Variational Active Space Selection with MC-PDFT

Daniel S. King,† Donald G. Truhlar,‡ and Laura Gagliardi*¶

†Department of Chemistry, University of Chicago, Chicago IL 60637
‡Department of Chemistry, Chemical Theory Group, and Supercomputing Institute, University of Minnesota, Minneapolis MN 55455
¶Department of Chemistry, Pritzker School of Molecular Engineering, James Franck Institute, Chicago Center for Theoretical Chemistry, University of Chicago, Chicago IL 60637

E-mail: lgagliardi@uchicago.edu

Abstract

The selection of a good set of active orbitals for modeling strongly correlated quantum states is difficult to automate as it is highly dependent on the state and molecule of interest. As such, although many approaches have been proposed with some success, no single approach has worked in all cases. Here, we propose an improved framework for automated selection in which (i) multiple wave functions based on different active-spaces are generated, and (ii) the resulting wave functions are chosen between by some means of facile evaluation. Using this framework, we propose a method in which (i) we construct different active space orbitals through diagonalization of a parameterized operator, and (ii) we choose the state with the lowest tPBE absolute energy from multiconfigurational pair density functional theory (MC-PDFT) averaged over the targeted states of interest. We test the method using density matrix renormalization

1
group (DMRG) wave functions with 40 active orbitals and bond index of 700, with no further orbital optimization following the active space selection. We find that with only four values of the parameter (i.e., with only four trial wave functions), we can obtain a mean unsigned error of only 0.19 eV for 199 vertical excitation energies in the QUESTDB database. Furthermore, the tPBE absolute energy proves robust in selecting between active spaces of very different sizes, over and above the wave function (CASCI/DMRG-CASCI) absolute energy. We believe this new framework is promising for the application of active space methods to chemical problems in a high-throughput fashion.

1 Introduction

The accurate treatment of excited electronic states of molecules is a long-standing and active area of research in computational chemistry.\(^1\)\(^-\)\(^18\) In particular, the accurate treatment of excited states is difficult when a single-determinant ground state provides a poor reference for computing the excited states (e.g., double excitations\(^19\),\(^20\) or strongly correlated systems\(^21\)-\(^25\)). A useful form of wave function for overcoming such difficulties is the complete active space configuration interaction (CASCI) trial function:

\[
|\Psi_{\text{CASCI}}\rangle = |22\ldots2\rangle \land \sum_{n_1n_2\ldots n_L} C_{n_1n_2\ldots n_L} |n_1n_2\ldots n_L\rangle \tag{1}
\]

in which \(|22\ldots2\rangle\) is a single Slater determinant consisting of doubly occupied orbitals (called inactive orbitals in the CASCI context), the \(C_{n_1n_2\ldots n_L}\) are coefficients, and the determinants \(|n_1n_2\ldots n_L\rangle\) span the space of all possible configurations obtained by distributing a fixed number of active electrons \(N_{\text{elec}}\) among \(L\) active orbitals. Each determinant \(|n_1n_2\ldots n_L\rangle\) 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 \(|22\ldots2\rangle\) 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 \(C_{n_1n_2\ldots n_L}\) in equation 1.\(^{27}\) Among the most successful approaches in this regard is the density matrix renormalization group (DMRG),\(^{28-40}\) in which the coefficients of equation 1 are approximated by the matrix product\(^{32}\)

\[
C_{n_1n_2\ldots n_L} = \sum_{ij\ldots(L-1)} A_{n_1i}^{n_1} A_{n_2j}^{n_2} A_{n_3k}^{n_3} \ldots A_{n_L(L-1)}^{n_L} \tag{2}
\]

In this equation, each possible occupation of each active orbital \(n_i\) is given 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.\(^{41}\)

The success of CAS-based methods relies heavily on the selection of the orbitals for the active space; 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)\(^{42}\) when used with a full-configuration-interaction solver or as DMRGSCF 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 and SA-DMRGSCF. However, we emphasize two issues with this approach: (i) the energetic optimization is prone to converging to local minima, and (ii) variational optimization of state-averaged CASCI energies may not be optimal for computing energy differences between states. 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 selection schemes and comes at a computational cost.

