Variational Monte Carlo for Quantum Excited States via Nested Low-Rank Approximation

Minchan Jeong[1*†], **J. Jon Ryu**[2*‡], **Se-Young Yun**[3], **Gregory W. Wornell**[4]
[1]Center for AI and Natural Sciences, KIAS, Seoul 02455, South Korea
[2]Department of CSE, Penn State University, University Park, PA 16802, USA
[3]Kim Jaechul Graduate School of AI, KAIST, Daejeon 34141, South Korea
[4]Department of EECS, MIT, Cambridge, MA 02139, USA
mcjeong@kias.re.kr, jongha@psu.edu, yunseyoung@kaist.ac.kr, gww@mit.edu

Abstract

The electronic Schrödinger equation encodes a molecule's electronic structure and quantum properties, but solving it scales exponentially with the number of electrons. Solving this equation with neural-network variational Monte Carlo (VMC) has reached chemical accuracy on the ground state of small molecules; extending this success to excited states is constrained by two existing paradigms. Penalty-based methods need K Laplacians per step but require tuning the overlap-penalty weights against the energy gaps they aim to compute. Natural excited states VMC (NES-VMC) is penalty-free, but it combines the K ansatzes into one determinant, whose local energy requires K^2 Laplacians per step. We resolve this trade-off via low-rank approximation (LoRA) reparameterization of the rank-K variational objective. The reparameterized objective admits a VMC-compatible stochastic gradient estimator via MCMC sampling from the per-state Born densities, and a sequential nesting scheme recovers the lowest K eigenstates in eigenvalue order. The resulting algorithm, NestedLoRA-VMC, inherits a global optimality guarantee from the LoRA principle. On the first-row atoms Li through Ne at $K = 10$, its accuracy is comparable to that of NES-VMC and of a state-of-the-art penalty-based method.

1 Introduction

Ab initio quantum chemistry aims to compute the electronic states of molecules by solving the Schrödinger equation $\mathcal{H}\psi = E\psi$. Accurate computation of these states remains challenging due to the curse of dimensionality: the wavefunction lives on the $3n_{\mathsf{el}}$-dimensional configuration space of n_{el} electrons, and the computational cost of traditional high-accuracy methods grows steeply with n_{el}. Recently, neural-network variational Monte Carlo (VMC) has emerged as a promising alternative, achieving chemical accuracy for the ground states of small molecules with architectures such as FermiNet [26], PauliNet [10], and Psiformer [38]. However, extending this success to excited states remains challenging. While minimizing the Rayleigh quotient naturally characterizes the ground state, computing multiple excited states requires an additional mechanism to recover mutually orthogonal eigenfunctions and identify their energy ordering. Moreover, such a mechanism must be compatible with VMC, where optimization relies on Monte Carlo samples drawn from the Born densities of the neural-network ansatzes.

*The two authors contributed equally to this work.
†Work done while a Ph.D. student at KAIST.
‡Work done while a postdoctoral researcher at MIT.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing excited-state VMC methods can be broadly categorized into two approaches. On the one hand, *penalty-based methods* [3, 23, 39, 37] minimize the sum of the Rayleigh quotients of multiple ansatzes, augmented with penalties on their pairwise overlaps to encourage orthogonality. While conceptually straightforward, these methods require careful tuning of the penalty weights. In particular, Wheeler et al. [39] show that recovering the desired eigenstates at the global minimum requires sufficiently large penalty weights relative to the energy gaps, which are themselves unknown quantities to be estimated. On the other hand, *natural excited states VMC* (NES-VMC) [27] takes an elegant, penalty-free approach by transforming the excited-state problem into a ground-state problem of an extended quantum system. By constructing a determinant from the individual ansatzes, NES-VMC implicitly enforces orthogonality without additional regularization. The individual eigenstates are then recovered by diagonalizing the Hamiltonian projected onto the learned subspace. However, this construction increases computational cost: evaluating its local energy requires K^2 Laplacian evaluations per optimization step for K states, compared to K for penalty-based methods. Thus, the existing approaches present a tradeoff between penalty tuning and computational efficiency (Table 1).

We pursue a different variational perspective based on the low-rank approximation (LoRA) principle. Specifically, we build on NestedLoRA [30], a general framework for learning ordered eigenfunctions of an operator using neural networks, without explicitly imposing orthogonality constraints or introducing penalty weights. This framework has been successfully applied to low-dimensional partial differential equations and representation learning [30], as well as stochastic dynamical systems [14]. Despite these appealing properties, the LoRA objective is not directly compatible with VMC. The main obstacle is that it depends on the norms of the learned functions, which involve high-dimensional integrals that cannot be directly estimated from samples of their normalized Born densities.

