automation of Active Space Selection for Multireference Methods via Machine Learning on Chemical Bond Dissociation. *J. Chem. Theory Comput.* **2020**, *16*, 2389–2399.

[82] Khedkar, A.; Roemelt, M. Extending the ASS1ST Active Space Selection Scheme to Large Molecules and Excited States. *J. Chem. Theory Comput.* **2020**,

[83] Tóth, Z.; Pulay, P. Comparison of Methods for Active Orbital Selection in Multiconfigurational Calculations. *J. Chem. Theory Comput.* **2020**, *16*, 7328–7341.

[84] Zou, J.; Niu, K.; Ma, H.; Li, S.; Fang, W. Automatic Selection of Active Orbitals From Generalized Valence Bond Orbitals. *J. Phys. Chem. A* **2020**, *124*, 8321–8329.

[85] Boguslawski, K.; Tecmer, P. Orbital Entanglement in Quantum Chemistry. *Int. J. Quantum Chem.* **2015**, *115*, 1289–1295.

[86] Stein, C. J.; Reiher, M. Automated Identification of Relevant Frontier Orbitals for Chemical Compounds and Processes. *Chimia (Aarau)* **2017**, *71*, 170–176.

[87] Pulay, P.; Hamilton, T. P. UHF Natural Orbitals for Defining and Starting MC-SCF Calculations. *J. Chem. Phys.* **1988**, *88*, 4926–4933.

[88] Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *J. Chem. Theory Comput.* **2018**, *14*, 4772–4779.

[89] Kendall, R. A.; Dunning Jr, T. H.; Harrison, R. J. Electron Affinities of the First-Row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

[90] Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; others PySCF: The Python-based Simulations of Chemistry Framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1340.

[91] Knowles, P. J.; Handy, N. C. A Determinant Based Full Configuration Interaction Program. *Comput. Phys. Commun.* **1989**, *54*, 75–83.

[92] Legeza, ; Sólyom, J. Optimizing the Density-Matrix Renormalization Group Method Using Quantum Information Entropy. *Phys. Rev. B* **2003**, *68*, 195116.

[93] Barcza, G.; Legeza, ; Marti, K. H.; Reiher, M. Quantum-Information Analysis of Electronic States of Different Molecular Structures. *Phys. Rev. A* **2011**, *83*, 012508.

[94] Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Rev. Mod. Phys.* **1960**, *32*, 300–302.

[95] Pipek, J.; Mezey, P. G. A Fast Intrinsic Localization Procedure Applicable for a b i n i t i o and Semiempirical Linear Combination of Atomic Orbital Wave Functions. *J. Chem. Phys.* **1989**, *90*, 4916–4926.

[96] Lehtola, S.; Jónsson, H. Pipek–Mezey Orbital Localization Using Various Partial Charge Estimates. *J. Chem. Theory Comput.* **2014**, *10*, 642–649.

[97] Edmiston, C.; Ruedenberg, K. Localized Atomic and Molecular Orbitals. *Rev. Mod. Phys.* **1963**, *35*, 457–464.

[98] Waskom, M.; the seaborn development team Mwaskom/Seaborn. 2020; `https://doi.org/10.5281/zenodo.592845`.

[99] Knizia, G.; Klein, J. E. Electron Flow in Reaction Mechanisms—revealed From First Principles. *Angew. Chem. Int. Ed.* **2015**, *54*, 5518–5522.

[100] Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. O. Density Matrix Averaged Atomic Natural Orbital (ANO) Basis Sets for Correlated Molecular Wave Functions. *Theor. Chim. Acta.* **1995**, *90*, 87–114.

[101] Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. Density Matrix Averaged Atomic Natural Orbital (ANO) Basis Sets for Correlated Molecular Wave Functions. *Theor. Chim. Acta.* **1990**, *77*, 291–306.

[102] Roos, B. O.; Lindh, R.; Malmqvist, P.-Å.; Veryazov, V.; Widmark, P.-O. Main Group Atoms and Dimers Studied With a New Relativistic ANO Basis Set. *J. Phys. Chem. A* **2004**, *108*, 2851–2858.

[103] Stein, C. J.; Reiher, M. Measuring Multi-Configurational Character by Orbital Entanglement. *Mol. Phys.* **2017**, *115*.

[104] Knizia, G. Intrinsic Atomic Orbitals: An Unbiased Bridge Between Quantum Theory and Chemical Concepts. *J. Chem. Theory Comput.* **2013**, *9*, 4834–4843.

[105] Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Courier Corporation, 2012.

[106] Liu, T.-Y. *Learning to Rank for Information Retrieval*; Springer Science & Business Media, 2011.

[107] Cheng, L.; Welborn, M.; Christensen, A. S.; Miller, T. F. A Universal Density Matrix Functional From Molecular Orbital-Based Machine Learning: Transferability Across Organic Molecules. *J. Chem. Phys.* **2019**, *150*, 131103.

[108] Townsend, J.; Vogiatzis, K. D. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *J. Phys. Chem. Lett.* **2019**, *10*, 4129–4135.

[109] Bernardi, F.; Olivucci, M.; Robb, M. A. Potential Energy Surface Crossings in Organic Photochemistry. *Chem. Soc. Rev.* **1996**, *25*, 321.

[110] Olivucci, M., Ed. *Computational Photochemistry*; Elsevier: Amsterdam, 2005.

[111] Navizet, I.; Liu, Y.-J.; Ferré, N.; Roca-Sanjuán, D.; Lindh, R. The Chemistry of Bioluminescence: An Analysis of Chemical Functionalities. *ChemPhysChem* **2011**, *12*, 3064–3076.

[112] Crespo-Otero, R.; Barbatti, M. Recent Advances and Perspectives on Nonadiabatic Mixed Quantum–Classical Dynamics. *Chem. Rev.* **2018**, *118*, 7026–7068.

[113] Dmitri, K., Svetlana, K., Yulun, H., Eds. *Computational Photocatalysis: Modeling of Photophysics and Photochemistry at Interfaces*; American Chemical Society: Washington, 2019.

[114] Segatta, F.; Cupellini, L.; Garavelli, M.; Mennucci, B. Quantum Chemical Modeling of the Photoinduced Activity of Multichromophoric Biosystems. *Chem. Rev.* **2019**, *119*, 9361.

[115] Mai, S.; González, L. Molecular Photochemistry: Recent Developments in Theory. *Angew. Chem. Int. Ed.* **2020**, *59*, 16832–16846.

[116] Piecuch, P.; Kowalski, K.; Pimienta, I. S. O.; Mcguire, M. J. Recent Advances in Electronic Structure Theory: Method of Moments of Coupled-Cluster Equations and Renormalized Coupled-Cluster Approaches. *Int. Rev. Phys. Chem.* **2002**, *21*, 527–655.

[117] Krylov, A. I. Spin-Flip Equation-Of-Motion Coupled-Cluster Electronic Structure Method for a Description of Excited States, Bond Breaking, Diradicals, and Triradicals. *Acc. Chem. Res.* **2006**, *39*, 83–91.

[118] González, L.; Escudero, D.; Serrano-Andrés, L. Progress and Challenges in the Calculation of Electronic Excited States. *ChemPhysChem* **2012**, *13*, 28–51.

[119] Sneskov, K.; Christiansen, O. Excited State Coupled Cluster Methods: Excited State Coupled Cluster Methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 566–584.

[120] Adamo, C.; Jacquemin, D. The Calculations of Excited-State Properties With Time-Dependent Density Functional Theory. *Chem. Soc. Rev.* **2013**, *42*, 845–856.

[121] Laurent, A. D.; Jacquemin, D. TD-DFT Benchmarks: A Review. *Int. J. Quantum Chem.* **2013**, *113*, 2019–2039.

