Fast and scalable quantum Monte Carlo simulations of electron-phonon models

Benjamin Cohen-Stead [1], Owen Bradley,[1] Cole Miles,[2] George Batrouni,[3,4,5] Richard Scalettar,[1] and Kipton Barros[6]

[1]*Department of Physics, University of California, Davis, California 95616, USA*
[2]*Department of Physics, Cornell University, Ithaca, New York 14853, USA*
[3]*Université Côte d'Azur, CNRS, Institut de Physique de Nice, (INPHYNI), 06103 Nice, France*
[4]*Centre for Quantum Technologies, National University of Singapore, 2 Science Drive 3, 117542 Singapore*
[5]*Department of Physics, National University of Singapore, 2 Science Drive 3, 117542 Singapore*
[6]*Theoretical Division and CNLS, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA*

We introduce methodologies for highly scalable quantum Monte Carlo simulations of electron-phonon models, and we report benchmark results for the Holstein model on the square lattice. The determinant quantum Monte Carlo (DQMC) method is a widely used tool for simulating simple electron-phonon models at finite temperatures, but it incurs a computational cost that scales cubically with system size. Alternatively, near-linear scaling with system size can be achieved with the hybrid Monte Carlo (HMC) method and an integral representation of the Fermion determinant. Here, we introduce a collection of methodologies that make such simulations even faster. To combat "stiffness" arising from the bosonic action, we review how Fourier acceleration can be combined with time-step splitting. To overcome phonon sampling barriers associated with strongly bound bipolaron formation, we design global Monte Carlo updates that approximately respect particle-hole symmetry. To accelerate the iterative linear solver, we introduce a preconditioner that becomes exact in the adiabatic limit of infinite atomic mass. Finally, we demonstrate how stochastic measurements can be accelerated using fast Fourier transforms. These methods are all complementary and, combined, may produce multiple orders of magnitude speedup, depending on model details.

I. INTRODUCTION

As a nonperturbative and controlled approach, quantum Monte Carlo (QMC) methods have been instrumental in advancing our understanding of interacting solid-state systems. In particular, the broad class of determinant QMC (DQMC) methods have proven highly effective in helping to characterize various correlated phases that arise as a result of interactions [1]. Perhaps most notably, DQMC has enabled the study of electron-electron interactions in the repulsive Hubbard model, where Mott insulator physics, magnetic order, unconventional superconductivity, and various additional correlation effects have been observed [2–10]. The sign problem, however, has severely limited our ability to simulate systems without particle-hole or other symmetries, giving rise to an effective computational cost that scales exponentially with system size and inverse temperature [11–17].

Electron-phonon models, however, are a family of Hamiltonian systems that typically evade the sign problem, while still playing an important role in describing the effect of interactions in solid-state systems. Electron-phonon interactions are essential in explaining a host of ordered phases in material systems, such as charge density wave (CDW) order in transition metal dichalcogenides and high temperature superconductivity in the bismuthates $Bi_{1-x}K_xBiO_3$ [18–26]. Significant effort has gone toward using DQMC to study Hamiltonian systems with electron-phonon interactions, in particular the Holstein and Su-Schrieffer-Heeger (SSH)

models [27–38]. Although there is no sign problem in such systems, low-temperature DQMC simulations of electron-phonon models can still be very expensive. Explicit evaluation of the Fermion determinant results in a computational cost that scales cubically with system size. Moreover, simulations of both the Holstein and SSH models suffer from significantly longer autocorrelation times than comparable DQMC simulations of the repulsive Hubbard model. While DQMC simulations of the Holstein model have been successfully accelerated using self-learning Monte Carlo techniques [39,40], these gains are ultimately limited by continuing to require the evaluation of the Fermion determinant ratio in the Monte Carlo accept or reject step.

In recent years substantial effort has gone toward developing improved methods for simulating electron-phonon models. Recent work has successfully reduced the computational cost to near linear-scaling in system size. Such scaling can be achieved by avoiding explicit calculation of the Fermion determinant, instead using iterative linear solvers for the sampling and measurement tasks. Applied to simulations of Holstein and SSH models, both Langevin [41–44] and hybrid Monte Carlo (HMC) [45] methods have proven to be highly effective. In this paper, we introduce several general and complementary techniques that can further reduce the overall costs of simulating electron-phonon models.

Many recent studies of the Holstein and SSH models have used the Langevin method [41–44]. The traditional Langevin

approach introduces a discretization error associated with the finite time step used to integrate the stochastic dynamics. Such error can, in principle, be eliminated by introducing an accept or reject step for each proposed Langevin update [46,47]. An alternative to the Langevin approach is HMC [48]. Originally developed for lattice gauge theory simulations, the method now finds applications well beyond physics, where HMC also goes by the name *Hamiltonian* Monte Carlo [49]. Interestingly, the Langevin method can be viewed as a special case of HMC, for which the Hamiltonian trajectory length consists of only a single time step [49]. Longer trajectories with persistent momentum can be advantageous, however, to reduce autocorrelation times [50].

As applied to QMC simulations, Langevin and HMC methods offer the promise of near linear-scaling with system size. The general framework is as follows: The aim is to sample a field x according to a probability weight that is proportional to a Fermion determinant $\det M(x)$. Seeking to avoid explicit calculation of this determinant, one instead uses a stochastic approximation scheme, which requires application of the Green function matrix $M^{-1}(x)$ to a vector. The matrix $M(x)$ is highly sparse, and very efficient to apply. Iterative linear solvers, such as conjugate gradient (CG), are effective if $M(x)$ is well conditioned for typical samples x. Good conditioning is not always guaranteed; previous studies of the Hubbard model have found that the condition number can sometimes grow exponentially (e.g., as a function of inverse temperature), making iterative solvers impractical [45,51,52]. Fortunately, for models of electron-phonon interactions, the condition number of $M(x)$ seems to be reasonably well controlled. Although traditional Langevin and HMC formulations have already been successfully applied to electron-phonon simulation, there are opportunities for substantial improvement, as we shall demonstrate in this paper.

In what follows, we will interweave our new algorithmic developments with benchmarks on a prototypical reference system: The square-lattice Holstein model, which we review in Sec. II. Our core framework for sampling the phonon field is HMC, which we review in Sec. III. This application of HMC is fairly sophisticated, involving both Fourier acceleration and time-step splitting to handle the highly disparate timescales that appear in the bosonic action.

At low temperatures, the sampling of phonons can still be hindered by the formation of tightly bound bipolarons. To combat this, we employ global Monte Carlo updates as described in Sec. IV. For example, by reflecting the entire phonon field ($x \rightarrow -x$) at a particular site, the configuration can "tunnel through" a possibly large action barrier. We achieve an improved acceptance rate for these moves by carefully formulating the effective action to respect known particle-hole symmetries of the original Hamiltonian, at least in certain limits. These global updates drastically reduce autocorrelation times, and mitigate ergodicity concerns associated with nodal surfaces (vanishing Fermion determinant) [43,45], while maintaining excellent scalability of the method.

All components of the simulation can be accelerated by reducing the cost of CG for the linear solves. In Sec. V we introduce a preconditioner that significantly reduces the required number of iterations for CG to converge. Specifically, we define the preconditioner $P(x)$ to have the same structure

as $M(x)$, but without fluctuations in imaginary time. Application of $P^{-1}(x)$ to a vector can be performed very efficiently through the careful use of the fast Fourier transform (FFT) and Chebyshev polynomial expansion.

It is important that the computational cost to perform measurements scales like the cost to collect phonon samples, i.e., near-linearly in system size. By Wick's theorem, all electronic measurements can be reduced to products of the single-particle Green function, and the latter can be sampled from the matrix elements $M^{-1}(x)$. It is therefore essential to be able to estimate elements of $M^{-1}(x)$ efficiently. For this we use stochastic techniques that involve applying $M^{-1}(x)$ to random vectors. Section VI describes how FFTs can be used to achieve near-linear scaling in system size, even when averaging correlation functions over all sites and imaginary-times.

II. THE HOLSTEIN MODEL AS A BENCHMARK SYSTEM

A. Model definition

The methods presented in this paper apply generally to models of electron-phonon interactions, including the SSH and Holstein models. For concreteness, we select the latter for our benchmarks. The Holstein Hamiltonian is [53]

$$\hat{H} = \hat{H}_{\text{el}} + \hat{H}_{\text{ph}} + \hat{H}_{\text{el-ph}}, \tag{1}$$

$$\hat{H}_{\text{el}} = -\sum_{i,j,\sigma} t_{ij} \hat{c}^{\dagger}_{i,\sigma} \hat{c}_{j,\sigma} - \mu \sum_{i,\sigma} \hat{n}_{i,\sigma}, \tag{2}$$

$$\hat{H}_{\text{ph}} = \frac{m_{\text{ph}} \omega_0^2}{2} \sum_i \hat{X}_i^2 + \frac{1}{2m_{\text{ph}}} \sum_i \hat{P}_i^2, \tag{3}$$

$$\hat{H}_{\text{el-ph}} = \alpha \sum_{i,\sigma} \hat{X}_i \left(\hat{n}_{i,\sigma} - \frac{1}{2} \right), \tag{4}$$

with the normalization $\hbar = 1$ applied throughout. The first term, $\hat{H}_{\text{el}}$, models the electron kinetic energy via the hopping strengths $t_{ij} = t_{ji}$, and controls electron filling through the chemical potential μ. As usual, $\hat{c}^{\dagger}_{i,\sigma} (\hat{c}_{i,\sigma})$ is the fermionic creation (annihilation) operator for an electron with spin σ, and $\hat{n}_{i,\sigma} = \hat{c}^{\dagger}_{i,\sigma} \hat{c}_{i,\sigma}$ is the electron number operator. The second term, $\hat{H}_{\text{ph}}$, describes a dispersionless phonon branch with energy ω_0 and mass m_{ph}, modeled via the canonical position and momentum operators $\hat{X}_i$ and $\hat{P}_i$ respectively. Henceforth, the atomic mass is normalized to one, $m_{\text{ph}} = 1$. The last term, $\hat{H}_{\text{el-ph}}$, introduces an electron-phonon coupling with strength α.

B. Benchmark parameters

Our methodology applies to models with arbitrary lattice type, hopping matrix, and electron filling fraction, but we must make some specific choices for our benchmarks. We select the square lattice Holstein model at half filling ($\mu = 0$). We include only a nearest neighbor electron hopping with amplitude $t_{ij} = 1$, which defines the basic unit of energy. For the square lattice, the noninteracting bandwidth is then $W = 8$. The discretization in imaginary time, which controls Suzuki-Trotter errors, will be $\Delta_\tau = 0.1$. Our benchmarks will vary over the number of lattice sites, N, and the inverse temperature, β. A useful reference energy scale is the

dimensionless electron-phonon coupling, $\lambda = \alpha^2/(\omega_0^2 W)$. We will consider two coupling strengths, $\lambda = 0.25$ or $\lambda = 0.60$, and two phonon frequencies $\omega_0 = 0.1$ and $\omega_0 = 1$.

For these Holstein systems, the stable phase at low temperatures and half-filling is charge-density-wave (CDW) order; electrons form a checkerboard pattern, spontaneously breaking the $\mathbb{Z}_2$ symmetry between sublattices. In the case of $\omega_0 = 1.0$ and $\lambda = 0.25$, the CDW transition temperature is $\beta_{cdw} \approx 6$ [41,54]. To detect this phase, we measure the (π, π) charge structure factor

$$S_{cdw} = \sum_{\mathbf{r}} (-1)^{r_x + r_y} C(\mathbf{r}), \qquad (5)$$

where

$$C(\mathbf{r}) = \frac{1}{N} \sum_{\mathbf{r}'} \langle \hat{n}_{\mathbf{r}'+\mathbf{r}} \hat{n}_{\mathbf{r}'} \rangle, \qquad (6)$$

is the real-space density-density correlations in $\hat{n}_{\mathbf{r}} = \hat{n}_{\mathbf{r},\uparrow} + \hat{n}_{\mathbf{r},\downarrow}$. Here we are using integers $\mathbf{r} = (r_x, r_y)$ to index sites on the square lattice, assuming periodic boundary conditions. Superconducting order, however, can be detected using the pair susceptibility

$$P_s = \frac{1}{N} \int_0^\beta \langle \hat{\Delta}(\tau) \hat{\Delta}^\dagger(0) \rangle d\tau, \qquad (7)$$

where $\hat{\Delta}(\tau) = \sum_{\mathbf{r}} \hat{c}_{\mathbf{r},\downarrow}(\tau) \hat{c}_{\mathbf{r},\uparrow}(\tau)$.

All results reported in this paper use HMC trajectories composed of $N_t = 100$ time steps (Sec. III). Except where noted, we will use Fourier acceleration with mass regularization $m_{reg} = \omega_0$ (Sec. III B 1), and time-step splitting with $n_t = 10$ (Sec. III B 2). We will use a varying number of thermalization and simulation HMC trial updates, denoted N_{therm} and N_{sim}, respectively, with measurements taken after each simulation update.

C. Path integral representation

To measure thermodynamic properties, one can formulate a path integral representation of the partition function. A full derivation is given in Appendix A, with the result

$$\begin{aligned} \mathcal{Z} &= \mathrm{tr}_{el\text{-}ph}\, e^{-\beta \hat{H}} \\ &\approx \int \mathcal{D}x\, e^{-(S_B - \Delta_\tau \alpha \sum_{i,\tau} x_{i,\tau})} (\det M)^2. \end{aligned} \qquad (8)$$

Here the inverse temperature β has been discretized into L_τ intervals of imaginary time, with $\Delta_\tau = \beta/L_\tau$. The integral goes over all sites i and imaginary times τ in the real phonon field $x_{i,\tau}$. The "bosonic action"

$$S_B = \frac{\Delta_\tau}{2} \sum_{i=1}^{N} \sum_{\tau=0}^{L_\tau - 1} \left[\omega_0^2 x_{i,\tau}^2 + \left(\frac{x_{i,\tau+1} - x_{i,\tau}}{\Delta_\tau} \right)^2 \right] \qquad (9)$$

describes dispersionless phonon modes, but can be readily generalized to include anharmonic terms and phonon dispersion [54,55]. The "Fermion determinant" involves the

$NL_\tau \times NL_\tau$ matrix,

$$M = \begin{pmatrix} I & & & & B_0 \\ -B_1 & I & & & \\ & -B_2 & \ddots & & \\ & & \ddots & \ddots & \\ & & & -B_{L_\tau - 1} & I \end{pmatrix}, \qquad (10)$$

composed of $N \times N$ blocks. The off-diagonal blocks are

$$B_\tau = e^{-\Delta_\tau V_\tau} e^{-\Delta_\tau K}, \qquad (11)$$

where the matrices

$$(V_\tau)_{ij} = \delta_{ij}(\alpha x_{i,\tau} - \mu), \quad K_{ij} = -t_{ij}, \qquad (12)$$

describe the electron-phonon coupling and the electron hopping, respectively. In this real-space basis, $e^{-\Delta_\tau V_\tau}$ is exactly diagonal, whereas $e^{-\Delta_\tau K} = I - \Delta_\tau K + \dots$ is highly sparse up to corrections of order Δ_τ^2. Note that one could alternatively formulate [1]

$$\det M = \det(I + B_{L_\tau - 1} \dots B_1 B_0), \qquad (13)$$

but we do not pursue that approach here.

An innovation in this work is to rewrite the partition function as

$$\mathcal{Z} \approx \int \mathcal{D}x\, e^{-S_B} [\det(M\Lambda)]^2, \qquad (14)$$

where $\Lambda(x)$ is any matrix that satisfies

$$\det \Lambda^2 = e^{\Delta_\tau \alpha \sum_{i,\tau} x_{i,\tau}}. \qquad (15)$$

Although Eqs. (8) and (14) are mathematically equivalent, this reformulation will have important consequences in Secs. II D and IV. The factor $\exp(\Delta_\tau \alpha \sum_{i,\tau} x_{i,\tau})$ originates from our choice to include the $-\alpha \sum_i \hat{X}_i/2$ term in Eq. (4), which effectively selects $\hat{X}_i = 0$ as the reflection point for particle-hole symmetry.

There are many possible choices for Λ. We select

$$\Lambda_{(i,\tau),(i',\tau')} = \delta_{i,i'} \delta_{\tau+1,\tau'} (2\delta_{\tau',0} - 1) e^{+\frac{\Delta_\tau \alpha}{2} x_{i,\tau'}}, \qquad (16)$$

with inverse

$$\Lambda^{-1}_{(i,\tau),(i',\tau')} = \delta_{i,i'} \delta_{\tau,\tau'+1} (2\delta_{\tau,0} - 1) e^{-\frac{\Delta_\tau \alpha}{2} x_{i,\tau}}, \qquad (17)$$

where the index $\tau = 0, 1, \dots L_\tau - 1$ is understood to be periodic in L_τ.

To collect equilibrium statistics, one samples the phonon field $x_{i,\tau}$, taking the positive-definite integrand in Eq. (14) to be the probability weight. Sampling $x_{i,\tau}$ is typically the dominant cost of a QMC code. A traditional DQMC code involves periodic evaluation of the matrix determinant of Eq. (13), at a cost of $\mathcal{O}(N^3)$ computational operations. In a careful DQMC implementation, this determinant may be calculated relatively infrequently, typically once per "full sweep" of Monte Carlo updates to each of the auxiliary field components, $x_{i,\tau}$ [56]. As we will next discuss in Sec. II D, the cost to sample the phonon field can still be significantly reduced, from cubic to approximately linear scaling with system size N.

