A Systematic Construction of Configuration Interaction Wavefunctions in the Complete CI Space

Andrew W. Prentice,[1] Jeremy P. Coe,[1, a)] and Martin J. Paterson[1, b)]
Institute of Chemical Sciences, School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, United Kingdom.

(Dated: 3 October 2019)

We introduce a *systematic* approach to construct configuration interaction (CI) wavefunctions through a variant of the Monte-Carlo CI (MCCI) method termed *systematic*-MCCI. Within this approach the entire interacting space is systematically considered in batches, with the most important configurations across all batches becoming potential additions to the wavefunction. We compare this method to MCCI and a novel *pruned*-FCI approach. For the ground state of neon, as described by the *cc-p*VTZ basis, we observe no apparent difference between *systematic*-MCCI, *pruned*-FCI and MCCI, with all recovering 99% of the correlation energy and producing a very similar wavefunction composition. We then consider the potential energy surface corresponding to the symmetric double hydrogen dissociation of water within a *cc-p*VDZ basis. Once again MCCI performs comparably to the *systematic* approaches. Despite *systematic*-MCCI having longer run times across the number of processors considered we do observe very good scalability. We then extend this comparison to the first A_1 excited energy of carbon monoxide using the *cc-p*VDZ basis where the MCCI methods perform similarly, approximating this aforementioned energy to within 0.1 eV despite vast reduction in the wavefunction size. Finally we consider the chromium dimer with the *cc-p*VTZ basis and 18 frozen orbitals. Here we find that the systematic approach avoids being trapped in the same local minimum of configuration space as MCCI, yet MCCI can reach a lower energy by repeating the calculation with more processors.

I. INTRODUCTION

The accurate description of electron-electron interactions has long been the crux of electronic-structure methods. In Hartree-Fock (HF) theory this repulsive force is reduced to an averaged one-electron problem, resulting in higher absolute energies and shorter predicted bond lengths, due to the *uncorrelated* motion of electrons. Within a full configuration interaction (FCI) approach, *electron correlation* can be accounted for exactly, albeit within the space spanned by the one-electron basis. Currently, such treatments are computationally intractable for all but the smallest of systems and basis sets, due to the rapidly increasing dimensionality of the Hamiltonian matrix ($\mathbf{H}^{FCI}$) as both the number of basis functions (M) and electrons (n) increase. To the best of our knowledge the largest number of configurations for the wavefunction in a published FCI calculation remains at 10^{10} for the nitrogen molecule[1] while exploratory calculations with a new parallel multi-configurational self-consistent field implementation have demonstrated one CI iteration using almost a trillion determinants.[2]

This limitation, and acknowledged sparseness of $\mathbf{H}^{FCI}$, has led to the development of approximate methods which consider only a portion of the configurational space, significantly reducing the computational costs, and thus allow for the application of these methods to a greater range of chemical problems whilst still reaching a sufficient level of accuracy, despite this vast reduction in the number of variationally optimizable parameters.

Elegant methods built around a single reference, such as coupled-cluster (CC) singles and doubles, have known problems when the single determinant approach breaks down and thus can deviate quite substantially from FCI results. The multi-reference (MR) method of complete active space self consistent field (CASSCF), attempts to recover this missing *static* correlation, from which MRCI or a perturbative approach may be employed to recover the remaining *dynamic* portion. However, CASSCF eventually suffers from similar scaling problems to FCI as the active space increases, and also requires considerable knowledge of both the system of interest and the problem at hand as the determination of an appropriate active space is crucial.

Systematically improvable approximations to the FCI energy have been created by considering the convergence of its many-body expansion. For example, the methods of incremental FCI,[3] and many-body expanded FCI[4,5] where multiple complete active space configuration interaction calculations are performed with increasing numbers of orbitals. This is inherently parallel and coupled with a screening protocol has allowed highly-accurate energies to be calculated without the FCI wavefunction.

To remove the need to diagonalize the Hamiltonian matrix, methods that use projector or diffusion Monte Carlo in configuration space[6,7] to allow improvable estimates of the FCI wavefunction and energy have been developed. Although ultimately still constrained by the scaling of the FCI wavefunction, it is highly parallelizable and requires less memory than FCI. The approach of FCIQMC has, for example, enabled[8] approximate cal-

a) Electronic mail: J.Coe@hw.ac.uk
b) Electronic mail: M.J.Paterson@hw.ac.uk

culations of excitations of butadiene when the FCI space has 10^{29} determinants.

There is renewed interest in the development and use of approaches that iteratively select configurations (selected CI) to build up compact wavefunctions that can describe MR problems sufficiently well. This includes methods based on using perturbation theory and selecting configurations by their coefficient in the perturbed wavefunction or their contribution in the perturbative correction to the energy. An early example of this was configuration interaction using a perturbative selection made iteratively (CIPSI).[9] A modern implementation of CIPSI within the Quantum Package 2.0 program[10] considers the perturbation of the energy to second order. This has been used to provide trial wavefunctions for highly accurate diffusion Monte Carlo calculations including formaldehyde[11] and FeS,[12] for benchmark excited state calculations on small molecules,[13] and has been built upon to create dressed perturbation calculations of cyanine dyes[14] or spin adapted wavefunctions.[15] CIPSI has also led to the creation of the adaptive sampling CI approach[16] that only uses those configurations with sufficiently large coefficients to generate members of the singles and doubles space for the perturbation calculation of the wavefunction to first order. Heat bath CI[17] uses an approximation as a simpler, faster alternative to the full first-order perturbation calculation of coefficients when selecting configurations and has been successfully used to model Cr_2. The approach of MC3I uses diffusion Monte Carlo in configuration space to sample the first-order correction to the wavefunction and has successfully been demonstrated on excited potential curves for C_2.[18]

As an alternative to perturbation theory, the recently developed approach of machine learning configuration interaction[19] uses an artificial neural network that learns on-the-fly to select important configurations in an iterative process. While configurations are chosen using their energy expectation value in the Λ-CI approach.[20] A combination of a perturbation estimate of the energy coupled with the coefficients in the configuration interaction wavefunction is used in the adaptive configuration interaction method.[21] This approach has been further developed[22] to allow the efficient calculation of vertical excitations of methylene, LiF and, when using an active space, polyenes.

Stochastic selection of configurations is used in the method of Monte Carlo configuration interaction (MCCI).[23–25] This iteratively constructs a compact wavefunction by eventually removing those configurations with an absolute coefficient less than a cutoff value (c_{min}). MCCI has been demonstrated to produce wavefunctions that use a very small fraction of the FCI space yet can successfully describe dissociation energies,[26] ground-state potential energy surfaces,[27] electronic excitation energies,[28] and transition metal dimers.[29] This method has been built upon to calculate multipole moments,[30] excited potential energy surfaces including conical intersections,[31] hyperpolarizabilities,[32] dissociation energies when using perturbative corrections,[33] X-ray absorption values,[34] energy levels in molecular tunnel junctions[35] and spin-orbit coupling.[36]

Although the compact wavefunctions resulting from MCCI have been demonstrated to capture properties of the FCI wavefunction with sufficient accuracy and reliability for small systems, whether these compact wavefunctions are near to optimal has not been investigated. This raises questions such as can we find a similarly sized wavefunction with a substantially lower energy or can we find a significantly reduced wavefunction with a similar energy? To systematically and exhaustively check all possible combinations of configurations is not computationally possible for the usual sizes of wavefunctions and configuration spaces encountered. Even if we only consider one iteration of a selected CI calculation and limit ourselves to finding the optimum 20 configurations, from the 1000 configurations that can interact with the existing wavefunction (singles and doubles space), then we still have $\binom{1000}{20} \approx 10^{41}$ combinations and a diagonalization to perform for each one. Yet in actual calculations the singles and doubles space is often orders of magnitude larger.

To approximate a *systematic* approach we have created the computationally tractable method of *systematic*-MCCI which considers all configurations, albeit not in every possible combination, as potential additions to the current wavefunction. *Systematic*-MCCI looks for the best i_{add} configurations to include by randomly ordering the singles and doubles space (N_{sd}) on every iteration then working through this list in batches (i_{batch}) to be added to the current set of configurations. Although this creates a large number of diagonalizations ($\sim N_{sd}/i_{batch}$) to find the overall i_{add} most important configurations on each iteration, this is a huge reduction compared with considering all combinations and furthermore the multiple diagonalizations can be run in parallel. Also as the N_{sd} is randomly ordered within each iteration different combinations will be considered as the computation progresses. Hence *systematic*-MCCI can be viewed as a hybrid approach that guarantees every candidate configuration is tested while avoiding the computational intractability of systematically considering every possible combination to be added on a single iteration.

As a further comparison we also consider pruning the normalized FCI wavefunction to remove all configurations with an absolute coefficient less than c_{min}. The *pruned*-FCI wavefunction is then found by diagonalization using the remaining configurations. Although this can only be applied to small systems and basis sets where FCI is possible, it takes into account all configurations in creating a compact wavefunction and contrasts with the MCCI approaches of iteratively building up the wavefunction.

In this paper we first discuss the computational methods and provide details of the *systematic*-MCCI approach. We then use the neon atom with the *cc-p*VTZ basis as a test case to calibrate the parameters for *system-*

atic-MCCI. This allows us to compare the accuracy and wavefunction size from the *systematic*-MCCI approach and *pruned*-FCI with MCCI for this system. The scalings of calculation time with the number of processors for *systematic*-MCCI and MCCI are also investigated. Next we turn to the potential curve for the double hydrogen dissociation of the water molecule, which includes varying levels of MR character as the bond length changes. This allows us to compare the accuracy of the MCCI potential curve and average size of the wavefunction with the systematic approaches. An investigation of the methods when dealing with an excited state of carbon monoxide in the *cc-p*VDZ basis is then presented. Finally the chromium dimer with the *cc-p*VTZ basis is considered at a particularly challenging geometry.