Contributions. We overcome this obstacle through a simple reparameterization, representing each function in the LoRA objective as a normalized neural-network ansatz multiplied by a trainable scale. This makes the resulting objective compatible with the Born sampling used in VMC. Building on this formulation, we propose NestedLoRA-VMC, which learns eigenstates directly in energy order through the *sequential nesting* technique [7, 30]. The proposed framework is penalty-free, requires no post-training diagonalization, and retains the global optimality characterization of the underlying LoRA principle. Moreover, NestedLoRA-VMC requires only K Laplacian evaluations per optimization step, matching the cost of penalty-based methods rather than the K^2 evaluations of NES-VMC. We evaluate NestedLoRA-VMC on first-row atoms from Li to Ne with $K = 10$ states and demonstrate accuracy comparable to NES-VMC and the state-of-the-art penalty-based method of Szabó et al. [37], without requiring gap-dependent penalty weights.

Organization. Section 2 introduces the problem and existing excited-state VMC approaches. Section 3 develops the proposed VMC-compatible LoRA formulation and its theoretical properties. Section 4 presents experimental results, and Section 5 concludes with limitations and future directions. Proofs and additional technical details are deferred to the appendix.

2 Preliminaries

Our goal is to recover the K lowest eigenpairs $\{(\varphi_k, E_k)\}_{k=1}^{K}$ of the electronic Hamiltonian $\mathcal{H}$ of a molecule on the space $\mathscr{H}$ of its wavefunctions, both defined below. We order them as $E_1 \le E_2 \le \cdots \le E_K$. Consider a molecule with n_{el} electrons at positions $\mathbf{r}_1, \ldots, \mathbf{r}_{n_{\mathsf{el}}} \in \mathbb{R}^3$ and n_{nu} fixed nuclei of charges Z_I at positions $\mathbf{R}_I$, $I \in [n_{\mathsf{nu}}]$. In the Born–Oppenheimer approximation and with $\mathbf{x} = (\mathbf{r}_1, \ldots, \mathbf{r}_{n_{\mathsf{el}}})$, the non-relativistic electronic Hamiltonian is

$$\mathcal{H} := \sum_{i=1}^{n_{\mathsf{el}}} \left\{ -\frac{1}{2}\nabla_{\mathbf{r}_i}^2 - \sum_{I=1}^{n_{\mathsf{nu}}} \frac{Z_I}{\|\mathbf{r}_i - \mathbf{R}_I\|} \right\} + \sum_{1 \le i < j \le n_{\mathsf{el}}} \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|},$$

and the time-independent Schrödinger equation reads $\mathcal{H}\psi = E\psi$. Each electron also has a spin, up or down. We fix the numbers $n_\uparrow$ and $n_\downarrow$ of spin-up and spin-down electrons and assign the spins to the electrons, so that a wavefunction is a function of $\mathbf{x}$ alone that must be antisymmetric under exchange of two electrons with the same spin (Appendix D.1.1). The neural-network architecture enforces this antisymmetry (Appendix H.1). The wavefunctions therefore live in the space $\mathscr{H}$ of complex-valued functions in $L^2(\mathbb{R}^{3n_{\mathsf{el}}})$ with this antisymmetry. As a closed subspace of $L^2(\mathbb{R}^{3n_{\mathsf{el}}})$, $\mathscr{H}$ is a Hilbert space with the standard inner product $\langle f \mid g \rangle := \int f^*(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x}$. Since $\mathcal{H}$ commutes

with permutations of same-spin electrons, it maps $\mathscr{H}$ into itself. It is moreover self-adjoint on $\mathscr{H}$, a property that requires its adjoint to have the same domain and that the Kato–Rellich theorem establishes despite the Coulomb singularities (Appendix D.1.3). We assume that $E_1, \dots, E_K$ lie strictly below the essential spectrum of $\mathcal{H}$, which begins at the first ionization limit (Appendix D.1.4). Since $\mathcal{H}$ has real coefficients in position space, its eigenstates can be chosen real-valued, and we restrict every ansatz to a real-valued function.[4]

