Two-dimensional quantum lattice models via mode optimized hybrid CPU-GPU density matrix renormalization group method

Andor Menczer,[1] Kornél Kapás,[1,2] Miklós Antal Werner,[1] and Örs Legeza[1,3,*]

[1]*Strongly Correlated Systems Lendület Research Group, Wigner Research Centre for Physics, H-1525 Budapest, Hungary*

[2]*Department of Theoretical Physics, Institute of Physics, Budapest University of Technology and Economics, Műegyetem rkp. 3, H-1111 Budapest, Hungary*

[3]*Institute for Advanced Study, Technical University of Munich, Lichtenbergstrasse 2a, 85748 Garching, Germany*

(Received 12 December 2023; revised 29 March 2024; accepted 3 May 2024; published 14 May 2024)

We present a hybrid numerical approach to simulate quantum many-body problems on two spatial dimensional quantum lattice models via the non-Abelian *ab initio* version of the density matrix renormalization group method on state-of-the-art high-performance computing infrastructures. We demonstrate that for the two-dimensional spinless fermion model and for the Hubbard model on torus geometry, altogether several orders of magnitude in computational time can be saved by performing calculations on an optimized basis and by utilizing hybrid CPU-multiGPU parallelization. At least an order of magnitude reduction in computational complexity results from mode optimization, while a further order of reduction in wall time is achieved by massive parallelization. Our results are measured directly in the number of floating point operations and seconds. A detailed scaling analysis of the obtained performance as a function of matrix ranks and as a function of system size up to 12×12 lattice topology is discussed. Our CPU-multiGPU model also tremendously accelerates the calculation of the one- and two-particle reduced density matrices, which can be used to construct various order parameters and trace quantum phase transitions with high fidelity.

DOI: 10.1103/PhysRevB.109.195148

I. INTRODUCTION

The simulation of strongly correlated quantum many-body problems in two and higher spatial dimensions poses a great challenge and is still part of active research in modern condensed-matter physics [1–3]. In spite of great efforts to generalize the density matrix renormalization group (DMRG) method [4–8], that is the most accurate algorithm for one-dimensional systems; there is still no universal tensor network state (TNS) solution that could be applied in higher spatial dimensions with an efficiency approaching the DMRG method for one-dimensional systems.

In fact, the DMRG method has serious limitations in higher spatial dimensions as even models with local interactions become longer ranged when mapped to the one-dimensional DMRG topology [9–11]. Other approaches such as the projected entangled pair states (PEPS) and its variants [12–14] or the tree tensor network state (TTNS) algorithms [15–19] are based on tensor networks that are already higher dimensional [14,20], but, on the other hand, they have much higher computational demands due to the significant increase in scaling exponents with increasing accuracy. Other alternatives, that are not based on TNS wave functions, also have limitations. Quantum Monte Carlo (QMC) and its extensions are hindered by the so-called negative sign problem [21–24], which generally arises for fermionic and frustrated models and can only be avoided for special model parameters [23,25]. Other approaches, such as dynamical mean-field theory (DMFT), density matrix embedding theory (DMET), etc., are based on a self-consistent simplification of the many-body Hamiltonian and are therefore nonvariational [1,26–28]. Nevertheless, besides their limitations, all of the above-mentioned methods also have their benefits and are commonly used—sometimes complementing each other—as state-of-the-art numerical methods for two- and higher-dimensional many-body quantum systems [1].

Entanglement scaling law—the so-called area law—for two- and higher-dimensional systems also indicates serious limitations for matrix product state (MPS) based methods [29,30]. Most of the DMRG-based analysis, however, has been performed for lattice models with local interactions and represented by localized basis sets, i.e., lattice models [29,31]. We also note that using periodic boundary conditions for such lattice models, which is usually beneficial due to the suppressed finite-size effects, may result in even higher entanglement [32]. To circumvent these difficulties, a commonly used approach in two dimensions is to consider systems with a rather narrow stripe or cylinder topology, where boundary conditions are periodic for the shorter side, while open for the long side [1,33,34]. This approach is sometimes supplemented by a transformation to momentum space in the shorter direction to also exploit the translational invariance [1,35,36]. In these calculations, however, the length of the shorter side of the system is usually rather limited ($L_\perp \lesssim 7-8$), which may results in stronger and spatially inhomogeneous finite-size effects.

The curse of high entanglement can be removed or at least reduced by transforming to a more general basis via nonlocal, one-particle unitary transformations. Using such

*legeza.ors@wigner.hu

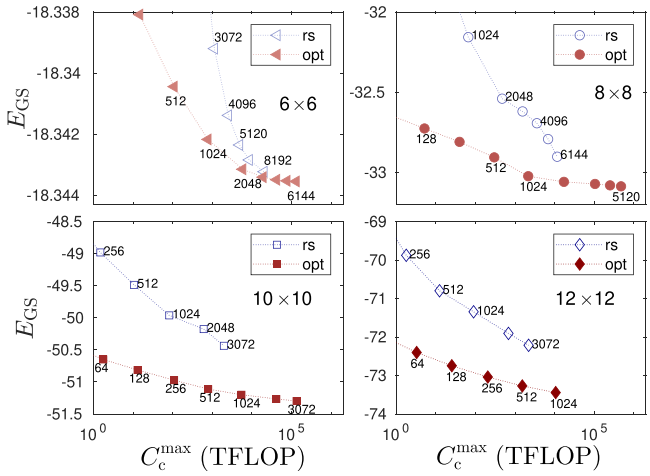

FIG. 1. Ground-state energy as a function of the maximum computational complexity, C_c^{max}, for the half-filled, $n \times n$ ($n = 6, 8, 10, 12$), two-dimensional spinless fermion model with $t = 1, t' = 0.4$, and $V = 0.8$ on a torus geometry using a real-space (rs) and optimized (opt) basis. Numbers next to some selected data points label the DMRG bond dimension.

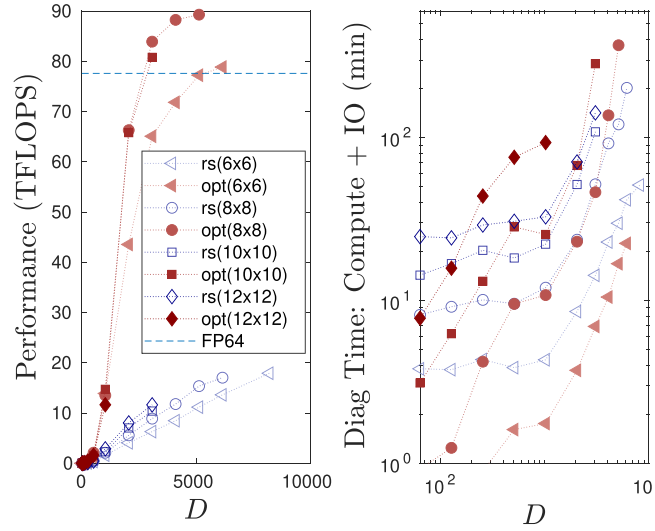

FIG. 2. Maximum performance in TFLOPS (left) and total diagonalization time including device-to-host (D2H) IO of nine DMRG sweeps in minutes (right) measured for the eight GPU accelerated diagonalization of the effective Hamiltonian as a function of the bond dimension using the real-space and the optimized basis for the model parameters given in Fig. 1. The estimated FP64 theoretical upper bound for eight Nvidia A100-PCIE-40GB GPU devices is shown by the dashed line. Note that the achieved performance above the FP64 theoretical upper bound is due to utilization of the highly specialized Nvidia tensor core units (TCUs).

nonlocal basis can alter the scaling properties and bring in new ingredients, especially for approximate, i.e., truncated solutions. An optimal unitary transformation can result in drastically decreased entanglement and, consequently, a large reduction of tensor ranks, also called the bond dimension [37,38]. In contrast, the nonlocality of the transformation can result in a Hamiltonian with long-range interactions instead of the original lattice model with short-range couplings only. The combined effect of these two opposing mechanisms determines the total computational cost of the calculation, the reduction of which, as much as possible, is mandatory for efficient simulations on classical computers.

This work is devoted to demonstrate the efficiency of our approach to target low-energy eigenstates of two and higher spatial dimensional quantum many-body systems by a generalized variant of the DMRG method. First, the conventional real-space basis is relaxed and a more general basis is employed and optimized via the fermionic mode transformation [37]. As a result, the computational complexity of the problem (number of floating point operations), C_c, for a given accuracy threshold can be reduced by an order of magnitude or even more. This is demonstrated in Fig. 1 for the two-dimensional spinless fermion model at half filling on a torus geometry up to lattice sizes of 12×12 with nearest- and next-nearest-neighbor hopping, $t = 1$ and $t' = 0.4$, respectively, and nearest-neighbor Coulomb interaction $V = 0.8$. This point in the $t' - V$ phase space was selected as a numerically very challenging one due to the vicinity of a quantum phase transition, as will be shown in Sec. III. The tremendous reduction, which also manifests itself in drastic saving in wall time, is based on the minimization of the artificial entanglement and correlation introduced by mapping the physical system to a given tensor network topology [38].

