JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 4109−4121

pubs.acs.org/JCTC

Phaseless Auxiliary-Field Quantum Monte Carlo on Graphical Processing Units

James Shee,*[,†] Evan J. Arthur,[‡] Shiwei Zhang,[¶] David R. Reichman,[†] and Richard A. Friesner[†]

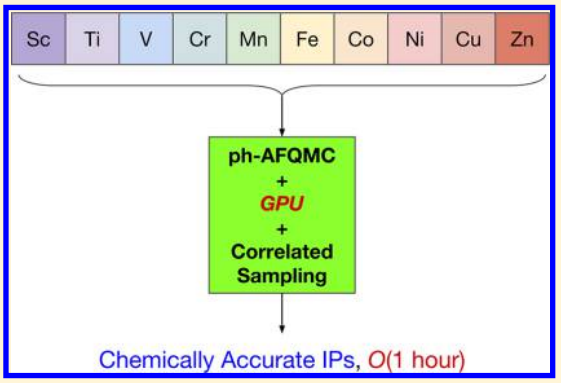
[†]Department of Chemistry, Columbia University, 3000 Broadway, New York, New York 10027, United States

[‡]Schrödinger Inc., 120 West 45th Street, New York, New York 10036, United States

[¶]Department of Physics, College of William and Mary, Williamsburg, Virginia 23187-8795, United States

Supporting Information

ABSTRACT: We present an implementation of phaseless Auxiliary-Field Quantum Monte Carlo (ph-AFQMC) utilizing graphical processing units (GPUs). The AFQMC method is recast in terms of matrix operations which are spread across thousands of processing cores and are executed in batches using custom Compute Unified Device Architecture kernels and the GPU-optimized cuBLAS matrix library. Algorithmic advances include a batched Sherman-Morrison-Woodbury algorithm to quickly update matrix determinants and inverses, density-fitting of the two-electron integrals, an energy algorithm involving a high-dimensional precomputed tensor, and the use of single-precision floating point arithmetic. These strategies accelerate ph-AFQMC calculations with both single- and multideterminant trial wave functions, though particularly dramatic wall-time reductions are achieved for the latter. For typical calculations we find speed-ups of roughly 2 orders of magnitude using just a single GPU card compared to a single modern CPU core. Furthermore, we achieve near-unity parallel efficiency using 8 GPU cards on a single node and can reach moderate system sizes via a local memory-slicing approach. We illustrate the robustness of our implementation on hydrogen chains of increasing length and through the calculation of all-electron ionization potentials of the first-row transition metal atoms. We compare long imaginary-time calculations utilizing a population control algorithm with our previously published correlated sampling approach and show that the latter improves not only the efficiency but also the accuracy of the computed ionization potentials. Taken together, the GPU implementation combined with correlated sampling provides a compelling computational method that will broaden the application of ph-AFQMC to the description of realistic correlated electronic systems.



INTRODUCTION

Auxiliary-Field Quantum Monte Carlo (AFQMC) is a computational method capable of predicting ground-state observables of chemical systems with very high accuracy. We refer the reader to refs 1−4 for core papers presenting the constrained-path and phaseless variants of AFQMC and review publications and to refs 5−24 for methodological advances and illustrative applications. For finite-sized systems such as molecules in Gaussian basis sets, AFQMC calculations scale with the fourth power of the system size, which compares favorably with traditional wave function methods such as second-order Møller−Plesset Perturbation Theory (MP2),[25] Coupled Cluster (CC) approaches,[26] and Complete Active Space methods such as CASSCF[27] and CASPT2.[28] However, the prefactor of a typical AFQMC calculation is relatively large. Recently we have introduced a correlated sampling (CS) approach for quantities involving energy differences which is capable of reducing computational prefactors by approximately an order of magnitude.[21] In this work we present a different but complementary strategy involving hardware optimization on graphical processing units (GPUs) which can drastically reduce the prefactors in calculations of general ground-state properties.

GPUs have several distinct advantages over traditional Central Processing Units (CPUs), including the ability to perform efficiently parallelized matrix operations both in serial and in "batches" and the use of single-precision (sp) floating-point arithmetic with significant gains in computational speed. We refer the reader to ref 29 for a lucid exposition of many general properties of GPU hardware. In recent years the use of GPUs has been extended well beyond traditional image visualization tasks into many fields such as machine learning[30] and molecular mechanics.[31] Of particular relevance to our work presented here is the progress in performing electronic structure calculations on GPUs. This hardware has been utilized to efficiently evaluate the integrals required in ab initio calculations,[32−35] to perform Hartree−Fock[36,37] (HF) and Density Functional Theory (DFT) calculations,[38−40] and to study model systems such as the Hubbard Model within the

Received: April 11, 2018
Published: June 13, 2018

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

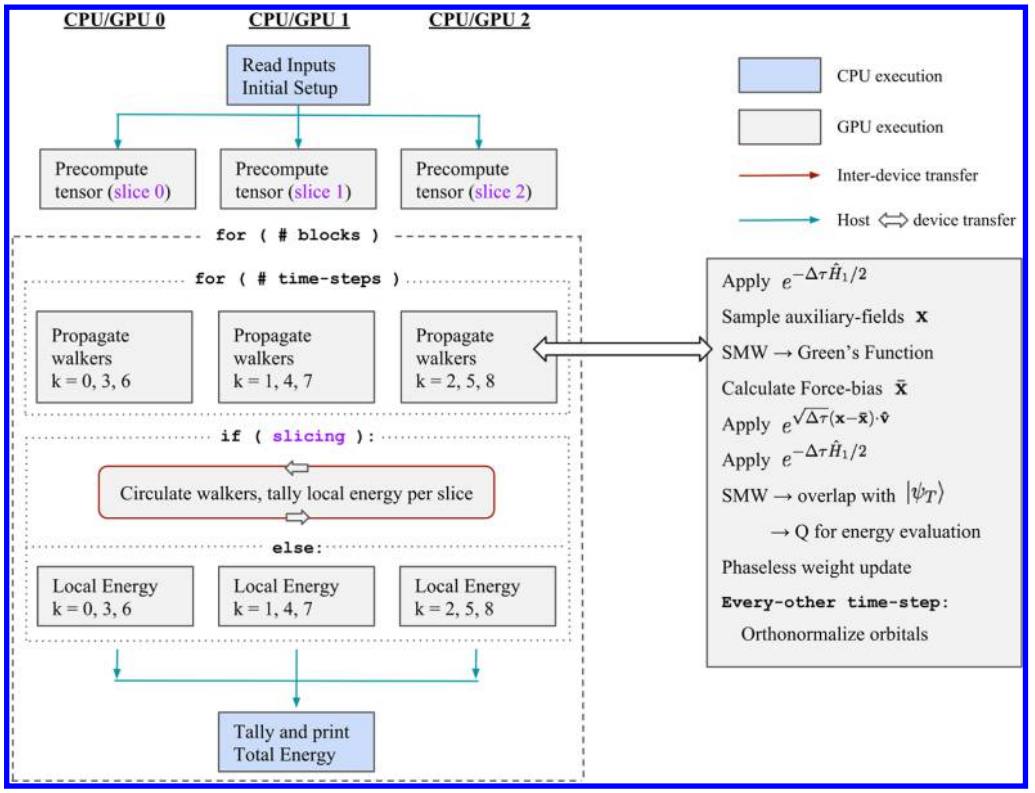


Figure 1. Flowchart of our AFQMC implementation, for 9 walkers (k = 0,...,8) and 3 CPU-GPU pairs (CPU/GPU 0, 1, 2). If the size of the precomputed tensor used in the energy evaluation exceeds the memory capacity of each device, each GPU precomputes and stores only a slice of the tensor, and the energy of a walker is computed by circulating the walker to all other GPUs and tallying the partial energies obtained from the resident slices.

dynamical cluster approximation[41] and the Ising model.[42,43] In addition there have been recent GPU implementations of MP2,[44−47] CC methods,[48] TDDFT,[49] Configuration Interaction (CI),[50,51] and CASSCF approaches.[52,53] Efficient algorithms to compute energy gradients[54,55] and tensor contractions[56] have also been developed.

With respect to Quantum Monte Carlo (QMC) methods, GPU implementations have been formulated primarily for real-space approaches. For example, Diffusion Monte Carlo (DMC) with sp arithmetic has been accelerated by a factor of ∼6× on a GPU versus a quad-core CPU.[57] A recent study employing a multi-GPU implementation has reported speed-ups of a factor of 10−15× relative to a quad-core CPU for Variational and Diffusion MC for real materials.[58] Very recently an open-source QMC suite, QMCPACK,[59] has released scalable implementations of real-space methods including Variational, Diffusion, and Reptation MC. An implementation of auxiliary-field QMC is mentioned in ref 59, although data illustrating its efficiency and accuracy is not yet available.

In this paper we detail our GPU implementation of the phaseless variant of auxiliary-field QMC (ph-AFQMC) and illustrate its performance and accuracy via calculations of the total energies of linear chains of hydrogen atoms and the all-electron ionization potentials (IPs) of the first-row transition metal (TM) atoms. We explicitly compare our GPU wall-times with CPU timings from a code of equivalent algorithmic sophistication. Speed-ups from the GPU port of 2 orders of magnitude are seen in large systems, with the potential for even greater reductions of the scaling prefactor depending on the system-size. The robustness and accuracy of our implementation are shown by comparing our calculated values to either exact numerical techniques or experiment.

Our paper is organized as follows: In Section II we provide a concise review of AFQMC and the phaseless constraint. In Section III we detail our GPU implementation and highlight significant algorithmic additions. In Sections IV and V we present timing and accuracy results for the hydrogen chains and TM IPs, respectively, and comment on the advantages of the correlated sampling approach. In Section VI we conclude with a summary of our results and a discussion of future work.

BRIEF OVERVIEW OF AFQMC

Detailed expositions of AFQMC and the phaseless constraint can be found elsewhere.[2,4,21] In this section we highlight only aspects that are directly relevant to this work.

