Diagrammatic Monte Carlo for electronic correlation in molecules: High-order many-body perturbation theory with low scaling

G. Bighin,[1,2] Q. P. Ho [ORCID],[2,3] M. Lemeshko,[2] and T. V. Tscherbul [ORCID][4]

[1]*Institut für Theoretische Physik, Universität Heidelberg, Philosophenweg 19, 69120 Heidelberg, Germany*
[2]*Institute of Science and Technology Austria (ISTA), Am Campus 1, 3400 Klosterneuburg, Austria*
[3]*Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong*
[4]*Department of Physics, University of Nevada, Reno, Nevada 89557, USA*

We present a low-scaling diagrammatic Monte Carlo approach to molecular correlation energies. Using combinatorial graph theory to encode many-body Hugenholtz diagrams, we sample the Møller-Plesset (MPn) perturbation series, obtaining accurate correlation energies up to $n = 5$, with quadratic scaling in the number of basis functions. Our technique reduces the computational complexity of the molecular many-fermion correlation problem, opening up the possibility of low-scaling, accurate stochastic computations for a wide class of many-body systems described by Hugenholtz diagrams.

I. INTRODUCTION

The many-electron correlation energy, defined as the difference between the true energy of a many-electron system and that obtained in the Hartree-Fock (HF) approximation [1–6], plays a central role in the theoretical description of a wide array of phenomena in chemistry, physics, and material science [1,6–12]. The development of highly efficient computational methods for calculating the correlation energy is thus an ultimate goal of modern electronic structure theory [2,3,7]. Many-body perturbation theory (MBPT) [1,5] and coupled-cluster (CC) theory [2] are the two primary methods for treating the effects of dynamic correlation, where a single HF state provides a qualitatively correct zeroth-order approximation to the electronic wave function [2]. Both MBPT and CC methods have been very successful in predicting the correlation energies of small- and medium-sized molecules [2,5], whereas second-order Møller-Plesset perturbation theory (MP2) [1–3] is the method of choice for calculating electron correlation effects in large systems involving thousands of atoms [4,5].

However, widespread application of MPn and CC methods to larger molecules is limited by the steep scaling of the computational cost (N^{n+3} in the case of MPn) with the number of spin orbitals N [4,5,13]. This problem has motivated the development of ingenious low-scaling methods [14–18]. Among those, several promising Monte Carlo (MC) techniques rely on stochastic sampling of configuration-interaction (CI) [19,20] and CC [21] expansions in imaginary time or performing real-space MC integration to obtain MPn energies [5,22–25].

In this paper, we introduce a stochastic approach to the many-electron correlation problem in molecules based on the powerful diagrammatic Monte Carlo (DiagMC) methodology [26–28], which uses direct sampling of the entire diagrammatic series for the many-electron correlation energy to obtain numerical results free of systematic bias. Originally developed in the context of quantum impurity problems [26,28], DiagMC has been applied with great success to a wide range of problems in quantum many-body physics, including exotic impurities with internal degrees of freedom [29–32], correlated lattice fermions [33], unitary Fermi gases [34], and nonequilibrium quantum dynamics [35]. Recent applications of the DiagMC approach have provided numerically exact correlation energies of the homogeneous electron gas [36] and of an infinite chain of hydrogen atoms [37]. Thus far, however, save for a very recent application to molecular quantum impurity problems at finite temperature [38], DiagMC has not been applied to calculate molecular correlation energies, likely due to the topological complexity of the underlying Hugenholtz diagrams.

Here, we overcome this problem by using combinatorial graph theory to encode Hugenholtz diagrams into adjacency matrices, a technique recently developed in nuclear physics [39,40]. This allows us to design general and efficient updates for sampling the diagrammatic expansions of MBPT using the Metropolis algorithm. Unlike full configuration-interaction MC [20], stochastic MPn theory in real space [5,22–24], or DiagMC for molecular quantum impurities [38], our DiagMC/MPn method evaluates the correlation energy directly based on a random walk in the space of Hugenholtz diagrams, rather than that of Slater determinants or in real space.

We apply our approach to calculate the correlation energies of small molecules, obtaining accurate MPn results up to $n = 5$ with low scaling, opening up the possibility of computing accurate dynamical correlation energies for much larger systems. Because our methodology only relies on graph theory, it can be easily extended beyond electronic structure theory to include the diagrammatic expansions that occur in, e.g., vibrational spectroscopy [41–43], crystal phonon perturbation theory [44–46], and nuclear physics [39,40].

II. MPn THEORY AND MATRIX ENCODING OF HUGENHOLTZ DIAGRAMS

In MPn theory [3–5], the nonrelativistic electronic Hamiltonian $\hat{H} = \sum \varepsilon_i \hat{c}_i^\dagger \hat{c}_i + \frac{1}{2} \sum_{ijkl} \langle ij||kl \rangle \hat{c}_i^\dagger \hat{c}_j^\dagger \hat{c}_l \hat{c}_k$ is partitioned into the mean-field reference Hamiltonian $\hat{H}_0 = \sum_i \varepsilon_i \hat{c}_i^\dagger \hat{c}_i$ plus a fluctuation potential $\hat{V} = \hat{H} - \hat{H}_0$, where ε_i are the HF orbital energies, $\hat{c}_i^\dagger$ ($\hat{c}_i$) are the creation (annihilation) operators for the electron in the ith HF spin orbital, respectively, and $\langle ij||kl \rangle$ are the antisymmetrized two-electron repulsion integrals (ERIs) [1]. The correlation energy is given by the Rayleigh-Schrödinger perturbation series including only the linked terms [3,47],

$$E_{\mathrm{c}} = E - E_{\mathrm{HF}} = \sum_{m=0} \langle \hat{V}(\hat{R}_0 \hat{V})^m \rangle_L, \tag{1}$$

where $\hat{R}_0 = (1 - |\Phi_0\rangle\langle\Phi_0|)(z - \hat{H}_0)^{-1}$ is the reduced resolvent operator for the HF reference state $|\Phi_0\rangle$ [3,48].