Next we show that another order of magnitude reduction in wall time can be achieved via massive parallelization utilizing state-of-the-art hardware and software solutions. Although the locality of the Hamiltonian is lost via mode optimization, leading to a significant increase in the so-called matrix product operator (MPO) bond dimension, our benchmark calculations using a hybrid CPU-multiGPU DMRG [39,40] demonstrate that the underlying computational power of high-performance computing (HPC) architecture can be utilized more efficiently for the optimized basis (see left panel of Fig. 2). Therefore, for a given DMRG bond dimension, the total wall time of the real and optimized basis can be brought much closer for a broad range of the bond dimension values, D, as is shown in the right panel of Fig. 2. Recalling from Fig. 1 that by using the optimized basis, the same accuracy can be achieved with much smaller D values; the overall reduction in wall time can be well over two orders of magnitude for a given accuracy (see, for example, $D = 512$ data sets for the 8×8 lattice model). Our method, therefore, has the potential to open a remarkable direction to study models in two and higher spatial dimensions.

The reduction in block entropy and the efficiency of mode optimization have already been analyzed in terms of the DMRG bond dimension for the spinless fermion model [38], but a detailed scaling analysis in terms of the underlying computational complexity (C_c), i.e., the competition between the the MPS bond and MPO dimensions, has remained relatively unexplored, although computational complexity is the purest measure that determines the efficiency of a numerical approach at the end. In addition, the suitability for massive parallelization can also bring in new ingredients when efficiency is discussed. Therefore, here we aim to fill this

gap by measuring C_c directly in the number of floating point operations (FLOP), wall time after massive parallelization in seconds and the effect of SU(2) spin symmetry on these quantities.

The setup of the paper is as follows. Section II is devoted to a brief overview of model systems and mode optimization. In Secs. III and IV, we present scaling analysis and numerical benchmark results obtained by large-scale mode optimized hybrid CPU-multiGPU DMRG calculations for the two-dimensional spinless fermion and for the Hubbard models, respectively. Our work closes with main conclusions and future perspectives.

II. THEORY

A. Model Hamiltonians

Our numerical framework to simulate the properties of a two-dimensional quantum lattice models relies on the DMRG method [41], which is a special variant of TNS algorithms [6,14,32,42–44], and on a very general form of the Hamiltonian operator,

$$\mathcal{H} = \sum_{ij\alpha\beta} T_{ij}^{\alpha\beta} c_{i\alpha}^\dagger c_{j\beta} + \sum_{ijkl\alpha\beta\gamma\delta} V_{ijkl}^{\alpha\beta\gamma\delta} c_{i\alpha}^\dagger c_{j\beta}^\dagger c_{k\gamma} c_{l\delta}, \quad (1)$$

implemented in our code [45], which can treat any form of nonlocal interactions related to two-particle scattering processes. The operators $c_{i\alpha}^\dagger$ or $c_{i\alpha}$ denote fermionic creation and annihilation operators [46], where indices i and α specify the mode and (generalized) spin information of the single-particle state, respectively. The general form (1) of the Hamiltonian allows for the simulation of strongly correlated quantum many-body systems in various discipline fields, such as condensed-matter physics, nuclear structure theory, or quantum chemistry even in the relativistic domain [5,6,14,32,41–44,47–54]. In the DMRG method, the eigenstate of the Hamiltonian is approximated by a matrix product state (MPS) wave function, which is optimized iteratively during the so-called DMRG sweeps [5,32]. Local (one-site or two-site) optimization steps lead to the iterative diagonalization of the so-called effective quantum many-body Hamiltonian [32] that is carried out by the Lánczos or Davidson algorithms and which corresponds usually to 85–90% of the total execution time. The effective many-body Hamiltonian is built from renormalized block operators that are formed via the course of the network contraction procedure which is responsible for another 5–10% of the total execution time. The dimension of the virtual indices of the MPS matrices, usually called the DMRG bond dimension D, determines the accuracy of the calculations and also the required computational complexity. The overall computational cost of the DMRG for models with generic two-body interactions described by Eq. (1) scales as D^3N^4, where N stands for the number of modes, i.e., for the system size. The memory requirement is proportional to D^2N^2. The details of the algorithm can be found in various review articles (see Refs. [5,6,14,32,42–44]).

When interactions are local, in the sense that interactions are restricted to nearest- or next-nearest neighbors of the underlying lattice only, the number of nonzero terms in Eq. (1) is largely reduced. For example, if one considers the two-dimensional spin-1/2 Hubbard model including only particle hopping t and on-site Coulomb interaction U, Eq. (1) simplifies to

$$\mathcal{H} = \sum_{\langle i,j\rangle\alpha} (t\, c_{i\alpha}^\dagger c_{j\alpha} + \text{H.c.}) + \sum_i U n_{i\alpha} n_{j\beta}, \quad (2)$$

where i and j denote lattice sites on a two-dimensional lattice topology, $\alpha, \beta \in \{\uparrow, \downarrow\}$ denote the spin of the states, and $\langle i, j\rangle$ mark nearest neighbors on the lattice. $c_{i,\sigma}^\dagger$ and $c_{i,\sigma}$ are the fermionic creation and annihilation operators, satisfying the canonical anticommutation relations $\{c_{i,\sigma}, c_{j,\sigma'}\} = 0$ and $\{c_{i,\sigma}^\dagger, c_{j,\sigma'}\} = \delta_{i,j}\delta_{\sigma,\sigma'}$. Here, $n_{i\alpha}$ is the occupation operator given as $n_{i\alpha} = c_{i\alpha}^\dagger c_{i\alpha}$. Another simple lattice model is the spinless fermionic model, where the spin degrees of freedom are also omitted and whose Hamiltonian reads

$$\mathcal{H} = \sum_{\langle i,j\rangle} (t\, c_i^\dagger c_j + \text{H.c.}) + \sum_{\langle i,j\rangle'} (t'\, c_i^\dagger c_j + \text{H.c.}) + \sum_{\langle i,j\rangle} V n_i n_j, \quad (3)$$

where $\langle i, j\rangle$ and $\langle i, j\rangle'$ denote nearest- and next-nearest neighbors, respectively, on a two-dimensional square lattice.

Boundary conditions play an important role for lattice models, as finite-size corrections take different scaling forms [55–58]. Usually, the fastest convergence towards the infinite-size limit can be achieved by enforcing periodic boundary condition in both directions. On the other hand, periodic boundary conditions impose serious drawbacks in the efficiency of the DMRG [10] because periodicity introduces long-range couplings between sites that are far away from each other in the one-dimensional MPS chain. In addition, when two-dimensional systems are mapped to the one-dimensional topology of the DMRG, local interactions also become nonlocal. Therefore, the Hamiltonian given by Eqs. (2) and (3) cannot be considered as a local one in a DMRG calculation. The corresponding MPO bond dimension for models in two and three spatial dimensions systematically increases with system size, albeit slower if compared to the general long-range model of Eq. (1). Nevertheless, the algorithmic nonlocality of two- and three-dimensional lattice models makes our approach, which is described below, increasingly beneficial in higher-dimensional lattices.

B. Joint optimization on the MPS and Grassman manifolds

The DMRG method provides the wave function in the MPS representation,

$$|\psi\rangle = \sum_{\substack{\alpha_1,\ldots,\alpha_d \\ \in\{0,\downarrow,\uparrow,\downarrow\uparrow\}}} A_{[1]}^{\alpha_1} \cdots A_{[d]}^{\alpha_d} |\alpha_1, \ldots, \alpha_d\rangle, \quad (4)$$

where the component tensors are $A_{[i]}^{\alpha_i} \in \mathbb{C}^{D_{i-1}\times D_i}$, with bond dimensions D_i, and $D_0 = D_d = 1$. Although every state vector can be factorized to an MPS form by applying consecutive Schmidt decompositions [59,60] using sufficiently large bond dimensions, the sufficient dimensions scale exponentially with the system size in the generic case. The restriction of the bond dimensions to a fixed value D confines the possible state vectors to a submanifold of the full state space. We can then approximate an eigenstate of the Hamiltonian

(1) within this submanifold by the DMRG algorithm, which, being an alternating least-squares method, optimizes the entries of the MPS tensors $A_{[i]}$ iteratively [12,43,61–63], leading to a variational treatment of the eigenvalue problem of the Hamiltonian (1).

Utilizing a single-particle unitary mode transformation $U \in \mathrm{U}(N)$, a linear transformation of a set of fermionic annihilation operators $\{c_{i,\sigma}\}$ to a new set $\{d_{i,\sigma}\}$ satisfying the canonical anticommutation relations can be obtained, i.e., $c_{i,\sigma} = \sum_{j=1}^{d} U_{i,j,\sigma} d_{j,\sigma}$. We note that in the presented system, it is not necessary to use different unitaries for spin up and down, $U_{i,j,\uparrow} = U_{i,j,\downarrow}$; however, the implementation is applicable for the unrestricted case too. Under this transformation, the many-body representation $G(U)$ can also be expressed on the Fock space [37], by which a fermionic wave function $|\psi(\mathbb{I})\rangle$ transforms to $|\psi(U)\rangle = G(U)^{\dagger}|\psi(\mathbb{I})\rangle$ and the Hamiltonian written in terms of the transformed modes by $H(U) = G(U)^{\dagger} H G(U)$. We note that generic Hamiltonians of the form given by Eq. (1) keep their form under such transformation, while the simpler Hamiltonians of Eqs. (2) and (3) lose their simpler structure and are transformed again to the generic form (1).