II. COMPUTATIONAL METHODOLOGY

A FCI wavefunction ($\left|\Psi^{FCI}\right\rangle$) can be represented as a linear combination of all potential n-electron Slater determinants (SDs). It is common to express each SD as an N-tuple excitation relative to the restricted closed-shell HF wavefunction ($\left|\Psi^{HF}\right\rangle$), as outlined below:

$$\left|\Psi^{FCI}\right\rangle = c^{HF}\left|\Psi^{HF}\right\rangle + \sum_{a,r}^{k,M-k} c_a^r \left|\Psi_a^r\right\rangle + \sum_{a<b,r<s}^{k,M-k} c_{ab}^{rs}\left|\Psi_{ab}^{rs}\right\rangle + \dots \quad (1)$$

where a and b run over all k occupied MOs, with r and s pertaining to the $M-k$ virtual MOs. Therefore, within this notation $\left|\Psi_a^r\right\rangle$ would correspond to a single excitation, substituting the a^{th} occupied MO with the r^{th} virtual MO, the subsequent coefficient of this configuration would be c_a^r.

A. Conventional Monte-Carlo Configuration Interaction

Within the context of this study we utilize the MCCI version 4 algorithm,[25] which builds upon a reference wavefunction through an iterative stochastic sampling procedure. It is common to transition from SDs to configuration state functions (CSFs) due to the reduction in optimizable parameters and also ensuring the resulting MCCI wavefunction $\left|\Psi^{MCCI}\right\rangle$ is a pure spin state; however, added complexity is introduced in creating $\mathbf{H}$ and the overlap matrix ($\mathbf{S}$) needs to be calculated as in MCCI different CSFs are not necessarily orthogonal.[24] This non-orthogonal relation between CSFs exists because they are created via a projection method followed by a random walk through the spin path diagram to ensure linear independence.[24] A brief overview of the algorithm follows, in which the superscripts relate to the specific iteration.

The beginning reference wavefunction ($\left|\Psi^{(0)}\right\rangle$) is augmented with a number of random, symmetry preserving, CSFs ($N_{new}^{(1)}$) which is approximately equal to the size of the previous wavefunction, via single and double-excitations, such that the initial modification of the wavefunction $\left|\Psi_{new}^{(1)}\right\rangle$ has the form of equation 2 - this process is referred to as *branching*. In early iterations, to ensure the wavefunction builds at a sufficient rate, N_{new} is varied such that the modified wavefunction contains 100 configurations. The wavefunction is checked for duplicate configurations which are subsequently removed.

$$\left|\Psi_{new}^{(1)}\right\rangle = c^{HF}\left|\Psi^{HF}\right\rangle + \sum_{k=1}^{N_{new}^{(1)}} c_k \left|\Psi_k\right\rangle \quad (2)$$

The next step involves *diagonalization* to determine the expansion coefficient of each CSF within this reduced basis by solving $\mathbf{H}^{(1)}\boldsymbol{c}^{(1)}=\mathbf{S}^{(1)}\boldsymbol{c}^{(1)}\mathcal{E}_{new}^{(1)}$, where $\boldsymbol{c}^{(1)}$ and $\mathcal{E}_{new}^{(1)}$ are the coefficient vector and the energy, respectively. The newly added CSFs are only retained if the relevant coefficient is larger than a user defined cut-off value, $|c_p| \geq c_{min}$, else discarded. Every P_f iterations a full-*prune* is implemented so that all CSFs contained are subject to this selection criterion as per equation 3.

$$\left|\Psi^{(P_f t)}\right\rangle = \underbrace{\left|\Psi^{(P_f t-1)}\right\rangle + \sum_{k=1}^{N_{new}^{(P_f t)}} c_k \left|\Psi_k\right\rangle}_{\text{All subject to } prune} \quad (3)$$

This *pruned* wavefunction ($\left|\Psi^{(1)}\right\rangle$) then forms the reference space for the next iteration, and the process of *branching*, *diagonalization* and *pruning* is repeated until convergence is achieved.

The convergence criteria is implemented on iterations which follow a full-*prune* only (those of the form $P_f t+1$, with $t = 1, 2, 3\dots$), and checks against both the energy ($conv_{\mathcal{E}}$) and number of configurations ($conv_l$). It should be noted that in both MCCI and the following *systematic* approach a full-*prune* iteration is always performed on the initial cycle, however, this point is not included in the convergence check. A moving average over L successive cycles is introduced.

$$\bar{\mathcal{E}}_b = \frac{1}{L}\sum_{t=b+1-L}^{b} \mathcal{E}_{new}^{(i=P_f t+1)} \quad (4)$$

The convergence criteria is fulfilled when the last J differences between $\bar{\mathcal{E}}_b$ and $\bar{\mathcal{E}}_{b-1}$ are lower than the convergence threshold.

$$\max_{i=b+1-J,\,b} |\bar{\mathcal{E}}_i - \bar{\mathcal{E}}_{i-1}| \;\leq conv.thres. \quad (5)$$

For this work we set $L = J = 3$ and for MCCI we run a full-*prune* every 10^{th} iteration ($P_f = 10$) this means that convergence checking begins on iteration 61, i.e., when there is sufficient previous data.

MCCI also has the ability to be performed in parallel across multiple processors (N_{proc}). Within this approach each processor independently performs a *branching*, *diagonalization*, and *pruning* procedure with $\sim N_{new}/N_{proc}$ new CSFs. The set of new configurations stored on each processor after this cycle are then shared amongst all others using MPI so they each contain the same wavefunction for the subsequent iteration, all duplicates that may be present are also removed.

B. Systematic Monte-Carlo Configuration Interaction

A truly *systematic* approach would require consideration of the entire single and double-excitation space, not solely a random assortment of coupled configurations, subjecting all to a pruning procedure and continuing within this regime for at least $n/2$ iterations. As the configurational space expands, and subsequently the potential excitation space, the aforementioned approach would suffer from a similar issue to FCI. A possible alternative to this, as mentioned in the introduction, is by evaluating all the possible ways of adding a certain number of configurations. However this replaces the problem of the *diagonalization* rapidly becoming computationally intractable with the issue of the *number of diagonalizations* quickly becoming computationally intractable. A third way would involve working through the randomly ordered entire excitation space in smaller, computationally viable, batches, storing only the largest weighted configurations when combined with the reference space.

In this work we explore such a way as a modified version of MCCI, referred to herein as *systematic*-MCCI (sMCCI). The algorithm is depicted in Fig. 1. As with MCCI the starting point is $\left|\Psi^{(0)}\right\rangle$, from which *branching* is performed once again ensuring symmetry is preserved, but now generating the entire single and double space (N_{sd}). This is achieved by looping through each configuration in the current set and storing all symmetry allowed single and double substitutions. However, as the same new configuration may be created from different configurations in the current set, whether using SDs or CSFs, we implement the approach of Ref. 37 to remove these duplicates using the quicksort algorithm. The set of singly and doubly excited configurations are then randomly ordered and placed into batches. The number of configurations allocated to each batch is specified by the parameter i_{batch}, generating $N_{sd}^{(1)}/i_{batch}$ batches and possibly one extra batch containing the remainder. Each batch is tested as an addition to the current reference wavefunction, where the coefficients are determined via an analogous procedure to that in MCCI, and the program keeps track of the best i_{add} configurations which are defined as those with the largest absolute coefficients. When all batches have been considered then the best i_{add} configurations are added to the current reference wavefunction, and the coefficients recomputed. As all singles and doubles are being considered then, rather than *pruning* every 10 iterations as in MCCI, all configurations present are subject to the *pruning* criterion after every iteration. However for the convergence check we can still vary P_f. Thus, i_{add} defines an upper limit to the number of configurations that can be added per iteration. This process is continued in an iterative fashion until the convergence check is satisfied.

The bottle-neck of sMCCI will thus be the stage at which approximately N_{sd}/i_{batch} *diagonalisazions* are to be performed sequentially, due to the increasing reference and potential excitation space. However, as this can be performed independently for each batch one can share this workload across multiple processors. Within this regime each processor would receive approximately an equal integer number of batches to work through, with one processor receiving a reduced batch containing any remainder. The parallel algorithm determines a lead processor that generates N_{sd} and subsequently shares the batches to the remaining processors, the lead processor also retains batches to be *diagonalized*. When all batches have been worked through each processor then sends its best i_{add} configurations to the lead processor, which then selects the overall best i_{add} configurations.

This sharing procedure ensures that results are independent of the number of processors, with each processor receiving a similar workload.

In addition to considering all possible interacting configurations the procedure would also eventually consider all possible combinations of i_{batch} configurations if run for long enough and N_{sd} is fixed. To quantify this we estimate how many iterations would be necessary for two configurations to have a high probability of being tested in the same batch. If we assume that i_{batch} exactly divides N_{sd} then we have $B = N_{sd}/i_{batch}$ batches. As the ordering of the batches and the ordering within a batch does not matter then the total number of ways to put the configurations into batches is

$$W_{total} = \frac{N_{sd}!}{B!(i_{batch}!)^B}, \tag{6}$$

while for two determinants always together (e.g. 1 and 2) the combinations are

$$W_{1,2} = \frac{(N_{sd}-2)!}{(B-1)!(i_{batch}!)^{B-1}(i_{batch}-2)!}. \tag{7}$$

The probability of finding the determinants together is then

$$P_{1,2} = \frac{W_{1,2}}{W_{total}} = \frac{i_{batch}(i_{batch}-1)B}{N_{sd}(N_{sd}-1)} = \frac{i_{batch}-1}{N_{sd}-1}. \tag{8}$$