Because of the above considerations, active-space selection remains a vigorous area of research, and several approaches have been proposed to date. The historically popular method involves chemical intuition with trial and error. However, this trial-and-error 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 active space selection. A tool called AVAS, developed by Sayfutyarova and coworkers, allows one to semiautomate the active-space selection by constructing 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, entanglement information from a large DMRG calculation, the quantitative accuracy of some physical observable such as the dipole moment, machine learning predictive models, and physically motivated equations based on information such as HF matrix elements. An assumption of all these approaches is that the key physics necessary to select the active space can be captured by the preliminary calculation (UHF, DMRG, etc.). However, as we will show, even when using large active spaces (e.g., with 40 orbitals) for small molecules, it is difficult to determine active spaces that are adequate for any given excitation with a single method.

In recent work, we employed one such automated approach, approximate pair coefficient (APC) active space selection, on the extensive QUESTDB database of accurate vertical excitation energies for small-to-medium-sized organic systems. Through this, we were able to carry out extensive benchmarking of post-CAS methods such as n-electron valence perturbation theory (NEVPT2) and multiconfigurational pair-density functional theory (MC-PDFT) using the translated PBE (tPBE) functional by using the active spaces generated by the automated approach. However, in order to separate errors in these post-CAS methods due to a poor active space from errors occurring even with good active spaces, we considered only the active spaces for which the state-averaged CASSCF (SA-CASSCF) result
was within 1.1 eV of the best estimate from the database. Depending on the active-space size and the basis set, this criterion was satisfied for 363–436 (68–82%) of the 532 excitations in the database. Although the APC scheme for these cases proved to be competitive with active spaces selected by hand,\textsuperscript{82–84} the very high error (sometimes greater than 5 eV) of the results obtained in the remaining active spaces (18–32% of the excitations) remained unacceptable for predictive work.

In the present study we try to provide a better framework for employing active space selection predictively. We proceed with the working assumption 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 \textit{choose} between active spaces generated with different methods. We envision a framework for CAS-based methods that involves (i) generating active orbitals with different methods (e.g., any of those mentioned or cited above) or with different parameters, (ii) solving the CI problem within the active spaces defined by these orbitals, and then (iii) choosing between the generated active-space wave functions by some means of facile evaluation. This is similar to what was done in our previous study, except that in that work the selection of CAS wave functions involved knowing the accurate excitation energies, but here the selection will use only information available from the CASSCF solutions themselves.

With the motivation of the previous paragraph, we present here a method for (a) efficiently generating a series of trial CAS wave functions for each excitation, and then (b) choosing between these wave functions with a facile evaluation criteria. To achieve (a), we select 40-orbital active spaces using APC active space selection on sets of orbitals generated through the diagonalization of different exchange-weighted operators $\lambda$; for (b), we select the CAS wave function which has the lowest MC-PDFT energy averaged over the two states involved in the transition, which we call variational selection with tPBE. We refer to the scheme in total as “$\lambda$APC-tPBE”, and find that on a test set of 199 vertical excitations from QUESTDB, this approach is able to reproduce the results of our previous benchmark of post-CAS methods.\textsuperscript{70}
with only 4 values of $\lambda$, and without the need to exclude any excitations due to poor active space selection. Furthermore, we show the robustness of variational selection with tPBE in selecting between active spaces of vastly different sizes, over and above variational selection using the wave function (CASCI/DMRG-CASCI) absolute energy.

2 Theory and Methods

In this section, we provide an overview of MC-PDFT$^{81}$ and hybrid MC-PDFT$^{85}$ and explain the active space selection scheme used in this work, which builds on our previous studies$^{67,70}$ We then provide details of the QUESTDB data and the DMRG calculations used for benchmarking the new active space selection scheme.

2.1 Multiconfiguration Pair-Density Functional Theory

The energy expression of multiconfiguration pair-density functional theory may be written as$^{81}$

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

where $V_{nn}$ is the nuclear repulsion, $i, j, k, l$ are orbital indices, $h_{ij}$ is a one-electron integral, $\gamma_{ij}$ 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 $\rho$ and the on-top pair density $\Pi$ and is called an on-top density functional. Recently, we have begun to explore different types of nonclassical-energy functionals derived from machine learning$^{86}$ or the density coherence,$^{87}$ and we refer the reader to a recent review.$^{17}$ However, all practical applications so far have employed a translated version of the PBE$^{88}$ 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, for which the energy expression becomes

\[ E_{\text{HMC-PDFT}} = X E_{\text{SA-CASSCF}} + (1 - X) E_{\text{MC-PDFT}} \]  

(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 \( X = 0.25 \), which is called tPBE0.