The different terms in the MPn series of Eq. (1) are most compactly represented by means of Hugenholtz diagrams [1,49]. Diagrams contributing to the n order consist of n labeled vertices, vertically aligned by convention. Each vertex corresponds to an ERI $\langle ij||kl \rangle$ and has two incoming and two outgoing lines, corresponding to either particle orbitals $a, b, c, \dots$ (upward lines) or hole orbitals $r, s, t, \dots$ (downward lines) [3]. Additionally, diagrams with different line orientations are considered distinct, a line cannot start and end on the same vertex, each diagram must consist of only one connected component, and its overall sign $s = (-1)^{h-l}$ depends on the number of holes h and of closed loops l. Each pair of adjacent vertices contributes the factor $(\sum_h \varepsilon_h - \sum_p \varepsilon_p)^{-1}$, where the sums run over the particles and holes crossing an imaginary horizontal line between the vertices. Finally, each diagram is scaled by 2^{-p}, p being the number of equivalent line pairs (i.e., codirected lines that start and end on the same vertex) [3]. The number of nth-order Hugenholtz diagrams grows factorially with n [50]. Representative $n = 2$–4 diagrams are shown in Fig. 1. Our DiagMC approach presented below directly samples the expansion of the correlation energy (1) in terms of Hugenholtz diagrams.

We now observe, following recent work on MBPT in nuclear physics [39,40], that, according to graph theory, Hugenholtz diagrams of order n can be conveniently encoded into $n \times n$ adjacency matrices that satisfy the following conditions: (i) A_{ij} can only take values 0, 1, 2, (ii) $\sum_i A_{ij} = 2 \, \forall j$, (iii) $\sum_j A_{ij} = 2 \, \forall i$, and (iv) $A_{ii} = 0 \, \forall i$. Figure 1 shows the adjacency matrix representations of selected MPn diagrams. We stress that in some contexts one names "diagram" the summed-over expression, after the sums over the hole and particle indices have been carried out. Here, we call "diagram" an expression depending on these indices, with no sum implied. We make this apparent by introducing appropriate subscript indices on the entries of the matrices of Fig. 1; the diagrammatic rules above, along with the convention we choose for the adjacency matrix, imply that entries below (above) the diagonal will carry hole (particle) indices, respectively. The core idea of the present paper is to stochastically sample these diagrams—through their matrix representation—at all orders, varying the topology of the diagram and the value of the
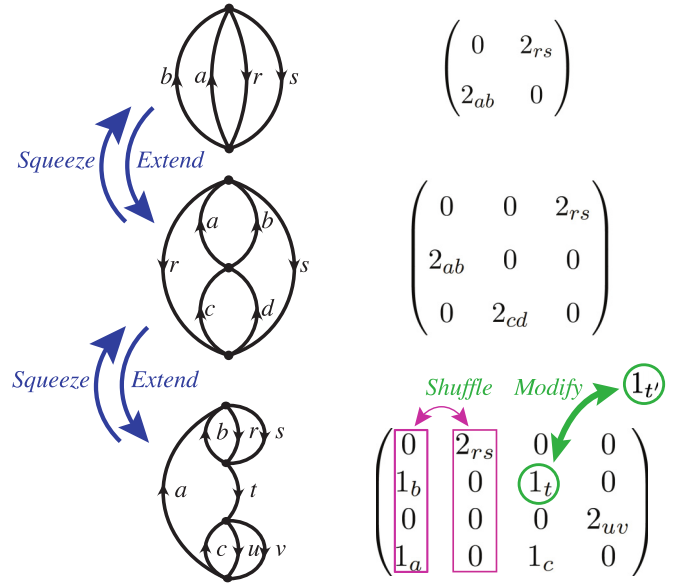

FIG. 1. Representative Hugenholtz diagrams contributing to the n order of MP theory with $n = 2$, 3, and 4 along with their corresponding adjacency matrices [39,40] shown to the right of each diagram. The different types of DiagMC updates are indicated by blue arrows (for *Extend* and *Squeeze*), by magenta arrows (for *Shuffle*), and by green arrows (for *Modify*) (see main text for details).

indices [26,51–53], converging, in the statistical sense, to the exact correlation energy.

For this purpose, we start by relaxing condition (iv) above, considering a larger set of matrices that have $A_{ii} \neq 0$. Within this extended configuration space $\mathcal{E}$, we distinguish between physical matrices, satisfying all four conditions, and unphysical ones, satisfying only conditions (i)–(iii) above. It can be shown (see Appendix A) that each matrix in $\mathcal{E}$ can be represented as the sum of two $N \times N$ permutation matrices P defined by the following conditions: (1) Each $P_{ij} = 0$ or 1, (2) $\sum_i P_{ij} = 1 \, \forall j$, and (3) $\sum_j P_{ij} = 1 \, \forall i$. The converse is also, more trivially, true: Two permutation matrices always sum to a matrix in $\mathcal{E}$. Therefore, the configuration space $\mathcal{E}$ consists essentially of two copies of the permutation matrix configuration space. As a consequence, we can just design a stochastic process sampling in the permutation matrix configuration space, subsequently "doubling" it to sample over $\mathcal{E}$.

III. DIAGRAMMATIC MONTE CARLO PROCEDURE

We now apply the DiagMC methodology [26,51] by devising a set of updates that can ergodically explore the space of permutation matrices. The *Extend*$_1$ update adds a row to the bottom and a column to the right of a permutation matrix, thereby going from order N to order $N + 1$. We begin by choosing a nonzero entry P_{ij} of the original matrix, setting it to zero, and subsequently "projecting" it onto the newly created column and row. More specifically, we add two new entries $P_{(N+1)j} = 1$ and $P_{i(N+1)} = 1$. Due to the conventions discussed above, $P_{i(N+1)}$ will carry a hole index, while $P_{(N+1)j}$ will carry a particle index. We then reuse the numerical value of the index of the erased entry P_{ij} as the index carried by one of the two new entries. Depending whether the old value

was a particle or hole index, we will need to choose from a discrete uniform random distribution a new hole or particle index, respectively. The probability for this update is then

$$\mathcal{P}_{\text{Extend}_1} = \begin{cases} \frac{1}{N\,n_{\text{p}}} & \text{if } i \leqslant j, \\ \frac{1}{N\,n_{\text{h}}} & \text{otherwise,} \end{cases} \quad (2)$$

where n_{h} (n_{p}) is the total number of hole (particle) orbitals in the basis set being used, respectively. For the complementary update, that we denote $Squeeze_1$, we need to remove the two elements on the last row and column. There is just one way of doing so. Then we need to restore the P_{ij} matrix element whose index might correspond either to a hole or to a particle state, and we can get the numerical value of that index from the index of one of the removed entries. The probability is then $\mathcal{P}_{\text{Squeeze}_1} = 1$.