As N_{sd} and i_{batch} are much greater than one and if we want the probability of 1 and 2 never occurring together to be less than 5% over K iterations then we can write

$$\left(1 - \frac{i_{batch}}{N_{sd}}\right)^K \lesssim 0.05. \tag{9}$$

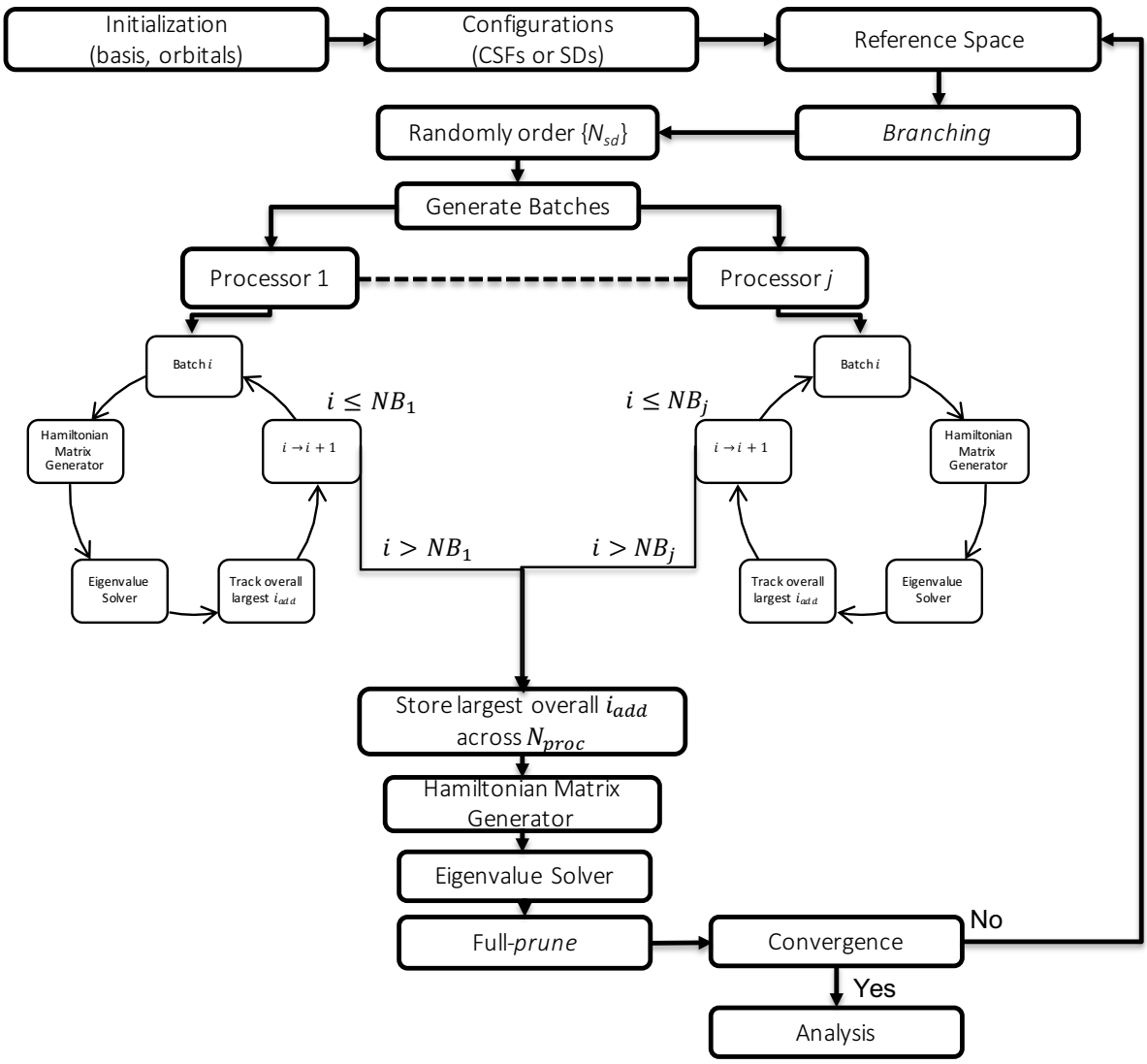

FIG. 1. General overview of sMCCI algorithm running in parallel. NB_1 and NB_j represent the number of batches on processor 1 and j respectively.

If $N_{sd} >> i_{batch}$ then we can use $\ln(1-x) \approx -x$ for small x to give

$$K \gtrsim \frac{3N_{sd}}{i_{batch}}. \tag{10}$$

For example with a batch size of 5000 and 10^5 interacting configurations then if we run for 60 iterations we would expect that any pair of configurations are eventually tested together. The number of iterations unfortunately increases linearly with the interacting space and we will see in the next section that increasing i_{batch} does not necessarily make the calculation more efficient.

C. Pruned FCI

The other systematic approach we consider is to start with the normalized FCI wavefunction, calculated using MOLPRO,[38] then any configurations whose absolute coefficients are less than c_{min} are removed. The *pruned*-FCI wavefunction (p-FCI) is then found by *diagonalization* using the remaining configurations. Although this can only be applied to small systems and basis sets, it offers an alternative systematic approach to building up the wavefunction and allows another way of investigating the optimality of the MCCI wavefunction.

All FCI computations were performed with MOLPRO[38] and truncated-CI computations performed with Gaussian 09 (Revision D.01).[39] At each geometry the HF orbitals were re-optimized, within the one-electron basis, with the one-electron and two-electron integrals obtained from MOLPRO.[38]

III. RESULTS & DISCUSSION

A. Neon

To initially establish a protocol for this systematic approach we consider the neon atom with a *cc-p*VTZ basis [*4s 3p 2d 1f*], implementing the frozen-core approximation for the lowest-energy molecular orbital. Neon is constrained to the D_{2h} point group with the ground state pertaining to A_g symmetry. The FCI energy of this system was found to be -128.802534 Hartree with $\left|\Psi^{FCI}\right\rangle$ containing 7.05×10^7 SDs, resulting in an $\mathcal{E}^{corr}$ of -169.849 kcal mol^{-1}. Truncated-CI calculations were also performed for comparison. In most instances we provide energies in terms of $\Delta\mathcal{E} = \mathcal{E}^{FCI} - \mathcal{E}^{method}$ and, unless stated otherwise, computations were run in parallel across 8 processors.

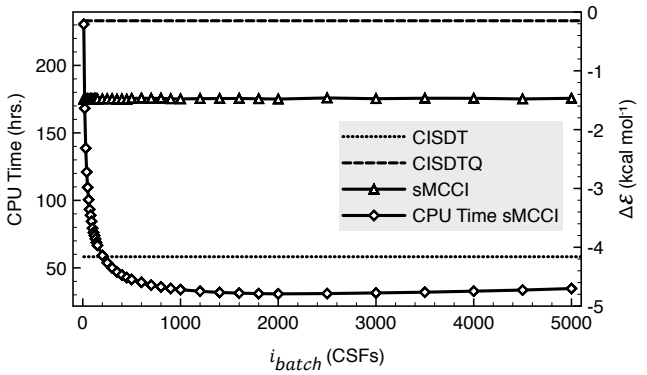

FIG. 2. CPU time (hrs.) and $\Delta\mathcal{E}$ (kcal mol^{-1}) as a function of i_{batch} (CSFs) for neon within a *cc-p*VTZ basis, $c_{min} = 10^{-4}$, $i_{add} = 100$ CSFs and allowed to run for 61 iterations. The methods of CISDT and CISDTQ also included. The CPU Time sMCCI data set is with respect to the x_1 axis with all others pertaining to x_2.

1. i_{batch} and i_{add}

The i_{batch} parameter was altered from $10 \rightarrow 20000$ CSFs, in various increments, whilst limiting the CSFs added per iteration to no more than 100 ($i_{add} = 100$ CSFs). Each computation was allowed to run for 61 iterations, with a c_{min} value of 10^{-4}, thus providing an upper configurational limit of 6101 CSFs on each wavefunction. For this inherently single-reference system the percentage of electron correlation recovered was 99.1 % for $\{i_{batch}\}$, with $\left|\Psi^{sMCCI}\right\rangle$ containing on average ($\bar{x}$) 5658 CSFs. The standard deviation (σ) of the energy and coefficient vector length (l) was 7.25×10^{-3} kcal mol^{-1} and 43 CSFs, respectively, highlighting the negligible dependence on i_{batch}. The sMCCI approach was found to outperform CISDT by 2.685 kcal mol^{-1}, but not CISDTQ, as the sMCCI energy was higher by 1.326 kcal mol^{-1}, however, these methods contain orders of magnitude more configurations. The CPU time on the other hand drastically differs across i_{batch}, initially lowering to a minimum around 2000 CSFs before steadily increasing (see Fig. 2).

We can also estimate the i_{batch} that would result in the fastest calculation by considering the scaling of the Davidson algorithm which is given as $O(kX^2)$ in, e.g., Ref. 40. Here X is the number of determinants and k is the number of Davidson vectors which are limited to a maximum of 30 in this work. If we assume that the implementation of the Davidson algorithm for CSFs in the MCCI code perfectly follows this scaling and any overheads are negligible, then for each iteration of sMCCI there are approximately $B^{(\nu)} = N_{sd}^{(\nu)}/i_{batch}^{(\nu)}$ *diagonalizations* each with $N^{(\nu)} + i_{batch}^{(\nu)}$ configurations, where $N^{(\nu)}$ represents the number of configurations contained within the reference space for iteration ν. The relative time increase upon going from iteration ν to $\nu + 1$ is given by the following:

$$T = \frac{(N^{(\nu+1)} + i_{batch}^{(\nu+1)})^2 B^{(\nu+1)}}{(N^{(\nu)} + i_{batch}^{(\nu)})^2 B^{(\nu)}}. \tag{11}$$

Hence the value of $i_{batch}^{(\nu+1)}$ that minimizes this relative time is found by setting the derivative

$$\frac{\partial T}{\partial i_{batch}^{(\nu+1)}} = \frac{N_{sd}^{(\nu+1)}}{B^{(\nu)}(N^{(\nu)} + i_{batch}^{(\nu)})^2}\left(1 - \frac{(N^{(\nu+1)})^2}{(i_{batch}^{(\nu+1)})^2}\right) \tag{12}$$

to zero. This leads to the general adaptive $i_{batch}^{(\nu+1)} = N^{(\nu+1)}$ as that expected to result in the fastest calculation. We find that this adaptive approach takes 32.3 processor hours so would reside amongst the shorter times shown in Fig. 2 thereby suggesting that this estimate is reasonable. The overall fastest calculation ($i_{batch} = 2000$ CSFs) required slightly less time at 30.7 processor hours and its order of magnitude fits in with the adaptive approach and the average size of the wavefunction given that the converged wavefunction consisted of around 5600 configurations. There is also the issue of the final energy being affected by the batch size but as shown previously in Fig. 2 the difference with FCI is essentially constant on the scale of the graph. As $i_{batch} = 2000$ CSFs gave the quickest time we therefore use this value in subsequent calculations.