In the course of our implementation of the DMRG algorithm, the unitary U is constructed iteratively from two-mode unitary operators by sweeping through the network. This implementation has the benefit that the optimal unitaries can be constructed parallel to the DMRG optimization of the eigenstate. At each micro-iteration step, the half-Rényi block entropy $S_{1/2}(\rho_{\{1,2,\ldots,k\}}) = 2\ln(\mathrm{Tr}\sqrt{\rho_{\{1,2,\ldots,k\}}})$ is minimized by a two-mode rotation. (Here, $\rho_{\{1,2,\ldots,k\}}$ is the density operator of the first k modes [64,65].) In practice, when turn to numerical simulation including mode optimization, it is favorable not to transform the operators themselves to keep robustness. Rather, it is practical to perform mode optimization in terms of the parameters $T_{ij}^{\alpha\beta}$ and $V_{ijkl}^{\alpha\beta\gamma\delta}$ in the Hamiltonian (1) [15].

The optimization based on local (two-mode) unitaries often converges to a suboptimal local minima, and therefore we combine it with a global reordering of modes also [37]. At the end of the last DMRG sweep, the one-mode entropies [66] s_i, the two-mode mutual information [67] $I_{i,j} := s_i + s_j - s_{i,j}$, the total correlation [68] $I_{\mathrm{tot}} = \sum_i s_i$, the correlation distance $I_{\mathrm{dist}} = \sum_{i,j} I_{i,j}|i-j|^2$, the one-particle reduced density matrix $\gamma_{i,j} = \langle c_j^{\dagger} c_i \rangle$, and the occupation number distribution $\langle n_i \rangle$ are calculated (where $i, j \in \{1, \ldots, d\}$). Here, $s_i = -\mathrm{Tr}(\rho_i \ln \rho_i)$ and $s_{i,j} = -\mathrm{Tr}(\rho_{i,j} \ln \rho_{i,j})$ are the von Neumann entropies of the one- and two-mode reduced density operators ρ_i and $\rho_{i,j}$ [69]. The eigenvalues and eigenvectors of the one-particle reduced density matrix γ define the natural occupation numbers λ_i and the natural orbitals (NO), respectively. An optimized ordering of modes along the tensor network is calculated from the mutual information $I_{i,j}$, using the Fiedler vector approach [70]; a new complete active space vector is calculated from the entropies s_i for the dynamically extended active space (DEAS) procedure [66]; and a new Hartree-Fock (HF) reference configuration is calculated from the occupations $\langle n_i \rangle$. These, together with the final rotated interaction matrices, are all used as inputs for the subsequent mode transformation macro iteration. To reach sufficient convergence of the optimized basis, several tens of macro iterations are performed, and each macro iteration contains 5–10 sweeps of DMRG and two-site unitary micro iterations with low-to-medium values of bond dimensions, $D \lesssim 256$. After an acceptable optimized mode set is found, high-precision (large-D) DMRG calculations are performed without further mode optimization to find eigenstates of the Hamiltonian with high accuracy. For more details on mode transformation and its applications we guide readers to previous works (see Refs. [37,38,71–73]). To fix the notation, we use D_{mo} and D_{opt} to refer to the bond dimension used during and after the mode optimization procedure, respectively.

III. SPINLESS FERMION MODEL

In this section, we present numerical results together with scaling analysis for the half-filled spinless fermion model [Eq. (3)] on a torus geometry including nearest- and next-nearest-neighbor hopping and nearest-neighbor Coulomb interactions. This simple model already imposes serious challenges to understand the nature of emerging quantum phases and phase transitions. When nearest-neighbor hopping is neglected, i.e., $t' = 0$ for half filling, the charge density wave order parameter C_{CDW} takes on a finite value for arbitrary small Coulomb interaction, as has already been shown in Ref. [38]. However, when next-nearest-neighbor hopping is also included, the two-dimensional Fermi surface becomes distorted and interesting physics emerges. In general, for small perturbations, the Fermi surface changes and umklapp scattering dies out, but the so-called perfect nesting survives, i.e., one segment of the Fermi surface can be connected to another segment of the Fermi surface via a reciprocal lattice vector [74]. For stronger interactions, however, even perfect nesting is destroyed. The nature of the Mott-Hubbard transition in the absence of umklapp scattering even in one spatial dimension is a subject of active research and has been explored only recently [75], in the context of the $1/r$-Hubbard model. In this model, there is only a single Fermi point and the gap opens linearly at a finite Coulomb interaction. When both local and nearest-neighbor Coulomb interactions are included, a Luttinger-liquid phase emerges in an extended region between the charge density wave (CDW) insulator and the Mott-Hubbard insulator phases [76]. Analogously, the two-dimensional $t - t' - V$ spinless model is also expected to possess a rich phase diagram.

In the current work, we focus mainly on the technical aspects of our numerical procedure. Thus, only the charge density wave order parameter and the quantum information motivated half-chain block entropy will be analyzed for a selected finite value of $t' = 0.4$ as a function of V. In addition, for a detailed scaling analysis of the computational complexity, we fix further model parameters to $t = 1$, $t' = 0.4$, and $V = 0.8$, as this point lies in the vicinity of a quantum phase transition, which poses a great challenge to acquire accurate numerical results. A detailed analysis of the two-dimensional phase diagram, together with further analyses of the emerging quantum phases, will be part of a subsequent publication [77].

Our software package [45] has been developed for more than three decades and it has a very flexible structure, leading to several highly specialized modules to solve computational tasks in condensed-matter theory, quantum chemistry,

nuclear physics, and quantum information theory, among many others. The main code parts are fully scalable with the number of computational units even on heterogeneous architectures. This is due to the used mathematical models for parallelization, based on self-managed thread scheduling and hierarchical task managements, together with a virtual memory management model based on graph theory, which minimizes redundancy between input-output (IO) operations, substantially increases spatial locality, and eliminates the over-reliance on driver level memory mappings [39,40,78]. Computational performance using the message phrasing protocol (MPI), openMP, multithreading, and GPU are monitored according to various criteria, together with resource usage and communication overhead. Altogether, these form the basis of our analysis discussed in the following sections.

A. Computational complexity: Real space versus optimized basis

When the full bond dimension is kept, i.e., no truncation is enforced, a unitary transformation on the Hamiltonian leaves the energy spectrum intact. For truncated bond dimension, however, better or worse representations can be obtained via mode optimization in terms of computational complexity. The combination of the DMRG with the self-consistent field (SCF) method based on the energy gradient would be a straightforward procedure to obtain an optimized basis, but it requires the calculation of the one- and two-particle reduced density matrices that are, unfortunately, computationally very expensive [15,42]. The Rényi-entropy-based optimization scheme based on entanglement reduction, on the other hand, is a much cheaper method, but in general there is no guarantee that it ultimately leads to energy minimization. Nevertheless, for a large class of problems, including those studied in this work, it provides a powerful tool to attack high-dimensional strongly correlated problems [37,38,71–73].

In Fig. 3, we show the result of the Rényi-entropy-based mode optimization procedure for a fixed bond dimension value, $D_{\text{mo}} = 80$, for the half-filled 12×12 two-dimensional spinless fermion model with $t = 1, t' = 0.4$, and $V = 0.8$ on a torus geometry. The systematic decrease in the ground-state energy together with the dramatic drop in the block entropy is obvious. Note that the ground-state energy obtained by mode transformation with $D_{\text{mo}} = 80$ is already far below the one calculated via the real-space basis with $D_{\text{rs}} = 3072$ (see Fig. 1). Therefore, the DMRG calculation for a given D_{opt} value using the optimized basis, which basis has been determined beforehand using a bond dimension $D_{\text{mo}} \ll D_{\text{rs}}$, is expected to provide a significantly more accurate result than with the conventional real-space basis. A typical outcome of such calculation is shown in Fig. 4, i.e., the convergence of the ground-state energy and the block entropy as a function of DMRG sweeping. Although the block entropy takes much larger values for large D_{opt} compared to the profile optimized with a very limited D_{mo} value, i.e., it captures more correlations, the resulting profile, nevertheless, remains almost symmetric. The ground-state energy and the maximum value of the block entropy are shown in Fig. 5 for the 6×6, 8×8, and 10×10 lattice sizes as a function of D. This clearly shows that the optimized modes allow one to reach the same

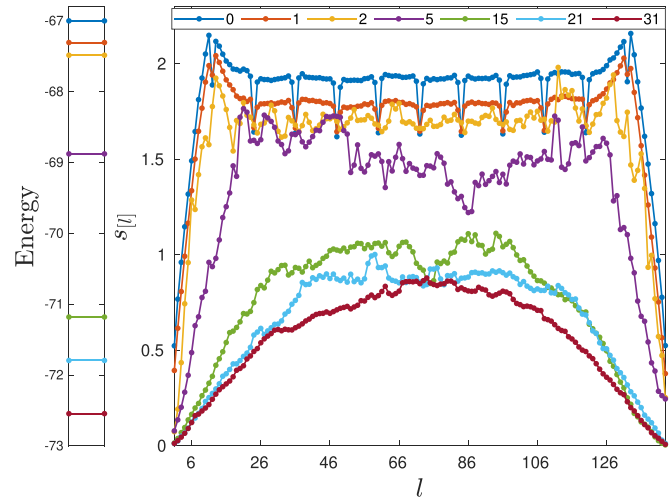

FIG. 3. Ground-state energy (left) and block entropy as a function of the left block size, ℓ (right), for some selected mode transformation macro-iteration cycles, 0,1,2,5,15,21,31, with fixed bond dimension $D_{\text{mo}} = 80$ for the half-filled 12×12 two-dimensional spinless fermionic model, with model parameters given in Fig. 1.

accuracy with significantly lower bond dimension values corresponding to much lower block entropy values.

