A Localized-Orbital Energy Evaluation for Auxiliary-Field Quantum Monte Carlo

John L. Weber,∗,† Hung Vuong,† Pierre A. Devlaminck,† James Shee,‡ Joonho Lee,† David R. Reichman,† and Richard A. Friesner∗,†

†Department of Chemistry, Columbia University, 3000 Broadway, New York, NY, 10027
‡Kenneth S. Pitzer Center for Theoretical Chemistry, Department of Chemistry, University of California, Berkeley, California 94720, USA

E-mail: jlw2245@columbia.edu; raf8@columbia.edu

Abstract

Phaseless Auxiliary-Field Quantum Monte Carlo (ph-AFQMC) has recently emerged as a promising method for the production of benchmark-level simulations of medium to large-sized molecules, due to its accuracy and favorable polynomial scaling with system size. Unfortunately the memory footprint of standard energy evaluation algorithms are non-trivial, which can significantly impact timings on graphical processing units (GPUs) where memory is limited. Previous attempts to reduce scaling by taking advantage of the low rank structure of the Coulombic integrals have been successful, but are significantly limited by high prefactors, rendering the utility limited to very large systems. Here, we present a complementary, cubic scaling route to reduce memory and computational scaling based on the low rank of the Coulombic interactions between localized orbitals, focusing on the application to phaseless AFQMC. We show that the error due to this approximation, which we term Localized Orbital AFQMC (LO-AFQMC), is systematic and controllable via a single variable, and is computationally favorable even for small systems. We present results demonstrating a robust
retention of accuracy versus both experiment and full ph-AFQMC for a variety of
test cases chosen for their potential difficulty for localized orbital based methods, in-
cluding the singlet-triplet gaps of polyacenes benzene through pentacene, the heats of
formation for a set of platonic hydrocarbon cages, and the total energy of ferrocene
(Fe(Cp)$_2$). Finally, we reproduce our previous result of the gas phase ionization energy
of Ni(Cp)$_2$, agreeing with full ph-AFQMC to within statistical error while using less
than a fifteenth of the computer time.

1 Introduction

Solving the electronic structure of large, strongly correlated molecules at zero temperature
represents a quintessential challenge of quantum chemistry. Quantum Monte Carlo meth-
ods, in particular projector-based Monte Carlo techniques such as Auxiliary-Field Quan-
tum Monte Carlo (AFQMC), have emerged as promising approaches capable of producing
benchmark results for small, strongly correlated systems.$^1$ This method relies on solving the
Schroedinger equation in imaginary time $\tau$, projecting out the exact ground state $|\phi_0\rangle$ under
the requirement that the initial state $|\phi_i\rangle$ has a non-zero overlap, $\langle \phi_i | \phi_0 \rangle \neq 0$

$$|\phi_0\rangle \propto \lim_{\tau \to \infty} e^{-\tau H}|\phi_i\rangle.$$

AFQMC does so by reformulating the many body propagator via a Hubbard-Stratonovich
transformation to an integral of a set of one body propagators coupled to auxiliary fields.
This large multi-dimensional integral is estimated by Monte Carlo sampling a random walk
of non-orthogonal Slater determinants, called walkers. Each walker can be propagated inde-
pendently, aside from occasional communication to maintain efficient sampling populations
(see later discussion on population control), and so this algorithm is nearly embarrassingly
parallel. This allows one to make effective use of graphical processing units (GPUs), in
favorable cases resulting in a $\simeq 200\times$ reduction in the prefactor. Aside from a few model
systems, however, one encounters a fermionic sign problem (or phase problem in the case of AFQMC). This leads to an exponential decrease in the signal to noise ratio, and thus an exponential increase in the required computational effort to obtain statistically meaningful results as the system grows. Algorithms for which the phase problem is uncontrolled, i.e. free projection AFQMC (fp-AFQMC), are formally exact, and the use of systematically better trial wavefunctions can reduce variance and allow convergence of exact results for small systems. In order to render AFQMC polynomially scaling, the phase problem can be avoided via the phaseless approximation (ph-AFQMC), where we remove any phase accumulated in the weight by imposing a boundary condition based on the trial wavefunction. This removes the phase problem and allows extended simulations of realistic materials, at the expense of a systematic bias with respect to the trial wavefunction.

The success of ph-AFQMC for \textit{ab initio} systems is in large part dependent upon the ability to systematically increase the quality of the results by increasing the quality of the trial wavefunction. Thus a fundamental question in the field centers on the optimal construction of trial functions, and there have been many studies exploring the space of quantum chemistry methods for efficient and accurate trial selection. In some cases, mean-field trial wavefunctions such as Hartree-Fock (HF) or Kohn-Sham Density Functional Theory (KS-DFT) are sufficient. In many cases, however, multi-determinant trials are necessary, particularly in transition metal containing complexes, with Configuration Interaction (CI) expansions including Complete Active Space Self-Consistent Field (CASSCF), Non-Orthogonal CI (NOCI), and Semi-stochastic Heat-bath CI (SHCI) having been used for this purpose. Interestingly, even in some transition metal systems which are not expected to exhibit strong static correlation effects, multi-determinant trials have been necessary for ph-AFQMC to converge to within experimental results. However, use of a multideterminantal trial function leads to increased costs in both generating the trial and using it in AFQMC propagation and energy evaluation. Addressing these issues is a key challenge inherent in the development of a scalable AFQMC approach with benchmark accuracy across a wide range of chemical
problems.

For transition metal containing molecules, we have found that it is often necessary to use a multi-determinantal trial such as one provided by a CASSCF wavefunction (AFQMC/CAS) with a limited number of determinants \(N_{\text{det}} \simeq 300 - 1000\), in order to converge the result to chemical accuracy or near chemical accuracy.\(^8\)-\(^10\) However, the use of AFQMC/CAS has been limited to small, single transition-metal systems (< 1000 basis functions, 200 electrons), in part due to the enhanced computational cost and scaling of the local energy evaluation. Optimization of this step, therefore, may allow for the extension of AFQMC to larger transition metal systems of importance in biology, catalysis, and materials science.

For single determinant trials, the propagation of a walker within ph-AFQMC algorithms exhibits cubic scaling, whereas the energy evaluation exhibits quartic scaling. The most expensive steps in the propagation are the formation of the propagator, as well as the formation of the a shift in the auxiliary fields necessary for efficient importance sampling,\(^16\) both of which scale as \(XM^2\), or asymptotically \(O(M^3)\), where \(M\) is the full basis dimension and \(X\) is defined as the number of auxiliary fields, \(X \simeq 4 - 10M\) for Cholesky decomposition thresholds between \(10^{-4}\) and \(10^{-5}\) Ha.\(^17\) In the most straightforward implementation,\(^18\) the energy evaluation has a cost of \(O(N^2M^2)\), where \(N\) is the number of electrons in the molecule. There have been numerous attempts to decrease the scaling of the energy evaluation for single determinant AFQMC. Many take advantage of the low-rank nature of the ERIs beyond that of the literature standard, namely the modified Cholesky decomposition. For example, Motta \textit{et al.} formed a low-rank compression of each individual Cholesky matrix for single determinant trials.\(^19\) Additionally, Tensor Hyper-Contraction (THC) has emerged as a cubically scaling method with only quadratically scaling memory usage, albeit with a large prefactor that limits the utility of the approach to systems with \(M > 2000\) to 4000.\(^20,21\) In addition to these low rank algorithms, Lee and Reichman applied a stochastic resolution of the identity (sRI) to reduce the scaling of the sum over Cholesky matrices, resulting in a cubic scaling algorithm with minimal overhead, although the accuracy has not been demon-
strated for correlated systems.\textsuperscript{22} It is also possible to take advantage of the sparsity of the Hamiltonian, where in a specified representation only elements above a certain threshold are explicitly stored. However, linear operations on sparse matrices are fairly inefficient, particularly on graphical processing units (GPUs), and thus doing so is only favorable in cases with extreme sparsity where nearly 90\% of the elements are removed, which is in our experience typically not the case for molecular systems.

When extending the trial to include multiple determinants, there are many choices of algorithms which take advantage of the excitation structure of CI expansions. The simplest involves using the Sherman Morrison Woodbury (SMW) algorithm to update the overlap between the walker and trial determinants.\textsuperscript{23} This approach significantly reduces the computational cost of multideterminantal propagation, but results in a scaling of $O(N_{\text{det}} N^2 M^2)$ for the energy evaluation, where $N_{\text{det}}$ is the number of determinants in the trial, resulting in a quickly intractable code for large determinant expansions. Recently, Mahajan \textit{et al.} have described an algorithm for the local energy evaluation of AFQMC with multi-determinant trials, which effectively reduces the scaling dependence on the number of determinants to $O(N M^3 + N_{\text{det}} M)$.\textsuperscript{2,15,24} This algorithm enables the use of extremely accurate trial wavefunctions based on large SHCI expansions, which in turn reduces the variance of free-projection AFQMC enough to treat significantly larger systems than previously possible. While this opens the door to the usage of larger determinant expansions, for large systems the energy evaluation is still quartic, and thus remains the effective bottleneck. It is possible to modify THC to use within this framework, but this is again subject to the associated large prefactor and is likely to be useful only for very large systems.