Based upon this, we now vary i_{add} from $10 \rightarrow 1000$ CSFs, once again subjecting to 61 full prune iterations with a c_{min} of 10^{-4}. As was to be expected increasing i_{add} lowers the resulting energy more per iteration, with the contrary observed in regards to the CPU time upon completion of iteration 61. However, for this increasing i_{add} we also observe a larger discrepancy between the maximum and computed size of the wavefunction from 61 iterations. The difference for $i_{add} = 100$ CSFs is 447

CSFs, increasing to 53,938 CSFs for $i_{add} = 1000$ CSFs. Hence many more configurations are pruned for larger i_{add}. This was due to convergence being achieved for an $i_{add} \geq 200$ CSFs well before iteration 61 therefore leading to essentially redundant cycles. When examining an i_{add} of 10 CSFs we find that the energy is still lowering, with a difference of 0.131 kcal mol^{-1}, between iteration 60 & 61. This can be attributed to adding no more than 10 CSFs per iteration, therefore only subtly correcting the energy at each stage and requiring more iterations to reach convergence. We observe a similar trend for an i_{add} of 100 CSFs, when reducing the plotted energy range (see inset of Fig. 3). As mentioned previously all other i_{add} values are well converged within 61 iterations, with the CPU time required to reach this decreasing with an increasing i_{add}: 9.47, 14.00 and 29.75 CPU hours for an i_{add} of 1000, 500 and 200 CSFs respectively.

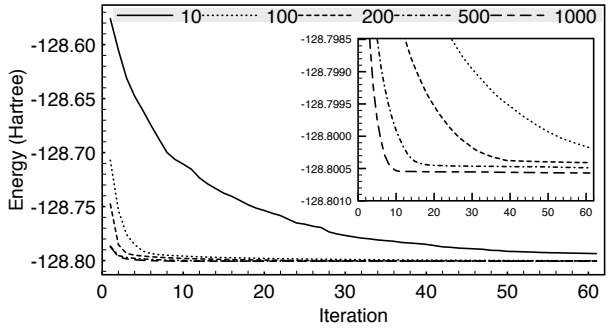

FIG. 3. Energy (Hartree) of each iteration as a function of i_{add} (CSFs) for neon within a *cc-p*VTZ basis, $c_{min} = 10^{-4}$, $i_{batch} = 2000$ CSFs and allowed to run for 61 iterations.

Based on these exploratory calculations we continue with i_{add} and i_{batch} values of 1000 and 2000 CSFs, respectively, implementing the convergence-check procedure outlined above at early stages of the computation. Herein, all values of L and J are held constant at 3 and $P_f = 1$, therefore the check could begin at iteration 7 but this version of sMCCI implements 10 iterations before checking for convergence. The value for $conv_{\mathcal{E}}$ was varied from 10^{-3} to 10^{-4} Hartree with $conv_l$ held fixed at 100 CSFs, this value of $conv_l$ was used throughout. The energy varied by less than 4×10^{-3} kcal mol^{-1}, recovering 99.3 % of $\mathcal{E}^{corr}$, despite the former taking 55% more CPU hours.

2. MCCI vs. sMCCI

Following on from this we compare our *systematic* approach to that of MCCI. The convergence criterion means that there are at least 7 full-*prune* iterations, which occur every 10^{th} iteration for MCCI and every iteration for sMCCI, before beginning the convergence check ensuring a level of consistency between the two methods, with $c_{min} = 10^{-4}$ and $conv_{\mathcal{E}} = 10^{-3}$ Hartree. We briefly explore the statistical nature, $\bar{x}$ and σ, of $\Delta\mathcal{E}$, l and run time (t) for both approaches. We initially explore subsets of 5, 10 and 20 individual runs.

For sMCCI we observe properties (Table I) that are essentially invariant between the various subsets, highlighting the high level of consistency between individual runs. In regards to MCCI we see in Table II that $\bar{x}_{\Delta\mathcal{E}}$ varies between samples to a greater extent in comparison to sMCCI. However these variations are still somewhat insignificant and $\sigma_{\Delta\mathcal{E}}$ was found to vary by only 3.8×10^{-2} and -2.2×10^{-2} kcal mol^{-1} for successive increasing sample sizes. As a result of this we allow a sample size of 20 to represent the population of both sMCCI and MCCI to make for a fair comparison.

TABLE I. Resultant sMCCI properties of varying sample size for neon with a *cc-p*VTZ basis, $i_{batch} = 2000$ CSFs, $i_{add} = 1000$ CSFs and $c_{min} = 10^{-4}$. All energies are in terms of kcal mol^{-1} and l pertains to CSFs.

Sample size	$\bar{x}_{\Delta\mathcal{E}}$	$\sigma_{\Delta\mathcal{E}}$	$\bar{x}_l$	σ_l	$\bar{x}_t$ (sec.)	σ_t (sec.)
5	-1.252	10^{-3}	6897	9	4273.8	18.9
10	-1.252	10^{-3}	6899	7	4281.3	17.8
20	-1.252	10^{-3}	6900	6	4276.8	18.0

TABLE II. Resultant MCCI properties of varying sample size for neon with a *cc-p*VTZ basis, $c_{min} = 10^{-4}$. All energies are in terms of kcal mol^{-1} and l pertains to CSFs.

Sample size	$\bar{x}_{\Delta\mathcal{E}}$	$\sigma_{\Delta\mathcal{E}}$	$\bar{x}_l$	σ_l	$\bar{x}_t$ (sec.)	σ_t (sec.)
5	-1.697	1.98×10^{-1}	6914	53	486.4	37.0
10	-1.719	2.36×10^{-1}	6926	57	497.7	39.6
20	-1.711	2.14×10^{-1}	6932	61	500.5	48.8

For the populated data despite containing 32 fewer CSFs sMCCI captures 4.59×10^{-1} kcal mol^{-1} more of $\mathcal{E}^{corr}$, indicating a slightly more compactness within the resulting wavefunction. This can be attributed to sMCCI considering the entire single and double space at each iteration, therefore it is more likely that all the important configurations will be located. As would be expected a systematic approach has a greater degree of consistency between individual runs, highlighted by a lower σ_l and $\sigma_{\Delta\mathcal{E}}$. This should not detract from the $\sigma_{\Delta\mathcal{E}}$ of 2.14×10^{-1} kcal mol^{-1} for MCCI as this is still a highly consistent spread of values for an inherently stochastic technique. The full consideration of N_{sd}, and minimal energetic decrease, comes at the expense of wall time as sMCCI takes 7.5 times longer than MCCI (cf. 4276 sec. and 500.5 respectively).

3. Scaling of sMCCI & MCCI

The scalability of the sMCCI and MCCI algorithm was then explored (see Fig. 4), ranging from 8 to 72 processors. In regards to our computational resources each node

contains 12 processors therefore we must utilize a multi-node algorithm for $N_{proc} > 12$, which involves additional communication of individual nodes with the head node, upon completion of each cycle, subsequently affecting the observed timings. This difference was tested on 12 processors and found to increase computational time by 4 seconds therefore we do not expect this to be an issue. As the relative timings of sMCCI are far larger than σ_t for 8 processors we compare a single run of each. However this is not the case for MCCI as the separation of successive wall times is well within the aforementioned σ_t, therefore we compare the results over 20 runs.

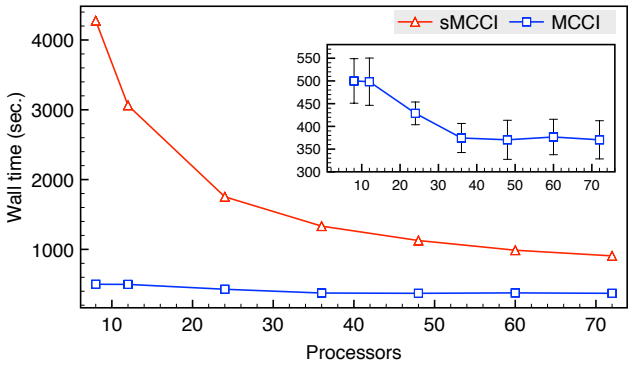

FIG. 4. Wall time (sec.) as a function of N_{proc} for sMCCI and MCCI computations of neon within a *cc*-*p*VTZ basis, $c_{min} = 10^{-4}$, $i_{batch} = 2000$ CSFs, $i_{add} = 1000$ CSFs and $conv_{\mathcal{E}} = 10^{-3}$ Hartree. Inset: an enlarged view of the y_1 axis between $300 \rightarrow 590$ seconds. The error bar relates to $\pm\sigma_t$.

Increasing the number of processors from 8 to 72 results in a wall time decrease of 78.8 and 26.0 % for sMCCI and MCCI respectively. For sMCCI we observe very promising scaling factors upon changing from 8 to 36 processors, with the total wall time reducing by 68.8 %. For MCCI we do not observe this level of scaling with no benefit for this system of going above 36 processors. As a result of the small run time difference when increasing the number of processors, as shown in Fig. 4, and the stochastic nature of the method it is entirely possible for slightly fewer processors to result in faster computations.

We note that the sMCCI energy is independent of the number of processors while the energy of MCCI may vary. This is because the equivalent of the batch size in MCCI depends on the number of processors. Of course both methods can give slightly different results from run to run, as discussed previously, unless the seed is fixed for the random number generator. We did not see that sMCCI could run in less time than MCCI when considering up to 72 processors, but the scaling suggested that this may be possible with a sufficiently large number of processors. However the removal of duplicates in N_{sd} may become the bottle-neck that prevents this.

4. SDs vs CSFs

To compare the methods of sMCCI and MCCI with that of *pruned*-FCI where the FCI wavefunction is calculated using MOLPRO[38] we must transition to SDs, as up until this point we have solely considered wavefunctions consisting of CSFs. When using SDs the resulting sMCCI energy is 4.64×10^{-1} kcal mol^{-1} larger than the CSF result, with the converged $\left|\Psi^{sMCCI}\right\rangle$ containing 2811 more configurations - all other computational parameters were kept the same as III A 2 and again averaged over 20 runs. The total wall time of the aforementioned computations also increased by 76.2 % upon this transition. In regards to MCCI, we observe a similar outcome as the energy from a SD approach is larger by 2.21×10^{-1} kcal mol^{-1} and contains 2338 more configurations. However the total wall time was found to decrease by 20 % when using SDs. For SDs the energy difference between sMCCI and MCCI is lowered to 2.15×10^{-1} kcal mol^{-1}, with the former now containing 441 more configurations (see Table III).

TABLE III. Resultant MCCI properties for a sample size of 20 for neon with a *cc*-*p*VTZ basis, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-4}$. All energies are in terms of kcal mol^{-1} and l pertains to SDs.