Our results are in agreement with the previous findings of Refs. [38,71–73], but the more general form of the Hamiltonian after mode optimization, given by Eq. (1), imposes serious technical challenges. Although the block entropy and thus the DMRG/MPS bond dimension are tremendously reduced, the number of terms in the Hamiltonian and, consequently, the bond dimension of its matrix product operator (MPO) representation increases significantly [compare Eqs. (2) and (3) with Eq. (1)]. Since the overall scaling of the computational complexity of the DMRG for the general

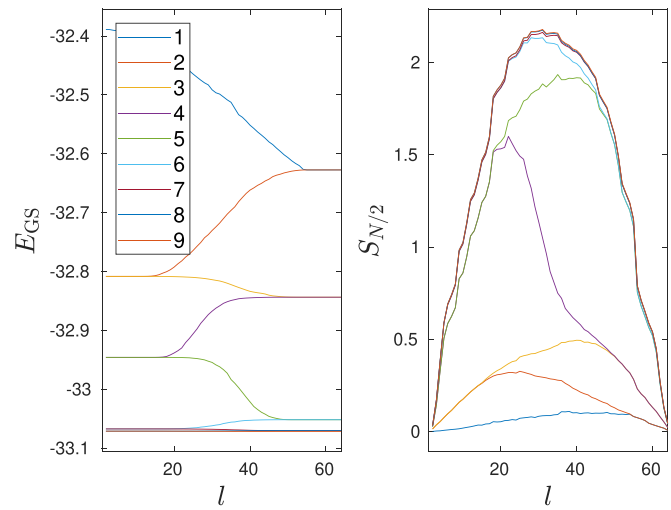

FIG. 4. Ground-state energy (left) and block entropy (right) as a function of left block size, ℓ, with increasing DMRG sweeps for the half-filled 8×8 two-dimensional spinless fermionic model parameters given in Fig. 1 using optimized basis, $D_{\text{l}}^{(\text{start})} = 1024$, $D_{\text{r}}^{(\text{start})} = 1024$, $D_{\text{opt}} = 4096$, and $\varepsilon_{\text{Davidson}} = 10^{-7}$.

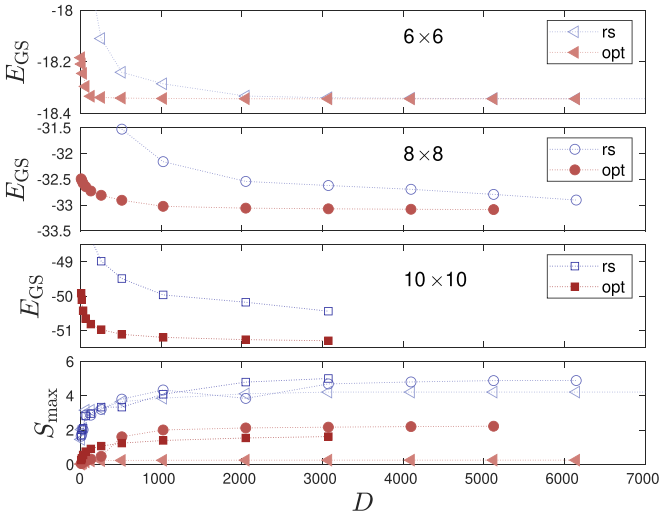

FIG. 5. Convergence of the ground-state energy and the block entropy as a function of bond dimension using the real-space and the optimized basis for 6×6, 8×8, and 10×10 lattices and for model parameters given in Fig. 1 and $\varepsilon_{\text{Davidson}} = 10^{-5}$.

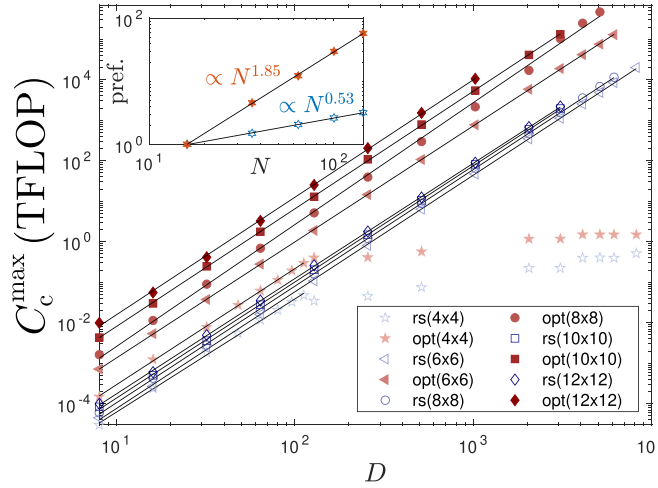

FIG. 6. Maximum value of the total computational complexity measured in TFLOP for diagonalization of the effective Hamiltonian as a function of bond dimension using the real-space and the optimized basis for model parameters given in Fig. 1 and $\varepsilon_{\text{Davidson}} = 10^{-5}$. The solid lines are the results of first-order polynomial fits leading to exponents $\nu_{\text{rs}}(4 \times 4) = 2.79$, $\nu_{\text{opt}}(4 \times 4) = 2.86$, $\nu_{\text{rs}}(6 \times 6) = 2.85$, $\nu_{\text{opt}}(6 \times 6) = 2.85$, $\nu_{\text{rs}}(8 \times 8) = 2.9$, $\nu_{\text{opt}}(8 \times 8) = 3.2$, $\nu_{\text{rs}}(10 \times 10) = 2.87$, $\nu_{\text{opt}}(10 \times 10) = 2.91$, $\nu_{\text{rs}}(12 \times 12) = 2.85$, and $\nu_{\text{opt}}(12 \times 12) = 2.90$. Inset: The scaling of the prefactor as a function of system size N, where the fitted exponents lead to 0.53 and 1.85 for the real-space and for the optimized basis, respectively.

form of the Hamiltonian given by Eq. (1) is D^3N^4, it remains mandatory to demonstrate if the reduction in D can overcompensate the increase in the exponent connected to N when the locality of the Hamiltonian is lost. Note that for the two- and higher-dimensional systems, the locality of the Hamiltonian is already lost when mapped to the one-dimensional DMRG topology, and thus the increase in the MPO bond dimension is less than for a one-dimensional model. Here we also remark that for a given DMRG sweep, the scaling is only D^3N^3 [48] and the extra factor related to the system size reflects the number of sweeps required to reach convergence. This latter one is highly problem dependent and the worst scenario holds for models obeying volume law [79].

In Fig. 6, the maximum value of the computational complexity (number of floating point operations), $C_c^{\max}$, measured via the diagonalization procedure in TFLOP (10^{12} FLOP), is summarized for the real-space and for the optimized basis as a function of bond dimension on a double-logarithmic scale for various system sizes and for the model parameters given in Fig. 5. Solid lines are the result of first-order polynomial fits. It is clearly seen that the data points lie along a straight line. Here we remark that for the 4×4 model, the exact solution is already recovered with $D = 2^7 = 128$ and thus a saturation in $C_c^{\max}$ becomes apparent for larger-D values up to $D = 2^{13}$. The fitted exponents for the real-space basis are in the range of 3 ± 0.2, confirming the expected D^3 scaling. Data points for the optimized basis also lie along straight lines corresponding to the fitted exponents in the range of 3 ± 0.2. Thus the overall scaling with D does not change due to mode optimization, as expected, but lines shift toward larger C_c values due to an increase in the MPO bond dimension. Switching from a real-space to mode optimized basis, the scaling of such prefactor is expected to change from $\sqrt{N}$ to N^2. In fact, as shown in the inset of Fig. 6, the fitted exponents of the prefactor, as a function of system size N, lead to 0.53 and 1.85 for the real-space and for the optimized basis, respectively. Therefore, naively, one could render the optimized basis behind the real-space basis, but after recalling that the same accuracy in energy can be reached with significantly lower-D values, this is no longer true. In fact, as can be seen in Fig. 1, the same accuracy in energy can be reached with $D_{\text{opt}} = 512$ as with $D_{\text{rs}} \simeq 3000$ for a 6×6 model and with $D_{\text{opt}} = 512$ as with $D_{\text{rs}} \simeq 5500$ for an 8×8 model, respectively [80]. Here, by D_{opt}, we refer to the bond dimension used after the mode optimization procedure. Mode optimization were performed at fixed bond dimension $D_{\text{mo}} = 80$. The saving in computational complexity significantly increases with system size; for example, for 10×10 and 12×12, this gets close to four orders of magnitude.

We conclude that the computational complexity for a given accuracy can be reduced by several orders of magnitude, depending on system size, via mode optimization. Since mode optimization is performed with low bond dimension, the computational complexity of the optimization procedure is negligible compared to the saving when the post-mode-transformation DMRG calculations are performed using the optimized basis. The data for 10×10 and 12×12 show that mode optimization can allow for accuracy that is far beyond the scope of real-space DMRG.