2.1 Variational Monte Carlo for ground state computation

VMC minimizes the Rayleigh quotient of a wavefunction ansatz ψ_θ with parameters θ,

$$\min_\theta \mathcal{L}(\theta), \quad \mathcal{L}(\theta) := \frac{\langle \psi_\theta \mid \mathcal{H}\psi_\theta \rangle}{\langle \psi_\theta \mid \psi_\theta \rangle} = \mathbb{E}_{p_\theta(\mathbf{x})}\big[\mathcal{E}_\theta(\mathbf{x})\big],$$

where $\mathcal{E}_\theta(\mathbf{x}) := \frac{\mathcal{H}\psi_\theta(\mathbf{x})}{\psi_\theta(\mathbf{x})}$ is called the *local energy*, and $p_\theta(\mathbf{x}) := \frac{|\psi_\theta(\mathbf{x})|^2}{\langle \psi_\theta \mid \psi_\theta \rangle}$ is the induced probability density from the wavefunction ansatz, called the *Born density*. Then the gradient of the Rayleigh quotient is

$$\nabla_\theta \mathcal{L}(\theta) = 2\mathbb{E}_{p_\theta(\mathbf{x})}\Big[\mathrm{Re}\Big((\mathcal{E}_\theta(\mathbf{x}) - \mathbb{E}_{p_\theta(\mathbf{x})}[\mathcal{E}_\theta(\mathbf{x})])\nabla_\theta \log \psi_\theta^*(\mathbf{x})\Big)\Big]. \tag{1}$$

See Lemma E.1 for its proof. The gradient (1) can be estimated unbiasedly from *walkers*, samples drawn from the Born density p_θ by a Markov chain Monte Carlo (MCMC) algorithm [4]. Since integrals over $\mathbb{R}^{3n_{\mathsf{el}}}$ are otherwise intractable, VMC needs an objective, or at least its gradient, written as expectations under the Born densities of the ansatzes; we call such an objective or gradient *VMC-compatible*.

2.2 Prior works on excited states computation

Two VMC-compatible approaches exist: penalty-based methods tune penalty weights against the energy gaps, and NES-VMC needs K^2 Laplacians per step.

(1) Penalty-Based Methods. Penalty-based methods [3, 23, 39, 37] keep the per-state Rayleigh quotient and add an explicit pairwise overlap penalty to enforce orthogonality:

$$\mathcal{L}_{\mathsf{penalty}}(\psi_{1:K}) = \sum_{k=1}^{K} \mathbb{E}_{p_k}[\mathcal{E}_k] + \sum_{k<\ell} \alpha_{k\ell}\, \Pi(\tilde{\mathsf{S}}_{k\ell}), \quad \tilde{\mathsf{S}}_{k\ell} := \frac{\langle \psi_k \mid \psi_\ell \rangle}{\|\psi_k\|\|\psi_\ell\|},$$

where $\mathcal{E}_k(\mathbf{x}) := \frac{\mathcal{H}\psi_k(\mathbf{x})}{\psi_k(\mathbf{x})}$ denotes the local energy of the k-th ansatz and $p_k(\mathbf{x}) \propto |\psi_k(\mathbf{x})|^2$ the corresponding Born density, $\alpha_{k\ell} > 0$ are penalty weights, and Π is a penalty kernel, such as $\Pi(s) = |s|^2$, which admits an unbiased estimator [23, 39]. Wheeler et al. [39] established an ensemble variational principle showing that an energy-weighted version of this objective recovers the ordered lowest K eigenfunctions of $\mathcal{H}$ once the overlap penalty exceeds a spectral-gap-dependent threshold (Appendix C.2). This threshold depends on the energy gaps, so the tuning depends on the very spectrum the optimization aims to compute. Based on this criterion, Szabó et al. [37] set each $\alpha_{k\ell}$ at every step to a global scale $\tilde{\alpha}$ times the larger of two exponential moving averages (EMAs) over training steps: that of the energy gap between states k and ℓ and that of the local-energy standard deviation of state k (Appendix C.2). Whether the global minimum gives the lowest K eigenstates thus depends on how accurate these estimates are.