Method	$\bar{x}_{\Delta\mathcal{E}}$	$\sigma_{\Delta\mathcal{E}}$	$\bar{x}_l$	σ_l	$\bar{x}_t$ (sec.)	σ_t (sec.)
sMCCI	-1.716	10^{-4}	9711	2	7531.1	29
MCCI	-1.931	8.5×10^{-2}	9270	46	400.7	23.9

5. Comparison with *pruned*-FCI

For a $c_{min} = 10^{-4}$ the *pruned*-FCI wavefunction contained 9679 SDs with a resulting $\Delta\mathcal{E}$ of -1.729 kcal mol^{-1}. This is very similar to the sMCCI and MCCI result. Hence for the neon atom efficiently building up a compact wavefunction with random selections is suggested to be comparable with the *systematic* but computationally expensive approach of beginning with the FCI wavefunction and *pruning* configurations. One approach to broadly compare the wavefunctions is by quantifying their multi-reference character. We use MR_{char} (see Refs. 29 and 41) which is defined as

$$MR_{char} = \sum_i |c_i|^2 - |c_i|^4 \tag{13}$$

where c_i are the normalized coefficients of the SDs in the wavefunction. MR_{char} ranges from zero, for a wavefunction consisting of a single determinant, to one as the number of important configurations increases. For this system we find that, in addition to the energies being very similar, the wavefunction's multi-reference character was 6.7×10^{-2} for all three methods.

The composition of each wavefunction was then explored in terms of the number and relative contributions of each substitution level. For sMCCI and MCCI we take one single run and not an average as was done previously. As shown in Fig. 5 the wavefunctions are dominated by double, triple and quadruple excited configurations. However, when looking at the importance of each substitution level it is essentially dominated by $\left|\Psi^{HF}\right\rangle$ with some contribution from double-excited configurations (see Fig. 6). The only major discrepancy between the wavefunctions is with respect to quadruple excited configurations as MCCI contained 423 and 407 fewer when compared with sMCCI and *pruned*-FCI, respectively.

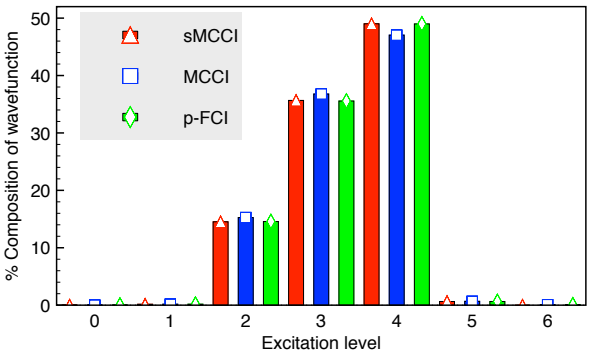

FIG. 5. The % composition of each wavefunction in terms of configurations from each substitution level for neon with a *cc-p*VTZ, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-4}$.

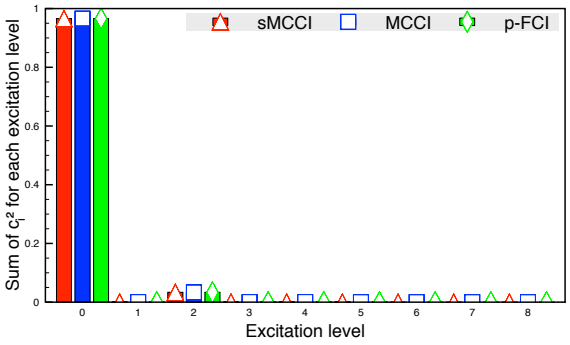

FIG. 6. The sum of c_i^2 for each substitution level present in the various wavefunctions for neon with a *cc-p*VTZ, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-4}$.

We also increased the c_{min} value by an order of magnitude to 10^{-3}. The *pruned*-FCI wavefunction resulted in an energy of -128.790810 Hartree with a vector length of 1137 SDs. The sMCCI and MCCI energies are both higher by 1.971×10^{-1} and 5.806×10^{-1} kcal mol^{-1}, once again the sMCCI wavefunction contained fewer configurations. Despite increasing the c_{min} to 10^{-3}, we still capture between 95.3 - 95.7 % of $\mathcal{E}^{corr}$, with a MR_{char} of 6.2×10^{-2}. For this c_{min} no method contains quintuply excited configurations and above. Once again the composition of each wavefunction is very similar across all methods (see Fig 7 and 8).

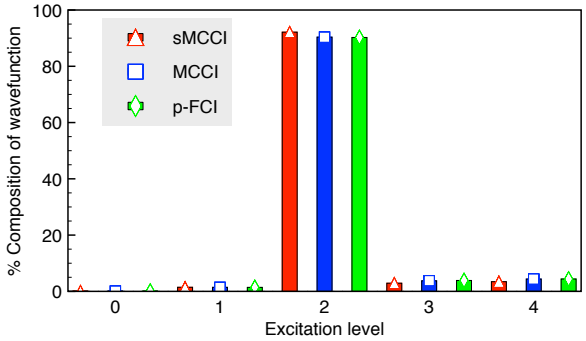

FIG. 7. The % composition of each wavefunction in terms of configurations from each substitution level for neon with a *cc-p*VTZ, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-3}$.

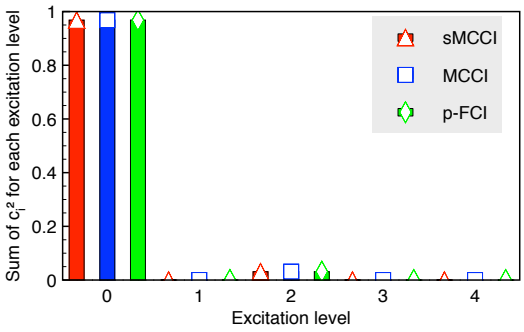

FIG. 8. The sum of c_i^2 for each substitution level present in the various wavefunctions for neon with a *cc-p*VTZ, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-3}$.

B. H_20

We now turn our attention to the double-hydrogen dissociation of water with a *cc-p*VDZ [*7s 4p 1d*] basis, once again invoking the frozen-core approximation for the lowest-energy MO. The distance between the oxygen and hydrogen atom (R_{OH}) was varied symmetrically from 1.0 to 4.6 Bohr, in 0.2 Bohr increments, with the H-O-H angle held constant at 104.5° therefore maintaining c_{2v} symmetry along the potential energy surface (PES). A previous study[37] provides FCI energies which we will use for comparison.

1. i_{batch} and i_{add} parameters

Initially the same protocol as section III A of varying i_{batch} and i_{add} was explored for an R_{OH} of 4.0 Bohr,

allowing us to investigate this dependence on a system expected to have a great deal more multi-reference character than neon. The FCI energy for this separation is -75.932598 Hartree with $\left|\Psi^{FCI}\right\rangle$ containing 1.96×10^7 SDs, giving an $\mathcal{E}^{corr}$ of -250.841 kcal mol^{-1}. For i_{batch} we once again observe no energy dependence for the range of values considered ($0 \rightarrow 2000$ CSFs). Fig. 9 displays the effect of varying i_{add} for sMCCI when $c_{min} = 10^{-4}$. The results are somewhat analogous to the neon system therefore we once again utilize an i_{batch} and i_{add} of 2000 and 1000 CSFs, with the convergence check beginning after iteration 10. When examining $conv_{\mathcal{E}}$ values of 10^{-3} and 10^{-4} Hartree we find an energy difference of 3.89×10^{-1} kcal mol^{-1}, with the latter having a CPU time 5.1 times that of the former. The respective wavefunctions for the convergence thresholds contained 10301 and 14901 CSFs.

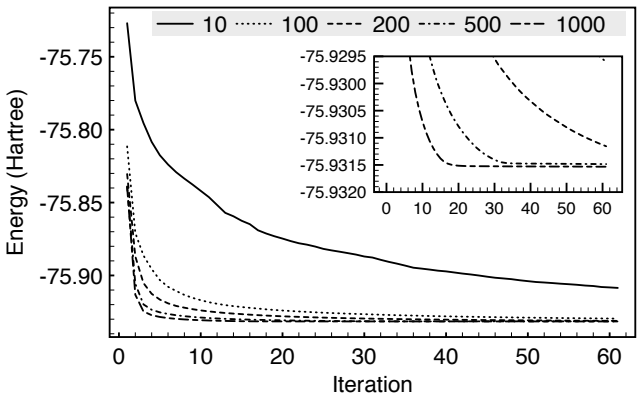

FIG. 9. Energy (Hartree) of each iteration as a function of i_{add} (CSFs) for H_2O ($R_{OH} = 4.0$ Bohr) within a *cc-p*VDZ basis, $c_{min} = 10^{-4}$, $i_{batch} = 2000$ CSFs and allowed to run for 61 iterations.

2. sMCCI vs. MCCI

For the 4.0 Bohr R_{OH} system we also explored the statistical nature of the MCCI and sMCCI results over 20 individual runs with $i_{add} = 1000$ CSFs, $i_{batch} = 2000$ CSFs, $conv_{\mathcal{E}} = 10^{-3}$ Hartree and $c_{min} = 10^{-4}$. As was the case with neon, sMCCI shows little variation between various subsets. In regards to MCCI the difference between $\sigma_{\Delta\mathcal{E}}$ was 1.2×10^{-2} kcal mol^{-1} for the samples containing 10 and 20 subsets , this value was lower than its neon counterpart, despite containing 8789 more configurations. Therefore we allow the subset of 20 computations to represent the population statistics.

Once again, sMCCI shows a greater degree of consistency between individual runs when compared to MCCI ($\sigma_{\Delta\mathcal{E}}$ of 10^{-3} and 1.9×10^{-2} kcal mol^{-1}, respectively). However, the resulting MCCI energy was found to be lower by 3.951×10^{-1} kcal mol^{-1}, containing 5409 more CSFs. The discrepancy in configurations can be attributed to the varying N_{new} added configurations in MCCI, and also starting the convergence procedure after at least 7 full prune iterations, equating to the 61^{st} cycle in MCCI. As the reference space increases so does N_{new}, therefore the wavefunction can increase at a far superior rate for MCCI at large values, contrary to sMCCI which has an upper limit of adding 1000 configurations per iteration. Upon lowering $conv_{\mathcal{E}}$ to 10^{-4} the MCCI energy varies by only 2.97×10^{-2} kcal mol^{-1}, far lower than the aforementioned sMCCI difference. At a $conv_{\mathcal{E}}$ of 10^{-4} both MCCI and sMCCI capture 99.7 % of the electron correlation energy, with the latter containing 1013 fewer CSFs.