[122] Faber, C.; Boulanger, P.; Attaccalite, C.; Duchemin, I.; Blase, X. Excited States Properties of Organic Molecules: From Density Functional Theory to the GW and Bethe-Salpeter Green's Function Formalisms. *Phil. Trans. Roy. Soc. A* **2014**, *372*, 20130271.

[123] Lischka, H.; Nachtigallová, D.; Aquino, A.; MacHado, F.; Barbatti, M. Multireference Approaches for Excited States of Molecules. *Chem. Rev.* **2018**, *118*, 7293.

[124] Ghosh, S.; Verma, P.; Cramer, C. J.; Gagliardi, L.; Truhlar, D. G. Combining Wave Function Methods With Density Functional Theory for Excited States. *Chem. Rev.* **2018**, *118*, 7249–7292.

[125] Blase, X.; Duchemin, I.; Jacquemin, D.; Loos, P.-F. The Bethe–Salpeter Equation Formalism: From Physics to Chemistry. *J. Phys. Chem. Lett.* **2020**, *11*, 7371–7382.

[126] Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2021**, *121*, 9873.

[127] Dral, P.; Barbatti, M. Molecular Excited States Through a Machine Learning Lens. *Nature Rev. Chem.* **2021**, *5*, 388.

[128] Loos, P.-F.; Boggio-Pasqua, M.; Scemama, A.; Caffarel, M.; Jacquemin, D. Reference Energies for Double Excitations. *J. Chem. Theory Comput.* **2019**, *15*, 1939–1956.

[129] Vancoillie, S.; Zhao, H.; Tran, V. T.; Hendrickx, M. F. A.; Pierloot, K. Multiconfigurational Second-Order Perturbation Theory Restricted Active Space (RASPT2) Studies on Mononuclear First-Row Transition-Metal Systems. *J. Chem. Theory Comput.* **2011**, *7*, 3961–3977.

[130] Zhou, C.; Gagliardi, L.; Truhlar, D. G. Multiconfiguration Pair-Density Functional Theory for Iron Porphyrin With CAS, RAS, and DMRG Active Spaces. *J. Phys. Chem. A* **2019**, *123*, 3389–3394.

[131] Blunt, N. S.; Mahajan, A.; Sharma, S. Efficient Multireference Perturbation Theory Without High-Order Reduced Density Matrices. *J. Chem. Phys.* **2020**, *153*, 164120.

[132] Beran, P.; Matoušek, M.; Hapka, M.; Pernal, K.; Veis, L. Density Matrix Renormalization Group With Dynamical Correlation via Adiabatic Connection. *J. Chem. Theory Comput.* **2021**, *17*, 7575–7585.

[133] Nakano, H. Quasidegenerate Perturbation Theory With Multiconfigurational Self-Consistent-Field Reference Functions. *J. Chem. Phys.* **1993**, *99*, 7983.

[134] Nakano, H. McSCF Reference Quasidegenerate Perturbation Theory With Epstein–Nesbet Partitioning. *Chem. Phys. Lett.* **1993**, *207*, 372.

[135] Andersson, K.; Malmqvist, P.-Å.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-Order Perturbation Theory With a CASSCF Reference Function. *J. Phys. Chem.* **1990**, *94*, 5483.

[136] Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. Second-order Perturbation Theory With a Complete Active Space Self-consistent Field Reference Function. *J. Chem. Phys.* **1992**, *96*, 1218.

[137] Hoyer, C. E.; Ghosh, S.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory Is as Accurate as CASPT2 for Electronic Excitation. *J. Phys. Chem. Lett.* **2016**, *7*, 586–591.

[138] Gagliardi, L.; Truhlar, D. G.; Li Manni, G.; Carlson, R. K.; Hoyer, C. E.; Bao, J. L. Multiconfiguration Pair-Density Functional Theory: A New Way To Treat Strongly Correlated Systems. *Acc. Chem. Res.* **2017**, *50*, 66–73.

[139] Zhang, D.; Hermes, M. R.; Gagliardi, L.; Truhlar, D. G. Multiconfiguration Density-Coherence Functional Theory. *J. Chem. Theory Comput.* **2021**, *17*, 2775–2782.

[140] Pandharkar, R.; Hermes, M. R.; Truhlar, D. G.; Gagliardi, L. A New Mixing of Nonlocal Exchange and Nonlocal Correlation With Multiconfiguration Pair-Density Functional Theory. *J. Phys. Chem. Lett.* **2020**, *11*, 10158–10163.

[141] Bao, J. L.; Odoh, S. O.; Gagliardi, L.; Truhlar, D. G. Predicting Bond Dissociation Energies of Transition-Metal Compounds by Multiconfiguration Pair-Density Functional Theory and Second-Order Perturbation Theory Based on Correlated Participating Orbitals and Separated Pairs. *J. Chem. Theory Comput.* **2017**, *13*, 616–626.

[142] Li, S. J.; Gagliardi, L.; Truhlar, D. G. Extended Separated-Pair Approximation for Transition Metal Potential Energy Curves. *J. Chem. Phys.* **2020**, *152*, 124118.

[143] Golub, P.; Antalik, A.; Veis, L.; Brabec, J. Machine Learning-Assisted Selection of Active Spaces for Strongly Correlated Transition Metal Systems. *J. Chem. Theory Comput.* **2021**, *17*, 6053–6072.

[144] Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron Affinities of the First-row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796.

[145] Woon, D. E.; Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum Through Argon. *J. Chem. Phys.* **1993**, *98*, 1358.

[146] Loos, P.-F.; Scemama, A.; Boggio-Pasqua, M.; Jacquemin, D. Mountaineering Strategy to Excited States: Highly Accurate Energies and Benchmarks for Exotic Molecules and Radicals. *J. Chem. Theory Comput.* **2020**, *16*, 3720–3736.

[147] Loos, P.-F.; Lipparini, F.; Boggio-Pasqua, M.; Scemama, A.; Jacquemin, D. A Mountaineering Strategy to Excited States: Highly Accurate Energies and Benchmarks for Medium Sized Molecules. *J. Chem. Theory Comput.* **2020**, *16*, 1711–1741.

[148] Véril, M.; Scemama, A.; Caffarel, M.; Lipparini, F.; Boggio-Pasqua, M.; Jacquemin, D.; Loos, P. QUESTDB: A Database of Highly Accurate Excitation Energies for the Electronic Structure Community. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*.

[149] MRH. `https://github.com/MatthewRHermes/mrh`, (accessed 2022-01-18).

[150] McKinney, W. Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. 2010; p 56.

[151] pandas development team, T. Pandas-Dev/Pandas: Pandas. 2020; `https://doi.org/10.5281/zenodo.3509134`.

[152] Efron, B. *Breakthroughs in Statistics*; Springer: New York, 1992; p 569.

[153] Tibshirani, R. J.; Efron, B. An Introduction to the Bootstrap. *Monogr. Statist. Appl. Probab.* **1993**, *57*, 1.

[154] Mooney, C. Z.; Mooney, C. F.; Mooney, C. L.; Duval, R. D.; Duvall, R. *Bootstrapping: A Nonparametric Approach to Statistical Inference*; Sage: Newbury Park, CA, 1993.

[155] Papajak, E.; Zheng, J.; Xu, X.; Leverentz, H. R.; Truhlar, D. G. Perspectives on Basis Sets Beautiful: Seasonal Plantings of Diffuse Basis Functions. *J. Chem. Theory Comput.* **2011**, *7*, 3027–3034.

[156] Dunning, T. H., Jr. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007.