(2) NES-VMC. Pfau et al. [27] propose an extended-system construction for the lowest K states. NES-VMC evaluates the K ansatzes $\psi_1, \dots, \psi_K$ on K electron configurations $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^K)$ and forms the matrix ansatz $\Psi(\mathbf{X}) \in \mathbb{C}^{K\times K}$ with $\Psi_{ik}(\mathbf{X}) = \psi_k(\mathbf{x}^i)$. The total ansatz is the determinant $\Phi(\mathbf{X}) := \det \Psi(\mathbf{X})$. NES-VMC minimizes the energy of this extended system, in which $\mathcal{H}$ acts on each of the K configurations $\mathbf{x}^i$ and the results are summed. This energy has two forms (Theorems C.1 and C.2 in Appendix C.1)

$$\mathcal{L}_{\mathsf{NES\text{-}VMC}}(\psi_{1:K}) = \mathrm{tr}(\mathsf{S}^{-1}\mathsf{H}) = \mathbb{E}_{\mathbf{X}\sim|\Phi|^2}\big[\mathrm{tr}(\mathsf{E}_{\mathsf{loc}}(\mathbf{X}))\big], \tag{2}$$

[4] This relies on the absence of external magnetic fields and of relativistic effects such as spin-orbit coupling, both outside the present scope.

Table 1: Comparison of NestedLoRA-VMC with the two existing excited-state VMC approaches. Theoretical scaling excludes spin-penalty overhead (Appendix I.2). †Measured by Szabó et al. [37, Figure 1] for $K \in \{1, \ldots, 30\}$ on Ne.

	NES-VMC [27]	**Penalty-based** [3, 37]	**NestedLoRA-VMC**
Objective	Energy of the extended system	Energies + overlap penalty	Rank-K approximation of $-\mathcal{H}$
Overlap-penalty weights	-	$\alpha_{k\ell} > \lvert E_k - E_\ell \rvert$, per pair	-
Sampling	One chain over Kn_{el} electrons	K chains over n_{el} electrons	K chains over n_{el} electrons
Laplacians per step	K^2	K	K
Theoretical scaling	$O(K^2 {n_{\mathsf{el}}}^4 + K^3)$	$O(K {n_{\mathsf{el}}}^4 + K^2 {n_{\mathsf{el}}}^3)$	$O(K {n_{\mathsf{el}}}^4 + K^2 {n_{\mathsf{el}}}^3)$
Wall-time scaling on Ne	$O(K^{2.5})^\dagger$	$O(K^{1.1})^\dagger$	$\approx$ penalty-based (Figure 6)
Output	Subspace, diagonalized post hoc	Ordered states	Ordered states

where $\mathsf{S} := [\langle \psi_k \mid \psi_\ell \rangle]_{k,\ell=1}^K$ and $\mathsf{H} := [\langle \psi_k \mid \mathcal{H}\psi_\ell \rangle]_{k,\ell=1}^K$ are the overlap and energy matrices, and $\mathsf{E}_{\mathsf{loc}}(\mathbf{X}) := \Psi(\mathbf{X})^{-1}(\mathcal{H}\Psi)(\mathbf{X})$ is the *total local energy matrix* with $(\mathcal{H}\Psi)(\mathbf{X}) := [\,(\mathcal{H}\psi_k)(\mathbf{x}^i)\,]_{i,k=1}^K$. The matrix $\mathsf{S}^{-1}\mathsf{H}$ represents $\mathcal{H}$ projected onto $\mathrm{span}\{\psi_{1:K}\}$, so the middle form is the sum of the eigenvalues of this projected operator, minimized with value $E_1 + \cdots + E_K$ by any K ansatzes whose span is that of $\varphi_1, \ldots, \varphi_K$. The right form of Eq. (2) is the ground-state VMC form on the K-fold configuration space under the Born density $|\Phi|^2/\langle \Phi \mid \Phi \rangle$ of the total ansatz, written $\mathbf{X} \sim |\Phi|^2$. The objective and its gradient are thus expectations under this density, so the objective is VMC-compatible. Since the minimum does not single out eigenfunctions within this span, NES-VMC orders and orthogonalizes the states in a post-processing step. This step diagonalizes the sample mean of $\mathsf{E}_{\mathsf{loc}}$, an estimate of $\mathsf{S}^{-1}\mathsf{H}$ (Proposition C.1), and reads the individual energies from the eigenvalues of this mean. With this procedure, NES-VMC recovers accurate vertical excitation energies, including double excitations, on molecules up to the size of benzene.