3. Scaling of sMCCI and MCCI

We compare computations employing an energy-convergence threshold of 10^{-4} Hartree, due to the significant discrepancy between the number of CSFs in the resulting MCCI wavefunctions for a threshold of 10^{-3} Hartree. We still observe a difference of 1013 CSFs for the lower convergence threshold, with the $\left|\Psi^{sMCCI}\right\rangle$ once again the more compact of the two, which must be remembered when comparing wall times of the MCCI methods. We compare a single run of increasing processors. Upon increasing the number of processors, once again from 8 to 72, we find wall times that decrease by 78.2 and 18.7 % for sMCCI and MCCI respectively.

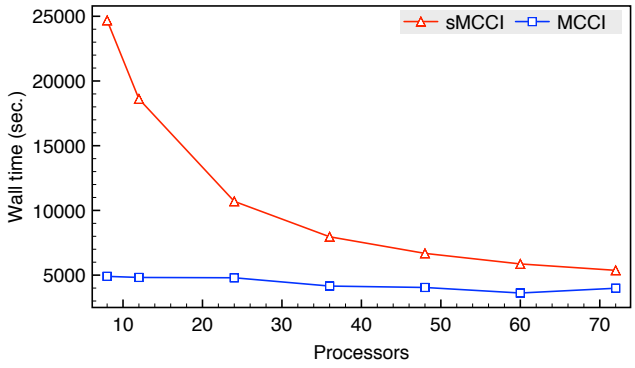

FIG. 10. Wall time (sec.) as a function of N_{proc} for sMCCI and MCCI computations of H_2O ($R_{OH} = 4.0$ Bohr) within a *cc-p*VDZ basis, $c_{min} = 10^{-4}$, $i_{batch} = 2000$ CSFs, $i_{add} = 1000$ CSFs and $conv_{\mathcal{E}} = 10^{-4}$ Hartree.

When running on 8 processors the sMCCI computation is 403 % longer than the MCCI counterpart (24680 and 4905 seconds, respectively). If we then compare the wall time differential when utilizing 72 processors, this significantly lowers to 35 % (5370 and 3987 seconds, respectively). Once again for MCCI we do not observe a smooth convergence as the number of processors are increased.

4. SDs vs CSFs

Once again we must transition from CSFs to SDs. We compare the results of a single computation at $conv_{\mathcal{E}}$ values of 10^{-3} and 10^{-4} Hartree, with a $c_{min} = 10^{-4}$. All resulting energies from wavefunctions containing CSFs are found to be lower in energy than their SD counterparts. For the computations using SDs we observe the same general trend, as was found for CSFs in Section III B 1 and III B 2, upon differing $conv_{\mathcal{E}}$. However, the energy difference between sMCCI at these varying convergence values was larger at 1.317 kcal mol^{-1}. At a $conv_{\mathcal{E}} = 10^{-4}$ Hartree sMCCI and MCCI have a $\Delta\mathcal{E}$ of -1.070 and -1.089 kcal mol^{-1} respectively.

5. Potential energy surface

We now calculate the PES as the bonds are symmetrically stretched using a c_{min} and $conv_{\mathcal{E}}$ of 10^{-3} for the method comparison.

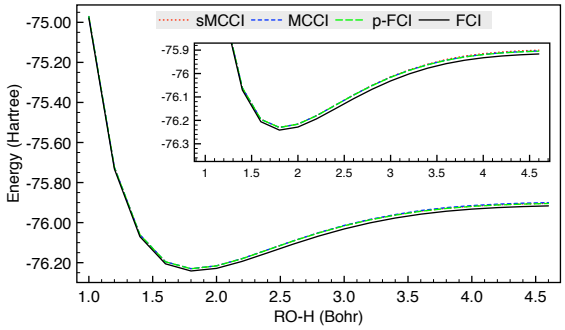

FIG. 11. The PES for the double-hydrogen dissociation of H_2O using the *cc-p*VDZ basis, one frozen orbital and 104.5° for the H-O-H angle when calculated with FCI, *pruned*-FCI, MCCI and sMCCI.

As shown in Fig. 11 all methods are found to be in good agreement with the FCI curve, with the % of $\mathcal{E}^{corr}$ recovered ranging between 94 - 98 %. In order to provide a quantitative account of the differences in accuracy between individual curves we compare the nonparallelity error (NPE) and $\sigma_{\Delta\mathcal{E}}$. The NPE is defined as[42] the difference between the largest and smallest absolute error for the calculated curve.

$$NPE = \max|\Delta\mathcal{E}| - \min|\Delta\mathcal{E}|. \tag{14}$$

The standard deviation of the energy difference with FCI, $\sigma_{\Delta\mathcal{E}}$, was introduced in Ref. 37 as a way of quantifying the error in a potential energy curve that incorporates the feature of the NPE that the curves may be shifted by a constant, but also takes into account all of the points used for the calculated PES. It is defined as

$$\sigma_{\Delta\mathcal{E}} = \sqrt{\frac{1}{d}\sum_{j=1}^{d}(\Delta\mathcal{E}_j - \overline{\Delta\mathcal{E}})^2}, \tag{15}$$

where there are d points in the potential curve and $\overline{\Delta\mathcal{E}}$ is the mean value for $\Delta\mathcal{E}$. Previously in this paper we used $\sigma_{\Delta\mathcal{E}}$ for multiple runs at a single geometry to indicate the consistency of the methods, but emphasize that now it is being used for single runs at a range of geometries to quantify the accuracy of the potential curve.

TABLE IV. NPE and $\sigma_{\Delta\mathcal{E}}$ for the various computational methods ($c_{\min} = 10^{-3}$), all points of the PES were included.

Method	NPE (kcal mol^{-1})	$\sigma_{\Delta\mathcal{E}}$ (kcal mol^{-1})	$\bar{x}_{\%\mathcal{E}^{corr}}$
sMCCI	5.262	1.607	95.32
MCCI	6.479	2.156	94.87
pruned FCI	4.971	1.501	95.46

In Table IV we see that the NPE is 1.217 kcal mol^{-1} lower for sMCCI when compared to MCCI with $\Delta\mathcal{E}$s also closer to one another by on average 5.49×10^{-1} kcal mol^{-1}. The most accurate PES is that approximated by a *pruned*-FCI approach as the lowest NPE and $\sigma_{\Delta\mathcal{E}}$ values are obtained. The *pruned*-FCI and sMCCI methods capture a similar % of $\mathcal{E}^{corr}$ over the PES.

For the three methods, at the equilibrium geometry the system is single reference with a MR character of 9.7×10^{-2} increasing to 0.72 for a separation of 4.0 Bohr. For these two geometries we also explore the composition of each converged wavefunction from the individual methods, see Fig. 12 and 13. For an R_{OH} of 1.8 Bohr, doubly-excited configurations consist of 89.8 - 92 % of the wavefunction, with the remaining percent essentially quadruples. However, as was the case with neon, in terms of importance we find that only the HF and double-excited configurations have any real contribution.

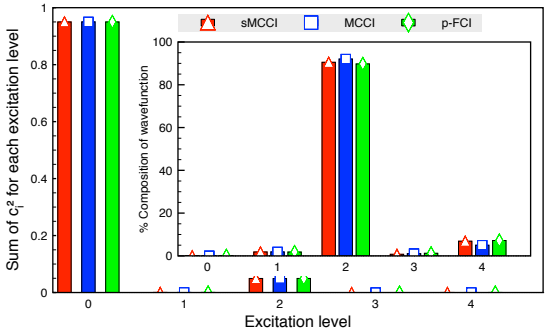

FIG. 12. The sum of c_i^2 for each substitution level present in the various wavefunctions for H_2O with an $R_{OH} = 1.8$ Bohr and *cc-p*VDZ basis, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-3}$. Inset: The % composition of each wavefunction in terms of configurations.

If we consider an $R_{OH} = 4.0$ Bohr, we observe the effect of the drastically increased MR character. For this separation, in terms of configuration amount, the wavefunctions consist of up to sextuple excited determinants. However, now the contribution of the HF wavefunction has been drastically lowered, from 95 % for the equilibrium geometry, to 50 %. For this increased separation

the double excited configurations are extremely important with a contribution of around 40 %. The singly, triply and quintuply excited configurations also have non-negligible contributions.

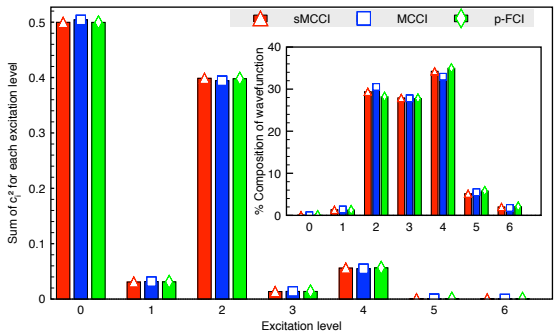

FIG. 13. The sum of c_i^2 for each substitution level present in the various wavefunctions for H_2O with an $R_{OH} = 4.0$ Bohr and *cc-p*VDZ basis, $i_{batch} = 2000$ SDs, $i_{add} = 1000$ SDs and $c_{min} = 10^{-3}$. Inset: The % composition of each wavefunction in terms of configurations.