### 2.2 \( \lambda \text{APC-tPBE} \)

The APC method is an approach for estimating the importance of including selected orbitals in the active space on the basis of HF exchange matrix elements \( K_{ii} \) and Fock matrix elements \( F_{ii} \). Recently, we have used an active space selection scheme built around this method to select CASSCI active spaces for all molecules in the QUESTDB dataset. Here, we modify this scheme in two key ways: (1) we orient the method towards selecting active spaces for DMRG-CASCI calculations (without orbital optimization) instead of CASSCF, and (2) instead of generating one active space for a given excitation, we generate several active spaces and choose between them variationally with tPBE. We call the new active space selection scheme "\( \lambda \text{APC-tPBE} \)", and it is composed of four steps: (i) orbital construction; (ii) APC entropy calculation; (iii) orbital selection with APC entropies; and (iv) variational selection with tPBE. In the following we explain these components step-by-step.

**Orbital Construction.** We start with the full set of orbitals obtained by a closed-shell HF calculation on the ground state, and we index the doubly occupied orbitals with \( i \) and the virtual orbitals with \( a \). We generate multiple trial active spaces for a calculation by diagonalizing the operator

\[ G = F - \lambda K \]  

(5)

with multiple values of the parameter \( \lambda \), where \( F \) and \( K \) are the Fock and exchange matrices.
generated from the HF density matrix. The operator is diagonalized within the space of all virtual orbitals. Once these virtual orbitals have been generated, the manifold of doubly occupied orbitals $i$ and virtual orbitals $a$ remains unchanged; the next steps serve to select 40 active orbitals and a set of inactive orbitals from this initial manifold.

**APC Entropy Calculation.** For a doubly occupied orbital $i$ and virtual orbital $a$, the approximate pair coefficient (APC) between these two orbitals is defined by

$$C_{ia} = \frac{0.5K_{aa}}{F_{aa} - F_{ii} + \sqrt{(0.5K_{aa})^2 + (F_{aa} - F_{ii})^2}}$$  \hspace{1cm} (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

$$S_i = -\frac{1}{1 + \sum_a C_{ia}^2} \ln \left( \frac{1}{1 + \sum_a C_{ia}^2} - \frac{\sum_a C_{ia}^2}{1 + \sum_a C_{ia}^2} \ln \left( \frac{\sum_a C_{ia}^2}{1 + \sum_a C_{ia}^2} \right) \right)$$  \hspace{1cm} (7)

and

$$S_a = -\frac{1}{1 + \sum_i C_{ia}^2} \ln \left( \frac{1}{1 + \sum_i C_{ia}^2} - \frac{\sum_i C_{ia}^2}{1 + \sum_i C_{ia}^2} \ln \left( \frac{\sum_i C_{ia}^2}{1 + \sum_i C_{ia}^2} \right) \right)$$  \hspace{1cm} (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, 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 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 $N = 2$, which is used in this work. The entropies obtained after the $N$ removals are now used for the rest of
the procedure without recalculation.

*Orbital Selection.* Considering the entire manifold of doubly occupied orbitals $i$ and virtual orbitals $a$, 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 in the CASCI calculation, whereas dropped virtual orbitals become secondary in the CASCI calculation.

As a result of the above steps, there are always 40 active orbitals in the active space, and the number of inactive orbitals is the number of doubly occupied orbitals dropped from the active space in the above selection stage. The number of active electrons 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.

*Variational Selection with tPBE.* The above steps are repeated for $N_{\text{trial}}$ values of $\lambda$, resulting in the generation of $N_{\text{trial}}$ trial active spaces. Wave functions are computed using each of the $N_{\text{trial}}$ active spaces with CASCI calculations carried out by DMRG. We then select the active space that has the lowest average tPBE energies between the ground state and the excited state of interest. This concludes the description of $\lambda$APC-tPBE.