The $Extend_2$ update adds one column and one row to a permutation matrix, and adds a new "1" entry on the diagonal, on the bottom right. This will always take us to the unphysical sector, and by convention the newly added entry will always carry a hole index. The value is then drawn from a uniform random distribution, and the probability is then $\mathcal{P}_{\text{Extend}_2} = 1/n_{\text{h}}$. The complementary $Squeeze_2$ update simply deletes the matrix element in the bottom right corner, returning to an $N \times N$ matrix. There are no probability distributions involved in this process, therefore one has $\mathcal{P}_{\text{Squeeze}_2} = 1$.

In the $Shuffle$ update, we first decide if we want to shuffle rows or columns. We then choose two random rows or columns and swap them. In doing so, the update might need to replace a hole index with a particle one or vice versa, thus requiring to draw numbers from a uniform distribution. However, since the update is clearly self-complementary, one does not need to keep track of the associated probabilities, since the acceptance ratio depends on weight ratios only. Lastly, we design a $Modify$ update, in which a nonzero hole or particle entry is selected and the associated index is changed to a different value chosen from a uniform distribution. This update is also self-complementary [26,51–53].

It is easily seen that the set of updates just introduced is ergodic. We then consider two permutation matrices and we apply the updates just introduced to each matrix at each MC step, with the constraint that the two matrices must always have the same dimension. In the spirit of DiagMC, we accept or reject the updates with a probability chosen so as to make the process satisfy a detailed balance condition [26,51–53]; this implies that in the long run the process will spend with each diagram a number of MC steps proportional to the diagram weight, allowing us to collect statistics about the ratio of energies at different orders. The process jumps back and forth between the physical and unphysical sectors, the latter not contributing to the sampled quantities [52,53]. We verified that at every order the fraction of physical diagrams is always substantially large, and moreover an arbitrary unphysical penalty dividing the weight of unphysical diagrams can help in tipping the balance towards the physical sector [52,53].

We finally note that there are several distinct ways in which a given adjacency matrix $\mathcal{A}$ can be represented as a sum of permutation matrices. We will call the number of such ways *the multiplicity* of $\mathcal{A}$. Since the multiplicity is not always

TABLE I. Calculated correlation energies (in μE_h) for BH, H_2O, and C_6H_6 in the 6-31G basis set compared with the reference MPn data from the PSI4 code.

Molecule	MP order n	DiagMC (this work)	Exact
BH	2	-38.983338 ± 0.007937	-38.993128
	3	-13.297859 ± 0.004215	-13.301207
	4	-5.726406 ± 0.022243	-5.728702
	5	-2.664634 ± 0.050187	-2.779645
H_2O	2	-129.050784 ± 0.025138	-129.053394
	3	-1.550238 ± 0.010480	-1.554750
	4	-5.110730 ± 0.052355	-5.247546
C_6H_6	3	-20.582538 ± 0.130253	-20.3382

one, some diagrams can be incorrectly "counted" more than once. To avoid the multiple counting we divide the weight associated to a matrix by its multiplicity. An algorithmic determination of the multiplicity is presented in Appendix B.

IV. RESULTS

A. Correlation energies and scaling

As a first application of the proposed DiagMC/MPn methodology, we carry out proof-of-principle computations on the CH_2, H_2O, and BH molecules [54] and compare the results with reference MPn calculations to assess the accuracy of the approach. This choice of molecules allows us to explore the performance of the DiagMC/MPn approach for different convergence patterns of the MPn series. While CH_2 and BH are type-A molecules, for which the series converges monotonically, H_2O belongs to type B, exhibiting oscillating convergence [55,56].

We observe that the DiagMC/MPn correlation energies listed in Table I and shown in Figs. 2(a) and 2(b), calculated using 10^{12} MC steps per data point, are in excellent agreement with the reference MPn data computed using PSI4 [57], which validates all of the elements of our DiagMC procedure described above. While the error in the DiagMC/MPn correlation energy depends on the molecule and basis set used, it exhibits the expected statistical scaling $1/\sqrt{N_{\text{MC}}}$ with the number of MC steps [see Fig. 2(c)]. Figures 2(d) and 3(b) show that, for a fixed N_{MC}, the error increases approximately linearly as the basis set size is increased from the smallest (STO-3G) to the largest (6-31**G). This leads to the overall $O(N^2)$ scaling of the computational effort in our approach with respect to the number of basis states, which makes it much more attractive computationally than conventional MPn (N^{n+3}). The traditional HF scales as N^4 and the integral transformation scales as N^5. While these scalings could become the bottleneck for large molecules, in our proof-of-principle calculations on small- and medium-sized molecules, the integral transformation is so fast that it does not contribute to the overall scaling. The correlation energy computed with a larger and more flexible correlation-consistent polarized valence triple zeta (cc-pVTZ) basis set is $4.915\,62 \pm 0.003\,66\mu E_h$, to be compared with the exact MP3 result of $4.870\,32\mu E_h$.

