

Removing Nodal and Support-Mismatch Pathologies in Variational Monte Carlo via Blurred Sampling

Zhou-Quan Wan^{1,*,†}, Roeland Wiersema^{1,*,‡} and Shiwei Zhang^{1,§}

Center for Computational Quantum Physics, Flatiron Institute, New York, New York 10010, USA



(Received 8 April 2026; revised 29 June 2026; accepted 10 August 2026; published 2 September 2026)

Variational Monte Carlo (VMC) is a powerful and fast-growing method for optimizing and evolving parametrized many-body wave functions, especially with modern neural-network quantum states. In practice, however, the stochastic estimators that form the backbone of the method can become unstable or biased due to the presence of nodes, a ubiquitous feature of quantum wave functions. In the continuum, this results in heavy-tailed estimators with potentially divergent variances, while in discrete Hilbert spaces the sampling distribution can miss parts of the support needed to form unbiased estimators. These statistical pathologies lead to unreliable optimization trajectories in stochastic reconfiguration or incorrect variational dynamics in time-dependent variational Monte Carlo (t-VMC) and severely limit the power of the numerical simulations. We introduce blurred sampling to address these difficulties. The method has a number of rigorous properties that make it well behaved, effective, and computationally efficient. Additionally, it is a postprocessing approach that can be used without modifying the underlying sampler and incurs only minimal overhead. We demonstrate its effectiveness in a broad range of problems—including ground-state optimization in both continuum and discrete settings as well as in time-dependent variational Monte Carlo, where it resolves previously identified failure modes—and further apply it to large-scale problems in spin dynamics. This work establishes a broadly applicable framework for robust VMC and t-VMC calculations.

DOI: [10.1103/jrm5-gv19](https://doi.org/10.1103/jrm5-gv19)

Subject Areas: Computational Physics,
Condensed Matter Physics,
Quantum Physics

I. INTRODUCTION

Variational Monte Carlo (VMC) provides a powerful framework for studying interacting quantum many-body systems by combining variational parametrizations of the wave function with stochastic sampling, thereby circumventing the exponential growth of Hilbert space [1,2]. Recent advances in highly expressive variational *Ansätze*—most notably, neural-network quantum states—together with rapid progress in machine-learning techniques and computational infrastructure, have substantially expanded the scope of VMC [3]. Within a unified Monte Carlo framework, VMC now enables accurate calculations of ground and excited states, real- and imaginary-time

dynamics, and finite-temperature properties [3–17]. As a result, VMC has found broad applications across condensed matter physics [18–29], quantum chemistry [30–36], cold-atom or electron gas systems [37–47], and quantum simulation [48–50].

Despite this versatility, the reliability of VMC ultimately hinges on the statistical properties of its Monte Carlo estimators. Central quantities entering stochastic reconfiguration and time-dependent variational dynamics—such as energy gradients and variational forces—are expressed as ratios of wave-function amplitudes [3,9,13,51]. Whenever the wave function develops nodes, these ratios become singular. Such nodal structures are ubiquitous in fermionic systems, frustrated magnets, and systems with sign structure enforced by symmetry or gauge fields. Near nodes, statistical fluctuations can become heavy-tailed and may even acquire infinite variance in continuum settings [52–55]. In discrete configuration spaces, a related but more severe pathology can arise: If the support of the wave function does not coincide with that of the Hamiltonian action, the resulting estimators remain biased even in the infinite-sample limit [56,57]. These statistical pathologies directly destabilize gradient-based optimization and variational time evolution, limiting the robustness and reliability of VMC.

*These authors contributed equally to this work.

†Contact author: zwan@flatironinstitute.org

‡Contact author: rwiersema@flatironinstitute.org

§Contact author: szhang@flatironinstitute.org

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

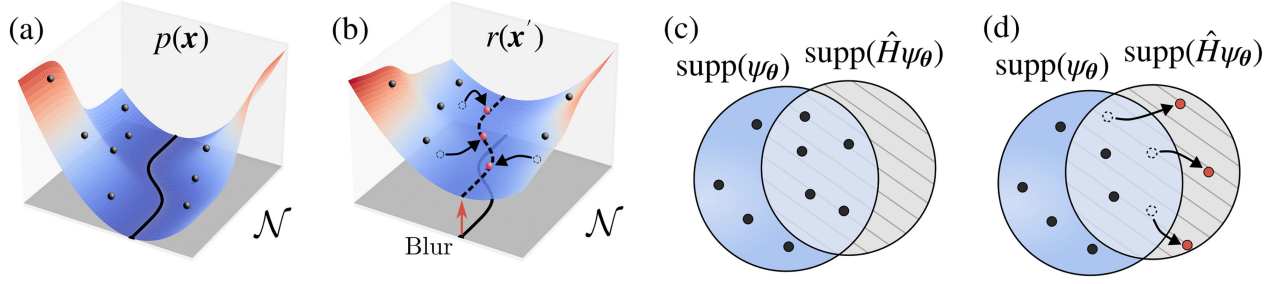


FIG. 1. Statistical pathologies and their resolution via blurred sampling. (a) Nodal hypersurfaces lead to divergences in ratio-type estimators due to vanishing probability density $p(\mathbf{x})$. (b) Blurred sampling locally perturbs the configurations, assigning finite probability to the original nodal set, regularizing the divergence. (c) The supports of ψ_θ and $\hat{H}\psi_\theta$ may not coincide (shaded region), resulting in a support-mismatch bias; see Eq. (4). (d) Blurred sampling exploits the connectivity of $\hat{H}$, allowing configurations in $\text{supp}(\psi_\theta)$ to access the mismatched region of $\text{supp}(\hat{H}\psi_\theta)$, eliminating the bias.

Considerable effort has, therefore, been devoted to mitigating these instabilities. Some approaches modify the estimator directly through explicit regularization [58] or variance-reduction techniques such as control variates [56,59,60]. More commonly, the problem is addressed through importance-sampling strategies that replace the original sampling distribution by a modified reference measure [61–66]. While these approaches can alleviate specific pathologies, they introduce competing requirements which are difficult to satisfy simultaneously. In practice, residual bias or variance amplification persists, and significant modification to the algorithm and additional computational overhead are often incurred. A broadly applicable and structurally robust solution to nodal instabilities in VMC remains lacking, which hinders progress in many application areas.

In this work, we introduce *blurred sampling* to address this challenge. Figure 1 provides a schematic illustration of the new approach. Rather than adding explicit regularization to the estimator or globally redefining the sampling distribution, blurred sampling applies a single local mixing step to Monte Carlo configurations as a postprocessing operation. This procedure induces an implicit reference measure that regularizes nodal singularities at their origin while leaving the underlying sampling dynamics unchanged. We show that the resulting reweighting factors are strictly bounded, which ensures a finite effective sample size and resolves both infinite-variance behavior and support-mismatch bias. Because the method is implemented as a postprocessing step, it preserves the computational scaling of standard VMC and, in discrete configuration spaces, can often be realized without additional wave-function evaluations. This design allows blurred sampling to integrate seamlessly with existing variational *Ansätze*, optimization schemes, real- and imaginary-time dynamics, and modern sampling architectures [67,68], including autoregressive models [69–75].

As we demonstrate below through theoretical analysis and numerical examples, blurred sampling restores stable and accurate simulations in regimes where conventional

VMC breaks down. It provides a broadly applicable and scalable framework for achieving reliable simulations in variational Monte Carlo and neural quantum states, which will enable a wide range of applications across disciplines in quantum science.

II. BACKGROUND AND PRELIMINARIES

A. Variational Monte Carlo

In the general VMC framework, the many-body state is approximated by a parametrized wave function $|\psi_\theta\rangle = \sum_{\mathbf{x}} \psi_\theta(\mathbf{x})|\mathbf{x}\rangle$ in the computational basis $\{|\mathbf{x}\rangle\}$, where the variational parameters $\theta \in \mathbb{C}^{N_{\text{para}}}$ are taken to be complex valued, with real parametrizations recovered as a special case, and the sum over $\mathbf{x}$ can be either discrete or continuous (integral). The optimization of the variational energy $E_\theta \equiv (\langle\psi_\theta|\hat{H}|\psi_\theta\rangle / \langle\psi_\theta|\psi_\theta\rangle)$ and the real- or imaginary-time dynamics are governed by two fundamental objects: the variational force $\mathbf{F} \equiv \partial_\theta E_\theta$ and the quantum geometric tensor (QGT) $\mathbf{S}$. They are defined as

$$F_i = \frac{\langle\partial_{\theta_i}\psi_\theta|\hat{H}|\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle} - \frac{\langle\partial_{\theta_i}\psi_\theta|\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle} \frac{\langle\psi_\theta|\hat{H}|\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle},$$

$$S_{ij} = \frac{\langle\partial_{\theta_i}\psi_\theta|\partial_{\theta_j}\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle} - \frac{\langle\partial_{\theta_i}\psi_\theta|\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle} \frac{\langle\psi_\theta|\partial_{\theta_j}\psi_\theta\rangle}{\langle\psi_\theta|\psi_\theta\rangle}. \quad (1)$$

These quantities determine the dynamics of parameters in stochastic reconfiguration (SR) and time-dependent VMC (t-VMC) through

$$S_{ij}\dot{\theta}_j = -\xi F_i, \quad (2)$$

which is known as the time-dependent variational principle (TDVP) equation [3,9,12,13,76]. The choice $\xi = 1$ corresponds to the imaginary-time evolution (i.e., SR), and $\xi = i$ yields real-time dynamics.

In VMC, these quantities are evaluated by Monte Carlo sampling from the Born distribution $p(\mathbf{x}) = |\psi_\theta(\mathbf{x})|^2 / \sum_{\mathbf{x}} |\psi_\theta(\mathbf{x})|^2$. The estimators take the covariance form

$$F_i^{\text{MC}} = \langle \mathcal{O}_i^*(\mathbf{x}) E_{\text{loc}}(\mathbf{x}) \rangle_p - \langle \mathcal{O}_i^*(\mathbf{x}) \rangle_p \langle E_{\text{loc}}(\mathbf{x}) \rangle_p, \\ S_{ij}^{\text{MC}} = \langle \mathcal{O}_i^*(\mathbf{x}) \mathcal{O}_j(\mathbf{x}) \rangle_p - \langle \mathcal{O}_i^*(\mathbf{x}) \rangle_p \langle \mathcal{O}_j(\mathbf{x}) \rangle_p, \quad (3)$$

where $\langle \cdot \rangle_p$ denotes expectation with respect to $p(\mathbf{x})$, $\mathcal{O}_i(\mathbf{x}) \equiv \partial_{\theta_i} \psi_{\theta}(\mathbf{x}) / \psi_{\theta}(\mathbf{x})$ is the logarithmic derivative, and $E_{\text{loc}}(\mathbf{x}) \equiv \langle \mathbf{x} | \hat{H} | \psi_{\theta} \rangle / \psi_{\theta}(\mathbf{x})$ is the local energy.

This representation makes the origin of statistical pathologies transparent: Both E_{loc} and $\mathcal{O}_i$ contain divisions by $\psi_{\theta}(\mathbf{x})$, so the corresponding ratio-type estimators can diverge in the vicinity of wave-function nodes, i.e., when $|\psi_{\theta}(\mathbf{x})| \rightarrow 0$. As a consequence, the stochastic estimators entering VMC can often become ill behaved, as we discuss further in the next section.

B. Statistical pathologies

These pathologies take different forms depending on the structure of the configuration space. We distinguish between continuum and discrete settings in turn.

1. Infinite variance in continuum systems

In continuum fermionic systems, the wave function $\psi_{\theta}(\mathbf{x})$ depends continuously on the particle coordinates. Fermionic statistics enforces the presence of nodal sets $\mathcal{N} \equiv \{\mathbf{x} | \psi_{\theta}(\mathbf{x}) = 0\}$ [see Fig. 1(a)], whose geometry has been extensively studied [77–80]. In the vicinity of a node, the wave function vanishes linearly in the normal direction, $|\psi_{\theta}| \propto d$, where d denotes the distance to the nodal set. Consequently, ratio-type quantities in Eq. (3) diverge as $|E_{\text{loc}}|, |\mathcal{O}_i| \propto 1/d$. Since the distribution scales as $p(\mathbf{x}) \propto d^2$ near the node, the resulting estimator can have finite expectation but infinite variance.

To illustrate the resulting pathology, we consider a minimal continuum fermionic system (two spinless fermions on a ring). As shown in Fig. 2(a), the distribution of the energy-gradient estimator exhibits a pronounced heavy tail [81–84]. More precisely, the tail follows $P(|X| > x) \sim Cx^{-\alpha}$ as $x \rightarrow \infty$, with exponent $\alpha \simeq 1.5$. Figure 2(b) shows that the variance fails to converge with increasing Monte Carlo runs. This behavior is generic. For real (complex) wave functions, the tail exponent of the estimator in Eq. (3) satisfies $\alpha = 3/2$ ($\alpha = 2$), reflecting the codimension of the nodal manifold [52]. Since $\alpha \leq 2$, the variance diverges, leading to anomalously slow and unstable Monte Carlo convergence—the so-called infinite-variance problem [52–55], i.e., $\epsilon \propto 1/N^{1/3}$ for $\alpha = 3/2$ and $\epsilon \propto \sqrt{\log(N)/N}$ for $\alpha = 2$ rather than $1/\sqrt{N}$ [83].

2. Bias in discrete configuration spaces

In discrete configuration spaces, the problem can become more severe. The support of $\psi_{\theta}(\mathbf{x})$ need not coincide with that of $\langle \mathbf{x} | \hat{H} | \psi_{\theta} \rangle$; see Fig. 1(c). Configurations that contribute to the exact force may, therefore, lie outside the sampled support $\text{supp}(\psi_{\theta}) = \{\mathbf{x} | \psi_{\theta}(\mathbf{x}) \neq 0\}$.

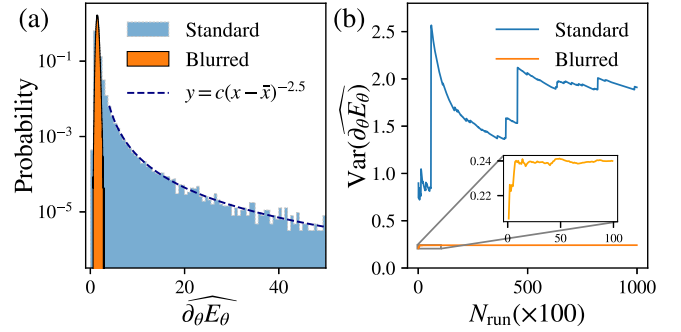


FIG. 2. Pedagogical example in the continuum. We consider two noninteracting spinless fermions on a ring, with the Hamiltonian $\hat{H} = -\frac{1}{2}(\partial_1^2 + \partial_2^2)$, and variational Ansatz $\psi_{\theta}(x_1, x_2) = \cos(\theta) \sin(x_1 - x_2) + \sin(\theta) \sin(2x_1 - 2x_2)$, $x_{1,2} \in [0, 2\pi]$. (a) Distribution of the gradient estimator based on 1000 samples at $\theta = \pi/4$. Standard sampling exhibits a heavy-tailed distribution with exponent $\alpha = 1.5$, while blurred sampling yields a finite-variance estimator. (b) Estimated variance versus number of runs; standard sampling does not converge.

As a result, the Monte Carlo estimator can remain biased even in the infinite-sample limit [56,57]. Explicitly,

$$F_i - F_i^{\text{MC}} = \frac{\sum_{\mathbf{x} \notin \text{supp}(\psi_{\theta})} \langle \partial_{\theta_i} \psi_{\theta} | \mathbf{x} \rangle \langle \mathbf{x} | \hat{H} | \psi_{\theta} \rangle}{\sum_{\mathbf{x}} |\psi_{\theta}(\mathbf{x})|^2}. \quad (4)$$

The right-hand side corresponds precisely to the contribution of configurations that are never sampled under p . We refer to this structural discrepancy as a *support-mismatch bias*. Even absent an absolute mismatch, when $\psi_{\theta}(\mathbf{x}) \ll \langle \mathbf{x} | \hat{H} | \psi_{\theta} \rangle$ on rarely sampled configurations, the ratio-type structure of the estimator can strongly amplify statistical fluctuations and render the simulation unstable. The effect of the support-mismatch bias is illustrated in Fig. 3, using a single-spin example with $\hat{H} = X$ and variational Ansatz $\psi_{\theta} = (\cos \theta, \sin \theta)$. As $\theta \rightarrow 0$, the gradient estimator exhibits increasingly large fluctuations and eventually develops a systematic bias. As a result, a simple ground-state optimization can lead to an incorrect answer. These issues are not specific to the force: The QGT, sharing the same ratio-type structure, is subject to an analogous bias and instabilities.

C. Requirements for a remedy

The infinite-variance problem and support-mismatch bias identified above present a number of challenges, compounded by the fact that VMC is typically applied in extremely high dimensions. A natural strategy for remedying nodes is to modify the sampling distribution via importance sampling [85]. By drawing samples $\mathbf{x}'$ from a reference distribution $r(\mathbf{x}')$, the variational force in Eq. (1) can be evaluated in a self-normalized importance-sampling (SNIS) form:

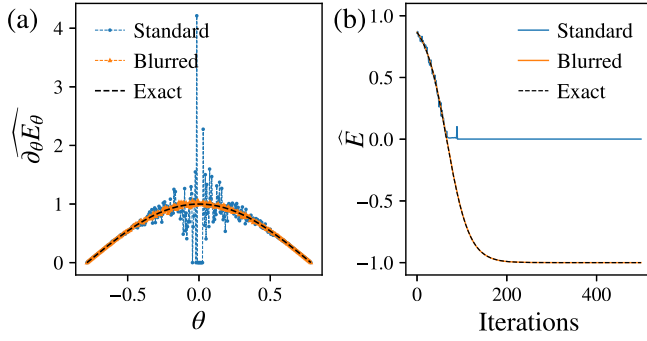


FIG. 3. Pedagogical example in discrete space. We consider a single-spin system with Hamiltonian $\hat{H} = X$ and $\psi_\theta = [\cos(\theta), \sin(\theta)]$. (a) Gradient estimator at different θ using 1000 samples. As $\theta \rightarrow 0$, standard sampling exhibits large fluctuations and develops a systematic bias, whereas blurred sampling remains stable. (b) Energy optimization via SR, initialized at $\theta = \pi/3$. Standard sampling can become trapped near $\theta \rightarrow 0$, while blurred sampling follows the exact imaginary-time trajectory.

$$F_i^{\text{SNIS}} = \frac{\langle \omega \mathcal{O}_i^* E_{\text{loc}} \rangle_r}{\langle \omega \rangle_r} - \frac{\langle \omega \mathcal{O}_i^* \rangle_r \langle \omega E_{\text{loc}} \rangle_r}{\langle \omega \rangle_r^2}, \quad (5)$$

where $\omega(\mathbf{x}) \equiv p(\mathbf{x})/r(\mathbf{x})$ is the *reweighting factor*. Note that divergences arising from rarely sampled configurations can now be compensated by the reweighting factor ω . In principle, infinite variance can be mitigated if r reduces the effective codimension of nodal set. Similarly, the support-mismatch bias in the discrete case can be resolved provided that the reference distribution satisfies $\text{supp}(\hat{H}\psi_\theta) \cup \text{supp}(\psi_\theta) \subseteq \text{supp}(r)$ for the forces and $\text{supp}[\partial_\theta \psi_\theta(\mathbf{x}) \partial_\theta \psi_\theta(\mathbf{x})] \cup \text{supp}(\psi_\theta) \subseteq \text{supp}(r)$ for the QGT. In both cases, the missing contributions are restored.