*Comparison to Previous Method.* The main difference of $\lambda$APC-tPBE from previous schemes is the generation of multiple trial active spaces by variation of the $\lambda$ parameter in equation 5 followed by variational selection from among the trial active spaces with tPBE. There are also two smaller changes motivated by the goal of choosing large active spaces for DMRG (rather than the previous goal of choosing smaller active spaces that are affordable with conventional solvers for the CASCI equations). First, we no longer limit selection to orbitals close in energy to the HOMO and LUMO; we instead select from a complete manifold of molecular orbitals. Second, instead of dropping orbitals until a size constraint is met based on the number of configurations in the CAS space, we impose a limit on the number of active orbitals. The reason for the latter change is because although the cost of conventional
CASCII scales directly with the number of configurations, the cost of DMRG depends more directly on the number of orbitals than on the number of configurations.\textsuperscript{35}

2.3 Benchmarking Data

We test the new active space selection scheme on a subset of the theoretical best estimates in the QUESTDB dataset\textsuperscript{70} of vertical excitation energies of small-to-medium-sized molecules. The subset is formed by excluding some QUESTDB excitations according the following conditions:

- 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).\textsuperscript{78}

- 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. Additionally, for ease of development and use of DMRG, we have excluded consideration of the 51 doublet excitations in the QUESTDB dataset. However, the proposed scheme is straightforwardly generalizable to these cases by use of ROHF orbitals, and we have addressed these cases in previous work.\textsuperscript{70} After eliminating data according to these criteria, we are left with a set of 199 excitations for testing the present method.

The $2^1 A_g$ State of Ethylene. Special attention is given to the theoretical best estimate listed in QUESTDB for the $2^1 A_g$ state of ethylene, which is characterized by Loos and and
coworkers as a valence \((\pi, \pi) \rightarrow (\pi^*, \pi^*)\) double excitation at roughly 12.15 eV, referencing a 2004 study by Barbatti et al.\cite{20,89} However, in the comprehensive 2014 study on the excited states of ethylene carried out by Feller et. al.,\cite{90} the \(2^1A_g\) state of ethylene is clearly characterized by both experiment and theory as a single \((\pi, 3p)\) Rydberg excitation at about 8.45 eV. Although we have been able to converge to the double excitation in the \(1^1A_g\) 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.\cite{90} 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.\cite{90}

### 2.4 Computational Methodology

Density matrix renormalization group calculations were carried out without re-optimization of orbitals (in particular, SA-DMRG-CASCI) by using the state-averaged DMRG-CASCI in BLOCK2\cite{91} as implemented in PySCF.\cite{92,93} This will usually be called DMRG in the rest of the article (except where we wish to emphasize the distinction from SCF calculations). The maximum bond dimension (number of renormalized states) was fixed to \(M = 700\). Calculations were done with the aug-cc-pVTZ basis\cite{94,95} as was used for the theoretical best estimates listed in the QUESTDB database. MC-PDFT calculations were carried out using a version of PySCF that incorporates the MRH code\cite{96} now available in PySCF-FORGE.\cite{97} 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.\cite{70}

As in our previous study, 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 \(1^1A_1\) and two states of symmetry \(1^1A_2\), as in our previous work\cite{70}). Final excitation energies were calculated by subtracting the ground state tPBE0 energy from the tPBE0 energy of the
excited state of interest (in the example, that would be the second root in \(^1\text{A}_2\) symmetry).

## 3 Results

30-Excitation Tests. We first show the robustness of the new \(\lambda\text{APC-}\text{tPBE}\) active space selection scheme by carrying out calculations for the 30 excitations for which our previous approach with the aug-cc-pVTZ basis had a \(\text{tPBE0}\) error greater than 0.55 eV; these 30 excitations involve a set of 17 molecules. We generated nine active spaces for each excitation (i.e., \(N_{\text{trial}} = 9\)) by the method explained in section 2.2. In particular we used \(\lambda\) equal to 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, and 2.0.

