The Ranked-Orbital Approach to Selecting Active Spaces

Daniel S. King and Laura Gagliardi*

Department of Chemistry, University of Chicago, Chicago IL

E-mail: lgagliardi@uchicago.edu

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. 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.

1 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,[1] gas storage,[2] and photochemistry.[3–5] 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.[1–5] However, one area where the high-throughput approach is poised to play a large role is in the development of new transition metal catalysts,[6,7] and these complexes are often strongly correlated and thus poorly described by single-reference methods such as DFT.[8–13] Furthermore, DFT suffers from an inability to describe open-shell systems without resorting to broken-symmetry solutions,[10] and this becomes particularly severe when multiple low-lying spin states are important to consider (e.g. in application to spin-crossover complexes[14,15]). 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.[13]

For these reasons, we expect that the expansion of reliable high-throughput computation to these problems will require the use of multireference approaches.[12] 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,[16] 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:

$$|\psi_{CAS}\rangle = \sum_{n_1 n_2 ... n_n} c_{n_1 n_2 ... n_n} \, |22...n_1 n_2 ... n_n 00...\rangle \tag{1}$$

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.[12] The number of alpha and beta electrons in the active space is conserved in all determinants in the expansion (equation 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 1 can be expressed through the number of "configuration state functions" (CSFs) as[17]

$$N_{CSF} = \binom{L}{\alpha}\binom{L}{\beta} - \binom{L}{\alpha+1}\binom{L}{\beta-1} \tag{2}$$

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 1 through exact diagonalization is known as CASCI, while optimizing the orbitals and the coefficients simultaneously is known as CASSCF.[18] Currently, the maximum active space that can be computed with these methods is about (20,20).[12] To expand beyond this limit, several methods exist for approximating the coefficients in equation 1, such as the density matrix renormalization group method (DMRG),[19–21] full configuration

interaction quantum Monte Carlo (FCIQMC),[22,23] and neural networks.[24] These approximate methods can handle up to hundreds of orbitals in the active space.[25] Often, the active orbitals are variationally optimized in tandem with the coefficients in equation 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[26] or n-electron valence perturbation theory (NEVPT2).[27] The addition of dynamical correlation through these methods limits calculations to only about 14 orbitals in the active space.[12]

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.[28] 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.[25,29–40]

A new approach for selecting active spaces that has gathered a lot of attention in recent years goes by the name of AutoCAS,[25,30] 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[41]

$$S_i = -\sum_{j=\{0,\uparrow,\downarrow,2\}} \rho^i_{jj} \ln \rho^i_{jj} \tag{3}$$

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:[41]

$$\rho^i = \sum_{\{\mathbf{n} \neq n^i\}} \langle \mathbf{n}|\psi\rangle \langle \psi|\mathbf{n}\rangle = \sum_{kj} c_k c_j^* \left(\sum_{\{\mathbf{n} \neq n^i\}} \langle \mathbf{n}|\mathbf{n}_k\rangle \langle \mathbf{n}_j|\mathbf{n}\rangle \right) \left|n_k^i\right\rangle \left\langle n_j^i\right| \tag{4}$$

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. The AutoCAS procedure published 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.[25] 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;[42] we refer here to this extended and more expensive scheme as AutoCAS+.

Here, we investigate the performance of the AutoCAS and AutoCAS+ 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.[33] We find that while the AutoCAS 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[33,43] and AVAS,[31] with similar results. To demonstrate the robustness of this approach for high-throughput computation, we carry out 1120 SA-CASSCF/NEVPT2 calculations using four different active space sizes and seven different localization schemes, 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

AutoCAS 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 AutoCAS procedure for these simple systems. Taking inspiration from molecular-orbital based machine learning (MOB-ML),[44] 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 Methods

Excitation Energies. Geometries and reference values for excitation energies were taken from the previous work of Bao and coworkers.[32,33] 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). After the active space is selected by various schemes, final SA-CASSCF/NEVPT2 calculations with the aug-cc-pVTZ basis[45] were performed using PySCF,[46] 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.

AutoCAS/AutoCAS+. Orbital entropies for orbitals generated in PySCF were calculated by interfacing with QCMaquis[21] via the FCIDUMP[47] file interface. DMRG calculations were initialized using the CIDEAS initial guess,[48] and information from this initial calculation was

used to employ an optimized Fiedler ordering[49] 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,[50] Pipek-Mezey with Löwdin charges (PM),[51,52] and Edmiston-Ruedenberg[53] (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,[43] but as recently noted is still capable of selecting good active spaces for many difficult systems.[39] 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[33] 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,[43]

$$S^{1/2}(D_\alpha + D_\beta)S^{1/2}(S^{1/2}C) = \sigma(S^{1/2}C) \tag{5}$$

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.[31] 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,[54] 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.[55]

3 Results and Discussion

3.1 The Limitations of Threshold Schemes

As a simple first test of the viability of AutoCAS and AutoCAS+ 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.[33] 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[56–58] basis (8s8p4d3f2g for carbon and 6s4p3d1f for hydrogen).[42] 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 AutoCAS and AutoCAS+ methods we used both canonical (HF) and Boys-localized orbitals (AutoCAS/HF, AutoCAS+/HF, AutoCAS/Boys and AutoCAS+/Boys).

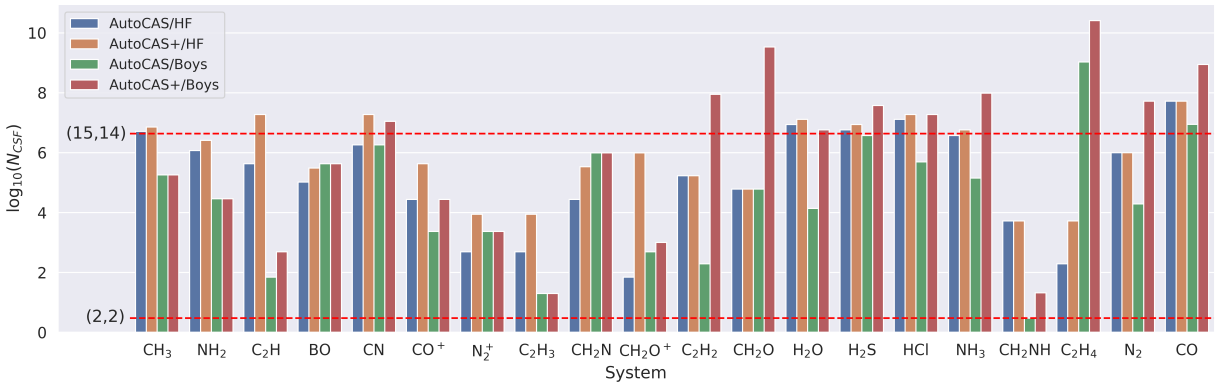

Figure 1: Comparison of the size of the active spaces selected by AutoCAS/HF, AutoCAS+/HF, AutoCAS/Boys, and AutoCAS+/Boys for the twenty different systems investigated, as plotted by the base-10 logarithm of the number of configurations in the selected active space ($\log_{10} N_{CSF}$). No method selects spaces for all systems under the affordable CASSCF/NEVPT2 limit of (15,14) (top horizontal dotted line).