It is important to note that, unlike in nuclear physics or Hubbard-type models, in chemical systems the crucial quantities are the energy *differences*. Because the total MPn energies
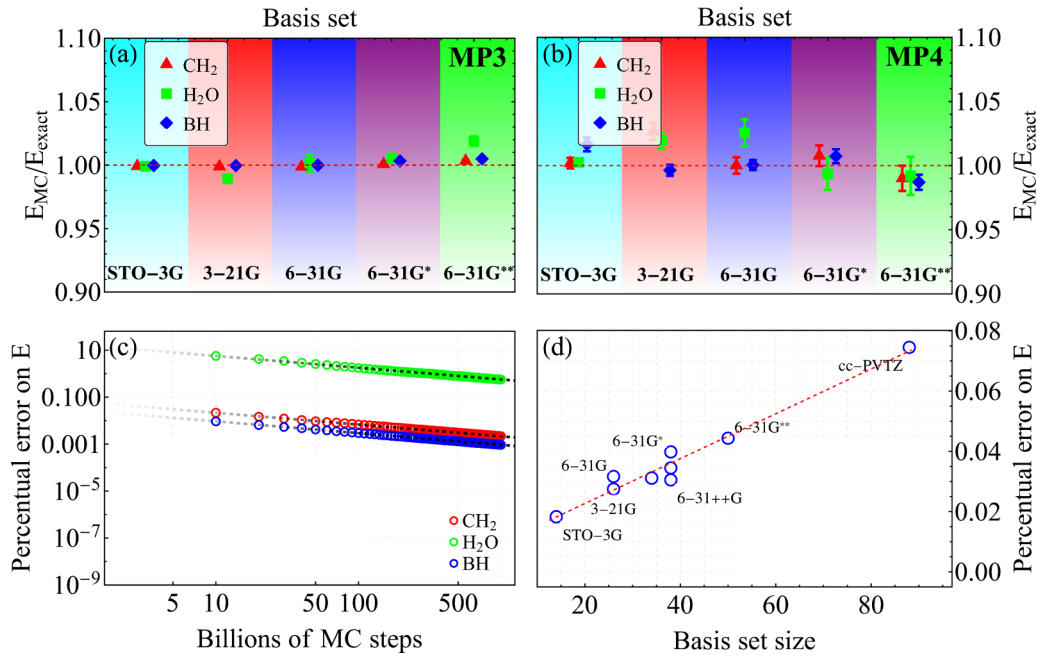
FIG. 2. (a), (b) MP3 and MP4 energies as calculated from our DiagMC procedure for CH_2, H_2O, and BH molecules and for the STO-3G, 3-21G, 6-31G, 6-31*G, 6-31**G basis sets with 10^{12} MC steps per data point. (c) Percentual error on the energy as a function of the number of MC steps; note the logarithmic scale. The black dashed lines guide the eye and correspond to the statistical $1/\sqrt{N_{MC}}$ scaling. (d) Percentual error at MP3 level as a function of basis size for BH with 10^{12} MC steps. The dashed line guides the eyes, highlighting the essentially linear scaling.

grow proportional to system size, reaching a given accuracy for the energy differences would require an additional $O(N^2)$ factor, bringing the overall scaling from $O(N^2)$ to $O(N^4)$, which is still lower than the traditional $O(N^{n+3})$ scaling of conventional MPn for $n \geqslant 3$. In particular, the highest-order
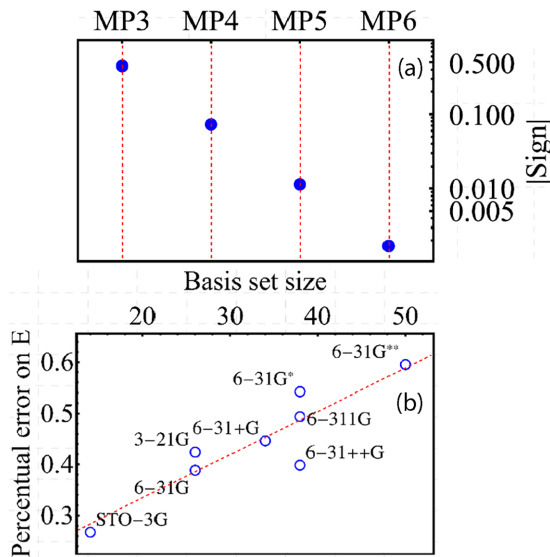

FIG. 3. (a) Expectation value of the sign as a function of the MP order. Notice the logarithmic scale on the vertical axis. (b) Percentual error at MP4 level as a function of basis size for BH with 10^{12} MC steps. The dashed line guides the eyes, highlighting the essentially linear scaling.

DiagMC/MP5 approach demonstrated here scales as $O(N^4)$ compared to the $O(N^8)$ scaling of conventional MP5. Further, for an important class of problems involving the calculation of noncovalent interaction energies, symmetry-adapted perturbation theory (SAPT) [58–60] can be used. SAPT gives the interaction energy directly in terms of the Brandow diagrams [60], which are closely related to the Hugenholtz diagrams considered in this work. DiagMC sampling of these diagrams would eliminate the additional $O(N^2)$ factor and reinstate the $O(N^2)$ scaling of SAPT correlation energies.

B. Sign problem

Some of the diagrams we sample have negative weight, therefore we sample with respect to the absolute value of the diagram weight $\mathcal{D}_\xi$ [61]. Doing so, we observe that the statistical error in DiagMC/MPn correlation energies grows significantly with increasing order n. This is due to the fermion sign problem, whereby the Hugenholtz diagrams with opposite signs cancel out, making it necessary to use an increasingly large number of MC steps to obtain a nonzero signal-to-noise ratio [62–64]. For instance, calculating the MP6 energy of NH using the 6-31G basis and 10^9 MC steps, we obtain a positive contribution of $0.342\,020 \pm 0.001\,172$ and a negative contribution of $0.342\,260 \pm 0.001\,26$, with greatly increased relative error once these are subtracted [65]. We have also investigated this analytically, verifying that several topologies are dominated by near-perfect cancellations.

This phenomenon, bearing a remarkable resemblance to the sign problem observed in other contexts

[27,33,34,53,66,67] is, however, distinct from—and less severe than—the one that plagues quantum many-body MC simulations of, e.g., Fermi-Hubbard models. There, one is interested in the thermodynamic limit, and the expectation value of the sign decreases exponentially with the size of the system [62–64]. In contrast, for finite-size molecules explored here, this expectation value is small, decreases exponentially with the perturbation theory order, but is always finite, as shown in Fig. 3(a), significantly reducing the acuity of the sign problem.

Table I shows that the sign problem becomes severe enough to strongly affect our results only at high MPn orders (MP6 for type-A molecules and MP5 for type-B molecules). Thus DiagMC/MPn yields fairly accurate correlation energies (comparable to those given by coupled-cluster theory with single, double, and perturbative triple excitations [CCSD(T)] for type-A molecules). To explore the severity of the sign problem for larger molecules, we performed DiagMC/MP3 simulations on benzene (C_6H_6) and did not observe any significant increase in statistical error compared to smaller molecules (see Table I). This indicates that the sign problem does not necessarily get more severe with increasing system size. In future work it would be important to explore the sign problem for larger molecules at higher MPn orders.