Here we report an alternative (and in some cases complementary) method, directly applicable to both single determinant and multideterminant CI expansion trials, which takes advantage of the low rank of the electron repulsion integrals (ERIs) when working in a localized orbital basis. It does so by compressing the block of the “half-rotated” ERI tensor corresponding to the interaction between each electron pair $[ij]$ using singular value decom-
position (SVD), which is systematically controllable by an energetic threshold in addition to being compatible with dense linear algebra routines. This results in a reduced memory scaling from quartic to cubic for this tensor, with a concomitant reduction of scaling for the energy evaluation for a single determinant. We implement and outline the use of these localized ERIs within the SMW algorithm for multideterminant CAS trials, resulting in an energy evaluation algorithm that scales as \( \langle M_{SVD} \rangle (N^2 M + N_{\text{det}} N^2) \), where \( \langle M_{SVD} \rangle \) represents the average rank of the compressed [ij] block, and we omit other values which are constant with system size, such as the maximum number of excitations from the reference for the set of determinants, \( \epsilon \). This low rank localized structure is additionally compatible with the modified generalized Wick’s theorem approach of Mahajan et al., and we outline one such algorithm in the SI, which results in a theoretical scaling of \( \langle M_{SVD} \rangle N^2 MA + N_{\text{det}} \), where \( A \) is the size of the active space. We move on to show the accuracy of Localized Orbital AFQMC (LO-AFQMC) in population control ph-AFQMC to within the statistical error of ph-AFQMC (\( \simeq 1 \text{ mHa} \)) for a variety of representative molecules of nontrivial size and complexity, including polyacenes, metallocenes, and a set of benchmark platonic hydrocarbon cages.

We emphasize that although we focus on the application of this localized orbital compression to a particular set of algorithms within ph-AFQMC, much of this work is applicable to other orbital based Quantum Monte Carlo methods. We expect that the relative simplicity of the approach described here will enable these extensions. Due to the many advantages of GPU parallelization, we make efforts in this work to tailor our algorithm towards GPU architectures by maintaining an emphasis on both the reduction of memory requirements and the ability to perform operations efficiently in parallel using standard CUDA libraries. The manuscript is organized as follows: In Sec. 2 we provide a brief overview of AFQMC (for a more in depth discussion see Ref 17), along with relevant multi-determinant algorithms for the local energy evaluation. In Sec. 3 we present the localized-orbital based algorithm, including some discussion of optimizing an implementation on GPU architectures. In Sec.
4 we then present the accuracy, memory usage, and timings for a series of test molecules. Lastly, we discuss the impact of such results, and conclude with our current outlook for AFQMC calculations in medium to large systems with strong correlation.

2 Theory

2.1 Overview of ph-AFQMC

Here we give a brief introduction to the framework of AFQMC; for a more in depth review of the method, we suggest some recent reviews.\textsuperscript{18,26} In AFQMC, initial states are propagated in imaginary time $\tau$ according to

$$\left| \phi(\tau + \Delta \tau) \right\rangle = e^{-\Delta \tau \hat{H}} \left| \phi(\tau) \right\rangle = \int d\mathbf{x} P(\mathbf{x}) \hat{B}(\mathbf{x}) \left| \phi(\tau) \right\rangle \approx \sum_w \hat{B}(\mathbf{x}_w) \left| \phi_w(\tau) \right\rangle,$$  \hspace{1cm} (2)

where $\hat{H}$ denotes the electronic Hamiltonian, $\hat{H} = \hat{H}_1 + \hat{H}_2 = \sum_{pq} h_{pq} c^\dagger_p c_q + \frac{1}{2} \sum_{pqrs} V_{pqrs} c^\dagger_p c^\dagger_q c_s c_r$, $\left| \phi_w \right\rangle$ denotes the Slater determinant associated with walker $w$, and $\hat{B} \left| \phi_w(\tau) \right\rangle = \left| \phi_w(\tau + \Delta \tau) \right\rangle$ is another Slater determinant with rotated orbitals as given by the Thouless theorem.\textsuperscript{27} $V_{pqrs}$ are the two electron integrals referred to as $(pr|qs)$ and $(pq|rs)$ in chemists and physicists notation, respectively. For long imaginary times, computing observables using this Monte Carlo representation of the wavefunction will recover ground-state properties as long as the initial wavefunction has a non-zero overlap with the ground state. Using a symmetric Suzuki-Trotter decomposition, we separate the one- and two-body terms in $\hat{H}$ with an error quadratic in the imaginary time $\tau$,

$$e^{-\Delta \tau (\hat{H}_1 + \hat{H}_2)} \approx e^{-\frac{\Delta \tau \hat{H}_1}{2}} e^{-\frac{\Delta \tau \hat{H}_2}{2}} e^{-\frac{\Delta \tau \hat{H}_1}{2}} + \mathcal{O}(\Delta \tau^3).$$  \hspace{1cm} (3)

In practice, we mitigate the Trotter error by restricting our calculations to a small timestep, $\Delta \tau$ (0.005 Ha\textsuperscript{-1} in this work). If we write the electronic two-body operator as a sum of one-
body operators squared, \( V_{pqrs} = \sum_\alpha L_{pr,\alpha} L_{qs,\alpha} \), which can be accomplished exactly via diagonalization or approximately via a density fitting or a modified Cholesky decomposition,\(^{28}\) we can then use the Hubbard-Stratonovich identity to convert the two-body operators into a multi-dimensional integral over a set of fluctuating “auxiliary-fields” \( x_\alpha \),

\[
e^{-\frac{\Delta \tau}{2} (\sum_\alpha L^2_\alpha)} = \prod_\alpha \int_\alpha^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2_\alpha}{2}} e^{\sqrt{\Delta \tau} x_\alpha L_\alpha} dx_\alpha + \mathcal{O}(\Delta \tau^2). \tag{4}
\]

It is this multi-dimensional integral on which we perform Monte Carlo sampling

\[
|\phi(\tau + \Delta \tau)\rangle = \prod_\alpha \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2_\alpha}{2}} e^{\sqrt{\Delta \tau} x_\alpha L_\alpha} dx_\alpha |\phi(\tau)\rangle = \int dx P(x) \hat{B}(x) |\phi(\tau)\rangle, \tag{5}
\]

where \( x \) is the vector of auxiliary fields. This Monte Carlo simulation can be reformulated as a branching, open-ended ensemble of random walkers \( w \) over the manifold of Slater determinants, each represented by a single Slater determinant \( \phi_{\tau,w} \) and corresponding weight \( W_{\tau,w} \) and overlap with the trial \( \langle \Phi_T | \Phi_{\tau,w} \rangle \). As each walker is propagated forward by \( \hat{B}(x_{\tau,w}) \), with the space of auxiliary fields \( x \) being sampled from the Gaussian probability defined in Eq. 4, the weights are updated according to the ratio of the new overlap with the trial to the old overlap

\[
|\Phi_{\tau+\Delta \tau,w}\rangle = \hat{B}(x_{\tau,w}) |\Phi_{\tau,w}\rangle, \tag{6}
\]

\[
W_{\tau+\Delta \tau,w} e^{i\theta_{\tau+\Delta \tau,w}} = \frac{\langle \Phi_T | \Phi_{\tau+\Delta \tau,w} \rangle}{\langle \Phi_T | \Phi_{\tau,w} \rangle} W_{\tau,w} e^{i\theta_{\tau,w}}. \tag{7}
\]

All theory thus far is formally exact; however, the fermionic phase problem leads to an exponential decrease in signal-to-noise ratio as the walkers are propagated, resulting in an exponentially growing population of walkers (and thus computational time) necessary to achieve a given statistical error. In order to mitigate this noise, we perform importance sampling by shifting the auxiliary fields to favor sampling in regions with high(er) overlap with the trial wavefunction.\(^{16,29}\) This is complemented by the phaseless constraint, in which we multiply each walker’s weight by a factor corresponding to projecting the accumulated
phase with respect to the trial wavefunction back onto the real axis, namely we multiply by \( \max(0, \cos(\Delta \theta)) \).

Within both free projection and phaseless AFQMC, there are numerous ways to perform the sampling over auxiliary fields. It is advantageous for computational efficiency to restrict the number of walkers to some set value. This is most typically done by periodically annihilating walkers with low weights and duplicating those with high weights with a probability proportional to the distance of the walker’s weight from one, and keeping the total number of walkers constant, referred to as Population Control (PC). PC introduces a bias which scales linearly with the inverse number of walkers, which is typically minimal when running with hundreds to a few thousands of walkers, and in the absence of the phase problem leads to stable calculations over hundreds of Ha\(^{-1}\). In this work we implement population control via a “comb” algorithm,\(^ {30,31} \) doing so every 20 time steps, along with measurements of energy. Additionally, the walkers themselves need to be periodically orthonormalized for numerical stability; we do so every two steps. A mean field subtraction is performed prior to propagation to reduce variance.\(^ {17} \)

As an alternative to PC, we have recently introduced a sampling approach based on direct calculation of energy differences between two relevant states (e.g. +2 and +3 states to compute ionization potentials) which is designated “correlated sampling” (CS).\(^ {32} \) For problems which can be effectively formulated to utilize CS, i.e. similar geometries, it can provide an attractive combination of speed and accuracy, at least for the systems tested to date.\(^ {8–10} \) While we do not report values for free projection or correlated sampling in this work, we note that as the local energy evaluation remains identical for both, we expect LO-AFQMC to exhibit similar results.
2.1.1 Evaluation of the Local Energy

The local energy associated with a walker is defined as

\[ E = \frac{\langle \phi_T | \hat{H} | \phi \rangle}{\langle \phi_T | \phi \rangle}. \tag{8} \]