The active space selections of AutoCAS and AutoCAS+ using the $0.1S_{max}$ threshold suggested by Stein and Reiher are shown in figure 1, plotted against the $\log_{10} N_{CSF}$ of configu-

rations in the selected active space, with N_{CSF} as calculated by equation 2. 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).[12] We find that for many systems the active spaces selected by the AutoCAS and AutoCAS+ 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,[59] the size of the selected active space varies heavily by system, orbital localization, and selection method (AutoCAS vs. AutoCAS+). 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 modern methods 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. AutoCAS+) selects an active space with many orders of magnitude more configurations than another scheme (e.g. AutoCAS), it makes it very difficult to meaningfully compare these results.
- Similarly, 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).
- Finally, the variability of the selected active space sizes makes it hard to automate. Spaces with orders of magnitude more configurations will require drastically different amounts of computational resources than smaller ones.

Despite these issues, on physical grounds the orbital entropies used by the AutoCAS and AutoCAS+ 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.[43] Here, we modify these threshold schemes procedurally in order to select consistent, flexible, and affordable active spaces for high-throughput computation.

3.2 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 $\max(N_{orbs}, N_{elec})$ and this space is converted to a maximum N_{CSF}^{MAX} via equation 2, 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), 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.
- AutoCAS, with arbitrary candidate orbitals ranked by their orbital entropy from DMRG
- AutoCAS+, 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):

$$S_i = \frac{1}{N} \sum_{n}^{N} \frac{S_{ni}}{S_n^{max}} \tag{6}$$

- 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.[33] 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.

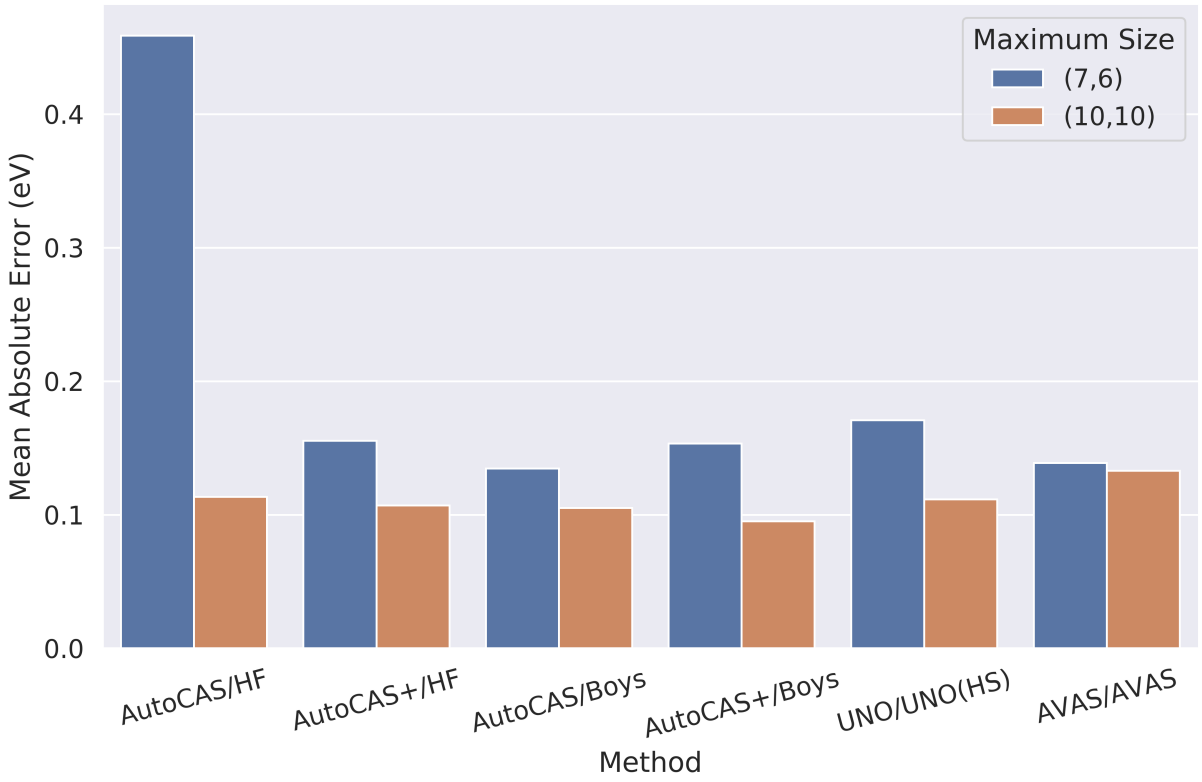

Figure 2: Performance of six different threshold schemes that have been modified by the ranked-orbital procedure at maximum active space sizes of (7,6) and (10,10). The ranked-orbital scheme allows for a meaningful comparison between active space selection schemes, orbital localization schemes, and active 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. 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 AutoCAS/Boys with a mean error of 0.13eV while the best method at max(10,10) is AutoCAS+/Boys with a mean error of 0.11eV. AutoCAS/HF performs quite poorly at the max(7,6) level, mostly due to a poor selection for CH_2O (supporting information).

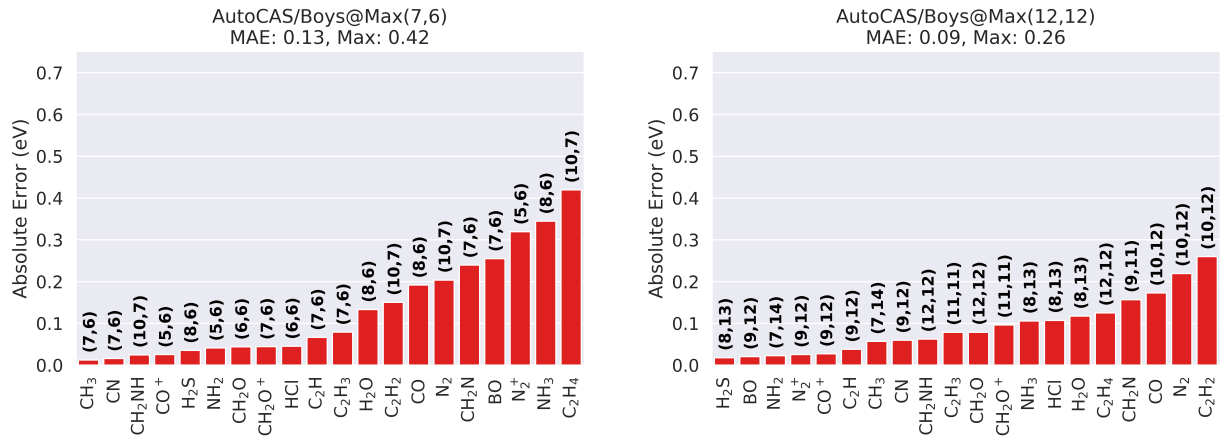

Figure 3: Performance of the modified ranked-orbital AutoCAS scheme at max(7,6) and max(12,12). The ranked-orbital procedure allows for flexibility in the chosen active space while being able to select calculations of a consistent and managable size, by fixing the maximum number of CSFs.

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 AutoCAS procedure using Boys orbitals at max(7,6) and max(12,12) is shown in figure 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 modern CAS solvers, and in line with other methods that require users to select the size of their

approximation such as CISD/CISDT/CISDTQ.