NES-VMC incurs three structural costs, all from the total ansatz $\Phi = \det \Psi$ through which it estimates (2). First, $\mathsf{E}_{\mathsf{loc}}$ contains a cross term $(\mathcal{H}\psi_k)(\mathbf{x}^i)$ for every pair (i, k), and these K^2 Laplacians per step dominate the wall time (Table 1). Second, the Born density of Φ couples all K states in a single Markov chain, whereas penalty-based methods and NestedLoRA-VMC sample each state in its own chain, in parallel. Lastly, the gradient flows through the matrix inverse $\Psi(\mathbf{X})^{-1}$, which can become ill-conditioned when two ansatzes are nearly linearly dependent. We elaborate each point in Appendix I.2 and survey the numerical-stability measures of Pfau et al. [27] in Appendix H.6. Appendix B presents more related work on excited states and on learning eigenfunctions with neural networks.

3 Proposed method: NestedLoRA-VMC

This section adapts the NestedLoRA framework, developed for general operator decomposition, to excited-state VMC. The resulting algorithm, NestedLoRA-VMC, is penalty-free and needs K Laplacians per step. It inherits a global optimality guarantee from the LoRA principle and learns the ordered eigenfunctions without explicit diagonalization. Sections 3.1 and 3.2 build the objective of NestedLoRA-VMC and its guarantee, and Sections 3.3 and 3.4 estimate it by Monte Carlo and compare its per-step cost with the two existing approaches.

3.1 Operator spectral decomposition via low-rank approximation

For a self-adjoint operator $\mathcal{T}$ that is bounded above, suppose that we wish to recover its top-K eigenvalues $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_K$ and corresponding orthonormal eigenfunctions $|\varphi_1\rangle, \ldots, |\varphi_K\rangle$. In our case, we will set $\mathcal{T} = -\mathcal{H}$ such that $\mu_k = -E_k$. Recently, Ryu et al. [30] proposed the low-rank approximation (LoRA) objective

$$\mathcal{L}_{\mathsf{LoRA}}(f_{1:K}) := -2\sum_{k=1}^{K} \langle f_k \mid \mathcal{T} f_k \rangle + \sum_{k=1}^{K}\sum_{\ell=1}^{K} \langle f_k \mid f_\ell \rangle \langle f_\ell \mid f_k \rangle, \tag{3}$$

as a training objective for parametric spectral decomposition. For a Hilbert–Schmidt operator $\mathcal{T}$, this objective is equivalent to the rank-K approximation error in Hilbert–Schmidt norm $\|\mathcal{T} -$

$\sum_{k=1}^{K} |f_k\rangle\langle f_k|\|_{\mathsf{HS}}^2$. If this $\mathcal{T}$ is moreover positive semidefinite and $\mu_K > \mu_{K+1}$, a global minimizer $f_{1:K}^\star$ of $\mathcal{L}_{\mathsf{LoRA}}(f_{1:K})$ recovers the top-K eigenspace of $\mathcal{T}$ as $\sum_{k=1}^{K} |f_k^\star\rangle\langle f_k^\star| = \sum_{k=1}^{K} \mu_k |\varphi_k\rangle\langle\varphi_k|$; for a self-adjoint $\mathcal{T}$ with an orthonormal basis of eigenfunctions, not necessarily compact or positive semidefinite, the minimizer recovers the top-$\min(K, r)$ eigenspace, where r is the number of positive eigenvalues [30, Theorems 3.1 and C.5].

While neither result covers our quantum chemistry case $\mathcal{T} = -\mathcal{H}$, which is unbounded and has a continuous spectrum of scattering states, we show that the LoRA principle is still applicable when the lowest K energies are eigenvalues below its essential spectrum (Appendix D.1.4). That is, a global minimizer recovers the top-K eigenspace as above whenever $\mu_K > \max(\mu_{K+1}, \sup \sigma_{\mathrm{ess}}(\mathcal{T}), 0)$, with $\sigma_{\mathrm{ess}}(\mathcal{T})$ the essential spectrum; the proof applies the Eckart–Young theorem to $\mathcal{T}$ projected onto the span of $f_{1:K}$, the only part of $\mathcal{T}$ on which the LoRA objective depends (Appendix D).

3.2 NestedLoRA-VMC

The LoRA objective requires a reparameterization to be *VMC-compatible*. Ryu et al. [30] further propose an optimization technique called *nesting* to learn ordered eigenfunctions without explicit diagonalization, and call the framework *NestedLoRA*.