The two-body contribution to the energy can be written as

\[ E_2 = \frac{1}{2} \sum_{pqrs}^M V_{pqrs} \frac{\langle \phi_T | a_p^\dagger a_q^\dagger a_s a_r | \phi \rangle}{\langle \phi_T | \phi \rangle}, \tag{9} \]

which, when the generalized Wick’s theorem is used, gives

\[ E_2 = \frac{1}{2} \sum_{pqrs}^M V_{pqrs} \sum_{\sigma \tau} (G_{p\sigma,q\tau} G_{q\tau,s\sigma} - G_{p\sigma,s\tau} G_{q\tau,r\sigma}), \tag{10} \]

where \( \sigma \) and \( \tau \) are spin indices. The equal-time Green’s function is defined as

\[ G_{p\sigma,q\tau} = \frac{\langle \phi_T | a_p^\dagger a_{q\tau} | \phi \rangle}{\langle \phi_T | \phi \rangle} = \delta_{\sigma \tau} [\Phi(\Phi^T \Phi)^{-1} \Phi_T]^T, \tag{11} \]

where \( \delta_{\sigma \tau} \) is the Kronecker delta function. Hereafter, indices \( i,j \) run over the number of electrons \( N \), and \( p,q,r,s \) index the entire set of basis functions \( M \). \( | \phi_T \rangle \), the trial wavefunction, constrains the paths of the random walk according to the phaseless constraint. \( \Phi \) is used to denote the matrix representation of a Slater determinant, with columns representing orbitals, e.g. linear combinations of orthonormal basis functions. These matrices have dimensions \( M \times N \) (\( M = \) basis size, \( N = \) number of electrons). In the case of a CASSCF trial, \( | \phi_T \rangle = \sum_d^{N_{det}} c_d \phi_T^d \), where every \( | \phi_T^d \rangle \) corresponds to a determinant with a maximum of \( \epsilon \) excitations from the reference determinant.

The expression for the two-body contribution to the local energy is given by

\[ E_2 = \sum_d^{N_{det}} \frac{c_d \langle \phi_T^d | \Phi \rangle}{\sum_d^{N_{det}} c_d \langle \phi_T^d | \Phi \rangle} \times \frac{1}{2} \sum_{pqrs}^M V_{pqrs} \sum_{\sigma \tau} (G_{p\sigma,q\tau}^d G_{q\tau,s\sigma}^d - G_{p\sigma,s\tau}^d G_{q\tau,r\sigma}^d), \tag{12} \]
where

\[ G_{pq,qr}^d = \frac{\langle \phi_T^d | a_p^+ a_{qr} | \phi \rangle}{\langle \phi_T^d | \phi \rangle} = \delta_{\sigma \tau} [\Phi (\Phi_T^d \Phi)^{-1} \Phi_T^d]_{qp}. \]  

(13)

In practice, the four index ERI tensor is represented as a three-index factorized tensor

\[ V_{pqrs} = \sum_{\alpha} X_{\alpha} L_{pr}^\alpha L_{qs}^\alpha, \]

where \( X \) ranges from \( 4M \) to \( 10M \) for Cholesky decompositions with thresholds between \( 10^{-4} \) and \( 10^{-6} \). A naive implementation of the local energy evaluation would thus scale as \( O(N_{\text{det}} M^4) \). In typical ph-AFQMC calculations the local energy must be evaluated approximately 4 to 6 million times; the energy evaluation can thus quickly become a computational bottleneck. However, there are many algorithmic tricks one can use that can significantly reduce the scaling of this step, allowing for the extension of AFQMC/CAS to treat large scale \( ab \text{ initio} \) systems in an accurate manner. We describe one such possibility for scaling reduction below.

### 2.2 Half-Rotated ERIs

We begin by recognizing that the trial wavefunction is known prior to propagation. We define an intermediate matrix, \( Q \), which partially excludes the dependence of the Green’s function on the trial

\[ G_{pq,qr}^d = \sum_i^N Q_{qi}^d \Phi_T^{d|p|}, \]

(14)

\[ Q_{qi}^d = [\Phi (\Phi_T^d \Phi)^{-1}]_{qi}. \]  

(15)

As both the trial wavefunction and ERIs are known, we can precompute the “half-rotated” cholesky vectors, \( L_{\alpha}^{vi} = \sum_p^M L_{\alpha}^{vi} \Phi_{T,pi} \), at a cost of \( XM^2 N_{\text{cas}} \) at the beginning of the calculation, where \( N_{\text{cas}} \) corresponds to the inactive occupied plus active orbitals (aka each orbital that can be occupied in a trial determinant). \( N_{\text{cas}} \) can thus be as small as \( N \) (for a single determinant) or as large as \( M \) for a full CI expansion; for the case of interest here, relatively limited CASSCF active space CI expansions, \( N_{\text{cas}} \simeq N \). Use of this precomputed tensor and
the Q intermediates results in an algorithm which scales as $N_{\text{det}}XMN^2$

$$E_2 = \sum_{d} \frac{N_{\text{det}}}{\sum_{d} N_{\text{det}}} c_d(\Phi_d|\Phi) \times \frac{1}{2} \sum_{ij} \sum_{rs} \sum_{\alpha} \sum_{\sigma\tau} \bar{L}_{ri}^{\alpha} \bar{L}_{sj}^{\alpha} \sum_{\sigma\tau} (Q_{r\sigma,i\sigma}^d Q_{s\tau,j\tau}^d - Q_{s\sigma,i\tau}^d Q_{r\tau,j\sigma}^d).$$

(16)

The cost of this energy evaluation can further be reduced via precomputation of the sum over $X$, resulting in $O(N_{\text{det}}M^2N^2)$ scaling at the cost of storing a $N_{\text{cas}}^2M^2$ dimensional tensor in memory, $Y_{ijrs} = \sum_{\alpha} \bar{L}_{ri}^{\alpha} \bar{L}_{sj}^{\alpha}$, giving

$$E_2 = \sum_{d} \frac{N_{\text{det}}}{\sum_{d} N_{\text{det}}} c_d(\Phi_d|\Phi) \times \frac{1}{2} \sum_{ij} \sum_{rs} Y_{ijrs} \sum_{\sigma\tau} (Q_{r\sigma,i\sigma}^d Q_{s\tau,j\tau}^d - Q_{s\sigma,i\tau}^d Q_{r\tau,j\sigma}^d).$$

(17)

Note that if $i$ or $j$ are in the active space of a CAS trial, they might not be present in every determinant. In order to effectively perform the sum over determinants in parallel, we find it advantageous to store the set of $Q^d$ matrices as if they include every orbital, substituting with zeros when this is not the case. While the HR-ERI algorithm is generally faster for small CI expansions and smaller systems in general, this approach can quickly present a bottleneck for large systems. This is especially apparent in calculations utilizing GPUs, due to the inability to store the half-rotated $Y_{ijrs}$ on a single GPU. This renders large (and expensive) interconnected GPU clusters a requirement, and results in the efficiency being primarily controlled by memory passing protocols, which can easily lead to major slowdowns for medium to large systems around $M$ greater than around 1000. When running on clusters with fast GPU to GPU transfer speeds, however, the effect can be minimal, and it is often then advantageous to precompute $Y_{ijrs}$ even in the case where it cannot fit on a single GPU.

### 2.3 LO-AFQMC Algorithms

We here present the general structure of localized ERIs, before outlining the application and use of localization to the HR-ERI algorithm.
2.3.1 Compression of the HR-ERI Tensor

We begin by noting that the structure of the half-rotated tensor, $Y_{ijrs}$, corresponds to $N_{cas}^2$ blocks of dimension $M \times M$ corresponding to the interaction integrals between the occupied orbital pair $[ij]$, which we denote $\{Y_{rs}^{[ij]}\}$. It is well known that the low-rank structure of the Coulombic integrals is best revealed when dealing with interactions between distinct localized orbitals. This low rank structure can be taken advantage of in numerous ways; we do so via a block-wise singular value decomposition for each localized orbital pair $[ij]$, which is then truncated according to a threshold beyond which singular values are discarded, here denoted $T_{SVD}$. The result of this procedure is an $N_{cas} \times N_{cas}$ list of rectangular matrices with dimensions $M \times M_{SVD}$ and $M_{SVD} \times M$ with $M_{SVD}$ being the truncated dimension. Thus, the expression for $Y_{rs}^{[ij]}$ may be written

$$Y_{rs}^{[ij]} \simeq \sum_{K}^{M_{SVD}} U_{rK}^{[ij]} \Sigma_{K}^{[ij]} V_{Ks}^{[ij]} = \sum_{K}^{M_{SVD}} \bar{U}_{rK}^{[ij]} \bar{V}_{Ks}^{[ij]},$$

(18)

where $U^{[ij]}$ and $V^{[ij]}$ are the set of unitary left and right singular vectors, and $\Sigma^{[ij]}$ are their associated singular values. For notational clarity, we omit the indices $[ij]$ from $M_{SVD}$ and $K$ throughout the remainder of the paper. At $M_{SVD} = M$ (i.e. $T_{SVD} = 0$), this expression becomes exact. This compression is performed once in the beginning of the simulation, at a cost of $O(N_{cas}^2 M^3)$, which is easily parallelized over GPUs for a speedup of $\frac{1}{N_{GPU}}$, and represents a minimal ($\simeq 1$ to 2\%) addition to the time of the calculations outlined in this study. We choose to allow $M_{SVD}$ to vary between $[ij]$ pairs, instead fixing $T_{SVD}$ to obtain a given accuracy. It is additionally possible to employ approximate versions of the singular value decomposition, which would reduce the scaling of this step significantly, although we did not explore this in this work. In practice, $\langle M_{SVD} \rangle$, the average compressed dimension, does not scale with the system size (and in fact decreases with a given $T_{SVD}$ as a function of the system size for a given class of molecules, as we demonstrate in section 3), and thus the memory scaling is effectively reduced from quartic to cubic in system size, namely $N_{cas}^2 M$. 