3.3 A High-Throughput Exam: 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 the excitation energies for the 20 small systems using SA-CASSCF/NEVPT2 and choosing from the lowest 30 orbitals in energy with the AutoCAS and AutoCAS+ ranked-orbital approaches using seven different localization schemes (two types of Boys, Pipek-Mezey, and Edmiston-Ruedenberg in addition to canonical orbitals, further described in the supporting information). 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); these differ in their number of CSFs by roughly an order of magnitude each. In total, this amounted to 1120 multireference 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 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 AutoCAS+ 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?

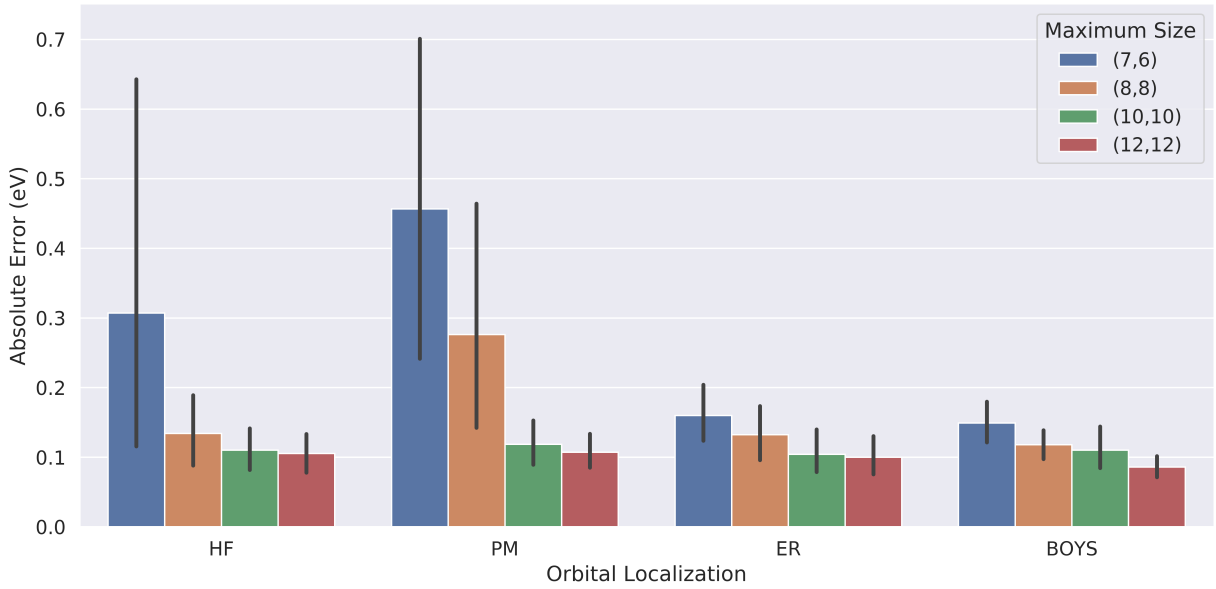

Figure 4: Performance of the ranked AutoCAS/AutoCAS+ procedures over different orbital localizations and maximum active space sizes, plotted by the error of their final CASSCF/NEVPT2 excitation energies from reference values; bootstrapped 95% confidence intervals are shown by vertical bars in black. 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 AutoCAS/AutoCAS+ procedures are shown in figure 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,[52] 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.

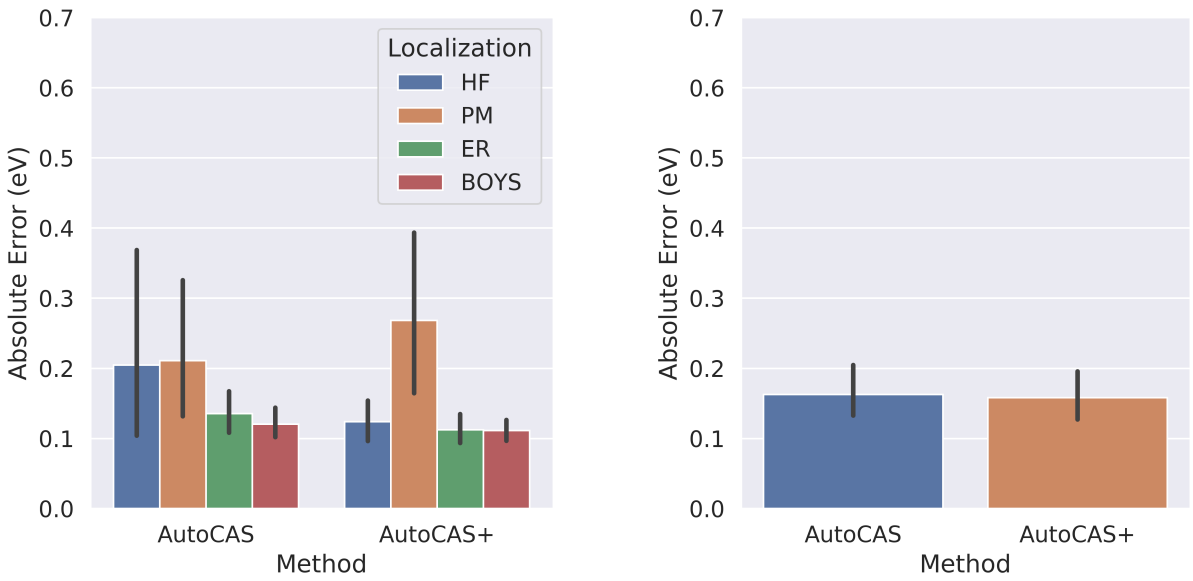

Figure 5: Left: Comparison of the AutoCAS and AutoCAS+ procedures by orbital localization. Right: Comparison of the AutoCAS and AutoCAS+ procedures overall. Confidence intervals at 95% are shown in black.

AutoCAS vs. AutoCAS+. Figure 5 shows the performance of the AutoCAS and AutoCAS+ 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 (AutoCAS+), 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 AutoCAS+ procedure 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 AutoCAS and AutoCAS+ 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).

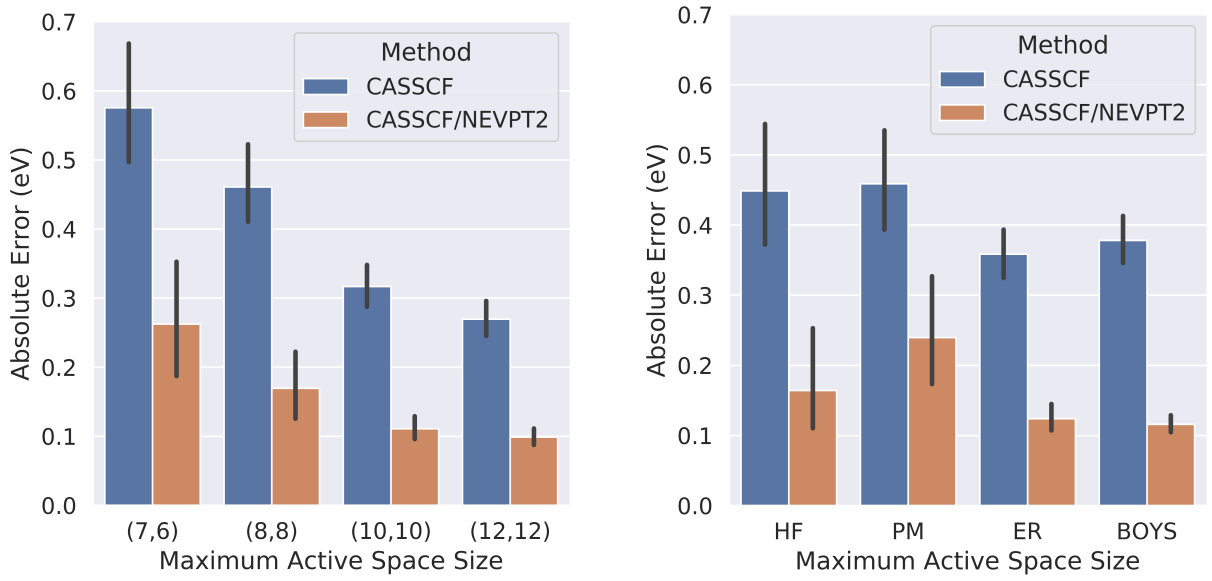

Figure 6: Left: Comparison of CASSCF vs. CASSCF/NEVPT2 by maximum active space size. Right: Comparison of CASSCF vs. CASSCF/NEVPT2 by orbital localization.