V. OUTLOOK AND CONCLUSIONS

We have demonstrated a low-scaling stochastic approach to calculating molecular electronic correlation energies based on DiagMC sampling of the MPn series. The approach samples the many-body electronic correlation energy directly using Hugenholtz diagrams, encoded in adjacency matrices using combinatorial graph theory [40]. Our DiagMC/MPn approach shares many of the attractive features with its antecedents in quantum many-body physics [26,27], such as low scaling and the ability to converge towards the exact result (the full CI limit). We demonstrate accurate results for the MPn correlation energies with $n \lesssim 5$. Thus, our low-scaling DiagMC/MPn methodology could be applied to a wide range of quantum chemical problems, where high-precision estimates of dynamical correlation energy are crucial, such as calculating intermolecular dispersion interactions [68].

We find that results for $n \gtrsim 5$ are affected by the sign problem, which, however, is significantly less severe than the sign problem encountered in the thermodynamic limit [62,63] due to the finite size of molecular systems. In future work, we plan to address this problem (see Appendix C) by adapting the recently developed CDet algorithm [69,70]. This would enable one to perform reliable extrapolations to the full CI limit [55], using, e.g., Padé approximants, resummation techniques, and Feenberg scaling [4], and to explore the convergence behavior of the MPn series for large molecules, currently outside of the reach of modern quantum chemistry techniques [56].

ACKNOWLEDGMENTS

APPENDIX A: ADJACENCY MATRICES AND PERMUTATION MATRICES

Here, we demonstrate the claim made in the main text, that each adjacency matrix corresponding to a Hugenholtz diagram can be written as the sum of two permutation matrices.

Definition 1. A square matrix is said to be a permutation matrix if its entries are either 1 or 0, and moreover, the sum of each row/column is exactly 1.

A adjacency matrix in the extended configuration space $\mathcal{E}$ defined in the main text is a 2-permutation matrix in the following definition.

Definition 2. A square matrix is said to be a 2-permutation matrix if its entries are either 2, 1, or 0, and the sum of each row/column is exactly 2.

Clearly, the sum of any two permutation matrices is a 2-permutation matrix. We will call such 2-permutation matrices decomposable. Our goal is to show that, in fact, all 2-permutation matrices are decomposable.

Proposition 3. Any 2-permutation matrix is decomposable, i.e., it can be expressed as a sum of two permutation matrices.

We will use Lemmas 4 and 5 below, which are straightforward.

Lemma 4. Let M be a 2-permutation matrix and M' obtained from M by deleting all rows and columns containing an entry of value 2. Then M is decomposable if and only if M' is.

This lemma allows us to reduce to the case where M only contains 1's. We will assume this from now on.

Lemma 5. Let M be a 2-permutation matrix and M' obtained from M by a series of row/column permutations. Then M is decomposable if and only if M' is.

Proof of Proposition 3. Let M be a 2-permutation matrix. We will show that M is decomposable. As mentioned above, we can, and we will, assume that M only contains 0's and 1's. By Lemma 5, it suffices to do so after a series of row and column permutations. More precisely, by doing row/column permutations, we will turn M into a new matrix with 1's on the diagonal. The decomposition of such a matrix into two permutation matrices is obvious.

Starting with the first row, by doing a row/column permutation, we can turn M into a new matrix such that the top left is given by

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Necessarily, anything else on the first column besides the first two entries are 0's like so,

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Now, for the second row, we know that there exists a nonzero entry. So by row/column permutation, we can turn M into a new matrix with the first two columns having one of the following two forms,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

Now it is clear how to proceed inductively. The main point is that suppose we are done with the first k rows, then the first k entries of the $(k + 1)$th row will have one of the following two forms,

$$[0 \quad 0 \quad \cdots \quad 0 \quad 1] \quad \text{or} \quad [0 \quad 0 \quad \cdots \quad 0 \quad 0], \quad \text{(A1)}$$

i.e., there is at most one 1 entry here, which means that we can find another 1 on the same row. Let $(k + 1, l)$ denote the coordinates of that entry. Note that $l \geqslant k + 1$.

The lth column will have another 1 at coordinate (k', l). If $k' < k + 1$, then we use column permutation to swap the lth column and the $(k + 1)$th column. If $k' > k + 1$, then we use row permutation to swap the k'th row and the $(k + 2)$th row, and then, use column permutation to swap the lth column and the $(k + 1)$th column. After doing this we see that the $(k + 2)$th row will have the form (A1) again, where, of course, in the first case, the 1 is on the $(k + 1)$ slot. Now, we can continue the process until we are done. In the end, our matrix will have the following block(s) along the diagonal (and 0's everywhere else),

$$B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}.$$

It is clear that this matrix is decomposable and the proof concludes. ∎

APPENDIX B: COUNTING MULTIPLICITIES

Above, we showed that a 2-permutation matrix A could be written as a sum of two permutation matrices. This was done by showing that by permuting rows and columns of A, we can bring it to a special form from which the decomposition of A into a sum of two permutation matrices could be obtained directly. To determine the multiplicity of a 2-permutation matrix, i.e., to determine the number of ways to do that, we

proceed similarly. Namely, since permuting rows and columns does not change the result, we can count the number of decompositions after A has been transformed to a special form.

In what follows, we will assume that A does not have 2-entries, as those are easy to deal with. From the proof of the previous Appendix, we see that by permuting rows and columns, any 2-permutation matrix A could be brought to block matrix, where the diagonal is given by blocks of the form

$$B = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}.$$

It thus remains to count the number of matrices such as B could be written as a sum of two permutation matrices. The final answer for A will be the product of all those numbers for all the blocks.

Suppose we want to write $B = B' + B''$, where B' and B'' are permutation matrices. We will now build B' column by column. For the first column, there are exactly two choices for B', corresponding to having 1 on the first or second row.

(1) Suppose we choose the 1 on the first row, we see that for the second column, we have exactly one choice: picking 1 on the second row (since otherwise B' will never have a 1 on the second row). Continuing this way, we see that there is exactly one way to choose the rest of B'.

(2) Suppose we choose the 1 on the second row, and arguing similarly, we also see that the rest of B' is determined uniquely. Thus, there is exactly one way to get the rest of B'.

Altogether, we see that there are exactly two ways to write B as a sum of two permutation matrices. Thus, the multiplicity of A is 2^n, where n is the number of blocks.

APPENDIX C: SIGN PROBLEM

In this Appendix, we examine the sign problem in Diag/MPn simulations and suggest a promising path towards its resolution building on recent advances in connected determinant DiagMC (CDet) [69]. As already stressed in the main text, we note from the outset that the sign problem in Diag/MPn is different from that encountered in previous fermionic DiagMC simulations (see, e.g., Ref. [69]). The latter are typically performed at finite temperature and in the thermodynamic limit, by sampling imaginary-time Feynman diagrams. In contrast, here we are interested in systems with a finite number of electrons at zero temperature described by the Hugenholtz diagrams.