13
2.3.2 LO-AFQMC in the HR-ERI Algorithm

We next explicitly express the energy for a CASSCF trial using the HR-ERI algorithm within the compressed framework using

$$E_2 = \sum_{d} c_d \langle \Phi_d | \Phi \rangle \times \frac{1}{2} \sum_{ij} M_{\text{SVD}} \sum_{rs} \bar{U}_{rK}^{[ij]} \bar{V}_{Ks}^{[ij]} (4Q_{d,i}^d Q_{d,j}^d - 2Q_{s,i}^d Q_{s,j}^d).$$

(19)

The complexity for a direct summation is now $O(N_{\text{det}} N^2 M \langle M_{\text{SVD}} \rangle)$, and for a single determinant trial this is straightforward and results in $O(N^2 M \langle M_{\text{SVD}} \rangle)$ scaling. Due to the introduction of a low rank index K, however, we can extend the savings for the half-rotated algorithm further with respect to the number of determinants by taking advantage of the structure of the CAS trial. In the HR-ERI algorithm we use a Sherman-Morrison-Woodbury (SMW) algorithm to update the overlap of each walker with the determinants of the trial. In other words, we can write the $N_{\text{det}}$ intermediate Green’s functions $Q_{ri}^d$ of Eq. 15 as a single contribution from the reference, along with rank $\epsilon$ corrections specific to each determinant

$$Q_{ri}^d = \Phi \ast (A + U^d V^{dT})^{-1} = \Phi A^{-1} - \Phi A^{-1} U^d (I + V^{dT} A^{-1} U^d)^{-1} V^{dT} A^{-1},$$

(20)

where $A = \Phi^0_T \Phi$, $\Phi^0_T$ is the reference determinant, and $U^d$ and $V^{dT}$ are $N$ by $\epsilon$ matrices corresponding to the rows which are changed via excitation from the reference for a given determinant. From here on $\epsilon$ will refer to the maximum number of excitations possible in the CAS trial, which is typically $\simeq 6$ and bounded by the size of the active space. The scaling of the formation of the set of all $Q$ matrices has now been transformed from $O(N_{\text{det}} N^2 M)$ to $O(N^2 M + N_{\text{det}} N \epsilon^2)$ when in the molecular orbital basis (see Sec. S4). If one formed the full set of $N_{\text{det}} N \times M$ matrices $Q_{ri}^d$, as in the full HR-ERI algorithm, the resulting operations in Eq. 19, computed in a loop over determinants, would scale as $O(N_{\text{det}} N^2 M \langle M_{\text{SVD}} \rangle)$. However, If we rearrange the sums in Eq. 19 so as to avoid the explicit computation of $Q_{ri}^d$, we can effectively replace another factor of the full basis dimension, $M$, with $\langle M_{\text{SVD}} \rangle$.  

14
Since the LO-ERI are specific to the pairs \([ij]\), we must operate on a column of the \(Q\) matrix, \(Q^d_{r,i}\), corresponding to the specific orbital \(i\),

\[
Q^d_{r,i} = \Phi A^{-1} - \Phi A^{-1}U^d(I + V^dT A^{-1}U^d)^{-1}V^dT A^{-1}.
\]  \(\text{(21)}\)

Note that the \(U^d\) matrix has a block structure of zero rows followed by an identity (or permutation matrix) of maximum dimensions of \(\epsilon \times \epsilon\), and so \(A^{-1}U^d\) is equal to the last \(\epsilon\) columns (or permutations thereof) of \(A^{-1}\) and requires no computation. Additionally, note that each \(Q\) has \(N\) total columns, whereas the index \(i\) in the energy evaluation (see Eq. 22), is summed over \(N_{\text{cas}}\); thus the index “\(i\)” in Eq. 22 corresponds to the index of the column of the \(d\)-th determinant in the original CAS MOs. This means that for a specific determinant, the \(i\)-th column might not exist - thus, to allow for efficient operations in parallel on GPUs (“batching”), which is most effective for processes with equivalent operation count, we assume that they do in memory, substituting with zeros if needed.

We can now combine Eqns. 17 and 21. This involves effectively performing the sum over full basis dimensions \(r\) and \(s\) first. We then form four vectors of dimension \(\langle M_{SVD}\rangle\),

\[
\bar{Q}^d_{K,L,i} = U^{[ij],T}_{rK} \cdot Q^d_{r,i}, \quad \bar{Q}^d_{K,R,i} = V^{[ij]}_{Ks} \cdot Q^d_{s,i}, \quad \bar{Q}^d_{K,L,j} = U^{[ij],T}_{rK} \cdot Q^d_{r,j}, \quad \bar{Q}^d_{K,R,j} = V^{[ij]}_{Ks} \cdot Q^d_{s,j},
\]

for every \([ij]\) pair and every determinant. If we multiply by the determinant coefficients and overlaps in this step, the resulting sums over \([ij]\), determinants, and \(K\), are simply a dot product (e.g. for the Coulomb interaction \(\bar{Q}^d_{K,L,i} * \bar{Q}^d_{K,R,j}\), scaling as \(O(N_{\text{det}}N^2\langle M_{SVD}\rangle)\)).

Explicitly, we write

\[
E_2 = \sum_{d} \frac{N_{\text{det}}}{\sum_{d}N_{\text{det}}} c_d \langle \Phi_d | \Phi \rangle \times \frac{1}{2} \sum_{ij} \sum_{K} (4\bar{Q}^d_{K,L,i} \bar{Q}^d_{K,R,j} - 2\bar{Q}^d_{K,L,i} \bar{Q}^d_{K,R,i}),
\]  \(\text{(22)}\)

\[
\bar{Q}^d_{K,L,i} = U^{[ij],T}_{rK} \cdot Q^d_{r,i} = U^{[ij]}_{K,i} \Phi A^{-1} - U^{[ij]}_{Kr} \Phi A^{-1}U^d(I + V^dT A^{-1}U^d)^{-1}V^dT A^{-1}.
\]  \(\text{(23)}\)

Consider now the formation of \(\bar{Q}^d_{K,L,i}\), for which the scaling of key intermediates has been outlined in Table 1. Note that for every \([ij]\) pair, we only require \(U^{[ij]}_{Kr}\) intersect with vectors
$A_i^{-1}$ and $A_j^{-1}$, instead of the entire $M \times M$ matrix $A^{-1}$. For the reference section, i.e. the first term of the right hand side of Eq. 23, the computational order is straightforward. Specifically, we store the full $\Phi A^{-1}$, then define vectors corresponding to $i$ and $j$, and lastly multiply $U_{Kr}^{ij}$ by these vectors. This set of steps scales as $N^2M\langle M_{SV D}\rangle$, and must occur four times, once for every $\bar{Q}$ in Eq. 22.

Table 1: List of key intermediates in the LO-AFQMC evaluation of the 2-body energy using SMW, including the memory and computational scaling. The nomenclature $\{X\}_{a,b,c}$ indicates that the scaling listed is for the formation of the set of intermediates $X$ for all valid indices $a$, $b$, $c$.

| Intermediate | Operation | Memory scaling | Computational scaling |
|--------------|-----------|----------------|----------------------|
| $\{\bar{Q}_{L,i}\}_{ij,d,K}$ | $\bar{Q}_{L}^{ij} \times S_i$ | $N_{det}N^2\langle M_{SV D}\rangle \epsilon$ | $N_{det}N^2\langle M_{SV D}\rangle \epsilon$ |
| $\{\bar{Q}_{L,i}\}_{ij,d,K}$ | $\bar{U}_{Kr}^{ij} \times (\Phi A^{-1}U_d)$ | $N^2\langle M_{SV D}\rangle \epsilon$ | $N^2\langle M_{SV D}\rangle \epsilon$ |
| $\{S\}_d$ | $(I + V^{T,d}A^{-1}U_d)^{-1} \times V^{T,d}A^{-1}$ | $N_{det}N \epsilon$ | $N_{det}N \epsilon^2$ |

For the SMW correction term (the rightmost term in Eq. 23), we must take advantage of the block identity form of $U_d$, namely that the first $N_{inact}$ (occupied orbitals which are not in the active space) columns of $\Phi A^{-1}$ are zero regardless of the determinant. Thus we need only compute $\bar{Q}_{L}^{\text{int}} = \bar{U}_{Kr}^{ij} \Phi A^{-1}_{act}$ once, which scales as $N^2M\langle M_{SV D}\rangle \epsilon$, where $A^{-1}_{act}$ is the last $\epsilon$ columns of $A^{-1}$. This must be performed twice, once for $\bar{U}_{Kr}^{ij}$ and once for $\bar{V}_{Ks}^{ij}$.