However, identifying a proper choice of $r(\mathbf{x})$ that is easy to sample, removes the nodes, and does not significantly modify the distribution elsewhere is very difficult. Because quantum problems typically involve extremely high-dimensional configuration spaces, $p(\mathbf{x})$ can vary by many tens of orders of magnitude across configurations [53]. This difficulty is further compounded by the fact that modifying the sampling distribution introduces a reweighting factor

$\omega = p/r$, whose fluctuations strongly affect the statistical efficiency of the estimator.

Existing approaches construct r in different ways. For example, overdispersed sampling schemes employ $r(\mathbf{x}) \propto |\psi(\mathbf{x})|^\alpha$ with $\alpha < 2$, which broadens the distribution and is straightforward to sample [63,64]. However, the resulting reweighting factor $\omega \propto |\psi|^{2-\alpha}$ exhibits increasingly large fluctuations as the system size grows, causing an exponential degradation of the effective sample size ($\text{ESS} \equiv \langle \omega \rangle_r^2 / \langle \omega^2 \rangle_r$) and preventing reliable scaling to larger systems (a quantitative example is given in Appendix B 3). Other methods introduce explicit regularization factors based on local-energy estimators or nodal-distance diagnostics [65,66], leading to modified distributions of the form $r(\mathbf{x}) = g(\mathbf{x})p(\mathbf{x})$, where $g(\mathbf{x})$ is constructed to diverge near the nodal set so as to compensate for the vanishing of $p(\mathbf{x})$. While such constructions can partially tame divergences, they typically rely on continuum-space diagnostics and substantially alter the target distribution. This increases algorithmic complexity, requires additional evaluations per configuration, and makes the overall computational cost difficult to control, particularly in the context of modern neural-network quantum states (NQS).

These examples illustrate that designing a suitable reference distribution involves balancing several competing objectives. An admissible reference distribution must simultaneously satisfy the following requirements:

- (i) It should have sufficiently broad support to regularize nodal singularities;
- (ii) it should control the fluctuations of $\omega(\mathbf{x})$ to prevent effective-sample-size collapse;
- (iii) it should remain computationally efficient for sampling and evaluation of the reweighting factor.

These requirements are generally in tension with one another, and no existing reference distribution satisfies all of them simultaneously in a generic VMC setting (see Table I).

III. BLURRED SAMPLING METHOD

Overcoming these challenges requires a different strategy. Conceptually inspired by the bridge-link method in

TABLE I. Comparison between different importance-sampling methods and the blurred sampling approach. The ESS scaling refers to how the effective sample size, defined in Eq. (12), scales with the system size N . The additional complexity quantifies the extra computational cost beyond standard sampling, measured in terms of additional wave-function evaluations, which typically constitute the main computational bottleneck of the algorithm. For the middle two methods, this cost is difficult to express purely in terms of wave-function evaluations; instead, we list the primary additional quantity required by each method. In discrete space, blurred sampling introduces zero overhead, whereas in the continuum the cost scales as N_{conn} .

Method	Infinite variance	Force bias	ESS scaling	Sampling	Additional complexity
Overdispersed sampling $ \psi_\theta ^\alpha$ [63,64]	Solved	Unsolved	$O(e^{-N})$	Modified	0
E_{loc} regularization [66]	Solved	Unsolved	$O(1)$	Modified	E_{loc}
Node distance regularization [65]	Solved	...	$O(1)$	Modified	$\nabla \psi_\theta$
Blurred sampling (this work)	Solved	Solved	$\geq 1 - q$	Postprocessing	0; N_{conn}

determinant quantum Monte Carlo [53,55], we introduce a local mixing scheme applied to sampled configurations as a postprocessing step. This construction, which we term *blurred sampling*, modifies configurations without redefining the underlying sampling distribution. Intuitively, it “blurs” sampled configurations, replacing the singular sampling measure with a locally mixed reference distribution that assigns finite weight near nodal regions. In effect, this produces an effective $r(\mathbf{x})$ that is coupled to $p(\mathbf{x})$ locally, allowing the reweighting factor $\omega(\mathbf{x})$ to remain well behaved as sampling traverses the widely varying landscape of $p(\mathbf{x})$.

A. Construction

Starting from a configuration $\mathbf{x} \sim p(\mathbf{x})$, a new configuration $\mathbf{x}'$ is generated according to a *blur kernel*, defined by the discrete Markov transition kernel

$$K(\mathbf{x}'|\mathbf{x}) = (1 - q)\delta_{\mathbf{x}'\mathbf{x}} + qK_{\text{off}}(\mathbf{x}'|\mathbf{x}), \quad (6)$$

where $0 \leq q < 1$ controls the strength of the blurring. Operationally, with probability $1 - q$ the configuration $\mathbf{x}$ is left unchanged, while with probability q it is updated according to the normalized, off-diagonal kernel $K_{\text{off}}(\mathbf{x}'|\mathbf{x})$, which satisfies $K_{\text{off}}(\mathbf{x}|\mathbf{x}) = 0$ and $\sum_{\mathbf{x}'} K_{\text{off}}(\mathbf{x}'|\mathbf{x}) = 1$ for any $\mathbf{x}$. This postprocessing step induces an implicit reference distribution for $\mathbf{x}'$:

$$r(\mathbf{x}') = \sum_{\mathbf{x}} K(\mathbf{x}'|\mathbf{x}) p(\mathbf{x}). \quad (7)$$

Because all contributions to $r(\mathbf{x}')$ are non-negative, we have $r(\mathbf{x}') > 0$ whenever there exists a configuration $\mathbf{x}$ with $K(\mathbf{x}'|\mathbf{x}) > 0$ and $p(\mathbf{x}) > 0$. The blurring operation, therefore, expands the support of the sampling distribution according to the connectivity structure of $K_{\text{off}}(\mathbf{x}'|\mathbf{x})$, satisfying requirement (i) in Sec. II C, given an appropriately chosen kernel $K(\mathbf{x}'|\mathbf{x})$. The blur kernel can be chosen in a variety of ways; below, we present representative constructions for continuum and discrete systems and illustrate their effectiveness using the two pedagogical examples introduced in Sec. II B.

1. Continuum systems

In continuum space where $\mathbf{x} = \{\vec{r}_1, \dots, \vec{r}_{n_e}\}$, singular behavior arises from the vanishing of the wave function along nodal manifolds. Locally, the wave function varies linearly along normal directions to the nodal set. It, therefore, suffices to introduce small local perturbations of configurations so as to assign nonzero weight to configurations in the nodal set. To cover all local coordinate directions, one simple way is to define the connected set

$$\mathcal{C}^{\text{out}}(\mathbf{x}) = \{\mathbf{x} \pm \epsilon \mathbf{e}_{i\alpha} | i = 1, \dots, n_e, \alpha = 1, \dots, d\}, \quad (8)$$

where $\mathbf{e}_{i\alpha}$ shifts the α th coordinate of particle i by one unit and ϵ is a small parameter to control the noise strength. We then use the blur kernel

$$K(\mathbf{x}'|\mathbf{x}) = (1 - q)\delta_{\mathbf{x}'\mathbf{x}} + \frac{q}{N_{\text{conn}}} \sum_{\mathbf{y} \in \mathcal{C}^{\text{out}}(\mathbf{x})} \delta_{\mathbf{x}'\mathbf{y}}, \quad (9)$$

where $N_{\text{conn}} = |\mathcal{C}^{\text{out}}(\mathbf{x})| = 2dn_e$. This means that, for each sampled configuration $\mathbf{x}$, with probability q one particle is randomly displaced by ϵ along one of the coordinate directions. The connectivity, therefore, remains sparse and scales linearly with system size. Since all dn_e local coordinate directions are covered, the procedure effectively smears nodal singularities [see Fig. 1(b)]. For the example shown in Fig. 2, blurred sampling ($q = 2/3$, $\epsilon = 0.2$) straightforwardly eliminates the infinite-variance problem.

2. Discrete configuration space

In discrete spaces, the bias of the force $\mathbf{F}$ originates from the support mismatch between ψ_θ and $\hat{H}\psi_\theta$. We, therefore, exploit the connectivity structure of the Hamiltonian to define the blur kernel. Specifically, we denote by

$$\mathcal{C}^{\text{out}}(\mathbf{x}) = \{\mathbf{y} \neq \mathbf{x} | \langle \mathbf{y} | \hat{H} | \mathbf{x} \rangle \neq 0\} \quad (10)$$

the set of configurations connected to $\mathbf{x}$ by $\hat{H}$ and adopt the same kernel as in Eq. (9). More refined kernels are also possible; for example, the transition probabilities can be weighted according to the matrix elements of $\hat{H}$, since many Monte Carlo quantities, such as the force and energy variance, involve the same matrix elements through E_{loc} .

With this construction, the support-mismatch bias is eliminated for the variational force $\mathbf{F}$. Since $\text{supp}(\psi_\theta) \cup \text{supp}(\hat{H}\psi_\theta) \subseteq \text{supp}(r)$, all configurations contributing to the force are sampled, and the resulting estimator becomes unbiased. In the example in Fig. 3, blurred sampling ($q = 0.5$) removes both the bias and the large fluctuations of the gradient estimator, leading to stable and smooth ground-state optimization.

B. Statistical guarantees and computational scaling

The blurred sampling method possesses two key statistical properties. First, the reweighting factor satisfies $\langle \omega(\mathbf{x}') \rangle_r = 1$ as a direct consequence of the Markov property (see Appendix A 1). As a result, the SNIS estimator in Eq. (5) has a standard importance-sampling form, and the following estimator (an analogous expression holds for QGT)

$$\hat{F}_i^{\text{blur}} = \frac{N[\omega]_r^2}{N-1} \times \left(\frac{[\omega \mathcal{O}_i^* E_{\text{loc}}]_r}{[\omega]_r} - \frac{[\omega \mathcal{O}_i^*]_r [\omega E_{\text{loc}}]_r}{[\omega]_r^2} \right) \quad (11)$$

is finite-sample unbiased, i.e., $\mathbb{E}[\hat{F}_i^{\text{blur}}] = \lim_{N \rightarrow \infty} \hat{F}_i^{\text{blur}}$ (see Appendix A 2). Here, $[\cdot]_r = (1/N) \sum_{\mathbf{x}'} \cdot(\mathbf{x}')$ denotes the

finite-sample average over samples drawn from r . This estimator differs from the SNIS estimator only by an overall factor $N[\omega]_r^2/(N-1)$. Consequently, in solving the TDVP equation in Eq. (2), this common factor cancels in the linear system. Blurred sampling can, therefore, be implemented operationally in a self-normalized form while retaining finite-sample unbiasedness.

Second, the fluctuations of $\omega(\mathbf{x}')$ are explicitly controlled. From the construction of the blur kernel in Eq. (6), the reweighting factor is bounded as $0 \leq \omega(\mathbf{x}') \leq 1/(1-q)$. This boundedness immediately implies a lower bound on the effective sample size (ESS) (see Appendix A 3):

$$\text{ESS}[\omega(\mathbf{x}')] \equiv \frac{\langle \omega(\mathbf{x}') \rangle_r^2}{\langle \omega(\mathbf{x}')^2 \rangle_r} \geq 1 - q. \quad (12)$$

Thus, for any fixed $0 < q < 1$, the effective sample size remains finite, preventing its exponential degradation with increasing system size. Consequently, the blurred sampling method avoids the finite-sample bias inherent to SNIS and satisfies requirement (ii) outlined in Sec. II C. Importantly, the method is robust to the precise choice of q : For a broad range of $q \sim O(1)$, the estimator remains stable, as demonstrated in the numerical results later.

The blurred sampling method also exhibits well-controlled computational cost. The blurring step is a *post-processing* procedure and does not require additional wave-function evaluations; hence, it does not inherit the computational complexity of the underlying *Ansätze* and can, therefore, be combined seamlessly with sophisticated sampling methods [67,68] and autoregressive networks [69–75]. The only additional cost, measured in units of wave-function evaluations (the dominant expense in VMC), arises from computing the reweighting factor

$$\omega(\mathbf{x}') = \frac{p(\mathbf{x}')}{(1-q)p(\mathbf{x}') + \sum_{\mathbf{x} \in \mathcal{C}^{\text{in}}(\mathbf{x}')} K(\mathbf{x}'|\mathbf{x})p(\mathbf{x})}, \quad (13)$$

where $\mathcal{C}^{\text{in}}(\mathbf{x}') = \{\mathbf{x} | K(\mathbf{x}'|\mathbf{x}) > 0\}$ denotes the set of configurations connected to $\mathbf{x}'$ through the kernel. Evaluating $\omega(\mathbf{x}')$, therefore, requires $|\mathcal{C}^{\text{in}}(\mathbf{x}')|$ wave-function evaluations. To preserve computational scalability, we restrict the blur kernel to be sparsely connected, such that $|\mathcal{C}^{\text{in}}(\mathbf{x}')|$ grows at most polynomially with system size. Under this constraint, blurred sampling does not alter the overall computational scaling of standard VMC, thereby satisfying requirement (iii) in Sec. II C.

In the continuum case, the kernel introduced above is symmetric and sparsely connected, with $|\mathcal{C}^{\text{in}}(\mathbf{x})| = |\mathcal{C}^{\text{out}}(\mathbf{x})| = 2dn_e$, which scales linearly with system size. In discrete configuration spaces, evaluating the reweighting factor in Eq. (13) requires amplitudes of the wave function only on configurations connected by $\hat{H}$. These amplitudes are already computed in the evaluation of the local energy. Consequently, blurred sampling in the discrete case introduces essentially no additional computational cost.

C. Generalizations

The constructions introduced above rely on a fixed blur kernel with a prescribed connectivity structure. While this deterministic choice suffices to eliminate the dominant pathologies—such as infinite variance and bias in the variational force—there remain situations in which a broader exploration of configuration space may be desirable. One possibility would be to employ more densely connected kernels, with the associated reweighting factor estimated through an auxiliary Monte Carlo procedure [53]. Although such constructions expand the accessible support, they involve an additional stochastic layer (alternative to the separate Monte Carlo integration, a different computational structure [86] has been explored).

This motivates a randomized generalization of blurred sampling that enlarges the effective support while preserving computational efficiency. Specifically, consider a family of sparse blur kernels $\{K_\lambda\}$ with a connectivity structure that is parametrized by λ , each satisfying the structure of Eq. (6). For each sample $\mathbf{x}_i \sim p(\mathbf{x})$, we independently draw $\lambda_i \sim f(\lambda)$ from a normalized density f and then generate $\mathbf{x}'_i \sim K_{\lambda_i}(\cdot|\mathbf{x}_i)$. Let $r_\lambda(\mathbf{x}') = \sum_{\mathbf{x}} p(\mathbf{x})K_\lambda(\mathbf{x}'|\mathbf{x})$ denote the blurred distribution corresponding to a fixed λ . We define the conditional reweighting factor $\omega_\lambda(\mathbf{x}') = [p(\mathbf{x}')/r_\lambda(\mathbf{x}')]$.

The estimator in Eq. (11) then applies analogously under randomized blurred sampling, with $[\omega]_r$ replaced by $[\omega_\lambda]_{f(\lambda)r_\lambda}$. Through the stochastic choice of K_λ , the sampling effectively explores a broader support while preserving finite-sample unbiasedness. Importantly, the weight is evaluated conditionally with respect to the realized kernel K_λ , so the computational cost is controlled by the sparsity of each individual kernel rather than by the averaged kernel. This randomized construction increases flexibility while preserving scalability. As illustrated in later examples, such flexibility can further mitigate instabilities in the QGT and enable more refined blur kernels like a Gaussian kernel in the continuum case. A detailed derivation of the randomized blurred sampling scheme is provided in Appendixes A 4 and A 5.

In summary, blurred sampling provides a structurally robust solution under the general importance-sampling scheme. Unlike overdispersed or explicitly regularized reference distributions (see Table I), it preserves the underlying sampling dynamics and avoids reweighting factors with system-size-dependent variance growth. Instead, it operates as a sparse postprocessing step that restores missing support and regularizes singular estimators while maintaining computational scalability.

IV. RESULTS

Having established the conceptual foundation of blurred sampling and validated it on pedagogical examples, we now turn to realistic settings. We first consider two ground-state optimization examples in Sec. IV A: the two-dimensional electron gas in continuum space and the

J_1 - J_2 model in discrete space. We show that blurred sampling removes the infinite-variance problem and the statistical pathologies in the Monte Carlo estimators, makes the optimization more stable, and allows faster convergence to the ground state. We then move on to t-VMC, where statistical pathologies have been explicitly identified in the literature [56,87]. In this setting, the correctness of the time evolution hinges directly on the unbiasedness and stability of the Monte Carlo estimators entering the TDVP equation. We show that blurred sampling resolves previously identified failure cases and scales efficiently to large-scale spin-dynamics simulations. Thus blurred sampling directly addresses a fundamental limitation in applications of t-VMC for quantum dynamics.

A. Ground-state optimization

1. Continuum-space example

Here, we consider one of the iconic examples in the continuum, the two-dimensional electron gas, with Hamiltonian

$$\hat{H} = -\frac{1}{2} \sum_{i=1}^N \nabla_i^2 + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + \text{bg}, \quad (14)$$

in a periodic cell, where bg denotes the background terms required to enforce charge neutrality. We use atomic units and follow conventional notations, with the electron density parametrized by $r_s = (1/\sqrt{\pi n})$, where $n = N/A$ is the particle density in a square cell of area A . We test a state-of-the-art neural quantum state, (MP)²NQS, as the trial wave function. See, e.g., Ref. [40] for additional details. The blur strength q and the blur distance ϵ are the two hyperparameters. We choose $q = 0.5$ and $\epsilon = 1$ here. The method remains effective even for small q over a broad range of values. A detailed analysis of the q dependence of the estimator variance is given in Appendix B 1.