The sign problem here manifests itself as a rapid (exponential or factorial) decline, as a function of the MP order n, of the average value of the sign of sampled diagrams, caused by a nearly perfect cancellation between the positive and negative diagrams [63,69]. This decline causes a rapid increase in statistical error, if the number of MC steps is kept fixed, for the quantity being calculated (in our case, the correlation energy) making DiagMC/MPn simulations for $n \gtrsim 6$ very computationally demanding. Figure 3(a) shows the average value of the sign as a function of MPn order. The sign problem

becomes quite relevant at the fourth order, hence we will focus on MP4 diagrams in the following.

To gain more insight into how sign cancellations occur between the individual Hugenholtz diagrams, we define the quantity

$$R_{\mathcal{S}} = \frac{|\sum_{i \in \mathcal{S}} \mathcal{D}_i|}{\sum_{i \in \mathcal{S}} |\mathcal{D}_i|}, \qquad (C1)$$

where $\mathcal{S}$ denotes a subset of all diagrams and $\mathcal{D}_i$ is the diagram weight associated to the diagram i, as discussed in the main text. From Eq. (C1) one immediately sees that $R_{\mathcal{S}}$ will be close to unity if no cancellations occur between the diagrams in the subset $\mathcal{S}$. In contrast, $R_{\mathcal{S}}$ will approach zero if the sum of the diagram in $\mathcal{S}$ has many cancellations, and will be exactly zero in the case of perfect cancellation. We note that these quantities are not additive, i.e., given two subsets of all diagrams, be they $\mathcal{S}_A$ and $\mathcal{S}_B$, the conditions $R_{\mathcal{S}_A} \sim 1$ and $R_{\mathcal{S}_B} \sim 1$ do not necessarily imply $R_{\mathcal{S}_A \cup \mathcal{S}_B} \sim 1$, since cancellations may occur between the diagrams when the subsets $\mathcal{S}_A$ and $\mathcal{S}_B$ are joined together.

We now wish to use Eq. (C1) to estimate the degree of cancellation between the different subsets of Hugenholtz diagrams, in order to propose a path towards a solution. Preliminarily, we assign a progressive integer number—let us call it the diagram ID—to each order 4 diagram topology in the following way: As discussed in the main text, the adjacency matrix of each diagram is the sum of two permutation matrices. At order 4 there are $4! = 24$ different permutation matrices, and after decomposing a diagram in two permutation matrices, we assign to each one an index—be they i and j—according to the canonical ordering of the permutation it represents. Finally, we assign to each diagram topology the integer number $24i + j$. It turns out that different numbers can correspond to the same adjacency matrix (see Appendix B on the multiplicity): In this case, we conventionally assign to a diagram topology the smallest integer, which will be the diagram ID. This establishes a one-to-one mapping between a subset of integers and the diagram topologies at order 4.

Having done so, in Fig. 4 we show a scatter plot of the R values for the distinct 39 fourth-order Hugenholtz diagram topologies, as a function of their diagram ID, with the subsets defined by all the diagrams with a fixed topology but different internal indices. We observe that the vast majority of the diagrams have R values above 0.2. Notable exceptions include diagram topologies with ID 233, 262, and 334, which correspond to the "single" diagram 4, and to the "triple" diagrams 23 and 24 in Fig. 5.6 of Ref. [3]. These diagram topologies have $R \lesssim 0.05$, signaling a large number of cancellations between positive and negative contributions, in turn meaning that the stochastic sum over the internal variables will be affected by a considerably large statistical error. Also, these diagram topologies contribute the most to the statistical error in the MP4 correlation energy, since the final error is dominated by the terms with the largest error.

This suggests that one could greatly mitigate the sign problem by identifying "problematic" topologies and replacing, for these topologies, the stochastic sums with actual sums. Extending this idea, one could imagine a mixed approach where the stochastic process moves through different
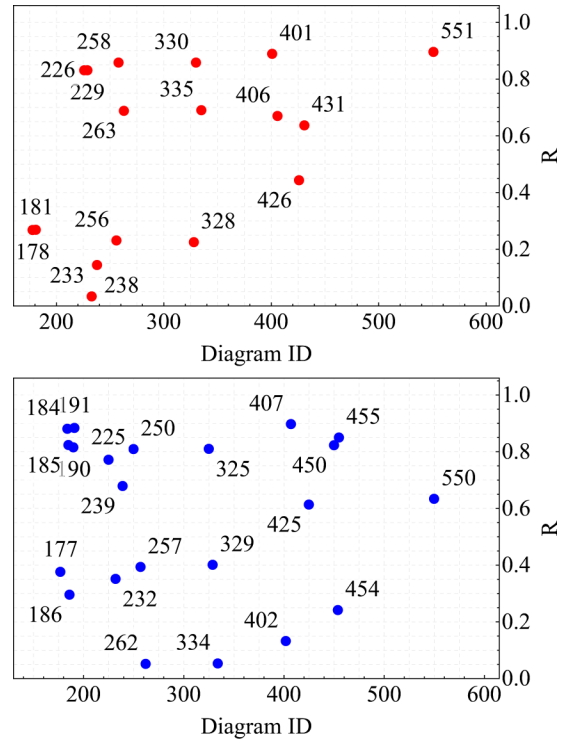

FIG. 4. The R factors defined by Eq. (C1) plotted vs diagram IDs for MP4 Hugenholtz diagrams summed over their internal indices for the BH molecule in the 6-31G basis. Diagram topologies with a total positive expectation value of the sign are plotted in the upper panel, and diagram topologies with a total negative expectation value of the sign in the lower panel.

topologies, whereas all sums over the internal indices are carried over exactly. Using the notation of Ref. [69], denoting with $\mathcal{T}$ the topology of a diagram, and with $\mathcal{X}$ all internal variables, this would correspond to sampling configurations $\mathcal{C} = (\mathcal{T})$ according to the distribution

$$\mathcal{P}(\mathcal{C}) = \left| \sum_{\mathcal{X}} \mathcal{D}(\mathcal{T}; \mathcal{X}) \right|, \qquad (C2)$$

motivated by cancellations between diagrams with the same topology, whereas standard diagrammatic Monte Carlo sample configurations $\mathcal{C} = (\mathcal{T}, \mathcal{X})$ according to the distribution

$$\mathcal{P}(\mathcal{C}) = |\mathcal{D}(\mathcal{T}; \mathcal{X})|. \qquad (C3)$$

It is interesting to note that this approach would be complementary to the CDet approach [71], which samples configurations $\mathcal{C} = (\mathcal{X})$ according to the distribution

$$\mathcal{P}(\mathcal{C}) = \left| \sum_{\mathcal{T}} \mathcal{D}(\mathcal{T}; \mathcal{X}) \right|, \qquad (C4)$$

giving rise to cancellation between diagrams with *different* topologies. The CDet strategy could be adapted to the present case, as well.