CASSCF vs. CASSCF/NEVPT2. Figure 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).

3.4 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:[31,36,39]

- Small singular values σ_i of the overlap matrix between the initial (selected) and final (optimized) active spaces in the CASSCF procedure,[31]

$$S_{change} = (C_{act}^{final})^{\dagger} S C_{act}^{initial} \tag{7}$$

$$= \sum_i \sigma_i u_i v_i^T \tag{8}$$

  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,[39,40] $E_{CASSCF} - E_{CASCI}$ or $\Delta E_{CASCI}^{CASSCF}$

- Large numbers of iterations/macro cycles undertaken by the CASSCF optimization procedure, N_{iter}.[40]

- Large absolute differences between the CASSCF and NEVPT2 excitation energies,[33] $|\Delta E_{NEVPT2} - \Delta E_{CASSCF}|$ or $|\Delta\Delta E_{CASSCF}^{NEVPT2}|$

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.

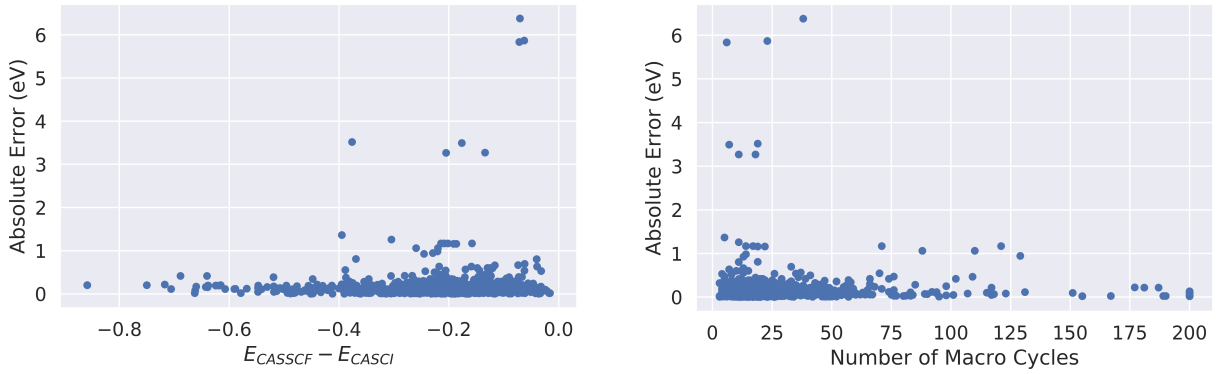

Figure 7: Left: Absolute errors with respect to the state-averaged $\Delta E_{CASCI}^{CASSCF}$. Right: Absolute errors with respect to the number of macro cycles in the CASSCF procedure, N_{iter}. Neither value has any significant correlation with the error of calculated excitation energies.

$\Delta E_{CASCI}^{CASSCF}$ and N_{iter}. Figure 7 shows the performance of the state-averaged $\Delta E_{CASCI}^{CASSCF}$ 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 $\Delta E_{CASCI}^{CASSCF}$ remains surprisingly consistent with maximum active space size (supporting information).

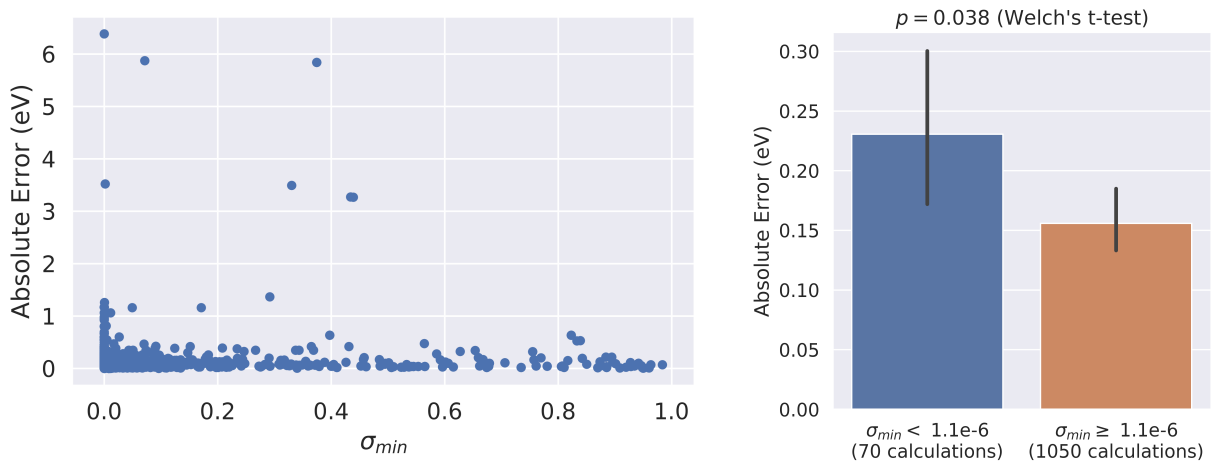

Figure 8: Left: Absolute errors with respect to the minimum singular value σ_{min} of the active space overlap matrix (equations 7 and 8). Right: Performance of the suggested threshold of 1.1e-6, which demonstrates a statistically significant difference between the two groups of calculations under Welch's t-test.[60]

Active Space Overlap. Small singular values of the active space overlap matrix (equations 7 and 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.[31] Figure 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. 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 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.