In Fig. 4(a), we show the optimization results for an $N = 45$, $r_s = 0.5$, spin-polarized system. We use the blur kernel defined in Eqs. (8) and (9). It is clear from Fig. 4(a) that blurred sampling accelerates the optimization process compared with the standard method. To confirm that the slowing down is caused by nodal singularities, we also track the energy variance, which is crucial for reliable estimates of the ground-state energy and variance extrapolations. As shown in Fig. 4(b), the energy-variance estimator under standard sampling exhibits repeated spikes, which disappear once blurred sampling is turned on. Figure 4(c) shows the single-sample distribution of $\omega(\mathbf{x})(E_{\text{loc}}(\mathbf{x}) - \bar{E})^2$, where $\bar{E}$ is the energy estimate from

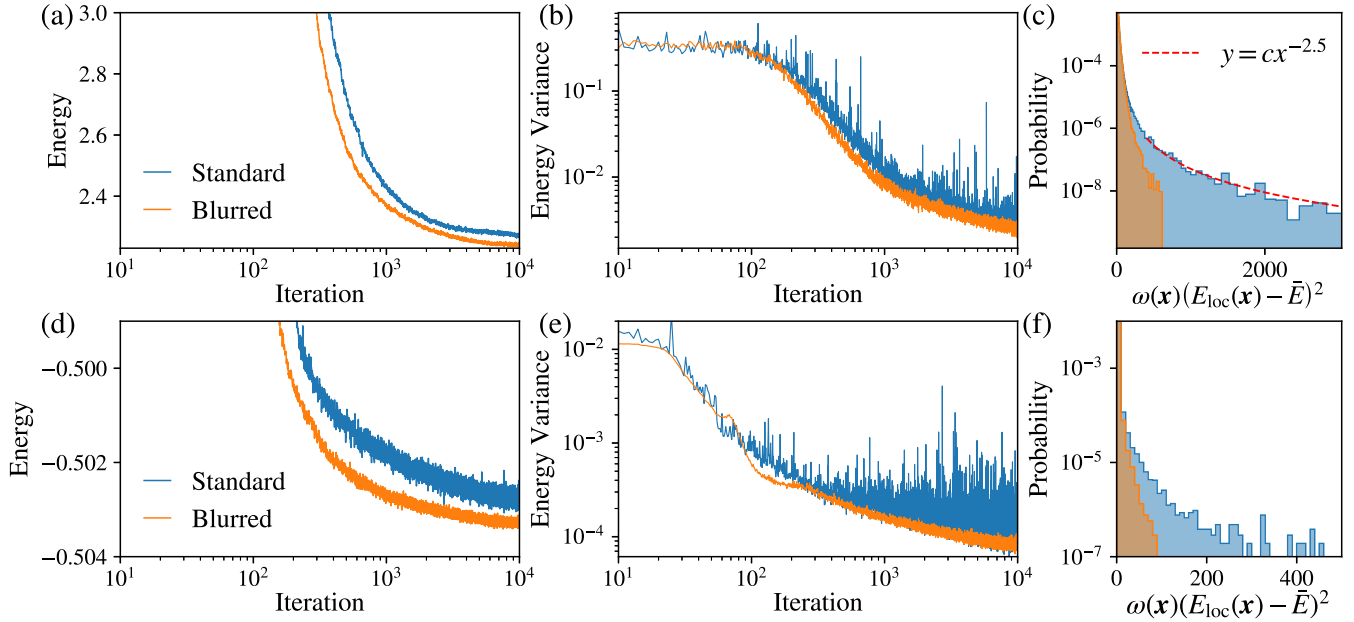


FIG. 4. Ground-state optimization tests. The top row shows a continuum example of the two-dimensional electron gas, with $N = 45$, $r_s = 0.5$, and full spin polarization: (a) variational energy per electron, E/N , in Hartree, as a function of optimization iteration for standard and blurred sampling; (b) corresponding energy variance during optimization, illustrating the suppression of large stochastic spikes by blurred sampling; (c) the distributions of the energy-variance estimator $\omega(\mathbf{x})(E_{\text{loc}}(\mathbf{x}) - \bar{E})^2$. In (c), the red dashed line is a power-law fit of the state obtained by standard sampling, confirming a heavy-tail distribution with exponent $\alpha = 3/2$, corresponding to a probability-density tail exponent of $5/2$. The bottom row shows a discrete example, the $J_1 - J_2$ model on a 6×6 lattice at the frustrated point $J_2/J_1 = 1/2$, with a similar study: (d) variational energy per site for optimization trajectories with standard sampling and blurred sampling; (e) corresponding variance of the energy during optimization; (f) distributions of the corresponding energy-variance estimators obtained at the end of the optimizations.

the full sample and $\omega(x) = 1$ for standard sampling. The expectation value of this quantity is the energy variance, with or without blurred sampling. As discussed in Sec. II B 1, standard sampling is expected to produce a heavy tail with exponent $\alpha = 3/2$, which is confirmed by the fit shown as the dashed red line in Fig. 4(c). With blurred sampling, the tail is strongly suppressed, the variance scaling is modified, and the central limit theorem is restored, effectively removing the infinite-variance problem caused by nodal pathologies.

2. Discrete-space example

For a discrete-space benchmark of ground-state optimization, we consider the frustrated J_1 - J_2 Heisenberg model, with Hamiltonian

$$\hat{H} = J_1 \sum_{\langle i,j \rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j + J_2 \sum_{\langle\langle i,j \rangle\rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j, \quad (15)$$

where $\langle i,j \rangle$ denotes nearest-neighbor pairs and $\langle\langle i,j \rangle\rangle$ denotes next-nearest-neighbor pairs. In our simulations, we study a 6×6 square lattice with periodic boundary conditions. We set $J_2/J_1 = 1/2$ and $J_1 = 1$, the most frustrated point, where the competition between J_1 and J_2 makes the optimization particularly challenging.

In this case, although the problem may not exhibit strict support mismatch, frustration can still induce strong fluctuations in the wave-function amplitude, making the estimators extremely noisy. We, therefore, apply blurred sampling using a Hamiltonian-induced kernel. As in the continuum benchmark, the blurred estimator is implemented as a postprocessing step that leaves the original sampling dynamics unchanged. We took $q = 0.5$ here, using 16384 samples. The optimizations use the SR method, with all hyperparameters kept identical between standard and blurred sampling. Details of the *Ansatz* and optimization are provided in Appendix C. As in the continuum example, blurred sampling improves the stability and accuracy of the optimization, as shown in Fig. 4(d). We also monitor the energy variance, shown in Fig. 4(e). Large spikes appear in the standard method, whereas blurred sampling removes them. Finally, the distribution in Fig. 4(f) shows that the standard estimator has a heavy tail, which is suppressed by blurred sampling as expected.

The above two examples on ground-state VMC show that blurred sampling is a simple and effective stabilization method for SR optimization in both continuous and discrete configuration spaces. The statistical fluctuations originating from nodal singularities in ratio-type estimators are a very general feature, which can be remedied by the approach discussed here. Because blurred sampling is implemented as a postprocessing step with only modest overhead, it can be dropped into existing SR work flows without changing the sampler or variational *Ansatz*. This suggests that blurred sampling can serve as a default in all

such calculations in the future whenever estimator noise or support-mismatch pathologies are a concern.

B. t-VMC: Previously identified difficult examples

In ground-state optimization, nodal and support-mismatch pathologies primarily manifest as heavy-tailed fluctuations, noisy estimators, and slow convergence. In t-VMC, however, they can qualitatively alter the projected dynamics through biased force and QGT estimators, leading to incorrect time evolution even in the infinite-sample limit. Here, we first revisit two benchmarks where standard t-VMC is known to fail due to bias in the force estimators.

1. Single spin

Following Ref. [56], we evolve $|\psi_\theta\rangle = \alpha|0\rangle + \beta|1\rangle$ under $\hat{H} = Y$. As the state approaches $|0\rangle$, standard sampling collapses due to a support-mismatch bias. Blurred sampling removes the bias and reproduces the exact dynamics, as shown in Fig. 5(a).

2. 2×2 Heisenberg quench

We next consider the anisotropic-bond Heisenberg model [87]:

$$\hat{H}_{\text{HM}} = \sum_{\langle i,j \rangle} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j + \Delta J \sum_{(i,j) \in \text{vert}} \hat{\mathbf{S}}_i \cdot \hat{\mathbf{S}}_j, \quad (16)$$

where the second term runs over the vertical bonds of the lattice. In Ref. [87], it was shown that a complex restricted

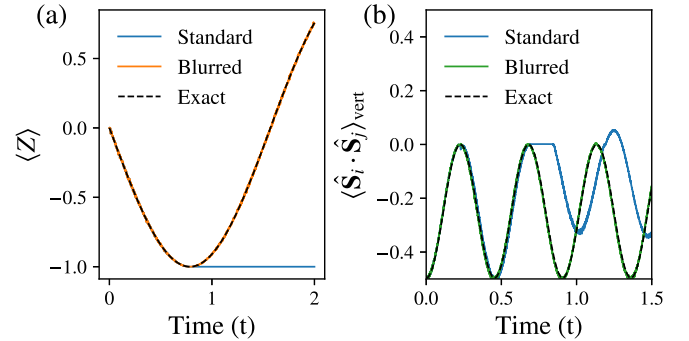


FIG. 5. t-VMC results for two known benchmark problems. (a) We use the variational state $|\psi_\theta\rangle = \alpha|0\rangle + \beta|1\rangle$, with $(\alpha, \beta) = (1, 1)$ at $t = 0$ and evolve it under $U(t) = \exp\{-itY\}$ with t-VMC. As shown in Ref. [56], the standard Monte Carlo estimator collapses due to finite-support bias as the variational state gets close to $|0\rangle$, causing the dynamics to diverge. With blurred sampling, the correct dynamics is recovered. (b) A complex RBM with a single hidden unit fails to provide the correct dynamics under t-VMC for the 2×2 Heisenberg Hamiltonian [87], as the spin correlation diverges from the exact result using the standard sampler. The blurred sampling method produces the correct dynamics. In both cases, we use $q = 0.5$ Hamiltonian-induced kernel for blurred sampling.

Boltzmann machine (RBM) with a single hidden parameter fails to reproduce the correct real-time dynamics under t-VMC for the 2×2 system. The initial state is taken as the ground state of $\hat{H}$ at $\Delta J = 0$ and is subsequently quenched to $\Delta J = -2$. The correlation $\sum_{(i,j) \in \text{vert}} \langle \hat{S}_i \cdot \hat{S}_j \rangle \equiv \langle \hat{S}_i \cdot \hat{S}_j \rangle_{\text{vert}}$ is measured as a function of time. Consistent with their findings, the standard Monte Carlo method breaks down due to support-mismatch bias, whereas blurred sampling successfully restores the correct dynamics, as shown in Fig. 5(b). (A more detailed analysis of this example is provided in Appendix B 4 as a pedagogical example.)

C. Solution of the parity-mixing problem

We now turn to a larger-scale problem that captures a physically relevant mechanism: symmetry-sector mixing during nonequilibrium dynamics. In many quantum quenches, the initial state that is being evolved may reside entirely within a specific symmetry sector, while the quench Hamiltonian couples different symmetry sectors to each other. The exact time evolution then necessarily involves coherent transfer of weight between them. Accurately describing such sector mixing is essential for capturing dynamical symmetry restoration, collective oscillations, and, more generally, symmetry-changing processes in nonequilibrium systems.

To construct a setting where this mechanism appears in a controlled manner, we consider an n -site, spin- $\frac{1}{2}$ system initialized in the equal superposition of all configurations $\mathbf{x} \in \{\pm 1\}^n$ with even Z parity:

$$|\psi_{\theta}^{\text{even}}\rangle = \sum_{\mathbf{x} \in \mathcal{X}_{\text{even}}} |\mathbf{x}\rangle, \quad (17)$$

where $\mathcal{X}_{\text{even}} = \{\mathbf{x} \in \{\pm 1\}^n : \hat{P}_Z |\mathbf{x}\rangle = +|\mathbf{x}\rangle\}$ and the Z -parity operator is defined as $\hat{P}_Z = \prod_{i=1}^n Z_i$. Here, $|\mathbf{x}\rangle$ denotes a computational-basis (Z -basis) state.

The system is evolved under the Hamiltonian of a transverse-field Ising model (TFIM) with periodic boundary conditions:

$$\hat{H}_{\text{TFIM}} = J \sum_{i=1}^n Z_i Z_{i+1} + h \sum_{i=1}^n X_i. \quad (18)$$

Crucially, the spin-flip terms in $\hat{H}_{\text{TFIM}}$ map even-parity configurations to odd-parity ones, so that $\hat{H}|\psi_{\theta}^{\text{even}}\rangle$ has support on $\mathcal{X}_{\text{odd}} = \{\mathbf{x} \in \{\pm 1\}^n : \hat{P}_Z |\mathbf{x}\rangle = -|\mathbf{x}\rangle\}$, while $|\psi_{\theta}^{\text{even}}\rangle$ itself has support only on $\mathcal{X}_{\text{even}}$. As a consequence, the force estimator becomes systematically biased, since Monte Carlo sampling under $p(\mathbf{x})$ cannot access configurations that carry nonzero weight in $\hat{H}_{\text{TFIM}}|\psi_{\theta}^{\text{even}}\rangle$.

In Fig. 6, we investigate the dynamics with t-VMC, by computing the time-dependent expectation value of the Z parity under real-time evolution $U(t) = \exp(-it\hat{H}_{\text{TFIM}})$,

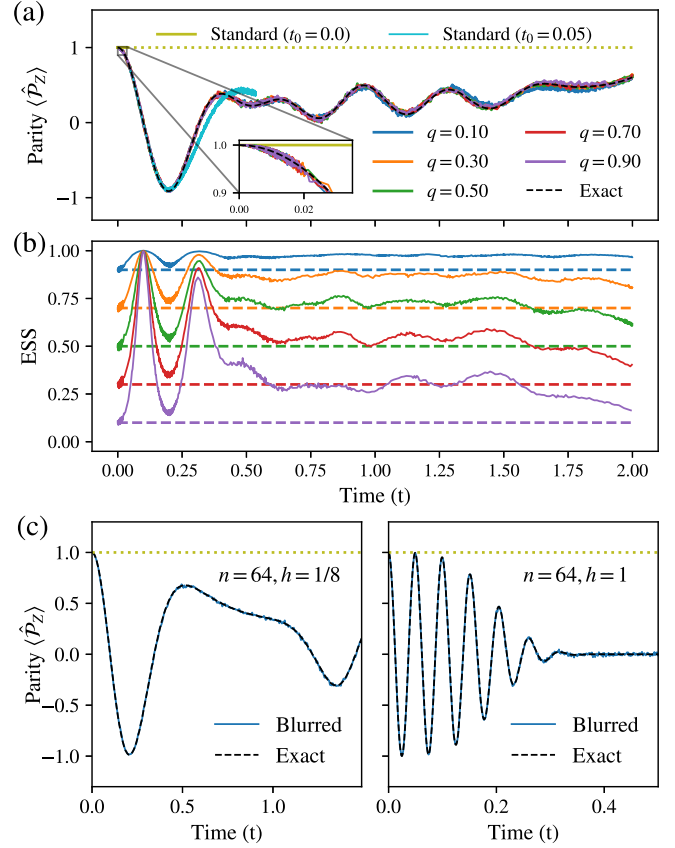


FIG. 6. Accurate simulations of spin parity mixing with blurred sampling in t-VMC. (a) Time evolution of the parity operator for the quench of the eight-spin state $|\psi_{\theta}^{\text{even}}\rangle$ under the TFIM Hamiltonian at $J = h = 1$. The variational state is an RBM with 32 hidden units. Different lines indicate simulations with different values of the blur strength q . The inset shows that starting the t-VMC simulation with the standard estimator at $t_0 = 0$ leads to the immediate freezing of the dynamics due to the bias of the force. The dotted line shows the expected behavior of the parity for the standard $t = 0$ dynamics. If we start the simulation at $t_0 = 0.05$, the standard estimator performs well for the initial ramp of the oscillation but diverges for longer times. (b) The effective sample size $\text{ESS} = \mathbb{E}[\omega(\mathbf{x}')^2] / \mathbb{E}[\omega(\mathbf{x}')^2]$. The dashed lines indicate the value $1 - q$. At initialization, the ESS is exactly $1 - q$. For all q , the ESS peaks whenever the parity goes through 0, indicating that the state is balanced between odd- and even-parity sectors. (c) Time evolution of the parity operator using a Gaussian-state Ansatz, with $n = 64$. We show the slow ($h = 1/8$) and fast ($h = 1$) quench on the left and right, respectively. The dotted line indicates the expected freezing of the t-VMC dynamics with the standard estimator. Blurred sampling uses the Hamiltonian-induced kernel with $q = 0.5$.

starting from $|\psi_{\theta}^{\text{even}}\rangle$. For the $n = 8$ system, we employ a complex RBM as the variational Ansatz. For this architecture, the initial state can be represented analytically (see Appendix D 1), and the full dynamics can be simulated exactly with Majorana fermions (see Appendix D 2). At $t = 0$, the standard Monte Carlo sampling draws configurations only from the even-parity sector. Consequently, the

parity remains constant, as evidenced by the flat curve in Fig. 6(a), and the TDVP dynamics is, incorrectly, frozen. This is a support-mismatch problem; accordingly, simply making the sampler more overdispersed does not restore the missing odd-parity sector, as shown in Appendix B 3.

One might expect that initializing the evolution slightly away from the pathological point, i.e., at a small $t_0 > 0$ where the state already contains a small odd-parity component, would restore the correct dynamics. However, we observe that standard sampling still fails: Although odd-sector configurations are now formally present, their statistical weight remains extremely small, leading to a severely biased and noisy force estimator. The resulting trajectory deviates from the exact solution at later times as shown in Fig. 6(a).

In contrast, blurred sampling restores access to the missing parity sector at every time step and faithfully reproduces the exact real-time dynamics over long times. Moreover, the results are robust with respect to the blur strength q , which we vary in the range $q \in [0.1, 0.9]$. This demonstrates the stability of the method against changes in its only hyperparameter. We also stress that all previous results presented in this work are similarly insensitive to the choice of q . This robustness stems from the bounded effective sample size and the structural elimination of estimator bias. As shown in Fig. 6(b), the effective sample size remains lower bounded by $1 - q$, in agreement with the theoretical expectation. A more detailed analysis of the signal-to-noise ratio of the force estimator and the ESS as functions of q is presented in Appendix B 2.

To demonstrate scalability, we extend the system size to $n = 64$ spins. For these larger systems, we find that the RBM *Ansatz* is no longer sufficiently expressive to accurately capture the long time evolution. To disentangle variational expressiveness from estimator bias in the t-VMC algorithm, we instead employ a Gaussian state *Ansatz* tailored to the structure of the problem, which is capable of representing the exact state at all times (see Appendix D 3). As shown in Fig. 6(c), blurred sampling accurately reproduces the parity-mixing dynamics for both a slow quench ($h = 1/8$) and a fast quench ($h = 1$), whereas the standard method fails in both cases. This demonstrates that our method remains stable and accurate at large system sizes and across different dynamical regimes.

D. Unbiased spin-relaxation dynamics

We now consider a physically distinct regime in which the variational state is initially highly localized in configuration space. Such situations naturally arise in spin-relaxation dynamics, where the system starts from a nearly classical product state and evolves under a many-body Hamiltonian. Strong localization of the wave function also appears in neural quantum state simulations of quantum circuits, where one typically evolves a single computational

basis state [48]. In all these cases, the state carries significant weight on only a small subset of configurations, and an accurate description of the time evolution requires efficient exploration of the surrounding states in the Hilbert space. They, thus, provide realistic and stringent tests of the new method in a large class of important applications.

We also use this problem to provide an illustration of how the flexibility of the blurred sampling framework can be exploited to further enlarge the effective sampling support. The Hamiltonian-induced blur ensures that $\text{supp}(\hat{H}\psi_\theta)$ is sampled and, therefore, removes the bias in the force estimator. However, the QGT can still suffer from a support mismatch when $\text{supp}[\partial_\theta \psi_\theta(\mathbf{x}) \partial_\theta \psi_\theta(\mathbf{x})] \not\subseteq \text{supp}(\hat{H}\psi_\theta) \cup \text{supp}(\psi_\theta)$ (see further discussion in Appendix A 5). Note that, unlike the bias in the forces, the QGT pathology depends explicitly on the chosen parametrization of the state through the derivatives of the wave function.

We adopt a parametrization of our variational *Ansatz* in which a symmetric complex RBM is multiplied with a factor $f_\phi(\mathbf{x})$, $\phi \in \mathbb{C}^{n+1}$:

$$\psi_\theta(\mathbf{x}) = f_\phi(\mathbf{x}) \psi_{\text{symm-RBM}}(\mathbf{x}).$$

Two forms of f_ϕ are considered to help illustrate the effect of the randomized blurred sampling. In the first, different spin sectors are parametrized independently in $f_\phi(\mathbf{x}) = \phi_{s(\mathbf{x})}$, while in the second we break the independence and let $f_\phi = \phi_1 + \phi_3$ when $s(\mathbf{x}) = 3$, thereby correlating $s = 1$ to the $s = 3$ spin sectors. (See Appendix E 2 for further details.) Below, we refer to the two parametrization schemes as *uncorrelated* and *correlated*, respectively.

The variational state is initialized in the fully polarized all-down product state, with exponentially small amplitudes assigned to all other configurations (see Appendix E 2 for more details). It is straightforward to see that, under the TFIM Hamiltonian ($J = h = 1$), the resulting state has significant support mismatch between the QGT estimator and the forces estimator.