The ground state of a many-body wave function $|\Phi\rangle$ can be obtained via imaginary-time propagation, i.e.

$$\lim_{N\to\infty}(e^{\Delta\tau(E_0-\hat{H})})^N|\Phi\rangle = |\Phi_0\rangle, \quad \text{if } \langle\Phi_0|\Phi\rangle \neq 0$$

The general electronic Hamiltonian is

$$\hat{H} = \sum_{ij}^{M} T_{ij}\sum_{\sigma} c_{i\sigma}^{\dagger}c_{j\sigma} + \frac{1}{2}\sum_{ijkl}^{M} V_{ijkl}\sum_{\sigma,\tau} c_{i\sigma}^{\dagger}c_{j\tau}^{\dagger}c_{l\tau}c_{k\sigma} \quad (1)$$

where M is the size of an orthonormal one-particle basis, σ and τ denote electron spin, and $c_{i\sigma}^{\dagger}$ and $c_{i\sigma}$ are the second-quantized Fermionic creation and annihilation operators. The matrix elements, V_{ijkl}, can be expressed as a sum over products of three-index quantities via Cholesky decomposition[13] (CD) or Density-Fitting[60] (DF) procedures, allowing eq 1 to be written

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

as the sum of all one-body operators plus a two-body operator of the form $\hat{H}_2 = -\frac{1}{2}\sum_\alpha \hat{v}_\alpha^2$.

After utilizing a Trotter-Suzuki decomposition[61] and the Hubbard-Stratonovich transformation written in the form[62,63]

$$e^{\frac{1}{2}\Delta\tau\hat{v}_\alpha^2} = \int_{-\infty}^{\infty} dx_\alpha \left(\frac{e^{-\frac{1}{2}x_\alpha^2}}{\sqrt{2\pi}}\right) e^{\sqrt{\Delta\tau}x_\alpha\hat{v}_\alpha} \tag{2}$$

the imaginary-time propagator can be expressed as an exponential of a one-body operator integrated over auxiliary-fields

$$e^{-\Delta\tau\hat{H}} = \int d\mathbf{x} P(\mathbf{x})\hat{B}(\mathbf{x}) \tag{3}$$

This integral is evaluated using a Monte Carlo technique in which the wave function is represented as a weighted sum of Slater determinants. A theorem due to Thouless[64] enables the propagation of the Slater determinants along "paths" defined by a set of auxiliary-fields, which in practice can be accomplished via matrix−matrix multiplications. Each auxiliary-field is shifted by a force-bias chosen to minimize the fluctuations in a walker's weight. With this choice the weights are updated after each propagation step by a factor proportional to $e^{-\Delta\tau E_L}$, where the local energy is defined as $E_L = \frac{\langle\psi_T|\hat{H}|\phi\rangle}{\langle\psi_T|\phi\rangle}$, and $|\psi_T\rangle$ is a trial wave function, typically taken as a single Slater determinant or a linear combination of determinants.

For the Coulomb interaction, the propagator in eq 3 has an imaginary component which rotates the walker orbitals and weights in the complex plane, leading inevitably to divergences and an exponentially decaying signal-to-noise ratio. The phaseless approximation can be used to constrain the weights to be real and positive. This involves taking the real-part of the local energy and multiplying the resulting weights by $\max\{0, \cos(\Delta\theta)\}$, where $\Delta\theta = Im\left\{\ln\frac{\langle\psi_T|\phi^{\tau+\Delta\tau}\rangle}{\langle\psi_T|\phi^{\tau}\rangle}\right\}$. The latter breaks the rotational symmetry of the random walk in the complex plane by choosing a unique gauge for $|\Phi_0\rangle$ (and eliminates walkers whose weights have undergone phase rotations in excess of $\pm\frac{\pi}{2}$). We note that the severity of the phaseless constraint can be systematically reduced with the use of trial wave functions which even more closely represent the true ground-state.

GPU IMPLEMENTATION

In contrast to traditional computing paradigms which utilize CPUs to execute all computing tasks, we employ a strategy in which CPUs offload a majority of the computational effort to one or more GPU cards. A typical GPU device has 4−12 GB of memory which is separate from that accessible by the CPU; therefore, data is usually allocated on both the host and device, and intercommunication between these different memory spaces requires the explicit copying of data back and forth. In this work great care is taken to minimize such transfers, and we create custom memory structures that organize memory addresses and facilitate switching between sp and double-precision (dp).

A flowchart outlining the ph-AFQMC algorithm with 3 CPU/GPU pairs is presented in Figure 1. First the root CPU reads in relevant quantities such as matrix elements of $\hat{H}$, overlap integrals, and the initial and trial wave functions and then completes a preliminary setup which includes transformations to an orthonormal basis, walker, and operator initializations. These quantities are sent to all devices, after which the tensor (or slices of it) used in the energy evaluation is precomputed directly on the GPUs. Throughout the entire sequence of functions involved in propagating a walker, all operations are performed on the devices, i.e. without any data transfers or operations involving the CPUs. We utilize NVIDIA's Compute Unified Device Architecture (CUDA) Basic Linear Algebra Subprograms (cuBLAS) library to execute, e.g., the matrix multiplications that propagate walker determinants by a one-body operator, and have supplemented this library with custom C/CUDA functions which can be classified roughly into two types. Those in the first constitute a matrix library of kernels which carry out, most notably, element-wise matrix additions and matrix sums (i.e., matrix → scalar). These are used frequently to compute the trace of a matrix product, which is utilized in the calculation of expectation values such as the force-bias and local energy. In addition to such library-type functions, we also wrote GPU kernels to sample auxiliary-fields, compute the force-bias, assemble and exponentiate one-body operators, carry out the Sherman-Morrison-Woodbury (SMW) updates, orthonormalize the orbitals of walker determinants, and measure the local energy. With the exception of the SMW and energy measurement functions, which we will subsequently detail, the GPU port of the above functions did not involve notable algorithmic improvements over our CPU implementation.

Once the code enters the loops shown in Figure 1, data need only be transferred from the current device when the energy is measured, which for our typical time-step choice of $\Delta\tau = 0.005$ Ha^{-1} happens once in 20 propagation steps. An explanation of the slicing variant of the energy algorithm will be presented later in Section IV. For now we simply wish to emphasize that throughout the majority of an AFQMC calculation data does not need to leave the devices. This is in large part why the current implementation leads to such pronounced speed-ups compared to our initial attempts to simply offload the matrix multiplications.

A relatively new addition to the cuBLAS library is so-called "batched" functions which perform many smaller operations simultaneously, e.g. a set of matrix−matrix multiplications or lower-upper (LU) decompositions. These batched functions are well-suited for operations that are individually too small to parallelize effectively across thousands of cores. We utilize this feature heavily in our implementation of SMW updates to quickly compute equal-time Green's functions when multi-determinant trial functions are used. We note that previous Diffusion MC studies have utilized similar SMW updates,[58,65,66] and ref 67 has also presented fast updates and memory-saving techniques for multideterminant CI trial wave functions in AFQMC. Given a reference matrix A, the following formulas are used to compute the determinants and inverses of matrices which differ from A by one or more rows or columns:

$$\begin{aligned}\det(A + U_i V_i^T) &= \det(I + V_i^T A^{-1} U_i)\det(A)\\ (A + U_i V_i^T)^{-1} &= A^{-1} - A^{-1}U_i(I + V_i^T A^{-1} U_i)^{-1} V_i^T A^{-1}\end{aligned} \tag{4}$$

In the context of ph-AFQMC, suppose we use a multi-determinant trial wave function, $|\Psi_T\rangle = \sum_{i=0} c_i|\psi_{T,i}\rangle$, where $\langle\psi_{T,i}|\psi_{T,j}\rangle = \delta_{ij}$. Then, for the k^{th} walker determinant $|\phi_k\rangle$, $A =$

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

$[\psi_{T,i=0}]^{\dagger}[\phi_k]$, where the square brackets denote a matrix representation, U_i and V_i are of dimension $N_\sigma \times E_i$, where N_σ is either the number of spin-up or spin-down electrons, and E_i is the number of excitations required to form the i^{th} configuration of the multideterminant expansion from the reference configuration. The determinant and inverse of the reference matrix corresponding to zero excitations ($i = 0$) are computed first for spin-up and spin-down configurations, followed by batched SMW updates for all $i \neq 0$. Subcubic scaling with respect to particle number is achieved since $E_i \ll N_\sigma$. Figure 2

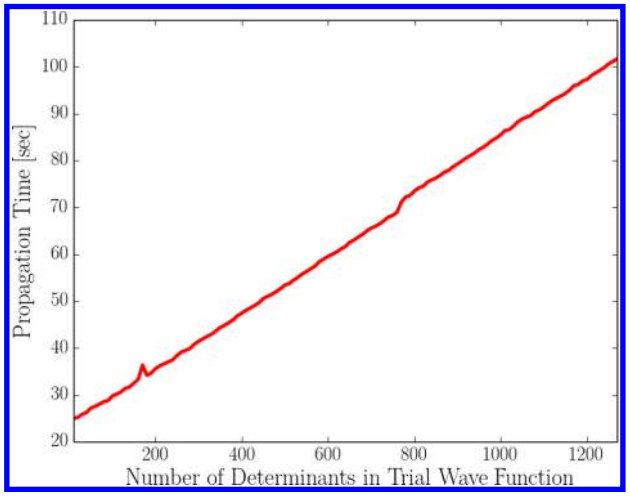


Figure 2. Propagation time vs the number of configurations in the CASSCF trial wave function for the Mn atom in the aug-cc-pwCVQZ-DK basis. Calculations use sp, a CD threshold of 10^{-4} Ha, 20 walkers, and imaginary-time trajectories of length 1 Ha^{-1} with a time step of $\Delta\tau = 0.005$ Ha^{-1}. Walker orthonormalization and local energy measurements were performed every 2 and 20 steps, respectively.

highlights the efficiency of our batched implementation of the SMW algorithm, for the Mn atom in the aug-cc-pwCVQZ-DK basis (185 basis functions, 25 electrons). "Propagation time" denotes the total wall-time minus the time spent on initial setup, e.g., memory allocation, input/output, and precomputation of the operators and required intermediates. Previously, going from, e.g., 10 to 1200 determinants would multiply the propagation time by a factor of 120. In contrast, our SMW algorithm reduces this to a mere factor of 3.9.