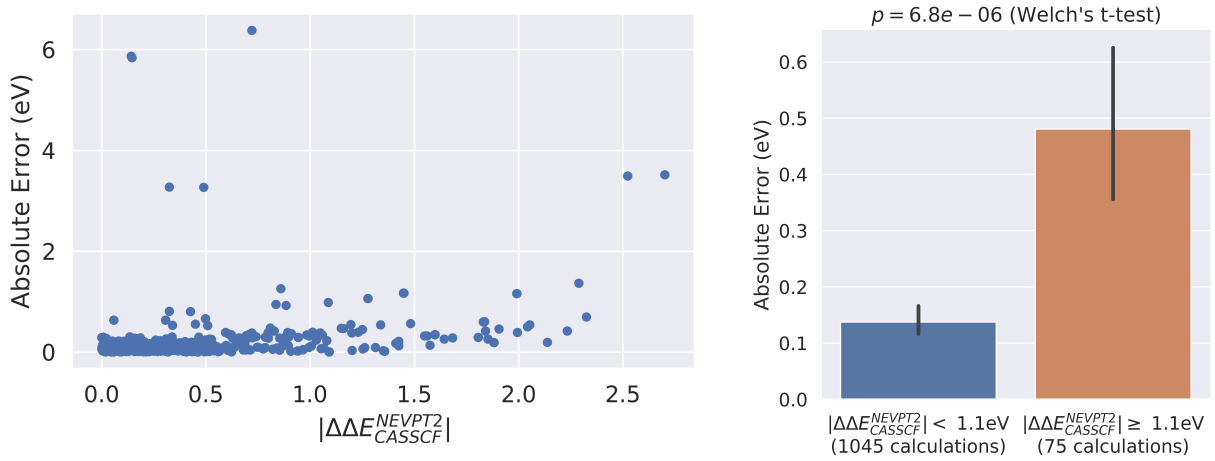

Figure 9: Left: Absolute errors with respect to $|\Delta\Delta E^{NEVPT2}_{CASSCF}|$. Right: Performance of the 1.1 eV threshold suggested by Bao and Truhlar, which demonstrates a significant difference between the two groups of calculations under Welch's t-test.[60]

$|\Delta\Delta E^{NEVPT2}_{CASSCF}|$. Bao and Truhlar suggested classifying an excitation energy result as "reliable" if $|\Delta\Delta E^{NEVPT2}_{CASSCF}| \leq 1.1$ eV.[33] Figure 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.

3.5 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 AutoCAS and AutoCAS+ 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

$$|\psi\rangle = |20\rangle + c\,|02\rangle \tag{9}$$

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 9), we can write the one-orbital reduced density matrix of the doubly occupied orbital, ρ^i, as roughly

$$\rho^i \approx \frac{1}{1+\sum_a c_{ia}^2}\left(|2\rangle\langle 2| + \sum_a c_{ia}^2\,|0\rangle\langle 0|\right) \tag{10}$$

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

$$\rho^a \approx \frac{1}{1+\sum_i c_{ia}^2}\left(|0\rangle\langle 0| + \sum_i c_{ia}^2\,|2\rangle\langle 2|\right) \tag{11}$$

Then, the entropy of a doubly occupied orbital i is approximated via equation 3 as

$$S^i \approx -\frac{1}{1+\sum_a c_{ia}^2}\ln\frac{1}{1+\sum_a c_{ia}^2} - \frac{\sum_a c_{ia}^2}{1+\sum_a c_{ia}^2}\ln\frac{\sum_a c_{ia}^2}{1+\sum_a c_{ia}^2} \tag{12}$$

and for a virtual orbital a as

$$S^a \approx -\frac{1}{1+\sum_i c_{ia}^2} \ln \frac{1}{1+\sum_i c_{ia}^2} - \frac{\sum_i c_{ia}^2}{1+\sum_i c_{ia}^2} \ln \frac{\sum_i c_{ia}^2}{1+\sum_i c_{ia}^2} \tag{13}$$

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 9), in which c is given exactly by the solution to the CI eigenvalue problem[61]

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

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,

$$c = -\frac{(12|12)}{\Delta + \sqrt{(12|12)^2 + \Delta^2}} \tag{14}$$

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 are the orbital energies. 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,

$$\epsilon_2 = h_{22} + J_{22} - 0.5K_{22} \tag{15}$$

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

$$\epsilon_2 = (2|2) + 2(11|22) - (12|12) \tag{16}$$

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

$$(12|12) \approx 0.5K_{22} \tag{17}$$

and similarly for an arbitrary doubly occupied orbital i and virtual orbital a as roughly

$$(ia|ia) \approx 0.5K_{aa} \tag{18}$$

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

$$c_{ia} = -\frac{0.5K_{aa}}{(\epsilon_a - \epsilon_i) + \sqrt{(0.5K_{aa})^2 + (\epsilon_a - \epsilon_i)^2}} \tag{19}$$

These coefficients are then gathered for each orbital and used in equations 12 and 13 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 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,[44] 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 DMRG calculations in the previous section, and a full description of this scheme is available in the supporting information; we refer to this scheme as "APCML".

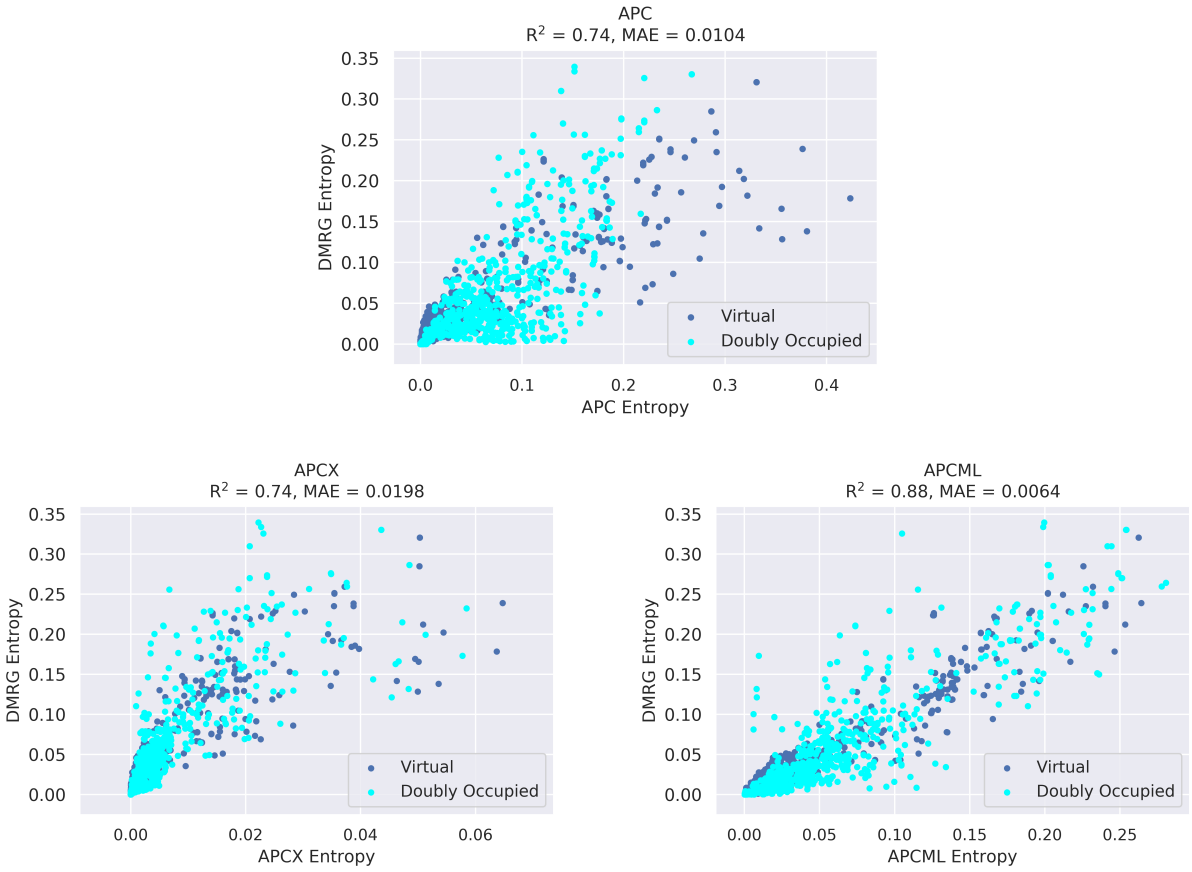

Figure 10: Top: APC entropies vs. DMRG entropies. Bottom: APCX and APCML entropies vs. DMRG entropies. The approximate pair coefficient (APC) approximation is a surprisingly accurate approximation of the orbital entropy for these simple systems.