We exploit the flexibility of randomized blurred sampling to further enlarge the effective sampling support. We introduce $\lambda = \{s_1, \dots, s_n\}$, $s_i = \pm 1$ as a flip mask and denote $\lambda \cdot \mathbf{x} = \{x_1 s_1, \dots, x_n s_n\}$ to be the configuration obtained by flipping the spins of $\mathbf{x}$ according to λ . The flip mask λ is sampled uniformly in different spin sectors. The parametrized blur kernel is defined as

$$K_\lambda(\mathbf{x}'|\mathbf{x}) = (1 - q_1 - q_2) \delta_{\mathbf{x}'\mathbf{x}} + \frac{q_1}{N_{\text{conn}}} \sum_{y \in \mathcal{C}^{\text{out}}(\mathbf{x})} \delta_{\mathbf{x}',y} + q_2 \delta_{\mathbf{x}',\lambda \cdot \mathbf{x}}, \quad (19)$$

where q_1 and q_2 control the strength of two distinct blurring mechanisms: q_1 introduces Hamiltonian-induced blurring as discussed in the previous section; q_2 introduces a flip via

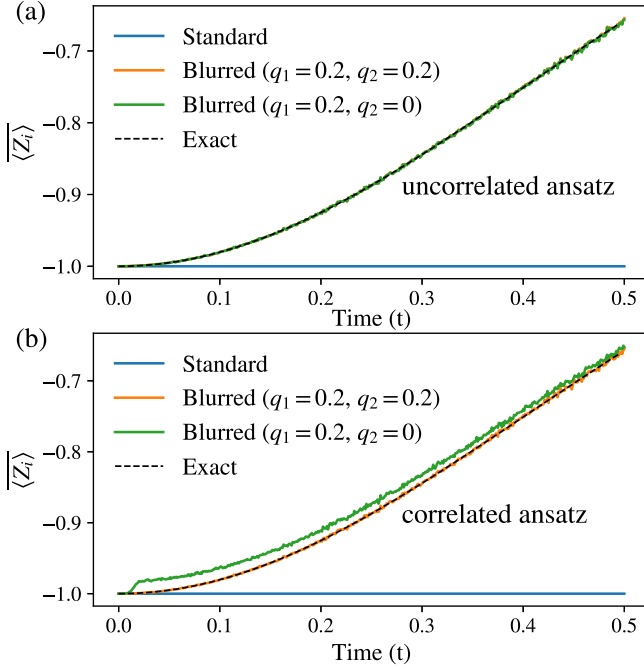


FIG. 7. Solution of the spin-relaxation problem. The time evolution of 32 spins in the TFIM ($J = \hbar = 1$) is simulated, initialized in the fully polarized all-down state. The magnetization, defined as the site-averaged expectation value $\langle Z_i \rangle$, is computed as a function of time and compared against exact results obtained with Majorana fermions (see Appendix E 1). (a) and (b) show the results for two different parametrizations—respectively, without and with correlations between parameters across different spin sectors. For standard t-VMC simulations, the dynamics remain trapped in the initial state for both parametrizations, as expected (see also Ref. [56]). In contrast, blurred sampling with only the Hamiltonian-induced blur ($q_2 = 0$) performs well for the uncorrelated parameterization (a) but deviates from the exact solution in the correlated case (b). Inclusion of randomized blurred sampling (with $q_1, q_2 \neq 0$) successfully captures the relaxation dynamics in both cases.

the sampled mask λ . For each fixed λ , the support of $r_\lambda = \sum_{\mathbf{x}} K_\lambda(\mathbf{x}'|\mathbf{x})p(\mathbf{x})$ always contains $\text{supp}(\hat{H}\psi_\theta)$, ensuring that the force estimator remains unbiased. The additional flip term effectively enlarges the accessible configuration space. Notice that $q_2 = 0$ recovers the Hamiltonian-induced blurred sampling of the previous section.

As shown in Figs. 7(a) and 7(b), the standard estimator fails to reproduce the correct dynamics under $\hat{H}_{\text{TFIM}}$ and remains trapped at the initial state. Interestingly, blurred sampling with only the Hamiltonian-induced term ($q_2 = 0$) already recovers the correct relaxation dynamics in Fig. 7(a), with *uncorrelated* parametrization. This is because of the independent parametrization of the different spin sectors. Once a mild regularization is applied to the QGT, the remaining tangent-space directions are sufficiently well conditioned to yield the correct parameter evolution, even without the additional randomized flips.

In Fig. 7(b), with the *correlated* parametrization, the Hamiltonian-induced blur alone ($q_2 = 0$) is no longer sufficient: The early-time dynamics becomes inaccurate due to the incomplete sampling of tangent-space directions. In this case, the inclusion of randomized blurred sampling restores the correct evolution. See Appendix A 5 for a detailed discussion of why randomized blurred sampling can yield the correct dynamics when the *Ansatz* is sufficiently expressive and Appendix B 5 for an analysis based on the QGT eigenvalues and variational forces.

V. DISCUSSION AND CONCLUDING REMARKS

In this work, we addressed one of the outstanding problems impeding a broad range of neural-network variational calculations in quantum systems, including ground and excited states, finite-temperature properties, and imaginary- and real-time dynamics. We introduced a flexible and general framework, blurred sampling, that regularizes both infinite-variance and support-mismatch pathologies in variational Monte Carlo with a simple sampling scheme which is straightforward to implement. By constructing an implicit reference distribution through a one-step local mixing of configurations, blurred sampling overcomes the challenge of establishing an explicit global reference measure and circumvents the competing requirements that limit conventional importance-sampling approaches.

A key feature of blurred sampling is that it operates as a postprocessing step. As a result, it preserves the original sampling dynamics and autocorrelation structure, making it directly compatible with existing VMC work flows. This distinguishes blurred sampling from methods that rely on modified sampling distributions or additional Markov chains, which can significantly complicate both implementation and analysis.

The generality of the approach suggests several directions for further development. While we focused on simple, physically motivated blur kernels, the framework readily accommodates alternative choices, including problem-specific kernels or adaptive constructions that optimize statistical efficiency. For example, one could tune the blurring strength q to target a fixed ESS. As long as the adaptive strategies depend only on information from previous steps, the method remains unbiased.

In the VMC setting, we show that blurred sampling addresses infinite-variance and support-mismatch pathologies, leading to stable optimization, a well-behaved energy-variance estimator, and reliable time evolution. More generally, blurred sampling offers a promising route for reducing statistical errors in a broad range of observables and may be extended to more sophisticated VMC-based methods, such as projected t-VMC. With the rapid growth of neural quantum state variational calculations, the potential gain in optimization effectiveness and efficiency can impact many applications across physics and chemistry.

Finally, although our analysis has been restricted to variational Monte Carlo, the underlying idea of implicit importance sampling through local configuration mixing is applicable more broadly. Potential extensions include projector quantum Monte Carlo methods, finite-temperature simulations, and classical Monte Carlo problems involving heavy-tailed observables. Exploring these directions can provide a unified perspective on variance regularization across a wide range of stochastic simulation techniques. It is also of interest to examine the applicability of this approach to other optimization problems in machine learning and related areas of artificial intelligence.

All numerical simulations on spin models were performed using NETKET [88,89]. Small-scale exact benchmarks were obtained with QUTIP [90], while larger-scale benchmarks were computed using the Majorana representation [91,92]. Further details of the numerical implementation are provided in the Appendixes.

ACKNOWLEDGMENTS

Z. W. is grateful for discussions with Miguel A. Morales, Conor Smith, and Markus Holzmann and especially thanks Ao Chen for insightful discussions on the lattice model simulations. R. W. thanks Filippo Vicentini and Anka van der Walle for discussions. The Flatiron Institute is a division of the Simons Foundation.

DATA AVAILABILITY

The data that support the findings of this article are openly available [93].

APPENDIX A: DETAILED DERIVATION AND PROPERTIES OF BLURRED SAMPLING

In this section, we derive the blurred sampling procedure and prove the properties stated in the main text. Blurred sampling does not modify the original sampler, which generates configurations $\mathbf{x}$ distributed according to $p(\mathbf{x})$. Throughout this section, we denote by $\mathbf{x}$ samples drawn from p , and by $\mathbf{x}'$ the configuration obtained after the postprocessing step.

1. Unit expectation of the reweighting factor

We first verify that the reweighting factors have unit expectation, i.e., $\langle \omega(\mathbf{x}') \rangle_r = 1$, for the general blur kernel defined in Eq. (6). Since $K(\mathbf{x}'|\mathbf{x})$ is a Markov kernel, $\sum_{\mathbf{x}'} r(\mathbf{x}') = \sum_{\mathbf{x}'} \sum_{\mathbf{x}} K(\mathbf{x}', \mathbf{x}) p(\mathbf{x}) = \sum_{\mathbf{x}} p(\mathbf{x})$, and, hence, r is normalized whenever p is normalized. The reweighting factor is defined as $\omega(\mathbf{x}') = p(\mathbf{x}')/r(\mathbf{x}')$. Therefore,

$$\langle \omega(\mathbf{x}') \rangle_r = \sum_{\mathbf{x}'} r(\mathbf{x}') p(\mathbf{x}') / r(\mathbf{x}') = \sum_{\mathbf{x}'} p(\mathbf{x}') = 1. \quad (\text{A1})$$

2. Finite-sample unbiased estimators

We first construct the finite-sample unbiased estimator for expectation values under the original distribution p . Since $\langle \omega \rangle_r = 1$, we have

$$\langle A \rangle_p = \frac{\langle \omega A \rangle_r}{\langle \omega \rangle_r} = \langle \omega A \rangle_r. \quad (\text{A2})$$

Therefore, the finite size estimator

$$\widehat{\langle A \rangle}_p = \frac{1}{N} \sum_{i=1}^N \omega(\mathbf{x}'_i) A(\mathbf{x}'_i) \equiv [\omega A]_r \quad (\text{A3})$$

is an unbiased estimator for $\langle A \rangle_p$ with $\mathbb{E}[\widehat{\langle A \rangle}_p] = \langle A \rangle_p$.

For covariance-type quantities $\langle\langle A, B \rangle\rangle_p \equiv \langle AB \rangle_p - \langle A \rangle_p \langle B \rangle_p$, we use the identities $\langle AB \rangle_p = \langle \omega AB \rangle_r$, $\langle A \rangle_p = \langle \omega A \rangle_r$, $\langle B \rangle_p = \langle \omega B \rangle_r$. Hence,

$$\langle\langle A, B \rangle\rangle_p = \langle \omega AB \rangle_r - \langle \omega A \rangle_r \langle \omega B \rangle_r. \quad (\text{A4})$$

We now use the identity $\langle \omega \rangle_r = 1$ to rewrite the first term as $\langle \omega AB \rangle_r = \langle \omega \rangle_r \langle \omega AB \rangle_r$, and add-subtract $\langle \omega^2 AB \rangle_r$:

$$\begin{aligned} \langle\langle A, B \rangle\rangle_p &= \langle \omega \rangle_r \langle \omega AB \rangle_r - \langle \omega^2 AB \rangle_r \\ &\quad - \langle \omega A \rangle_r \langle \omega B \rangle_r + \langle \omega^2 AB \rangle_r \\ &= \langle \omega A, \omega B \rangle_r - \langle \omega, \omega AB \rangle_r. \end{aligned} \quad (\text{A5})$$

It is well known that the finite-sample unbiased estimator of the covariance $\langle\langle X_1, X_2 \rangle\rangle_r$ is

$$\widehat{\langle\langle X_1, X_2 \rangle\rangle}_r = \frac{N}{N-1} ([X_1 X_2]_r - [X_1]_r [X_2]_r). \quad (\text{A6})$$

Therefore, we have that

$$\begin{aligned} \widehat{\langle\langle A, B \rangle\rangle}_p &= \widehat{\langle\langle \omega A, \omega B \rangle\rangle}_r - \widehat{\langle\langle \omega, \omega AB \rangle\rangle}_r \\ &= \frac{N}{N-1} \times ([\omega]_r [\omega AB]_r - [\omega A]_r [\omega B]_r) \\ &= \frac{N [\omega]_r^2}{N-1} \times \left(\frac{[\omega AB]_r}{[\omega]_r} - \frac{[\omega A]_r}{[\omega]_r} \frac{[\omega B]_r}{[\omega]_r} \right) \end{aligned} \quad (\text{A7})$$

is the finite-sample unbiased estimator for $\langle\langle A, B \rangle\rangle_p$.

In practice, one typically employs the self-normalized estimator [i.e., the right-hand side of the last line in Eq. (A7)], which is a biased estimator. However, this bias manifests itself as a common multiplicative factor affecting all covariancelike quantities. Since the SR or TDVP update depends only on ratios of such quantities (e.g., through $S^{-1} \mathbf{F}$), this common factor cancels in the parameter update. Consequently, the resulting time evolution is unaffected by this overall factor.

3. Bound of the reweighting factor and effective sample size

Consider the distribution obtained after marginalizing the kernel K in Eq. (6):

$$\begin{aligned} r(\mathbf{x}') &= \sum_{\mathbf{x}} K(\mathbf{x}'|\mathbf{x}) p(\mathbf{x}) \\ &= \sum_{\mathbf{x}} ((1-q)\delta_{\mathbf{x}',\mathbf{x}} p(\mathbf{x}) + qK_{\text{off}}(\mathbf{x}'|\mathbf{x}) p(\mathbf{x})) \\ &= (1-q)p(\mathbf{x}') + q \sum_{\mathbf{x}} K_{\text{off}}(\mathbf{x}'|\mathbf{x}) p(\mathbf{x}). \end{aligned} \quad (\text{A8})$$

Since $K_{\text{off}}(\mathbf{x}'|\mathbf{x}) p(\mathbf{x}) \geq 0$, we see that

$$r(\mathbf{x}') \geq (1-q)p(\mathbf{x}'). \quad (\text{A9})$$

As a result, the reweighting factor is bounded by

$$\omega(\mathbf{x}') = \frac{p(\mathbf{x}')}{r(\mathbf{x}')} \leq \frac{p(\mathbf{x}')}{(1-q)p(\mathbf{x}')} = \frac{1}{1-q}. \quad (\text{A10})$$

Using the fact that $p(\mathbf{x}') \geq 0$, we conclude that the reweighting factor is bounded as

$$0 \leq \omega(\mathbf{x}') \leq \frac{1}{1-q}. \quad (\text{A11})$$

Since the reweighting factors satisfy $\omega(\mathbf{x}) \in [0, 1/(1-q)]$, we have the pointwise inequality

$$\omega(\mathbf{x})^2 \leq \frac{1}{1-q} \omega(\mathbf{x}). \quad (\text{A12})$$

Taking expectations with respect to r yields

$$\langle \omega^2(\mathbf{x}') \rangle_r \leq \frac{1}{1-q} \langle \omega(\mathbf{x}') \rangle_r. \quad (\text{A13})$$

Using $\langle \omega(\mathbf{x}') \rangle_r = 1$, we obtain

$$\text{ESS}[\omega] = \langle \omega(\mathbf{x}') \rangle_r^2 / \langle \omega(\mathbf{x}')^2 \rangle_r \geq 1-q. \quad (\text{A14})$$

4. Randomized blur construction

For the randomized generalization of blurred sampling, we first recall the construction. Let K_λ be a family of blur kernels parametrized by λ , where each kernel satisfies the structure of Eq. (6) and has sparse connectivity. For each sample $\mathbf{x}_i \sim p(\mathbf{x})$, we independently draw $\lambda_i \sim f(\lambda)$ from a normalized density f and then generate $\mathbf{x}'_i \sim K_{\lambda_i}(\cdot|\mathbf{x}_i)$. Let $r_\lambda(\mathbf{x}') = \sum_{\mathbf{x}} p(\mathbf{x}) K_\lambda(\mathbf{x}'|\mathbf{x})$ denote the blurred distribution corresponding to a fixed λ . We define the conditional reweighting factor $\omega_\lambda(\mathbf{x}') = [p(\mathbf{x}')/r_\lambda(\mathbf{x}')]_r$. Then, the estimator $[\omega_\lambda A]_{f(\lambda)r_\lambda}$ remains finite-sample unbiased and equal to $\langle A \rangle_p$:

$$\langle A(\mathbf{x}') \rangle_p = \langle \langle \omega_\lambda(\mathbf{x}') A(\mathbf{x}') \rangle_{r_\lambda} \rangle_{f(\lambda)} = \langle \omega_\lambda(\mathbf{x}') A(\mathbf{x}') \rangle_{f(\lambda)r_\lambda}, \quad (\text{A15})$$

provided that no support-mismatch pathology occurs. Similarly, the estimator in Eq. (11) also applies under randomized blurred sampling, with $[\omega]_r$ replaced by $[\omega_\lambda]_{f(\lambda)r_\lambda(\mathbf{x}')}$. As a direct application, the probabilistic blur construction can be used to refine the continuum kernel introduced in Eq. (8), where the displacement magnitude ϵ and coordinate directions $\mathbf{e}_\alpha$ are treated as random parameters λ . By sampling these quantities from suitable distributions, the resulting blur approaches an effective Gaussian kernel. In particular, this randomization restores rotational symmetry at the level of the induced proposal distribution.

5. QGT estimator from randomized blurred sampling

Although the marginal distribution of $\mathbf{x}'$ is $r(\mathbf{x}') = \sum_{\lambda} f(\lambda) r_\lambda(\mathbf{x}')$, which coincides with that induced by the average kernel $K_{\text{mix}}(\mathbf{x}'|\mathbf{x}) = \sum_{\lambda} f(\lambda) K_\lambda(\mathbf{x}'|\mathbf{x})$, the present construction differs in how the reweighting factor is evaluated. In our scheme, the weight is evaluated conditionally on the realized value of λ , i.e., with respect to $r_\lambda(\mathbf{x}')$ rather than the marginal $r(\mathbf{x}')$. As a result, the computational cost remains governed by the sparsity of each individual kernel K_λ rather than by the mixed kernel K_{mix} , which can be difficult to calculate or require an auxiliary Monte Carlo calculation. This randomized formulation, therefore, represents a trade-off: While it does not fully expand the support at the level of individual blur kernels, it preserves sparsity and remains computationally feasible in regimes where the deterministic construction becomes costly. As demonstrated in Sec. IV D, randomized blurred sampling successfully resolves cases in which Hamiltonian-induced blurred sampling fails to cure the QGT instability.

While the randomized formulation preserves sparsity and improves robustness in practice, its impact on the QGT estimator can be characterized more precisely. In particular, although the marginal distribution $r(\mathbf{x}')$ may explore the full Hilbert space, the reweighting factor is evaluated conditionally on the realized value of λ , namely with respect to $r_\lambda(\mathbf{x}')$. This conditioning leads to a modified structure of the QGT estimator, which we now make explicit. For convenience, we denote by supp_λ the support of the conditional distribution $r_\lambda(\mathbf{x}')$. We further introduce the indicator function

$$\mathcal{I}_\lambda(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in \text{supp}_\lambda, \\ 0, & \text{otherwise.} \end{cases} \quad (\text{A16})$$

In the infinite-sample limit, the QGT estimator obtained from blurred sampling with a fixed realization of λ is given by (the sum is only on supp_λ as we have samples only there)

$$\begin{aligned}
S_{ij}(\lambda) &= \frac{\sum_{\mathbf{x} \in \text{supp}_\lambda} r_\lambda(\mathbf{x}) \frac{\partial_{\theta_i} \psi_\theta^*(\mathbf{x}) \partial_{\theta_j} \psi_\theta(\mathbf{x})}{r_\lambda(\mathbf{x})}}{\sum_{\mathbf{x} \in \text{supp}_\lambda} r_\lambda(\mathbf{x}) \frac{|\psi_\theta(\mathbf{x})|^2}{r_\lambda(\mathbf{x})}} - \frac{\left[\sum_{\mathbf{x} \in \text{supp}_\lambda} r_\lambda(\mathbf{x}) \frac{\partial_{\theta_i} \psi_\theta^*(\mathbf{x}) \psi_\theta(\mathbf{x})}{r_\lambda(\mathbf{x})} \right] \left[\sum_{\mathbf{x} \in \text{supp}_\lambda} r_\lambda(\mathbf{x}) \frac{\psi_\theta^*(\mathbf{x}) \partial_{\theta_j} \psi_\theta(\mathbf{x})}{r_\lambda(\mathbf{x})} \right]}{\left[\sum_{\mathbf{x} \in \text{supp}_\lambda} r_\lambda(\mathbf{x}) \frac{|\psi_\theta(\mathbf{x})|^2}{r_\lambda(\mathbf{x})} \right]^2} \\
&= \frac{\sum_{\mathbf{x} \in \text{supp}_\lambda} \langle \partial_{\theta_i} \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} - \frac{\langle \partial_{\theta_i} \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \psi_\theta \rangle} \\
&= \frac{\sum_{\mathbf{x}} \mathcal{I}_\lambda(\mathbf{x}) \langle \partial_{\theta_i} \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} - \frac{\langle \partial_{\theta_i} \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \psi_\theta \rangle}, \tag{A17}
\end{aligned}$$

where we use the fact that $\text{supp}(\psi_\theta) \subseteq \text{supp}_\lambda$ for any $q < 1$ since the blur kernel retains the original configuration with nonzero probability, which implies that $\langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda = 1$ for all $\mathbf{x} \in \text{supp}(\psi_\theta)$. Consequently, the second term is unaffected by the blur and remains independent of λ . Averaging over λ , therefore, gives

$$S_{ij}^{\text{r-blur}} = \langle S_{ij}(\lambda) \rangle_\lambda = \frac{\sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \langle \partial_{\theta_i} \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} - \frac{\langle \partial_{\theta_i} \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \psi_\theta \rangle}. \tag{A18}$$

Compared with the exact QGT in Eq. (1), the deviation is

$$\Delta S_{ij} = \frac{\sum_{\mathbf{x}} \langle 1 - \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \langle \partial_{\theta_i} \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle}. \tag{A19}$$

Therefore, the randomized blurred QGT differs from the exact one only through configurations for which $\langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda < 1$.