Figure 1: Left: Averaged squared distance from the centroid \(<r_c \cdot r_c>\) 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 1 shows trends in the orbital character of the selected active orbitals as \(\lambda\) is varied from 0 to 2 for the 17 molecules present in the 30-excitation test subset. The left side of Figure 1 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 $\lambda$ lead to significantly different average diffuse character of the selected orbitals, with the most diffuse character for $\lambda = 0.75$. The right side of Figure 1 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 $\lambda$ 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, while $\lambda = 0.75$ converges to the lower-energy singly excited Rydberg state (Supporting Information).

We next examine excitation energies calculated by tPBE0. To prevent confusion, we stress that although the $\lambda$APC-tPBE selection scheme employs the tPBE functional, 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 and as also discussed below.).

The left side of Figure 2 shows the absolute error of the tPBE0 calculations of the excitation energies with active spaces generated by the nine values of $\lambda$. The figure shows that no single value of $\lambda$ yields accurate results for all 30 cases. For each value of $\lambda$, 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 $\lambda$ that give an absolute error less than 0.55 eV. This motivated the use of a variational scheme to find the best value of $\lambda$ for each case.
As mentioned above, the criterion we use in the λAPC-tPBE selection scheme is to choose the active space that gives the lowest average tPBE absolute energy between the ground state and the excited state under investigation. Next we compare using the tPBE energy for variational selection to using the wave function energy, which can be called λAPC-CASCI. The right side of Figure 2 compares these two approaches to random selection 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 CASCI does little better than random chance, which does very poorly. The mean absolute error of the tPBE0 excitation energies is 1.24 eV with CASCI 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 CASCI selection to 0.63 eV with tPBE selection. As a hybrid between tPBE and CASCI, 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\cite{41} or the sum of occupation number deviations from zero or 2 are unable to select well between the different values of \(\lambda\) (see the Supporting Information for details of these tests).

199-Excitation Tests. We next consider the performance of \(\lambda\)APC-tPBE on the entire set of 199 excitations in the QUESTDB database that meet the selection criteria of section 2.3. For this larger test, we used only four values of \(\lambda\) (i.e., \(N_{\text{trial}} = 4\)) to generate active spaces: \(\lambda = 0.25, 0.5, 0.75,\) and 1.25. These values of \(\lambda\) 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 3 shows the mean absolute errors achieved by CASCI, tPBE, and tPBE0 transition energy calculations with \(\lambda\)APC-tPBE active-space selection for the full set of 199 excitations. These results are compared to our previous benchmark for the subset that excluded poor active spaces (those 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:

- CASCI: 0.47 eV presently vs. 0.37 eV previously.
- tPBE: 0.20 eV presently vs. 0.22 eV previously.
- tPBE0: 0.18 eV presently vs. 0.19 eV previously.

We note that the performance of CASSCF is slightly worse, as might have been expected due to the bias of the previous benchmark in excluding 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. Without excluding poor active spaces, our previous scheme returns a tPBE0 mean absolute error of

https://doi.org/10.26434/chemrxiv-2023-9b3g4 ORCID: https://orcid.org/0000-0003-0208-5274 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0
error of 0.40 eV, more than twice as large as for the subset.

Figure 3: Left: Mean absolute errors achieved by CASCI, tPBE, and tPBE0 on the 199-excitation test set with active spaces selected by λAPC-tPBE compared to the active spaces used in our previous benchmark (after eliminating 39 poor active spaces). Right: Comparison of number of wave functions variationally selected with tPBE vs. number of wave functions variationally selected with CASCI at each value of λ.

The key to the success of tPBE selection as compared to CASCI selection seems to be that it chooses lower values of λ. The right panel of Figure 3 shows the frequency with which each value of λ was chosen in the tPBE selection compared to selection with CASCI. 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 CASCI, for which the selected values of λ are instead clustered heavily around λ = 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 λ = 2. In that case we found that the selections by CASCI are clustered around λ = 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, but here not excluding
any wave function due to poor active spaces. We use tPBE to variationally select among five trial active-space wave functions: the previous SA-CASSCF wave functions, which have relatively small active spaces (typically having about 12 active orbitals) and four SA-DMRG-CASCI wave functions generated with the four values of $\lambda$ (large active spaces with 40 active orbitals and $M = 700$). We label this broader selection scheme as (\lambda APC+CASSCF)-tPBE, and the left panel of Figure 4 shows the results of this scheme compared to simply using the previous SA-CASSCF wave functions (again, not excluding any active spaces due to poor selection). Although the performance of (\lambda APC+CASSCF)-tPBE is slightly reduced compared to \lambda APC-tPBE alone (0.21 eV tPBE0 error vs. 0.19 eV), variational selection with tPBE is able to robustly discriminate against the outlier SA-CASSCF wave functions.