Note that a similar path integral formulation can be derived for the SSH model. There, however, the phonon position operators $\hat{X}_i$ modulate the electron hopping term, such that

the matrices K_τ gain a dependence on the phonon fields $x_{i,\tau}$ [20,36–38,43,45,57].

D. Sampling the phonon field at approximately linear scaling cost

Given a nonsingular matrix A of dimension D, its determinant can be formulated as an integral,

$$|\det A| = (2\pi)^{-D/2} \int \mathcal{D}\Phi \, e^{-\frac{1}{2}\Phi^T (A^T A)^{-1}\Phi}, \qquad (18)$$

where each component of the vector Φ is understood to be integrated over the entire real line.

We twice apply Eq. (18) to Eq. (14), introducing an integral for each of the two Fermion determinants. Taking

$$A(x) = M(x)\Lambda(x), \qquad (19)$$

the partition function becomes

$$\mathcal{Z} \approx (2\pi)^{NL_\tau} \int \mathcal{D}\Phi_\uparrow \mathcal{D}\Phi_\downarrow \mathcal{D}x \, e^{-S(x,\Phi_\sigma)}. \qquad (20)$$

In place of the matrix determinants, there is now a "fermionic" contribution to the action,

$$S(x, \Phi_\sigma) = S_B(x) + S_F(x, \Phi_\sigma), \qquad (21)$$

with

$$
\begin{aligned}
S_F(x, \Phi_\sigma) &= \frac{1}{2}\sum_\sigma \Phi_\sigma^T (A^T A)^{-1}\Phi_\sigma \\
&= \frac{1}{2}\sum_\sigma |A^{-T}\Phi_\sigma|^2.
\end{aligned}
\qquad (22)
$$

Now we must sample the two auxiliary fields $\Phi_{\{\uparrow,\downarrow\}}$ in addition to the phonon field x, according to the joint distribution $P(x, \Phi_\sigma) \propto \exp(-S)$. With the Gibbs sampling method, one alternately updates x and Φ_σ according to the conditional distributions $P(x|\Phi_\sigma)$ and $P(\Phi_\sigma|x)$, respectively.

Holding x fixed, observe that

$$P(\Phi_\sigma|x) \propto e^{-S_F} = e^{-\frac{1}{2}\sum_\sigma |R_\sigma|^2}, \qquad (23)$$

where the vector $R_\sigma = A^{-T}\Phi_\sigma$ is found to be Gaussian distributed. Therefore, to sample Φ_σ at fixed x, one may first sample Gaussian R_σ, and then assign

$$\Phi_\sigma = A^T(x) R_\sigma. \qquad (24)$$

Because Φ_σ is randomly sampled, it is convenient to treat it as an arbitrary, fixed vector. Alternatively, we can view $\Phi_\sigma(x, R_\sigma)$ as a deterministic function of x provided that the random sample R_σ is also supplied.

Sampling x at fixed Φ_σ is the primary numerical challenge. In the Metropolis Monte Carlo approach, one proposes an update $x \to x'$ and accepts it with probability,

$$P(x \to x') = \min(1, e^{-\Delta S}), \qquad (25)$$

where

$$\Delta S = S(x', \Phi_\sigma) - S(x, \Phi_\sigma). \qquad (26)$$

Sophisticated methods for proposing updates include HMC (Sec. III) and reflection and swap updates (Sec. IV).

Calculating the acceptance probability requires evaluating the change in action,

$$\Delta S = \Delta S_B + \Delta S_F. \qquad (27)$$

The bosonic part can be readily calculated from Eq. (9). The fermionic part is given by Eq. (22),

$$\Delta S_F = S_F(x', \Phi_\sigma) - S_F(x, \Phi_\sigma). \qquad (28)$$

The recipe for sampling the auxiliary field $\Phi_\sigma = \Phi_\sigma(x, R_\sigma)$ is given by Eq. (24), and involves the *initial* phonon configuration x. Substituting into Eq. (22) yields $S_F(x, \Phi_\sigma) = \frac{1}{2}\sum_\sigma |R_\sigma|^2$.

It remains nontrivial to calculate

$$S_F(x', \Phi_\sigma) = \frac{1}{2}\sum_\sigma \Phi_\sigma^T \Psi_\sigma, \qquad (29)$$

where

$$
\begin{aligned}
\Psi_\sigma &= (A^T A)^{-1}\Phi_\sigma \\
&= \Lambda^{-1}(M^T M)^{-1}\Lambda^{-T}\Phi_\sigma,
\end{aligned}
\qquad (30)
$$

and the matrices M and Λ are understood to be evaluated at the new phonon field, x'. The vector $b = \Lambda^{-T}\Phi_\sigma$, for each σ, can be readily calculated using Eq. (17).

To solve iteratively for the vector

$$v = (M^T M)^{-1} b, \qquad (31)$$

one can use the conjugate gradient (CG) method [58]. After n iterations, CG optimally approximates $v_n \approx v$ from within the nth Krylov space, i.e., the vector space spanned by basis vectors $(M^T M)^j b$ for $j = 0, 1, \ldots n$. Given the solution v, the action S_F in Eq. (29) can be evaluated by noting that $\Phi_\sigma^T \Psi_\sigma = b^T v$.

CG requires repeated multiplication by $M^T M$. Applying M and M^T to a vector is very efficient due to the block sparsity structure in Eq. (10). The off-diagonal blocks B_τ inside M involve the exponential of the tight-binding hopping matrix K. To apply efficiently $e^{-\Delta_\tau K}$ to a vector, one may approximately factorize this exponential as a chain of sparse operators using the minimal split checkerboard method [59], which remains valid up to errors of order $O(\Delta_\tau^2)$ [60]. This allows us to apply B_τ to a vector of like dimension at a cost that scales linearly with system size N.

The rate of CG convergence is determined by the condition number of $M^T M$, i.e., the ratio of largest to smallest eigenvalues (as a function of the fluctuating phonon field). In previous QMC studies on the Hubbard model, the analogous condition number was found to increase rapidly with inverse temperature and system size at moderate coupling [52]. Fortunately, for electron-phonon models at moderate parameter values ($\lambda \lesssim 1$ and $\omega_0 \lesssim t$) the condition number is observed to increase only very slowly with β and N [45]. We observe that larger phonon frequency ($\omega_0 \gtrsim t$) coincides with larger condition number and slower CG convergence. This will be reflected by the benchmarks in this paper, for which CG typically converges in hundreds of iterations or fewer. Furthermore, the required number of CG iterations can be significantly reduced by using a carefully designed preconditioning matrix, as we will describe in Sec. V.

III. HMC SAMPLING OF THE PHONON FIELD

A. Review of HMC

Hybrid Monte Carlo (HMC) was originally developed in the lattice gauge theory community [48] and has since proven broadly useful for statistical sampling of continuous variables [49]. In particular, it is a powerful method for sampling the phonon field x in electron-phonon models [45,61].

In HMC a fictitious momentum $p_{i,\tau}$ is introduced that is dynamically conjugate to $x_{i,\tau}$. Specifically, a Hamiltonian,

$$H(x, p) = S(x) + \frac{p^T \mathcal{M}^{-1} p}{2}, \tag{32}$$

is defined that can be interpreted as the sum of "potential" and "kinetic" energies. The dynamical mass $\mathcal{M}$ can be any positive-definite matrix, independent of x and p. Recall that the action $S(x)$ is implicitly dependent on the auxiliary field Φ_σ; we omit this dependence because Φ_σ is treated as fixed for purposes of sampling x.

The corresponding Hamiltonian equations of motion are

$$\dot{p} = -\frac{\partial H}{\partial x} = -\frac{\partial S}{\partial x}, \tag{33}$$

$$\dot{x} = \frac{\partial H}{\partial p} = \mathcal{M}^{-1} p. \tag{34}$$

The dynamics is time-reversible, energy conserving, and symplectic (phase space volume conserving). These properties make it well suited for proposing updates to the phonon field. We use a variant of HMC consisting of the following three steps:

Step (1) of HMC samples p from the equilibrium Boltzmann distribution, proportional to $\exp(-p^T \mathcal{M}^{-1} p/2)$. This is achieved by sampling components $R_{i,\tau}$ from a standard Gaussian distribution, and then setting

$$p = \sqrt{\mathcal{M}} R. \tag{35}$$

Step (2) of HMC integrates the Hamiltonian dynamics for N_t integration time steps. We use the leapfrog method,

$$p_{t+1/2} = p_t - \frac{\Delta t}{2} \frac{\partial S}{\partial x_t}, \tag{36}$$

$$x_{t+1} = x_t + \Delta t \mathcal{M}^{-1} p_{t+1/2}, \tag{37}$$

$$p_{t+1} = p_{t+1/2} - \frac{\Delta t}{2} \frac{\partial S}{\partial x_{t+1}}, \tag{38}$$

where Δt denotes the integration step size. Note that when performing leapfrog integration steps sequentially, only a single evaluation of $\partial S/\partial x$ must be performed per time step. This is because the final half-step momentum update $p_{t+1/2} \to p_{t+1}$ can be merged with the initial one from the next time step, $p_{t'} \to p_{t'+1/2}$, where $t' = t + 1$. The leapfrog integration scheme is exactly time-reversible and symplectic. One integration step is accurate to $\mathcal{O}(\Delta t^3)$ and, in the absence of numerical instability, total energy is conserved to order $\mathcal{O}(\Delta t^2)$ for arbitrarily long trajectories [62–64]. Any symplectic integration scheme could be used in place of leapfrog; the second-order Omelyan integrator is an especially promising alternative [65,66].

For this paper we will fix $N_t = 100$. Future work would likely benefit from randomizing the length of each HMC trajectory; doing so has been observed to reduce decorrelation times and would mitigate certain ergodicity concerns [67]. For certain types of sampling problems, e.g., sampling in the vicinity of a critical point, one should also consider the use of much longer HMC trajectories [50].

Step (3) of HMC is to accept (or reject) the dynamically evolved configuration x' according to the Metropolis probability, Eq. (25). HMC exactly satisfies detailed balance, and the proof depends crucially on the leapfrog integrator being time-reversible and symplectic [48,49]. An acceptance rate of order one can be maintained by taking the time step to scale only very weakly with system size ($\Delta t \sim N^{-1/4}$) [68]. Higher-order symplectic integrators are also possible, and come even closer to allowing constant Δt, independent of system size [62]. Extensive discussion about selecting a good Δt value is presented in Ref. [49], including evidence that a 65% acceptance rate may be close to optimal in many situations. In practice, however, we typically choose a more conservative Δt such that the acceptance rate is much closer to one. This minimizes the danger of hitting numerical instabilities, which can lead to significant slow-downs, and can be difficult to anticipate over a wide space of physical parameters.

Numerical integration requires evaluation of the fictitious force $-\partial S/\partial x$ at each time step. Specifically, one must calculate

$$\frac{\partial S}{\partial x_{i,\tau}} = \frac{\partial S_B}{\partial x_{i,\tau}} + \frac{\partial S_F}{\partial x_{i,\tau}}. \tag{39}$$

The bosonic part is

$$\frac{\partial S_B}{\partial x_{i,\tau}} = \Delta_\tau \left(\omega_0^2 x_{i,\tau} - \frac{x_{i,\tau+1} - 2x_{i,\tau} + x_{i,\tau-1}}{\Delta_\tau^2} \right). \tag{40}$$

For the fermionic part, we must calculate

$$\frac{\partial S_F}{\partial x_{i,\tau}} = \frac{1}{2} \sum_\sigma \Phi_\sigma^T \frac{\partial (A^T A)^{-1}}{\partial x} \Phi_\sigma, \tag{41}$$

where Φ_σ is fixed throughout the dynamical trajectory. Using the general matrix identity $dC^{-1} = -C^{-1}(dC)C^{-1}$, we find

$$\frac{\partial S_F}{\partial x_{i,\tau}} = -\sum_\sigma \Psi_\sigma^T A^T \frac{\partial A}{\partial x_{i,\tau}} \Psi_\sigma, \tag{42}$$

where $\Psi_\sigma = (A^T A)^{-1} \Phi_\sigma$ evolves as a function of x over the dynamical trajectory, with Φ_σ held fixed. As with the calculation of ΔS_F in Eq. (27), the numerically expensive task is to calculate $\Psi_\sigma = (A^T A)^{-1} \Phi_\sigma$, for which we use the CG algorithm.

Given Ψ_σ, we must also apply the highly sparse matrix

$$\frac{\partial A_x}{\partial x_{i,\tau}} = \frac{\partial M}{\partial x_{i,\tau}} \Lambda + M \frac{\partial \Lambda}{\partial x_{i,\tau}}, \tag{43}$$

for each index (i, τ) of the phonon field. Differentiating Λ in Eq. (16) is straightforward. The derivative of M in Eq. (10) with respect to $x_{i,\tau}$ involves only a single nonzero $N \times N$ block matrix. In the Holstein model, we use

$$\frac{\partial B_{\tau'}}{\partial x_{i,\tau}} = \delta_{\tau,\tau'} \left(\frac{\partial}{\partial x_{i,\tau}} e^{-\Delta_\tau V_\tau} \right) e^{-\Delta_\tau K}, \tag{44}$$

where V_τ is diagonal, so that its exponential is easy to construct and differentiate.

The situation is a bit more complicated for the SSH model, where the $x_{i,\tau}$-dependence appears inside the hopping matrix K_τ, which is *not* diagonal. In this case, we may exploit the checkerboard factorization [59] of $e^{-\Delta_\tau K_\tau}$, and use the product rule to differentiate each of the sparse matrix factors one-by-one. If implemented carefully, then the cost to evaluate all $N L_\tau$ forces $-\partial S/\partial x_{i,\tau}$ remains of the same order as the cost to evaluate the scalar S. That this is generically possible follows from the concepts of reverse-mode automatic differentiation [69].

B. Resolving disparate timescales in the bosonic action

One of the challenges encountered when simulating electron-phonon models is that the bosonic action gives rise to a large disparity of timescales in the Hamiltonian dynamics. Here we will present two established approaches for unifying these dynamical timescales.

The bosonic part of the Hamiltonian dynamics decouples in the Fourier basis. To see this, we will employ the discrete Fourier transform in imaginary time,

$$\hat{f}_\omega = \frac{1}{\sqrt{L_\tau}} \sum_{\tau=0}^{L_\tau-1} e^{-\frac{2\pi i}{L_\tau}\omega\tau} f_\tau, \tag{45}$$

where the integer index ω is effectively periodic mod L_τ. The Fourier transform may be represented by an $L_\tau \times L_\tau$ unitary matrix,

$$\mathcal{F}_{\omega,\tau} = \frac{1}{\sqrt{L_\tau}} e^{-\frac{2\pi i}{L_\tau}\omega\tau}, \tag{46}$$

such that $\hat{f} = \mathcal{F}f$.

Consider the bosonic force defined in Eq. (40),

$$f_{i,\tau} = -\partial S_{\mathrm{B}}/\partial x_{i,\tau}. \tag{47}$$

Its Fourier transform is

$$\hat{f}_{i,\omega} = -\tilde{Q}_{\omega,\omega}\hat{x}_{i,\omega}, \tag{48}$$

where $\hat{x} = \mathcal{F}x$ and

$$\tilde{Q}_{\omega,\omega} = \Delta_\tau\left[\omega_0^2 + \frac{4}{\Delta_\tau^2}\sin^2\left(\frac{2\pi\omega}{L_\tau}\right)\right]. \tag{49}$$

We may interpret $\tilde{Q}_{\omega,\omega}$ as the elements of a diagonal matrix $\tilde{Q}$ in the Fourier basis. In the original basis,

$$\partial S_{\mathrm{B}}/\partial x = Qx, \tag{50}$$

where $Q = \mathcal{F}^{-1}\tilde{Q}\mathcal{F}$.

The diagonal matrix element $\tilde{Q}_{\omega,\omega}$ gives the force acting on the Fourier mode $\hat{x}_\omega$. The extreme cases are $\omega = \pm L_\tau/2$ and $\omega = 0$, for which $\tilde{Q}_{\omega,\omega}/\Delta_\tau$ takes the values $\omega_0^2 + 4/\Delta_\tau^2$ and ω_0^2 respectively. The ratio of force magnitudes for the fastest and slowest dynamical modes is then

$$1 + \frac{4}{\omega_0^2\Delta_\tau^2} \gg 1, \tag{51}$$

which diverges in the continuum limit, $\Delta_\tau \to 0$. Typically Δ_τ is of order 0.1, and the physically relevant phonon frequencies are order $\omega_0 \sim 0.1$.

Numerical integration of the Hamiltonian dynamics will be limited to small time steps to resolve the dynamics of the fast modes, $\omega \sim \pm L_\tau/2$. Unfortunately, this means that a very large number of time steps $N_t \propto \mathcal{O}(4/\omega_0^2\Delta_\tau^2)$ is required to reach the dynamical timescale in which the slow modes, $\omega \sim 0$, can meaningfully evolve.

1. Dynamical mass matrix

Here we describe the method of Fourier acceleration, by which a careful selection of the dynamical mass matrix $\mathcal{M}$ can counteract the widely varying bosonic force scales appearing in Eq. (49) [41,70].

The Hamiltonian dynamics of Eqs. (33) and (34) may be written $\ddot{x} = -\mathcal{M}^{-1}\partial S/\partial x$. The characteristic scaling for fermionic forces is

$$\partial S_{\mathrm{F}}/\partial x \sim \Delta_\tau. \tag{52}$$

This is expected because Δ_τ enters into S_{F} only through the scaled phonon field, $y_{i,\tau} \equiv \Delta_\tau x_{i,\tau}$. The chain rule $\partial S_{\mathrm{F}}/\partial x = (\partial S_{\mathrm{F}}/\partial y)(\partial y/\partial x)$ then suggests linear scaling in Δ_τ.