While in Fig. 3, for simplicity, a fixed bond dimension $D_{\text{mo}} = 80$ has been used for the mode optimization procedure, in practice, usually we start mode optimization with an even lower bond dimension value, performing a finite number of mode transformation macro-iteration cycles, each comprising DMRG calculations with a finite number of sweeps, and this procedure is repeated by systematically increasing the bond dimension [38,73]. This approach further reduces the

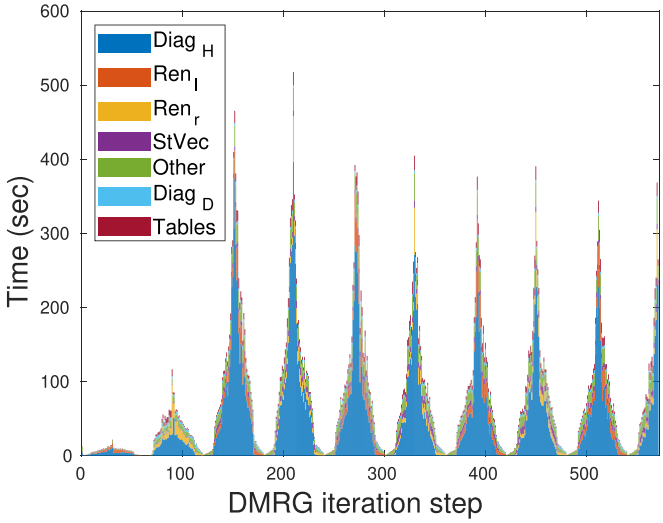

FIG. 7. Contribution of the most expensive functions to the total time measured in seconds as a function of the DMRG iteration step for model and DMRG parameters given in Fig. 1. The calculation has been performed on a dual AMD EPYC 7702 CPUs with 2×64 cores compiled with eight Nvidia A100-SXM4-40GB devices.

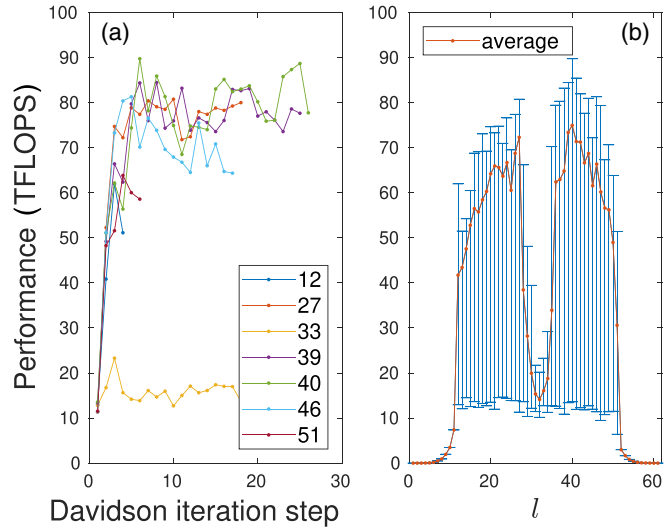

FIG. 8. (a) Performance in TFLOPS measured via the eight GPU accelerated Davidson diagonalization procedure for some selected DMRG iterations steps of the fifth DMRG sweep for the model and DMRG parameters given in Fig. 1. Numbers in the legend stand for the size of the left block size, l. (b) Minimum, maximum, and average performance measured in TFLOPS for the fifth DMRG sweep as a function of left block size, l, for the model and DMRG parameters given in Fig. 1.

calculation time of the optimization procedure and speeds up convergence at initial sweeps by avoiding sticking to local minima of the block entropies. A typical setting that we used is $D_{\text{mo}} = [8, 16, 64, 80]$, $\text{Sweep}_{\text{mo}} = [11, 11, 9, 7]$, and $\text{Iter}_{\text{mo}} = [9, 9, 7, 5]$.

B. Performance: Real space versus optimized basis

Our arguments and conclusion presented in Sec. III A can be further improved if we also consider the underlying power in state-of-the-art hardware and software technologies, i.e., massive parallelization via message passing interface (MPI) and graphical processing units (GPUs). Here we demonstrate that our hybrid CPU-multiGPU solution [39,40] has the potential to significantly reduce the effective exponent connected to D, thus further reducing the overall wall time of the calculations drastically.

First we present a detailed performance analysis for a single DMRG calculation for the 8×8 lattice, for the model parameters given in Fig. 1, using the Rényi-entropy optimized basis, $D_{\text{l}}^{(\text{start})} = 1024$, $D_{\text{r}}^{(\text{start})} = 1024$, $D = 4096$, and $\varepsilon_{\text{Davidson}} = 10^{-7}$. From a technical point of view, for the spinless fermion model, only the total particle number can be used as a conserved quantum number, and thus the underlying DMRG matrix and tensor algebra decompose into fewer but larger sectors compared to spinful models [39,40]. In Fig. 7, we show the contribution of the most expensive functions to the total time measured in seconds as a function of the DMRG iteration step. Here, Diag_H denotes the time spent on the diagonalization of the effective Hamiltonian, and Ren_{l} and Ren_{r} are the renormalization step for the left and right block via the forward and backward sweep, respectively. The overhead to prepare the quantum number decomposed based matrix and tensor algebra is labeled as Tables; the diagonalization of the reduced density matrix, i.e., the singular value decomposition (SVD) step, by Diag_D; the wave-function transformation to generate the starting vector for the diagonalization step by StVec; and all other utility components by Other. The calculation has been performed on dual AMD EPYC 7702 CPUs with 2×64 cores compiled with eight Nvidia A100-SXM4-40GB devices. It is clear that even after full GPU parallelization, 80–90% of the total time in a given DMRG iteration step is allocated for the diagonalization of the effective Hamiltonian. We also note that a factor of 25–30 speedup is already gained with respect to a fully parallelized CPU only implementation [39]. Therefore, in the following, we restrict our analysis to the diagonalization procedure. Here we remark that via the sweeping procedure, the renormalized operators, contracted network components, must be stored and reused in subsequent iteration steps, which IO time can be significant if disks are used for that purpose.

To provide more insight, in Fig. 8(a), we present the performance measured in TFLOPS via the Davidson diagonalization procedure for some selected DMRG iteration steps of the fifth DMRG sweep for the model and the DMRG parameters given in Fig. 1. The numbers in the legend stand for the left block size, l. It is clearly seen that the number of matrix-vector operations fluctuates heavily depending on the DMRG iteration steps, while the performance for a given DMRG iteration step reaches high values with increasing Davidson steps. Here we remark that for the first few Davidson iteration steps, the IO overhead to transfer data from host to device (H2D IO) and the overhead for our dynamic scheduler to be optimized [40] reduces the performance to lower initial values. The corresponding minimum, maximum, and average performance values are collected and presented in Fig. 8(b). Note that the FP64 limit of Nvidia [81], 76 TFLOPS, is reached by most of the iteration steps and the average performance is affected

only marginally by the low initial performance values of the Davidson procedure.

To complete our analysis, in Fig. 2, we compare the largest performance and wall time as a function of DMRG bond dimension for calculations obtained with the real-space and with the optimized basis. For the real-space DMRG, the MPO bond dimension is much lower, and therefore the efficient matrix and tensor algebra with SBATCH [40] cannot be utilized as efficiently as for the optimized basis. This leads to a relatively poor performance as, even for the eight GPU accelerated diagonalization procedure, the largest performance is about 10–14 TFLOPS (see Fig. 2, left panel). In contrast to this, for the optimized basis, the underlying GPU power can be utilized very efficiently. The sharp increase in performance up the FP64 ceiling is obvious, while a saturation for larger-D values becomes apparent. Nevertheless, our results already demonstrate the utilization of the highly specialized Nvidia tensor core units (TCUs) [40]. For larger system sizes, the sudden increase in performance gets even more pronounced, as has also been demonstrated for quantum chemical calculations [40].

The significant difference in the reachable performance for the real-space and for the optimized basis also determines the total wall time. The total time spent on the eight GPU accelerated diagonalization of the effective Hamiltonian including host to device (H2D) and device to host (D2H) IO measured for nine DMRG sweeps is presented in the right panel of Fig. 2 on a double-logarithmic scale as a function of the bond dimension for the $n \times n$ half-filled spinless model with $n = 6, 8, 10, 12$, with the model and DMRG parameters given in Fig. 1. It is apparent in the figure that via massive parallelization, the total diagonalization time for nine sweeps for a given D value and system size is almost the same for the real-space and for the optimized basis. For the 6×6 lattice size, the calculations are even faster for the optimized basis; for 8×8, they are almost the same; while for large-N calculations with the same D, it gets more expensive, and thus the difference increases in system size. Recalling, however, that for a 6×6 lattice, the same accuracy is obtained with $D = 512$ for the optimized basis taking two minutes as with $D = 3000$ for the real-space basis taking 20 minutes, one order of magnitude reduction can already be achieved in wall time for this small system size. Similarly, for 8×8, the corresponding numbers are $D = 512$ and seven minutes for the optimized basis, while the matching calculation in energy with $D \simeq 5500$ takes about 200 minutes, i.e., the reduction of wall time is almost two orders of magnitude. For larger system sizes, such comparison is not possible as the real-space approach has been unable to reach the same accuracy as the optimized basis has already reached for the lowest considered bond dimension, $D_{\text{opt}} = 64$. Nevertheless, our data indicate that the speedup is over two orders of magnitude in these cases. Note that data are measured for the eight GPU accelerated implementation for all system sizes.

Here we also remark that the saturation in performance will be lifted and even eliminated via our new kernel also utilizing the Nvidia fast device to device (D2D) NVLINK and message phrase interface (MPI) [78]. Therefore, further reduction in wall time for the optimized basis can be achieved for a much broader range of the bond dimension.