[157] Peterson, K. A.; Dunning, T. H. Accurate Correlation Consistent Basis Sets for Molecular Core–valence Correlation Effects: The Second Row Atoms Al–Ar, and the First Row Atoms B–Ne Revisited. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

[158] Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. Basis-Set Convergence in Correlated Calculations on Ne, N_2, and H_2O. *Chem. Phys. Lett.* **1998**, *286*, 243.

[159] Schapiro, I.; Sivalingam, K.; Neese, F. Assessment of N-Electron Valence State Perturbation Theory for Vertical Excitation Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3567–3580.

[160] Sarkar, R.; Loos, P.-F.; Boggio-Pasqua, M.; Jacquemin, D. Assessing the Performances of CASPT2 and NEVPT2 for Vertical Excitation Energies. *arXiv:2111.15386 [cond-mat, physics:physics]* **2021**, arXiv: 2111.15386.

[161] Loos, P.-F.; Jacquemin, D. Is ADC(3) as Accurate as CC3 for Valence and Rydberg Transition Energies? *J. Phys. Chem. Lett.* **2020**, *11*, 974–980.

[162] Sand, A. M.; Truhlar, D. G.; Gagliardi, L. Efficient Algorithm for Multiconfiguration Pair-Density Functional Theory With Application to the Heterolytic Dissociation Energy of Ferrocene. *J. Chem. Phys.* **2017**, *146*, 034101.

[163] Loos, P.-F.; Lipparini, F.; Boggio-Pasqua, M.; Scemama, A.; Jacquemin, D. A Mountaineering Strategy to Excited States: Highly Accurate Energies and Benchmarks for Medium Sized Molecules. *J. Chem. Theory Comput.* **2020**, *16*, 1711–1741.

[164] Pernal, K. Exact and Approximate Adiabatic Connection Formulae for the Correlation Energy in Multireference Ground and Excited States. *J. Chem. Phys.* **2018**, *149*, 204101.

[165] Chatterjee, K.; Sokolov, A. Y. Extended Second-Order Multireference Algebraic Diagrammatic Construction Theory for Charged Excitations. *J. Chem. Theory Comput.* **2020**, *16*, 6343–6357.

[166] Roemelt, M.; Krewald, V.; Pantazis, D. A. Exchange Coupling Interactions From the Density Matrix Renormalization Group and N-Electron Valence Perturbation Theory: Application to a Biomimetic Mixed-Valence Manganese Complex. *J. Chem. Theory Comput.* **2018**, *14*, 166–179.

[167] Levine, B. G.; Esch, M. P.; Fales, B. S.; Hardwick, D. T.; Peng, W.-T.; Shu, Y. Conical Intersections at the Nanoscale: Molecular Ideas for Materials. *Annu. Rev. Phys. Chem.* **2019**, *70*, 21–43.

[168] Izsák, R. Single-Reference Coupled Cluster Methods for Computing Excitation Energies in Large Molecules: The Efficiency and Accuracy of Approximations. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, *10*, e1445.

[169] Eriksen, J. J. The Shape of Full Configuration Interaction to Come. *J. Phys. Chem. Lett.* **2020**, *12*, 418–432.

[170] Casanova, D. Restricted Active Space Configuration Interaction Methods for Strong Correlation: Recent Developments. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, e1561.

[171] Shu, Y.; Truhlar, D. G. Doubly Excited Character or Static Correlation of the Reference State in the Controversial 2^1A_g State of Trans-Butadiene? *J. Am. Chem. Soc.* **2017**, *139*, 13770–13778.

[172] Jacquemin, D.; Zhao, Y.; Valero, R.; Adamo, C.; Ciofini, I.; Truhlar, D. G. Verdict: Time-Dependent Density Functional Theory "Not Guilty" of Large Errors for Cyanines. *J. Chem. Theory Comput.* **2012**, *8*, 1255–1259.

[173] Eriksen, J. J. The Shape of Full Configuration Interaction to Come. *J. Phys. Chem. Lett.* **2021**, *12*, 418–432.

[174] White, S. R.; Martin, R. L. Ab Initio Quantum Chemistry Using the Density Matrix Renormalization Group. *J. Chem. Phys.* **1999**, *110*, 4127–4130.

[175] Chan, G. K.-L.; Sharma, S. The Density Matrix Renormalization Group in Quantum Chemistry. *Annu. Rev. Phys. Chem.* **2011**, *62*, 465–481.

[176] Schollwöck, U. The Density-Matrix Renormalization Group in the Age of Matrix Product States. *Ann. Phys. (N. Y.)* **2011**, *326*, 96–192.

[177] Marti, K. H.; Reiher, M. New Electron Correlation Theories for Transition Metal Chemistry. *Phys. Chem. Chem. Phys* **2011**, *13*, 6750–6759.

[178] Olivares-Amaya, R.; Hu, W.; Nakatani, N.; Sharma, S.; Yang, J.; Chan, G. K.-L. The Ab-Initio Density Matrix Renormalization Group in Practice. *J. Chem. Phys.* **2015**, *142*, 034102.

[179] Szalay, S.; Pfeffer, M.; Murg, V.; Barcza, G.; Verstraete, F.; Schneider, R.; Legeza, Tensor Product Methods and Entanglement Optimization for Ab Initio Quantum Chemistry. *Int. J. Quantum Chem.* **2015**, *115*, 1342–1391.

[180] Yanai, T.; Kurashige, Y.; Mizukami, W.; Chalupskỳ, J.; Lan, T. N.; Saitow, M. Density Matrix Renormalization Group for Ab Initio Calculations and Associated Dynamic Correlation Methods: A Review of Theory and Applications. *Int. J. Quantum Chem.* **2015**, *115*, 283–299.

[181] Chan, G. K.-L.; Keselman, A.; Nakatani, N.; Li, Z.; White, S. R. Matrix Product Operators, Matrix Product States, and Ab Initio Density Matrix Renormalization Group Algorithms. *J. Chem. Phys.* **2016**, *145*, 014102.

[182] Knecht, S.; Hedegård, E. D.; Keller, S.; Kovyrshin, A.; Ma, Y.; Muolo, A.; Stein, C. J.; Reiher, M. New Approaches for Ab Initio Calculations of Molecules With Strong Electron Correlation. *Chimia (Aarau)* **2016**, *70*, 244–251.

[183] Baiardi, A.; Reiher, M. The Density Matrix Renormalization Group in Chemistry and Molecular Physics: Recent Developments and New Challenges. *J. Chem. Phys.* **2020**, *152*, 040903.

[184] Zgid, D.; Nooijen, M. The Density Matrix Renormalization Group Self-Consistent Field Method: Orbital Optimization with the Density Matrix Renormalization Group Method in the Active Space. *J. Chem. Phys.* **2008**, *128*, 144116.

[185] Ghosh, D.; Hachmann, J.; Yanai, T.; Chan, G. K.-L. Orbital Optimization in the Density Matrix Renormalization Group, With Applications to Polyenes and -Carotene. *J. Chem. Phys.* **2008**, *128*, 144117.

[186] Ruedenberg, K.; Cheung, L. M.; Elbert, S. T. McSCF Optimization Through Combined Use of Natural Orbitals and the Brillouin-Levy-Berthier Theorem. *Int. J. Quantum Chem.* **1979**, *16*, 1069–1101.

[187] Freitag, L.; Ma, Y.; Baiardi, A.; Knecht, S.; Reiher, M. Approximate Analytical Gradients and Nonadiabatic Couplings for the State-Average Density Matrix Renormalization Group Self-Consistent-Field Method. *J. Chem. Theory Comput.* **2019**, *15*, 6724–6737.

[188] Roos, B. O. In *Computational Photochemisty*; Olivucci, M., Ed.; Elsevier: Amsterdam, 2005; pp 317–348.