Per Eq. (49), the bosonic forces also typically scale like Δ_τ when ω is small. However, for the large Fourier modes $\omega \sim \pm L_\tau/2$, we find instead

$$\partial S_{\mathrm{B}}/\partial x \sim \Delta_\tau^{-1}, \tag{53}$$

which will typically dominate other contributions to the total force. One may therefore consider the idealized limit of a purely bosonic action, $S(x) = S_{\mathrm{B}}(x)$, which is approximately valid for the large ω modes. Using Eq. (50), the dynamics for purely bosonic forces is

$$\ddot{x} = -\mathcal{M}^{-1}Qx \qquad (S = S_{\mathrm{B}}), \tag{54}$$

If we were to select $\mathcal{M} = Q/\omega_0^2$, then the dynamics would become $\ddot{\hat{x}} = -\omega_0^2\hat{x}$, which describes a system of noninteracting harmonic oscillators, all sharing the *same* period, $2\pi/\omega_0$. This would be the ideal choice of $\mathcal{M}$ if the assumption $S = S_{\mathrm{B}}$ were perfect.

The true action S is not purely bosonic, and it can be advantageous to introduce a regularization m_{reg} that weakens the effect of $\mathcal{M}$ when acting on small ω. We define diagonal matrix elements [41],

$$\tilde{\mathcal{M}}_{\omega,\omega} = \Delta_\tau\left[\frac{m_{\mathrm{reg}}^2 + \omega_0^2 + \frac{4}{\Delta_\tau^2}\sin^2\left(\frac{2\pi\omega}{L_\tau}\right)}{m_{\mathrm{reg}}^2 + \omega_0^2}\right], \tag{55}$$

as the Fourier representation of the dynamical mass matrix,

$$\mathcal{M} = \mathcal{F}^{-1}\tilde{\mathcal{M}}\mathcal{F}. \tag{56}$$

For small frequencies ω (or infinite regularization m_{reg}) the mass matrix is approximately constant, $\mathcal{M} \approx \Delta_\tau$ consistent with the scaling of fermionic forces, Eq. (52). For large frequencies, $\omega \sim L_\tau/2$, however, a finite regularization m_{reg} is irrelevant, and we find $\mathcal{M} \approx Q/\omega_0^2$. Comparing with Eq. (54), the high-frequency modes are found to behave like harmonic oscillators with an ω-independent force-scale that is again consistent with Eq. (52).

The effectiveness of Fourier acceleration depends on the degree to which a clean separation of scales can be found. Typically Δ_τ will be sufficiently small such that there is a
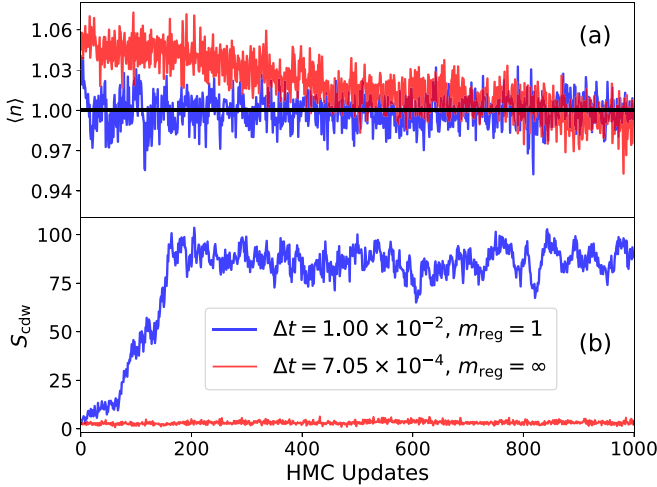
FIG. 1. Equilibration process for $\omega_0 = 1.0$, $\lambda = 0.25$, $\beta = 8$, and $N = 256$. Panel (a) displays the time history for the density $\langle n \rangle$. Panel (b) displays the time history for the structure factor S_{cdw}. Compares results for two simulations started from the same initial configuration that use different dynamical mass matrices $\mathcal{M}$. The time steps Δt are chosen so that the highest frequency mode in both simulations evolves on the same effective timescale.

range of Fourier modes for which S_B is the dominant contribution to the action S.

Our convention for the dynamical mass matrix $\mathcal{M}$ deviates somewhat from previous work [41]. The present convention aims to decouple the integration time step Δt from the discretization in imaginary time Δ_τ, such that the two parameters may be varied independently. In other words, one "unit of integration time" should produce an approximately fixed amount of decorrelation in the phonon field, independent of Δ_τ.

Figure 1 compares the equilibration process for two simulations of a Holstein model in the CDW phase, one using $m_{reg} = \omega_0$ and $\Delta t = 1 \times 10^{-2}$ shown in blue, the other using $m_{reg} = \infty$ and $\Delta t = 7.05 \times 10^{-4}$, shown in red. These Δt have been selected such that the highest frequency dynamical mode $\omega = L_\tau / 2$ evolves on the same timescales in both simulations.

Figure 1(a) shows the time history of sampled densities $\langle n \rangle$ for each simulation. While the measured densities in the simulation using $m_{reg} = \omega_0$ almost immediately begin fluctuating about $\langle n \rangle = 1$, in simulations using $m_{reg} = \infty$ the density only gradually approaches half-filling. The discrepancy between the two simulations is even more obvious when we look at the time series for S_{cdw} shown in Fig. 1(b). While the simulation using $m_{reg} = \omega_0$ rapidly equilibrates to CDW order in roughly ~ 150 updates, the $m_{reg} = \infty$ simulation shows no perceptible indication of thermalization toward CDW order.

2. Time-step splitting

A complementary strategy to handle the disparate timescales associated with the bosonic action is time-step splitting [49,65]. Typically, $\partial S_B / \partial x$ is much less expensive to evaluate than $\partial S_F / \partial x$. One may modify the leapfrog integration method of Eqs. (36)–(38) to use multiple, smaller integration time steps $\Delta t' = \Delta t / n_t$ using the bosonic force

Algorithm 1. Time-step Splitting HMC Update

Record initial state: x_i
Directly sample auxiliary field: $\Phi_\sigma := A^T(x_i)R_\sigma$
Directly sample momentum: $p_i := \sqrt{\mathcal{M}}R$
Calculate initial energy: $H_i := H(x_i, p_i)$
for $t \in 1 \ldots N_t$ **do**
    $p := p - \frac{\Delta t}{2}\frac{\partial S_F}{\partial x}$
    for $t' \in 1 \ldots n_t$ **do**
        $p := p - \frac{\Delta t'}{2}\frac{\partial S_B}{\partial x}$
        $x := x + \Delta t' \mathcal{M}^{-1}p$
        $p := p - \frac{\Delta t'}{2}\frac{\partial S_B}{\partial x}$
    end for
    $p := p - \frac{\Delta t}{2}\frac{\partial S_F}{\partial x}$
end for
Calculate final energy: $H_f := H(x_f, p_f)$
Acceptance probability: $P := \min(1, e^{-(H_f - H_i)})$
Sample r uniform in (0,1)
if $r < P$ **then**
    Accept final phonon field configuration x_f
else
    Revert to initial phonon field configuration x_i
endif

alone. After taking n_t of these sub time steps, a full time step Δt is performed using the fermionic force alone. The final leapfrog integrator is shown in Algorithm I, and can be derived by a symmetric operator splitting procedure. Like the original leapfrog algorithm, it is exactly time-reversible and symplectic.

Figure 2 demonstrates the practical benefit of time-step splitting by showing how the HMC acceptance probability varies with the number n_t of sub time steps. To isolate the impact of time-step splitting, we disabled Fourier acceleration by effectively setting $m_{reg} = \infty$. The measured acceptance rate is zero until $n_t \geqslant 4$, at which point it rapidly saturates to a value of $\sim 95\%$ once $n_t \geqslant 5$. This result illustrates a sharp stability limit: When $n_t < 4$, the corresponding value of $\Delta t'$ is too large to resolve the fastest Fourier modes, $\hat{x}_{\omega = L/2}$, which causes a dynamical instability and uncontrolled error. When n_t
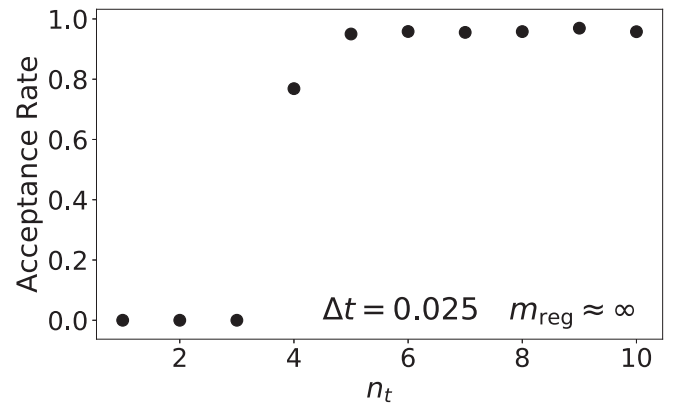

FIG. 2. HMC acceptance rate versus n_t for $\omega_0 = 1$, $\lambda = 0.25$, $\beta = 4$, and $L = 16$. Note the sharp transition at $n_t \geqslant 4$ sub time steps. For this test we disabled Fourier acceleration, effectively taking $m_{reg} = \infty$ such that $\mathcal{M} = \Delta_\tau$.

increases beyond a certain point, the corresponding values of $\Delta t'$ are sufficiently small to stabilize the S_B driven dynamics.

C. Summary of an HMC update

Algorithm I shows the pseudocode for one HMC trial update.

We remark that although methods of Fourier acceleration and time-step splitting aim to solve a similar problem, they employ different mechanisms. The dynamical mass matrix $\mathcal{M}$ of Eq. (56) was derived by analyzing a noninteracting system, and effectively slows down the dynamics of high-frequency Fourier modes. It is effective for handling Fourier modes for which the force contribution from S_B dominates. In contrast, time-step splitting works by focusing more computational effort on integrating the bosonic forces, and allows the high frequency modes to evolve on their natural, faster timescale. If the cost to calculate $\partial S_B/\partial x$ were truly negligible (relative to $\partial S_F/\partial x$), then we could take n_t sufficiently large to completely resolve the highest frequency dynamical modes arising from S_B, and Fourier acceleration could be disabled ($m_{reg} \to \infty$). Empirically, we find a combination of the two methods to be most effective. As such, for the rest of our benchmarks we perform HMC updates with $\Delta t = \omega_0^{-1}/100$, $N_t = 100$, $n_t = 10$, and $m_{reg} = \omega_0$.

IV. REFLECTION AND SWAP UPDATES

Simulations of Holstein models can suffer from diverging decorrelation times (effective ergodicity breaking) as a result of the effective phonon mediated electron-electron attraction. The strength of this attractive interaction between electrons is approximately parameterized by $U_{eff} = -\alpha^2/\omega_0^2 = -\lambda W$ [34], where W is the noninteracting bandwidth. Large dimensionless coupling λ gives rise to "heavy" bipolaron physics [33,71]. In this case, it is energetically favorable for the system to have either 0 or 2 electrons on a site, corresponding to the phonon position $\hat{X}$ being displaced in the positive or negative directions, respectively [cf. Eq. (4)]. The energy penalty at $x_{i,\tau} = 0$ roughly corresponds to the unfavorable condition of a single electron residing on the site, and is approximately proportional to $U_{eff}/2$. In the context of QMC, we aim to sample fluctuations in the phonon field $x_{i,\tau}$, with the action $S(x)$ exhibiting a strong repulsion around $x_{i,\tau} = 0$. When λ is large, this action barrier effectively traps the sign of the phonon field at each site i.

To overcome this effective trapping, one may employ additional types of Monte Carlo updates. We consider *reflection* updates to flip the phonon field $x_i \to -x_i$ on a single site i (at all imaginary times), and *swap* updates to exchange the phonon field $(x_i, x_j) \to (x_j, x_i)$ of neighboring sites. Similar updates have previously been shown to be effective in DQMC simulations of Hubbard and Holstein models [72,73]. A subtle difficulty arises, however, when attempting to use such global moves in the context of fixed auxiliary fields Φ_σ (cf. Sec. II D). Here we demonstrate how the introduction of the Λ matrix in the path integral formulation of Eq. (14) dramatically increases the acceptance rates for these global moves.
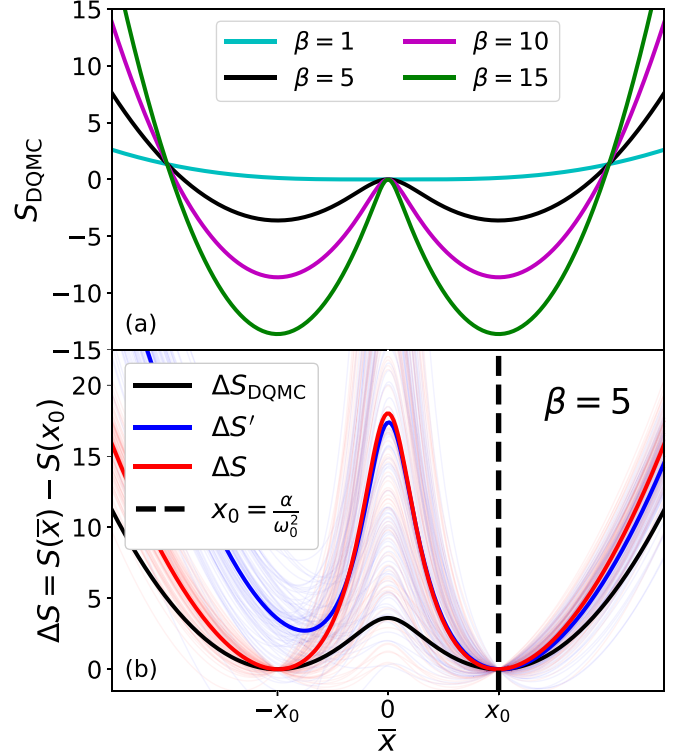

FIG. 3. (a) $S_{DQMC}(\bar{x})$ for the single-site Holstein model at half-filling ($t_{ij} = 0$, $\omega_0 = 1$, $\alpha = \sqrt{2}$, $\mu = 0$), plotted as a function of the phonon field $\bar{x}$ with imaginary-time fluctuations suppressed. (b) Change in action under the proposed move $x_0 \to \bar{x}$, where $x_0 = \alpha/\omega_0^2$. Bold blue and red lines represent the average over 100 vectors Φ_σ, sampled according to Eq. (24) with $\bar{x} = x_0$. With imaginary-time fluctuations suppressed, ΔS is exactly symmetric, whereas $\Delta S'$ is not.

To develop intuition, we consider the single-site limit ($t_{ij} = 0$) of the Holstein model at half filling ($\mu = 0$), which satisfies an exact particle-hole symmetry. In this limit a particle-hole transformation is realized by

$$\hat{X} \to -\hat{X}, \quad \hat{c} \to \hat{c}^\dagger. \tag{57}$$

This transforms $\hat{n} \to 1 - \hat{n}$, yet leaves the Hamiltonian $\hat{H}$ in Eq. (1) invariant.

In a traditional DQMC code, the phonon field would be sampled according to the weight $\exp(-S_{DQMC})$ appearing in Eq. (14), where

$$S_{DQMC} = S_B - 2\ln(e^{\beta\alpha\bar{x}/2}\det M) \tag{58}$$

and $\bar{x} = \sum_\tau x_\tau/L_\tau$. In the single site limit, each B_τ becomes scalar, and we can evaluate Eq. (13) as

$$\det M = 1 + e^{-\Delta\tau\sum_{\tau=0}^{L_\tau}V_\tau} = 1 + e^{-\beta(\alpha\bar{x}-\mu)}. \tag{59}$$

Taking $\mu = 0$, it follows

$$S_{DQMC} = S_B - 2\ln\cosh(\beta\alpha\bar{x}/2), \tag{60}$$

up to an irrelevant constant shift.

Let us momentarily ignore fluctuations in imaginary time, which is justifiable at small ω_0. By replacing $x_\tau \to \bar{x}$, the bosonic action becomes $S_B \to \beta\omega_0^2\bar{x}^2/2$. Figure 3(a) plots the resulting $S_{DQMC}(\bar{x})$. As the inverse temperature β increases,
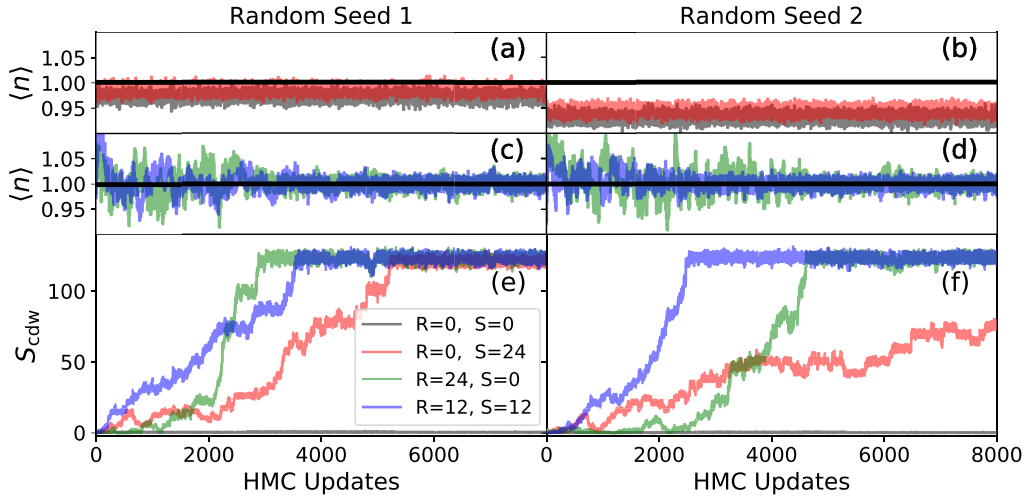
FIG. 4. Equilibration of observables with $\omega_0 = 0.1$, $\lambda = 0.6$, $\beta = 9$, and $L = 16$. The presence of strongly bound bipolarons leads to large sampling barriers, and simulation with HMC updates alone ($S = 0, R = 0$) shows a failure to equilibrate on accessible simulation scales. Including also swap updates ($S > 0, R = 0$) allows S_{cdw} to partially relax, but effective broken ergodicity is still observable from measurements of $\langle n \rangle$. Reflection updates ($R > 0$) are crucial to realizing fast decorrelation times in all observables.

a double-well structure emerges, and the action barrier at $\bar{x} = 0$ poses a practical problem for sampling. Equation (60) ensures the exact symmetry $S_{DQMC}(x) = S_{DQMC}(-x)$, even in the presence of imaginary-time fluctuations, such that reflection moves would always be accepted given this choice of action.