We have developed a GPU-optimized algorithm for evaluating the local two-electron energy of a walker. 4-Index tensors are precomputed once at the start of a simulation, which, in the spin-free and single-determinant trial case (for simplicity), are of the form

$$Y_{ijab} = \sum_{kl}\sum_{\alpha} L^{\alpha}_{ik}L^{\alpha}_{jl}[\psi^{\dagger}_{T,\mathrm{ak}}\psi^{\dagger}_{T,\mathrm{bl}} - \psi^{\dagger}_{T,\mathrm{al}}\psi^{\dagger}_{T,\mathrm{bk}}] \tag{5}$$

where the L^{α} arise from decomposing the two-electron integrals via CD or DF ($V_{ijkl} = \sum_{\alpha} L^{\alpha}_{ik}L^{\alpha}_{jl}$), indices i,j,k,l run from 1 to M (basis size), while indices a,b run from 1 to N (number of electrons). ψ_T is a matrix with columns composed of the orbitals in the trial function. Importantly, the sums over k,l and over auxiliary-fields α, the number of which typically scales as $2-10M$, need only be computed once at the start of the simulation. The energy is evaluated by pairing indices $i,a \rightarrow \gamma$ and $j,b \rightarrow \delta$, thereby flattening the 4-index tensor in (5) to a 2-index tensor and then performing the following contraction

$$E_{2e}[\phi] = Tr(QY) = \sum_{\gamma\delta} Q_{\gamma\delta}Y_{\gamma\delta} \tag{6}$$

where $Q_{\gamma\delta} = [\phi(\psi_T^{\dagger}\phi)^{-1}]_{\gamma}[\phi(\psi_T^{\dagger}\phi)^{-1}]_{\delta}$. The Q and Y matrices are of size $MN \times MN$, and hence the energy evaluation dominates the scaling, with respect to system size, of both the required memory and run-time of an AFQMC calculation. In Section IV we describe a strategy to split the memory burden among multiple GPU cards on a single node and suggest approaches to attain additional scalability in Section VI. In what follows we illustrate the efficiency of our current implementation.

Table 1 shows select performance metrics, from an analysis using NVIDIA's nvprof code profiler, for the GPU kernels

Table 1. Efficiency Metrics of the GPU Kernels Involved in Our Energy Algorithm[a]

kernel	% run-time	% comput util	reg/thread	% mem util (type)	% occ
CGEMM	88.2	95−100	84	50−60 (shared)	25
El. MatMul	6.8	<10	12	80−90 (device)	91.6
matrix sum	2.3	<10	16	80−90 (device)	96.1

[a]For each kernel we show the percentage of total run-time spent in that kernel, compute utilization as a percentage of peak compute performance, the number of registers per thread, memory utilization as a percentage of peak bandwidth (shown only for the memory type exhibiting the highest utilization), and the occupancy, i.e. the percentage of available warps (a group of 32 threads) that are active.

involved in our energy algorithm. While these metrics can, in general, vary widely depending on the particulars of both the device architecture and the description of the chemical system under study (e.g., choices of basis and trial function), we chose to optimize our code's performance for ph-AFQMC calculations using large multideterminant trials, in light of our interest in strongly correlated systems. As before, we show data for the Mn atom in the aug-cc-pwCVQZ-DK basis using sp, with 1200 determinants in the trial function (as is used to calculate the IP reported in Section V). Nearly 90% of the wall-time is spent in CGEMM from the cuBLAS library, which we use to compute a quantity analogous to Q in eq 6 but generalized to the case of a multideterminant trial function. It appears that our custom element-wise matrix multiplication and matrix sum kernels, while at peak warp utilization, are limited by the device memory (DRAM) bandwidth. Additional fine-tuning of the latter kernels' usage of the memory hierarchy will at best result in a small improvement in the overall performance of the energy function for this system (~1.1×, from Amdahl's law), given that the majority of the time is spent executing the highly optimized CGEMM kernel.

Finally we introduce the use of DF[68] in AFQMC calculations, where effective densities $\overline{\rho}_{ij}(\mathbf{r})$ are fit to auxiliary basis functions, $\chi(\mathbf{r})$:

$$V_{ijkl} = \int d\mathbf{r}_1 d\mathbf{r}_2 \phi_i(\mathbf{r}_1)\phi_j(\mathbf{r}_1)\frac{1}{r_{12}}\phi_k(\mathbf{r}_2)\phi_l(\mathbf{r}_2) \tag{7}$$

$$\sim \int d\mathbf{r}_1 d\mathbf{r}_2 \overline{\rho}_{ij}(\mathbf{r}_1)\frac{1}{r_{12}}\phi_k(\mathbf{r}_2)\phi_l(\mathbf{r}_2) \tag{8}$$

$$= \sum_{\nu} d^{ij}_{\nu}(\nu|kl) \tag{9}$$

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

The last equality follows from inserting $\bar{\rho}_{ij}(\mathbf{r}_1) = \sum_\nu d^{ij}_\nu \chi_\nu(\mathbf{r}_1)$ and defining the three-center integrals

$$(\nu|\mathrm{kl}) = \int d\mathbf{r}_1 d\mathbf{r}_2 \chi_\nu(\mathbf{r}_1)\frac{1}{r_{12}}\phi_k(\mathbf{r}_2)\phi_l(\mathbf{r}_2)$$

The expansion coefficients can be chosen such that $d^{ij}_\nu = \sum_\mu (ij|\mu)\mathbf{J}^{-1}_{\mu\nu}$, where $\mathbf{J}_{\mu\nu} = \int d\mathbf{r}_1 d\mathbf{r}_2 \chi_\mu(\mathbf{r}_1)\frac{1}{r_{12}}\chi_\nu(\mathbf{r}_2)$. Expressing $\mathbf{J}^{-1}_{\mu\nu}$ as a contraction over a third index allows the two-electron integrals to be written in a form suitable for AFQMC (cf. Section II):

$$V_{ijkl} = \sum_\alpha \left(\sum_\mu (ij|\mu)\mathbf{J}^{-1/2}_{\mu\alpha}\right)\left(\sum_\nu \mathbf{J}^{-1/2}_{\alpha\nu}(\nu|kl)\right) = \sum_\alpha L^\alpha_{ij} L^\alpha_{kl} \tag{10}$$

The number of terms in the sum over α is equal to the number of auxiliary-fields sampled by each walker in AFQMC, which via DF is typically reduced to ~2M. As a result the calculation of the force-bias and the assembly of the one-body operator in eq 2 can be done faster, and fewer L matrices (each with M^2 elements) need to be stored in memory relative to when CD is used. In addition, the smaller number of auxiliary-fields generally leads to a reduction in statistical noise. The accuracy of the DF approximation will be assessed in Section IV.

■ ILLUSTRATION WITH HYDROGEN CHAINS

In this section we explore the effects on both computational efficiency and accuracy due to the use of sp vs dp and DF vs CD for linear chains of hydrogen atoms. These systems have played an important role in benchmarking new theories of correlated electronic materials.[8,22,69−73] While these systems do not capture many nuances of more realistic molecular systems, they are nevertheless a useful prototype capable of (1) yielding wall-time and scaling insights due to the ability to systematically increase the system size, (2) providing an atomistic analogue of well-studied model systems such as the Heisenberg and Hubbard models albeit with a more realistic description of long-range Coulomb interactions, while (3) exhibiting strong static correlation at large bond lengths.

Computational Details. For all hydrogen chain calculations we use the cc-pVDZ basis, for which there is abundant benchmark data.[22] In this basis there are 5 basis functions per electron, a notably smaller number than used in typical molecular calculations. The Weigend Coulomb-fitting basis set[74] is employed as the auxiliary basis for DF, and CDs in this section employ a threshold of 10^{-5}Ha (as chosen in ref 22).

We use PySCF[75] to compute all inputs required of our ph-AFQMC code. Unless otherwise specified we use an imaginary-time step of 0.005 Ha^{-1}. Walker orbitals are orthonormalized after every two propagation steps, to preserve the antisymmetry of the walker configurations and also to keep the magnitude of orbital coefficients and associated quantities as small as possible (thus extending the accuracy of sp). We employ the hybrid method of ph-AFQMC[12] to minimize evaluations of the local energy, which is measured every 0.1 Ha^{-1}. The total number of walkers is fixed throughout each simulation, and when required we use a population control (PC) algorithm at intervals of 0.1 Ha^{-1}. Long imaginary-time runs utilizing PC use a reblocking analysis[76] to obtain statistical errors uncontaminated by autocorrelation. All calculations are run on NVIDIA GeForce GTX 1080 GPUs, with Intel Xeon E5-2620 v4 CPUs running at a maximum of 2.10 GHz. Further details about the CPU and GPU hardware can be found in the Supporting Information.

Timings. Employing an unrestricted HF trial for ph-AFQMC has been shown to produce very accurate energies for hydrogen chains near their equilibrium bond lengths.[22] Using a bond length of 1.880(2) Bohr as given by the Density Matrix Renormalization Group (DMRG) in the cc-pVDZ basis, we compare propagation times using a single GPU card for an increasing number of hydrogen atoms. Sample propagation times for several variants of precision and means of decomposing the two-electron terms are shown in Figure 3.

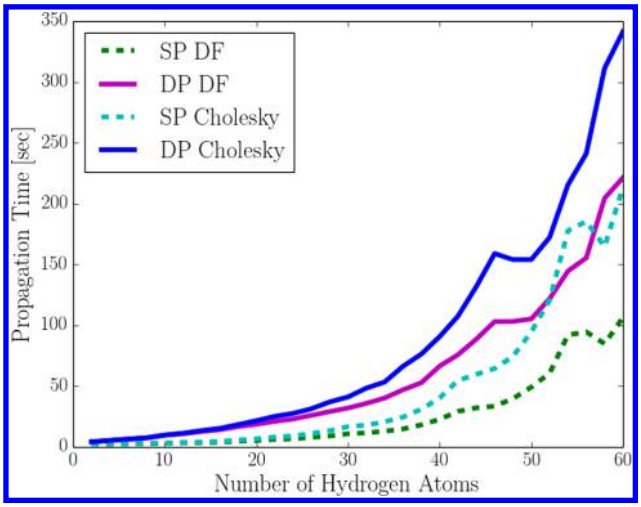


Figure 3. Propagation time using 1 GPU for hydrogen chains of varying lengths, comparing two types of two-electron integral decompositions, DF vs CD with a 10^{-5} cutoff, within both sp and dp. UHF trial functions are used, and 24 walkers are propagated for an imaginary-time segment of length 1 Ha^{-1}.

For H_{60} DF is 2.0× faster than CD in sp and 1.5× faster in dp. Sp is 2.1× faster than dp when DF is used and 1.6× faster using CD. Generally we find that the relative speed-ups afforded by sp over dp, and DF over CD, increase with system size. The nonmonotonicity of the propagation times vs system size is a unique and rather unexpected artifact of the GPU architecture, and we observe that the sp (dotted) and dp (lines) trajectories move together, suggesting a different treatment of sp and dp at the hardware level. The GPUs used in this work can perform sp and dp floating point operations at a maximum of 8876 and 277.36 GigaFLOP/s, respectively. Our observed speed-up going from dp to sp is significantly less than what these peak metrics would imply. This is because for the H chain sizes investigated here with single-determinant trials the GPU performance is not compute limited but rather bound by the device memory bandwidth. This suggests additional speed-ups can be expected for calculations of this type, and we plan to pursue further memory optimization in the near future.