[189] Bofill, J. M.; Pulay, P. The Unrestricted Natural Orbital–complete Active Space (UNO–CAS) Method: An Inexpensive Alternative to the Complete Active Space–self-consistent-field (CAS–SCF) Method. *J. Chem. Phys.* **1989**, *90*, 3637–3646.

[190] Schmidt, M. W.; Gordon, M. S. The Construction and Interpretation of MCSCF Wavefunctions. *Annu Rev Phys Chem* **1998**, *49*, 233–266.

[191] Keller, S.; Boguslawski, K.; Janowski, T.; Reiher, M.; Pulay, P. Selection of Active Spaces for Multiconfigurational Wavefunctions. *J Chem Phys* **2015**, *142*, 244104.

[192] Gieseking, R. L. A New Release of MOPAC Incorporating the INDO/S Semiempirical Model With CI Excited States. *J. Comput. Chem.* **2021**, *42*, 365–378.

[193] Khedkar, A.; Roemelt, M. Modern Multireference Methods and Their Application in Transition Metal Chemistry. *Phys. Chem. Chem. Phys.* **2021**, *23*, 17097–17112.

[194] Lei, Y.; Suo, B.; Liu, W. iCAS: Imposed Automatic Selection and Localization of Complete Active Spaces. *J. Chem. Theory Comput.* **2021**, *17*, 4846–4859.

[195] Levine, B. G.; Durden, A. S.; Esch, M. P.; Liang, F.; Shu, Y. CAS Without SCF—Why to Use CASCI and Where to Get the Orbitals. *J. Chem. Phys.* **2021**, *154*, 090902.

[196] Oakley, M. S.; Gagliardi, L.; Truhlar, D. G. Multiconfiguration Pair-Density Functional Theory for Transition Metal Silicide Bond Dissociation Energies, Bond Lengths, and State Orderings. *Molecules* **2021**, *26*, 2881.

[197] Weser, O.; Guther, K.; Ghanem, K.; Li Manni, G. Stochastic Generalized Active Space Self-Consistent Field: Theory and Application. *J. Chem. Theory Comput.* **2021**, *18*, 251–272.

[198] Cheng, Y.; Xie, Z.; Ma, H. Post-Density Matrix Renormalization Group Methods for Describing Dynamic Electron Correlation With Large Active Spaces. *J. Phys. Chem. Lett.* **2022**, *13*, 904–915.

[199] Bensberg, M.; Reiher, M. Corresponding Active Orbital Spaces Along Chemical Reaction Paths. *J. Phys. Chem. Lett.* **2023**, *14*, 2112–2118.

[200] Golub, P.; Antalik, A.; Beran, P.; Brabec, J. Mutual Information Prediction for Strongly Correlated Systems. *Chem. Phys. Lett.* **2023**, 140297.

[201] Kaufold, B. W.; Chintala, N.; Pandeya, P.; Dong, S. S. Automated Active Space Selection With Dipole Moments. *J. Chem. Theory Comput.* **2023**, *19*, 2469–2483.

[202] Serwatka, T.; Roy, P.-N. Optimized Basis Sets for DMRG Calculations of Quantum Chains of Rotating Water Molecules. *J. Chem. Phys.* **2023**, *158*, 214103.

[203] Stein, C. J.; von Burg, V.; Reiher, M. The Delicate Balance of Static and Dynamic Electron Correlation. *J. Chem. Theory Comput.* **2016**, *12*, 3764–3773.

[204] Sarkar, R.; Loos, P.-F.; Boggio-Pasqua, M.; Jacquemin, D. Assessing the Performances of CASPT2 and NEVPT2 for Vertical Excitation Energies. *J. Chem. Theory Comput.* **2022**, *18*, 2418–2436.

[205] Zhai, H.; Chan, G. K.-L. Low Communication High Performance Ab Initio Density Matrix Renormalization Group Algorithms. *J. Chem. Phys.* **2021**, *154*, 224116.

[206] Roothaan, C. New Developments in Molecular Orbital Theory. *Rev. Mod. Phys.* **1951**, *23*, 69–89.

[207] Roothaan, C. C. J. Self-Consistent Field Theory for Open Shells of Electronic Systems. *Rev. Mod. Phys.* **1960**, *32*, 179.

[208] Kendall, R. A.; Dunning, T. H.; Harrison, R. J. Electron Affinities of the First-row Atoms Revisited. Systematic Basis Sets and Wave Functions. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

[209] Woon, D. E.; Dunning, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. III. The Atoms Aluminum Through Argon. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

[210] Tsuchimochi, T.; Scuseria, G. E. Communication: ROHF Theory Made Simple. *J. Chem. Phys.* **2010**, *133*, 141102.

[211] Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L. PySCF: The Python-Based Simulations of Chemistry Framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1340.

[212] Sun, Q. et al. Recent Developments in the PySCF Program Package. *J. Chem. Phys.* **2020**, *153*, 024109.

[213] Pyscf-Forge. `https://github.com/pyscf/pyscf-forge`, (accessed 2023-09-15).

[214] Barbatti, M.; Paier, J.; Lischka, H. Photochemistry of Ethylene: A Multireference Configuration Interaction Investigation of the Excited-State Energy Surfaces. *J. Chem. Phys.* **2004**, *121*, 11614–11624.

[215] Feller, D.; Peterson, K. A.; Davidson, E. R. A Systematic Approach to Vertically Excited States of Ethylene Using Configuration Interaction and Coupled Cluster Techniques. *J. Chem. Phys.* **2014**, *141*, 104302.

[216] Sharma, P.; Truhlar, D. G.; Gagliardi, L. Multiconfiguration Pair-Density Functional Theory Investigation of the Electronic Spectrum of MnO_4. *J. Chem. Phys.* **2018**, *148*, 124305.

[217] Kohn, W.; Becke, A. D.; Parr, R. G. Density Functional Theory of Electronic Structure. *J. Phys. Chem.* **1996**, *100*, 12974–12980.

[218] Scuseria, G. E.; Staroverov, V. N. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 669–724.

[219] Cohen, A. J.; Mori-Sanchez, P.; Yang, W. Insights Into Current Limitations of Density Functional Theory. *Science* **2008**, *321*, 792–794.

[220] Verma, P.; Truhlar, D. G. Status and Challenges of Density Functional Theory. *Trends Chem.* **2020**, *2*, 302–318.

[221] Nakano, H.; Nakajima, T.; Tsuneda, T.; Hirao, K. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 507–557.

[222] Roos, B. O. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 725–764.

[223] Gordon, M. S.; Schmidt, M. W. In *Theory and Applications of Computational Chemistry*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 1167–1189.

[224] Hirao, K. Multireference Møller—Plesset Method. *Chem. Phys. Lett.* **1992**, *190*, 374–380.

[225] Brown, F. B.; Shavitt, I.; Shepard, R. Multireference Configuration Interaction Treatment of Potential Energy Surfaces: Symmetric Dissociation of H2O in a Double-Zeta Basis. *Chem. Phys. Lett.* **1984**, *105*, 363–369.

[226] Werner, H.; Knowles, P. J. An Efficient Internally Contracted Multiconfiguration–reference Configuration Interaction Method. *J. Chem. Phys.* **1988**, *89*, 5803–5814.

[227] Snyder, J. C.; Rupp, M.; Hansen, K.; Müller, K.-R.; Burke, K. Finding Density Functionals With Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 253002.

[228] Snyder, J. C.; Rupp, M.; Hansen, K.; Blooston, L.; Müller, K.-R.; Burke, K. Orbital-Free Bond Breaking via Machine Learning. *J. Chem. Phys.* **2013**, *139*, 224104.

[229] Snyder, J. C.; Rupp, M.; Müller, K.-R.; Burke, K. Nonlinear Gradient Denoising: Finding Accurate Extrema From Inaccurate Functional Derivatives. *Int. J. Quantum Chem.* **2015**, *115*, 1102–1114.