Curiously, the $x \leftrightarrow -x$ symmetry is missing from the action of Eq. (21) that we actually use for sampling the phonons. Specifically, $S(x, \Phi_\sigma) \neq S(-x, \Phi_\sigma)$ at fixed Φ_σ. As a practical consequence, the proposal of a global update $x \to -x$ at fixed Φ_σ may lead to very low Monte Carlo acceptance rates, Eq. (25), unless the action is carefully constructed.

To demonstrate how Monte Carlo acceptance rates can suffer, we consider two Φ_σ-dependent actions, S and S'. The first we have already defined in Eq. (21),

$$S = S_B + \frac{1}{2} \sum_\sigma |A^{-T} \Phi_\sigma|^2, \tag{61}$$

where $A = M\Lambda$. The second follows from Eq. (8) and would, more traditionally, be used for the Holstein model,

$$S' = S_B - \beta \alpha \bar{x} + \frac{1}{2} \sum_\sigma |M^{-T} \Phi_\sigma|^2. \tag{62}$$

Both actions are statistically valid—integration over the auxiliary fields yields the correct distribution for x,

$$\int \mathcal{D}\Phi_\sigma e^{-S} \propto \int \mathcal{D}\Phi_\sigma e^{-S'} \propto e^{-S_{DQMC}}. \tag{63}$$

However, the two actions produce very different acceptance rates for global Monte Carlo moves. Figure 3(b) demonstrates this by plotting ΔS and $\Delta S'$ for a proposed update $x_0 \to \bar{x}$, with imaginary-time fluctuations suppressed. For concreteness we selected the initial condition $x_0 = \alpha/\omega_0^2$, but the choice does not qualitatively affect our conclusions. Each thin curve is plotted using a different randomly sampled Φ_σ, drawn from the exponential distributions $\exp[-S(x_0, \Phi_\sigma)]$ or

$\exp[-S'(x_0, \Phi_\sigma)]$ in the case of ΔS (red) or $\Delta S'$ (blue), respectively.

From Fig. 3(b), it is apparent that the action S has the symmetry

$$\Delta S(\bar{x}) = \Delta S(-\bar{x}). \tag{64}$$

This is an exact result for the single-site, adiabatic limit of the Holstein model (see Appendix B). The action S', however, has a very different qualitative behavior. Here, the proposed update $x_0 \to -x_0$ imposes a very large action cost $\Delta S'$ for nearly all auxiliary field samples, Φ_σ.

The qualitative difference between ΔS and $\Delta S'$ has a profound effect on the Metropolis acceptance rate, Eq. (25), for phonon reflections $x \to -x$. We quantify this through numerical experiments using the single-site Holstein model at half filling, with moderate parameters $\omega_0 = 1$, $\alpha = 2$, and $\beta = 4$. If we used the full action S_{DQMC}, then the proposed move $x \to -x$ would have a 100% acceptance probability, which follows from particle-hole symmetry, and is the ideal behavior. If the naive action S' is used, then the Metropolis acceptance rate for a reflection update is only $\sim 2\%$, averaged over random samples of Φ_σ. If the action S is used instead of S', then particle-hole symmetry is statistically restored in the sense of Eq. (64), and the acceptance rate for reflection updates goes up to $\sim 68\%$ (larger is better). We will continue to use the action S throughout the rest of this paper. Although the action S_{DQMC} yields the highest acceptance rate, it incurs a computational cost that scales cubically with system size. The action S maintains a fairly high acceptance rate while retaining near-linear scaling of the overall method.

The use of reflection and swap updates provides tremendous speedups in practical studies of the Holstein model going beyond the single-site limit. Figure 4 shows the equilibration process for a Holstein model on a $N = 16^2$ square lattice. We used a relatively large coupling $\lambda = 0.6$, such that on-site action barriers are large. At inverse temperature $\beta = 9$, the system is in a robust CDW phase. We ran the same simulation twice using two different random seeds, shown in the left

and right columns. With $\mu = 0$, we know the system is at half-filling, $\langle n \rangle = 1.0$. However, in practice this correct filling fraction is only reliably observed when reflection updates are enabled. For S_{cdw}, both reflection and swap updates help reduce decorrelation times. In practice, using some combination of reflection and swap updates makes sense, with reflection updates being crucial for the system to converge properly to the correct filling.

In addition to reducing decorrelation times, reflection and swap updates also help ameliorate a concern of ergodicity breaking [43,45,60]. If the phonon configuration x only smoothly evolves under the Hamiltonian dynamics (Sec. III), then it would be formally impossible to cross the nodal surface where $\det M = 0$, for which the action (S_{DQMC} or S) diverges. To be sure that we are sampling the entire space of phonon configurations, for which $\det M$ may change sign, we should also incorporate some discontinuous Monte Carlo updates that allow for jumps across nodal surfaces. The reflection and swap updates proposed in this section are therefore a good complement to pure HMC sampling.

V. PRECONDITIONING

A. Preconditioner algorithm

Each iteration of HMC requires solving the linear system in Eq. (31),

$$M^T M v = b, \tag{65}$$

for the unknown v. The required number of CG iterations to reach a fixed level of accuracy scales approximately like the condition number of M (equivalently, the square root of the condition number of $M^T M$).

Convergence can be accelerated if a good preconditioner P is available. One can solve for u in

$$P^{-T} M^T M P^{-1} u = P^{-T} b \tag{66}$$

and then determine $v = P^{-1} u$. This is advantageous if MP^{-1} has a smaller condition number than M, and if P^{-1} can be efficiently to applied to a vector. In practice, each iteration of preconditioned CG requires one matrix-vector multiplication using $M^T M$, and one using $(P^T P)^{-1}$ [58].

A good preconditioner frequently benefits from problem-specific insight. For the Holstein model we make use of the fact that the τ-fluctuations in the phonon fields are damped due to the contribution to the total action S from the bosonic action S_B in the sampling weight $\exp(-S)$. It follows that the imaginary-time fluctuations of the block matrices B_τ should be relatively small. Inspired by this, we propose a preconditioner P that retains the sparsity structure of M, Eq. (10), but with fluctuations in τ effectively "averaged out." Specifically, we define

$$P = \begin{pmatrix} I & & & & & \bar{B} \\ -\bar{B} & I & & & & \\ & -\bar{B} & I & & & \\ & & & \ddots & & \\ & & & & -\bar{B} & I \end{pmatrix}, \tag{67}$$

where

$$\bar{B} = \frac{1}{L_\tau} \sum_{\tau=0}^{L_\tau - 1} B_\tau = e^{-\Delta_\tau \bar{V}} e^{-\Delta_\tau K} \tag{68}$$

and $\bar{V}$ is defined to satisfy

$$e^{-\Delta_\tau \bar{V}} = \frac{1}{L_\tau} \sum_{\tau=0}^{L_\tau - 1} e^{-\Delta_\tau V_\tau}. \tag{69}$$

This preconditioner P can be interpreted as describing a semi-classical system for which imaginary-time fluctuations are suppressed. We emphasize, however, that our purpose with preconditioning is to solve the full Holstein model without any approximation.

Careful benchmarks of P as a preconditioner to M will be presented in Sec. V B. Here, we can briefly provide some intuition about why it should work. For small Δ_τ, we have $B_\tau \approx I - \Delta_\tau (V_\tau + K)$. At this order of approximation, $\bar{V}$ becomes the imaginary time average over V_τ. The bandwidth of the hopping matrix K on the square lattice is 8, whereas fluctuations in diagonal elements $(\bar{V} - V_\tau)_{ii} = \alpha(\bar{x}_i - x_{i,\tau})$ are typically order 1 or smaller for the models considered in this work (fluctuations are controlled by ω_0 when the dimensionless coupling λ is held fixed). The relatively small magnitude of these V_τ fluctuations suggests that P should be a good approximation to M for the dominant part of the eigenspectrum (larger eigenvalues). We note, however, that P is frequently observed to be *ineffective* at capturing eigenvectors associated with the smallest eigenvalues of M, and this seems to be the biggest limiting factor in the utility of P as a preconditioner.

An important, but nonobvious, property of this preconditioner is that the matrix-vector product $P^{-1} v$ can be evaluated very efficiently. To demonstrate this, we will first show that the matrix P becomes exactly block diagonal after an appropriate Fourier transformation in the imaginary time τ index.

The block structure of M in Eq. (10) treats $\tau = 0$ as a special case. To make all of B_τ appear symmetrically, we introduce a unitary matrix,

$$\Theta_{\tau, \tau'} = \delta_{\tau, \tau'} e^{-\pi i \tau / L_\tau}. \tag{70}$$

Observe that the matrix $\Theta M \Theta^\dagger$ has the same sparsity structure as M, but a factor of $-e^{-\pi i / L}$ appears in front of each B_τ, and the block B_0 is no longer a special case.

Next we may employ the discrete Fourier transformation $\mathcal{F}_{\omega, \tau}$ defined in Eq. (46). Under a combined change of basis, M becomes

$$\tilde{M} = \mathcal{U} M \mathcal{U}^\dagger, \tag{71}$$

where

$$\mathcal{U} = \mathcal{F} \Theta \tag{72}$$

is unitary, with matrix elements given by

$$\mathcal{U}_{\omega, \tau} = \frac{1}{\sqrt{L_\tau}} e^{-\frac{2\pi i}{L_\tau} \tau (\omega + 1/2)}. \tag{73}$$

By construction, the indices τ and ω range from 0 to $L_\tau - 1$. It is interesting to observe, however, that the extension of τ would naturally introduce *antiperiodic* boundary conditions

$(\mathcal{U}_{\omega,\tau+L_\tau} = -\mathcal{U}_{\omega,\tau})$, allowing ω to be interpreted as indexing Matsubara frequencies.

Explicit calculation gives the $N \times N$ blocks of $\tilde{M}$ as

$$\tilde{M}_{\omega,\omega'} = \delta_{\omega,\omega'} I - e^{-i\phi_{\omega'}} \frac{\hat{B}_{\omega-\omega'}}{\sqrt{L_\tau}}, \tag{74}$$

where

$$\phi_\omega = \frac{2\pi}{L_\tau}\left(\omega + \frac{1}{2}\right), \tag{75}$$

$$\hat{B}_\omega = \frac{1}{\sqrt{L_\tau}} \sum_{\tau=0}^{L_\tau-1} e^{-\frac{2\pi i}{L_\tau}\tau\omega} B_\tau. \tag{76}$$

We emphasize that $\tilde{M}$ is an exact representation of M, but in a different basis. When the fluctuations in imaginary time are small, $\tilde{M}$ is dominated by its diagonal blocks,

$$\tilde{M}_{\omega,\omega} = I - e^{-i\phi_\omega}\bar{B}, \tag{77}$$

where $\bar{B} = \hat{B}_0/\sqrt{L_\tau}$ coincides with Eq. (68).

We may *define* the preconditioner to be block diagonal in the Fourier basis,

$$\tilde{P} = \mathrm{diag}(\tilde{M}). \tag{78}$$

Transforming back to the original basis,

$$P = \mathcal{U}^\dagger \tilde{P} \mathcal{U}, \tag{79}$$

makes contact with the equivalent definition in Eq. (67).

To apply the preconditioner to a vector v, we must evaluate

$$P^{-1}v = \mathcal{U}^\dagger \tilde{P}^{-1}\mathcal{U}v. \tag{80}$$

The action of $\mathcal{U}$ and $\mathcal{U}^\dagger$ can be efficiently implemented using a fast Fourier transform (FFT). Because $\tilde{P}$ is block diagonal, its inverse is also block diagonal,

$$\tilde{P}^{-1}_{\omega,\omega'} = \delta_{\omega,\omega'} \tilde{M}^{-1}_{\omega,\omega}. \tag{81}$$

Therefore, applying $\tilde{P}^{-1}$ to a $(N \times L_\tau)$-dimensional vector $\hat{v} = \mathcal{U}v$ is equivalent to applying each of the $\tilde{M}^{-1}_{\omega,\omega}$ blocks to the corresponding N-dimensional sub-vector $\hat{v}_\omega$. In Appendix C we describe how the kernel polynomial method (KPM) [74] can be used to carry out efficiently these matrix-vector multiplications. The key idea is to approximate each of $\tilde{M}^{-1}_{\omega,\omega}$ using a numerically stable Chebyshev series expansion in powers of the matrix $\bar{B}$.

B. Preconditioner speedup

Here we present results that demonstrate the utility of our preconditioner P, while also providing insight into the scaling of HMC with both system size N and inverse temperature β. The overwhelming computational cost in HMC is repeatedly solving the linear system Eq. (31) for varying realizations of the phonon field $x_{i,\tau}$. If the number of CG iterations required to find a solution is independent of N, then the total simulation cost would scale near linearly with N.

In all cases, we terminate the CG iterations when the relative magnitude of the residual error,

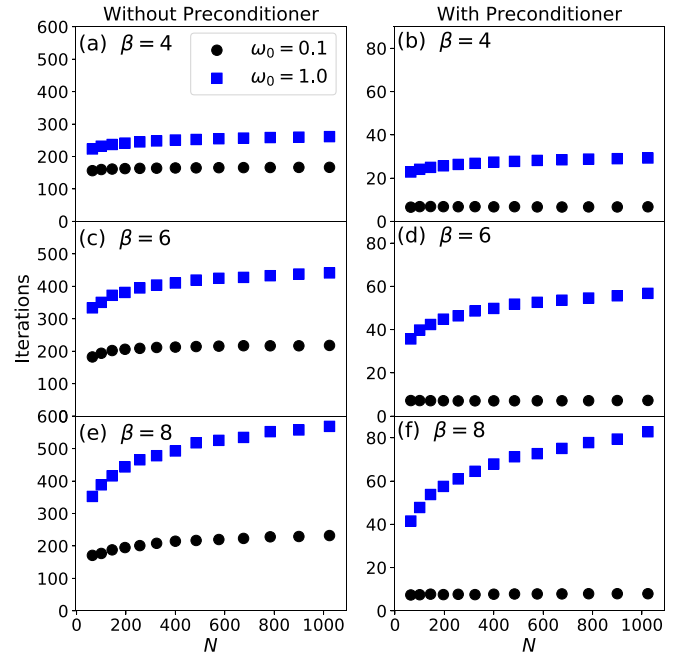$$\epsilon = |b - M^T M v|/|b| \tag{82}$$

FIG. 5. Average CG iteration count as a function of system size N for $\lambda = 0.25$. Comparing the left and right columns, the preconditioner significantly reduces the iteration count.

becomes less than a threshold value $\epsilon_{\max}$. When calculating ΔS_F in Eq. (27) to accept or reject a Monte Carlo update, we use $\epsilon_{\max} = 10^{-10}$. When calculating $\partial S_F/\partial x$ in Eq. (42) we use $\epsilon_{\max} = 10^{-5}$.

We benchmark using Holstein models of various systems sizes at two phonon frequencies $\omega_0 = 0.1$ and $\omega_0 = 1.0$, both with dimensionless coupling $\lambda = 0.25$. Figure 5 shows the average iteration count as a function of the number of lattice sites, N. For all temperatures and lattice sizes, the $\omega_0 = 0.1$ simulations require fewer CG iterations than comparable $\omega_0 = 1.0$ simulations. For $\omega_0 = 0.1$ without the preconditioner, the iteration count only weakly depends on system size. However, with the preconditioner the iteration count becomes nearly independent of system size, and is decreased by more than a factor of 20. For $\omega_0 = 1.0$, the growth of CG iteration count as a function of system size remains sub-linear. Introducing the preconditioner does not change the qualitative structure of this dependence, but still reduces the iteration count by more than a factor of 5 in all cases.

The dependence of iteration count on β is further explored in Fig. 6. For both $\omega_0 = 0.1$ and $\omega_0 = 1.0$, we observe a sharp jump in the order parameter S_{cdw} as the temperature is lowered, indicating that both systems order into a CDW phase. In the lower panel we see the average iteration count versus β. Both with and without the preconditioner, in the case of $\omega_0 = 1.0$ the iteration count increases monotonically with β. Simulations with $\omega_0 = 0.1$ have two qualitatively different behaviors: With preconditioning, the iteration count is relatively flat, but without preconditioning, the iteration count has a local maximum near where we would estimate the transition temperature to be based on S_{cdw}.