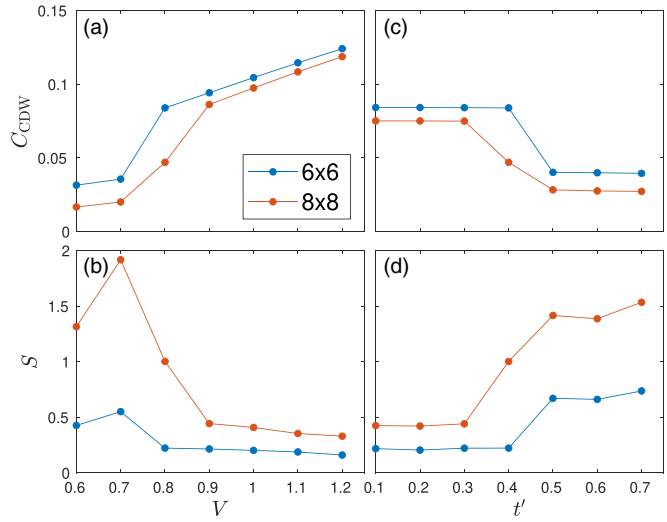

FIG. 9. (a) Charge density wave order parameter C_{CDW} and (b) half-chain block entropy obtained via the optimized basis for the 6×6 and for the 8×8 spinless fermion models on a torus geometry as a function of V for finite $t' = 0.4$. (c),(d) The same, but as a function of t' for $V = 0.8$.

C. Quantum phase transition at finite Coulomb interaction

In addition to ground-state energies, our procedure also allows us to calculate the expectation values of operators with much higher accuracy than can be done in the real-space basis. As discussed in Ref. [38], correlation functions, and one- and two-body reduced density matrices, can be obtained very accurately in the optimized basis and can then be transformed back to the original, real-space basis via a sequence of rotations using unitary matrices calculated and stored during mode optimization. Here we remark that the calculation of the computationally very demanding 4-index tensor of the two-body reduced density matrix, $\Gamma_{ijkl} = \langle c_i^\dagger c_j^\dagger c_k c_l \rangle$, can also be tremendously accelerated via our hybrid CPU-multiGPU kernels. By taking a linear combination of proper matrix elements of the back-rotated reduced density matrices, the CDW order parameter, $C_{\text{CDW}} = \frac{1}{N^2} \langle (n_{\text{even}} - n_{\text{odd}})^2 \rangle$, can, for example, be calculated, where $n_{\text{even/odd}}$ stands for the total particle number in the even/odd sublattices. This quantity is shown in Fig. 9, together with the half-chain block entropy for the optimized basis as a function of V for finite $t' = 0.4$ and as a function of t' for $V = 0.8$ for the 6×6 and 8×8 spinless fermion model on torus geometry. A sudden change in C_{CDW} is apparent at a finite value of V that shifts slightly from $V_c = 0.7$ to $V_c = 0.8$ with increasing system size. The large jump in the block entropy also indicates the anomalous behavior at finite V_c. Similar behavior is observed as a function of t' at $t' \simeq 0.4$ for $V = 0.8$. Further numerical and analytic analysis of the phase diagram is beyond the scope of the present work due to its very complex structure and will be part of a subsequent work [77].

IV. HUBBARD MODEL

In this section, we present scaling analysis and benchmark results obtained via the mode optimized non-Abelian SU(2)

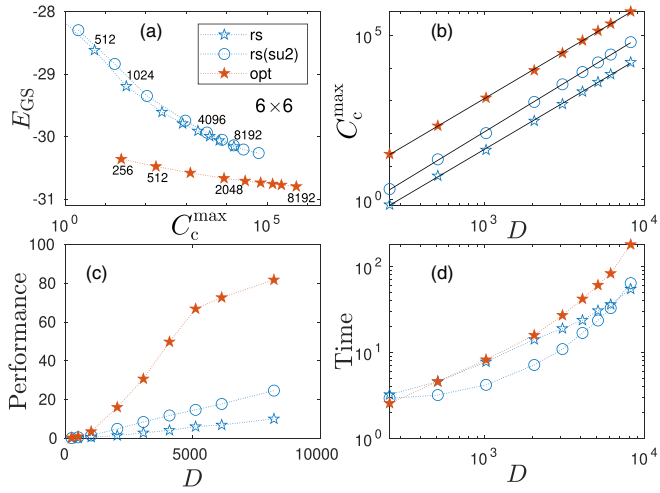

FIG. 10. Ground-state energy as a function of (a) computational complexity, (b) computational complexity in TFLOP, (c) largest performance in TFLOPS, and (d) total wall time of nine sweeps in minutes, as a function of bond dimension obtained via the real-space basis with U(1) and SU(2) spin symmetry and by optimized modes with U(1) symmetry for the half-filled 6×6 Hubbard model at $U = 4$ on a torus geometry. The fitted exponents in (b) are $\nu_{rs} = 2.96$ and $\nu_{opt} = 2.87$. For the SU(2) spin adapted calculations, D stands for the number of renormalized multiplets.

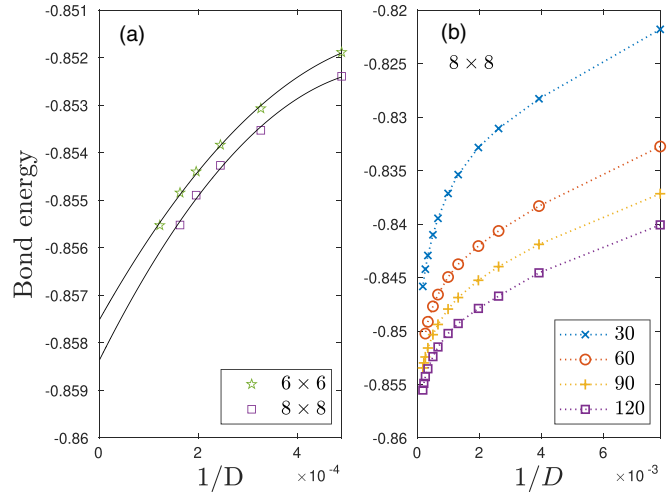

FIG. 11. (a) Bond energy as a function of inverse bond dimension for the optimized modes with U(1) symmetry for the half-filled 6×6 and 8×8 Hubbard models at $U = 4$ on a torus geometry. The solid lines are second-order polynomial fits. (b) Bond energy as a function of inverse bond dimension using the optimized basis via 30, 60, 90, and 120 macromode optimization cycles for the 8×8 system size.

spin adapted hybrid CPU-multiGPU DMRG method for the half-filled two-dimensional Hubbard model [Eq. (2)] on a torus geometry. Besides the total charge, the spin quantum number can also be kept, which leads to a significant increase in the number of independent tasks and to smaller sector sizes. In addition, the SU(2) spin symmetry can also be utilized, leading to a tremendous increase in the effective performance without additional computational overhead [40]. In this latter case, the DMRG matrix and tensor algebra are reformulated according to non-Abelian quantum numbers, and thus the bond dimension D stands for the number of renormalized multiplets. When the corresponding U(1) bond dimension is also indicated, it is explicitly denoted as $D^{\text{eff}}_{U(1)}$.

Here, we perform mode optimization using U(1) symmetries only, by either enforcing the same unitary for both spin components, and thus preserving the SU(2) symmetry of the Hamiltonian, or by optimizing two different unitary matrices for the two spin components independently. We remark that the SU(2) spin adapted version of the mode optimization procedure is part of our current developments. The post-mode-transformation DMRG calculations with large-D values can be performed using U(1) and SU(2) symmetries. In Fig. 10, we summarize our results for the 6×6 half-filled Hubbard model at $U = 4$ using the real-space basis with U(1) and SU(2) spin symmetries and via the mode optimized basis using U(1) symmetries only. The mode optimization has been performed iteratively up to $D_{mo} = 128$. The $U = 4$ interaction strength was selected as there are several benchmark results available for torus geometry in the literature obtained by various methods [82–84]. First we remark that although the SU(2) spin adapted version for the real-space basis provides lower energy than the U(1) implementation for a given bond dimension [Fig. 10(a)], it corresponds to enlarged computational complexity [Fig. 10(b)]. This is due to the larger number of, and more dense, sectors. Nevertheless, it can be parallelized more efficiently resulting in larger performance values as a function of D [Fig. 10(c)]. Therefore, the overall wall time can be scaled below the U(1) implementation for a broad range of the bond dimension [Fig. 10(d)].

When mode optimization is performed, we have found that a significantly larger reduction in block entropy, together with a significantly lower energy, can be achieved when the unitary matrices are optimized independently for the two spin components. Although the SU(2) symmetry is broken for such unrestricted optimization procedure, the low ground-state energy could not be recovered via the post-mode-transformation DMRG calculations even using SU(2) symmetry and large bond dimension when the same angles were enforced during the optimization procedure. We note that with $D_{opt} = 256$, the DMRG provides lower ground-state energy for the optimized basis than for the real-space basis with $D = 8192$ together with SU(2) spin symmetry, i.e., for an effective U(1) bond dimension around 25 000 [Fig. 10(a)]. In addition, a very reliably scaling with inverse bond dimension can be achieved for the optimized modes (see Fig. 11), leading to an extrapolated bond energy, $E_{GS}/N = -0.8575(2)$. In Table I, we summarize our results for the bond energy together with reference data from the literature obtained by various methods. It is evident that our method provides bond energy in the range of the reference data sets and, using optimized modes, a very reliable $1/D$ extrapolation can be achieved. Our result is, in fact, in close agreement with the bond energy obtained by auxiliary-field quantum Monte Carlo (AFQMC) [82]. We also remark that unlike DMRG, most of the reference methods are nonvariational.