If one generates the $N_{det}$ intermediates of size $\epsilon \times N$, corresponding to $\{S\}_d = (I + V^{d,T}A^{-1}U_d)^{-1}V^{d,T}A^{-1}$ (see details in Sec. S4), it is straightforward to determine the column of these intermediates corresponding to $i$ for each determinant and then multiply the stored $\bar{Q}_{L}^{\text{int}}$ by these column vectors and subtract them from the reference value. This series of steps then scales as $N_{det}N^2\langle M_{SV D}\rangle \epsilon$. The total scaling of the HR-ERI energy evaluation is thus reduced from $(N_{det}N^2M\langle M_{SV D}\rangle + N_{det}N^2\langle M_{SV D}\rangle)$ to $(N^2M\langle M_{SV D}\rangle(\epsilon+1) + N_{det}N^2\langle M_{SV D}\rangle \epsilon)$. Keeping only dimensions that scale with the system size, this is thus an asymptotically $O(N^2M + N_{det}N^2)$ scaling algorithm. While the use of SMW to speed up calculations has been well documented, we additionally provide details with respect to the generation of the Green’s function in the MO basis, for 1-body energetic terms and the evaluation of the force bias in Sec. S4.
3 Results

Here we present results illustrating the efficiency and accuracy of the LO-AFQMC approach. The specifics of the localization method used were not observed to significantly affect results, thus only a Foster-Boys localization scheme was used to produce the results shown here.\textsuperscript{34} Only the inactive orbitals were localized; localization of the active orbitals after CASSCF iterations reduced the efficiency of the CI expansion, resulting in orders of magnitude more determinants being required in the CI expansion to produce the same total sum of CI weight, and thus a loss of efficiency.

We first investigate dodecane in a cc-pVDZ basis as a medium sized test case. In order to estimate the error due to compression of the HR-ERI tensor, we take the difference in energies between that arising from the full and localized HR-ERI tensors to evaluate the local energy of a single walker (Figure 1). The error estimated in this manner converges with respect to the threshold $T_{SVD}$, with $T_{SVD} = 0$ reproducing the full AFQMC result. For ph-AFQMC, our goal is simply to reduce systematic errors to below the statistical error that one is seeking, which is generally around 1 mHa for molecular systems. This is achieved for a single energy evaluation for dodecane around $T_{SVD} = 5 \times 10^{-4}$, with an overall compression efficiency $(\frac{M-\langle M_{SVD}\rangle}{M} * 100)$ of 96 percent. We note that the reduction in error with $T_{SVD}$ does not occur monotonically. Error reduction does decrease systematically, however, and the magnitude of fluctuations at any given threshold are proportional to the threshold itself. To assess the increased scaling due to this compression, we performed calculations for 16 walkers on a single GPU, and present the speedup versus storing and using the full HR-ERI tensor. Even the tightest threshold tested ($T_{SVD} = 10^{-5}$) is over 15x faster than our HR-ERI code for a single energy evaluation.
Figure 1: Convergence of the LO-AFQMC error with regards to $T_{SVD}$ for dodecane in the cc-pVDZ basis. The error vs regular AFQMC (top, blue) converges to within 1 mHa (red shaded region) around $T_{SVD} = 0.0005$. Note that values to the left have a tighter threshold, and that the x axis is in a logarithmic scale. The compression efficiency (middle, green) ranges between 82 and 97 %, with an intermediate $T_{SVD} = 10^{-4}$ threshold reducing $M$ by 91.75%. The reduced computational scaling due to the compression is reflected in the speedup vs regular HR-ERI (bottom, red) of the energy evaluation for 16 walkers, which even for the tightest threshold tested ($T_{SVD} = 10^{-5}$) is over 15x faster than our HR-ERI code, with $T_{SVD} = 10^{-4}$ being $\simeq 31$ times faster. These calculations were performed on a single NVIDIA GeForce RTX 3090.

To illustrate the necessity of using localized orbitals, we compare LO-AFQMC ($T_{SVD} = 10^{-4}$) using both localized and non-localized orbitals for dodecane (Fig. 2). It can be seen that the degree of compression is significantly reduced when using the canonical basis, highlighting the intuitive fact that localized and distant orbital pairs exhibit low rank blocks within the HR-ERIs, similar to observations in other local correlation methods.\textsuperscript{35}
Figure 2: A histogram of $M_{SV_D}$ for LO-AFQMC in the canonical (orange) and localized (blue) molecular orbital bases for dodecane with $T_{SV_D} = 10^{-4}$. Localization results in an increase in compression efficiency from 48.89% to 91.75%. Note that the full basis size $M = 298$.

3.1 Metallocenes

While our results for dodecane are promising, it is necessary to test whether or not this localization scheme works for more challenging molecules where AFQMC may be a method of choice. Recently, we presented ph-AFQMC data on the gas phase ionization energy of a series of metallocenes (Fig. 3), for which we found it necessary to use multideterminant CASSCF trials. Here, we use calculations on ferrocene as benchmark results for LO-AFQMC, using geometries from Ref. 10. Population control calculations for Fe(III)(Cp)$_2$ were run with
a series of $T_{SVD}$ values and compare our results to the full AFQMC total electronic energies in a cc-pVTZ-dkh basis (Table 2). We again find $T_{SVD} = 10^{-4}$ to be a good choice to reduce the error in the total energy to the level of statistical error, while still maintaining a compression efficiency of 77.5%. We hereafter then use $T_{SVD} = 10^{-4}$ as the default threshold for the remaining results, although one might need to increase the threshold if one desires an increased accuracy.

Table 2: Errors in total energy for ferrocene (cc-pVTZ-dkh) vs an identical AFQMC calculation with full ERIs for a series of SVD truncation thresholds $T_{SVD}$. $T_{SVD} = 10^{-4}$ provides a good balance between memory efficiency and accuracy.

| $T_{SVD}$   | $\Delta E$ (Ha) | % Compression |
|------------|-----------------|---------------|
| 0.00005    | 0.0003(13)      | 70.84         |
| 0.0001     | 0.0010(10)      | 77.50         |
| 0.00025    | 0.0013(15)      | 85.04         |
| 0.0005     | 0.0034(11)      | 89.52         |
| 0.00075    | 0.0049(11)      | 91.64         |
| 0.001      | 0.0086(11)      | 92.91         |

3.2 Polyacenes

![Chemical structure of polyacenes 1-5](image)

While the performance for metallocenes is reasonable, the localized orbital approximation is likely to have a greater effect on the accuracy of ph-AFQMC (as well as a better compression efficiency) as more electron pair interactions are approximated. Thus it is important to assess the scaling of the error of the approximation with respect to the system size. To this end, the singlet-triplet (ST) gaps of the linear polyacenes provide a set of aromatic,
quasi one-dimensional systems for which ph-AFQMC has already been shown to be robust.\textsuperscript{4} As these systems exhibit significant delocalization throughout their $\pi$ systems, we expect the calculation of this quantity to provide a stringent test of the accuracy and efficiency of LO-AFQMC. Minimal CASSCF trials (6e 6o) were obtained for each system in the cc-pVTZ basis, for both singlet and triplet spin states using the 3 highest and 3 lowest energy occupied and unoccupied RHF orbitals, respectively, following Ref. 7 for benzene. Geometries were taken from Ref. 4.

In order to evaluate the error due to the LO approximation, both regular AFQMC and LO-AFQMC with $T_{SVD} = 10^{-4}$ were run for the entire polyacene set. 1656 walkers were run for 300 Ha$^{-1}$, maintaining 99.5\% of the CI weight for each trial wavefunction, and sampling the same auxiliary fields to attempt to remove discrepancies due to rare sampling events. Table 3 shows the total energies for each system, as well as the LO error, defined as the difference between the total energy derived from AFQMC and that from LO-AFQMC. In all cases, the LO error remains below 1 kcal/mol, and additionally below that of the statistical error of AFQMC, rendering LO-AFQMC statistically equivalent to AFQMC for these cases. As expected, as the system size increases, the efficacy of localization increases, with pentacene exhibiting a 93.4\% compression. Note that the error does not necessarily cancel between similar species, and additionally does not scale with system size.

Table 3: Total energies (Ha) for the singlet ($S_0$) and triplet ($T_1$) states of the series of polyacenes, as calculated by AFQMC and an identical LO-AFQMC calculation with $T_{SVD} = 10^{-4}$.

| System  | State | % Comp. | AFQMC (Ha) | LO-AFQMC (Ha) | LO Error (kcal/mol) |
|---------|-------|---------|------------|---------------|---------------------|
| benzene | $S_0$ | 67      | -231.9932(5) | -231.9928(4) | 0.2±0.4            |
|         | $T_1$ | 69      | -231.8436(6) | -231.8439(5) | -0.2±0.5           |
| naphthalene | $S_0$ | 81      | -385.4733(6) | -385.4739(6) | -0.4±0.5           |
|         | $T_1$ | 81      | -385.3682(7) | -385.3674(6) | 0.5±0.6            |
| anthracene | $S_0$ | 86      | -538.9477(8) | -538.9488(8) | -0.7±0.7           |
|         | $T_1$ | 87      | -538.8735(10)| -538.8741(7) | -0.4±0.8           |
| tetracene | $S_0$ | 90      | -692.4214(8) | -692.4207(9) | 0.4±0.8            |
|         | $T_1$ | 91      | -692.3684(13)| -692.3684(10)| 0.1±1              |
| pentacene | $S_0$ | 94      | -845.8904(14)| -845.8918(11)| -0.9±1.1           |
|         | $T_1$ | 93      | -845.8555(8) | -845.8552(13)| 0.2±0.9            |
The AFQMC and LO-AFQMC singlet triplet gaps for the entire series are presented in Table 4 and compared to previous ph-AFQMC calculations as well as experiment. In most cases the use of an CASSCF trial, even the minimal 6e 6o ones used here, results in a smaller error compared to experiment, with an MAE of $1.3 \pm 0.8$ kcal/mol vs the previously reported ph-AFQMC (using an unrestricted Kohn-Sham (UKS) trial) MAE of 2.8 kcal/mol. LO-AFQMC retains this accuracy with an MAD of $1.8 \pm 0.8$ kcal/mol versus experiment, and the LO error ranging from 0.4 to 1.1 kcal/mol, and always within the inherent statistical error of AFQMC.