Figure 4: Left: Violin plots comparing the distribution of errors for three kinds of energy calculation (CASCI, tPBE, and tPBE0) on the 199-excitation test set when using tPBE to select among the previous SA-CASSCF wave functions and the four SA-DMRG-CASCI wave functions (using active-space orbitals selected using the new scheme presented in this paper with $\lambda = 0.25, 0.5, 0.75, 1.25$) 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 CASCI to select. Note that the selection among five trial wave functions is labeled in the plot as (\lambda APC+CASSCF)-tPBE, and the previously generated SA-CASSCF wave functions are labeled as "Previous Benchmark."
The right panel of Figure 4 again compares the distribution of wave functions variationally selected (among these five trial wave functions) by tPBE to those selected by CASCI. The figure shows that, although the active-space wave functions selected by (λAPC + CASSCF)-tPBE have significantly smaller mean absolute errors (0.21 vs. 0.40 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 CASCI yields mainly wave functions generated with high values of λ and hardly any of the SA-CASSCF wave functions from our previous work.\textsuperscript{70}

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 multi-configurational wave functions. We obtain good results both with the selection among four DMRG wave functions and with the selection among five trial wave functions, although in the latter case the DMRG wave function 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.

4 Concluding Remarks

The goal of this work was to develop an automatic scheme for selecting active spaces for the calculation of vertical excitation energies in a predictive fashion. Towards this goal, we have presented an active space selection scheme in which active orbitals are generated by applying entropy calculations based on approximate pair coefficients to orbitals obtained by Hartree-Fock calculations and by diagonalization of a parameterized exchange-weighted
operator in the space of HF virtual orbitals. By using multiple values of the parameter, we generate multiple sets of active and inactive orbitals. Then, for each of the orbital sets, we calculate a DMRG wave function, and we variationally choose among these trial wave functions the one that gives the lowest average absolute energy between the ground state and the excited state under consideration. We refer to this scheme as λAPC-tPBE, and we find that, even with only four trial wave functions, it performs robustly at selecting good active spaces for the QUESTDB database of vertical excitation energies. Furthermore, if we enlarge the trial set of wave functions to include both SA-CASSCF wave functions with small active spaces and the new DMRG wave functions with large active spaces (40 active orbitals), and we chose 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 tPBE variational selection is able to choose robustly between active spaces of vastly different sizes, over and above variational selection using the wave function (CASCI/DMRG-CASCI) energy.

In summary, we have developed both a method of automatically generating active spaces and a variational method for choosing which one will perform best for vertical excitation energies. Our results show that the new method can 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, variational selection with tPBE appears robust enough in selecting between wave functions, and so a mixture of different active space construction schemes in tandem with tPBE selection seems to be a promising approach.

Although we have proposed construction of active spaces via the diagonalization of $F - \lambda K$ in the space of Hartree-Fock virtual orbitals followed by APC-based entropy calculations, the procedure of variational tPBE selection among trial wave functions can be combined with other schemes for constructing multiconfiguration wave functions. We note that although the scheme has proved successful for the excitations shown here, we expect that
applying the scheme to metal-containing systems and adiabatic excitations may require additional development. Nevertheless, a specific strength of this approach for applications with a large number of orbitals is that it forgoes the costly step of multiconfigurational orbital optimization.

5 Acknowledgments

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.

Supporting Information Available

Figures referred to in the main text and additional data analysis.

References


(20) Loos, P.-F.; Boggio-Pasqua, M.; Scemama, A.; Caffarel, M.; Jacquemin, D. Reference


(37) Yanai, T.; Kurashige, Y.; Mizukami, W.; Chalupský, J.; Lan, T. N.; Saitow, M. Density matrix renormalization group for ab initio Calculations and associated dynamic corre-


(64) 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. *The Journal of Chemical Physics* 2021, 154, 090902.


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