Importantly, this deviation does not necessarily invalidate the resulting dynamics. We now show that, provided the variational *Ansatz* is sufficiently expressive, randomized blurred sampling yields the exact projected TDVP evolution, despite modifying the QGT. The key observation is that the induced modification can be interpreted as a positive deformation of the Hilbert-space metric underlying the TDVP derivation.

Recall that the TDVP equation in Eq. (2) arises from minimizing the Hilbert-space distance between the variational tangent vector and the exact time-evolution direction. In the tangent space, these directions are given by

$$\begin{aligned}
|\dot{\psi}_\theta\rangle_{\text{var}} &= \left(1 - \frac{|\psi_\theta\rangle\langle\psi_\theta|}{\langle\psi_\theta|\psi_\theta\rangle} \right) \frac{\sum_i |\partial_{\theta_i} \psi_\theta\rangle \dot{\theta}_i}{\sqrt{\langle\psi_\theta|\psi_\theta\rangle}}, \\
|\dot{\psi}_\theta\rangle_{\text{exact}} &= \left(1 - \frac{|\psi_\theta\rangle\langle\psi_\theta|}{\langle\psi_\theta|\psi_\theta\rangle} \right) \frac{-\xi \hat{H} |\psi_\theta\rangle}{\sqrt{\langle\psi_\theta|\psi_\theta\rangle}}. \tag{A20}
\end{aligned}$$

The TDVP equation follows from minimizing the distance between these two tangent vectors:

$$\mathcal{L} = \| |\dot{\psi}_\theta\rangle_{\text{var}} - |\dot{\psi}_\theta\rangle_{\text{exact}} \|, \tag{A21}$$

where the distance is defined through the squared Hilbert-space norm $\| \cdot \| \equiv \sum_{\mathbf{x}} | \cdot(\mathbf{x}) |^2$.

If we replace the standard squared Hilbert-space norm by the weighted squared norm $\| \cdot \|' \equiv \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda | \cdot(\mathbf{x}) |^2$, the resulting TDVP equation reproduces precisely the QGT obtained from randomized blurred sampling. To see this, consider the corresponding distance

$$\begin{aligned}
\mathcal{L}' &= \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \left| \langle \mathbf{x} | \dot{\psi}_\theta \rangle_{\text{var}} - \langle \mathbf{x} | \dot{\psi}_\theta \rangle_{\text{exact}} \right|^2 \\
&= \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \left| \xi \left(\frac{\langle \mathbf{x} | \hat{H} | \psi_\theta \rangle}{\sqrt{\langle \psi_\theta | \psi_\theta \rangle}} - \frac{\langle \mathbf{x} | \psi_\theta \rangle \langle \psi_\theta | \hat{H} | \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle^{3/2}} \right) + \sum_i \left(\frac{\langle \mathbf{x} | \partial_{\theta_i} \psi_\theta \rangle}{\sqrt{\langle \psi_\theta | \psi_\theta \rangle}} - \frac{\langle \mathbf{x} | \psi_\theta \rangle \langle \psi_\theta | \partial_{\theta_i} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle^{3/2}} \right) \dot{\theta}_i \right|^2. \tag{A22}
\end{aligned}$$

Let $A_i(\mathbf{x}) \equiv (\langle \mathbf{x} | \partial_{\theta_i} \psi_\theta \rangle / \sqrt{\langle \psi_\theta | \psi_\theta \rangle}) - (\langle \mathbf{x} | \psi_\theta \rangle \langle \psi_\theta | \partial_{\theta_i} \psi_\theta \rangle / \langle \psi_\theta | \psi_\theta \rangle^{3/2})$ and $B(\mathbf{x}) \equiv [(\langle \mathbf{x} | \hat{H} | \psi_\theta \rangle / \sqrt{\langle \psi_\theta | \psi_\theta \rangle}) - (\langle \mathbf{x} | \psi_\theta \rangle \langle \psi_\theta | \hat{H} | \psi_\theta \rangle / \langle \psi_\theta | \psi_\theta \rangle^{3/2})]$; we have

$$\mathcal{L}' = \dot{\boldsymbol{\theta}}^\dagger \mathbf{S}' \dot{\boldsymbol{\theta}} + \xi \dot{\boldsymbol{\theta}}^\dagger \mathbf{F}' + \xi^* \mathbf{F}'^\dagger \dot{\boldsymbol{\theta}} + \text{const}, \tag{A23}$$

where $S'_{ij} = \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda A_i^*(\mathbf{x}) A_j(\mathbf{x})$ and $F'_i = \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda A_i^*(\mathbf{x}) B(\mathbf{x})$. Clearly, minimization of $\mathcal{L}'$ leads to the TDVP equation with $\mathbf{S}'$ as QGT and $\mathbf{F}'$ as variational force. Therefore, we have

$$S'_{ij} = \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda A_i^*(\mathbf{x}) A_j(\mathbf{x}) \tag{A24}$$

$$\begin{aligned}
&= \frac{\sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \langle \partial_{\theta_i} \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle} - \frac{\langle \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle^2} \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \langle \partial_{\theta_i} \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \psi_\theta \rangle \\
&\quad - \frac{\langle \partial_{\theta_i} \psi_\theta | \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle^2} \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda \langle \psi_\theta | \mathbf{x} \rangle \langle \mathbf{x} | \partial_{\theta_j} \psi_\theta \rangle + \frac{\langle \partial_{\theta_i} \psi_\theta | \psi_\theta \rangle \langle \psi_\theta | \partial_{\theta_j} \psi_\theta \rangle}{\langle \psi_\theta | \psi_\theta \rangle^2} \sum_{\mathbf{x}} \langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda |\langle \mathbf{x} | \psi_\theta \rangle|^2.
\end{aligned} \tag{A25}$$

Using the fact that $\langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda = 1$ for all $\mathbf{x} \in \text{supp}(\psi_\theta)$, the factors $\langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda$ do not affect any term containing $\langle \mathbf{x} | \psi_\theta \rangle$. The last three terms reduce to their standard Hilbert-space expressions, and S'_{ij} coincides exactly with the randomized blurred QGT $S_{ij}^{\text{r-blur}}$ in Eq. (A18).

Randomized blurred sampling can, therefore, be viewed as a deformation of the Hilbert-space metric underlying the TDVP. By construction, $\langle \mathcal{I}_\lambda(\mathbf{x}) \rangle_\lambda > 0$ for all $\mathbf{x}$, so the weighted norm remains strictly positive definite. Consequently,

$$\| |\psi_\theta\rangle_{\text{var}} - |\psi_\theta\rangle_{\text{exact}} \| = 0 \Leftrightarrow \| |\psi_\theta\rangle_{\text{var}} - |\psi_\theta\rangle_{\text{exact}} \|' = 0. \tag{A26}$$

Therefore, if the variational *Ansatz* is sufficiently expressive—namely, if there exists a parameter velocity $\dot{\theta}^*$ such that $\| |\psi_\theta\rangle_{\text{var}} - |\psi_\theta\rangle_{\text{exact}} \| = 0$ —then the same $\dot{\theta}^*$ also solves the TDVP equation defined with the modified norm, $\| |\psi_\theta\rangle_{\text{var}} - |\psi_\theta\rangle_{\text{exact}} \|' = 0$ and vice versa. Hence, randomized blurred sampling yields the correct parameter dynamics whenever the variational *Ansatz* is sufficiently expressive.

In contrast, under support mismatch the standard VMC QGT corresponds to a deformation of the Hilbert-space metric with vanishing weights on configurations that are not sampled. The resulting metric, therefore, develops null directions and becomes degenerate, leading to potentially spurious dynamics even when the variational *Ansatz* is sufficiently expressive. Randomized blurred sampling

removes this degeneracy. Although the resulting QGT differs from the exact one, the induced metric remains strictly positive definite and spans the correct tangent directions. In this sense, randomized blurred sampling acts as a natural regularization of the QGT, restoring stability while preserving the correct projected dynamics. A demonstration in the spin relaxation problem is presented in Appendix B 5.

APPENDIX B: SUPPLEMENTARY RESULTS

In this section, we present additional numerical results supporting the main text. We examine several estimator-level diagnostics, including the variance, signal-to-noise ratio, and effective sample size, to illustrate how blurred sampling addresses nodal and support-mismatch pathologies and remains robust over a broad range of q . We also compare blurred sampling with overdispersed sampling, showing that the latter does not resolve the support-mismatch pathology and can suffer from an exponentially small effective sample size. Finally, we provide additional diagnostics for representative dynamical failures, which help identify cases where biased force or QGT estimates become important.

1. Variance and ESS diagnostics for electron-gas system

We quantify the effect of the blur kernel by varying the blur strength q and measuring the variance of the energy-variance estimator in the electron-gas system. As shown in Fig. 8, the variance is large at $q = 0$, where standard sampling suffers from the nodal singularity, but drops

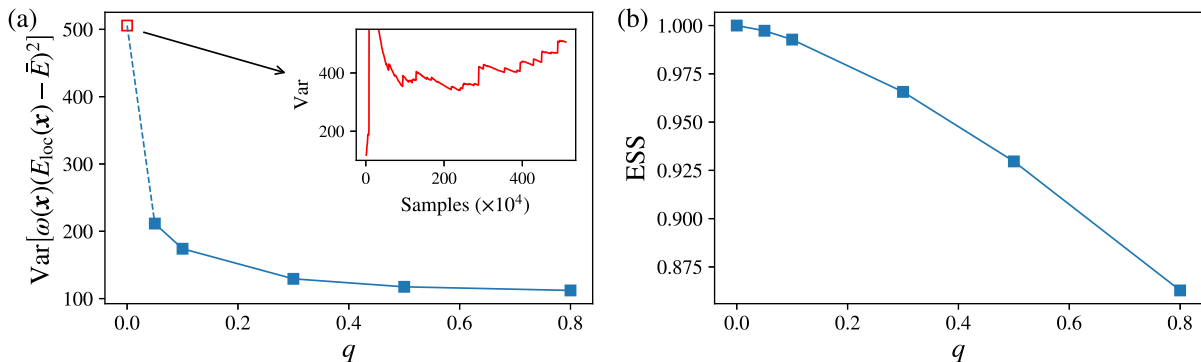


FIG. 8. (a) Variance of the energy-variance estimator as a function of the blur strength q . Once q is turned on, the variance of $\omega(\mathbf{x})(E_{\text{loc}}(\mathbf{x}) - \bar{E})^2$ drops rapidly. The variance is estimated using 512×10^4 samples. At $q = 0$, the variance is expected to diverge theoretically. The inset shows that the estimated variance increases with the number of samples, and the large but finite value reported here reflects the finite sample size. (b) Effective sample size as a function of q . The ESS decreases as q increases. In these results, we use the same state as in Fig. 4(c) and $\epsilon = r_s$.

sharply once blurred sampling is introduced. This is consistent with the theoretical expectation that the standard estimator has divergent variance; the large but finite value observed in Fig. 8 is a finite-sample remnant of this divergence. Any nonzero blur regularizes the singularity and restores ordinary central-limit-theorem scaling. The variance then remains stable over a broad range of q , indicating that the method does not require fine-tuning of the blur strength. Over the same range, the ESS remains high, showing that the variance reduction is not offset by large reweighting fluctuations.

2. Signal-to-noise and ESS diagnostics for parity-mixing problem

As an additional diagnostic for the parity-mixing problem in Fig. 6, we examine how the signal-to-noise ratio (SNR) of the force estimator depends on the blur strength q . We focus on the variational state at $t = 0.2$, where the parity is $\hat{P}_Z \approx -1$ and support-mismatch effects are expected to be pronounced. For each force component, we estimate

$$\text{SNR}(\hat{F}_i) = \frac{\langle \hat{F}_i(\mathbf{x}) \rangle_r^2}{\langle \hat{F}_i(\mathbf{x})^2 \rangle_r - \langle \hat{F}_i(\mathbf{x}) \rangle_r^2},$$

as a function of q with $\hat{F}_i(\mathbf{x}) \equiv \omega(\mathbf{x})[\mathcal{O}_i^*(\mathbf{x}) - \bar{\mathcal{O}}_i^*] \times [E_{\text{loc}}(\mathbf{x}) - \bar{E}]$. As shown in Fig. 9(a), the SNR is strongly suppressed at $q = 0$, reflecting the pathological variance of the standard estimator. With blurred sampling, the SNR increases rapidly and remains stable over an intermediate range of q , approximately between $q = 0.1$ and $q = 0.5$. For larger q , the SNR decreases again because the importance weights fluctuate more strongly, consistent with the reduction of the ESS shown in Fig. 9(b). Nevertheless, for all values of q , the ESS remains lower bounded by $1 - q$, in agreement with Eq. (A14). These results suggest that

adaptively choosing q to target an $\mathcal{O}(1)$ ESS could be a promising strategy.

3. Comparison with overdispersed sampling

Here, we investigate the effectiveness of overdispersed sampling method [63,64], a simple importance sampling method that replace the Born distribution by an overdispersed reference distribution, $r_\alpha(\mathbf{x}) \propto |\psi_\theta(\mathbf{x})|^\alpha$, $0 \leq \alpha < 2$, with the standard VMC method recovered at $\alpha = 2$. For simplicity, we omit the normalization factors in the following discussion. The corresponding reweighting factor is, thus, $\omega_\alpha(\mathbf{x}) = p(\mathbf{x})/r_\alpha(\mathbf{x}) = |\psi_\theta(\mathbf{x})|^{2-\alpha}$.

Since r_α vanishes more slowly than $p(\mathbf{x}) = |\psi_\theta(\mathbf{x})|^2$ near nodes, it samples configurations in the vicinity of nodal regions more frequently than conventional sampling. In this sense, overdispersed sampling is a direct importance-sampling approach: One modifies the Markov chain so that rare but statistically important configurations are visited more often and then reweights observables back to the target Born distribution. Choosing $\alpha < 2$ can reduce or remove the infinite-variance pathology by assigning more sampling weight to the nodal region.

However, this improvement comes with two important limitations. First, overdispersed sampling does not resolve support-mismatch bias in discrete configuration spaces, since the support of the distribution is unchanged. Second, the reweighting factor is a global function of the wave-function amplitude. Since $\log |\psi_\theta(\mathbf{x})|$ is typically extensive in system size, the fluctuations of $\log \omega_\alpha(\mathbf{x}) = (2 - \alpha) \log |\psi_\theta(\mathbf{x})|$ also grow with system size. Consequently, the effective sample size $\text{ESS}_\alpha = (\langle \omega_\alpha \rangle_{r_\alpha}^2 / \langle \omega_\alpha^2 \rangle_{r_\alpha})$ can decrease exponentially with system size for fixed $\alpha < 2$. For independent subsystems, this decay is strictly exponential in system size. This is the central trade-off of overdispersed sampling: Reducing nodal singularities requires moving α away from 2, but doing so increases

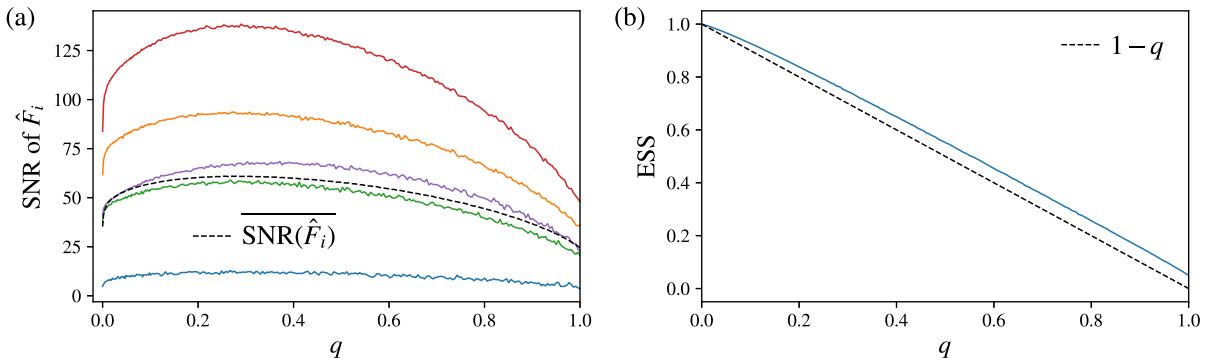


FIG. 9. (a) The SNR of variational forces $\hat{F}_i$ at time $t = 0.2$ for the spin-parity quench in Fig. 6, shown for different values of q . We show five representative components of the force vector as well as the average over all 592 parameters of the model. The SNR is significantly amplified when q is turned on and remains stable over a broad range of $q \in [0.1, 0.5]$. (b) The effective sample size of $\omega(\mathbf{x})$ as a function of q , which is always larger than $1 - q$ (the dashed line), as bounded by Eq. (A14). Both the SNR of $\hat{F}_i$ and the ESS are estimated from $N_s = 262144$ samples for each q .

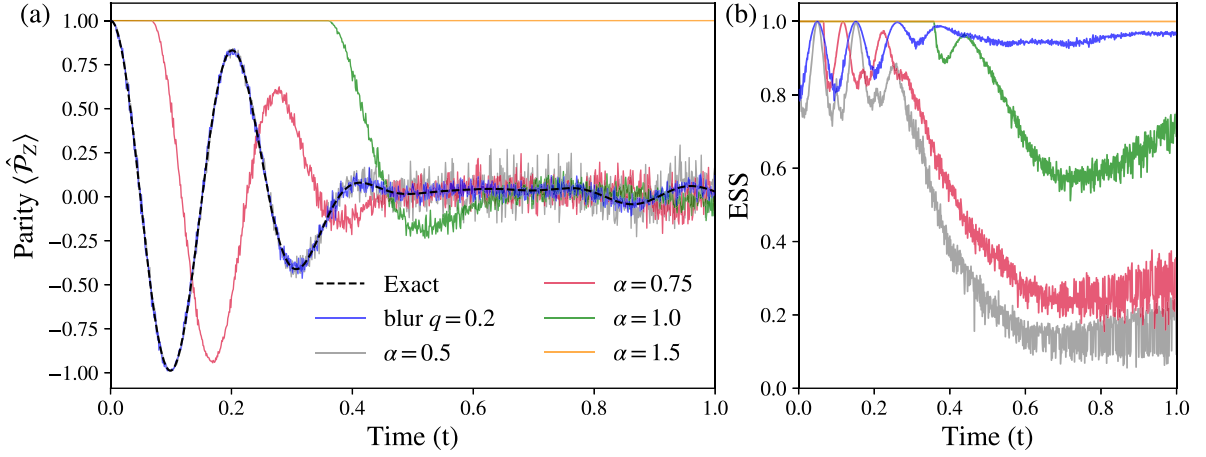


FIG. 10. Comparison of overdispersed and blurred sampling in the parity-mixing problem. (a) The Z-parity dynamics simulated using the overdispersed sampling method with $\alpha = 0.5, 0.75, 1.0, 1.5$ and the blurred sampling method with $q = 0.2$. The dashed line shows the exact solution. With $\alpha > 0.5$, the overdispersed sampling results remain frozen at the initial value, whereas the blurred sampling follows the exact solution. (b) ESS for the different methods during time evolution. The overdispersed sampling method exhibits low ESS at late times, whereas the blurred sampling method remains above the bound $1 - q = 0.8$. The simulations are performed for a 16-spin system with $h = 1$ and using the Gaussian Ansatz. We use 1024 samples and Heun integrator with $\delta t = 10^{-3}$.

the fluctuations of the reweighting factor away from the nodes.