Table 4: ST gaps in kcal/mol for the polyacenes (cc-pVTZ). The experimental values and ph-AFQMC with a UKS trial were obtained from Ref 4; the experimental values are corrected to compare to electronic energies from AFQMC using vibrational free energy corrections from B3LYP as described in Ref 4.

| System     | CAS/AFQMC | CAS/LO-AFQMC | LO Error | UKS/AFQMC  | Expt.  |
|------------|-----------|--------------|----------|-------------|--------|
| benzene    | 93.9±0.5  | 93.4±0.4     | -0.4±0.6 | N/A         |        |
| naphthalene| 65.9±0.6  | 66.8±0.5     | 0.9±0.8  | 68(1.2)     | 64.4   |
| anthracene | 46.6±0.8  | 46.9±0.7     | 0.4±1.1  | 46.2(1.2)   | 45.4   |
| tetracene  | 33.2±1    | 32.9±0.8     | -0.4±1.3 | 34(1.6)     | 31.2   |
| pentacene  | 21.9±1    | 23±1.1       | 1.1±1.4  | 25.2(1.6)   | 21.3   |
| MAD vs Expt.| 1.3±0.8  | 1.8±0.8      |          | 2.8±1.4     |        |

As we sample walkers along the imaginary time trajectory, there is a small variation in the LO error. To quantify this we take a sample of around 100,000 walkers for the case of benzene, and evaluate the local energy error due to compression (Fig. 5). The error follows a normal distribution at each timestep, which then varies slightly over imaginary time and results in an additional small ($\simeq 0.1$ kcal/mol) increase in the statistical error.
To investigate whether the insensitivity of the LO error to the size of the molecule was consistent across imaginary time (i.e. independent of the particular set of walkers we ran an energy evaluation using the non-compressed HR-ERI tensor at a single time step (somewhat arbitrarily chosen to be at 280 Ha$^{-1}$), using an identical set of walkers as the LO-AFQMC calculation. The resulting compression errors for the singlet total energies (Fig. 6) do in fact scale near linearly with system size (M), as was originally expected for systems with similar electronic structure. For all systems, the localization error is less than 1 kcal/mol, although for the case of pentacene the error due to compression as estimated at a single time step is actually larger than the statistical error after reblocking over imaginary time. This suggests that the error due to localization, while averaging out to a consistently small value, can vary significantly depending on the stochastic paths taken by the set of walkers. Due to this, we believe it may be prudent for applications which depend on the accuracy of the total energy to tighten $T_{SVD}$ as the system size grows, although the scaling of $T_{SVD}$ was not necessary for any of the applications presented in this work. As the efficacy of compression demonstrably increases as the system size increases, we expect the overall savings for percent reduction in memory to remain sizeable, should one tighten $T_{SVD}$ for larger systems.
3.3 Platonic Hydrocarbons

While the polyacenes are a good test to observe the scaling of localization error in a quasi one-dimensional setting, it is useful to test three-dimensional systems. Here, we choose a set of hydrocarbon cage molecules in the shape of the platonic solids, for which benchmark CCSD(T) heats of formation and geometries are available. Total heats of formation for LO-AFQMC using both cc-pVDZ and cc-pVTZ basis sets are presented in Table 5, along with LO-AFQMC errors versus full, standard ph-AFQMC calculations and the compression efficiency. Although AFQMC at the cc-pVTZ level reaches the bounds of chemical accuracy versus the cited CCSD(T)/CBS values for the smaller systems, the errors versus the reference CCSD(T) energies increase with the system size. This is not, however, attributed to the LO approximation; the error versus a full ph-AFQMC calculation was in all cases less than the statistical error of ph-AFQMC (LO Error in Table 5). Two tests (Fig. 7) using LO-AFQMC and full ph-AFQMC to measure the energy of specific sets of walkers at imaginary times 2 Ha$^{-1}$ and 250 Ha$^{-1}$, representing the beginning and end of an AFQMC simulation respectively, suggest that the linear increase in LO error seen in the polyacenes is indeed reproduced in the platonic hydrocarbons, with some expected variation in error between
timesteps. It thus appears that the averaging over imaginary time can again lead to some favorable cancellation of LO error, which is further supported by a small test calculation in the case of C_8H_8 in section S5. Again, as before, the efficacy of the compression increases with system size, with C_{20}H_{20} exhibiting 84\% compression.

Table 5: Total Atomization Energies in kcal/mol for the platonic hydrocarbon cages; CCSD(T)/CBS values were taken from Ref. 36. LO Error refers to the difference between the LO-AFQMC result and an identical ph-AFQMC calculation run without compression.

|         | CCSD(T) | LO-AFQMC | LO Error | % Comp. |
|---------|---------|----------|----------|---------|
| C_4H_4  | 793.9   | 794.22±0.34 | 0±0.48   | 50.4    |
| C_6H_6  | 1254.32 | 1255.06±0.49 | -0.02±0.69 | 67.1    |
| C_8H_8  | 1706.99 | 1708.58±0.46 | -0.15±0.64 | 72.1    |
| C_{10}H_{10} | 2196.46 | 2201.26±0.5  | 0.76±0.71  | 82.4    |
| C_{12}H_{12} | 2718.29 | 2724.36±0.56 | -0.23±0.8  | 85.9    |
| C_{20}H_{20} | 4621.46 | 4633.52±0.8  | -0.42±1.41 | 84.2    |

Figure 7: Scaling of the error of LO-AFQMC vs full ph-AFQMC, estimated at a time step of 2 Ha^{-1} (a) and 250 Ha^{-1} (b), with respect to the basis size (cc-pVTZ). Error bars are not plotted since the error is evaluated at a single timestep. Note that the error for each time step scales linearly, although there is notable variation as the walker distributions change.

3.4 GPU implementation and timing

LO-AFQMC effectively mediates the memory bottlenecks apparent in the HR-ERI algorithm, in addition to reducing the computational complexity and scaling with respect to system size. However, we have not discussed to what degree the actual wall time can be
reduced on a GPU system, as for these system sizes wall time is, more often than not, limited by the ability to perform operations in parallel. Since we are dealing with dense linear algebra, it is possible to perform most operations in parallel for all determinants (aka “batch” operations over determinants). As we allow $M_{SVD}$ to vary between each pair $[ij]$, great care must be taken to enable efficient simultaneous batching of operations, especially if one intends to use standard cuBLAS libraries which are designed for calculations with consistent dimension. By storing the compressed ERIs contiguously in GPU memory, we were able to batch nearly all operations over the total number of electron pairs and determinants simultaneously, resulting in a significant gain in efficiency. Some details are given in section S2. Additionally, as each walker is undergoing an identical set of operations, we were able to simultaneously evaluate the local energy for all walkers stored on a GPU at once. Such batching is ineffective in the HR-ERI algorithm due to the large memory requirements (and thus inability to store intermediates for each walker), but with the reduced memory requirements, batching LO-AFQMC over walkers significantly improves efficiency, as was previously reported for alternate algorithms in Ref 21. To illustrate this, we ran LO-AFQMC for Fe(Cp)$_2$ in the cc-pVTZ-DK basis, with 200 determinants, 1656 walkers, and propagated to 200 Ha$^{-1}$, while modifying the number of walkers for which operations are performed in parallel on a single GPU. The resulting timings in GPU-hours can be seen in Fig. 8; at 16 walkers performed in parallel, the total GPU-hours is reduced by over 3 times versus in series, to 65.9 GPU-hours.
Figure 8: Effect of performing operations for increasing numbers of walkers in parallel on a single GPU, for Fe(Cp)$_2$ in the cc-pVTZ-DK basis (508 basis functions, 200 determinants). The GPU-hours represent the time necessary to propagate 1656 walkers over 200 Ha$^{-1}$, evaluating the energy 2000 times in total per walker, on the Summit supercomputer at Oak Ridge National Laboratory.