[230] Vu, K.; Snyder, J. C.; Li, L.; Rupp, M.; Chen, B. F.; Khelif, T.; Müller, K.-R.; Burke, K. Understanding Kernel Ridge Regression: Common Behaviors From Simple Functions to Density Functionals. *Int. J. Quantum Chem.* **2015**, *115*, 1115–1128.

[231] Fritz, M.; Fernández-Serra, M.; Soler, J. M. Optimization of an Exchange-Correlation Density Functional for Water. *J. Chem. Phys.* **2016**, *144*, 224101.

[232] Li, L.; Baker, T. E.; White, S. R.; Burke, K.; others Pure Density Functional for Strong Correlation and the Thermodynamic Limit From Machine Learning. *Phys. Rev. B* **2016**, *94*, 245129.

[233] Li, L.; Snyder, J. C.; Pelaschier, I. M.; Huang, J.; Niranjan, U.-N.; Duncan, P.; Rupp, M.; Müller, K.-R.; Burke, K. Understanding Machine-Learned Density Functionals. *Int. J. Quantum Chem.* **2016**, *116*, 819–833.

[234] Yao, K.; Parkhill, J. Kinetic Energy of Hydrocarbons as a Function of Electron Density and Convolutional Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 1139–1147.

[235] Kolb, B.; Lentz, L. C.; Kolpak, A. M. Discovering Charge Density Functionals and Structure-Property Relationships With PROPhet: A General Framework for Coupling Machine Learning and First-Principles Methods. *Sci. Rep.* **2017**, *7*, 1–9.

[236] Liu, Q.; Wang, J.; Du, P.; Hu, L.; Zheng, X.; Chen, G. Improving the Performance of Long-Range-Corrected Exchange-Correlation Functional With an Embedded Neural Network. *J. Phys. Chem. A* **2017**, *121*, 7273–7281.

[237] Hollingsworth, J.; Li, L.; Baker, T. E.; Burke, K. Can Exact Conditions Improve Machine-Learned Density Functionals? *J. Chem. Phys.* **2018**, *148*, 241743.

[238] Ji, H.; Jung, Y. A Local Environment Descriptor for Machine-Learned Density Functional Theory at the Generalized Gradient Approximation Level. *J. Chem. Phys.* **2018**, *148*, 241742.

[239] Seino, J.; Kageyama, R.; Fujinami, M.; Ikabata, Y.; Nakai, H. Semi-Local Machine-Learned Kinetic Energy Density Functional With Third-Order Gradients of Electron Density. *J. Chem. Phys.* **2018**, *148*, 241705.

[240] Custódio, C. A.; Filletti, É. R.; França, V. V. Artificial Neural Networks for Density-Functional Optimizations in Fermionic Systems. *Sci. Rep.* **2019**, *9*, 1–7.

[241] Lei, X.; Medford, A. J. Design and Analysis of Machine Learning Exchange-Correlation Functionals via Rotationally Invariant Convolutional Descriptors. *Phys. Rev. Mater.* **2019**, *3*, 063801.

[242] Ma, J.; Zhang, P.; Tan, Y.; Ghosh, A. W.; Chern, G.-W. Machine Learning Electron Correlation in a Disordered Medium. *Phys. Rev. B* **2019**, *99*, 085118.

[243] Nudejima, T.; Ikabata, Y.; Seino, J.; Yoshikawa, T.; Nakai, H. Machine-Learned Electron Correlation Model Based on Correlation Energy Density at Complete Basis Set Limit. *J. Chem. Phys.* **2019**, *151*, 024104.

[244] Ryczko, K.; Strubbe, D. A.; Tamblyn, I. Deep Learning and Density-Functional Theory. *Phys. Rev. A* **2019**, *100*, 022512.

[245] Schmidt, J.; Benavides-Riveros, C. L.; Marques, M. A. Machine Learning the Physical Nonlocal Exchange–Correlation Functional of Density-Functional Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 6425–6431.

[246] Zhou, Y.; Wu, J.; Chen, S.; Chen, G. Toward the Exact Exchange–Correlation Potential: A Three-Dimensional Convolutional Neural Network Construct. *J. Phys. Chem. Lett.* **2019**, *10*, 7264–7269.

[247] Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.-R.; Burke, K. Quantum Chemical Accuracy From Density Functional Approximations via Machine Learning. *Nat. Commun.* **2020**, *11*, 5223.

[248] Manzhos, S. Machine Learning for the Solution of the Schrödinger Equation. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 013002.

[249] Meyer, R.; Weichselbaum, M.; Hauser, A. W. Machine Learning Approaches Toward Orbital-Free Density Functional Theory: Simultaneous Training on the Kinetic Energy Density Functional and Its Functional Derivative. *J. Chem. Theory Comput.* **2020**, *16*, 5685–5694.

[250] Nagai, R.; Akashi, R.; Sugino, O. Completing Density Functional Theory by Machine Learning Hidden Messages From Molecules. *Npj Comput. Mater.* **2020**, *6*, 1–8.

[251] Ryabov, A.; Akhatov, I.; Zhilyaev, P. Neural Network Interpolation of Exchange-Correlation Functional. *Sci. Rep.* **2020**, *10*, 1–7.

[252] Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS: A Comprehensive Data-Driven Approach Toward Chemically Accurate Density Functional Theory. *J. Chem. Theory Comput.* **2020**, *17*, 170–181.

[253] Dick, S.; Fernandez-Serra, M. Machine Learning Accurate Exchange and Correlation Functionals of the Electronic Density. *Nat. Commun.* **2020**, *11*.

[254] Kalita, B.; Li, L.; McCarty, R. J.; Burke, K. Learning to Approximate Density Functionals. *Acc. Chem. Res.* **2021**, *54*, 818–826.

[255] Li, L.; Hoyer, S.; Pederson, R.; Sun, R.; Cubuk, E. D.; Riley, P.; Burke, K.; others Kohn-Sham Equations as Regularizer: Building Prior Knowledge Into Machine-Learned Physics. *Phys. Rev. Lett.* **2021**, *126*, 036401.

[256] Nelson, J.; Tiwari, R.; Sanvito, S. Machine-Learning Semi-Local Density Functional Theory for Many-Body Lattice Models at Zero and Finite Temperature. *arXiv preprint arXiv:2103.05510* **2021**, Preprint.

[257] Ryczko, K.; Wetzel, S. J.; Melko, R. G.; Tamblyn, I. Orbital-Free Density Functional Theory With Small Datasets and Deep Learning. *arXiv preprint arXiv:2104.05408* **2021**, Preprint.

[258] Chen, Y.; Zhang, L.; Wang, H.; E, W. DeePKS: A Comprehensive Data-Driven Approach Toward Chemically Accurate Density Functional Theory. *J. Chem. Theory Comput.* **2021**, *17*, 170–181.

[259] Margraf, J. T.; Reuter, K. Pure Non-Local Machine-Learned Density Functional Theory for Electron Correlation. *Nat. Commun.* **2021**, *12*.

[260] Schwilk, M.; Tahchieva, D. N.; von Lilienfeld, O. A. Large Yet Bounded: Spin Gap Ranges in Carbenes. *arXiv:2004.10600 [physics]* **2020**, arXiv: 2004.10600.

[261] Werner, H.-J.; Knowles, P. J. An Efficient Internally Contracted Multiconfiguration–Reference Configuration Interaction Method. *J. Chem. Phys.* **1988**, *89*, 5803–5814.

[262] Knowles, P. J.; Werner, H.-J. An Efficient Method for the Evaluation of Coupling Coefficients in Configuration Interaction Calculations. *Chem. Phys. Lett.* **1988**, *145*, 514–522.