We illustrate these problems in Fig. 10 by repeating similar simulations for the parity-mixing problem as in

Fig. 6. For overdispersed sampling with $\alpha = 1.5$, the dynamics is frozen because of the support-mismatch problem. For $\alpha = 0.75, 1.0$, the dynamics are frozen at early times and recover only after some time. This is

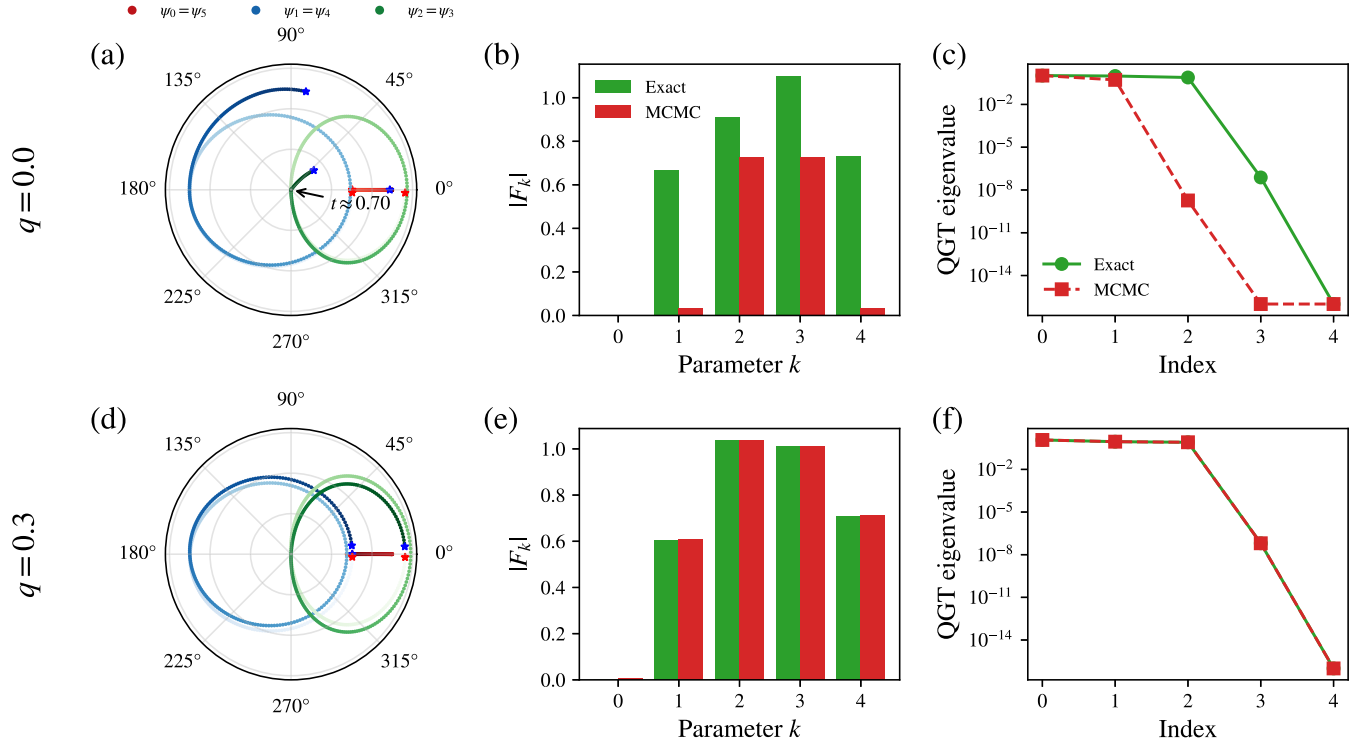


FIG. 11. Diagnostics for the 2×2 Heisenberg quench at the breakdown time $t \approx 0.70$, comparing standard t-VMC ($q = 0$, top row) with blurred sampling ($q = 0.3$, bottom row). (a),(d) Trajectories of the three independent wave-function amplitudes $\psi_0 = \psi_5$, $\psi_1 = \psi_4$, and $\psi_2 = \psi_3$ in the complex plane; red and blue stars mark the start and end of the dynamics, respectively. The dynamics repeatedly crosses the node, but around $t \approx 0.70$ the standard sampler gets stuck at the node and the dynamics diverges. (b),(e) Magnitude of the force-vector components $|F_k|$, computed exactly by full enumeration of the Hilbert space (green) and estimated from Markov chain Monte Carlo (MCMC) samples (red). (c),(f) Corresponding eigenvalue spectra of the QGT, exact (green) versus sampled (red).

because the initial wave function does not have exact zeros; since we always add small perturbations to the parameters, overdispersed sampling amplifies these components. Only when α is very small do the dynamics recover at the beginning. We also compare with blurred sampling with $q = 0.2$, which clearly gives the correct dynamics. In Fig. 10(b), we show the ESS for different methods. As α decreases, the ESS becomes smaller at late times, while blurred sampling remains lower bounded by $1 - q$.

4. Breakdown of the Heisenberg dynamics

Here, we study the breakdown of the 2×2 Heisenberg dynamics in Fig. 5(b) near $t \approx 0.7$ in more detail. In Figs. 11(a) and 11(d), we show the trajectories of the wave function amplitudes during the time evolution, after fixing the global phase by setting ψ_0 to be real and normalizing the wave function. At the breakdown point of $t \approx 0.7$, parameter ψ_2 (which is equal to ψ_3 by symmetry) crosses zero in the complex plane, leading to a collapse of both the force and QGT estimators as shown in Figs. 11(b) and 11(c). Blurred sampling resolves this problem, as illustrated in Figs. 11(e) and 11(f).

5. QGT bias and support of the forces in spin-relaxation problem

In this section, we argue why the Hamiltonian-induced blur sampling in Fig. 7(a) still produces the correct dynamics despite the QGT support-mismatch problem and why it fails in Fig. 7(b). In Fig. 12, we calculate the eigenvalue spectrum of the QGT and investigate the support of the force vector in the corresponding eigenbasis. We see in Fig. 12(a) that, without correlation in the *Ansatz* (first row), the Hamiltonian-induced blur ($q_2 = 0$, second column) produces a QGT with extremely limited support, compared to the randomized blur ($q_2 = 0.2$, first column). This should result in a divergence in the dynamics from the correct result, but we see in Fig. 7(a) that this is not the case. The reason for this can be found in Fig. 12(b). We see that both blurred sampling schemes produce unbiased force vectors that are strongly peaked on a single mode for $t = 0$ and $t = 0.01$, with the support on other modes having $|\rho_k| < 10^{-8}$. Because of this lack of support, the bias in the QGT does not affect the final force vector. However, when we look at the same quantities for the correlated *Ansatz* (second row), a different picture emerges. For $t = 0.01$, we

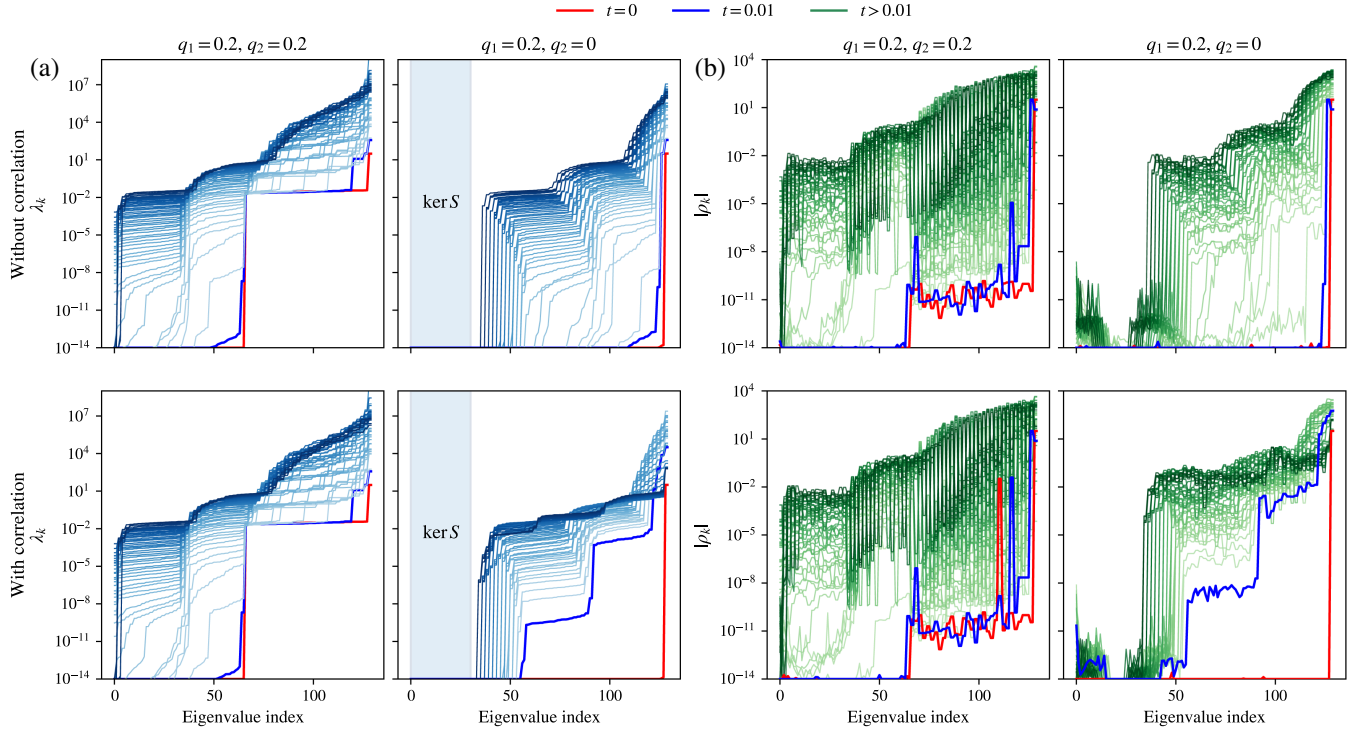


FIG. 12. Comparison of the support of the QGT and the force vectors for the t-VMC dynamics of Fig. 7 under blurred sampling. The first row corresponds to data for variational *Ansatz* without correlation, whereas the second row corresponds to a variational *Ansatz* with correlation between the spin sectors. (a) We diagonalize the QGT $S = V\Lambda V^\dagger$ and show the sorted eigenvalues λ_k over the duration of the dynamics, with lighter colors indicating earlier times $t \leq 0.5$. We highlight the times $t = 0$ and $t = 0.01$ with red and blue lines, respectively. Note that, for the Hamiltonian-induced blur ($q_2 = 0$, second column), the QGT has a large kernel for the total duration of the dynamics. (b) We transform the force vector to the eigenbasis of the QGT: $\rho = V^\dagger F$ and show the components ρ_k for the four cases studied here. The components are not sorted in magnitude but correspond directly to the ordering of the eigenvalues in (a).

see that the force vector has significant support on a large number of eigenmodes, which now get amplified incorrectly due to the QGT bias. The consequences are immediate: The dynamics in Fig. 7(b) diverges around this point.

APPENDIX C: DETAILS OF GROUND-STATE OPTIMIZATION RESULTS

1. Numerical details of the two-dimensional electron gas

The wave-function *Ansatz* used for the two-dimensional electron gas benchmark is the (MP)²NQS *Ansatz* [40], a neural backflow wave function. For a configuration $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)$, the trial wave function is written as

$$\Psi(\mathbf{R}) = \det[\phi_\alpha(\tilde{\mathbf{r}}_i)]_{\alpha,i=1}^N \times \exp\left(\mathcal{U}_{\text{NN}}(\{\tilde{\mathbf{r}}_i\}) - \sum_{i<j} u(r_{ij})\right), \quad (\text{C1})$$

where ϕ_α are the single-particle orbitals, $\tilde{\mathbf{r}}_i$ are the backflow-transformed coordinates, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, and $u(r)$ is the Jastrow factor, a symmetric two-body correlation function that captures short-range correlation effects and enforces the correct cusp structure in the wave function, which we parametrize using the form $u(r) = [(2c_{\text{cusp}}j_d^2e^{-j_br})/1 + 2j_bj_d][(1 - e^{-r/j_d})/r]$, where j_b and j_d are variational parameters and $c_{\text{cusp}} = 1/3$ sets the electron-electron cusp strength [94]. The $\mathcal{U}_{\text{NN}}(\{\tilde{\mathbf{r}}_i\})$ is the neural-network Jastrow factor. The single-particle orbitals ϕ_α are expanded in a plane-wave basis as

$$\phi_\alpha(\mathbf{r}) = \sum_{\mathbf{k} \in \mathcal{K}} A_{\alpha,\mathbf{k}} \cos(\mathbf{k} \cdot \mathbf{r}) + \sum_{\mathbf{k} \in \mathcal{K}} B_{\alpha,\mathbf{k}} \sin(\mathbf{k} \cdot \mathbf{r}). \quad (\text{C2})$$

Here, $A_{\alpha,\mathbf{k}}$ and $B_{\alpha,\mathbf{k}}$ are real variational parameters. The wave vectors $\mathbf{k}$ are chosen from a truncated closed-shell momentum set $\mathcal{K}$, with $|\mathcal{K}| = 97$, close to $2N$. Since the orbitals are represented in a real sine-cosine basis, the sum over $\mathbf{k}$ is taken over only one representative from each $\pm\mathbf{k}$ pair, avoiding double counting.

The backflow coordinates are defined by $\tilde{\mathbf{r}}_i = \mathbf{r}_i + \xi_i(\mathbf{R})$, where the displacement field $\xi_i(\mathbf{R})$ is produced by a message-passing neural network that depends symmetrically on the full electronic configuration. The neural network follows the same construction as Ref. [40], with one- and two-body streams h_i and h_{ij} , periodic pair features. In our calculations, we use a setup with $d_1 = 16$ (one-body channels), $d_2 = 8$ (two-body channels), one message-passing layer (MLP), MLP width 16, attention width 16, and Gaussian error linear unit (GELU) activations. The neural Jastrow term uses one hidden layer with width 16.

In the simulations, we use 512 Monte Carlo samples per optimization step. The learning rate is initialized at 10^{-3} and decayed as $\text{lr}/(1 + \text{step}/1000)$. The wave function is optimized using the SR method, with a diagonal

regularization λI to the S -matrix where $\lambda = 10^{-3}$. For Monte Carlo sampling, we use local Metropolis updates. Between two consecutive optimization steps, we perform 20 Metropolis steps to reduce autocorrelation. The proposal step size is adjusted adaptively during optimization to keep the harmonic average of acceptance rate close to 0.5. The simulations with and without blurred sampling are initialized from the same state and use identical hyperparameters and sampling settings. The only difference is that, in the blurred-sampling runs, the blurred-sampling postprocessing step is applied after every block of 20 local Metropolis updates. The additional computational overhead of blurred sampling is modest. On four NVIDIA A100 GPUs, the standard calculation takes approximately 0.57 s per optimization iteration, while the blurred-sampling calculation takes approximately 0.76 s per iteration. This corresponds to a roughly 33% increase in wall time, consistent with the expected constant-factor overhead from the postprocessing step.

2. Numerical details of the J_1 - J_2 model

We parametrize the variational wave function $\log \psi_\theta(\mathbf{x})$ of a spin configuration $\mathbf{x}$ on an $L \times L$ lattice using a patched Vision Transformer architecture, following Refs. [95,96], together with the factored attention mechanism introduced in Refs. [97,98]. The lattice, with $N = L^2$ sites, is partitioned into

$$n = \frac{N}{b^2} = \left(\frac{L}{b}\right)^2 \quad (\text{C3})$$

patches of linear size b . Each flattened patch is linearly embedded into a d -dimensional token $x_i \in \mathbb{R}^d$. The resulting sequence is processed by N_ℓ transformer blocks of the form

$$x'_i = x_i + \text{Attn}(\text{LN}(x))_i, \quad x''_i = x'_i + \text{FF}(\text{LN}(x'))_i, \quad (\text{C4})$$

where LN denotes layer normalization. The feedforward block FF is a two-layer MLP with intermediate width $4d$, mapping $d \rightarrow 4d \rightarrow d$, and using a GELU nonlinearity.

In the factored-attention layer, the usual query-key mechanism is replaced by a learned, translation-invariant mixing kernel. For each head $a = 1, \dots, H$, the values

$$v_i^{(a)} = W_v^{(a)} x_i \quad (\text{C5})$$

are mixed according to

$$o_i^{(a)} = \sum_{j=1}^n J_{i-j}^{(a)} v_j^{(a)}. \quad (\text{C6})$$

Here, $i - j$ denotes the relative displacement on the periodic patch lattice. We parametrize the kernel as

$$J_{\mathbf{r}}^{(a)} = p_{\mathbf{r}}^{(a)} \text{softmax}_{\mathbf{r}}(-\alpha_a d_{\mathbf{r}}), \quad (\text{C7})$$

where $p_{\mathbf{r}}^{(a)}$ is a learned per-head kernel amplitude, $d_{\mathbf{r}}$ is the periodic Euclidean distance between patches, and α_a is a learned per-head length-scale parameter initialized at $\alpha_a = 1$. This parametrization biases the initial kernel toward nearby patches while retaining fully learnable translation-invariant weights. The outputs from all heads are concatenated and projected back to dimension d .

After the final transformer block, the tokens are summed and normalized:

$$z = \text{LN}\left(\sum_{i=1}^n x_i^{(\text{final})}\right), \quad (\text{C8})$$

and two independent dense projections produce real and imaginary features:

$$y_{\text{R}} = \text{LN}(W_{\text{R}}z), \quad y_{\text{I}} = \text{LN}(W_{\text{I}}z). \quad (\text{C9})$$

The complex log-amplitude is then given by

$$\log \psi(\sigma) = \sum_k \log \cosh(y_{\text{R},k} + i y_{\text{I},k}). \quad (\text{C10})$$

For the optimization shown in Fig. 4(d), we use patch size $b = 3$, $N_{\ell} = 2$ transformer layers, $H = 12$ factored-attention heads, and embedding dimension $d = 72$. For the system size considered here, this gives a variational *Ansatz* with 117048 trainable parameters.

During optimization, we regularize the stochastic-reconfiguration linear system by shifting the S -matrix estimator in Eq. (3) by λI , with $\lambda = 10^{-3}$. The learning rate is annealed with a cosine-decay schedule from $\eta = 10^{-2}$ to $\eta = 10^{-3}$ over 10^4 optimization steps.