To showcase the speedups possible due to the LO-AFQMC algorithm, as well as the stability of our previous results with respect to use of the LO-AFQMC algorithm, we additionally re-ran the calculations in both cc-pVTZ-DK and cc-pVQZ-DK for the adiabatic ionization energy of Ni(Cp)$_2$ from Ref. 10 using $T_{SVD} = 10^{-4}$. For a significantly reduced cost, we were able to reproduce our previous result in the complete basis set (CBS) limit to within statistical error (-141.86±1.76 kcal/mol versus our previous -143.39±1.64 kcal/mol), which is additionally within error bars of the experimental result (-143.8±1.5 kcal/mol). To put this in perspective, our previous ph-AFQMC calculation of the Ni(III)(Cp)$_2$ species in the QZ basis (954 basis functions) required 1615 GPU-hours, whereas LO-AFQMC took 102 GPU hours to propagate the same number of walkers the same distance in imaginary time, reproducing the energy to within statistical error in 6.3% of the time. Of those 102 GPU hours, approximately 18.9% were spent in evaluating the energy (both 1- and 2-body terms), slightly more than the approximately 18.7% that were spent forming the propagation matrix ($B = \sum_{\alpha} x_{\alpha} L_{\alpha}$, scaling as $X M^2$). Although the formation of $B$ is $\approx 20$ times faster than the energy evaluation, it must be performed every step, whereas we only measure the energy once every 20 steps. The compression rates of the neutral/cationic Ni(Cp)$_2$ species
were 84/86% and 86/89% for the cc-pVTZ-DK and cc-pVQZ-DK bases, respectively, slightly higher than that of Fe(Cp)$_2$.

### 4 Conclusions

We have presented a localized orbital compression of the ERIs as a means to reduce the memory footprint and computational scaling of the energy evaluation in AFQMC calculations. Our algorithm has been demonstrated to be efficient and robust for a variety of difficult test systems in manner which is independent of cancellation of error between two electronic states. Although here we restricted our focus to the HR-ERI algorithm of phaseless AFQMC, the approach and approximation errors reported are general.

As the value of interest is not the controlled error of SVD (namely the $L^2$ norm of the matrix error), but rather the total energy estimated using the compressed integrals, we have taken care to estimate the error due to our localization approximation in multiple ways. We compare to full ph-AFQMC for a single energy evaluation, a weighted average over walkers at a single time step, and full population control AFQMC calculations averaged over imaginary time. In all cases we observe controllable if not entirely monotonic errors for each metric. In addition, we compare our results to experiment and CCSD(T) calculations, highlighting that the accuracy of ph-AFQMC with respect to these external benchmarks remains unaffected due to the LO approximation.

While the scaling of the algorithms outlined is dependent on the number of electron pairs $N^2$ (i.e. $O(N^2(M + N_{\text{det}}))$), the formal asymptotic rank of the ERIs is expected to be quadratic $O(M^2)$, and so we foresee the possibility to improve upon this localization procedure significantly for larger systems. In particular, one might envision screening pairs of electrons, and only including those that contribute meaningfully to the total energy. This would asymptotically reduce the number of pairs to $N$ from $N^2$ and thus reduce the memory scaling of this tensor to quadratic, namely $O(NM)$, as well as producing a corresponding
reduction in computational effort.

While LO-AFQMC significantly reduces the computational effort required for ph-AFQMC, it is still prohibitively expensive compared to, say, density functional theory (DFT). We thus do not expect LO-AFQMC to directly compete with fast quantum chemistry methods. However, the performance of DFT for transition metal containing molecules has been sub-standard, in no small part due to the lack of trainable reference data for relevant cases. AFQMC can obtain reference values on which to train DFT functionals, and when extended in viable system size via methods like LO-AFQMC, the task of producing a large benchmark dataset with a variety of represented chemistries begins to become feasible. This philosophy is akin to some recent work in the coupled cluster community, where pair-natural orbital (PNO) based localization approaches can reduce the scaling of CCSD(T) to near-linear, with respectable accuracy,\textsuperscript{37–41} and some benchmark datasets on transition metals have been recently produced using DLPNO-CCSD(T).\textsuperscript{42,43} However, for statically correlated systems, or even otherwise, the accuracy of full CCSD(T) can come into question,\textsuperscript{44} and alternate benchmark-level methods such as AFQMC are necessary to provide complementary or even more accurate results.\textsuperscript{8–10} It is worth noting that a PNO-based ph-AFQMC approach using approximate PNOs from MP2 was initially attempted, but was found to be inaccurate upon projection in imaginary time, presumably due to the approximate nature of the PNO transformation. One advantage of directly compressing the HR-ERI tensor as described here is that it is not dependent on any lower level methods, and so can effectively reproduce full AFQMC values. Even so, the idea to directly solve for a low rank structure for the ERIs is not new to AFQMC, and is shared by tensor hypercontraction and the low-rank compression of the cholesky vectors in Ref. 19. The LO-AFQMC algorithm presented here stands out in utility by virtue of it’s relative simplicity, high performance, and low prefactor, which depending on the degree of electronic localization within a given molecule can lead to better performance for systems of nearly all sizes.

Based on the timing results above for the metallocenes test cases, we can estimate the cost
of generating AFQMC benchmark data. As an example of a potentially useful benchmark
data set, consider the set of experimental redox potentials for octahedral transition metal
complexes in solution assembled in Ref. 45. The number of heavy atoms in these systems
ranges from 7 (e.g. $\text{M(H}_2\text{O)}_6$) to approximately 45; however, most of the molecules have
25 or fewer heavy atoms, and the data set could be restricted to those cases. The cost
of running 50 such calculations at the TZ/QZ level would be on the order of 70,000 GPU
hours on the Summit supercomputer. A gas phase comparison of DFT and AFQMC results
would enable the development of an improved DFT functional for transition metal containing
systems; using such a functional, the continuum model needed to compute redox potentials
in solution could then be fit to the experimental data assembled in Ref. 45.

The ability to reproduce full ph-AFQMC for considerably larger molecules than have
heretofore been amenable to AFQMC calculations represents a major step towards obtaining
benchmark-quality simulations of large transition metal complexes (and other challenging
systems) relevant to biology and materials science, such as the water splitting cluster of Pho-
tosystem II (PSII). Hundreds to thousands of calculations will be required to obtain relative
energies of different metal (Mn) spin and charge states, and different oxygen protonation
states, that are needed to precisely specify the intermediates in the Kok cycle, the catalytic
process in which the energy of four photons is used to convert water into oxygen and hydro-
gen gas.\footnote{46} A minimal PSII model would contain approximately 50 heavy atoms (based on the
smallest possible truncations of the relevant amino acids). The cost of a single calculation at
the TZ/QZ level would be on the order of 10,000 GPU hours. While a cost of this magnitude
would not be sufficient to evaluate every possible model for the S states of the Kok cycle,
50-100 selected (perhaps on the basis of DFT energetics) individual states could readily be
investigated, enabling benchmark relative energies to be determined and compared with vari-
ous DFT approaches and with experiment. Application to a system of this size would likely
benefit from some of the improvements proposed above, such as screening distant pairs.

The above applications presume that suitable trial functions can be generated for the
systems to be studied, and that a moderate number of determinants (in the 300-1000 range that we have used to date in our AFQMC modeling of metal-containing systems) can produce accurate results. For the redox potential benchmarks, this assumption is likely to be valid, although there could be surprises for particularly challenging individual cases. For the PSII model, which contains four Mn atoms and one Ca, the requirements for a trial function are far from clear. If an ultralarge number of determinants are needed, it may be necessary to switch to alternate algorithms for better scaling with $N_{\text{det}}$ (although, as pointed out above, the LO formulation described in this paper, or some alternate, will still be necessary to make such calculations tractable).

5 Acknowledgements

JLW thanks Benjamin Rudshteyn and Shiwei Zhang for many discussions on AFQMC, and Evan Arthur for help with GPU programming. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC05-00OR22725. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Calculations used the XSEDE resource Expanse at the SDSC through allocation ID COL151. JLW was funded in part by the Columbia Center for Computational Electrochemistry (CCCE). JS acknowledges funding from the National Institute of General Medical Sciences of the National Institutes of Health under award number F32GM142231.

6 Associated Content

The supporting information includes:

- Description of the algorithm of Ref. 2 and application of the LO approximation
• Additional Details of the GPU implementation.

• Discussion of the computational savings due to propagating in the Molecular Orbital Basis.

• Outline of the SMW algorithm to form the Green’s function in the MO basis

• Data tracking the LO error cancellation for C₈H₈ over imaginary time

This information is available free of charge via the Internet at http://pubs.acs.org
References

(1) Williams, K. T.; Yao, Y.; Li, J.; Chen, L.; Shi, H.; Motta, M.; Niu, C.; Ray, U.; Guo, S.; Anderson, R. J.; et. al., Direct comparison of many-body methods for realistic electronic Hamiltonians. Physical Review X 2020, 10, 011041.

(2) Mahajan, A.; Sharma, S. Taming the Sign Problem in Auxiliary-Field Quantum Monte Carlo Using Accurate Wave Functions. Journal of Chemical Theory and Computation 2021, 17, 4786–4798.

(3) Zhang, S.; Krakauer, H.; Al-Saidi, W. A.; Suewattana, M. Quantum simulations of realistic systems by auxiliary fields. Computer physics communications 2005, 169, 394–399.

(4) Shee, J.; Arthur, E. J.; Zhang, S.; Reichman, D. R.; Friesner, R. A. Singlet-Triplet Energy Gaps of Organic Biradicals and Polyacenes with Auxiliary-Field Quantum Monte Carlo. J. Chem. Theory Comput. 2019, 15, 4924–4932, PMID: 31381324.

(5) Weber, J. L.; Churchill, E. M.; Jockusch, S.; Arthur, E. J.; Pun, A. B.; Zhang, S.; Friesner, R. A.; Campos, L. M.; Reichman, D. R.; Shee, J. In silico prediction of annihilators for triplet–triplet annihilation upconversion via auxiliary-field quantum Monte Carlo. Chemical Science 2021, 12, 1068–1079.