In Table 2 we benchmark the performance of our GPU implementation for multideterminant trial functions with CD. We compare against our latest CPU code, which utilizes the same SMW algorithm but without the batching scheme and with cuBLAS kernels replaced by calls to equivalent functions in the Intel Math Kernel Library (MKL). The energy algorithm implemented in the CPU code also utilizes the precomputed tensor shown in eq 5. Furthermore, the CPU code defines the same C structure types for matrices in sp and dp and uses

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

Table 2. Propagation Times (in s) for an H_{50} Chain with a Varying Number of Determinants That Comprise the Trial Wave Function[a]

	N_{det} = 2	N_{det} = 50	N_{det} = 100	N_{det} = 500	N_{det} = 1000
GPU sp	99.9	105.3	111.8	158.0	218.7
CPU dp	8775.2	15019.7	22993.7	79713.3	148544.2
speed-up	87.8×	142.7×	205.6×	504.6×	679.1×

[a]We use CD with a 10^{-5} cutoff, and show the speed-up of a single GPU in sp over a single CPU in dp.

analogous algorithms to, e.g., copy and exponentiate these matrix structures. We believe that for these reasons a fair comparison between our CPU and GPU codes can be made. The GPU-accelerated code in sp achieves large speed-ups ranging from 87.2× with two determinants to 670.1× with 1000 determinants, compared to our CPU code in dp. Importantly we find that the relative speed-up increases with the number of determinants present in the trial function. This is due to efficient batched processing in the evaluation of mixed-expectation values involving the trial wave function.

To parallelize across GPU cards on a single node, we divide the total number of walkers into subsets which are independently propagated and measured on different GPU cards. We use Open Multi-Processing (OpenMP) to achieve shared-memory parallelization of the CPU threads, and to each CPU thread we associate a partner GPU device. Figure 4 highlights the near-unity parallel efficiency of our implementation, defined as the multi-GPU speed-up over 1 GPU divided by the number of GPUs utilized.

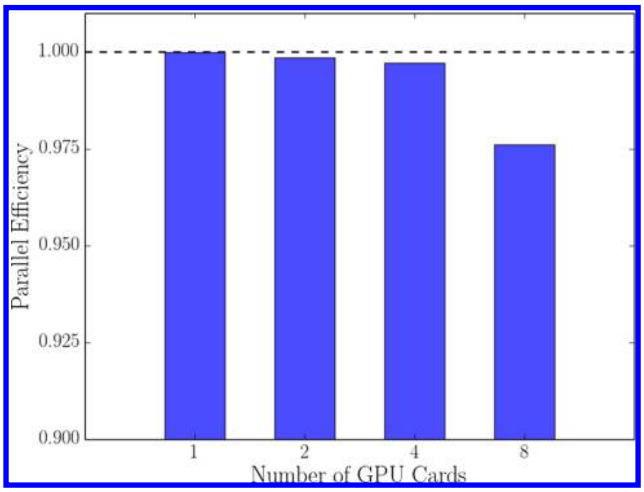


Figure 4. Parallel efficiency of our ph-AFQMC code illustrated on H_{50}. We use a CASSCF trial wave function with 44 determinants and 800 walkers propagated for 0.5 Ha^{-1} with $\Delta\tau$ = 0.01 Ha^{-1}.

To treat larger system sizes we have implemented a local memory strategy which spreads slices of the 4-dimensional tensors in eq 5 across 8 cards for the entire simulation. At the intervals where the energy is measured, the random walkers propagated on, e.g., GPU 0 are sent to GPUs 1−7 to compute the components of the two-electron energy derived from the elements stored locally on GPUs 1−7. This is done simultaneously for walkers on all GPUs, after which the energy components are gathered and tallied for each walker. The nodes utilized have 8 GPU cards each with 8 GB of RAM. Using DF and sp, this local memory-slicing algorithm enables us to treat systems as large as H_{100} in the double-ζ basis (M = 500, N = 100).

Accuracy. To benchmark the accuracy of our algorithm when sp and DF are used, we compute the total energy of H_{50} in the cc-pVDZ basis with a bond length of 1.8 Bohr and compare with results from a recent study[22] presenting data from state-of-the-art methods including ph-AFQMC, DMRG, and restricted CCSD(T), among others. DMRG is essentially exact for one-dimensional systems,[69] and RCCSD(T) is expected to provide a high level of accuracy as the bond-length is near its equilibrium value.[77]

Our GPU results are shown in Table 3 along with the previously published data. The differences in the total energies

Table 3. Total Electronic Energies [Ha] of H_{50} at R = 1.8 Bohr in the cc-pVDZ Basis[a]

	electronic energy	propagation time
GPU sp DF	−125.2107(7)	14.6
GPU dp DF	−125.2119(6)	31.4
GPU sp Chol	−125.2239(6)	27.8
GPU dp Chol	−125.2246(5)	45.2
CPU dp ref 22	−125.2242(8)	
RCCSD(T) ref 22	−125.2067	
DMRG ref 22	−125.2210(1)	

[a]Propagation times [hours] are presented using 8 CPU/GPU pairs. We use 1000 walkers propagated for a length of 200 Ha^{-1} (including equilibration).

of sp vs dp for our GPU calculations are 1.2(9) mHa for DF and 0.7(8) mHa for CD. Importantly, both of these are smaller than the resolution required for chemical accuracy (1.6 mHa), confirming that for this system size we can take advantage of the hardware-optimized sp arithmetic on the GPU without incurring a significant loss of accuracy.

DF produces about half the number of auxiliary-fields compared with CD (550 vs 1105), reducing propagation times by a factor of 1.9 for sp and 1.4 for dp. In terms of the resulting accuracy, it is well-known that while the DF decomposition may not be sufficient to produce total energies within chemical accuracy, it can recover sub-mHa accuracy in the calculation of relative energies.[78,79] Indeed, we find that for the total energy of H_{50} DF differs from CD with a 10^{-5} cutoff by 13.2(9) mHa in sp and 12.7(8) mHa in dp, respectively. Yet to put these errors into context we note in passing that DF ph-AFQMC in both sp and dp produces total energies for this system that are closer to the DMRG reference by ∼4 mHa than RCCSD(T), known to many as the "gold standard" of quantum chemistry.[80,81]

To conclude this section we illustrate the capacity of DF and more aggressive CD truncation thresholds to recover chemically accurate energy differences for the deprotonation of methanol. Table 4 shows errors of ∼3 mHa for the total energies of the neutral and deprotonated species; however, the deviation of the energy difference from that of the most stringent CD cutoff is negligible, taking statistical errors into account. In this molecular case, compared to H_{50}, we find a more pronounced reduction in the number of auxiliary-fields, implying a ∼4× speed-up (vs ∼2× for the hydrogen chain) over CD with a 10^{-5} threshold.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

Table 4. Accuracy of DF and CDs with Various Cutoffs for the Deprotonation Energy of Methanol[a]

	N_{AFs}	MeOH	MeO^-	ΔE	$\Delta E - \Delta E_{CD10^{-6}}$
DF	142	−155.8503(4)	−148.5340(7)	0.6777(8)	−0.0004(10)
CD 10^{-2}	187	−155.8144(6)	−148.4985(4)	0.6772(7)	−0.0008(10)
CD 10^{-3}	385	−155.8531(4)	−148.5357(4)	0.6787(6)	0.0007(9)
CD 10^{-4}	471	−155.8542(4)	−148.5366(4)	0.6790(6)	0.0009(9)
CD 10^{-5}	617	−155.8544(4)	−148.5371(3)	0.6787(6)	0.0007(8)
CD 10^{-6}	855	−155.8533(5)	−148.5366(4)	0.6780(6)	0

[a]Sp is used, and long imaginary-time trajectories are stabilized with PC. N_{AFs} denotes the resulting number of auxiliary-fields.

Table 5. Target Electron Configurations and Spin-Multiplicities (2S + 1), from Refs 82 and 83

system	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn
neutral	$4s^23d^1$	$4s^23d^2$	$4s^23d^3$	$4s^13d^5$	$4s^23d^5$	$4s^23d^6$	$4s^23d^7$	$4s^23d^8$	$4s^13d^{10}$	$4s^23d^{10}$
spin mult	2	3	4	7	6	5	4	3	2	1
cation	$4s^13d^1$	$4s^13d^2$	$3d^4$	$3d^5$	$4s^13d^5$	$4s^13d^6$	$3d^8$	$3d^9$	$3d^{10}$	$4s^13d^{10}$
spin mult	3	4	5	6	7	6	3	2	1	2

IPS OF TRANSITION METAL ATOMS

In this section we compute the IPs of the first-row TM atoms correlating *all* electrons and compare the calculated ph-AFQMC results to experiment and previous electronic structure calculations.

Computational Details. Our computational protocol begins with a restricted (open-shell) HF calculation. We visually inspect the occupied orbitals of this solution to ensure that the electron configurations shown in Table 5 are obtained. For some atomic species, HF provides a qualitatively incorrect description of the single-particle orbital occupancies, requiring initialization from custom density matrices to converge subsequent HF calculations to the target ground-state configurations. We note that for the V^+ cation the initial density matrix guess was constructed with the $L = 2$ orbital unoccupied. In all cases the canonical HF orbitals are used to initialize a restricted CASSCF calculation.

All ph-AFQMC calculations in this section use a CD cutoff of 10^{-4}. We utilize basis sets that have been optimized to account for scalar relativistic effects,[83] and use the spin-free exact two-component approach[84,85] to decouple the electronic degrees of freedom from the Dirac equation. This approximation produces one-body terms which we simply add to the nonrelativistic Hamiltonian in eq 1.

To compare calculations in finite basis sets to experiments we extrapolate the correlation energies to the Complete Basis Set (CBS) limit using two data points fit to $1/x^3$ (x = 3,4 for TZ,QZ).[15,83,86] We confirmed for a subset of the atoms that the inclusion of the aug-cc-pwCV5Z-DK energies did not significantly change the extrapolated results, consistent with ref 83. Following ref 82 and our own observation that the HF energies converge relatively quickly in this sequence of basis sets, we use the 5Z value for the CBS HF energies.