APPENDIX D: DETAILS OF THE PARITY-MIXING PROBLEM IN TFIM

We study the dynamics of the initial state chosen as the equal-weight superposition of computational-basis states with even Z parity:

$$|\psi_0\rangle = \sum_{\mathbf{x}} \frac{\hat{P}_Z + 1}{2} |\mathbf{x}\rangle, \quad (\text{D1})$$

where $\hat{P}_Z \equiv \prod_{i=1}^N Z_i$ denotes the Z -parity operator and $|\mathbf{x}\rangle$ are computational-basis states with $x_i \in \{\pm 1\}$. The time evolution is governed by the TFIM Hamiltonian

$$\hat{H}_{\text{TFIM}} = J \sum_{i=1}^N Z_i Z_{i+1} + h \sum_{i=1}^N X_i, \quad (\text{D2})$$

where periodic boundary conditions are assumed. In this setup, we are interested in how the Z parity evolves under

the dynamics, characterized by

$$\mathcal{P}_Z(t) = \frac{\langle \psi_0 | e^{i\hat{H}_{\text{TFIM}}t} \hat{P}_Z e^{-i\hat{H}_{\text{TFIM}}t} | \psi_0 \rangle}{\langle \psi_0 | \psi_0 \rangle}. \quad (\text{D3})$$

In the following, we describe the variational *Ansätze* employed—including the RBM and Gaussian *Ansatz*—and present the analytic solution.

1. RBM *Ansatz* and its initialization

We employ a RBM with complex-valued parameters as the variational *Ansatz*. The wave function is given by

$$\psi_{\text{RBM}}(\mathbf{x}) = \exp\left(\sum_{i=1}^N a_i x_i\right) \prod_{j=1}^M 2 \cosh\left(b_j + \sum_{i=1}^N W_{ij} x_i\right), \quad (\text{D4})$$

where a_i , b_j , and W_{ij} are complex parameters and M denotes the number of hidden units. We define the *Ansatz* size as $\alpha \equiv M/N$. The initialization is particularly simple. The action of the parity projector on a computational-basis state can be written as

$$\frac{\hat{P}_Z + 1}{2} |\mathbf{x}\rangle = \cos\left(\frac{\pi}{4} \left(N - \sum_{i=1}^N x_i\right)\right)^2 |\mathbf{x}\rangle. \quad (\text{D5})$$

Choosing the RBM parameters as $a_i = 0$, $b_1 = b_2 = i\pi N/4$, and $W_{i1} = W_{i2} = -i(\pi/4)$ (with all other parameters set to zero), the RBM wave function exactly reproduces these amplitudes, thereby initializing the variational state to the chosen parity sector. For numerical stability, we add a small perturbation of the order of 10^{-3} to the initial parameters, which ensures that the logarithmic derivatives remain well defined.

2. Exact solution via mapping to Majorana fermions

It is well known that the one-dimensional TFIM can be mapped to a system of free fermions. Here, we follow the mapping of Ref. [92] and directly introduce $2N$ Majorana fermions γ_{2i} and γ_{2i+1} . The spin operators are represented as

$$Z_{i+1} = (-i)^i \gamma_0 \gamma_1 \gamma_2 \cdots \gamma_{2i}, \quad X_{i+1} = -i \gamma_{2i} \gamma_{2i+1}. \quad (\text{D6})$$

One then obtains

$$Z_i Z_{i+1} = \begin{cases} -i \gamma_{2i-1} \gamma_{2i}, & \text{for } i < N, \\ i \hat{P}_X \gamma_{2N-1} \gamma_0, & \text{for } i = N, \end{cases} \quad (\text{D7})$$

where $\hat{P}_X = \prod X_i$ is the conserved X -parity operator. For even system size N , the initial state $|\psi_0\rangle$ lies in the sector $\hat{P}_X = 1$. Restricting to this sector, the Hamiltonian becomes

$$\hat{H}_{\text{TFIM}} = -iJ \sum_{i=0}^{N-2} \gamma_{2i+1} \gamma_{2i+2} + iJ \gamma_{2N-1} \gamma_0 - i h \sum_i \gamma_{2i} \gamma_{2i+1}. \quad (\text{D8})$$

In this representation, the Z-parity operator reads

$$\mathcal{P} = (-i)^{N/2} \prod_{i=0}^{N/2-1} \gamma_{4i+1} \gamma_{4i+2}. \quad (\text{D9})$$

Although the Hamiltonian is quadratic in Majorana operators, so that the time-evolution operator is a fermionic Gaussian operator (a match-gate circuit), the initial state $|\psi_0\rangle$ is not Gaussian. We, therefore, exploit the symmetry of the Z-parity operator. Using the anticommutation relation $\{X_i, \hat{\mathcal{P}}_Z\} = 0$, which implies

$$\hat{\mathcal{P}}_Z \hat{H}_{\text{TFIM}}(J, h) = \hat{H}_{\text{TFIM}}(J, -h) \hat{\mathcal{P}}_Z, \quad (\text{D10})$$

we obtain

$$\begin{aligned} |\psi(t)\rangle &\equiv e^{-i\hat{H}_{\text{TFIM}}(J, h)t} |\psi_0\rangle \\ &= e^{-i\hat{H}_{\text{TFIM}}(J, h)t} \frac{\hat{\mathcal{P}}_Z + 1}{\sqrt{2}} |+\rangle \\ &= \frac{1}{\sqrt{2}} e^{-i\hat{H}_{\text{TFIM}}(J, h)t} |+\rangle + \frac{\hat{\mathcal{P}}_Z}{\sqrt{2}} e^{-i\hat{H}_{\text{TFIM}}(J, -h)t} |+\rangle. \end{aligned} \quad (\text{D11})$$

Since the state $|+\rangle$ is Gaussian—being the ground state of the quadratic Hamiltonian $-\sum_i X_i = \sum_i i\gamma_{2i} \gamma_{2i+1}$ —its time evolution under the quadratic TFIM Hamiltonian remains Gaussian. We denote $e^{-i\hat{H}_{\text{TFIM}}(J, h)t} |+\rangle$ as $|\psi_G^+(J, h, t)\rangle$. The full wave function can then be written as

$$|\psi(t)\rangle = \frac{1}{\sqrt{2}} |\psi_G^+(J, h, t)\rangle + \frac{\hat{\mathcal{P}}_Z}{\sqrt{2}} |\psi_G^+(J, -h, t)\rangle. \quad (\text{D12})$$

The time-dependent Z parity becomes

$$\begin{aligned} \mathcal{P}_Z(t) &= \text{Re} \langle \psi_G^+(J, -h, t) | \psi_G^+(J, h, t) \rangle \\ &\quad + \frac{1}{2} \langle \psi_G^+(J, h, t) | \hat{\mathcal{P}}_Z | \psi_G^+(J, h, t) \rangle \\ &\quad + \frac{1}{2} \langle \psi_G^+(J, -h, t) | \hat{\mathcal{P}}_Z | \psi_G^+(J, -h, t) \rangle. \end{aligned} \quad (\text{D13})$$

The problem, thus, reduces to evaluating (i) the overlap between two Gaussian states and (ii) the expectation value of $\hat{\mathcal{P}}_Z$ in a Gaussian state. The overlap of two general fermionic Gaussian states [without U(1) symmetry] can be computed using Pfaffian techniques; we refer the interested reader to Ref. [91] for details. The operator $\hat{\mathcal{P}}_Z$ is a product of Majorana operators [cf. Eq. (D9)], and its expectation value

in a Gaussian state can be expressed as a Pfaffian of the corresponding covariance matrix $G_{i,j} \equiv \langle \gamma_i \gamma_j \rangle$ as

$$\left\langle \prod_i \gamma_{\alpha_i} \right\rangle = \text{pf}[G_{\alpha_i, \alpha_j}]. \quad (\text{D14})$$

Therefore, the time evolution of the Z parity can be computed analytically for system sizes far beyond the reach of exact diagonalization.

3. Gaussian Ansatz and its initialization

As shown in the main text, the RBM *Ansatz* combined with blurred sampling performs well up to system sizes $N \sim 8$, where we employ a relatively large network with $\alpha = 4$. For larger systems, however, the expressiveness of the *Ansatz* becomes the primary bottleneck of the t-VMC approach. To disentangle this limitation from other algorithmic aspects of t-VMC, we consider an alternative Gaussian *Ansatz*. This *Ansatz* permits efficient evaluation of wave-function amplitudes and is capable of representing the exact time-evolved state $|\psi(t)\rangle$ at all times:

$$|\psi_{\text{Gaussian}}\rangle = |G^1\rangle + \hat{\mathcal{P}}_Z |G^2\rangle, \quad (\text{D15})$$

where $|G^1\rangle$ and $|G^2\rangle$ are Gaussian states parametrized as

$$|G^{1,2}\rangle = e^{\theta_{ij}^{1,2} \gamma_i \gamma_j} |+\rangle, \quad (\text{D16})$$

with parameters $\theta^{1,2}$ being $2N \times 2N$ antisymmetric real matrices, each containing $N(2N-1)$ independent parameters.

The remaining task is to evaluate the amplitudes in the computational basis:

$$\langle \mathbf{x} | \psi_{\text{Gaussian}} \rangle = \langle \mathbf{x} | G^1 \rangle + \hat{\mathcal{P}}_Z \langle \mathbf{x} | G^2 \rangle. \quad (\text{D17})$$

However, the basis state $|\mathbf{x}\rangle$ is not Gaussian. Importantly, the wave function should preserve the X parity, so that $\langle \mathbf{x} | \psi_{\text{Gaussian}} \rangle$ should be the same as $\langle \hat{\mathcal{P}}_X \mathbf{x} | \psi_{\text{Gaussian}} \rangle$. Consequently, we may work in the symmetric combination

$$|\mathbf{x}\rangle_+ \equiv \frac{1}{\sqrt{2}} |\mathbf{x}\rangle + \frac{1}{\sqrt{2}} |\hat{\mathcal{P}}_X \mathbf{x}\rangle, \quad (\text{D18})$$

which is a Gaussian state. In fact, $|\mathbf{x}\rangle_+$ is the ground state of the quadratic Majorana Hamiltonian

$$i \sum_{i=0}^{N-2} x_{i+1} x_{i+2} \gamma_{2i+1} \gamma_{2i+2} - i x_N x_1 \gamma_{2N-1} \gamma_0. \quad (\text{D19})$$

This observation allows us to reformulate the amplitude evaluation entirely in terms of overlaps between Gaussian states, thereby enabling an efficient implementation within the t-VMC framework. For numerical stability at

initialization, we add a small perturbation of the order of 10^{-3} to the parameter $\theta_{ij}^{1,2} = 0$ to ensure that the logarithmic derivatives remain well defined.

APPENDIX E: DETAILS OF THE SPIN-RELAXATION PROBLEM IN THE TFIM

We study the real-time dynamics of the total Z magnetization $\hat{Z} = \sum_{i=0}^{N-1} Z_i$, starting from a computational-basis product state $|\mathbf{x}_0\rangle$ under the TFIM Hamiltonian. The time-dependent expectation value is

$$\mathcal{Z}(t) = \langle \mathbf{x}_0 | e^{i\hat{H}_{\text{TFIM}}(J,h)t} \hat{Z} e^{-i\hat{H}_{\text{TFIM}}(J,h)t} | \mathbf{x}_0 \rangle. \quad (\text{E1})$$

1. Exact solution via mapping to Majorana fermions

Using the anticommutation relation $\{X_i, Z_i\} = 0$, one obtains $Z_i e^{-i\hat{H}_{\text{TFIM}}(J,h)t} = e^{-i(\hat{H}_{\text{TFIM}}(J,h)-2hX_i)t} Z_i$, which allows us to rewrite

$$\mathcal{Z}(t) = \sum_i \langle \mathbf{x}_0 | e^{i\hat{H}(J,h)t} e^{-i(\hat{H}(J,h)-2hX_i)t} Z_i | \mathbf{x}_0 \rangle. \quad (\text{E2})$$

Since $|\mathbf{x}_0\rangle$ is an eigenstate of every Z_i with eigenvalue $\mathbf{x}_0(i) \in \{\pm 1\}$, we obtain

$$\mathcal{Z}(t) = \sum_i \mathbf{x}_0(i) \langle \mathbf{x}_0 | e^{i\hat{H}(J,h)t} e^{-i(\hat{H}(J,h)-2hX_i)t} | \mathbf{x}_0 \rangle. \quad (\text{E3})$$

To proceed, we decompose the initial state into definite X -parity sectors, $|\mathbf{x}_0\rangle = (1/\sqrt{2})(|\mathbf{x}_0^+\rangle + |\mathbf{x}_0^-\rangle)$, which yields

$$\begin{aligned} \mathcal{Z}(t) = & \frac{1}{2} \sum_i \mathbf{x}_0(i) (\langle \mathbf{x}_0^+ | e^{i\hat{H}(J,h)t} e^{-i(\hat{H}(J,h)-2hX_i)t} | \mathbf{x}_0^+ \rangle \\ & + \langle \mathbf{x}_0^- | e^{i\hat{H}(J,h)t} e^{-i(\hat{H}(J,h)-2hX_i)t} | \mathbf{x}_0^- \rangle). \end{aligned} \quad (\text{E4})$$

In this form, each contribution has a well-defined X parity and can, therefore, be evaluated efficiently using the Majorana fermionic representation of the TFIM.

2. Dressed RBM Ansatz and its initialization

In the simulations, we employ a complex RBM with translation symmetry, denoted by $\psi_{\text{symm-RBM}}(\mathbf{x})$ (see the definition in NETKET [89]). For this symmetric RBM, we fix the visible bias parameter to zero and use a hidden-unit density $\alpha = 1$. When the RBM parameters are initialized close to zero (approximately 10^{-3}), the resulting wave function $\psi_{\text{symm-RBM}}(\mathbf{x})$ approximates a uniform superposition over configurations. However, this *Ansatz* alone cannot represent the desired initial state, which has support only on the fully polarized all-down configuration.

To incorporate this structure, we augment the symmetric RBM with a sector-dependent factor $f_\phi(\mathbf{x})$, parametrized by $n+1$ complex parameters $\phi \in \mathbb{C}^{n+1}$:

$$\psi_\theta(\mathbf{x}) = f_\phi(\mathbf{x}) \psi_{\text{symm-RBM}}(\mathbf{x}), \quad f_\phi(\mathbf{x}) = \phi_{s(\mathbf{x})}. \quad (\text{E5})$$

Here, $s(\mathbf{x}) \equiv (\sum_i x_i + n)/2$ denotes the spin sector of configuration $\mathbf{x}$. This construction assigns an independent complex amplitude to each spin sector, allowing the *Ansatz* to selectively enhance or suppress entire sectors. In particular, setting $\phi_0 = 1$ and $\phi_s = 0$ for $s \neq 0$ yields a wave function supported entirely on the fully polarized all-down configuration $\mathbf{x} = (-1, -1, \dots, -1)$, corresponding to the $s = 0$ sector. In practice, to avoid exact zeros that may lead to numerical instabilities, we initialize $\phi_s = \epsilon = 10^{-5} \times 2^{-n/2}$ for $s \neq 0$ while fixing $\phi_0 = 1$ throughout the simulation. Since $f_\phi(\mathbf{x})$ is independent across spin sectors in this parametrization, we refer to it as the *uncorrelated* parametrization.

To further expose the bias of the quantum geometric tensor, we introduce correlations between different spin sectors in the parametrization. In particular, we couple the single-spin-up sector to another sector, thereby breaking the independence of sector amplitudes. Concretely, we define

$$f_\phi(\mathbf{x}) = \begin{cases} \phi_1 + \phi_3 & \text{if } s(\mathbf{x}) = 3, \\ \phi_{s(\mathbf{x})} & \text{otherwise,} \end{cases}$$

so that the amplitude of the $s = 3$ sector explicitly depends on ϕ_1 , introducing correlations between the $s = 1$ and $s = 3$ sectors. We refer to this as a *correlated* parametrization.

Under the TFIM Hamiltonian, the resulting state has the following support structure:

$$\begin{aligned} \text{supp}(\psi_\theta) &\approx \{\mathbf{x} \in \{\pm 1\}^n : s(\mathbf{x}) = 0\}, \\ \text{supp}(\hat{H}_{\text{TFIM}} \psi_\theta) &\approx \{\mathbf{x} \in \{\pm 1\}^n : s(\mathbf{x}) \in \{0, 1\}\}, \\ \text{supp}[\partial_{\phi_i} \psi_\theta(\mathbf{x}) \partial_{\phi_j} \psi_\theta(\mathbf{x})] &= \{\mathbf{x} \in \{\pm 1\}^n\}. \end{aligned}$$

This reveals a pronounced support-mismatch problem in both variational force and QGT estimators. In the following, we give further results to discuss the effectiveness of the blurred sampling method under both parametrizations.

APPENDIX F: NUMERICAL DETAILS OF T-VMC SIMULATIONS

The most important part of t-VMC involves the regularization of the $\mathcal{S}$ -matrix in Eq. (2) so that it can be stably inverted. For this, we use the regularizer by Schmitt [9] with the hyperparameters $\text{SNR}_{\text{cutoff}} = 2$, $\text{rcond} = 10^{-14}$ and $\text{rcond_smooth} = 10^{-10}$ for every experiment reported in this work. For sampling, we use Metropolis-Hastings MCMC with local updates depending on the Hamiltonian. In Table II, we report the hyperparameters of the simulations. We emphasize that these are always kept exactly the same for the experiments with standard sampling and blurred sampling; hence, any reported numerics differ only by the sampling scheme used. The code to reproduce all

TABLE II. Hyperparameters of t-VMC simulations.

Figure	Ansatz $ \psi_\theta\rangle$	MCMC	Integrator	Miscellaneous
Fig. 5(a)	$\alpha 0\rangle + \beta 1\rangle$	50% single flip 50% double flip $N_s = 8192$	Adaptive RK45 $\text{rtol} = 10^{-4}$ $\text{min } \delta t = 10^{-5}$ $\text{max } \delta t = 10^{-2}$	
Fig. 5(b)	RBM $\alpha = 1/4$ See Eq. (D4)	Two-site exchange $N_s = 16384$	Heun $\delta t = 10^{-3}$	
Figs. 6(a) and 6(b)	RBM $\alpha = 4$ See Eq. (D4)	50% single flip 50% double flip $N_s = 4096$	Adaptive RK45 $\text{rtol} = 10^{-4}$ $\text{min } \delta t = 10^{-5}$ $\text{max } \delta t = 10^{-2}$	
Fig. 6(c) $h = 1/8$	Gaussian Ansatz See Eq. (D17)	50% single flip 50% double flip $N_s = 8192$	Heun $\delta t = 4 \times 10^{-3}$	Used JAXMg [99] for distributed eigensolve Sampled $ \psi_\theta ^2$ in 32-bit precision [100]
Fig. 6(c) $h = 1$	Gaussian Ansatz See Eq. (D17)	50% single flip 50% double flip $N_s = 8192$	Heun $\delta t = 10^{-3}$	Used JAXMg [99] for distributed eigensolve Sampled $ \psi_\theta ^2$ in 32-bit precision [100]
Figs. 7(a) and 7(b)	Dressed RBM $\alpha = 1$ See Eq. (E5)	50% single flip 50% double flip $N_s = 4096$	Heun $\delta t = 10^{-3}$	

figures in this work can be found in Ref. [93]. Finally, we note that the simulations reported in Figs. 2, 3, 5, 6(a), 6(b), and 7 were performed on a single NVIDIA RTX 6000 taking at most an hour of wall time. For Fig. 6(c), we used four NVIDIA H100s, with simulations lasting several hours at most.