The preconditioner significantly reduces the average iteration count for both $\omega_0 = 0.1$ and 1, but the effect is more
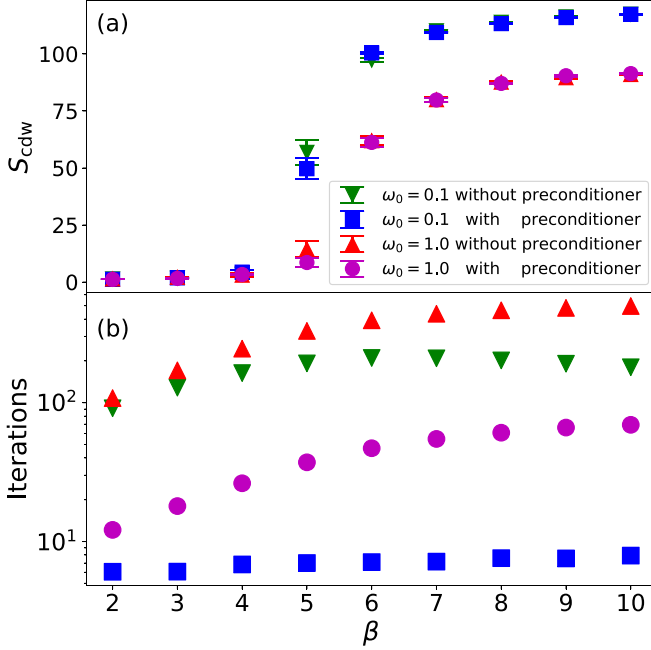
FIG. 6. $S_{\rm cdw}$ and average CG iteration count as functions of β for $\lambda = 0.25$ and $N = 16^2$. For both $\omega_0 = 0.1$ and $\omega_0 = 1.0$, the system goes through a CDW transition as the temperature is lowered. In the case of $\omega = 1.0$ the known transition temperature is approximately $\beta_{\rm cdw} \approx 6$.

pronounced for smaller ω_0, where imaginary time fluctuations are smaller. In the adiabatic limit, corresponding to the atomic mass going to infinity, the fluctuations in τ would vanish, and the preconditioner would become perfect. The adiabatic limit can equivalently be reached by sending the phonon frequency to zero ($\omega_0 \to 0$) while holding λ fixed.

The *practical* benefit of preconditioning depends strongly on the numerical cost C_P to apply the preconditioner P^{-1} to a vector. The natural reference scale is C_M, the cost to apply the unpreconditioned matrix M to a vector. In our implementation, we measure $(C_P + C_M)/C_M \approx 4$, approximately independent of model details (see Sec. C 5 for a theoretical analysis). At $\omega_0 = 0.1$, preconditioning reduces the iteration count by about a factor of 20, yielding an effective speedup of order $20/4 = 5$.

Wall-clock times for the simulation results reported in Figs. 5 and 6 can be found in Appendix D. The results confirm that the computational cost scales near-linearly in both system size and inverse temperature β. Furthermore, the speedups due to preconditioning are very close to the estimates given above.

VI. STOCHASTIC MEASUREMENTS WITH FFT ACCELERATION

In a traditional determinant QMC code, measurements of the Green function are obtained by explicit construction of the matrix M^{-1}. This cubic-scaling cost can be avoided by using stochastic techniques to estimate individual matrix elements. We review these methods, and then demonstrate how to efficiently average Green function elements over all space and imaginary times by using the FFT algorithm. Finally, we

will introduce a strategy to reduce the relatively large stochastic errors that appear when forming stochastic estimates of multiple-point correlation functions.

A. Measurements in QMC

A fundamental observable in QMC simulation is the time-ordered, single-particle Green function,

$$
\mathcal{G}_{i,j}(\tau) = \begin{cases} \langle \hat{c}_i(\tau) \hat{c}_j^\dagger(0) \rangle, & 0 \leqslant \tau < \beta, \\ -\langle \hat{c}_j^\dagger(0) \hat{c}_i(\tau) \rangle, & -\beta \leqslant \tau < 0, \end{cases} \tag{83}
$$

where $\hat{c}_i(\tau) \equiv e^{\tau H} \hat{c}_i e^{-\tau H}$ denotes evolution of the electron annihilation operator in continuous imaginary time τ. Multipoint correlation functions can be expressed as sums of products of single-particle Green functions via Wick's theorem [56,75]. Given an equilibrium sample of the phonon field, the matrix $G = M^{-1}$ provides an unbiased estimate of the Green function,

$$
\mathcal{G}_{i,j}(\tau) \approx G_{(i,l),(j,l')}, \tag{84}
$$

where $\tau = \Delta_\tau \cdot (l - l')$ satisfies $-\beta < \tau < \beta$. In what follows we will revert to using the symbol $\tau = 0, 1 \ldots, (L_\tau - 1)$ as a matrix index instead of a continuous imaginary time.

B. Stochastic approximation of the Green function

In a traditional determinant QMC code, one would explicitly calculate the full matrix $G = M^{-1}$ at a cost that scales cubically in system size. To reduce this cost, we instead employ the unbiased stochastic estimator

$$
G \approx (G\xi)\xi^T, \tag{85}
$$

for a random vector ξ with components that satisfy $\langle \xi_i \rangle = 0$ and $\langle \xi_i \xi_j \rangle = \delta_{i,j}$. For example, each component ξ_i may be sampled from a Gaussian distribution, or uniformly from $\{\pm 1\}$. The bold symbol $\boldsymbol{i}$ represents a combined site and imaginary-time index, (i, τ). Equation (85) may be viewed as a generalization the Hutchinson trace estimator $\mathrm{Tr}\, G \approx \xi^\dagger G \xi$ [76]. Various strategies are possible to reduce the stochastic error [77,78].

The vector $v = G\xi$ can be calculated iteratively at a cost that scales near-linearly with system size. For example, one may solve the linear system $M^T M v = M^T \xi$ using CG with preconditioning (cf. Sec. V).

Once $G\xi$ is known, individual matrix elements can be efficiently approximated,

$$
G_{i,j} \approx (G\xi)_i \xi_j. \tag{86}
$$

For products of Green functions elements, we may use

$$
G_{i,j} G_{k,l} \approx (G\xi)_i \xi_j (G\xi')_k \xi'_l. \tag{87}
$$

This product of estimators remains an unbiased estimator provided that the random vectors ξ and ξ' are mutually independent.

C. Averaging over space and imaginary time using FFTs

To improve the quality of statistical estimates, it is frequently desirable to average Green function elements over all

space and imaginary-time,

$$\mathcal{G}_{\mathbf{\Delta}} \approx \frac{1}{\mathcal{N}} \sum_i G^{\mathrm{ext}}_{(i+\mathbf{\Delta}),i}, \tag{88}$$

where $\mathcal{N} = N L_\tau$. The symbol $\mathbf{\Delta}$ indicates a displacement in both position and imaginary-time. The matrix G^{ext} will be defined below as an extension of G that accounts for antiperiodicity of imaginary time. Using direct summation, the total cost to calculate $\mathcal{G}_{\mathbf{\Delta}}$ for every possible displacement $\mathbf{\Delta}$ would scale like $\mathcal{O}(\mathcal{N}^2)$. However, we will describe a method using FFTs that reduces the cost to approximately $\mathcal{O}(\mathcal{N} \ln \mathcal{N})$.

Consider a finite, D-dimensional lattice with periodic boundary conditions. For a Bravais lattice, each site can be labeled by integer coordinates, $0 \leqslant n_d < L_d$, where L_d is the linear system size for dimension d. The combined index $\boldsymbol{i} = (n_1, \dots, n_D, \tau)$ can then be interpreted as integer coordinates for both space and imaginary-time; the index $(\boldsymbol{i} + \mathbf{\Delta})$ can be interpreted as a displacement of all $(D+1)$ coordinates. We must be careful, however, with boundary conditions. The Green function is *antiperiodic* in continuous imaginary time, $\mathcal{G}_{i,j}(\tau + \beta) = -\mathcal{G}_{i,j}(\tau)$. To encode this antiperiodicity in matrix elements, we define

$$G^{\mathrm{ext}} = QGQ^T = \begin{bmatrix} G & -G \\ -G & G \end{bmatrix}, \tag{89}$$

where

$$Q = \begin{bmatrix} I \\ -I \end{bmatrix}. \tag{90}$$

The extended matrix G^{ext} effectively doubles the range of the imaginary time index, $0 \leqslant \tau < 2L_\tau$, such that space *and* imaginary time indices become periodic,

$$n_d + L_d \equiv n_d, \quad \tau + 2L_\tau \equiv \tau. \tag{91}$$

Using Eq. (85), we obtain a stochastic approximation for the time averaged Green function elements,

$$\mathcal{G}_{\mathbf{\Delta}} \approx \frac{1}{\mathcal{N}} \sum_i (QG\xi\xi^T Q^T)_{(i+\mathbf{\Delta}),i}$$
$$= \frac{1}{\mathcal{N}} \sum_i a_i b_{i+\mathbf{\Delta}}, \tag{92}$$

involving the vectors

$$a = Q\xi = \begin{bmatrix} \xi \\ -\xi \end{bmatrix}, \quad b = QG\xi = \begin{bmatrix} G\xi \\ -G\xi \end{bmatrix}. \tag{93}$$

This can be written

$$\mathcal{G}_{\mathbf{\Delta}} \approx \frac{1}{\mathcal{N}} (a \star b)_{\mathbf{\Delta}}, \tag{94}$$

where $a \star b$ denotes the circular cross-correlation. Like the convolution operation, it can be expressed using ordinary multiplication in Fourier space,

$$(a \star b)_{\mathbf{\Delta}} = \mathcal{F}^{-1}\{\mathcal{F}[a]^* \mathcal{F}[b]\}_{\mathbf{\Delta}}. \tag{95}$$

Here, $\mathcal{F}$ denotes the $(D+1)$-dimensional discrete Fourier transform. This formulation allows using the FFT algorithm to estimate $\mathcal{G}_{\mathbf{\Delta}}$ at near-linear scaling cost.

In the QMC context, Wick's theorem ensures that multipoint correlation functions can always be reduced to products of ordinary Green functions. The latter can be estimated using a product of independent stochastic approximations, as in Eq. (87). Here, again, we can accelerate space and imaginary-time averages using FFTs. In the case of four-point measurements, Wick's theorem produces three types of Green function products. The first is

$$\sum_i G_{i+\mathbf{\Delta},i} G_{i+\mathbf{\Delta},i} \approx \sum_i (G\xi\xi^T)_{i+\mathbf{\Delta},i} (G\xi'\xi'^T)_{i+\mathbf{\Delta},i}$$
$$= \sum_i [\xi_i \xi'_i][(G\xi)_{i+\mathbf{\Delta}} (G\xi')_{i+\mathbf{\Delta}}], \tag{96}$$

which is again recognized as a cross correlation $\star$. This can be expressed compactly by introducing $\odot$ to denote elementwise multiplication of vectors,

$$\sum_i G_{i+\mathbf{\Delta},i} G_{i+\mathbf{\Delta},i} \approx [(\xi \odot \xi') \star (G\xi \odot G\xi')]_{\mathbf{\Delta}}, \tag{97}$$

The other two averages that appear for four-point measures can be expressed similarly,

$$\sum_i G_{i+\mathbf{\Delta},i+\mathbf{\Delta}} G_{i,i} \approx [(\xi \odot G\xi) \star (\xi' \odot G\xi')]_{\mathbf{\Delta}}, \tag{98}$$

$$\sum_i G_{i+\mathbf{\Delta},i} G_{i,i+\mathbf{\Delta}} \approx [(\xi \odot G\xi') \star (\xi' \odot G\xi)]_{\mathbf{\Delta}}. \tag{99}$$

D. Reducing stochastic error in multipoint correlation function estimates

To reduce the stochastic error in Eq. (86), we may average over a collection of random vectors, $[\xi_1, \dots \xi_{N_{\mathrm{rv}}}]$,

$$G_{i,j} \approx \frac{1}{N_{\mathrm{rv}}} \sum_{n=1}^{N_{\mathrm{rv}}} (G\xi_n \xi_n^\dagger)_{i,j}. \tag{100}$$

A similar strategy could be used to replace Eq. (87) with an average over N_{rv} independent estimates.

A significant reduction in error is possible by averaging over all $\binom{N_{\mathrm{rv}}}{2} = N_{\mathrm{rv}}(N_{\mathrm{rv}} - 1)/2$ pairs of random vectors,

$$G_{i,j} G_{k,l} \approx \binom{N_{\mathrm{rv}}}{2}^{-1} \sum_{n<m} (G\xi_n \xi_n^\dagger)_{i,j} (G\xi_m \xi_m^\dagger)_{k,l}. \tag{101}$$

This improved estimator is an average of unbiased estimators and therefore remains unbiased. Furthermore, if N_{rv} is much smaller than the vector dimension $\mathcal{N}$, then these $\binom{N_{\mathrm{rv}}}{2} \approx N_{\mathrm{rv}}^2$ estimates are approximately mutually independent. It follows that the stochastic error in Eq. (101) decays approximately like N_{rv}^{-1}. This scheme is advantageous because, for moderate N_{rv}, the dominant computational cost is calculating the N_{rv} matrix-vector products $[G\xi_1, \dots G\xi_{N_{\mathrm{rv}}}]$. There remains the task of evaluating the sum over all pairs $n \neq m$. For each pair, we must evaluate cross-correlations as in Eq. (95), but the required FFTs are relatively fast.

Figure 7 demonstrates how the improved stochastic approximator in Eq. (101) can significantly reduce error bars for certain observables in QMC simulation. Measurements and corresponding estimated errors are plotted as a function of N_{rv}. For the observables $\langle n \rangle$ and S_{cdw}, the error appears largely independent N_{rv}; in these two cases, the dominant source of statistical error seems to be limited by the effective number of independent phonon configurations sampled.
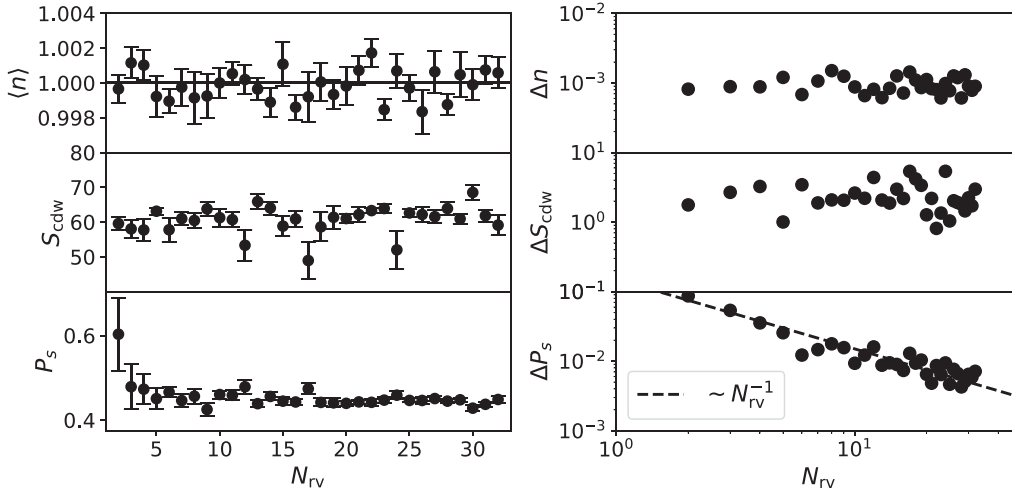
FIG. 7. Left: QMC measurements, employing stochastic Green function estimation with N_{rv} random vectors. Right: Error for each measured quantity. Simulations were performed using $\omega = 1$, $\lambda = 0.25$, $\beta = 6$, $L = 16$, $N_{therm} = 1000$ and $N_{sim} = 2000$.

For the observable P_s, however, we find the error ΔP_s to depend strongly on the quality of the stochastic estimator, controlled by N_{rv}. The observed scaling $\Delta P_s \sim N_{rv}^{-1}$ matches the theoretical expectation for stochastic error in Eq. (101). This indicates that the stochastic measurements are the primary source of error in P_s.

It is also important to consider the relative computational cost of measurements as N_{rv} increases. Figure 8 plots the time spent making measurements t_{meas}, relative to the total simulation time t_{total}, versus N_{rv}. Even at the maximum value of $N_{rv} = 32$ tested, the time spent making measurements is significantly less than half the total run-time. The fact that the t_{meas}/t_{total} grows linearly at large N_{rv} indicates that calculating the matrix-vector products $G\xi_n$ is the dominant computational cost in the measurement process. The curvature at small N_{rv} is a result of t_{total} including the overhead time spent equilibrating the system before measurements begin. A practical limitation on N_{rv} may be memory usage, since Eq. (101) requires
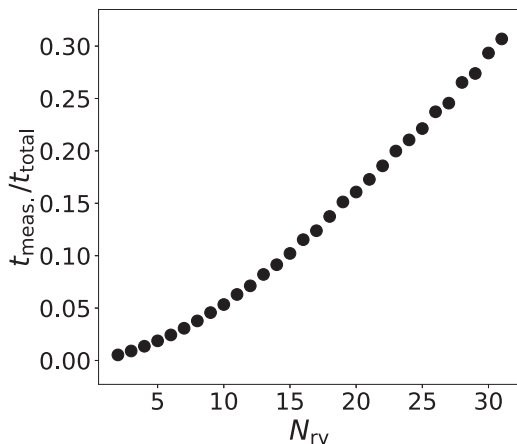

FIG. 8. Wall-clock time spent taking measurements t_{meas} relative to the total run-time t_{total} as a function of N_{rv}. Simulation parameters are the same as in Fig. 7.

that all vectors $[\xi_1, \ldots \xi_{N_{rv}}]$ and $[G\xi_1, \ldots G\xi_{N_{rv}}]$ be stored simultaneously.