C. Carbon Monoxide

We now consider the first excited state of A_1 symmetry for carbon monoxide ($^1\Sigma^+$), within c_{2v} symmetry, with a *cc-p*VDZ basis [*6s 4p 2d*] and the two lowest energy MOs doubly occupied in all configurations. The experimental bond length for the excited state was used (2.116 Bohr). At this geometry, the FCI energy of the ground and excited A_1 state was -113.055014 and -112.666416 Hartree respectively, giving an excitation energy of 10.574 eV. For the A_1 excited state we ensure that the wavefunction always contains all configurations from the converged MCCI A_1 ground state wavefunctions to prevent instabilities in the excited state calculation arising from the *pruning* of important ground-state configurations. For a c_{min} of 10^{-3}, $conv_{\mathcal{E}} = 10^{-3}$ Hartree, $i_{batch} = 2000$ CSFs and $i_{add} = 1000$ CSFs, the ground state MCCI and sMCCI wavefunctions contained 1770 and 1919 CSFs with the latter 6.45×10^{-1} kcal mol^{-1} lower in energy. For the excited state both wavefunctions increase in size, by a factor of 1.6 for MCCI and 1.5 for sMCCI, with sMCCI again the lower of the two by 4.06×10^{-1} kcal mol^{-1}. Both sMCCI and MCCI provide good approximations to the exact excitation energy of 10.574 eV, only slightly overestimating this quantity by less than one tenth of an eV. The respective absolute excitation value of the sMCCI and MCCI approaches are 10.658 and 10.668 eV. We then employed the approach of recomputing the *corrected* ground state energy within the space spanned by the excited state wavefunction. Within this approach the MCCI and sMCCI gave excitation energies of 10.745 and 10.744 eV respectively, due to the lowering of the A_1 ground state energy via inclusion of the important excited CSFs. In order to investigate the MR character we employ SDs, with all other parameters analogous to that above. The MCCI and sMCCI wavefunctions contained 3104 and 3168 SDs respectively, with the latter 4.32×10^{-1} kcal mol^{-1} lower in energy. Once again in regards to the A_1 excited state both wavefunctions increase in size by a factor of 1.5, with sMCCI again the lower of the two by 2.03×10^{-1} kcal mol^{-1}. We do not observe the same agreement as that of the CSF variant with the excitation energy underestimated by > 1.6 eV. In regards to the MR character the excited state is far larger than that of the ground state with values of 0.17 and 0.80 respectively. Upon *correction* of the ground state energy this was found to slightly improve the excitation energies with values of 9.039 and 9.045 eV for MCCI and sMCCI, respectively. Upon lowering the c_{min} to 10^{-4} the approximated excitation energies lower by 1.73×10^{-1} and 1.77×10^{-1} eV for MCCI and sMCCI respectively, despite vastly increased wavefunctions. Applying the aforementioned correction results in excitation energies of 8.805 and 8.831 eV for this lower c_{min}.

D. Chromium dimer

The potential energy curve of the chromium dimer was investigated with MCCI in Ref. 29 using the *cc-p*VDZ and *cc-p*VTZ basis sets with 18 frozen orbitals. For the *cc-p*VTZ results, a cutoff of $c_{min} = 2 \times 10^{-4}$ was used. However for a larger cutoff and a stretched bond length of 2.75 Å we have found that the *cc-p*VTZ result can get trapped in a local minimum of configuration space. We attribute this to the large configuration space for *cc-p*VTZ ($\sim 10^{18}$ determinants) and the very strongly multi-reference character that was demonstrated for this system.[29]

This calculation is therefore very appropriate as a test of *systematic*-MCCI where the consideration of all single and double substitutions should reduce the chance of being trapped in a local minimum of configuration space. We use CSFs, $c_{min} = 5 \times 10^{-4}$ (with the same value for the convergence threshold) and initially implement the calculations on 12 processors with $i_{\text{add}} = 100$ CSFs. We see in Fig. 14 that the *systematic* approach reaches a lower energy than MCCI in less time. The *systematic* approach did not run until convergence due to the number of singles and doubles surpassing the maximum allocated for this calculation (200 million). This occurred after 28 iterations compared with the 1044 iterations for MCCI to reach convergence. The final MCCI wavefunction used 34629 configurations while that of sMCCI needed only 2588. This demonstrates how the *systematic* approach can produce wavefunctions that are more accurate and more compact in less time. For balance we emphasize that it is possible that multiple runs of MCCI would yield lower energies however such an approach is less feasible for this system due to the long computation time.

We next increase the number of processors to 144 for sMCCI and 36 for standard MCCI. We implement these

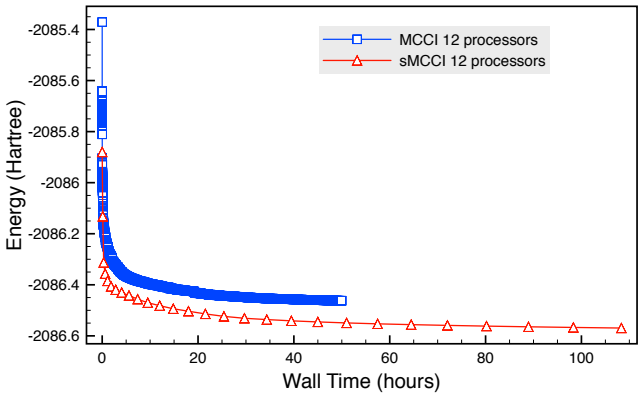

FIG. 14. Energy (Hartree) against time (hours) for MCCI and *systematic*-MCCI 12 processor calculations of the chromium dimer at a bond length of 2.75 Å using CSFs, the *cc-p*VTZ basis with 18 frozen orbitals and a cutoff of $c_{min} = 5 \times 10^{-4}$.

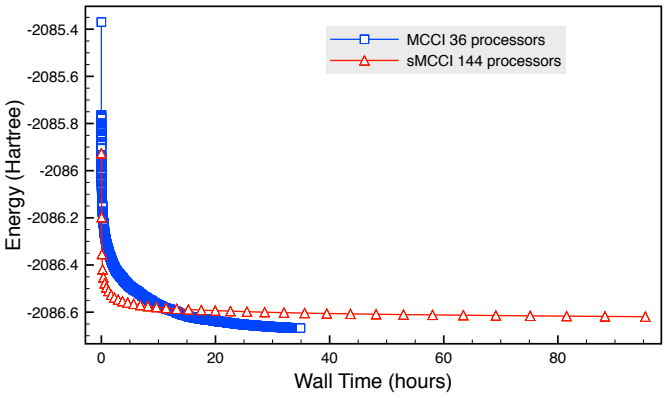

FIG. 15. Energy (Hartree) against time (hours) for MCCI and *systematic*-MCCI calculations using 36 and 144 processors respectively when applied to the chromium dimer at a bond length of 2.75 Å using CSFs, the *cc-p*VTZ basis with 18 frozen orbitals and a cutoff of $c_{min} = 5 \times 10^{-4}$.

larger parallel calculations on one of the EPSRC Tier-2 National HPC facilities (Cirrus) and we now set i_{add} to 200 CSFs with the aim of getting more improvement in the sMCCI energy per iteration. This increase in i_{add} means that by iteration 28 the energy is now -2086.606 Hartree compared with -2086.569 Hartree for $i_{\text{add}} = 100$ CSFs. Due to allocating more space for the singles and doubles then the calculation continues until it runs out of time on iteration 38 when there are more than half a billion singles and doubles configurations. However as there are twelve times as many processors then this calculation is faster than the previous $i_{\text{add}} = 100$ CSFs result. Fig. 15 shows that despite the larger i_{add} lowering the final energy for the *systematic* calculation and the MCCI energy being higher in the early stages, the converged MCCI energy is now lower than the *systematic* result and takes less time even when using one quarter of the processors. The systematic result now only used 7181 configurations while MCCI, similarly to 12 processors, needed 1113 iterations and 33507 configurations.

Due to the size of the singles and doubles space growing so fast then large enough numbers of processors would eventually confer an advantage to *systematic*-MCCI. However for the parallel calculations considered here the lowest energy was due to MCCI for a run on 36 processors confirming again that the stochastic selection of configurations in MCCI can provide a high quality result. Yet the highest energy also occurred for the result of an MCCI calculation on 12 processors which was noticeably improved upon in a relatively short time and with significantly fewer configurations when using the *systematic*-MCCI approach for this system. Coupled with the more predictable behavior of *systematic*-MCCI this suggests that the current incarnation of the approach can be used to indicate if a single MCCI result is reliable or if further, and more costly, MCCI calculations are necessary.

IV. CONCLUSION

We introduced the approach of *systematic*-Monte Carlo configuration interaction (sMCCI) where all interacting configurations are randomly ordered then tested in batches as additions to the current wavefunction, albeit not in every possible combination, in iterated *diagonalizations* to build up a compact wavefunction. In contrast to this selected CI scheme, we looked at removing configurations from the full configuration interaction (FCI) wavefunction if their absolute coefficient is less than a cutoff to give a *pruned*-FCI approach. These systematic approaches allowed us to investigate whether the stochastically generated MCCI wavefunction appears close to optimum for a range of systems.

The neon atom in the *cc-p*VTZ basis was first used to calibrate sMCCI. There it was found that a batch size of 2000 and adding 1000 configurations on each iteration was most efficient. The order of magnitude of the batch size was in line with our estimate based on the scaling of the *diagonalizations*. We found that when using the systematic approaches the energy, wavefunction size and multi-reference character were similar to MCCI. The form of the approximate wavefunctions were also comparable which we analyzed by looking at the weighted excitation level of the configurations. sMCCI exhibited much better scaling with the number of processors although it was not demonstrated to be faster than MCCI when employing up to 72 processors.

We next considered the water molecule in a *cc-p*VDZ basis and calculated its potential energy curve when both bonds are symmetrically stretched. The FCI curve was well approximated by all the methods and when quantifying the accuracy *pruned*-FCI had the lowest error followed by sMCCI but there was not a strong difference between the approximate methods. For the equilibrium geometry and a stretched bond length we found all

three approaches gave very similar wavefunctions, containing $\sim$ 1600 and $\sim$ 2700 configurations respectively. The MR_{char} of these geometries were also found to be 9.7×10^{-2} and 0.72.

For the first A_1 excitation energy of carbon monoxide in a *cc-p*VDZ basis, we then found that the energy was similar whether sMCCI or MCCI was used and in good agreement to the exact excitation energy.

Finally we looked at a particularly challenging geometry of the already difficult electronic structure problem of the chromium dimer. With the *cc-p*VTZ basis and 18 frozen orbitals, we found when using a reasonable cutoff that the systematic approach could result in a noticeably lower energy in less time than MCCI despite using less than a tenth of the configurations. This used 12 processors, but when we increased the number of processors to 36 for MCCI then the MCCI result was lower in energy than the systematic value.

We have seen that by stochastically building up a compact wavefunction, MCCI generally produces energies and wavefunctions that are similar to the two systematic approaches created here to select configurations. This suggests that, for most of the systems considered here (covering equilibrium geometries, stretched bonds and excited states) the MCCI wavefunction was likely to be reasonably close to the optimum wavefunction for the number of configurations used. The exception to this was the chromium dimer where *systematic*-MCCI avoided being trapped in the same local minimum of configuration space as MCCI when running on 12 processors. However another run of MCCI with more processors reached a lower energy in less time than sMCCI.

Although *systematic*-MCCI shows better parallel scaling than MCCI it was not demonstrated to be faster overall at this stage in its development. Hence we are not suggesting that the systematic approach supplants the stochastic method, rather that it can be used to demonstrate the likely optimality of the MCCI wavefunction.