In conclusion, it is worth mentioning that significant cancellations can occur when the diagrams with different topologies are added together. For example, while the individual $R_{\mathcal{S}_i}$ values for the single diagram topologies $\mathrm{S}_1 - \mathrm{S}_4$ as

defined in Fig. 5.6 of Ref. [3] are all above 0.5, we have $R_{S_1 \cup \cdots \cup S_4} = 0.016$. This motivates our ongoing studies aimed at further characterizing which diagrams give rise to most of

the cancellations in the MPn series, order by order, with potential far-reaching applications both for analytical and stochastic techniques.

[1] A. Szabo and N. S. Ostlund, *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*, Dover Books on Chemistry (Dover, New York, 1996).

[2] R. J. Bartlett and M. Musiał, Rev. Mod. Phys. **79**, 291 (2007).

[3] I. Shavitt and R. J. Bartlett, *Many-Body Methods in Chemistry and Physics: MBPT and Coupled-Cluster Theory* (Cambridge University Press, Cambridge, UK, 2009).

[4] D. Cremer, WIREs Comput. Mol. Sci. **1**, 509 (2011).

[5] S. Hirata, X. He, M. R. Hermes, and S. Y. Willow, J. Phys. Chem. A **118**, 655 (2014).

[6] D. P. Tew, W. Klopper, and T. Helgaker, J. Comput. Chem. **28**, 1307 (2007).

[7] R. Martin, L. Reining, and D. M. Ceperley, *Interacting Electrons. Theory and Computational Approaches* (Cambridge University Press, Cambridge, U.K., 2006).

[8] X. He, L. Fusti-Molnar, G. Cui, and K. M. Merz, J. Phys. Chem. B **113**, 5290 (2009).

[9] S. Gozem, M. Huntress, I. Schapiro, R. Lindh, A. A. Granovsky, C. Angeli, and M. Olivucci, J. Chem. Theory Comput. **8**, 4069 (2012).

[10] L. Liu, J. Liu, and T. J. Martinez, J. Phys. Chem. B **120**, 1940 (2016).

[11] G. K.-L. Chan and S. Sharma, Annu. Rev. Phys. Chem. **62**, 465 (2011).

[12] L. Mueck, Nat. Chem. **7**, 361 (2015).

[13] L. E. Ratcliff, S. Mohr, G. Huhs, T. Deutsch, M. Masella, and L. Genovese, WIREs Comput. Mol. Sci. **7**, e1290 (2017).

[14] A. F. Izmaylov and G. E. Scuseria, Phys. Chem. Chem. Phys. **10**, 3421 (2008).

[15] P. Pulay, Chem. Phys. Lett. **100**, 151 (1983).

[16] P. Y. Ayala, K. N. Kudin, and G. E. Scuseria, J. Chem. Phys. **115**, 9698 (2001).

[17] H.-J. Werner, F. R. Manby, and P. J. Knowles, J. Chem. Phys. **118**, 8149 (2003).

[18] D. Neuhauser, E. Rabani, and R. Baer, J. Chem. Theory Comput. **9**, 24 (2013).

[19] A. J. W. Thom and A. Alavi, Phys. Rev. Lett. **99**, 143001 (2007).

[20] G. H. Booth, A. J. W. Thom, and A. Alavi, J. Chem. Phys. **131**, 054106 (2009).

[21] C. J. C. Scott, R. Di Remigio, T. D. Crawford, and A. J. W. Thom, J. Phys. Chem. Lett. **10**, 925 (2019).

[22] S. Y. Willow, K. S. Kim, and S. Hirata, J. Chem. Phys. **137**, 204122 (2012).

[23] S. Y. Willow and S. Hirata, J. Chem. Phys. **140**, 024111 (2014).

[24] Z. Li, J. Chem. Phys. **151**, 244114 (2019).

[25] A. E. Doran and S. Hirata, J. Chem. Phys. **154**, 134114 (2021).

[26] N. V. Prokof'ev and B. V. Svistunov, Phys. Rev. Lett. **81**, 2514 (1998).

[27] K. Van Houcke, E. Kozik, N. Prokof'ev, and B. Svistunov, Phys. Procedia **6**, 95 (2010).

[28] E. Gull, A. J. Millis, A. I. Lichtenstein, A. N. Rubtsov, M. Troyer, and P. Werner, Rev. Mod. Phys. **83**, 349 (2011).

[29] G. Bighin and M. Lemeshko, Phys. Rev. B **96**, 085410 (2017).

[30] G. Bighin, T. V. Tscherbul, and M. Lemeshko, Phys. Rev. Lett. **121**, 165301 (2018).

[31] X. Li, G. Bighin, E. Yakaboylu, and M. Lemeshko, Mol. Phys. **117**, 1981 (2019).

[32] X. Li, E. Yakaboylu, G. Bighin, R. Schmidt, M. Lemeshko, and A. Deuchert, J. Chem. Phys. **152**, 164302 (2020).

[33] E. Kozik, K. Van Houcke, E. Gull, L. Pollet, N. Prokof'ev, B. Svistunov, and M. Troyer, Europhys. Lett. **90**, 10004 (2010).

[34] K. Van Houcke, F. Werner, E. Kozik, N. Prokof'ev, B. Svistunov, M. J. H. Ku, A. T. Sommer, L. W. Cheuk, A. Schirotzek, and M. W. Zwierlein, Nat. Phys. **8**, 366 (2012).

[35] G. Cohen, E. Gull, D. R. Reichman, and A. J. Millis, Phys. Rev. Lett. **115**, 266802 (2015).