[263] Knowles, P. J.; Werner, H.-J. Internally Contracted Multiconfiguration-Reference Configuration Interaction Calculations for Excited States. *Theor. Chim. Acta.* **1992**, *84*, 95–103.

[264] Shiozaki, T.; Knizia, G.; Werner, H.-J. Explicitly Correlated Multireference Configuration Interaction: MRCI-F12. *J. Chem. Phys.* **2011**, *134*, 034113.

[265] Chen, Y.; Zhang, L.; Wang, H.; E, W. Ground State Energy Functional With Hartree–Fock Efficiency and Chemical Accuracy. *J. Phys. Chem. A* **2020**, *124*, 7155–7165.

[266] Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

[267] Paszke, A. et al. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.

[268] Dick, S. Semodi/Neuralxc. `https://github.com/semodi/neuralxc`, (accessed 2021-04-15).

[269] Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (Gelus). *arXiv:1606.08415 [cs.LG]* **2016**, Preprint.

[270] Qiao, Z.; Welborn, M.; Anandkumar, A.; Manby, F. R.; Miller, T. F. OrbNet: Deep Learning for Quantum Chemistry Using Symmetry-Adapted Atomic-Orbital Features. *J. Chem. Phys.* **2020**, *153*, 124111.

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

[272] Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs.LG]* **2014**, Preprint.

[273] Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA, 2019; pp 2623–2631.

[274] MatthewRHermes/mrh. `https://github.com/MatthewRHermes/mrh`, (accessed 2021-04-15).

[275] Sharma, P.; Bernales, V.; Truhlar, D. G.; Gagliardi, L. Valence * Excitations in Benzene Studied by Multiconfiguration Pair-Density Functional Theory. *J. Phys. Chem. Lett.* **2019**, *10*, 75–81.

[276] Stoneburner, S. J.; Truhlar, D. G.; Gagliardi, L. MC-PDFT Can Calculate Singlet–triplet Splittings of Organic Diradicals. *J. Chem. Phys.* **2018**, *148*, 064108.

[277] Doering, J. P. Low-Energy Electron-Impact Study of the First, Second, and Third Triplet States of Benzene. *J. Chem. Phys.* **1969**, *51*, 2866–2870.

[278] Stoneburner, S. J.; Shen, J.; Ajala, A. O.; Piecuch, P.; Truhlar, D. G.; Gagliardi, L. Systematic Design of Active Spaces for Multi-Reference Calculations of Singlet–triplet Gaps of Organic Diradicals, With Benchmarks Against Doubly Electron-Attached Coupled-Cluster Data. *J. Chem. Phys.* **2017**, *147*, 164120.

[279] Cheng, L.; Kovachki, N. B.; Welborn, M.; Miller, T. F. Regression Clustering for Improved Accuracy and Training Costs With Molecular-Orbital-Based Machine Learning. *J. Chem. Theory Comput.* **2019**, *15*, 6668–6677.

[280] Brockherde, F.; Vogt, L.; Li, L.; Tuckerman, M. E.; Burke, K.; Müller, K.-R. Bypassing the Kohn-Sham Equations With Machine Learning. *Nat. Commun.* **2017**, *8*, 1–10.

[281] Fabrizio, A.; Grisafi, A.; Meyer, B.; Ceriotti, M.; Corminboeuf, C. Electron Density Learning of Non-Covalent Systems. *Chem. Sci.* **2019**, *10*, 9424–9432.

[282] Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2019**, *5*, 57–64.

[283] Kamal, D.; Chandrasekaran, A.; Batra, R.; Ramprasad, R. A Charge Density Prediction Model for Hydrocarbons Using Deep Neural Networks. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025003.

[284] Cuevas-Zuviría, B.; Pacios, L. F. Machine Learning of Analytical Electron Density in Large Molecules Through Message-Passing. *J. Chem. Inf. Model.* **2021**, *61*, 2658–2666.

[285] Fabrizio, A.; Briling, K. R.; Girardier, D. D.; Corminboeuf, C. Learning On-Top: Regressing the On-Top Pair Density for Real-Space Visualization of Electron Correlation. *J. Chem. Phys.* **2020**, *153*, 204111.

[286] Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J. Data-Driven Approaches Can Overcome the Cost–Accuracy Trade-Off in Multireference Diagnostics. *J. Chem. Theory Comput.* **2020**,

[287] Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J. Semi-Supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost. *J. Phys. Chem. Lett.* **2020**, *11*, 6640–6648.

[288] Shee, J.; Loipersberger, M.; Hait, D.; Lee, J.; Head-Gordon, M. Revealing the Nature of Electron Correlation in Transition Metal Complexes With Symmetry Breaking and Chemical Intuition. *J. Chem. Phys.* **2021**, *154*, 194109.

[289] Janet, J. P.; Kulik, H. J. Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure–Property Relationships. *J. Phys. Chem. A* **2017**, *121*, 8939–8954.

[290] Paul Janet, J.; J. Kulik, H. Predicting Electronic Structure Properties of Transition Metal Complexes With Neural Networks. *Chem. Sci.* **2017**, *8*, 5137–5152.

[291] Westermayr, J.; Marquetand, P. Machine Learning and Excited-State Molecular Dynamics. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 043001.

[292] Sifain, A. E.; Lystrom, L.; Messerly, R. A.; Smith, J. S.; Nebgen, B.; Barros, K.; Tretiak, S.; Lubbers, N.; Gifford, B. J. Predicting Phosphorescence Energies and Inferring Wavefunction Localization With Machine Learning. *Chem. Sci.* **2021**,

[293] Glaser, C. Beiträge Zur Kenntniss Des Acetenylbenzols. *Ber. Dtsch. Chem. Ges.* **1869**, *2*, 422–424.

[294] Glaser, C. Untersuchungen Über Einige Derivate Der Zimmtsäure. *Justus Liebigs Ann. Chem.* **1870**, *154*, 137–171.

[295] Siemsen, P.; Livingston, R. C.; Diederich, F. Acetylenic Coupling: A Powerful Tool in Molecular Construction. *Angew. Chem. Int. Ed.* **2000**, *39*, 2632–2657.

[296] Wendlandt, A. E.; Suess, A. M.; Stahl, S. S. Copper-Catalyzed Aerobic Oxidative C–H Functionalizations: Trends and Mechanistic Insights. *Angew. Chem. Int. Ed.* **2011**, *50*, 11062–11087.

[297] Casitas, A.; Ribas, X. The Role of Organometallic Copper (III) Complexes in Homogeneous Catalysis. *Chem. Sci.* **2013**, *4*, 2301–2318.

[298] Allen, S. E.; Walvoord, R. R.; Padilla-Salinas, R.; Kozlowski, M. C. Aerobic Copper-Catalyzed Organic Reactions. *Chem. Rev.* **2013**, *113*, 6234–6458.

[299] Trammell, R.; Rajabimoghadam, K.; Garcia-Bosch, I. Copper-Promoted Functionalization of Organic Molecules: From Biologically Relevant Cu/O2 Model Systems to Organometallic Transformations. *Chem. Rev.* **2019**, *119*, 2954–3031.

[300] Suess, A. M.; Ertem, M. Z.; Cramer, C. J.; Stahl, S. S. Divergence Between Organometallic and Single-Electron-Transfer Mechanisms in Copper (II)-mediated Aerobic C–H Oxidation. *J. Am. Chem. Soc.* **2013**, *135*, 9797–9804.

[301] McCann, S. D.; Stahl, S. S. Copper-Catalyzed Aerobic Oxidations of Organic Molecules: Pathways for Two-Electron Oxidation With a Four-Electron Oxidant and a One-Electron Redox-Active Catalyst. *Acc. Chem. Res.* **2015**, *48*, 1756–1766.