We close our analysis by providing benchmark data for the 8×8 half-filled Hubbard model (see Fig. 11) together with

TABLE I. Bond energy for the half-filled 6×6 Hubbard model for $U = 4$ on a torus geometry obtained by various methods. Our notation reads DMRG[D_{rs}, Symmetry] and DMRG[D_{mo}, D_{opt}, Symmetry], respectively. The maximum effective bond dimension is also included in the last column. Note that initiator similarity transformed full configuration interaction quantum Monte Carlo (iST-FCIQMC), AFQMC, projected DMET (p-DMET), and DMET are nonvariational methods.

Bond energy	Method	$\max(D^{\text{eff}}_{U(1)})$
−0.85625(30)	iST-FCIQMC [83]	
−0.85736(25)	AFQMC [82]	
−0.79703	unrestricted Hartree-Fock [84]	
−0.86792	p-DMET [84]	
−0.86856	DMET [84]	
−0.83894	DMRG[6140,SU(2)]	19748
−0.84067	DMRG[8192,SU(2)]	25088
−0.84333	DMRG[128,256,U(1)]	256
−0.85553	DMRG[128,8192,U(1)]	8192
−0.8575(2)	DMRG extrapolated	

extrapolation leading to $E_{\text{GS}}/N = -0.8584(2)$. This value is close, but off by 10^{-3} compared to the bond energy obtained by AFQMC [82]. Note, however, that by increasing D_{opt} further, one could get even closer to the error-free solution and obtain better extrapolation; it would simply require more computational resources. On the other hand, the rate of convergence significantly depends on the quality of the optimized basis, i.e., there is an initial error obtained by mode optimization with fixed rank D_{mo} if a fully converged data set is not used. This affects the $1/D_{\text{opt}}$ extrapolation as well, i.e., lower energies with the same D_{opt} can be reached if a better basis is used. This is demonstrated in Fig. 11(b) for the 8×8 system size using optimized modes obtained via 30, 60, 90, and 120 macromode optimization cycles.

To gain more insights into how mode optimization is influenced by the nature of the quantum many-body wave function, we have repeated our analysis for attractive interactions. By taking a negative value, $U = -4$, we find that the two angles during mode optimization for the two modes take on very close values. This is not a surprise as for attractive interaction, particles with opposite spins like to form pairs, which is in favor of preserving SU(2) symmetry. Accordingly, the expectation value of the occupation number operator of both the spin-up and spin-down components oscillates with the same profile along the lattice sites. In contrast, the ground state of the repulsive ($U > 0$) model is characterized by antiferromagnetic correlations, in which electrons of differing spin are in different sublattices (as also confirmed by our numerics through the $\langle n_{i\uparrow/\downarrow} \rangle$ expectation values). Enforcing a spin-independent spatial form of fermionic modes during the mode optimization is not optimal for such states, leading to reduced performance for the repulsive case. We have found, however, that there is very stable and fast convergence in the attractive case with enforced SU(2) symmetry and obtain much lower ground-state energies for a given bond dimension. Therefore, we conclude that the optimal unitary transformations that correspond to independent modes depend on the structure of the ground state. Nevertheless, our procedure works in general.

Further algorithmic developments to boost the convergence of the mode optimization protocol and to lower such initial error, together with additional analysis including other fillings and topologies, are under investigation and will be part of our subsequent work. Nevertheless, our procedure is variational, free of sign problem, and thus can be applied at general fillings and for frustrated models too. Finally, an important advantage of the mode optimized basis is that it can be parallelized very efficiently, reaching again 90 TFLOPS in our benchmark calculations [Fig. 10(c)] and reducing wall time to the same range as the real-space calculation with U(1) symmetry [Fig. 10(d)]. In addition, improving performance further via our CPU-multiGPU kernel together with the MPI protocol and fast Nvidia D2D NVLINK will reduce the total execution time even more for larger system sizes and D values.

V. CONCLUSION

In this work, we have presented a hybrid numerical approach to simulate quantum many-body problems on two spatial dimensional quantum lattice models via the non-Abelian *ab initio* version of the density matrix renormalization group method on state-of-the-art high-performance computing infrastructures. We demonstrated for the two-dimensional spinless fermion model and for the Hubbard model on torus geometry that altogether, several orders of magnitude in computational time can be saved for a given accuracy by performing calculations on an optimized basis and by utilizing hybrid CPU-multiGPU parallelization. At least an order of magnitude reduction in computational complexity, measured directly in FLOP, results from mode optimization, while an order of magnitude reduction in wall time is achieved by massive parallelization. A detailed scaling analysis of the obtained performance measured in FLOPS as a function of matrix ranks and as a function of system size up to 12×12 lattice topology for the spinless model revealed that more efficient parallelization can be gained for the optimized modes. As a result, even though the computational complexity increases via mode optimization, the overall wall time can be reduced to similar magnitude as for the real-space basis for a broad range of bond dimension values.

For the Hubbard model, we also analyzed the scaling of the various quantities for the real-space basis using U(1) symmetry only and for the SU(2) spin adapted version. For the mode optimization, we have found that a much better basis can be obtained by optimizing unitary matrices for two spin components independently. After such unrestricted optimization protocol, we managed to provide new DMRG lower bond energy values and reliable extrapolations for the nontruncated limit with respect to previous attempts which are, in fact, in good agreement with the AFQMC reference data set.

Our numerical results, however, show that mode optimization is a more delicate issue for the Hubbard model and it requires further developments to boost convergence. Nevertheless, mode optimization is crucial to obtain a quasi-optimal basis, i.e., to reduce initial error for an optimization on a fixed rank manifold. It is worthwhile to note an imbalance in current developments in information technology, namely, the number of computational units increases at a much higher rate than the available size of memory on HPC infrastructures. Therefore,

it is more important to reduce the rank of the matrices and tensors than to minimize the computational complexity. The latter can be handled more easily via massive parallelization.

Finally, our approach is variational, free of sign problem, and can be applied to more general fillings and topologies and for frustrated systems as well. It can also be further improved by utilizing message phrase passing (MPI) protocols and Nvidia fast device-to-device NVLINK communication. These latter aspects are part of our current developments.

ACKNOWLEDGMENTS

The authors acknowledge useful discussions with Tamás Kozsik, Reinhard Noack, Florian Gebhard, and Jenő Sólyom. This work has been supported by the Hungarian National Research, Development and Innovation Office (NKFIH) through Grants No. K134983 and No. TKP2021-NVA-04, by the Quantum Information National Laboratory of Hungary. Ö.L. acknowledges financial support by the Hans Fischer Senior Fellowship programme funded by the Technical University of Munich - Institute for Advanced Study and by the Center for Scalable and Predictive Methods for Excitation and Correlated phenomena (SPEC), funded as part of the Computational Chemical Sciences Program FWP 70942 by the US Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, Division of Chemical Sciences, Geosciences, and Biosciences at Pacific Northwest National Laboratory. M.A.W. has also been supported by the Janos Bolyai Research Scholarship of the Hungarian Academy of Sciences. The simulations were performed on the national supercomputer HPE Apollo Hawk at the High Performance Computing Center Stuttgart (HLRS) under Grant No. MPTNS/44246, and at the Wigner Scientific Computational Laboratory (WSCLAB).