Figure 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. $\sigma = 0.033$).

Surprisingly, we find that the APC approximation performs significantly better than APCX in approximating the DMRG entropies, indicating a fortunate cancellation of error. 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.

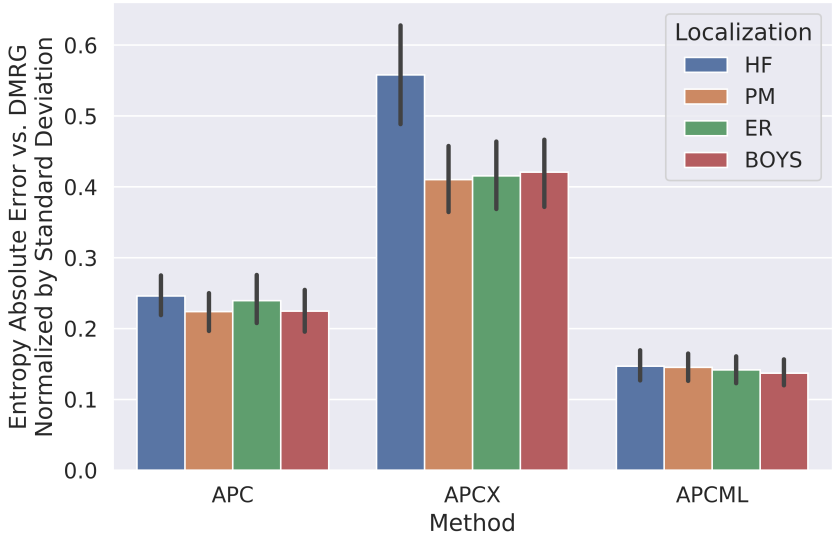

Figure 11: Error of different approximate methods for the DMRG entropy vs. the DMRG values, normalized by the standard deviation of entropy values in that orbital type (supporting information); bootstrapped 95% confidence intervals are shown by vertical bars in black. Surprisingly, there is no significant 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 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)).

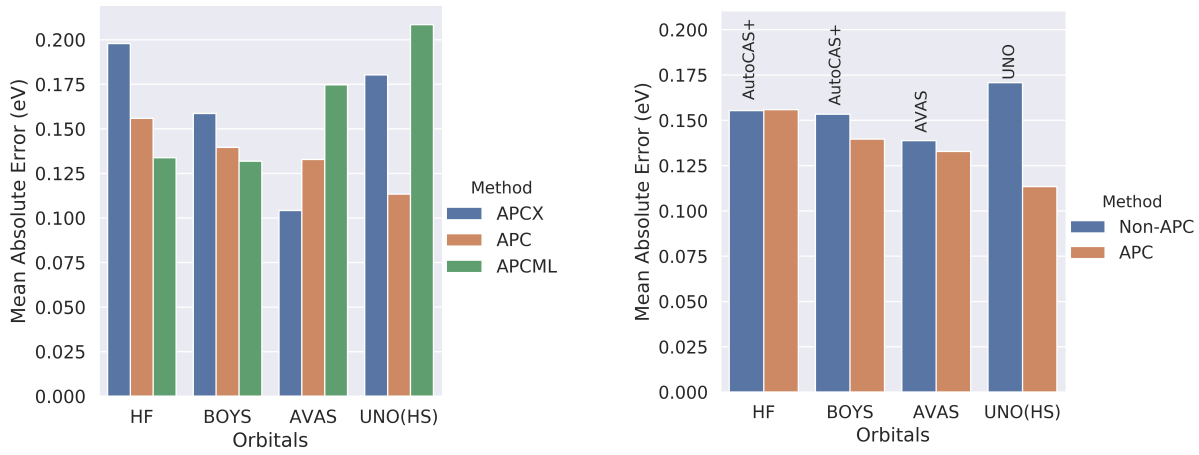

Figure 12: Left: Performance of APC/APCX/APCML selection on orbitals from different active space selection schemes at the max(7,6) level. Right: Performance of APC selection vs. non-APC selection at the max(7,6) level.

Figure 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[33]), 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,[36] who focused on learning the entropy for transition metal systems in much more difficult cases. While not easily comparable due to the difficulty of the systems studied, the approach here manages a 5-6x lower mean absolute error with an order of magnitude less training data, in addition to it being much less expensive due to its featurization from solely the HF matrices and not from molecular orbital integrals. Regardless, 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[62] 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,[44,63,64] and we hope that the APC framework developed here helps to gain insight into these models.