Although N_{rv} appears to have little impact on some observables, it seems reasonable to set $N_{rv} \gtrsim 10$ in most cases, given the negligible computational costs.

VII. CONCLUSION

This paper introduces a set of algorithms that collectively enable highly scalable, finite temperature simulations of electron-phonon models such as the Holstein and SSH models. Traditionally, such studies would be performed using DQMC, but that approach is limited in two important respects.

First, with a computational cost that scales cubically with system size, DQMC simulations of the Holstein model have been restricted to lattices of no more than a few hundred sites. As a result, DQMC studies of the Holstein model have typically been confined to relatively simple geometries in one or two dimensions. In the HMC approach explored in this paper, we replace each Fermion determinant $\det M(x)$ that appears in DQMC with a Gaussian integral over a newly introduced auxiliary field Φ_σ (Sec. II D). This field must be multiplied by the inverse matrix $M^{-1}(x)$; for this, we use the iterative conjugate gradient (CG) method, with a computational cost that scales near-linearly with system size. As a result, it becomes possible to simulate lattice sizes a full order of magnitude larger than is possible with DQMC. We accelerate CG convergence by introducing a preconditioner P that retains the structure of $M(x)$, but discards fluctuations in imaginary time (Sec. V). These advances open the door to studying both more complicated multiband models in two dimensions, as well as three-dimensional systems.

Second, DQMC simulations rely on a local updating scheme that results in long autocorrelation times that increase with decreasing phonon frequency. This has restricted DQMC simulations to systems where the phonon energy is comparable to the hopping amplitude, $\omega_0 \sim t$. However, in most real materials the relative phonon energy is much smaller, $\omega_0 \ll t$. We address this limitation by using HMC to update efficiently

the entire phonon field simultaneously. To do so, we employ a Hamiltonian dynamics with a carefully defined dynamical mass matrix that slows down the modes with highest frequency in imaginary time, which counteracts stiffness in the bosonic action S_B (Sec. III B 1). Additionally, we employ a time-step splitting algorithm (Sec. III B 2) that uses a smaller time step to integrate the bosonic forces $-\partial S_F/\partial x$, relative to the time step for the fermionic forces. As a result, we are able to simulate efficiently electron-phonon models with small phonon frequencies, which are of greatest physical relevance for real materials.

At moderate to strong electron-phonon coupling, simulations of the Holstein model also suffer from long autocorrelation times as a result of the phonon-mediated, electron-electron binding. We introduce two additional types of Monte Carlo updates, termed reflection and swap updates, to address this issue. While similar types of updates have been employed in DQMC simulations of the Holstein model, we are able to do so while maintaining near linear scaling with system size.

Finally, we introduce techniques for efficiently measuring correlation functions. Elements of the matrix $M^{-1}(x)$ can be estimated stochastically, provide samples of the single-particle Green's function. It is frequently desirable to average correlation measurements over both real space and imaginary time to reduce the error. A straightforward approach to performing this average results in a computational cost that scales as $\mathcal{O}(N^2 L_\tau^2)$, which would violate our target of near linear-scaling cost. To recover the desired scaling, we formulated the real space and imaginary time averages as cross-correlations (with periodic boundaries), which enables their efficient evaluation using FFTs. As a consequence, measurements come almost "for free," relative to the computational work required to sample the phonon field.

Electron-phonon interactions play an important role in describing emergent behaviors that occur in certain strongly interacting materials. The methods outlined in this paper allow for the efficient simulation of electron-phonon models over a much greater range of system sizes and parameter regimes, than was previously possible. This capability makes accessible the study of many new material systems where electron-phonon interactions are believed to play a prominent role in determining the low energy physics.

ACKNOWLEDGMENTS

The code is open source and available on Github [81].

APPENDIX A: REVIEW OF PATH INTEGRAL FORMALISM

Here we review how the partition function for the Holstein model,

$$\mathcal{Z} = \text{tr}_{\text{el-ph}} \, e^{-\beta \hat{H}}, \tag{A1}$$

can be formulated as a path integral over phonon fields. The trace is over the combined Fock space for both electron and phonon operators. The Suzuki-Trotter approximation yields [79]

$$\mathcal{Z} \approx \text{tr}_{\text{el-ph}} \left[e^{-\frac{\Delta_\tau}{2} \hat{H}_{\text{el-ph}}} e^{-\Delta_\tau (\hat{H}_{\text{el}} + \hat{H}_{\text{ph}})} e^{-\frac{\Delta_\tau}{2} \hat{H}_{\text{el-ph}}} \right]^{L_\tau}$$
$$= \text{tr}_{\text{el-ph}} \left[e^{-\Delta_\tau \hat{H}_{\text{el-ph}}} e^{-\Delta_\tau \hat{H}_{\text{el}}} e^{-\Delta_\tau \hat{H}_{\text{ph}}} \right]^{L_\tau}, \tag{A2}$$

where $\beta = \Delta_\tau L_\tau$ is the discretization in imaginary time. This approximation is valid to order $\mathcal{O}(\Delta_\tau^2)$. In the second step we used the fact that $\hat{H}_{\text{ph}}$ and $\hat{H}_{\text{el}}$ commute, and the cyclic property of the trace.

The next step is to evaluate the phonon trace in the position basis. This is done by repeatedly inserting the identity operator $\int d^N x \, |x\rangle \langle x|$, where $|x\rangle = |x_1, x_2, \ldots x_N\rangle$ denotes an entire real-space phonon configuration, such that the integral is understood to be over all sites. Using $\langle x_\tau | x_{\tau+1} \rangle = \delta(x_\tau - x_{\tau+1})$, the result is

$$\mathcal{Z} \approx \text{tr}_{\text{el}} \int \mathcal{D}x \prod_{\tau=0}^{L_\tau-1} e^{-\Delta_\tau \hat{H}_{\text{el-ph}}(x_\tau)} e^{-\Delta_\tau \hat{H}_{\text{el}}} \langle x_\tau | e^{-\Delta_\tau \hat{H}_{\text{ph}}} | x_{\tau+1} \rangle, \tag{A3}$$

where the differential $\mathcal{D}x$ indicates a path integral over all phonon fields $x_{i,\tau}$. $\hat{H}_{\text{el-ph}}(x_\tau)$ denotes the operator $\hat{H}_{\text{el-ph}}$ with the replacement $\hat{X} \mapsto x_\tau$, subject to the periodic boundary condition $x_{L_\tau} \equiv x_0$. Next we write

$$\mathcal{Z} \approx \text{tr}_{\text{el}} \int \mathcal{D}x \, e^{-S_B} \prod_{\tau=0}^{L_\tau-1} e^{-\Delta_\tau \hat{H}_{\text{el-ph}}(x_\tau)} e^{-\Delta_\tau \hat{H}_{\text{el}}}, \tag{A4}$$

where

$$e^{-S_B} = \prod_{\tau=0}^{L_\tau-1} \langle x_\tau | e^{-\Delta_\tau \hat{H}_{\text{ph}}} | x_{\tau+1} \rangle. \tag{A5}$$

Again using a symmetric operator splitting,

$$e^{-\Delta_\tau \hat{H}_{\text{ph}}} \approx e^{-\Delta_\tau \frac{\omega_0^2}{4} \hat{X}^2} e^{-\Delta_\tau \frac{1}{2} \hat{P}^2} e^{-\Delta_\tau \frac{\omega_0^2}{4} \hat{X}^2}, \tag{A6}$$

we find

$$\langle x_\tau | e^{-\Delta_\tau \hat{H}_{\text{ph}}} | x_{\tau+1} \rangle \approx e^{-\frac{\Delta_\tau \omega_0^2}{4}(x_\tau^2 + x_{\tau+1}^2)} \langle x_\tau | e^{-\Delta_\tau \frac{1}{2} \hat{P}^2} | x_{\tau+1} \rangle, \tag{A7}$$

which is locally valid to $\mathcal{O}(\Delta_\tau^3)$. In this notation, we are treating x_τ and $\hat{P}$ as N-component vectors. The second factor can be evaluated by inserting a complete set of momentum

states,

$$\langle x_\tau | e^{-\Delta_\tau \frac{1}{2} \hat{p}^2} | x_{\tau+1} \rangle = \int d^N p \, \langle x_\tau | p \rangle e^{-\Delta_\tau \frac{1}{2} p^2} \langle p | x_{\tau+1} \rangle$$

$$= \int d^N p \, e^{-\frac{\Delta_\tau}{2} p^2 + i p \cdot (x_{\tau+1} - x_\tau)}$$

$$\propto e^{-\frac{\Delta_\tau}{2} \left(\frac{x_{\tau+1} - x_\tau}{\Delta_\tau} \right)^2}. \tag{A8}$$

Combining Eqs. (A4)–(A8), and recalling that $x_{L_\tau} = x_0$, we arrive at the "bosonic action" for the phonons,

$$S_B \approx \Delta_\tau \sum_{i=1}^{N} \sum_{\tau=0}^{L_\tau-1} \left[\frac{1}{2} \omega_0^2 x_{i,\tau}^2 + \frac{(x_{i,\tau+1} - x_{i,\tau})^2}{2\Delta_\tau^2} \right] + \text{const.} \tag{A9}$$

This approximation is valid to order $\mathcal{O}(\Delta_\tau^2)$ because we have chained the approximation in Eq. (A7) order $1/\Delta_\tau$ times.

With some algebraic rearrangement, the partition function in Eq. (A4) may be written

$$\mathcal{Z} \approx \int \mathcal{D}x \, e^{-(S_B - \Delta_\tau \alpha \sum_{i,\tau} x_{i,\tau})} \text{tr}_{\text{el}} \prod_{\tau=0}^{L_\tau-1} \prod_{\sigma=\uparrow,\downarrow} e^{-\Delta_\tau \hat{V}_{\tau,\sigma}} e^{-\Delta_\tau \hat{K}_\sigma},$$

where

$$\hat{V}_{\tau,\sigma} = \sum_i (\alpha x_{i,\tau} - \mu) \hat{n}_{i,\sigma}, \tag{A10}$$

$$\hat{K}_\sigma = -\sum_{ij} t_{ij} \hat{c}_{i,\sigma}^\dagger \hat{c}_{j,\sigma}, \tag{A11}$$

are purely quadratic in the Fermions, making it possible to evaluate the remaining electron trace. Since the two spin sectors are not coupled, the result is [1]

$$\text{tr}_{\text{el}} \prod_{\tau=0}^{L_\tau-1} \prod_{\sigma=\uparrow,\downarrow} e^{-\Delta_\tau \hat{V}_{\tau,\sigma}} e^{-\Delta_\tau \hat{K}_\sigma} = (\det M)^2.$$

where M is a $NL_\tau \times NL_\tau$ matrix, conveniently expressed in block form,

$$M(x) = \begin{pmatrix} I & & & & B_0 \\ -B_1 & I & & & \\ & -B_2 & \ddots & & \\ & & \ddots & \ddots & \\ & & & -B_{L_\tau-1} & I \end{pmatrix}. \tag{A12}$$

I is the $N \times N$ identity matrix, and

$$B_\tau = e^{-\Delta_\tau V_\tau} e^{-\Delta_\tau K}.$$

The V_τ and K are matrix counterparts of the Fock-space operators of Eqs. (A10) and (A11), with elements

$$(V_\tau)_{ij} = \delta_{ij} (\alpha x_{i,\tau} - \mu), \quad K_{ij} = -t_{ij}.$$

Putting together the pieces, the partition function may be approximated,

$$\mathcal{Z} \approx \int \mathcal{D}x \, e^{-(S_B - \Delta_\tau \alpha \sum_{i,\tau} x_{i,\tau})} (\det M)^2, \tag{A13}$$

which is valid up to an error of order $\mathcal{O}(\Delta_\tau^2)$.

APPENDIX B: STATISTICAL SYMMETRY OF THE ACTION

Here we demonstrate how the particle-hole symmetry of the single-site Holstein model at half-filling emerges in the action $S(x, \Phi_\sigma)$ of Eq. (21), provided that imaginary-time fluctuations can be ignored.

Consider the change in action

$$\Delta S(x) = S(x) - S(x_0), \tag{B1}$$

for a move $x_0 \to x$. For particle-hole symmetry to be respected, we should find

$$\Delta S(x) \stackrel{?}{=} \Delta S(-x), \tag{B2}$$

such that MC proposals $x_0 \to x$ and $x_0 \to -x$ would be accepted with equal probability. This condition is equivalent to vanishing

$$\delta S = S(-x) - S(x). \tag{B3}$$

Observe that the starting configuration x_0 is irrelevant. Let us now investigate the condition $\delta S = 0$.

The bosonic action $S_B(x)$ defined in Eq. (9) is symmetric at half filling, but symmetry *breaking* may arise from the fermionic action $S_F(x, \Phi_\sigma)$ defined in Eq. (22). The result is

$$\delta S = \frac{1}{2} \sum_\sigma \Phi_\sigma^T \left(D_{-x}^{-1} - D_x^{-1} \right) \Phi_\sigma, \tag{B4}$$

where

$$D_x = A_x^T A_x, \tag{B5}$$

and the auxiliary field Φ_σ is arbitrary. If $D_x = D_{-x}$, then $\delta S = 0$, and the particle-hole symmetry of Eq. (B2) would be satisfied.

We now show that D_x indeed satisfies this symmetry in the special case of the adiabatic limit of the single-site Holstein model at half-filling ($\mu = 0$). Without the hopping matrix K, the block matrices $B_\tau = e^{-\Delta_\tau \alpha x_\tau}$ become effectively scalar. In the absence of imaginary-time fluctuations, we replace $B_\tau \to \bar{B} = e^{-\Delta_\tau \alpha \bar{x}}$. Next, we explicitly calculate $A^T = \Lambda^T M^T$ using Eqs. (10) and (16),

$$A_{\bar{x}}^T = \begin{pmatrix} \bar{B}^{1/2} & & & & \bar{B}^{-1/2} \\ -\bar{B}^{-1/2} & \bar{B}^{1/2} & & & \\ & -\bar{B}^{-1/2} & \ddots & & \\ & & \ddots & \ddots & \\ & & & -\bar{B}^{-1/2} & \bar{B}^{1/2} \end{pmatrix}. \tag{B6}$$

The subscript $\bar{x}$ emphasizes our neglect of imaginary-time fluctuations. It follows,

$$D_{\bar{x}} = \begin{pmatrix} \bar{B} + \bar{B}^{-1} & -I & & & I \\ -I & \bar{B} + \bar{B}^{-1} & \ddots & & \\ & -I & \ddots & & \\ & & \ddots & \ddots & -I \\ I & & & -I & \bar{B} + \bar{B}^{-1} \end{pmatrix}, \tag{B7}$$

The transformation $\bar{x} \to -\bar{x}$ corresponds to $\bar{B} \to \bar{B}^{-1}$. We conclude $D_{\bar{x}} = D_{-\bar{x}}$, as claimed, which implies particle-hole symmetry of the action, Eq. (B2). The result is exact in the

adiabatic limit (infinite atomic mass), for which imaginary-time fluctuations can be ignored.

APPENDIX C: PRECONDITIONER IMPLEMENTATION

In Sec. V we described a preconditioner P that is block diagonal in the Fourier space representation. Along the diagonal, its $N \times N$ blocks have the form

$$\tilde{M}_{\omega,\omega} = I - e^{-i\phi_\omega}\bar{B}, \tag{C1}$$

where

$$\phi_\omega = \frac{2\pi}{L_\tau}\left(\omega + \frac{1}{2}\right), \quad \bar{B} = e^{-\Delta_\tau \bar{V}}e^{-\Delta_\tau K}, \tag{C2}$$

and both $\bar{V}$ and K are Hermitian matrices. Applying P^{-1} to a vector requires application of the $N \times N$ matrices $\tilde{M}_{\omega,\omega}^{-1}$, for all indices $\omega = 0, 1, \dots L_\tau - 1$. Here we describe how the kernel polynomial method (KPM) [74] may be used to perform these matrix-vector products efficiently. This approach systematically approximates each matrix $\tilde{M}_{\omega,\omega}^{-1}$ in polynomials of $\bar{B}$.

A first observation is that the matrices $e^{-\Delta_\tau \bar{V}}$ and $e^{-\Delta_\tau K}$ in their exact forms are positive definite and Hermitian. From this, we can guarantee that all eigenvalues of $\bar{B}$ are real [80]. The checkerboard approximation to $e^{-\Delta_\tau K}$ slightly violates Hermiticity, but even in this case, we have observed empirically that the eigenvalues of $\bar{B}$ remain exactly real in the context of our QMC simulations.

A second observation is that the eigenvalues $\bar{b}$ of $\bar{B}$ are bounded near 1,

$$\bar{b}_{\min} \leqslant \bar{b} \leqslant \bar{b}_{\max}, \tag{C3}$$

otherwise Δ_τ would not be sufficiently small for the Suzuki-Trotter expansion to be meaningful. In the Holstein model, K will typically have a much larger spectral magnitude than $\bar{V}$, so we can get the correct scaling with the approximation $\bar{B} \approx e^{-\Delta_\tau K}$. On the square lattice with hopping $t = 1$, the extreme eigenvalues of K are ± 4. Given our choice of $\Delta_\tau = 0.1$, the extreme eigenvalues will be of order $\exp(\pm\Delta_\tau 4)$, namely, $\bar{b}_{\min} \approx 0.7$ and $\bar{b}_{\max} \approx 1.6$.