[1] B.-X. Zheng, C.-M. Chung, P. Corboz, G. Ehlers, M.-P. Qin, R. M. Noack, H. Shi, S. R. White, S. Zhang, and G. K.-L. Chan, Science **358**, 1155 (2017).
[2] D. P. Arovas, E. Berg, S. A. Kivelson, and S. Raghu, Annu. Rev. Condens. Matter Phys. **13**, 239 (2022).
[3] M. Qin, T. Schäfer, S. Andergassen, P. Corboz, and E. Gull, Annu. Rev. Condens. Matter Phys. **13**, 275 (2022).
[4] S. R. White and R. M. Noack, Phys. Rev. Lett. **68**, 3487 (1992).
[5] U. Schollwöck, Rev. Mod. Phys. **77**, 259 (2005).
[6] R. M. Noack and S. R. Manmana, AIP Conf. Proc. **789**, 93 (2005).
[7] K. A. Hallberg, Adv. Phys. **55**, 477 (2006).
[8] F. Verstraete, T. Nishino, U. Schollwöck, M. C. Bañuls, G. K. Chan, and M. E. Stoudenmire, Nat. Rev. Phys. **5**, 273 (2023).
[9] S. Liang and H. Pang, Phys. Rev. B **49**, 9214 (1994).
[10] R. M. Noack, S. R. White, and D. J. Scalapino, in *Computer Simulation Studies in Condensed-Matter Physics VII: Proceedings of the Seventh Workshop Athens, GA, USA, 28 February–4 March 1994*, edited by D. P. Landau, K. K. Mon, and H.-B. Schüttler (Springer, New York, 1994), pp. 85–98.
[11] S. R. White and D. J. Scalapino, Phys. Rev. Lett. **91**, 136403 (2003).
[12] F. Verstraete and J. I. Cirac, arXiv:cond-mat/0407066.
[13] F. Verstraete, V. Murg, and J. Cirac, Adv. Phys. **57**, 143 (2008).
[14] R. Orús, Ann. Phys. **349**, 117 (2014).
[15] V. Murg, F. Verstraete, Ö. Legeza, and R. M. Noack, Phys. Rev. B **82**, 205105 (2010).
[16] L. Tagliacozzo, G. Evenbly, and G. Vidal, Phys. Rev. B **80**, 235127 (2009).
[17] K. Gunst, F. Verstraete, S. Wouters, Ö. Legeza, and D. Van Neck, J. Chem. Theory Comput. **14**, 2026 (2018).
[18] K. Gunst, F. Verstraete, and D. V. Neck, J. Chem. Theory Comput. **15**, 2996 (2019).
[19] N. Nakatani and G. K.-L. Chan, J. Chem. Phys. **138**, 134113 (2013).
[20] R. Orús, Nat. Rev. Phys. **1**, 538 (2019).
[21] A. Himeda, T. Kato, and M. Ogata, Phys. Rev. Lett. **88**, 117001 (2002).
[22] C.-C. Chang and S. Zhang, Phys. Rev. Lett. **104**, 116402 (2010).
[23] F. F. Assaad and I. F. Herbut, Phys. Rev. X **3**, 031010 (2013).
[24] J. Kolorenč and L. Mitas, Rep. Prog. Phys. **74**, 026502 (2011).
[25] Z. Li, J. Li, N. S. Dattani, C. J. Umrigar, and G. K.-L. Chan, J. Chem. Phys. **150**, 024302 (2019).
[26] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, Rev. Mod. Phys. **68**, 13 (1996).
[27] G. Kotliar, S. Y. Savrasov, K. Haule, V. S. Oudovenko, O. Parcollet, and C. A. Marianetti, Rev. Mod. Phys. **78**, 865 (2006).
[28] G. Knizia and G. K.-L. Chan, J. Chem. Theory Comput. **9**, 1428 (2013).
[29] J. Eisert, M. Cramer, and M. B. Plenio, Rev. Mod. Phys. **82**, 277 (2010).
[30] E. M. Stoudenmire and S. R. White, Annu. Rev. Condens. Matter Phys. **3**, 111 (2012).
[31] P. Calabrese and J. Cardy, J. Stat. Mech.: Theory Expt. (2004) P06002.
[32] U. Schollwöck, Ann. Phys. **326**, 96 (2011).
[33] Y.-F. Jiang, J. Zaanen, T. P. Devereaux, and H.-C. Jiang, Phys. Rev. Res. **2**, 033073 (2020).
[34] Y. Huang and D. N. Sheng, Phys. Rev. X **12**, 031009 (2022).
[35] J. Motruk, M. P. Zaletel, R. S. K. Mong, and F. Pollmann, Phys. Rev. B **93**, 155139 (2016).
[36] G. Ehlers, S. R. White, and R. M. Noack, Phys. Rev. B **95**, 125125 (2017).
[37] C. Krumnow, L. Veis, Ö. Legeza, and J. Eisert, Phys. Rev. Lett. **117**, 210402 (2016).
[38] C. Krumnow, L. Veis, J. Eisert, and Ö. Legeza, Phys. Rev. B **104**, 075137 (2021).
[39] A. Menczer and Örs Legeza, arXiv:2309.16724.
[40] A. Menczer and Örs Legeza, arXiv:2305.05581.
[41] S. R. White, Phys. Rev. Lett. **69**, 2863 (1992).
[42] G. K.-L. Chan, J. J. Dorando, D. Ghosh, J. Hachmann, E. Neuscamman, H. Wang, and T. Yanai, in *Frontiers in Quantum Systems in Chemistry and Physics*, Progress in Theoretical Chemistry and Physics, Vol. 18, edited by S. Wilson, P. J. Grout, J. Maruani, G. Delgado-Barrio, and P. Piecuch (Springer, Netherlands, 2008).

[43] S. Szalay, M. Pfeffer, V. Murg, G. Barcza, F. Verstraete, R. Schneider, and Ö. Legeza, Intl. J. Quantum Chem. **115**, 1342 (2015).

[44] A. Baiardi and M. Reiher, J. Chem. Phys. **152**, 040903 (2020).

[45] Ö. Legeza, L. Veis, and T. Mosoni, DMRG-Budapest, A program for condensed matter, quantum chemical, and nuclear shell DMRG calculations (unpublished).

[46] Operators $c^{\dagger}_{i\alpha}$ and $c_{i\alpha}$ can also denote spin ladder operators in our DMRG code, but fermionic mode optimization can only be employed for fermionic systems.

[47] T. Xiang, Phys. Rev. B **53**, R10445 (1996).

[48] S. R. White and R. L. Martin, J. Chem. Phys. **110**, 4127 (1999).

[49] S. Knecht, Ö. Legeza, and M. Reiher, J. Chem. Phys. **140**, 041101 (2014).

[50] J. Dukelsky and S. Pittel, Rep. Prog. Phys. **67**, 513 (2004).

[51] Ö. Legeza, L. Veis, A. Poves, and J. Dukelsky, Phys. Rev. C **92**, 051303(R) (2015).

[52] Ö. Legeza and C. Schilling, Phys. Rev. A **97**, 052105 (2018).

[53] I. Shapir, A. Hamo, S. Pecker, C. P. Moca, Ö. Legeza, G. Zarand, and S. Ilani, Science **364**, 870 (2019).

[54] G. Barcza, V. Ivády, T. Szilvási, M. Vörös, L. Veis, Á. Gali, and Ö. Legeza, J. Chem. Theory Comput. **17**, 1143 (2021).

[55] G. G. Cabrera and R. Jullien, Phys. Rev. B **35**, 7062 (1987).

[56] H. Nakano, Y. Takahashi, and M. Imada, J. Phys. Soc. Jpn. **76**, 034705 (2007).

[57] N. Shibata and C. Hotta, Phys. Rev. B **84**, 115116 (2011).

[58] S. Sorella, Phys. Rev. B **91**, 241116(R) (2015).

[59] E. Schmidt, Math. Ann. **63**, 433 (1907).

[60] G. Vidal, Phys. Rev. Lett. **91**, 147902 (2003).

[61] S. Östlund and S. Rommer, Phys. Rev. Lett. **75**, 3537 (1995).

[62] F. Verstraete, D. Porras, and J. I. Cirac, Phys. Rev. Lett. **93**, 227205 (2004).

[63] Ö. Legeza, T. Rohwedder, R. Schneider, and Sz. Szalay, in *Many-Electron Approaches in Physics, Chemistry and Mathematics*, Mathematical Physics Studies, edited by V. Bach and L. Delle Site (Springer International, Switzerland, 2014), pp. 53–76.

[64] S. Szalay, Z. Zimborás, M. Máté, G. Barcza, C. Schilling, and Ö. Legeza, J. Phys. A: Math. Theor. **54**, 393001 (2021).

[65] K. Boguslawski, P. Tecmer, G. Barcza, Ö. Legeza, and M. Reiher, J. Chem. Theory Comput. **9**, 2959 (2013).

[66] Ö. Legeza and J. Sólyom, Phys. Rev. B **68**, 195116 (2003).

[67] J. Rissler, R. M. Noack, and S. R. White, Chem. Phys. **323**, 519 (2006).

[68] Ö. Legeza and J. Sólyom, Phys. Rev. B **70**, 205118 (2004).

[69] S. Szalay, Phys. Rev. A **92**, 042329 (2015).

[70] G. Barcza, Ö. Legeza, K. H. Marti, and M. Reiher, Phys. Rev. A **83**, 012508 (2011).

[71] C. Krumnow, Ö. Legeza, and J. Eisert, arXiv:1904.11999.

[72] M. Máté, K. Petrov, S. Szalay, and Ö. Legeza, J. Math. Chem. **61**, 362 (2023).

[73] K. Petrov, A. Ganyecz, Z. Benedek, A. Olasz, G. Barcza, and Ö. Legeza, in *Advances in Methods and Applications of Quantum Systems in Chemistry, Physics, and Biology Selected Proceedings of QSCP-XXV Conference (Torun, Poland, June 2022)*, edited by I. Grabowski, K. Słowik, J. Maruani, E. J. Brändas (Springer, Cham, 2024).

[74] J. Sólyom, *Fundamentals of the Physics of Solids, Vol. 3, Normal, Broken-Symmetry, and Correlated Systems* (Springer, New York, 2011).

[75] F. Gebhard and O. Legeza, Phys. Rev. B **104**, 245118 (2021).

[76] F. Gebhard, K. Bauerbach, and Ö. Legeza, Phys. Rev. B **108**, 205130 (2023).

[77] R. M. Noack, K. Kapas, M. A. Werner, A. Menczer, M. Bergman, and Ö. Legeza (unpublished).

[78] A. Menczer and Örs Legeza, Petaflops density matrix renormalization group method (unpublished).

[79] G. Ehlers, J. Sólyom, Ö. Legeza, and R. M. Noack, Phys. Rev. B **92**, 235116 (2015).

[80] Here, by D_{opt}, we refer to the bond dimension used after the mode optimization procedure. Mode optimizations were performed at fixed bond dimension $D_{\text{mo}} = 80$.

[81] NVIDIA, https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf (unpublished).

[82] M. Qin, H. Shi, and S. Zhang, Phys. Rev. B **94**, 085103 (2016).

[83] W. Dobrautz, H. Luo, and A. Alavi, Phys. Rev. B **99**, 075119 (2019).

[84] X. Wu, Z.-H. Cui, Y. Tong, M. Lindsey, G. K.-L. Chan, and L. Lin, J. Chem. Phys. **151**, 064108 (2019).