3.5.1 Case Study: Selecting Orbitals for Benzene

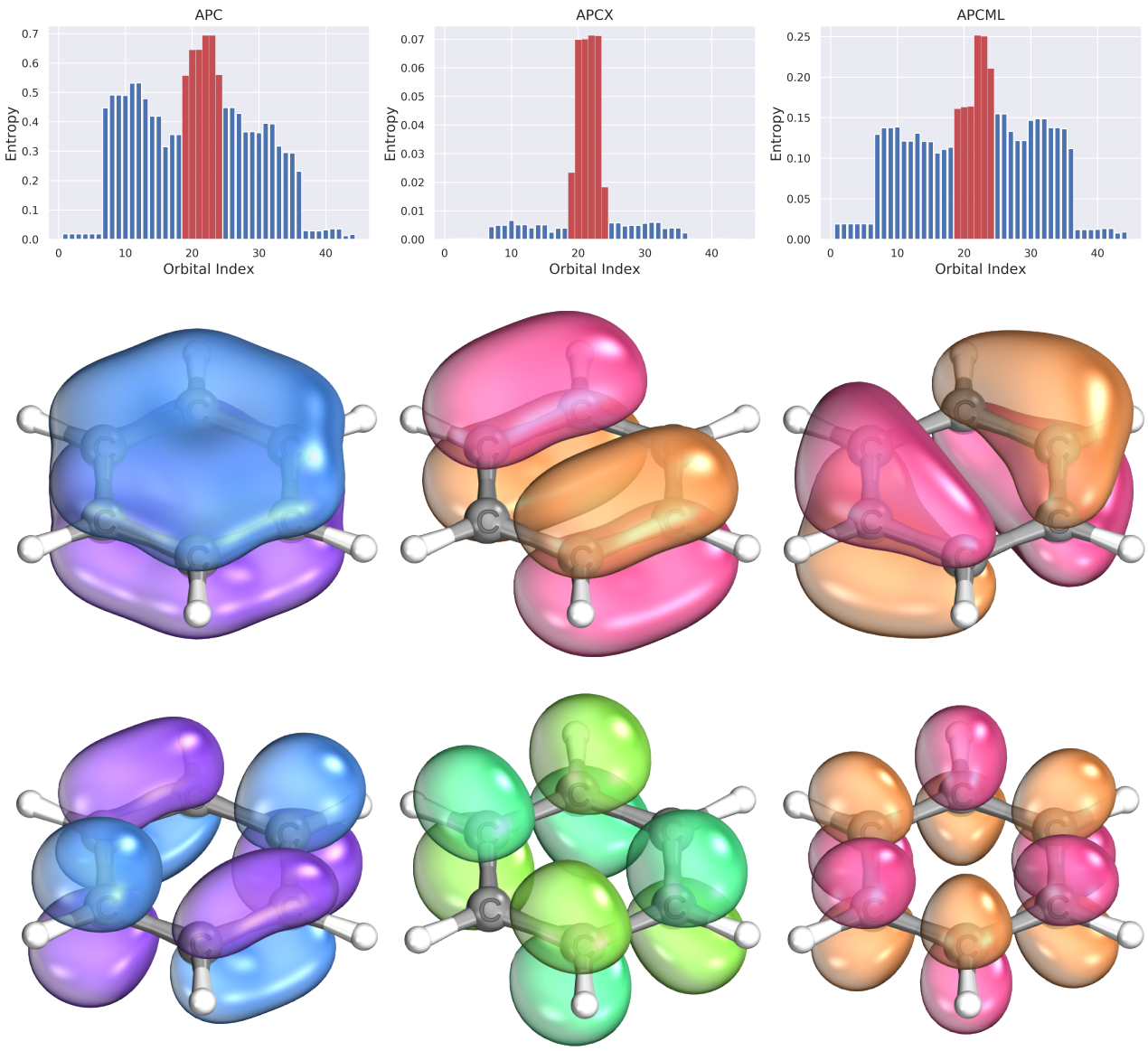

Figure 13: Top: APC, APCX, and APCML predicted entropies for all doubly occupied orbitals and the first 23 ground-state UNO orbitals highest in occupation number for the benzene geometry of Bao and Truhlar,[33] with orbitals indexed by increasing occupation number (the HOMO is orbital 21). Middle: Orbitals 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 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.[33] 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).[39,65] It is clearly seen that the APCX scheme 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, all three schemes are able to rank the same six orbitals as UNO 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.

4 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 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 AutoCAS is comparable to AutoCAS+ in performance when localized orbitals are used. Additionally, we analyzed the effectiveness of methods for estimating the error of CASSCF/NEVPT2 results, including active space overlap, N_{iter}, $\Delta E^{CASSCF}_{CASCI}$, and $|\Delta\Delta E^{NEVPT2}_{CASSCF}|$. Among these, we find $|\Delta\Delta E^{NEVPT2}_{CASSCF}|$ 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. APC entropies appear to be able to select good active spaces over many different types of orbitals, and APC-selected high-spin UNO orbitals appears to be a very effective approach for calculating the excitation energies of small molecules. 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.

Acknowledgement

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.

Supporting Information Available

Featurization and hyperparameters for APCML, further characterization and data concerning the 1120 CASSCF/NEVPT2 calculations, and performance plots of non-APC and APC selection schemes for different types of orbitals.

References

(1) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO 2 reduction and H 2 evolution. *Nature Catalysis* **2018**, *1*, 696–703.

(2) 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: Iden-

tifying the Performance Limits of Physical Hydrogen Storage. *Chemistry of Materials* **2017**, *29*, 2844–2854.

(3) Teunissen, J. L.; De Proft, F.; De Vleeschouwer, F. Tuning the HOMO–LUMO Energy Gap of Small Diamondoids Using Inverse Molecular Design. *Journal of Chemical Theory and Computation* **2017**, *13*, 1351–1365.

(4) Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *The Journal of Physical Chemistry Letters* **2013**, *4*, 1613–1623.

(5) Shu, Y.; Levine, B. G. Simulated evolution of fluorophores for light emitting diodes. *The Journal of Chemical Physics* **2015**, *142*, 104104.

(6) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catalysis* **2020**, *10*, 2354–2377.

(7) 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. *Chemical Reviews* **2019**, *119*, 2453–2523.

(8) Cramer, C. J.; Truhlar, D. G. Density functional theory for transition metals and transition metal chemistry. *Physical Chemistry Chemical Physics* **2009**, *11*, 10757.

(9) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for Density Functional Theory. *Chemical Reviews* **2012**, *112*, 289–320.

(10) Yu, H. S.; Li, S. L.; Truhlar, D. G. Perspective: Kohn-Sham density functional theory descending a staircase. *The Journal of Chemical Physics* **2016**, *145*, 130901, Publisher: American Institute of Physics.

(11) Becke, A. D. Perspective: Fifty years of density-functional theory in chemical physics. *The Journal of Chemical Physics* **2014**, *140*, 18A301.

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

(13) 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. *Chemical Reviews* **2018**, *118*, 7293–7361, Publisher: American Chemical Society.

(14) Ashley, D. C.; Jakubikova, E. Ironing out the photochemical and spin-crossover behavior of Fe(II) coordination compounds with computational chemistry. *Coordination Chemistry Reviews* **2017**, *337*, 97–111.

(15) 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. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1064–1071.

(16) Duan, C.; Liu, F.; Nandy, A.; Kulik, H. J. Semi-Supervised Machine Learning Enables the Robust Detection of Multireference Character at Low Cost. *The Journal of Physical Chemistry Letters* **2020**, *11*, 6640–6648.

(17) Aquilante, F.; Autschbach, J.; Carlson, R. K.; Chibotaru, L. F.; Delcey, M. G.; Vico, L. D.; Galván, I. F.; Ferré, N.; Frutos, L. M.; Gagliardi, L.; Garavelli, M.; Giussani, A.; Hoyer, C. E.; Manni, G. L.; Lischka, H.; Ma, D.; Malmqvist, P. ; Müller, T.; Nenov, A.; Olivucci, M.; Pedersen, T. B.; Peng, D.; Plasser, F.; Pritchard, B.; Reiher, M.; Rivalta, I.; Schapiro, I.; Segarra-Martí, J.; Stenrup, M.; Truhlar, D. G.; Ungur, L.; Valentini, A.; Vancoillie, S.; Veryazov, V.; Vysotskiy, V. P.; Weingart, O.; Zapata, F.; Lindh, R. Molcas 8: New capabilities for multiconfigurational quantum chemical calculations across the periodic table. *Journal of Computational Chemistry* **2016**, *37*, 506–541, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.24221.

(18) Roos, B. O.; Taylor, P. R.; Sigbahn, P. E. A complete active space SCF method

(CASSCF) using a density matrix formulated super-CI approach. *Chemical Physics* **1980**, *48*, 157–173.

(19) White, S. R. Density matrix formulation for quantum renormalization groups. *Physical Review Letters* **1992**, *69*, 2863–2866.

(20) White, S. R. Density-matrix algorithms for quantum renormalization groups. *Physical Review B* **1993**, *48*, 10345–10356, Publisher: American Physical Society.

(21) Keller, S.; Dolfi, M.; Troyer, M.; Reiher, M. An efficient matrix product operator representation of the quantum chemical Hamiltonian. *The Journal of Chemical Physics* **2015**, *143*, 244118, Publisher: American Institute of Physics.

(22) 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. *Journal of Chemical Theory and Computation* **2016**, *12*, 1245–1258, Publisher: American Chemical Society.

(23) 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. *The Journal of Chemical Physics* **2009**, *131*, 054106, Publisher: American Institute of Physics.

(24) Yang, P.-J.; Sugiyama, M.; Tsuda, K.; Yanai, T. Artificial Neural Networks Applied as Molecular Wave Function Solvers. *Journal of Chemical Theory and Computation* **2020**,

(25) Stein, C. J.; Reiher, M. autoCAS: A Program for Fully Automated Multiconfigurational Calculations. *Journal of Computational Chemistry* **2019**, *40*, 2216–2226.