(6) Lee, J.; Malone, F. D.; Morales, M. A. Utilizing Essential Symmetry Breaking in Auxiliary-Field Quantum Monte Carlo: Application to the Spin Gaps of the C_{36} Fullerene and an Iron Porphyrin Model Complex. arXiv preprint arXiv:2001.05109 2020.

(7) Lee, J.; Malone, F. D.; Reichman, D. R. The performance of phaseless auxiliary-field quantum Monte Carlo on the ground state electronic energy of benzene. The Journal of Chemical Physics 2020, 153, 126101.

(8) Shee, J.; Rudshteyn, B.; Arthur, E. J.; Zhang, S.; Reichman, D. R.; Friesner, R. A. On Achieving High Accuracy in Quantum Chemical Calculations of 3d Transition Metal-Containing Systems: A Comparison of Auxiliary-Field Quantum Monte Carlo with Coupled Cluster, Density Functional Theory, and Experiment for Diatomic Molecules. Journal of chemical theory and computation 2019, 15, 2346–2358.

(9) Rudshteyn, B.; Coskun, D.; Weber, J. L.; Arthur, E. J.; Zhang, S.; Reichman, D. R.; Friesner, R. A.; Shee, J. Predicting Ligand-Dissociation Energies of 3d Coordination Complexes with Auxiliary-Field Quantum Monte Carlo. Journal of Chemical Theory and Computation 2020.

(10) Rudshteyn, B.; Weber, J.; Coskun, D.; Devlaminck, P. A.; Zhang, S.; Reichman, D. R.; Shee, J.; Friesner, R. Calculation of Metallocene Ionization Potentials via Auxiliary Field Quantum Monte Carlo: Towards Benchmark Quantum Chemistry for Transition Metals. 2021.

(11) Purwanto, W.; Zhang, S.; Krakauer, H. Auxiliary-Field Quantum Monte Carlo Calculations of the Molybdenum Dimer. J. Chem. Phys. 2016, 144, 244306.

(12) Purwanto, W.; Zhang, S.; Krakauer, H. An auxiliary-field quantum Monte Carlo study of the chromium dimer. J. Chem. Phys. 2015, 142, 064302.

(13) Purwanto, W.; Zhang, S.; Krakauer, H. Excited state calculations using phaseless auxiliary-field quantum Monte Carlo: Potential energy curves of low-lying C 2 singlet states. The Journal of chemical physics 2009, 130, 094107.

(14) Landinez Borda, E. J.; Gomez, J.; Morales, M. A. Non-orthogonal multi-Slater determinant expansions in auxiliary field quantum Monte Carlo. The Journal of chemical physics 2019, 150, 074105.
(15) Mahajan, A.; Lee, J.; Sharma, S. Selected configuration interaction wave functions in phaseless auxiliary field quantum Monte Carlo. *arXiv preprint arXiv:2201.10458* 2022.

(16) Zhang, S.; Krakauer, H. Quantum Monte Carlo Method using Phase-Free Random Walks with Slater Determinants. *Phys. Rev. Lett.* 2003, 90, 136401.

(17) Motta, M.; Zhang, S. Computation of ground-state properties in molecular systems: back-propagation with auxiliary-field quantum Monte Carlo. *J. Chem. Theory Comput.* 2017, 13, 5367–5378.

(18) Shi, H.; Zhang, S. Some recent developments in auxiliary-field quantum Monte Carlo for real materials. *The Journal of Chemical Physics* 2021, 154, 024107.

(19) Motta, M.; Shee, J.; Zhang, S.; Chan, G. K.-L. Efficient ab initio auxiliary-field quantum Monte Carlo calculations in Gaussian bases via low-rank tensor decomposition. *Journal of chemical theory and computation* 2019.

(20) Malone, F. D.; Zhang, S.; Morales, M. A. Overcoming the Memory Bottleneck in Auxiliary Field Quantum Monte Carlo Simulations with Interpolative Separable Density Fitting. *Journal of chemical theory and computation* 2018, 15, 256–264.

(21) Malone, F. D.; Zhang, S.; Morales, M. A. Accelerating Auxiliary-Field Quantum Monte Carlo Simulations of Solids with Graphical Processing Units. *Journal of Chemical Theory and Computation* 2020, 16, 4286–4297.

(22) Lee, J.; Reichman, D. R. Stochastic resolution-of-the-identity auxiliary-field quantum Monte Carlo: Scaling reduction without overhead. *The Journal of Chemical Physics* 2020, 153, 044131.

(23) Shee, J.; Arthur, E. J.; Zhang, S.; Reichman, D. R.; Friesner, R. A. Phaseless Auxiliary-Field Quantum Monte Carlo on Graphical Processing Units. *J. Chem. Theor. Comput.* 2018,

(24) Mahajan, A.; Sharma, S. Efficient local energy evaluation for multi-Slater wave functions in orbital space quantum Monte Carlo. *The Journal of Chemical Physics* 2020, 153, 194108.

(25) NVIDIA, CUDA, release: 11.4.2. 2021; [https://docs.nvidia.com/cuda](https://docs.nvidia.com/cuda).

(26) Motta, M.; Zhang, S. Ab initio Computations of Molecular Systems by the Auxiliary-Field Quantum Monte Carlo Method. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 2018, 8, e1364.

(27) Thouless, D. J. Stability Conditions and Nuclear Rotations in the Hartree-Fock Theory. *Nucl. Phys.* 1960, 21, 225–232.

(28) Purwanto, W.; Krakauer, H.; Virgus, Y.; Zhang, S. Assessing weak hydrogen binding on Ca+ centers: An accurate many-body study with large basis sets. *The Journal of chemical physics* 2011, 135, 164105.

(29) Rom, N.; Fattal, E.; Gupta, A. K.; Carter, E. A.; Neuhauser, D. Shifted-contour auxiliary-field Monte Carlo for molecular electronic structure. *The Journal of chemical physics* 1998, 109, 8241–8248.

(30) Booth, T.; Gubernatis, J. Monte Carlo determination of multiple extremal eigenpairs. *Physical Review E* 2009, 80, 046704.

(31) Buonaura, M. C.; Sorella, S. Numerical study of the two-dimensional Heisenberg model using a Green function Monte Carlo technique with a fixed number of walkers. *Physical Review B* 1998, 57, 11446.
(32) Shee, J.; Zhang, S.; Reichman, D. R.; Friesner, R. A. Chemical Transformations Approaching Chemical Accuracy via Correlated Sampling in Auxiliary-Field Quantum Monte Carlo. *J. Chem. Theory Comput.* **2017**, *13*, 2667–2680.

(33) Neese, F.; Wennmohs, F.; Hansen, A. Efficient and accurate local approximations to coupled-electron pair approaches: An attempt to revive the pair natural orbital method. *The Journal of chemical physics* **2009**, *130*, 114108.

(34) Foster, J.; Boys, S. Canonical configurational interaction procedure. *Reviews of Modern Physics* **1960**, *32*, 300.

(35) Neese, F.; Hansen, A.; Wennmohs, F.; Grimme, S. Accurate theoretical chemistry with coupled pair models. *Accounts of chemical research* **2009**, *42*, 641–648.

(36) Karton, A.; Schreiner, P. R.; Martin, J. M. Heats of formation of platonic hydrocarbon cages by means of high-level thermochemical procedures. *Journal of computational chemistry* **2016**, *37*, 49–58.

(37) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138*, 034106.

(38) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *The Journal of chemical physics* **2013**, *139*, 134101.

(39) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps?A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *The Journal of chemical physics* **2016**, *144*, 024109.

(40) Datta, D.; Kossmann, S.; Neese, F. Analytic energy derivatives for the calculation of the first-order molecular properties using the domain-based local pair-natural orbital coupled-cluster theory. *The Journal of Chemical Physics* **2016**, *145*, 114101.

(41) Saitow, M.; Becker, U.; Riplinger, C.; Valeev, E. F.; Neese, F. A new near-linear scaling, efficient and accurate, open-shell domain-based local pair natural orbital coupled cluster singles and doubles theory. *J. Chem. Phys.* **2017**, *146*, 164105.

(42) Dohm, S.; Hansen, A.; Steinmetz, M.; Grimme, S.; Checinski, M. P. Comprehensive thermochemical benchmark set of realistic closed-shell metal organic reactions. *Journal of chemical theory and computation* **2018**, *14*, 2596–2608.

(43) Maurer, L. R.; Bursch, M.; Grimme, S.; Hansen, A. Assessing Density Functional Theory for Chemically Relevant Open-Shell Transition Metal Reactions. *Journal of Chemical Theory and Computation* **2021**, *17*, 6134–6151.

(44) Hait, D.; Tubman, N. M.; Levine, D. S.; Whaley, K. B.; Head-Gordon, M. What levels of coupled cluster theory are appropriate for transition metal systems? a study using near-exact quantum chemical values for 3d transition metal binary compounds. *Journal of chemical theory and computation* **2019**, *15*, 5370–5385.

(45) Hughes, T. F.; Friesner, R. A. Development of accurate DFT methods for computing redox potentials of transition metal complexes: results for model complexes and application to cytochrome P450. *Journal of chemical theory and computation* **2012**, *8*, 442–459.

(46) Chen, H.; Dismukes, G. C.; Case, D. A. Resolving ambiguous protonation and oxidation states in the oxygen evolving complex of photosystem II. *The Journal of Physical Chemistry B* **2018**, *122*, 8654–8664.
7 TOC Figure