It will be convenient to define a rescaled matrix,

$$A = 2(\bar{B} - \bar{b}_{\min})/\Delta\bar{b} - 1, \tag{C4}$$

with $\Delta\bar{b} = \bar{b}_{\max} - \bar{b}_{\min}$. The eigenvalues y of A satisfy $-1 \leqslant y \leqslant 1$. This will allow us to approximate

$$\tilde{M}_{\omega,\omega}^{-1} = (1 - e^{-i\phi_\omega}\bar{B})^{-1} = f_\omega(A), \tag{C5}$$

using Chebyshev polynomials in A. We may view

$$f_\omega(y) = (1 - e^{-i\phi_\omega}\bar{b})^{-1}, \tag{C6}$$

as a scalar function that acts on the eigenvalues y of A, which are related to the eigenvalues $\bar{b}$ of $\bar{B}$ via

$$y = 2(\bar{b} - \bar{b}_{\min})/\Delta\bar{b} - 1. \tag{C7}$$

1. Chebyshev polynomial approximation

An arbitrary scalar function $f(y)$ may be expanded in the basis of Chebyshev polynomials,

$$f(y) = \sum_{m=0}^{\infty} c_m T_m(y), \tag{C8}$$

valid for $-1 \leqslant y \leqslant 1$. In this domain, the Chebyshev polynomials can be written $T_m(y) = \cos(m \arccos y)$, such that the coefficients c_m may be interpreted as the cosine transform of f in the variable $\theta = \arccos(y)$.

The Chebyshev polynomials satisfy a generalized orthogonality relation,

$$\int_{-1}^{+1} w(y)T_m(y)T_{m'}(y)dy = q_m \delta_{m,m'}, \tag{C9}$$

where

$$w(y) = (1 - y^2)^{-1/2},$$
$$q_m = \frac{\pi}{2}(1 + \delta_{m,0}).$$

The expansion coefficients are then given by

$$c_m = \frac{1}{q_m}\int_{-1}^{+1} w(y)T_m(y)f(y)dy. \tag{C10}$$

Usually a closed form solution for c_m is not available, but one can use Chebyshev-Gauss quadrature to obtain a good approximation,

$$c_m \approx \frac{\pi}{q_m N_Q}\sum_{n=0}^{N_Q-1} \cos(m\theta_n)f(\cos\theta_n), \tag{C11}$$

where N_Q is the number of quadrature points, and $\theta_n = \pi(n + \frac{1}{2})/N_Q$ are the abscissas. A fast Fourier transform can be used to calculate all coefficients c_m efficiently [74].

The utility of the expansion in Eq. (C8) is that we can obtain a good approximation by truncating

$$f(y) \approx \sum_{m=0}^{N_P-1} g_m c_m T_m(y), \tag{C12}$$

at an appropriate polynomial order N_P. Here one has the option to introduce damping factors g_m associated with a *kernel*. The damping factors should be close to 1 for $m \ll N_P$ and may decay to 0 as $m \to N_P$. An appropriately selected kernel guarantees uniform convergence of the Chebyshev series, avoiding numerical artifacts such as Gibbs oscillations. In our application, we are working with the smooth functions in Eq. (C6), and we will simply set $g_m = 1$.

For a given polynomial order N_P, we find it sufficient to use $N_Q = 2N_P$ quadrature points to approximate the expansion coefficients c_m in Eq. (C11).

2. Selecting the polynomial order

Figure 9 illustrates Chebyshev approximation of the real and imaginary parts of $f_\omega = (1 - e^{-i\phi_\omega}\bar{b})^{-1}$ for various polynomial orders N_P. Angles ϕ_ω near zero give rise to sharper features in f_ω, which require a larger polynomial order N_P to resolve.

We will use the convention that the angle ϕ_ω is between 0 and π. This effectively restricts our attention to $0 \leqslant \omega < L_\tau/2$, which is possible due to the symmetry $f_{L_\tau-\omega-1}(\bar{b}) = f_\omega^*(\bar{b})$.

In practice, we can achieve a good polynomial approximation using the heuristic

$$N_P = \lfloor \Delta\bar{b}(a_1\phi_\omega^{-1} + a_2)\rfloor, \tag{C13}$$
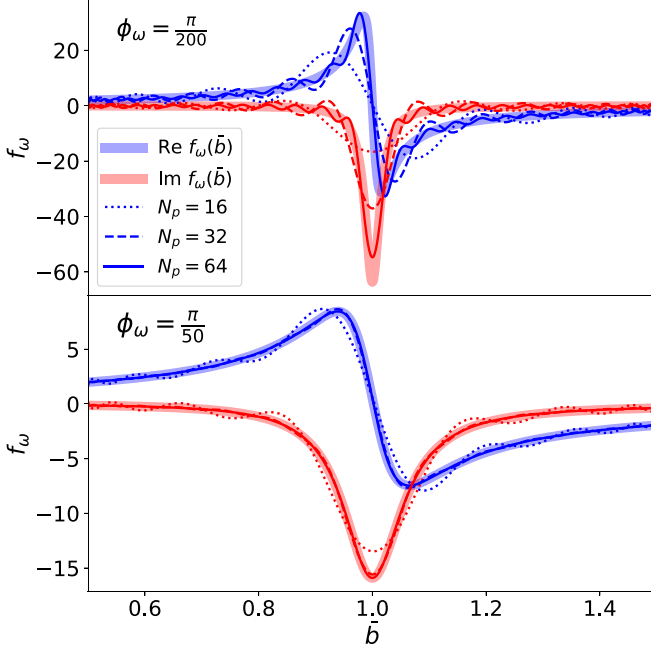
FIG. 9. Chebyshev polynomial approximation of $f_\omega = (1 - e^{-i\phi_\omega}\bar{b})^{-1}$ on a given interval $\bar{b}_{\min} \leqslant \bar{b} \leqslant \bar{b}_{\max}$. To resolve the sharp features in f_ω for small angles ϕ_ω, the polynomial order should scale like $N_P \sim \phi_\omega^{-1}$.

where $\lfloor \cdot \rfloor$ denotes the floor function; the coefficients a_1 and a_2 are both of order 1 and independent of system details (temperature, etc.). Note that the polynomial order N_P scales linearly with the range $\Delta \bar{b} = \bar{b}_{\max} - \bar{b}_{\min}$ over which an approximation is required. Observe that the polynomial order N_P decays rapidly when ω moves away from zero, such that the *typical* value of N_P is of order 1.

3. Using KPM to evaluate matrix-vector products

We wish to apply the matrix

$$\tilde{M}_{\omega,\omega}^{-1} = (I - e^{-i\phi_\omega}\bar{B})^{-1} = f_\omega(A), \tag{C14}$$

to a vector, where A is a rescaling of $\bar{B}$ as defined in Eq. (C4). Using the truncated Chebyshev expansion, we may approximate

$$\tilde{M}_{\omega,\omega}^{-1} \approx \sum_{m=0}^{N_P-1} c_m T_m(A). \tag{C15}$$

The expansion order N_p and scalar coefficients c_m, given in Eq. (C10), implicitly depend on ϕ_ω, $\bar{b}_{\min}$, and $\bar{b}_{\max}$.

A key result from KPM is that the task of evaluating the matrix-vector product,

$$\tilde{M}_{\omega,\omega}^{-1} u \approx \sum_{m=0}^{N_P-1} c_m T_m(A)u = \sum_{m=0}^{N_P-1} c_m \alpha_m, \tag{C16}$$

does *not* require explicit construction of the dense matrix $\tilde{M}_{\omega,\omega}^{-1}$. Instead, we will iteratively calculate the vectors

$$\alpha_m = T_m(A)u. \tag{C17}$$

The Chebyshev polynomials satisfy a two-term recurrence relation,

$$T_{m+1}(A) = 2AT_m(A) - T_{m-1}(A). \tag{C18}$$

Multiplying by u on the right yields an explicit scheme for computing α_m,

$$\alpha_{m+1} = 2A\alpha_m - \alpha_{m-1}, \tag{C19}$$

beginning with

$$\alpha_0 = u, \quad \alpha_1 = Au. \tag{C20}$$

As the vectors α_m become available, they are accumulated into the right-hand side of Eq. (C16), eventually giving the desired matrix-vector product.

4. A full recipe for the preconditioner

Here we summarize all steps needed to apply the preconditioner in Eq. (67) efficiently. Our task is to evaluate the matrix-vector product,

$$P^{-1}v = \mathcal{U}^\dagger \tilde{P}^{-1}\mathcal{U}v. \tag{C21}$$

The unitary matrix $\mathcal{U}$ is defined in Eq. (72) and can be efficiently applied with an FFT. The matrix $\tilde{P}$ is zero except for its diagonal blocks $\tilde{M}_{\omega,\omega}$, which are given by Eq. (C1). The main challenge is to apply the $N \times N$ matrix $\tilde{M}_{\omega,\omega}^{-1}$ to a vector. We must do so for each index ω.

The matrix $\tilde{M}_{\omega,\omega}$ is a function of $\bar{B} = e^{-\Delta_\tau \bar{V}} e^{-\Delta_\tau K}$. If we can find numbers $\bar{b}_{\min}$ and $\bar{b}_{\max}$ that *assuredly* bound all eigenvalues of $\bar{B}$, then we may approximate $\tilde{M}_{\omega,\omega}^{-1}$ as in Eq. (C15).

To estimate $\bar{b}_{\max}$, we may use the Arnoldi iteration, repeatedly applying the matrix $\bar{B}$ to an initial random vector. This method produces an upper Hessenberg matrix, which serves as a low-rank approximation to $\bar{B}$. After about 20 iterations, the largest eigenvalue of this Hessenberg matrix (increased by 5%, to be safe) provides a suitable estimate of $\bar{b}_{\max}$. For numerical stability reasons, we estimate $\bar{b}_{\min}$ by applying the Arnoldi iteration to $\bar{B}^{-1} = e^{\Delta_\tau K} e^{\Delta_\tau \bar{V}}$, estimating its maximum eigenvalue and then taking the inverse. This is possible because, just like for $\bar{B}$, we are able to apply $\bar{B}^{-1}$ to a vector efficiently.

Given the approximation in Eq. (C15), we can efficiently calculate $\tilde{M}_{\omega,\omega}^{-1} u$ using Eq. (C16), where the vectors $\alpha_m = T_m(A)u$ are iteratively calculated using the Chebyshev recurrence in Eq. (C19).

The appropriate polynomial order N_P depends on the index ω. A reasonable choice is given in Eq. (C13).

5. Scaling of costs

The calculation of the matrix-vector product in Eq. (C16) requires $N_P - 1$ matrix-vector multiplications involving $\bar{B}$, where N_P depends on ω via Eq. (C13). Since the indices ω and $L - \omega - 1$ are effectively equivalent, we restrict attention to $0 \leqslant \omega < L_\tau/2$. We can sum over all such ω values to count

the total number of required matrix-vector multiplications

$$N_{\text{mat-vec}} = 2 \sum_{\omega=0}^{L_\tau/2-1} [N_P(\omega) - 1]$$

$$= 2 \sum_{\omega=0}^{L_\tau/2-1} \lfloor \Delta \bar{b} (a_1 \phi_\omega^{-1} + a_2) \rfloor - L_\tau. \quad \text{(C22)}$$

The factor of 2 accounts for the skipped indices, $L_\tau/2 \leqslant \omega < L_\tau$. Removing the floor function is justified when ω is order 1, such that ϕ_ω^{-1} is order L_τ [cf. Eq. (C2)], and in general produces an *upper* bound,

$$N_{\text{mat-vec}} \leqslant 2\Delta \bar{b} \left(a_1 \sum_{\omega=0}^{L_\tau/2-1} \phi_\omega^{-1} + a_2 L_\tau/2 \right) - L_\tau. \quad \text{(C23)}$$

We can explicitly evaluate the sum,

$$\sum_{\omega=0}^{L_\tau/2-1} (\omega + 1/2)^{-1} = \ln 4 + \gamma + \psi(L_\tau/2 + 1/2), \quad \text{(C24)}$$

where $\gamma = 0.577\ldots$ is the Euler-Mascheroni constant and $\psi(x) = \ln x + \mathcal{O}(x^{-1})$ is the digamma function. To a good approximation, the upper bound is

$$N_{\text{mat-vec}} \lesssim L_\tau \Delta \bar{b} \left[\frac{a_1}{\pi} (\gamma + \ln 2L_\tau) + a_2 \right] - L_\tau. \quad \text{(C25)}$$

Typically $a_1 = a_2 = 1$ and $\Delta \bar{b} \approx 1$. For, say, $L_\tau = 200$ (corresponding to inverse temperature $\beta = 20$ at $\Delta_\tau = 0.1$), the bound of Eq. (C25) gives

$$N_{\text{mat-vec}}/L_\tau \lesssim 2.1, \quad \text{(C26)}$$

whereas direct numerical evaluation of the sum yields $N_{\text{mat-vec}}/L_\tau = 1.6$. We infer that the bound of Eq. (C25) is in general a fairly tight one.

Note that L_τ applications of the matrix $\bar{B} = \exp(-\Delta_\tau \bar{V}) \exp(-\Delta_\tau K)$ are equivalent to the work required to apply the matrix M in Eq. (10). It follows that the task of applying the preconditioner in the Fourier basis, $\tilde{P}^{-1}$, is about two times more expensive than applying M. To apply $P^{-1} = \mathcal{U}^\dagger \tilde{P}^{-1} \mathcal{U}$, we additionally require two FFTs. For the benchmarks performed in this paper, we measured numerically that the total cost to apply P^{-1} is about three times greater than the cost to apply M.

APPENDIX D: SIMULATION TIME VERSUS SYSTEM SIZE AND INVERSE TEMPERATURE

In this Appendix we report the wall clock time for a full simulation, as a function of both system size N and inverse temperature β. Each simulation was performed using only a single core of an Intel i7-4770 and i7-2600 processor (no parallelism).

All simulations used to generate results in this Appendix were for Holstein systems with a dimensionless electron-phonon coupling of $\lambda = 0.25$. Each simulation performed $N_{\text{therm}} = 1000$ HMC updates to equilibrate the system, followed by an additional $N_{\text{sim}} = 2000$ HMC updates. Each HMC update consisted of $N_t = 100$ time steps, and each time step requires two CG solves. Each HMC update was
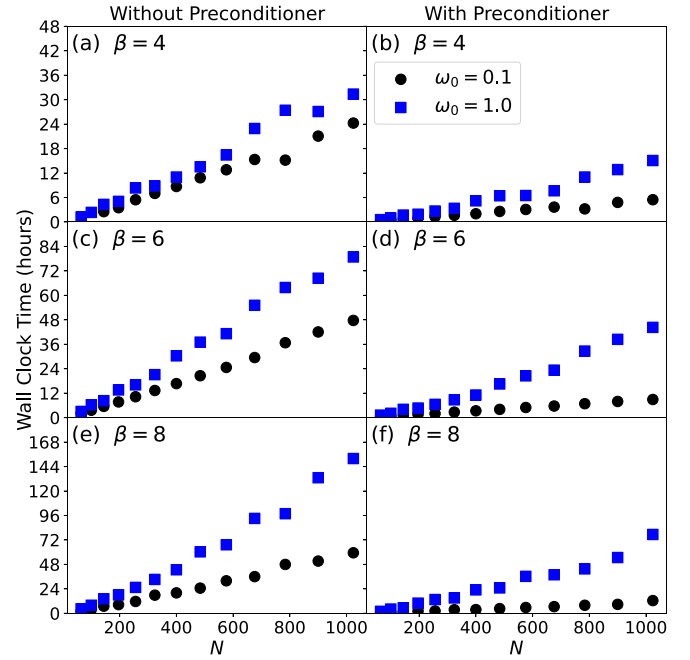

FIG. 10. Wall clock time for a full simulation (including over 650k CG solves) as a function of system size N. The corresponding average iteration counts per CG solve are shown in Fig. 5.

followed by 4 reflection and 4 swap updates, requiring $8 + 8$ CG solves. Additionally, a total of $N_{\text{sim}} = 2000$ measurements were taken, each requiring $N_{\text{rv}} = 10$ CG solves. In total, the simulation involved approximately 668k CG solves, which comprise the dominant computational cost. This simulation run-time was sufficient to achieve very accurate statistics, as demonstrated by the S_{CDW} measurements shown in Fig. 6(a).

Figure 10 displays the total simulation wall clock time as a function of N, and corresponds to Fig. 5, which shows the average iteration count per CG solve. In all panels we see that the wall clock timescales in an approximately linear fashion
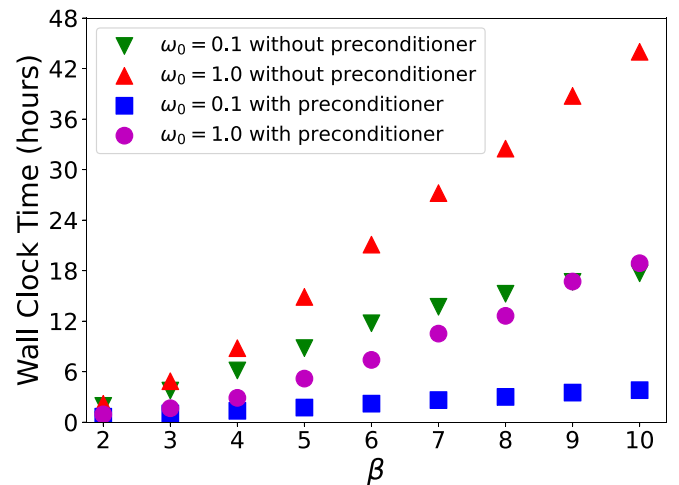

FIG. 11. Wall clock time for a full simulation as a function of inverse temperature β with system size $N = 256$. The corresponding average iteration counts are shown in Fig. 6.

with N. Empirical fitting of the wall clock time to a power law curve in N yields an exponent between 1.0 and 1.3 in all cases. Additionally, we see that the preconditioner uniformly decreases the simulation time, although the relative speedup is more significant at $\omega_0 = 0.1$ than $\omega_0 = 1.0$.