Furthermore, as have we seen for the chromium dimer, it can also be used to indicate relatively quickly if MCCI has become trapped in a local minimum of configuration space and therefore repeated, more costly, MCCI calculations are necessary to create a wavefunction of sufficient accuracy. In the future we wish to explore the large scale parallelization of sMCCI, due to the degree of scalability observed in the aforementioned neon and water computations, and also alternative approaches to generating the entire single and doubles space that currently must be stored in available memory at each iteration. This would allow larger systems and configurational spaces to be explored.

ACKNOWLEDGMENTS

We thank the EPSRC for funding through the platform grant EP/P001459/1.

V. REFERENCES

[1] E. Rossi, G. L. Bendazzoli, S. Evangelisti, and D. Maynau, "A full-configuration benchmark for the N_2 molecule," Chem. Phys. Lett. **310**, 530 (1999).

[2] K. D. Vogiatzis, D. Ma, J. Olsen, L. Gagliardi, and W. A. de Jong, "Pushing configuration-interaction to the limit: Towards massively parallel MCSCF calculations," J. Chem. Phys. **147**, 184111 (2017).

[3] P. M. Zimmerman, "Incremental full configuration interaction," J. Chem. Phys. **146**, 104102 (2017).

[4] J. J. Eriksen, F. Lipparini, and J. Gauss, "Virtual Orbital Many-Body Expansions: A Possible Route towards the Full Configuration Interaction Limit," J. Phys. Chem. Lett. **8**, 4633 (2017).

[5] J. J. Eriksen and J. Gauss, "Many-Body Expanded Full Configuration Interaction. I. Weakly Correlated Regime," J. Chem. Theory Comput. **14**, 5180 (2018).

[6] Y. Ohtsuka and S. Nagase, "Projector Monte Carlo method based on configuration state functions. Test applications to the H_4 system and dissociation of LiH," Chem. Phys. Lett. **463**, 431 (2008).

[7] G. H. Booth, A. J. W. Thom, and A. Alavi, "Fermion Monte Carlo without fixed nodes: A game of life, death, and annihilation in Slater determinant space," J. Chem. Phys. **131**, 054106 (2009).

[8] C. Daday, S. Smart, G. H. Booth, A. Alavi, and C. Filippi, "Full Configuration Interaction Excitations of Ethene and Butadiene: Resolution of an Ancient Question," J. Chem. Theory Comput. **8**, 4441–4451 (2012).

[9] B. Huron, J. P. Malrieu, and P. Rancurel, "Iterative perturbation calculations of ground and excited state energies from multiconfigurational zeroth-order wavefunctions," J. Chem. Phys. **58**, 5745 (1973).

[10] Y. Garniron, T. Applencourt, K. Gasperich, A. Benali, A. Ferté, J. Paquier, B. Pradines, R. Assaraf, P. Reinhardt, J. Toulouse, P. Barbaresco, N. Renon, G. David, J.-P. Malrieu, M. Véril, M. Caffarel, P.-F. Loos, E. Giner, and A. Scemama, "Quantum Package 2.0: An Open-Source Determinant-Driven Suite of Programs," J. Chem. Theory Comput. **15**, 3591 (2019).

[11] A. Scemama, A. Benali, D. Jacquemin, M. Caffarel, and P.-F. Loos, "Excitation energies from diffusion Monte Carlo using selected configuration interaction nodes," J. Chem. Phys. **149**, 034108 (2018).

[12] A. Scemama, Y. Garniron, M. Caffarel, and P.-F. Loos, "Deterministic Construction of Nodal Surfaces within Quantum Monte Carlo: The Case of FeS," J. Chem. Theory Comput. **14**, 1395 (2018).

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

[14] Y. Garniron, A. Scemama, E. Giner, M. Caffarel, and P.-F. Loos, "Selected configuration interaction dressed by perturbation," J. Chem. Phys. **149**, 064103 (2018).

[15] T. Applencourt, K. Gasperich, and A. Scemama, "Spin adaptation with determinant-based selected configuration interaction," arXiv:1812.06902 (2018).

[16] N. M. Tubman, J. Lee, T. Y. Takeshita, M. Head-Gordon, and K. Birgitta Whaley, "A deterministic alternative to the full configuration interaction quantum Monte Carlo method," J. Chem. Phys. **145**, 044112 (2016).

[17] A. A. Holmes, N. M. Tubman, and C. J. Umrigar, "Heat-Bath Configuration Interaction: An Efficient Selected Configuration Interaction Algorithm Inspired by Heat-Bath Sampling," J. Chem. Theory Comput. **12**, 3674 (2016).

[18] Y. Ohtsuka and J. Hasegawa, "Selected configuration interaction using sampled first-order corrections to wave functions," J. Chem. Phys. **147**, 034102 (2017).

[19] J. P. Coe, "Machine learning configuration interaction," J. Chem. Theory Comput. **14**, 5739 (2018).

[20]F. A. Evangelista, "Adaptive multiconfigurational wave functions," J. Chem. Phys. **140**, 124114 (2014).
[21]J. B. Schriber and F. A. Evangelista, "Communication: An adaptive configuration interaction approach for strongly correlated electrons with tunable accuracy," J. Chem. Phys. **144**, 161106 (2016).
[22]J. B. Schriber and F. A. Evangelista, "Adaptive Configuration Interaction for Computing Challenging Electronic Excited States with Tunable Accuracy," J. Chem. Theory Comput. **13**, 5354 (2017).
[23]J. C. Greer, "Estimating full configuration interaction limits from a Monte Carlo selection of the expansion space," J. Chem. Phys. **103**, 1821 (1995).
[24]J. C. Greer, "Monte Carlo Configuration Interaction," J. Comp. Phys. **146**, 181 (1998).
[25]L. Tong, M. Nolan, T. Cheng, and J. C. Greer, "A Monte Carlo configuration generation computer program for the calculation of electronic states of atoms, molecules, and quantum dots," Comp. Phys. Comm. **131**, 142 (2000), see https://github.com/MCCI/mcci.
[26]J. C. Greer, "Consistent treatment of correlation effects in molecular dissociation studies using randomly chosen configurations," J. Chem. Phys. **103**, 7996 (1995).
[27]J. P. Coe, D. J. Taylor, and M. J. Paterson, "Calculations of potential energy surfaces using Monte Carlo configuration interaction," J. Chem. Phys. **137**, 194111 (2012).
[28]W. Győrffy, R. J. Bartlett, and J. C. Greer, "Monte Carlo configuration interaction predictions for the electronic spectra of Ne, CH_2, C_2, N_2, and H_2O compared to full configuration interaction calculations," J. Chem. Phys. **129**, 064103 (2008).
[29]J. P. Coe, P. Murphy, and M. J. Paterson, "Applying Monte Carlo configuration interaction to transition metal dimers: Exploring the balance between static and dynamic correlation," Chem. Phys. Lett. **604**, 46 (2014).
[30]J. P. Coe, D. J. Taylor, and M. J. Paterson, "Monte Carlo configuration interaction applied to multipole moments, ionization energies, and electron affinities," J. Comput. Chem. **34**, 1083 (2013).
[31]J. P. Coe and M. J. Paterson, "State-averaged Monte Carlo configuration interaction applied to electronically excited states," J. Chem. Phys **139**, 154103 (2013).
[32]J. P. Coe and M. J. Paterson, "Approaching exact hyperpolarizabilities via sum-over-states Monte Carlo configuration interaction," J. Chem. Phys **141**, 124118 (2014).
[33]T. P. Kelly, A. Perera, R. J. Bartlett, and J. C. Greer, "Monte Carlo configuration interaction with perturbation corrections for dissociation energies of first row diatomic molecules: C_2, N_2, O_2, CO, and NO," J. Chem. Phys. **140**, 084114 (2014).
[34]J. P. Coe and M. J. Paterson, "Multireference X-ray emission and absorption spectroscopy calculations from Monte Carlo configuration interaction," Theor. Chem. Acc. **134**, 58 (2015).
[35]M. Szepieniec, I. Yeriskin, and J. C. Greer, "Quasiparticle energies and lifetimes in a metallic chain model of a tunnel junction," J. Chem. Phys. **138**, 144105 (2013).
[36]P. Murphy, J. P. Coe, and M. J. Paterson, "Development of spin-orbit coupling for stochastic configuration interaction techniques," J. Comp. Chem. **39**, 319 (2018).
[37]J. P. Coe and M. J. Paterson, "Development of Monte Carlo configuration interaction: Natural orbitals and second-order perturbation theory," J. Chem. Phys. **137**, 204108 (2012).
[38]H.-J.Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, P. Celani, W. Györffy, D. Kats, T. Korona, R. Lindh, A. Mitrushenkov, G. Rauhut, K. R. Shamasundar, T. B. Adler, R. D. Amos, A. Bernhardsson, A. Berning, D. L. Cooper, M. J. O. Deegan, A. J. Dobbyn, F. Eckert, E. Goll, C. Hampel, A. Hesselmann, G. Hetzer, T. Hrenar, G. Jansen, C. Köppl, Y. Liu, A. W. Lloyd, R. A. Mata, A. J. May, S. J. McNicholas, W. Meyer, M. E. Mura, A. Nicklass, D. P. O'Neill, P. Palmieri, D. Peng, K. Pflüger, R. Pitzer, M. Reiher, T. Shiozaki, H. Stoll, A. J. Stone, R. Tarroni, T. Thorsteinsson, and M. Wang, "Molpro, version 2015.1, a package of ab initio programs," (2015), see.
[39]M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, "*Gaussian 09 Revision D.01*," (2016), *Gaussian Inc. Wallingford CT*.
[40]D. Zuev, E. Vecharynski, C. Yang, N. Orms, and A. I. Krylov, "New Algorithms for Iterative Matrix-Free Eigensolvers in Quantum Chemistry," J. Comp. Chem. **36**, 273 (2015).
[41]J. P. Coe and M. J. Paterson, "Investigating Multireference Character and Correlation in Quantum Chemistry," J. Chem. Theory Comput. **11**, 4189 (2015).
[42]X. Li and J. Paldus, "Comparison of the open-shell state-universal and state-selective coupled-cluster theories: H_4 and H_8 models," The Journal of Chemical Physics **103**, 1024–1034 (1995).