The Trotter error due to finite imaginary-time discretization can be extrapolated to 0 using progressively smaller time steps. Here we use $\Delta\tau$ = 0.005, 0.01, and 0.02 Ha^{-1}. For Co through Zn we compared the CBS estimate from such an extrapolation with values from the smallest time step only, $\Delta\tau$ = 0.005 Ha^{-1}. In the latter approach we observe a substantial yet systematic cancellation of error, and CBS estimates of equivalent accuracy compared to the 3-point extrapolation approach are shown in Table 6. In light of this data we use only $\Delta\tau$ = 0.005 Ha^{-1} for all calculations.

Table 6. Comparison of CBS IPs [eV] for Co, Ni, Cu, and Zn with $\Delta\tau \to 0$ vs $\Delta\tau$ = 0.005 Ha^{-1} Computed in sp with ph-AFQMC/PC

	Co	Ni	Cu	Zn
expt	7.87	7.59	7.73	9.39
$\Delta\tau \to 0$	7.87(3)	7.61(3)	7.54(3)	9.33(4)
$\Delta\tau$ = 0.005 Ha^{-1}	7.89(3)	7.59(3)	7.55(3)	9.37(3)

Details of the CS procedure can be found in ref 21. In short, we run a set of independent calculations called repeats, each of which uses a distinct random number seed to propagate both the neutral and cationic species such that pairs of walkers sample the same auxiliary-fields. After an initial equilibration period, cumulative averages of the energy difference are computed along each of the imaginary-time trajectories and are averaged among the set of repeats to obtain an estimate of statistical error. In the present case, stochastic error cancellation leads to a pronounced reduction in the variance of the IPs, and convergence at very short imaginary times can be achieved when the standard error drops below the target error tolerance and upon visual observation of a plateau in the measured quantity. We will show in the next section that this CS approach leads not only to significant reductions in computational cost, relative to the uncorrelated approach but also to systematically improved accuracy.

Results and Discussion. Tables 7 and 8 summarize our results for the all-electron IPs of the first-row TM atoms. We show values obtained from both PC and CS ph-AFQMC approaches and compare with experimental and CCSD(T) values.

Table 7. Calculated ph-AFQMC IPs [eV] in the CBS Limit Computed with sp and $\Delta\tau$ = 0.005 Ha^{-1}, Compared with Experimental and CCSD(T) Values[b]

	Sc	Ti	V	Cr	Mn
ph-AFQMC/PC	6.51(1)	6.71(2)	6.74(1)	6.75(2)	7.41(2)
ph-AFQMC/CS	6.52(3)	6.80(3)	6.74(3)	6.74(3)	7.45(3)
expt	6.56	6.83	6.73	6.77	7.43
CCSD(T)[a]	6.54	6.81	6.73	6.79	7.42

[a]Reference 83. [b]Experimental IPs have spin-orbit contributions removed.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

Table 8. Same as Table 7, but for Atoms in the Right-Half of the Row

	Fe	Co	Ni	Cu	Zn
ph-AFQMC/PC	7.86(2)	7.89(3)	7.59(3)	7.55(3)	9.37(3)
ph-AFQMC/CS	7.89(2)	7.87(3)	7.61(2)	7.68(3)	9.37(3)
expt	7.90	7.87	7.59	7.73	9.39
CCSD(T)	7.89	7.88	7.59	7.72	9.37

The active spaces employed for the neutral and cationic species in the TZ and QZ basis sets are described in detail in Tables 9 and 10. In general, the use of truncated CASSCF trial wave functions in ph-AFQMC involves subtleties that require careful consideration, since the CASSCF calculation itself can become an expensive preprocessing step when large active spaces are required, and the truncation breaks size extensivity. This approach is viable if the ph-AFQMC result converges quickly with trial wave functions generated from active spaces much smaller than the full Hilbert space. For atoms and molecules this is typically the case, and an internal validation procedure within ph-AFQMC can be employed involving a series of calculations using various active space and truncation cutoffs. In particular, for Fe−Zn we started by including the 4*s* and 3*d* electrons in an active space composed of 13 active orbitals. While the resulting truncated CASSCF trial wave functions produced sufficiently accurate ph-AFQMC/PC results in the CBS and $\Delta\tau \rightarrow 0$ limits for Co, Ni, and Zn, 18 orbitals were required in the case of Fe. The improvement in the IP resulting from the inclusion of a second shell of *d* orbitals in the CASSCF active space is a manifestation of the so-called "double-shell" effect.[87,88] We find that this effect is less pronounced in the case of all-electron ph-AFQMC since the application of $e^{-\tau\hat{H}}$ to walker configurations can explore the space of excitations into virtual *d* orbitals even if such excitations are not represented in the trial function.

For the left half of the first row of transition metals in the periodic table, Sc−Mn, we designate the 3*p* electrons as active in addition to the 4*s* and 3*d* electrons. In an effort to maintain consistency (i.e., to include HF virtuals of similar character in the initial guesses for the CASSCF procedure) among all atoms in the row, for those in the left-half we start with 16 active orbitals. This produced accurate ph-AFQMC/PC results for Sc and Cr. For V we noticed a sharp drop in energy in both the neutral and cationic species going from 16 to 19 active orbitals; for Mn an accurate IP required the replacement of three 5*p* orbitals with five 4*d* in the CASSCF active space to accommodate the double-shell effect.

The case of Cu proves to be particularly challenging and illustrates an additional merit of the CS approach. A trial function with 18 active orbitals approaches the memory limit of traditional CASSCF solvers but is still insufficient to produce results of the desired accuracy within ph-AFQMC/PC. With additional active orbitals, approximate CASSCF solvers utilizing DMRG[89] did converge, but only a subset of the resulting configurations and CI coefficients could be accessed with the current implementation of selected CI in PySCF. Even with moderate selection cutoffs, when such a wave function was used as a trial function in ph-AFQMC, we found a significant increase in statistical error, in addition to larger deviations of the resulting IP from experiment.

In contrast to regular ph-AFQMC/PC, which stabilizes long imaginary-time trajectories, a key advantage of the CS approach is that averaging among independent repeats at short times allows for not only a vast variance reduction when the auxiliary fields are correlated but also the ability to converge measurements of the energy difference *before* the full onset of the bias that results from the phaseless constraint. Even though the phaseless approximation is made after each time step, the walker weights at early times stay relatively closer to their true unconstrained values than at long times when the phaseless constraint has fully equilibrated. To illustrate this we plot the IP of Cu in the TZ basis at short imaginary-times in Figure 5. At longer imaginary times (not shown) the CS IP appears to approach the ph-AFQMC/PC result (albeit with substantial noise due to the absence of PC), yet from 2 to 7 Ha^{-1} ph-AFQMC/CS unambiguously converges on an answer consistent with iFCI-QMC, which is expected to be very accurate here.[82] Moreover, this value after CBS extrapolation is within range of chemical accuracy with respect to experiment.

The case of Ni is also quite remarkable. Both CS and PC methods produce IPs consistent with the experimental value and each other in the CBS limit; however, a detailed comparison with CCSD(T) values in each basis set, shown in Table 11, reveals that this agreement is due to fortuitous cancellations of error. While the CCSD(T) values approach the CBS limit from above, the ph-AFQMC/PC values approach the same value from below. ph-AFQMC/CS calculations, on the other hand, produce statistically consistent results with CCSD(T) in each basis and in the CBS limit.

For the case of Ti, having observed a quick equilibration time in the PC run with 16 active orbitals we chose to use CS as a much cheaper alternative to further ncreasing the size of the active space. We note, however, that this alternative may not always be feasible, e.g. when a poor trial function results in long equilibration times. Generally, for all atoms in this work CS results exhibit equivalent or better accuracy compared to the conventional method of running ph-AFQMC calculations with PC. Moreover, the ability to consistently produce chemically accurate results while using sp is reassuring, given that the total energies involved in these calculations are on the order of −1000 Ha. This implies that mHa energy scales require precision out to at least 7 significant figures, which would be stretching the typical capabilities of sp arithmetic in deterministic algorithms. In the future any differences in the sensitivity of stochastic vs deterministic algorithms to

Table 9. Number of Active Electrons and Orbitals in the CASSCF Trial Wave Functions for the Cation/Neutral Species and the Number of Determinants Kept in the ph-AFQMC Trial Function Accounting for 99.5% of the CI Weight[b]

	Sc	Ti	V	Cr	Mn
active space	8/9e,16o	9/10e,16o	10/11e,19o	11/12e,16o	12/13e,18o[a]
N_{dets} TZ	146/224	240/442	366/751	303/271	423/584
N_{dets} QZ	143/439	293/388	300/903	92/262	852/1266

[a]Three 5*p* orbitals replaced by five 4*d* orbitals in the active space. [b]For all species in this table the 3p electrons are active.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

Table 10. Same as Table 9, but for Cu and Zn with 99.0% of the CI Weight Retained

	Fe	Co	Ni	Cu	Zn
active space	7/8e,18o	8/9e,13o	9/10e,13o	10/11e,18o	11/12e,13o
N_{dets} TZ	23/227	210/85	138/156	374/322	299/518
N_{dets} QZ	23/121	237/66	159/161	504/507	277/526

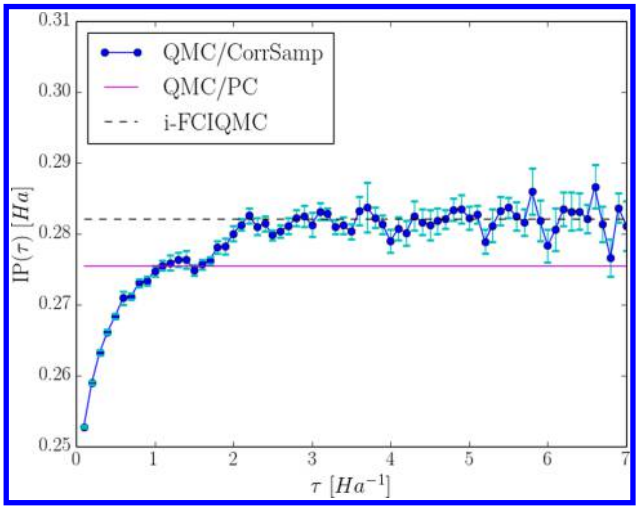


Figure 5. Comparison of the IP of Cu as a function of imaginary-time produced from ph-AFQMC/CS in the aug-cc-pwCVQZ-DK basis compared to the regular ph-AFQMC/PC result in the same basis. The i-FCIQMC result in the aug-cc-pVQZ-DK basis is indicated by a dashed line.