- [1] W. L. McMillan, *Ground state of liquid He⁴*, *Phys. Rev.* **138**, A442 (1965).
- [2] D. M. Ceperley and B. J. Alder, *Ground state of the electron gas by a stochastic method*, *Phys. Rev. Lett.* **45**, 566 (1980).
- [3] G. Carleo and M. Troyer, *Solving the quantum many-body problem with artificial neural networks*, *Science* **355**, 602 (2017).
- [4] K. Choo, G. Carleo, N. Regnault, and T. Neupert, *Symmetries and many-body excitations with neural-network quantum states*, *Phys. Rev. Lett.* **121**, 167204 (2018).
- [5] S. Pathak, B. Busemeyer, J. N. B. Rodrigues, and L. K. Wagner, *Excited states in variational Monte Carlo using a penalty method*, *J. Chem. Phys.* **154**, 034101 (2021).
- [6] M. T. Entwistle, Z. Schätzle, P. A. Erdman, J. Hermann, and F. Noé, *Electronic excited states in deep variational Monte Carlo*, *Nat. Commun.* **14**, 274 (2023).
- [7] H. Lange, F. Döschl, J. Carrasquilla, and A. Bohrdt, *Neural network approach to quasiparticle dispersions in doped antiferromagnets*, *Commun. Phys.* **7**, 187 (2024).
- [8] G. Carleo, F. Becca, L. Sanchez-Palencia, S. Sorella, and M. Fabrizio, *Light-cone effect and supersonic correlations in one- and two-dimensional bosonic superfluids*, *Phys. Rev. A* **89**, 031602(R) (2014).

- [9] M. Schmitt and M. Heyl, *Quantum many-body dynamics in two dimensions with artificial neural networks*, *Phys. Rev. Lett.* **125**, 100503 (2020).
- [10] I. L. Gutiérrez and C. B. Mendl, *Real time evolution with neural-network quantum states*, *Quantum* **6**, 627 (2022).
- [11] M. Schmitt, M. M. Rams, J. Dziarmaga, M. Heyl, and W. H. Zurek, *Quantum phase transition dynamics in the two-dimensional transverse-field Ising model*, *Sci. Adv.* **8**, eabl6850 (2022).
- [12] M. Schmitt and M. Heyl, *Simulating dynamics of correlated matter with neural quantum states*, *arXiv:2506.03124*.
- [13] S. Sorella, *Green function Monte Carlo with stochastic reconfiguration*, *Phys. Rev. Lett.* **80**, 4558 (1998).
- [14] N. Irikura and H. Saito, *Neural-network quantum states at finite temperature*, *Phys. Rev. Res.* **2**, 013284 (2020).
- [15] H. Xie, Z.-H. Li, H. Wang, L. Zhang, and L. Wang, *Deep variational free energy approach to dense hydrogen*, *Phys. Rev. Lett.* **131**, 126501 (2023).
- [16] J. Nys, Z. Denis, and G. Carleo, *Real-time quantum dynamics of thermal states with neural thermofields*, *Phys. Rev. B* **109**, 235120 (2024).
- [17] J. Nys and J. Carrasquilla, *Fermionic neural Gibbs states*, *arXiv:2512.04663*.
- [18] D. Luo and B. K. Clark, *Backflow transformations via neural networks for quantum many-body wave functions*, *Phys. Rev. Lett.* **122**, 226401 (2019).
- [19] Y. Nomura and M. Imada, *Dirac-type nodal spin liquid revealed by refined quantum many-body solver using neural-network wave function, correlation ratio, and level spectroscopy*, *Phys. Rev. X* **11**, 031034 (2021).
- [20] W.-Y. Liu, Y.-Z. Huang, S.-S. Gong, and Z.-C. Gu, *Accurate simulation for finite projected entangled pair states in two dimensions*, *Phys. Rev. B* **103**, 235155 (2021).

- [21] J.R. Moreno, G. Carleo, A. Georges, and J. Stokes, *Fermionic wave functions from neural-network constrained hidden states*, *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2122059119 (2022).
- [22] M. Medvidović and J.R. Moreno, *Neural-network quantum states for many-body physics*, *Eur. Phys. J. Plus* **139**, 1 (2024).
- [23] A. Chen and M. Heyl, *Empowering deep neural quantum states through efficient optimization*, *Nat. Phys.* **20**, 1476 (2024).
- [24] H. Lange, A. Van de Walle, A. Abedinnia, and A. Bohrdt, *From architectures to applications: A review of neural quantum states*, *Quantum Sci. Technol.* **9**, 040501 (2024).
- [25] Y. Gu, W. Li, H. Lin, B. Zhan, R. Li, Y. Huang, D. He, Y. Wu, T. Xiang, M. Qin, L. Wang, and D. Lv, *Solving the hubbard model with neural quantum states*, *arXiv:2507.02644*.
- [26] W.-Y. Liu, H. Zhai, R. Peng, Z.-C. Gu, and G. K.-L. Chan, *Accurate simulation of the Hubbard model with finite fermionic projected entangled pair states*, *Phys. Rev. Lett.* **134**, 256502 (2025).
- [27] A. Chen, Z.-Q. Wan, A. Sengupta, A. Georges, and C. Roth, *Neural network-augmented Pfaffian wave-functions for scalable simulations of interacting fermions*, *arXiv:2507.10705*.
- [28] X. Li, Y. Chen, B. Li, H. Chen, F. Wu, J. Chen, and W. Ren, *Deep learning sheds light on integer and fractional topological insulators*, *arXiv:2503.11756*.
- [29] L. L. Viteritti, R. Rende, S. Sachdev, and G. Carleo, *Approaching the thermodynamic limit with neural-network quantum states*, *arXiv:2602.02665*.
- [30] J. Hermann, Z. Schätzle, and F. Noé, *Deep-neural-network solution of the electronic Schrödinger equation*, *Nat. Chem.* **12**, 891 (2020).
- [31] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, *Ab initio solution of the many-electron Schrödinger equation with deep neural networks*, *Phys. Rev. Res.* **2**, 033429 (2020).
- [32] K. Choo, A. Mezzacapo, and G. Carleo, *Fermionic neural-network states for ab-initio electronic structure*, *Nat. Commun.* **11**, 2368 (2020).
- [33] J. Hermann, J. Spencer, K. Choo, A. Mezzacapo, W. M. C. Foulkes, D. Pfau, G. Carleo, and F. Noé, *Ab initio quantum chemistry with neural-network wavefunctions*, *Nat. Rev. Chem.* **7**, 692 (2023).
- [34] I. von Glehn, J. S. Spencer, and D. Pfau, *A self-attention ansatz for ab-initio quantum chemistry*, *arXiv:2211.13672*.
- [35] H. Shang, C. Guo, Y. Wu, Z. Li, and J. Yang, *Solving Schrödinger equation with a language model*, *arXiv:2307.09343*.
- [36] D. Pfau, S. Axelrod, H. Sutterud, I. von Glehn, and J. S. Spencer, *Accurate computation of quantum excited states with neural networks*, *Science* **385**, eadn0137 (2024).
- [37] J. Kim, G. Pescia, B. Fore, J. Nys, G. Carleo, S. Gandolfi, M. Hjorth-Jensen, and A. Lovato, *Neural-network quantum states for ultra-cold Fermi gases*, *Commun. Phys.* **7**, 148 (2024).
- [38] M. Wilson, S. Moroni, M. Holzmann, N. Gao, F. Wudarski, T. Vegge, and A. Bhowmik, *Neural network ansatz for periodic wave functions and the homogeneous electron gas*, *Phys. Rev. B* **107**, 235139 (2023).
- [39] G. Cassella, H. Sutterud, S. Azadi, N. D. Drummond, D. Pfau, J. S. Spencer, and W. M. C. Foulkes, *Discovering quantum phase transitions with fermionic neural networks*, *Phys. Rev. Lett.* **130**, 036401 (2023).
- [40] C. Smith, Y. Chen, R. Levy, Y. Yang, M. A. Morales, and S. Zhang, *Unified variational approach description of ground-state phases of the two-dimensional electron gas*, *Phys. Rev. Lett.* **133**, 266504 (2024).
- [41] G. Pescia, J. Nys, J. Kim, A. Lovato, and G. Carleo, *Message-passing neural quantum states for the homogeneous electron gas*, *Phys. Rev. B* **110**, 035108 (2024).
- [42] M. Geier, K. Nazaryan, T. Zaklaman, and L. Fu, *Self-attention neural network for solving correlated electron problems in solids*, *Phys. Rev. B* **112**, 045119 (2025).
- [43] Y. Teng, D. D. Dai, and L. Fu, *Solving the fractional quantum Hall problem with self-attention neural network*, *Phys. Rev. B* **111**, 205117 (2025).
- [44] Y. Qian, T. Zhao, J. Zhang, T. Xiang, X. Li, and J. Chen, *Describing Landau level mixing in fractional quantum Hall states with deep learning*, *Phys. Rev. Lett.* **134**, 176503 (2025).
- [45] C.-T. Li, T. Ong, M. Geier, H. Lin, and L. Fu, *Attention is all you need to solve chiral superconductivity*, *arXiv:2509.03683*.
- [46] Z. Ge, C. Smith, Z. He, Y. Yang, Q. Li, H.-L. Kim, Z. Xiang, J. Xiao, W. Zhou, S. Kahn *et al.*, *Visualizing the impact of quenched disorder on 2D electron Wigner solids*, *Nature (London)* **654**, 902 (2026).
- [47] C. Smith, Y. Yang, Z.-Q. Wan, Y. Chen, M. A. Morales, and S. Zhang, *Neural network discovery of paired Wigner crystals in artificial graphene*, *arXiv:2602.16798*.
- [48] M. Medvidović and G. Carleo, *Classical variational simulation of the quantum approximate optimization algorithm*, *npj Quantum Inf.* **7**, 101 (2021).
- [49] A. D. King, A. Nocera, M. M. Rams, J. Dziarmaga, R. Wiersema, W. Bernoudy, J. Raymond, N. Kaushal, N. Heinsdorf, R. Harris *et al.*, *Beyond-classical computation in quantum simulation*, *Science* **388**, 199 (2025).
- [50] R. Haghshenas, E. Chertkov, M. Mills, W. Kadow, S.-H. Lin, Y.-H. Chen, C. Cade, I. Niesen, T. Begusić, M. S. Rudolph *et al.*, *Digital quantum magnetism on a trapped-ion quantum computer*, *Nature (London)* **653**, 56 (2026).
- [51] S. Sorella, *Generalized Lanczos algorithm for variational quantum Monte Carlo*, *Phys. Rev. B* **64**, 024512 (2001).
- [52] J. R. Trail, *Heavy-tailed random error in quantum Monte Carlo*, *Phys. Rev. E* **77**, 016703 (2008).
- [53] H. Shi and S. Zhang, *Infinite variance in fermion quantum Monte Carlo calculations*, *Phys. Rev. E* **93**, 033303 (2016).
- [54] P. L. Ríos and G. J. Conduit, *Tail-regression estimator for heavy-tailed distributions of known tail indices and its application to continuum quantum Monte Carlo data*, *Phys. Rev. E* **99**, 063312 (2019).
- [55] Z.-Q. Wan and S. Zhang, *Addressing the infinite variance problem in fermionic Monte Carlo simulations: Retrospective error remediation and the exact bridge link method*, *Phys. Rev. E* **112**, 065309 (2025).
- [56] A. Sinibaldi, C. Giuliani, G. Carleo, and F. Vicentini, *Unbiasing time-dependent variational Monte Carlo by projected quantum evolution*, *Quantum* **7**, 1131 (2023).

- [57] D. Hofmann, G. Fabiani, J. H. Mentink, G. Carleo, and M. A. Sentef, *Role of stochastic noise and generalization error in the time propagation of neural-network quantum states*, *SciPost Phys.* **12**, 165 (2022).
- [58] S. Pathak and L. K. Wagner, *A light weight regularization for wave function parameter gradients in quantum Monte Carlo*, *AIP Adv.* **10**, 085213 (2020).
- [59] L. Gravina, V. Savona, and F. Vicentini, *Neural projected quantum dynamics: A systematic study*, *Quantum* **9**, 1803 (2025).
- [60] D. Linteau, S. Moroni, G. Carleo, and M. Holzmman, *Variance reduction for forces and pressure in variational Monte Carlo*, *arXiv:2603.14521*.
- [61] P. J. Robinson, S. D. Pineda Flores, and E. Neuscamman, *Excitation variance matching with limited configuration interaction expansions in variational Monte Carlo*, *J. Chem. Phys.* **147**, 164114 (2017).
- [62] L. Otis, I. M. Craig, and E. Neuscamman, *A hybrid approach to excited-state-specific variational Monte Carlo and doubly excited states*, *J. Chem. Phys.* **153**, 234105 (2020).
- [63] K. Inui, Y. Kato, and Y. Motome, *Determinant-free fermionic wave function using feed-forward neural networks*, *Phys. Rev. Res.* **3**, 043126 (2021).
- [64] A. Misery, L. Gravina, A. Santini, and F. Vicentini, *Looking elsewhere: Improving variational Monte Carlo gradients by importance sampling*, *Mach. Learn.* **7**, 015035 (2026).
- [65] C. Attaccalite and S. Sorella, *Stable liquid hydrogen at high pressure by a novel ab initio molecular-dynamics calculation*, *Phys. Rev. Lett.* **100**, 114501 (2008).
- [66] J. R. Trail, *Alternative sampling for variational quantum Monte Carlo*, *Phys. Rev. E* **77**, 016704 (2008).
- [67] A. Malyshev, M. Schmitt, and A. I. Lvovsky, *Neural quantum states and peaked molecular wave functions: Curse or blessing?*, *arXiv:2408.07625*.
- [68] A.-J. Liu and B. K. Clark, *Efficient optimization of neural network backflow for ab initio quantum chemistry*, *Phys. Rev. B* **112**, 155162 (2025).
- [69] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, *Neural autoregressive distribution estimation*, *J. Mach. Learn. Res.* **17**, 1 (2016), <http://jmlr.org/papers/v17/16-272.html>.
- [70] M. Hibat-Allah, M. Ganahl, L. E. Hayward, R. G. Melko, and J. Carrasquilla, *Recurrent neural network wave functions*, *Phys. Rev. Res.* **2**, 023358 (2020).
- [71] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua, *Deep autoregressive models for the efficient variational simulation of many-body quantum systems*, *Phys. Rev. Lett.* **124**, 020503 (2020).
- [72] S. Humeniuk, Y. Wan, and L. Wang, *Autoregressive neural Slater-Jastrow ansatz for variational Monte Carlo simulation*, *SciPost Phys.* **14**, 171 (2023).
- [73] K. Sprague and S. Czischek, *Variational Monte Carlo with large patched transformers*, *Commun. Phys.* **7**, 90 (2024).
- [74] M. S. Moss, R. Wiersema, M. Hibat-Allah, J. Carrasquilla, and R. G. Melko, *Leveraging recurrence in neural network wavefunctions for large-scale simulations of heisenberg antiferromagnets on the square lattice*, *Phys. Rev. B* **112**, 134450 (2025).
- [75] M. S. Moss, R. Wiersema, M. Hibat-Allah, J. Carrasquilla, and R. G. Melko, *Leveraging recurrence in neural network wavefunctions for large-scale simulations of heisenberg antiferromagnets on the triangular lattice*, *Phys. Rev. B* **112**, 134449 (2025).
- [76] G. Carleo, F. Becca, M. Schiró, and M. Fabrizio, *Localization and glassy dynamics of many-body quantum systems*, *Sci. Rep.* **2**, 243 (2012).
- [77] D. M. Ceperley, *Fermion nodes*, *J. Stat. Phys.* **63**, 1237 (1991).
- [78] W. M. C. Foulkes, L. Mitás, R. J. Needs, and G. Rajagopal, *Quantum Monte Carlo simulations of solids*, *Rev. Mod. Phys.* **73**, 33 (2001).
- [79] L. Mitás, *Structure of fermion nodes and nodal cells*, *Phys. Rev. Lett.* **96**, 240402 (2006).
- [80] R. J. Needs, M. D. Towler, N. D. Drummond, and P. L. Ríos, *Continuum variational and diffusion quantum Monte Carlo calculations*, *J. Phys. Condens. Matter* **22**, 023201 (2009).
- [81] P. Lévy, *Calcul des probabilités* (Gauthier-Villars, 1925).
- [82] B. Gnedenko and A. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley Mathematical Series (Addison-Wesley, Reading, MA, 1954).
- [83] J. Nolan, *Univariate Stable Distributions: Models for Heavy Tailed Data*, Springer Series in Operations Research and Financial Engineering (Springer International Publishing, New York, 2020).
- [84] J.-P. Bouchaud and A. Georges, *Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications*, *Phys. Rep.* **195**, 127 (1990).
- [85] R. Y. Rubinstein, *Simulation and the Monte Carlo Method* (Wiley, New York, 1981).
- [86] A. Alexandru, P. F. Bedaque, A. Carosso, and H. Oh, *Infinite variance problem in fermion models*, *Phys. Rev. D* **107**, 094502 (2023).
- [87] H. Vrcan and J. H. Mentink, *Instability of explicit time integration for strongly quenched dynamics with neural quantum states*, *arXiv:2507.17421*.
- [88] G. Carleo, K. Choo, D. Hofmann, J. E. Smith, T. Westerhout, F. Alet, E. J. Davis, S. Efthymiou, I. Glasser, S.-H. Lin, M. Mauri, G. Mazzola, C. B. Pereira, and F. Vicentini, *NETKET: A machine learning toolkit for many-body quantum systems*, *SoftwareX* **10**, 100311 (2019).
- [89] F. Vicentini, D. Hofmann, A. Szabó, D. Wu, C. Roth, C. Giuliani, G. Pescia, J. Nys, V. Vargas-Calderón, N. Astrakhantsev, and G. Carleo, *NETKET3: Machine learning toolbox for many-body quantum systems*, *SciPost Phys. Codebases* **7** (2022).
- [90] N. Lambert, E. Giguère, P. Menczel, B. Li, P. Hopf, G. Su'arez, M. Gali, J. Lishman, R. Gadhvi, R. Agarwal, A. Galicia, N. Shammah, P. Nation, J. R. Johansson, S. Ahmed, S. Cross, A. Pitchford, and F. Nori, *QUTIP5: The quantum toolbox in Python*, *Phys. Rep.* **1153**, 1 (2026).
- [91] Z.-Y. Han, Z.-Q. Wan, and H. Yao, *Pfaffian quantum Monte Carlo: Solution to Majorana sign ambiguity and applications*, *arXiv:2408.10311*.
- [92] Z.-Q. Wan, X.-D. Dai, and G.-Y. Zhu, *Revisiting nishimori multicriticality through the lens of information measures*, *Phys. Rev. Res.* **8**, 023059 (2026).

- [93] https://github.com/therooler/nqs_blurred_sampling.
- [94] D. Ceperley, G. V. Chester, and M. H. Kalos, *Monte Carlo simulation of a many-fermion study*, *Phys. Rev. B* **16**, 3081 (1977).
- [95] L. L. Viteritti, R. Rende, and F. Becca, *Transformer variational wave functions for frustrated quantum spin systems*, *Phys. Rev. Lett.* **130**, 236401 (2023).
- [96] L. L. Viteritti, R. Rende, A. Parola, S. Goldt, and F. Becca, *Transformer wave function for two dimensional frustrated magnets: Emergence of a spin-liquid phase in the Shastry-Sutherland model*, *Phys. Rev. B* **111**, 134411 (2025).
- [97] R. Rende, F. Gerace, A. Laio, and S. Goldt, *Mapping of attention mechanisms to a generalized Potts model*, *Phys. Rev. Res.* **6**, 023057 (2024).
- [98] R. Rende and L. Loris Viteritti, *Are queries and keys always relevant? a case study on transformer wave functions*, *Mach. Learn.* **6**, 010501 (2025).
- [99] R. Wiersema, *Jaxmg: A multi-GPU linear solver in jax*, [arXiv:2601.14466](https://arxiv.org/abs/2601.14466).
- [100] M. Solinas, A. Valenti, N. Bou-Rabee, and R. Wiersema, *Neural quantum states in mixed precision*, [arXiv:2601.20782](https://arxiv.org/abs/2601.20782).