In similar fashion, Fig. 11 shows the wall clock time versus β, and should be compared with Fig. 6(b), which reports the average iteration count per CG solve. Once again we see that the preconditioner strictly reduces total simulation times, and that the wall clock timescales near linearly with β.

[1] R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, Monte Carlo calculations of coupled Boson-Fermion systems, Phys. Rev. D **24**, 2278 (1981).

[2] E. W. Huang, C. B. Mendl, S. Liu, S. Johnston, H.-C. Jiang, B. Moritz, and T. P. Devereaux, Numerical evidence of fluctuating stripes in the normal state of high-Tc cuprate superconductors, Science **358**, 1161 (2017).

[3] E. W. Huang, C. B. Mendl, H.-C. Jiang, B. Moritz, and T. P. Devereaux, Stripe order from the perspective of the Hubbard model, npj Quantum Mater. **3**, 22 (2018).

[4] S. Li, A. Nocera, U. Kumar, and S. Johnston, Particle-hole asymmetry in the dynamical spin and charge responses of corner-shared 1D cuprates, Commun. Phys. **4**, 217 (2021).

[5] S. Sorella, The phase diagram of the Hubbard model by variational auxiliary field quantum Monte Carlo, arXiv:2101.07045.

[6] D. P. Arovas, E. Berg, S. A. Kivelson, and S. Raghu, The Hubbard model, Annu. Rev. Condens. Matter Phys. **13**, 239 (2022).

[7] M. Qin, T. Schäfer, S. Andergassen, P. Corboz, and E. Gull, The Hubbard model: A computational perspective, Annu. Rev. Condens. Matter Phys. **13**, 275 (2022).

[8] D. J. Scalapino, A common thread: The pairing interaction for unconventional superconductors, Rev. Mod. Phys. **84**, 1383 (2012).

[9] E. Y. Loh, J. E. Gubernatis, R. T. Scalettar, S. R. White, D. J. Scalapino, and R. L. Sugar, Numerical stability and the sign problem in the determinant quantum Monte Carlo method, Int. J. Mod. Phys. C **16**, 1319 (2005).

[10] S. R. White, D. J. Scalapino, R. L. Sugar, E. Y. Loh, J. E. Gubernatis, and R. T. Scalettar, Numerical study of the two-dimensional Hubbard model, Phys. Rev. B **40**, 506 (1989).

[11] E. Y. Loh, Jr., J. E. Gubernatis, R. T. Scalettar, S. R. White, D. J. Scalapino, and R. L. Sugar, Sign problem in the numerical simulation of many-electron systems, Phys. Rev. B **41**, 9301 (1990).

[12] E. Berg, M. A. Metlitski, and S. Sachdev, Sign-problem–free quantum Monte Carlo of the onset of antiferromagnetism in metals, Science **338**, 1606 (2012).

[13] S. Chandrasekharan, Fermion bag approach to lattice field theories, Phys. Rev. D **82**, 025007 (2010).

[14] Z.-X. Li, Y.-F. Jiang, and H. Yao, Majorana-Time-Reversal Symmetries: A Fundamental Principle for Sign-Problem-Free Quantum Monte Carlo Simulations, Phys. Rev. Lett. **117**, 267002 (2016).

[15] C. Wu and S.-C. Zhang, Sufficient condition for absence of the sign problem in the fermionic quantum Monte Carlo algorithm, Phys. Rev. B **71**, 155115 (2005).

[16] R. Levy and B. K. Clark, Mitigating the Sign Problem through Basis Rotations, Phys. Rev. Lett. **126**, 216401 (2021).

[17] S. Tarat, B. Xiao, R. Mondaini, and R. T. Scalettar, Deconvolving the components of the sign problem, Phys. Rev. B **105**, 045107 (2022).

[18] M. Jiang, G. A. Sawatzky, M. Berciu, and S. Johnston, Polaron and bipolaron tendencies in a semiclassical model for hole-doped bismuthates, Phys. Rev. B **103**, 115129 (2021).

[19] Z. Li, G. Antonius, M. Wu, H. Felipe, and S. G. Louie, Electron-Phonon Coupling from *Ab Initio* Linear-Response Theory within the *GW* Method: Correlation-Enhanced Interactions and Superconductivity in $Ba_{1-x}K_xBiO_3$, Phys. Rev. Lett. **122**, 186402 (2019).

[20] S. Li and S. Johnston, Quantum Monte Carlo study of lattice polarons in the two-dimensional three-orbital Su-Schrieffer-Heeger model, npj Quantum Mater. **5**, 40 (2020).

[21] A. W. Sleight, Bismuthates: $BaBiO_3$ and related superconducting phases, Phys. C **514**, 152 (2015).

[22] X. Xi, L. Zhao, Z. Wang, H. Berger, L. Forró, J. Shan, and K. F. Mak, Strongly enhanced charge-density-wave order in monolayer nbse 2, Nat. Nanotechnol. **10**, 765 (2015).

[23] P. Chen, Y.-H. Chan, X.-Y. Fang, Y. Zhang, M.-Y. Chou, S.-K. Mo, Z. Hussain, A.-V. Fedorov, and T.-C. Chiang, Charge density wave transition in single-layer titanium diselenide, Nat. Commun. **6**, 8943 (2015).

[24] P. Chen, Y.-H. Chan, M.-H. Wong, X.-Y. Fang, M.-Y. Chou, S.-K. Mo, Z. Hussain, A.-V. Fedorov, and T.-C. Chiang, Dimensional effects on the charge density waves in ultrathin films of $TiSe_2$, Nano Lett. **16**, 6331 (2016).

[25] C. H. P. Wen, H. C. Xu, Q. Yao, R. Peng, X. H. Niu, Q. Y. Chen, Z. T. Liu, D. W. Shen, Q. Song, X. Lou, Y. F. Fang, X. S. Liu, Y. H. Song, Y. J. Jiao, T. F. Duan, H. H. Wen, P. Dudin, G. Kotliar, Z. P. Yin, and D. L. Feng, Unveiling the Superconducting Mechanism of $Ba_{0.51}K_{0.49}BiO_3$, Phys. Rev. Lett. **121**, 117002 (2018).

[26] K. Foyevtsova, A. Khazraie, I. Elfimov, and G. A. Sawatzky, Hybridization effects and bond disproportionation in the bismuth perovskites, Phys. Rev. B **91**, 121114(R) (2015).

[27] O. Bradley, G. G. Batrouni, and R. T. Scalettar, Superconductivity and charge density wave order in the two-dimensional Holstein model, Phys. Rev. B **103**, 235104 (2021).

[28] B. Cohen-Stead, N. C. Costa, E. Khatami, and R. T. Scalettar, Effect of strain on charge density wave order in the Holstein model, Phys. Rev. B **100**, 045125 (2019).

[29] C. Feng, H. Guo, and R. T. Scalettar, Charge density waves on a half-filled decorated honeycomb lattice, Phys. Rev. B **101**, 205103 (2020).

[30] Z.-X. Li, M. L. Cohen, and D.-H. Lee, Enhancement of superconductivity by frustrating the charge order, Phys. Rev. B **100**, 245105 (2019).

[31] P. Niyaz, J. E. Gubernatis, R. T. Scalettar, and C. Y. Fong, Charge-density-wave-gap formation in the two-dimensional Holstein model at half-filling, Phys. Rev. B **48**, 16011 (1993).

[32] R. M. Noack, D. J. Scalapino, and R. T. Scalettar, Charge-Density-Wave and Pairing Susceptibilities in a Two-Dimensional Electron-Phonon Model, Phys. Rev. Lett. **66**, 778 (1991).

[33] B. Nosarzewski, E. W. Huang, P. M. Dee, I. Esterlis, B. Moritz, S. A. Kivelson, S Johnston, and T. P. Devereaux, Superconductivity, charge density waves, and bipolarons in the Holstein model, Phys. Rev. B **103**, 235156 (2021).

[34] M. Vekić, R. M. Noack, and S. R. White, Charge-density waves versus superconductivity in the Holstein model with next-nearest-neighbor hopping, Phys. Rev. B **46**, 271 (1992).

[35] Y.-X. Zhang, W.-T. Chiu, N. C. Costa, G. G. Batrouni, and R. T. Scalettar, Charge Order in the Holstein Model on a Honeycomb Lattice, Phys. Rev. Lett. **122**, 077602 (2019).

[36] X. Cai, Z.-X. Li, and H. Yao, Antiferromagnetism Induced by Bond Su-Schrieffer-Heeger Electron-Phonon Coupling: A Quantum Monte Carlo Study, Phys. Rev. Lett. **127**, 247203 (2021).

[37] C. Feng, B. Xing, D. Poletti, R. Scalettar, and G. Batrouni, Phase diagram of the Su-Schrieffer-Heeger-Hubbard model on a square lattice, arXiv:2109.09206.

[38] B. Xing, W.-T. Chiu, D. Poletti, R. T. Scalettar, and G. Batrouni, Quantum Monte Carlo Simulations of the 2D Su-Schrieffer-Heeger Model, Phys. Rev. Lett. **126**, 017601 (2021).

[39] S. Li, P. M. Dee, E. Khatami, and S. Johnston, Accelerating lattice quantum Monte Carlo simulations using artificial neural networks: Application to the Holstein model, Phys. Rev. B **100**, 020302(R) (2019).

[40] X. Y. Xu, Y. Qi, J. Liu, L. Fu, and Z. Y. Meng, Self-learning quantum Monte Carlo method in interacting fermion systems, Phys. Rev. B **96**, 041119(R) (2017).

[41] G. G. Batrouni and R. T. Scalettar, Langevin simulations of a long-range electron-phonon model, Phys. Rev. B **99**, 035114 (2019).

[42] B. Cohen-Stead, K. Barros, Z. Y. Meng, C. Chen, R. T. Scalettar, and G. G. Batrouni, Langevin simulations of the half-filled cubic Holstein model, Phys. Rev. B **102**, 161108(R) (2020).

[43] A. Götz, S. Beyl, M. Hohenadler, and F. F. Assaad, Langevin dynamics simulations of the two-dimensional Su-Schrieffer-Heeger model, Phys. Rev. B **105**, 085151 (2022).

[44] S. Karakuzu, K. Seki, and S. Sorella, Solution of the sign problem for the half-filled Hubbard-Holstein model, Phys. Rev. B **98**, 201108(R) (2018).

[45] S. Beyl, F. Goth, and F. F. Assaad, Revisiting the hybrid quantum Monte Carlo method for Hubbard and electron-phonon models, Phys. Rev. B **97**, 085144 (2018).

[46] J. Besag, Comments on "Representations of knowledge in complex systems" by U. Grenander and M. I. Miller, J. R. Stat. Soc. Ser. B **56**, 591 (1994).

[47] P. J. Rossky, J. D. Doll, and H. L. Friedman, Brownian dynamics as smart Monte Carlo simulation, J. Chem. Phys. **69**, 4628 (1978).

[48] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, Hybrid Monte Carlo, Phys. Lett. B **195**, 216 (1987).

[49] R. Neal, MCMC Using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo*, edited by S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (Chapman and Hall/CRC, Boca Raton, FL, 2011), Vol. 20116022.

[50] A. D. Kennedy and B. Pendleton, Cost of the generalised hybrid Monte Carlo algorithm for free field theory, Nucl. Phys. B **607**, 456 (2001).

[51] R. T. Scalettar, D. J. Scalapino, R. L. Sugar, and D Toussaint, Hybrid molecular-dynamics algorithm for the numerical simulation of many-electron systems, Phys. Rev. B **36**, 8632 (1987).

[52] Z. Bai, W. Chen, R. Scalettar, and I. Yamazaki, Numerical methods for quantum Monte Carlo simulations of the Hubbard model, in *Multi-Scale Phenomena In Complex Fluids: Modeling, Analysis and Numerical Simulation* (World Scientific, Singapore, 2009), pp. 1–110.

[53] Th. Holstein, Studies of polaron motion: Part I. The molecular-crystal model, Ann. Phys. **8**, 325 (1959).

[54] N. C. Costa, T. Blommel, W.-T. Chiu, G. Batrouni, and R. T. Scalettar, Phonon Dispersion and the Competition Between Pairing and Charge Order, Phys. Rev. Lett. **120**, 187003 (2018).

[55] G. Paleari, F. Hébert, B. Cohen-Stead, K. Barros, R. T. Scalettar, and G. G. Batrouni, Quantum Monte Carlo study of an anharmonic Holstein model, Phys. Rev. B **103**, 195117 (2021).

[56] J. Gubernatis, N. Kawashima, and P. Werner, *Quantum Monte Carlo Methods* (Cambridge University Press, Cambridge, UK, 2016).

[57] W. P. Su, J. R. Schrieffer, and A. J. Heeger, Solitons in Polyacetylene, Phys. Rev. Lett. **42**, 1698 (1979).

[58] Y. Saad, *Iterative Methods for Sparse Linear Systems* (SIAM, 2003).

[59] C.-R. Lee, Minimal split checkerboard method for exponentiating sparse matrices and its applications in quantum statistical mechanics, SIAM J. Sci. Comput. **35**, C143 (2013).

[60] S. Beyl, Hybrid Quantum Monte Carlo for Condensed Matter Models, Ph.D. thesis, Universität Würzburg, 2020.

[61] R. T. Scalettar, N. E. Bickers, and D. J. Scalapino, Quantum Monte Carlo studies of the Holstein model, in *Computer Simulation Studies in Condensed Matter Physics* (Springer, Berlin, 1988), pp. 166–171.

[62] M. A. Clark and A. D. Kennedy, Speeding up HMC with better integrators, arXiv:0710.3611.

[63] A. D. Kennedy, P. J. Silva, and M. A. Clark, Shadow Hamiltonians, poisson brackets, and gauge theories, Phys. Rev. D **87**, 034511 (2013).

[64] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Number 14 (Cambridge University Press, Cambridge, UK, 2004).

[65] J. C. Sexton and D. H. Weingarten, Hamiltonian evolution for the hybrid Monte Carlo algorithm, Nucl. Phys. B **380**, 665 (1992).

[66] T. Takaishi and P. de Forcrand, Testing and tuning symplectic integrators for the hybrid Monte Carlo algorithm in lattice QCD, Phys. Rev. E **73**, 036706 (2006).

[67] P. B. Mackenze, An improved hybrid Monte Carlo method, Phys. Lett. B **226**, 369 (1989).

[68] M. Creutz, Global Monte Carlo algorithms for many-fermion systems, Phys. Rev. D **38**, 1228 (1988).

[69] A. Griewank, On automatic differentiation, in *Mathematical Programming: Recent Developments and Applications* (Kluwer Academic Publishers, Amsterdam, 1989), pp. 83–108.

[70] G. G. Batrouni, G. R. Katz, A. S. Kronfeld, G. P. Lepage, B. Svetitsky, and K. G. Wilson, Langevin simulations of lattice field theories, Phys. Rev. D **32**, 2736 (1985).

[71] I. Esterlis, S. A. Kivelson, and D. J. Scalapino, Pseudogap crossover in the electron-phonon system, Phys. Rev. B **99**, 174516 (2019).

[72] S. Johnston, E. A. Nowadnick, Y. F. Kung, B. Moritz, R. T. Scalettar, and T. P. Devereaux, Determinant quantum Monte Carlo study of the two-dimensional single-band Hubbard-Holstein model, Phys. Rev. B **87**, 235133 (2013).

[73] R. T. Scalettar, R. M. Noack, and R. R. P. Singh, Ergodicity at large couplings with the determinant Monte Carlo algorithm, Phys. Rev. B **44**, 10502 (1991).

[74] A. Weiße, G. Wellein, A. Alvermann, and H. Fehske, The kernel polynomial method, Rev. Mod. Phys. **78**, 275 (2006).

[75] F. F. Assaad and H. G. Evertz, World-line and determinantal quantum Monte Carlo methods for spins, phonons, and electrons, in *Computational Many-Particle Physics* (Springer, Berlin, 2008), pp. 277–356.

[76] M. F. Hutchinson, A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines, Commun. Stat.-Simul. Comput. **18**, 1059 (1989).

[77] J. M. Tang and Y. Saad, A probing method for computing the diagonal of a matrix inverse, Numer. Linear Alg. Appl. **19**, 485 (2012).

[78] Z. Wang, G.-W. Chern, C. D. Batista, and K. Barros, Gradient-based stochastic estimation of the density matrix, J. Chem. Phys. **148**, 094107 (2018).

[79] H. F. Trotter, On the product of semi-groups of operators, Proc. Am. Math. Soc. **10**, 545 (1959).

[80] M. P. Drazin and E. V. Haynsworth, Criteria for the reality of matrix eigenvalues, Math. Z. **78**, 449 (1962).

[81] https://github.com/cohensbw/ElPhDynamics.git.