(26) 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. *Chemical Physics* **1982**, *66*, 197–207.

(27) Angeli, C.; Cimiraglia, R.; Evangelisti, S.; Leininger, T.; Malrieu, J.-P. Introduction of n -electron valence states for multireference perturbation theory. *The Journal of Chemical Physics* **2001**, *114*, 10252–10264.

(28) Veryazov, V.; Malmqvist, P. ; Roos, B. O. How to select active space for multiconfigurational quantum chemistry? *International Journal of Quantum Chemistry* **2011**, *111*, 3329–3338.

(29) 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. *Journal of Chemical Theory and Computation* **2016**, *12*, 4274–4283.

(30) Stein, C. J.; Reiher, M. Automated Selection of Active Orbital Spaces. *Journal of Chemical Theory and Computation* **2016**, *12*, 1760–1771, Publisher: American Chemical Society.

(31) Sayfutyarova, E. R.; Sun, Q.; Chan, G. K.-L.; Knizia, G. Automated Construction of Molecular Active Spaces from Atomic Valence Orbitals. *Journal of Chemical Theory and Computation* **2017**, *13*, 4063–4078.

(32) 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. *Journal of Chemical Theory and Computation* **2018**, *14*, 2017–2025.

(33) Bao, J. J.; Truhlar, D. G. Automatic Active Space Selection for Calculating Electronic Excitation Energies Based on High-Spin Unrestricted Hartree–Fock Orbitals. *Journal of Chemical Theory and Computation* **2019**, *15*, 5308–5318.

(34) Khedkar, A.; Roemelt, M. Active Space Selection Based on Natural Orbital Occupation Numbers from n-Electron Valence Perturbation Theory. *Journal of Chemical Theory and Computation* **2019**, *15*, 3522–3536, Publisher: American Chemical Society.

(35) Sayfutyarova, E. R.; Hammes-Schiffer, S. Constructing Molecular -Orbital Active Spaces for Multireference Calculations of Conjugated Systems. *Journal of Chemical Theory and Computation* **2019**, *15*, 1679–1689.

(36) Golub, P.; Antalik, A.; Veis, L.; Brabec, J. Automatic selection of active spaces for strongly correlated systems using machine learning algorithms. *arXiv:2011.14715 [physics]* **2020**, arXiv: 2011.14715.

(37) 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. *Journal of Chemical Theory and Computation* **2020**, *16*, 2389–2399, Publisher: American Chemical Society.

(38) Khedkar, A.; Roemelt, M. Extending the ASS1ST active space selection scheme to large molecules and excited states. *Journal of Chemical Theory and Computation* **2020**,

(39) Tóth, Z.; Pulay, P. Comparison of Methods for Active Orbital Selection in Multiconfigurational Calculations. *Journal of Chemical Theory and Computation* **2020**,

(40) Zou, J.; Niu, K.; Ma, H.; Li, S.; Fang, W. Automatic Selection of Active Orbitals from Generalized Valence Bond Orbitals. *The Journal of Physical Chemistry A* **2020**, *124*, 8321–8329.

(41) Boguslawski, K.; Tecmer, P. Orbital entanglement in quantum chemistry. *International Journal of Quantum Chemistry* **2015**, *115*, 1289–1295, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qua.24832.

(42) Stein, C.; Reiher, M. Automated Identification of Relevant Frontier Orbitals for Chemical Compounds and Processes. *CHIMIA International Journal for Chemistry* **2017**, *71*, 170–176.

(43) Pulay, P.; Hamilton, T. P. UHF natural orbitals for defining and starting MC-SCF calculations. *The Journal of Chemical Physics* **1988**, *88*, 4926–4933.

(44) Welborn, M.; Cheng, L.; Miller, T. F. Transferability in Machine Learning for Electronic Structure via the Molecular Orbital Basis. *Journal of Chemical Theory and Computation* **2018**, *14*, 4772–4779.

(45) Kendall, R. A.; Dunning Jr, T. H.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *The Journal of chemical physics* **1992**, *96*, 6796–6806.

(46) 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. P y SCF: the Python-based simulations of chemistry framework. *WIREs Computational Molecular Science* **2018**, *8*.

(47) Knowles, P. J.; Handy, N. C. A determinant based full configuration interaction program. *Computer Physics Communications* **1989**, *54*, 75–83.

(48) Legeza, ; Sólyom, J. Optimizing the density-matrix renormalization group method using quantum information entropy. *Physical Review B* **2003**, *68*.

(49) Barcza, G.; Legeza, ; Marti, K. H.; Reiher, M. Quantum-information analysis of electronic states of different molecular structures. *Physical Review A* **2011**, *83*, 012508, Publisher: American Physical Society.

(50) Foster, J. M.; Boys, S. F. Canonical Configurational Interaction Procedure. *Reviews of Modern Physics* **1960**, *32*, 300–302.

(51) 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. *The Journal of Chemical Physics* **1989**, *90*, 4916–4926.

(52) Lehtola, S.; Jónsson, H. Pipek–Mezey Orbital Localization Using Various Partial Charge Estimates. *Journal of Chemical Theory and Computation* **2014**, *10*, 642–649, Publisher: American Chemical Society.

(53) Edmiston, C.; Ruedenberg, K. Localized Atomic and Molecular Orbitals. *Reviews of Modern Physics* **1963**, *35*, 457–464.

(54) Waskom, M.; the seaborn development team, mwaskom/seaborn. 2020; `https://doi.org/10.5281/zenodo.592845`.

(55) Knizia, G.; Klein, J. E. Electron Flow in reaction mechanisms—revealed from first principles. *Angewandte Chemie International Edition* **2015**, *54*, 5518–5522.

(56) Pierloot, K.; Dumez, B.; Widmark, P.-O.; Roos, B. O. Density matrix averaged atomic natural orbital (ANO) basis sets for correlated molecular wave functions. *Theoretica chimica acta* **1995**, *90*, 87–114.

(57) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. Density matrix averaged atomic natural orbital (ANO) basis sets for correlated molecular wave functions. *Theoretica chimica acta* **1990**, *77*, 291–306.

(58) 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. *The Journal of Physical Chemistry A* **2004**, *108*, 2851–2858.

(59) Stein, C. J.; Reiher, M. Measuring multi-configurational character by orbital entanglement. *Molecular Physics* **2017**, *115*, 2110–2119, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00268976.2017.1288934.

(60) Welch, B. L. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika* **1947**, *34*, 28–35.

(61) Szabo, A.; Ostlund, N. S. *Modern quantum chemistry: introduction to advanced electronic structure theory*; Courier Corporation, 2012.

(62) Liu, T.-Y. *Learning to rank for information retrieval*; Springer Science & Business Media, 2011.

(63) 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. *The Journal of Chemical Physics* **2019**, *150*, 131103.

(64) Townsend, J.; Vogiatzis, K. D. Data-Driven Acceleration of the Coupled-Cluster Singles and Doubles Iterative Solver. *The Journal of Physical Chemistry Letters* **2019**, *10*, 4129–4135.

(65) Tóth, Z.; Pulay, P. Finding symmetry breaking Hartree-Fock solutions: The case of triplet instability. *The Journal of Chemical Physics* **2016**, *145*, 164102.

Graphical TOC Entry