[302] King, A. E.; Brunold, T. C.; Stahl, S. S. Mechanistic Study of Copper-Catalyzed Aerobic Oxidative Coupling of Arylboronic Esters and Methanol: Insights Into an Organometallic Oxidase Reaction. *J. Am. Chem. Soc.* **2009**, *131*, 5044–5045.

[303] Jover, J.; Spuhler, P.; Zhao, L.; McArdle, C.; Maseras, F. Toward a Mechanistic Understanding of Oxidative Homocoupling: The Glaser–Hay Reaction. *Catalysis Science & Technology* **2014**, *4*, 4200–4209.

[304] Funes-Ardoiz, I.; Maseras, F. Oxidative Coupling Mechanisms: Current State of Understanding. *ACS Catal.* **2018**, *8*, 1161–1172.

[305] Fomina, L.; Vazquez, B.; Tkatchouk, E.; Fomine, S. The Glaser Reaction Mechanism. A DFT Study. *Tetrahedron* **2002**, *58*, 6741–6747.

[306] Bohlmann, F.; Schönowsky, H.; Inhoffen, E.; Grau, G. Polyacetylenverbindungen, LII. Über Den Mechanismus Der Oxydativen Dimerisierung Von Acetylenverbindungen. *Chem. Ber.* **1964**, *97*, 794–800.

[307] Qi, X.; Bai, R.; Zhu, L.; Jin, R.; Lei, A.; Lan, Y. Mechanism of Synergistic Cu(II)/Cu(I)-Mediated Alkyne Coupling: Dinuclear 1,2-Reductive Elimination After Minimum Energy Crossing Point. *J. Org. Chem.* **2016**, *81*, 1654–1660.

[308] Ryan, M. C.; Martinelli, J. R.; Stahl, S. S. Cu-Catalyzed Aerobic Oxidative N–N Coupling of Carbazoles and Diarylamines Including Selective Cross-Coupling. *J. Am. Chem. Soc.* **2018**, *140*, 9074–9077.

[309] Ryan, M. C.; Kim, Y. J.; Gerken, J. B.; Wang, F.; Aristov, M. M.; Martinelli, J. R.; Stahl, S. S. Mechanistic Insights Into Copper-Catalyzed Aerobic Oxidative Coupling of N–N Bonds. *Chem. Sci.* **2020**, *11*, 1170–1175.

[310] Wang, F.; Gerken, J. B.; Bates, D. M.; Kim, Y. J.; Stahl, S. S. Electrochemical Strategy for Hydrazine Synthesis: Development and Overpotential Analysis of Methods for Oxidative N–N Coupling of an Ammonia Surrogate. *J. Am. Chem. Soc.* **2020**, *142*, 12349–12356.

[311] Hayashi, H.; Kainoh, A.; Katayama, M.; Kawasaki, K.; Okazaki, T. Hydrazine Production From Ammonia via Azine. *Ind. Eng. Chem., Prod. Res. Dev.* **1976**, *15*, 299–303.

[312] Hayashi, H. Hydrazine Synthesis by a Catalytic Oxidation Process. *Catal. Rev. - Sci. Eng.* **1990**, *32*, 229–277.

[313] Halfen, J. A.; Mahapatra, S.; Wilkinson, E. C.; Kaderli, S.; Young Jr, V. G.; Que Jr, L.; Zuberbühler, A. D.; Tolman, W. B. Reversible Cleavage and Formation of the Dioxygen OO Bond Within a Dicopper Complex. *Science* **1996**, *271*, 1397–1400.

[314] Mirica, L. M.; Ottenwaelder, X.; Stack, T. D. P. Structure and Spectroscopy of Copper-Dioxygen Complexes. *Chem. Rev.* **2004**, *104*, 1013–1046.

[315] Lewis, E. A.; Tolman, W. B. Reactivity of Dioxygen- Copper Systems. *Chem. Rev.* **2004**, *104*, 1047–1076.

[316] Elwell, C. E.; Gagnon, N. L.; Neisen, B. D.; Dhar, D.; Spaeth, A. D.; Yee, G. M.; Tolman, W. B. Copper–Oxygen Complexes Revisited: Structures, Spectroscopy, and Reactivity. *Chem. Rev.* **2017**, *117*, 2059–2107.

[317] Frisch, M. e.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H. Gaussian 16. 2016.

[318] Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys* **2005**, *7*, 3297–3305.

[319] Weigend, F. Accurate Coulomb-Fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys* **2006**, *8*, 1057–1065.

[320] Brotherton, T.; Lynn, J. The Synthesis and Chemistry of Cyanogen. *Chem. Rev.* **1959**, *59*, 841–883.

[321] Lu, J.; Dreisinger, D.; Cooper, W. Thermodynamics of the Aqueous Copper–Cyanide System. *Hydrometallurgy* **2002**, *66*, 23–36.

[322] Jayasooriya, I. U.; Palmer, R.; Ng, K.; Khachemoune, N. L.; Bertke, J. A.; Warren, T. H.; others Copper (Ii) Ketimides in Sp 3 C–H Amination. *Chem. Sci.* **2021**, *12*, 15733–15738.

[323] Ahmed, M. E.; Raghibi Boroujeni, M.; Ghosh, P.; Greene, C.; Kundu, S.; Bertke, J. A.; Warren, T. H. Electrocatalytic Ammonia Oxidation by a Low-Coordinate Copper Complex. *J. Am. Chem. Soc.* **2022**, *144*, 21136–21145.

[324] Gu, N. X.; Oyala, P. H.; Peters, J. C. Hydrazine Formation via Coupling of a Nickel (III)–NH_2 Radical. *Angew. Chem. Int. Ed.* **2021**, *60*, 4009–4013.

[325] DiMucci, I. M.; Lukens, J. T.; Chatterjee, S.; Carsch, K. M.; Titus, C. J.; Lee, S. J.; Nordlund, D.; Betley, T. A.; MacMillan, S. N.; Lancaster, K. M. The Myth of D8 Copper (III). *J. Am. Chem. Soc.* **2019**, *141*, 18508–18520.

[326] Dhar, D.; Yee, G. M.; Spaeth, A. D.; Boyce, D. W.; Zhang, H.; Dereli, B.; Cramer, C. J.; Tolman, W. B. Perturbing the Copper (III)–Hydroxide Unit Through Ligand Structural Variation. *J. Am. Chem. Soc.* **2016**, *138*, 356–368.

[327] Wu, T.; MacMillan, S. N.; Rajabimoghadam, K.; Siegler, M. A.; Lancaster, K. M.; Garcia-Bosch, I. Structure, Spectroscopy, and Reactivity of a Mononuclear Copper Hydroxide Complex in Three Molecular Oxidation States. *J. Am. Chem. Soc.* **2020**, *142*, 12265–12276.

[328] Bower, J. K.; Cypcar, A. D.; Henriquez, B.; Stieber, S. C. E.; Zhang, S. C (Sp3)–H Fluorination With a Copper (II)/(III) Redox Couple. *J. Am. Chem. Soc.* **2020**, *142*, 8514–8521.

[329] Bower, J. K.; Cypcar, A. D.; Henriquez, B.; Stieber, S. C. E.; Zhang, S. Correction to "C (Sp3)–H Fluorination With a Copper (II)/(III) Redox Couple". *J. Am. Chem. Soc.* **2022**, *144*, 6118–6119.

[330] Barnett, S. M.; Goldberg, K. I.; Mayer, J. M. A Soluble Copper–Bipyridine Water-Oxidation Electrocatalyst. *Nat. Chem.* **2012**, *4*, 498–502.