Table 11. Comparison of ph-AFQMC IPs for Ni, As Obtained with CS and Regular ph-AFQMC/PC, with CCSD(T) in Triple- and Quadruple-Zeta Basis Sets and in the CBS Limit[83a]

	CCSD(T)	ΔQMC/CS	ΔQMC/PC
TZ	7.68	7.68(1)	7.56(2)
QZ	7.63	7.64(1)	7.57(1)
CBS	7.59	7.61(2)	7.59(3)

[a]The CCSD(T) values were obtained with a composite method, namely cc-pVxZ-DK results plus a core−valence correction, which is the difference in the cc-pwCVxZ-DK basis of CCSD(T) calculations with active spaces defined by 3*s*3*p*3*d*4*s* and 3*d*4*s* orbitals (x = T,Q). QMC results used the aug-cc-pwCVxZ-DK basis sets.

numerical precision could be explored further using half-precision; however, on our current GPUs the peak performance of half-precision in GigaFLOP/s is 128× slower than sp and 4× slower than dp.

Table 12 shows the total propagation-times required to produce the final IPs in Tables 7 and 8, which account for calculations of the total energies of the neutral and cation in the TZ and QZ basis sets. 8-Core CPU times are estimated by scaling the propagation time of a small 20 walker system propagated for 1 Ha^{-1} by the required factors to reproduce the parameters of the GPU/PC calculations, i.e. 2000 walkers propagated in the TZ/QZ bases for 120/130 Ha^{-1} for Cr and Fe and for 200/230 Ha^{-1} for Cu. We assume perfect parallel efficiency in projecting our single-core CPU estimates to 8-cores. Obtaining comparable error bars using CS with our GPU code required the propagation of 200 walkers in the TZ/QZ bases for 5/3 Ha^{-1} for Cr, 4/3 Ha^{-1} for Cr, and 10/6 Ha^{-1} for Cu. For Cr and Cu we use 16 repeats in both basis sets, and for Fe we found that only 5/8 repeats in TZ/QZ are needed. We note that the larger wall-times for Cu are due to 1) the larger number of both particles and determinants in the trial functions employed and 2) the relatively poor trial function (compared to the exact ground state) which leads to longer propagation and equilibration times in the PC and CS methods, respectively. It may be the case that the relatively small atomic radius of Cu results in larger dynamical correlations compared to the rest of the atoms in the row, which are unaccounted for in the CASSCF trial wave functions (this explanation is consistent with the relative difficulty we encountered previously in calculating the electron affinity of the flourine atom[21]). The corresponding speed-ups for these selected atoms are shown in Table 13.

Table 12. Total Propagation Times [h] Required To Produce the Final All-Electron ph-AFQMC IPs in This Work

	Cr	Fe	Cu
est. 8-core CPU PC dp	9780	6970	34200
8-card GPU PC sp	50.5	47.5	136.4
8-card GPU CS sp	2.4	0.9	6.0

Table 13. Speed-ups Corresponding to the Timings in Table 12[a]

	Cr	Fe	Cu
GPU PC vs CPU PC	194×	138×	250×
GPU CS vs GPU PC	21×	53×	23×
GPU CS vs CPU PC	4100×	7700×	5700×

[a]All CPU/GPU calculations use dp/sp respectively.

We conclude this section with a few remarks. Currently we use a simple combing method[90] to implement PC. More sophisticated schemes are possible which may improve the statistical accuracy of the calculation. While this would slightly reduce the wall-times for the ph-AFQMC/PC method, the accuracy of the results with respect to experiments will be unchanged, since any bias due to PC vanishes when a large population (∼2000 walkers) is used. At the time of writing, auxiliary basis sets optimized for the scalar relativistic Hamiltonian and DK basis sets used in this work are not publicly available. The ability to use a DF decomposition would certainly provide additional speed-ups, although its effect on accuracy remains to be tested for these TM systems. Finally, we note that the capacity of our GPU code to treat $O(1000)$ determinants in the trial wave function will likely enable the accurate study of many strongly correlated systems. We anticipate that fewer determinants will be needed for metal−ligand complexes (in which the ligand is a nonmetal), as TM atoms typically exhibit larger static correlation effects than most coordinated complexes. In addition, the use of symmetry constraints in the CASSCF calculations will greatly reduce the number of configurations in the CI expansions.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

CONCLUSIONS AND OUTLOOK

We have designed a GPU implementation of ph-AFQMC for single- and multideterminant trial wave functions which can drastically reduce the scaling prefactor in realistic electronic structure calculations with near-unity parallel efficiency. Our strategy utilizes new batched SMW and energy algorithms, along with the ability to use sp and the DF decomposition. We validate performance enhancements with ph-AFQMC calculations of linear chains of hydrogen atoms and the atomic IPs of Sc through Zn, finding speed-ups relative to the CPU in dp of 2 orders of magnitude and which increase with the number of determinants in the trial wave function. For H_{50} and TM IPs, sp is sufficient to produce accuracy on the scale of 1 kcal/mol with respect to exact methods and experiment, respectively. In this work we also demonstrate that our previously outlined CS approach to ph-AFQMC enables both additional speed-ups of an order of magnitude, as well as the ability to converge measurements before the full onset of the bias due to the phaseless constraint. For all TM atoms, CS produces equivalent and often more accurate IPs and in a fraction of the wall-time.

While we have shown that the code segment which scales most steeply with system size, i.e. the energy algorithm, has been implemented with a very high level of device utilization, we anticipate that calculations on small systems, and especially those employing single-determinant trial functions, can still be substantially accelerated by additional tuning iterations in which the various utilization metrics from nvprof are prioritized. However, it must be stressed that the optimal choice of parameters (e.g., grid and block sizes) and memory strategy to most efficiently utilize the device architecture will depend on the particular choice of hardware. For this reason we postpone these fine-tuning optimizations until a target application on a large-scale computing cluster is ascertained.

We are optimistic that in the near future the investigation of many large, realistic systems will be feasible with ph-AFQMC. In what follows we anticipate issues of scalability and describe the possible solutions we envision. While our current implementation exclusively uses NVIDIA hardware with CUDA and cuBLAS, it would be straightforward to adapt it to a more universal standard, e.g., Open Computing Language (OpenCL) and associated BLAS packages. To enable large-scale calculations that efficiently utilize available High-Performance Computing clusters, we have designed a simple and scalable scheme to parallelize *across* GPU nodes for cases in which all required data for a ph-AFQMC calculation can be stored on a single node. Once a small population is equilibrated, walker data can be copied to all available nodes and used to initialize independent trajectories on separate nodes, each with a different random number seed. These subtrajectories can later be combined into a single trajectory from which averages and error bars can be obtained.

We note that our current memory limitation is rather artificial in the sense that GPU architectures and computing capabilities are improving at a rapid pace, suggesting that the memory capacity of GPU cards will continue to increase. Also, we are currently working on an implementation that uses Message Passing Interface (MPI) to extend our local memory-slice scheme to multiple nodes. We do not anticipate that transfer costs will be prohibitive since interprocess communication need only take place when the energy is measured (or when PC is performed), which occurs once in 20 time steps for our typical choice of $\Delta\tau = 0.005\ \mathrm{Ha}^{-1}$. This overhead is, moreover, a worthwhile trade-off since the pooling of GPU memory from multiple nodes will enable ph-AFQMC calculations of large systems. At the time of writing, NVIDIA's NVLink boasts transfer speeds of ∼300 GB/s between Tesla V100 GPUs, and we expect that future improvements in device-to-device and host-to-device transfer speeds will further reduce the overhead associated with MPI communication or possibly other strategies utilizing CPU memory to store the high-dimensional tensors.

The incorporation of additional theoretical approaches that would capture the same level of accuracy with diminished computational cost are also currently under consideration. One possibility is the use of Canonical Transcorrelation theory to produce an effective Hamiltonian with the electron cusps analytically removed,[91,92] which is related to the Jastrow factor in DMC. We anticipate that this will reduce the number of basis functions required to reach the CBS limit.[93,94] In addition, one may exploit real-space locality in the context of electronic excitations, which also has the potential to drastically reduce both the current scaling and memory demands.[95−97] Lastly, we note that when sp is insufficient, mixed-precision matrix strategies, which are well-studied and relatively straightforward to implement,[44,58,98] may be employed.

Combining the speed-ups due to the GPU and CS, we now have a robust and efficient computational protocol that is approximately 3 orders of magnitude faster than previous AFQMC procedures. This will enable routine ph-AFQMC calculations of a variety of chemically relevant properties with an unprecedented level of throughput and systematically improvable accuracy. Once the memory bottleneck is alleviated with the strategies mentioned above, systems that previously were inaccessible to study with ph-AFQMC will soon be within reach. Future targets include low-energy redox- and spin- states of catalytic metalloporphyrins[99,100] and iron−sulfur clusters[101] and the computation of pK_as for the oxygen evolving complex of Photosystem II.[102] Our method may also open the door to the accurate and fully *ab initio* investigation of strongly correlated solids such as high-temperature superconducting materials.[103,104] These and other targets of study will be the subject of future investigations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00342.

Detailed specifications of the GPU and CPU utilized in this work (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: js4564@columbia.edu.

ORCID

James Shee: 0000-0001-8333-8151

Richard A. Friesner: 0000-0002-1708-9342

Funding

D.R.R. acknowledges funding from NSF CHE-1464802, and S.Z. acknowledges funding from NSF DMR-1409510.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

J.S. gratefully acknowledges Mario Motta and Hao Shi for providing scripts to initialize HF calculations of hydrogen chains and V^+, respectively, and Qiming Sun for help with PySCF. J.S. would also like to thank Roald Hoffmann and Kirk A. Peterson for insightful discussions about TM atoms.