[36] K. Chen and K. Haule, Nat. Commun. **10**, 3725 (2019).

[37] M. Motta, D. M. Ceperley, G. K.-L. Chan, J. A. Gomez, E. Gull, S. Guo, C. A. Jiménez-Hoyos, T. N. Lan, J. Li, F. Ma, A. J. Millis, N. V. Prokof'ev, U. Ray, G. E. Scuseria, S. Sorella, E. M. Stoudenmire, Q. Sun, I. S. Tupitsyn, S. R. White, D. Zgid, and S. Zhang (Simons Collaboration on the Many-Electron Problem), Phys. Rev. X **7**, 031059 (2017).

[38] J. Li, M. Wallerberger, and E. Gull, Phys. Rev. Res. **2**, 033211 (2020).

[39] A. Tichai, Many-body perturbation theory for *ab initio* nuclear structure, Ph.D. thesis, University of Darmstadt, 2017.

[40] P. Arthuis, T. Duguet, A. Tichai, R. D. Lasseri, and J. P. Ebran, Comput. Phys. Commun. **240**, 202 (2019).

[41] M. R. Hermes and S. Hirata, J. Chem. Phys. **139**, 034111 (2013).

[42] M. R. Hermes and S. Hirata, J. Chem. Phys. **141**, 084105 (2014).

[43] S. Hirata, M. R. Hermes, J. Simons, and J. V. Ortiz, J. Chem. Theory Comput. **11**, 1595 (2015).

[44] R. A. Cowley, Adv. Phys. **12**, 421 (1963).

[45] V. V. Goldman, G. K. Horton, and M. L. Klein, Phys. Rev. Lett. **21**, 1527 (1968).

[46] T. R. Koehler, Phys. Rev. Lett. **22**, 777 (1969).

[47] J. Goldstone and N. F. Mott, Proc. R. Soc. London, Ser. A **239**, 267 (1957).

[48] W. Kutzelnigg, Int. J. Quantum Chem. **109**, 3858 (2009).

[49] J. W. Negele and H. Orland, *Quantum Many-Particle Systems*, Frontiers in Physics (Addison-Wesley, Boston, 1988).

[50] P. D. Stevenson, Number of labeled Hugenholtz diagrams with *n* nodes, in *The On-line Encyclopedia of Integer Sequences* (2020), https://oeis.org/A064732.

[51] A. S. Mishchenko, N. V. Prokof'ev, A. Sakamoto, and B. V. Svistunov, Phys. Rev. B **62**, 6317 (2000).

[52] J. Gukelberger, E. Kozik, and H. Hafermann, Phys. Rev. B **96**, 035152 (2017).

[53] K. Van Houcke, F. Werner, T. Ohgoe, N. V. Prokof'ev, and B. V. Svistunov, Phys. Rev. B **99**, 035140 (2019).

[54] The source code and the configuration files needed for generating all the data presented in this paper can be found at https://github.com/bighin/mpn.

[55] D. Cremer and Z. He, J. Phys. Chem. **100**, 6173 (1996).

[56] M. L. Leininger, W. D. Allen, H. F. Schaefer, and C. D. Sherrill, J. Chem. Phys. **112**, 9213 (2000).

[57] R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer, K. Patkowski *et al.*, J. Chem. Theory Comput. **13**, 3185 (2017).

[58] K. Szalewicz, WIREs Comput. Mol. Sci. **2**, 254 (2012).

[59] P. S. Żuchowski and R. Moszynski, J. Chem. Theory Comput. **19**, 1177 (2023).

[60] S. Rybak, B. Jeziorski, and K. Szalewicz, J. Chem. Phys. **95**, 6576 (1991).

[61] E. Gull and M. Troyer, in *Strongly Correlated Systems*, edited by A. Avella and F. Mancini (Springer, Berlin, 2013).

[62] D. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, Cambridge, U.K., 2000).

[63] E. Y. Loh, J. E. Gubernatis, R. T. Scalettar, S. R. White, D. J. Scalapino, and R. L. Sugar, Phys. Rev. B **41**, 9301 (1990).

[64] S. Chandrasekharan and U.-J. Wiese, Phys. Rev. Lett. **83**, 3116 (1999).

[65] All the errors on MC quantities we report throughout this paper have been determined by means of a jackknife analysis, using the ALEA library from the ALPSCORE package [72–74].

[66] N. Prokof'ev and B. Svistunov, Phys. Rev. B **77**, 020408(R) (2008).

[67] J. Vlietinck, J. Ryckebusch, and K. Van Houcke, Phys. Rev. B **89**, 085119 (2014).

[68] A. Heßelmann, J. Chem. Phys. **151**, 114105 (2019).

[69] R. Rossi, N. Prokof'ev, B. Svistunov, K. Van Houcke, and F. Werner, Europhys. Lett. **118**, 10004 (2017).

[70] R. Rossi, T. Ohgoe, E. Kozik, N. Prokof'ev, B. Svistunov, K. Van Houcke, and F. Werner, Phys. Rev. Lett. **121**, 130406 (2018).

[71] R. Rossi, Phys. Rev. Lett. **119**, 045701 (2017).

[72] B. Bauer, L. D. Carr, H. G. Evertz, A. Feiguin, J. Freire, S. Fuchs, L. Gamper, J. Gukelberger, E. Gull, S. Guertler, A. Hehn, R. Igarashi, S. V. Isakov, D. Koop, P. N. Ma, P. Mates, H. Matsuo, O. Parcollet, G. Pawłowski, J. D. Picon *et al.*, J. Stat. Mech.: Theory Exp. (2011) P05001.

[73] A. Gaenko, A. Antipov, G. Carcassi, T. Chen, X. Chen, Q. Dong, L. Gamper, J. Gukelberger, R. Igarashi, S. Iskakov, M. Könz, J. LeBlanc, R. Levy, P. Ma, J. Paki, H. Shinaoka, S. Todo, M. Troyer, and E. Gull, Comput. Phys. Commun. **213**, 235 (2017).

[74] M. Wallerberger, S. Iskakov, A. Gaenko, J. Kleinhenz, I. Krivenko, R. Levy, J. Li, H. Shinaoka, S. Todo, T. Chen, X. Chen, J. P. F. LeBlanc, J. E. Paki, H. Terletska, M. Troyer, and E. Gull, arXiv:1811.08331.