[331] Du, J.; Chen, Z.; Ye, S.; Wiley, B. J.; Meyer, T. J. Copper as a Robust and Transparent Electrocatalyst for Water Oxidation. *Angew. Chem. Int. Ed.* **2015**, *54*, 2073–2078.

[332] Koepke, S. J.; Light, K. M.; VanNatta, P. E.; Wiley, K. M.; Kieber-Emmons, M. T. Electrocatalytic Water Oxidation by a Homogeneous Copper Catalyst Disfavors Single-Site Mechanisms. *J. Am. Chem. Soc.* **2017**, *139*, 8586–8600.

[333] Henson, M. J.; Mukherjee, P.; Root, D. E.; Stack, T.; Solomon, E. I. Spectroscopic and Electronic Structural Studies of the Cu (III) 2 Bis--Oxo Core and Its Relation to the Side-On Peroxo-Bridged Dimer. *J. Am. Chem. Soc.* **1999**, *121*, 10332–10345.

[334] Garcia-Bosch, I.; Cowley, R. E.; Díaz, D. E.; Peterson, R. L.; Solomon, E. I.; Karlin, K. D. Substrate and Lewis Acid Coordination Promote O–O Bond Cleavage of an Unreactive L2CuII2 (O22–) Species to Form L2CuIII2 (O) 2 Cores With Enhanced Oxidative Reactivity. *J. Am. Chem. Soc.* **2017**, *139*, 3186–3195.

[335] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1*, 1–7.

[336] Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching Coupled Cluster Accuracy With a General-Purpose Neural Network Potential Through Transfer Learning. *Nat. Comm.* **2019**, *10*, 2903.

[337] Harvey, J. N. On the Accuracy of Density Functional Theory in Transition Metal Chemistry. *Annu. Rep. Prog. Chem., Sect. C: Phys. Chem.* **2006**, *102*, 203–226.

[338] Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Fractional Spins and Static Correlation Error in Density Functional Theory. *J. Chem. Phys.* **2008**, *129*, 121104.

[339] Fuchs, M.; Niquet, Y.-M.; Gonze, X.; Burke, K. Describing Static Correlation in Bond Dissociation by Kohn–Sham Density Functional Theory. *J. Chem. Phys.* **2005**, *122*, 094116.

[340] Hollett, J. W.; Gill, P. M. W. The Two Faces of Static Correlation. *J. Chem. Phys.* **2011**, *134*, 114111.

[341] Bulik, I. W.; Henderson, T. M.; Scuseria, G. E. Can Single-Reference Coupled Cluster Theory Describe Static Correlation? *J. Chem. Theory Comput.* **2015**, *11*, 3171–3179.

[342] Zhao, Q.; Savoie, B. M. Algorithmic Explorations of Unimolecular and Bimolecular Reaction Spaces. *Angew. Chem. Int. Ed.* **2022**, *61*, e202210693.

[343] Vitillo, J. G.; Cramer, C. J.; Gagliardi, L. Multireference Methods Are Realistic and Useful Tools for Modeling Catalysis. *Isr. J. Chem.* **2022**, *62*, e202100136.

[344] Zhang, J.; Hu, T.; Lv, H.; Dong, C. H-Abstraction Mechanisms in Oxidation Reaction of Methane and Hydrogen: A CASPT2 Study. *Int. J. Hydrogen Energy* **2016**, *41*, 12722–12729.

[345] Francés-Monerris, A.; Segarra-Martí, J.; Merchán, M.; Roca-Sanjuán, D. Complete-Active-Space Second-Order Perturbation Theory (CASPT2//CASSCF) Study of the Dissociative Electron Attachment in Canonical DNA Nucleobases Caused by Low-Energy Electrons (0-3 eV). *J. Chem. Phys.* **2015**, *143*, 215101.

[346] Guan, P.-J.; Fang, W.-H. The Combined CASPT2 and CASSCF Studies on Photolysis of 3-Thienyldiazomethane and Subsequent Reactions. *Theor. Chem. Acc.* **2014**, *133*, 1532.

[347] Sand, A. M.; Kidder, K. M.; Truhlar, D. G.; Gagliardi, L. Calculation of Chemical Reaction Barrier Heights by Multiconfiguration Pair-Density Functional Theory With Correlated Participating Orbitals. *J. Phys. Chem. A* **2019**, *123*, 9809–9817.

[348] Arenas, J. F.; Otero, J. C.; Peláez, D.; Soto, J. CASPT2 Study of the Decomposition of Nitrosomethane and Its Tautomerization Reactions in the Ground and Low-Lying Excited States. *J. Org. Chem.* **2006**, *71*, 983–991.

[349] Talotta, F.; González, L.; Boggio-Pasqua, M. CASPT2 Potential Energy Curves for NO Dissociation in a Ruthenium Nitrosyl Complex. *Molecules* **2020**, *25*, 2613.

[350] Vancoillie, S.; Malmqvist, P. ; Veryazov, V. Potential Energy Surface of the Chromium Dimer Re-Re-Revisited With Multiconfigurational Perturbation Theory. *J. Chem. Theory Comput.* **2016**, *12*, 1647–1655.

[351] Bensberg, M.; Reiher, M. Corresponding Active Orbital Spaces Along Chemical Reaction Paths. *J. Phys. Chem. Lett.* **2023**, *14*, 2112–2118.

[352] Andersson, K.; Malmqvist, P. A.; Roos, B. O.; Sadlej, A. J.; Wolinski, K. Second-Order Perturbation Theory With a CASSCF Reference Function. *J. Phys. Chem.* **1990**, *94*, 5483–5488.

[353] Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. Second-Order Perturbation Theory With a Complete Active Space Self-Consistent Field Reference Function. *J. Chem. Phys.* **1992**, *96*, 1218–1226.

[354] Song, S.; Vuckovic, S.; Sim, E.; Burke, K. Density-Corrected DFT Explained: Questions and Answers. *J. Chem. Theory Comput.* **2022**, *18*, 817–827.

[355] Lei, Y.; Suo, B.; Liu, W. iCAS: Imposed Automatic Selection and Localization of Complete Active Spaces. *J. Chem. Theory Comput.* **2021**, *17*, 4846–4859.

[356] King, D. S.; Truhlar, D. G.; Gagliardi, L. Variational Active Space Selection with Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2023**, *19*, 8118–8128.

[357] Zhao, Q.; Vaddadi, S. M.; Woulfe, M.; Ogunfowora, L. A.; Garimella, S. S.; Isayev, O.; Savoie, B. M. Comprehensive Exploration of Graphically Defined Reaction Spaces. *Sci. Data* **2023**, *10*, 145.

[358] Kaufold, B. W.; Chintala, N.; Pandeya, P.; Dong, S. S. Automated Active Space Selection With Dipole Moments. *J. Chem. Theory Comput.* **2023**, *19*, 2469–2483.

[359] Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K.-L. PySCF: The Python-Based Simulations of Chemistry Framework. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1340.

[360] Sun, Q. et al. Recent Developments in the PySCF Program Package. *J. Chem. Phys.* **2020**, *153*, 024109.

[361] Dunning Jr, T. H. Gaussian Basis Sets for Use in Correlated Molecular Calculations. I. The Atoms Boron Through Neon and Hydrogen. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

[362] Peterson, K. A.; Dunning Jr, T. H. Accurate Correlation Consistent Basis Sets for Molecular Core–Valence Correlation Effects: The Second Row Atoms Al–Ar, and the First Row Atoms B–Ne Revisited. *J. Chem. Phys.* **2002**, *117*, 10548–10560.

[363] Sharma, P.; Truhlar, D. G.; Gagliardi, L. Active Space Dependence in Multiconfiguration Pair-Density Functional Theory. *J. Chem. Theory Comput.* **2018**, *14*, 660–669.