■ REFERENCES

(1) Zhang, S.; Carlson, J.; Gubernatis, J. E. Constrained path quantum Monte Carlo method for fermion ground states. *Phys. Rev. Lett.* **1995**, *74*, 3652.
(2) Zhang, S.; Krakauer, H. Quantum Monte Carlo method using phase-free random walks with Slater determinants. *Phys. Rev. Lett.* **2003**, *90*, 136401.
(3) Zhang, S. 15 Auxiliary-Field Quantum Monte Carlo for Correlated Electron Systems. *Emergent Phenomena in Correlated Matter: Autumn School Organized by the Forschungszentrum Jülich and the German Research School for Simulation Sciences at Forschungszentrum Jülich 23−27 September 2013; Lecture Notes of the Autumn School Correlated Electrons 2013*; 2013; Vol. 3.
(4) Motta, M.; Zhang, S. Ab initio computations of molecular systems by the auxiliary-field quantum Monte Carlo method. arXiv preprint arXiv:1711.02242, 2017.
(5) Al-Saidi, W.; Zhang, S.; Krakauer, H. Auxiliary-field quantum Monte Carlo calculations of molecular systems with a Gaussian basis. *J. Chem. Phys.* **2006**, *124*, 224101.
(6) Al-Saidi, W.; Krakauer, H.; Zhang, S. Auxiliary-field quantum Monte Carlo study of TiO and MnO molecules. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *73*, 075103.
(7) Al-Saidi, W.; Krakauer, H.; Zhang, S. Auxiliary-field quantum Monte Carlo study of first-and second-row post-d elements. *J. Chem. Phys.* **2006**, *125*, 154110.
(8) Al-Saidi, W. A.; Zhang, S.; Krakauer, H. Bond breaking with auxiliary-field quantum Monte Carlo. *J. Chem. Phys.* **2007**, *127*, 144101.
(9) Suewattana, M.; Purwanto, W.; Zhang, S.; Krakauer, H.; Walter, E. J. Phaseless auxiliary-field quantum Monte Carlo calculations with plane waves and pseudopotentials: Applications to atoms and molecules. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *75*, 245123.
(10) Al-Saidi, W.; Krakauer, H.; Zhang, S. A study of H + H_2 and several H-bonded molecules by phaseless auxiliary-field quantum Monte Carlo with plane wave and Gaussian basis sets. *J. Chem. Phys.* **2007**, *126*, 194105.
(11) Purwanto, W.; Al-Saidi, W.; Krakauer, H.; Zhang, S. Eliminating spin contamination in auxiliary-field quantum Monte Carlo: Realistic potential energy curve of F_2. *J. Chem. Phys.* **2008**, *128*, 114309.
(12) Purwanto, W.; Krakauer, H.; Zhang, S. Pressure-induced diamond to β-tin transition in bulk silicon: A quantum Monte Carlo study. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *80*, 214116.
(13) Purwanto, W.; Krakauer, H.; Virgus, Y.; Zhang, S. Assessing weak hydrogen binding on Ca^+ centers: An accurate many-body study with large basis sets. *J. Chem. Phys.* **2011**, *135*, 164105.
(14) Virgus, Y.; Purwanto, W.; Krakauer, H.; Zhang, S. Ab initio many-body study of cobalt adatoms adsorbed on graphene. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2012**, *86*, 241406.
(15) Purwanto, W.; Zhang, S.; Krakauer, H. Frozen-orbital and downfolding calculations with auxiliary-field quantum Monte Carlo. *J. Chem. Theory Comput.* **2013**, *9*, 4825−4833.
(16) Virgus, Y.; Purwanto, W.; Krakauer, H.; Zhang, S. Stability, energetics, and magnetic states of cobalt adatoms on graphene. *Phys. Rev. Lett.* **2014**, *113*, 175502.
(17) Purwanto, W.; Zhang, S.; Krakauer, H. An auxiliary-field quantum Monte Carlo study of the chromium dimer. *J. Chem. Phys.* **2015**, *142*, 064302.
(18) Ma, F.; Purwanto, W.; Zhang, S.; Krakauer, H. Quantum Monte Carlo calculations in solids with downfolded Hamiltonians. *Phys. Rev. Lett.* **2015**, *114*, 226401.
(19) Purwanto, W.; Zhang, S.; Krakauer, H. Auxiliary-field quantum Monte Carlo calculations of the molybdenum dimer. *J. Chem. Phys.* **2016**, *144*, 244306.
(20) Ma, F.; Zhang, S.; Krakauer, H. Auxiliary-field quantum Monte Carlo calculations with multiple-projector pseudopotentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2017**, *95*, 165103.
(21) 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.
(22) Motta, M.; et al. Towards the Solution of the Many-Electron Problem in Real Materials: Equation of State of the Hydrogen Chain with State-of-the-Art Many-Body Methods. *Phys. Rev. X* **2017**, *7*, 031059.
(23) 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.
(24) Zheng, B.-X.; Chung, C.-M.; Corboz, P.; Ehlers, G.; Qin, M.-P.; Noack, R. M.; Shi, H.; White, S. R.; Zhang, S.; Chan, G. K.-L. Stripe order in the underdoped region of the two-dimensional Hubbard model. *Science* **2017**, *358*, 1155−1160.
(25) Møller, C.; Plesset, M. S. Note on an approximation treatment for many-electron systems. *Phys. Rev.* **1934**, *46*, 618.
(26) Bartlett, R. J.; Musiał, M. Coupled-cluster theory in quantum chemistry. *Rev. Mod. Phys.* **2007**, *79*, 291.
(27) Roos, B. O. The Complete Active Space Self-Consistent Field Method and its Applications in Electronic Structure Calculations. *Adv. Chem. Phys.* **2007**, 399−445.
(28) Andersson, K.; Malmqvist, P.-Å.; Roos, B. O. Second-order perturbation theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* **1992**, *96*, 1218−1226.
(29) Ufimtsev, I. S.; Martinez, T. J. Quantum chemistry on graphical processing units. 2. Direct self-consistent-field implementation. *J. Chem. Theory Comput.* **2009**, *5*, 1004−1015.
(30) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436.
(31) Anderson, J. A.; Lorenz, C. D.; Travesset, A. General purpose molecular dynamics simulations fully implemented on graphics processing units. *J. Comput. Phys.* **2008**, *227*, 5342−5359.
(32) Ufimtsev, I. S.; Martinez, T. J. Quantum chemistry on graphical processing units. 1. Strategies for two-electron integral evaluation. *J. Chem. Theory Comput.* **2008**, *4*, 222−231.
(33) Luehr, N.; Ufimtsev, I. S.; Martínez, T. J. Dynamic precision for electron repulsion integral evaluation on graphical processing units (GPUs). *J. Chem. Theory Comput.* **2011**, *7*, 949−954.
(34) Kalinowski, J.; Wennmohs, F.; Neese, F. Arbitrary Angular Momentum Electron Repulsion Integrals with Graphical Processing Units: Application to the Resolution of Identity Hartree-Fock Method. *J. Chem. Theory Comput.* **2017**, *13*, 3160.
(35) Kussmann, J.; Ochsenfeld, C. Hybrid CPU/GPU integral engine for strong-scaling ab initio methods. *J. Chem. Theory Comput.* **2017**, *13*, 3153−3159.
(36) Ufimtsev, I. S.; Martinez, T. J. Graphical processing units for quantum chemistry. *Comput. Sci. Eng.* **2008**, *10*, 26−34.
(37) Yoshikawa, T.; Nakai, H. Linear-scaling self-consistent field calculations based on divide-and-conquer method using resolution-of-identity approximation on graphical processing units. *J. Comput. Chem.* **2015**, *36*, 164−170.
(38) Yasuda, K. Accelerating density functional calculations with graphics processing unit. *J. Chem. Theory Comput.* **2008**, *4*, 1230−1236.
(39) Genovese, L.; Ospici, M.; Deutsch, T.; Méhaut, J.-F.; Neelov, A.; Goedecker, S. Density functional theory calculation on many-cores hybrid central processing unit-graphic processing unit architectures. *J. Chem. Phys.* **2009**, *131*, 034103.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

(40) Hacene, M.; Anciaux-Sedrakian, A.; Rozanska, X.; Klahr, D.; Guignon, T.; Fleurat-Lessard, P. Accelerating VASP electronic structure calculations using graphic processing units. *J. Comput. Chem.* **2012**, *33*, 2581−2589.
(41) Meredith, J. S.; Alvarez, G.; Maier, T. A.; Schulthess, T. C.; Vetter, J. S. Accuracy and performance of graphics processors: A Quantum Monte Carlo application case study. *Parallel Comput.* **2009**, *35*, 151−163.
(42) Levy, T.; Cohen, G.; Rabani, E. Simulating lattice spin models on graphics processing units. *J. Chem. Theory Comput.* **2010**, *6*, 3293−3301.
(43) Block, B.; Virnau, P.; Preis, T. Multi-GPU accelerated multi-spin Monte Carlo simulations of the 2D Ising model. *Comput. Phys. Commun.* **2010**, *181*, 1549−1556.
(44) Olivares-Amaya, R.; Watson, M. A.; Edgar, R. G.; Vogt, L.; Shao, Y.; Aspuru-Guzik, A. Accelerating correlated quantum chemistry calculations using graphical processing units and a mixed precision matrix multiplication library. *J. Chem. Theory Comput.* **2010**, *6*, 135−144.
(45) Doran, A. E.; Hirata, S. Monte Carlo MP2 on Many Graphical Processing Units. *J. Chem. Theory Comput.* **2016**, *12*, 4821−4832.
(46) Song, C.; Martínez, T. J. Atomic orbital-based SOS-MP2 with tensor hypercontraction. I. GPU-based tensor construction and exploiting sparsity. *J. Chem. Phys.* **2016**, *144*, 174111.
(47) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. Accelerating Resolution-of-the-Identity Second-Order Møller- Plesset Quantum Chemistry Calculations with Graphical Processing Units. *J. Phys. Chem. A* **2008**, *112*, 2049−2057.
(48) DePrince, A. E., III; Hammond, J. R. Coupled cluster theory on graphics processing units I. The coupled cluster doubles method. *J. Chem. Theory Comput.* **2011**, *7*, 1287−1295.
(49) Isborn, C. M.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J. Excited-state electronic structure with configuration interaction singles and Tamm-Dancoff time-dependent density functional theory on graphical processing units. *J. Chem. Theory Comput.* **2011**, *7*, 1814−1823.
(50) Fales, B. S.; Levine, B. G. Nanoscale multireference quantum chemistry: Full configuration interaction on graphical processing units. *J. Chem. Theory Comput.* **2015**, *11*, 4708−4716.
(51) Fales, B. S.; Shu, Y.; Levine, B. G.; Hohenstein, E. G. Complete active space configuration interaction from state-averaged configuration interaction singles natural orbitals: Analytic first derivatives and derivative coupling vectors. *J. Chem. Phys.* **2017**, *147*, 094104.
(52) Hohenstein, E. G.; Luehr, N.; Ufimtsev, I. S.; Martínez, T. J. An atomic orbital-based formulation of the complete active space self-consistent field method on graphical processing units. *J. Chem. Phys.* **2015**, *142*, 224103.
(53) Snyder, J. W., Jr; Fales, B. S.; Hohenstein, E. G.; Levine, B. G.; Martínez, T. J. A direct-compatible formulation of the coupled perturbed complete active space self-consistent field equations on graphical processing units. *J. Chem. Phys.* **2017**, *146*, 174113.
(54) Ufimtsev, I. S.; Martinez, T. J. Quantum chemistry on graphical processing units. 3. Analytical energy gradients, geometry optimization, and first principles molecular dynamics. *J. Chem. Theory Comput.* **2009**, *5*, 2619−2628.
(55) Song, C.; Martínez, T. J. Analytical gradients for tensor hypercontracted MP2 and SOS-MP2 on graphical processing units. *J. Chem. Phys.* **2017**, *147*, 161723.
(56) Kaliman, I. A.; Krylov, A. I. New algorithm for tensor contractions on multi-core CPUs, GPUs, and accelerators enables CCSD and EOM-CCSD calculations with over 1000 basis functions on a single compute node. *J. Comput. Chem.* **2017**, *38*, 842−853.
(57) Anderson, A. G.; Goddard, W. A.; Schröder, P. Quantum Monte Carlo on graphical processing units. *Comput. Phys. Commun.* **2007**, *177*, 298−306.
(58) Esler, K.; Kim, J.; Ceperley, D.; Shulenburger, L. Accelerating quantum Monte Carlo simulations of real materials on GPU clusters. *Comput. Sci. Eng.* **2012**, *14*, 40−51.
(59) Kim, J. QMCPACK: An open source ab initio Quantum Monte Carlo package for the electronic structure of atoms, molecules, and solids. *J. Phys.: Condens. Matter* **2018**, *30*, 195901.
(60) Whitten, J. L. Coulombic potential energy integrals and approximations. *J. Chem. Phys.* **1973**, *58*, 4496−4501.
(61) Trotter, H. F. On the product of semi-groups of operators. *Proc. Am. Math. Soc.* **1959**, *10*, 545−551.
(62) Stratonovich, R. A method for the computation of quantum distribution functions. *Dokl. Akad. Nauk SSSR* **1957**, *115*, 1097−1100.
(63) Hubbard, J. Calculation of Partition Functions. *Phys. Rev. Lett.* **1959**, *3*, 77−78.
(64) Thouless, D. Stability conditions and nuclear rotations in the Hartree-Fock theory. *Nucl. Phys.* **1960**, *21*, 225−232.
(65) Clark, B. K.; Morales, M. A.; McMinis, J.; Kim, J.; Scuseria, G. E. Computing the energy of a water molecule using multi-determinants: A simple, efficient algorithm. *J. Chem. Phys.* **2011**, *135*, 244105.
(66) McDaniel, T.; D'Azevedo, E.; Li, Y.; Wong, K.; Kent, P. Delayed Slater determinant update algorithms for high efficiency quantum Monte Carlo. *J. Chem. Phys.* **2017**, *147*, 174107.
(67) Shi, H.; Zhang, S. Accelerating the use of multi-determinant trial wave functions in auxiliary-field quantum Monte Carlo calculations. 2018, Manuscript in preparation.
(68) Werner, H.-J.; Manby, F. R.; Knowles, P. J. Fast linear scaling second-order Møller-Plesset perturbation theory (MP2) using local and density fitting approximations. *J. Chem. Phys.* **2003**, *118*, 8149−8160.
(69) Hachmann, J.; Cardoen, W.; Chan, G. K.-L. Multireference correlation in long molecules with the quadratic scaling density matrix renormalization group. *J. Chem. Phys.* **2006**, *125*, 144101.
(70) Tsuchimochi, T.; Scuseria, G. E. Strong correlations via constrained-pairing mean-field theory. *J. Chem. Phys.* **2009**, *131*, 121102.
(71) Sinitskiy, A. V.; Greenman, L.; Mazziotti, D. A. Strong correlation in hydrogen chains and lattices using the variational two-electron reduced density matrix method. *J. Chem. Phys.* **2010**, *133*, 014104.
(72) Lin, N.; Marianetti, C.; Millis, A. J.; Reichman, D. R. Dynamical mean-field theory for quantum chemistry. *Phys. Rev. Lett.* **2011**, *106*, 096402.
(73) Stella, L.; Attaccalite, C.; Sorella, S.; Rubio, A. Strong electronic correlation in the hydrogen chain: A variational Monte Carlo study. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2011**, *84*, 245117.
(74) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057−1065.
(75) Sun, Q.; Berkelbach, T. C.; Blunt, N. S.; Booth, G. H.; Guo, S.; Li, Z.; Liu, J.; McClain, J. D.; Sayfutyarova, E. R.; Sharma, S.; Wouters, S.; Chan, G. K. PySCF: the Python based simulations of chemistry framework. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1340.
(76) Flyvbjerg, H.; Petersen, H. G. Error estimates on averages of correlated data. *J. Chem. Phys.* **1989**, *91*, 461−466.
(77) Purvis, G. D., III; Bartlett, R. J. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.* **1982**, *76*, 1910−1918.
(78) Krisiloff, D. B.; Krauter, C. M.; Ricci, F. J.; Carter, E. A. Density fitting and cholesky decomposition of the two-electron integrals in local multireference configuration interaction theory. *J. Chem. Theory Comput.* **2015**, *11*, 5242−5251.
(79) Aquilante, F.; Lindh, R.; Bondo Pedersen, T. Unbiased auxiliary basis sets for accurate two-electron integral approximations. *J. Chem. Phys.* **2007**, *127*, 114107.
(80) Ramabhadran, R. O.; Raghavachari, K. Extrapolation to the gold-standard in quantum chemistry: Computationally efficient and accurate CCSD (T) energies for large molecules using an automated thermochemical hierarchy. *J. Chem. Theory Comput.* **2013**, *9*, 3986−3994.
(81) Pople, J. A. Nobel lecture: Quantum chemical models. *Rev. Mod. Phys.* **1999**, *71*, 1267.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121

(82) Thomas, R. E.; Booth, G. H.; Alavi, A. Accurate Ab initio calculation of ionization potentials of the first-row transition metals with the configuration-interaction quantum Monte Carlo technique. *Phys. Rev. Lett.* **2015**, *114*, 033001.
(83) Balabanov, N. B.; Peterson, K. A. Systematically convergent basis sets for transition metals. I. All-electron correlation consistent basis sets for the 3 d elements Sc-Zn. *J. Chem. Phys.* **2005**, *123*, 064107.
(84) Kutzelnigg, W.; Liu, W. Quasirelativistic theory equivalent to fully relativistic theory. *J. Chem. Phys.* **2005**, *123*, 241102.
(85) Peng, D.; Reiher, M. Exact decoupling of the relativistic Fock operator. *Theor. Chem. Acc.* **2012**, *131*, 1081.
(86) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. Basis-set convergence of correlated calculations on water. *J. Chem. Phys.* **1997**, *106*, 9639−9646.
(87) Bauschlicher, C. W.; Siegbahn, P.; Pettersson, L. G. The atomic states of nickel. *Theor. Chim. Acta* **1988**, *74*, 479−491.
(88) Andersson, K.; Roos, B. O. Excitation energies in the nickel atom studied with the complete active space SCF method and second-order perturbation theory. *Chem. Phys. Lett.* **1992**, *191*, 507−514.
(89) Sun, Q.; Yang, J.; Chan, G. K.-L. A general second order complete active space self-consistent-field solver for large-scale systems. *Chem. Phys. Lett.* **2017**, *683*, 291−299.
(90) Nguyen, H.; Shi, H.; Xu, J.; Zhang, S. CPMC-Lab: A Matlab package for Constrained Path Monte Carlo calculations. *Comput. Phys. Commun.* **2014**, *185*, 3344−3357.
(91) Yanai, T.; Shiozaki, T. Canonical transcorrelated theory with projected Slater-type geminals. *J. Chem. Phys.* **2012**, *136*, 084107.
(92) Nooijen, M.; Bartlett, R. J. Elimination of Coulombic infinities through transformation of the Hamiltonian. *J. Chem. Phys.* **1998**, *109*, 8232−8240.
(93) Sharma, S.; Yanai, T.; Booth, G. H.; Umrigar, C.; Chan, G. K.-L. Spectroscopic accuracy directly from quantum chemistry: Application to ground and excited states of beryllium dimer. *J. Chem. Phys.* **2014**, *140*, 104112.
(94) Knizia, G.; Adler, T. B.; Werner, H.-J. Simplified CCSD (T)-F12 methods: Theory and benchmarks. *J. Chem. Phys.* **2009**, *130*, 054104.
(95) Murphy, R. B.; Beachy, M. D.; Friesner, R. A.; Ringnalda, M. N. Pseudospectral localized Mo/ller-Plesset methods: Theory and calculation of conformational energies. *J. Chem. Phys.* **1995**, *103*, 1481−1490.
(96) 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.
(97) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138*, 034106.
(98) Göddeke, D.; Strzodka, R.; Turek, S. Performance and accuracy of hardware-oriented native-, emulated-and mixed-precision solvers in FEM simulations. *Int. J. Parallel, Emergent Distrib. Syst.* **2007**, *22*, 221−256.
(99) Bykov, D.; Neese, F. Six-electron reduction of nitrite to ammonia by cytochrome c nitrite reductase: insights from density functional theory studies. *Inorg. Chem.* **2015**, *54*, 9303−9316.
(100) Shen, J.; Kolb, M. J.; Göttle, A. J.; Koper, M. T. DFT study on the mechanism of the electrochemical reduction of CO2 catalyzed by cobalt porphyrins. *J. Phys. Chem. C* **2016**, *120*, 15714−15721.
(101) Sharma, S.; Sivalingam, K.; Neese, F.; Chan, G. K.-L. Low-energy spectrum of iron-sulfur clusters directly from many-particle quantum mechanics. *Nat. Chem.* **2014**, *6*, 927.
(102) Askerka, M.; Brudvig, G. W.; Batista, V. S. The O2-evolving complex of photosystem II: recent insights from quantum mechanics/molecular mechanics (QM/MM), extended X-ray absorption fine structure (EXAFS), and femtosecond X-ray crystallography data. *Acc. Chem. Res.* **2017**, *50*, 41−48.
(103) Orenstein, J.; Millis, A. Advances in the physics of high-temperature superconductivity. *Science* **2000**, *288*, 468−474.
(104) Lee, P. A.; Nagaosa, N.; Wen, X.-G. Doping a Mott insulator: Physics of high-temperature superconductivity. *Rev. Mod. Phys.* **2006**, *78*, 17.

DOI: 10.1021/acs.jctc.8b00342
J. Chem. Theory Comput. 2018, 14, 4109−4121