Reparameterization. The key idea for VMC-compatibility is to reparameterize each function f_k in (3) as $\sqrt{\lambda_k}\,\psi_k/\|\psi_k\|$, with the ansatzes $\psi_{1:K}$ parameterized as neural networks with parameters θ and $\lambda_{1:K} \in \mathbb{R}^K_{\geq 0}$ as additional trainable scale parameters. Under the sequential nesting introduced below, each λ_k plays the role of the k-th eigenvalue estimate. This yields the objective

$$\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:K}, \lambda_{1:K}) := -2\sum_{k=1}^{K} \lambda_k \mathsf{R}_{kk}^{\mathcal{T}} + \sum_{k,\ell=1}^{K} \lambda_k \lambda_\ell \mathsf{R}_{k\ell}^{I} \mathsf{R}_{\ell k}^{I},$$

where we define the generic functional $\mathsf{R}_{k\ell}^{\mathcal{O}} := \frac{\langle \psi_k \mid \mathcal{O}\psi_\ell \rangle}{\langle \psi_k \mid \psi_k \rangle} = \mathbb{E}_{p_k}[\frac{\mathcal{O}\psi_\ell}{\psi_k}]$ for a symmetric operator $\mathcal{O} \in \{\mathcal{T}, I\}$. The objective is VMC-compatible: drawing electronic configurations $\mathbf{X}$ from the product distribution $|\psi_1(\mathbf{x}^1)|^2 \cdots |\psi_K(\mathbf{x}^K)|^2$ yields unbiased estimates of the objective and its gradient $\nabla_\theta \mathcal{J}_{\mathsf{LoRA}}$ (Theorem E.1).

Sequential nesting for learning ordered excited states. As in Ryu et al. [31], we order the learned eigenfunctions by sequential nesting. Sequential nesting updates each network k by the gradient with respect to ψ_k of the order-k objective $\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k})$, the LoRA objective of the first k states, and updates all K networks at once. This update can be read as K nested problems solved together: if states $1, \ldots, k-1$ are the first $k-1$ eigenpairs, the order-k objective is minimized by $\psi_k \propto \varphi_k$ and $\lambda_k = \mu_k$, so the fixed point of the update is the ordered eigenpairs (Appendix D.4).

We define the sequential nesting gradient as

$$\nabla_\theta^{\mathsf{seq}} := \frac{1}{4}\sum_{k=1}^{K} \langle \nabla_\theta \psi_k \mid \partial_{\psi_k} \mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k}) \rangle = \sum_{k=1}^{K} \lambda_k \,\mathrm{Re}\Big\{ -\tilde{\Delta}_{kk}^{\mathcal{T}} + \sum_{\ell=1}^{K} \mathsf{M}_{k\ell}^{\mathsf{seq}} \lambda_\ell \mathsf{R}_{\ell k}^{I} \tilde{\Delta}_{k\ell}^{I} \Big\}, \quad (4)$$

where ∂_{ψ_k} is the functional derivative with respect to ψ_k, $\mathsf{M}_{k\ell}^{\mathsf{seq}} = \mathbb{1}\{k > \ell\}$ is the strictly lower-triangular mask and $\tilde{\Delta}_{k\ell}^{\mathcal{O}} := \mathrm{Cov}_{p_k}(\mathcal{O}\psi_\ell/\psi_k, \nabla_\theta \log \psi_k^*)$ is the covariance, under p_k, of the operator action $\mathcal{O}\psi_\ell/\psi_k$ against the per-state score $\nabla_\theta \log \psi_k^*$ (using score in the same log-derivative sense as in eq. (1)), for $\mathcal{O} \in \{\mathcal{T}, I\}$. Appendices E.1.2 and E.1.4 derive Eq. (4). For real-valued ansatzes the gradient is VMC-compatible, a sum over states of expectations under the Born densities p_k, estimated by averages over samples $\mathbf{x} \sim p_k$, for each $k \in [K]$:

$$\nabla_\theta^{\mathsf{seq}} = \sum_{k=1}^{K} \mathbb{E}_{p_k(\mathbf{x})}\Big[(g_k(\mathbf{x}) - \mathbb{E}_{p_k}[g_k])\, \nabla_\theta \log |\psi_k| \Big],$$

$$g_k(\mathbf{x}) := \lambda_k \left(-\frac{\mathcal{T}\psi_k(\mathbf{x})}{\psi_k(\mathbf{x})} + \sum_{\ell=1}^{K} \mathsf{M}_{k\ell}^{\mathsf{seq}}\, \lambda_\ell\, \mathsf{R}_{\ell k}^{I}\, \frac{\psi_\ell(\mathbf{x})}{\psi_k(\mathbf{x})} \right). \quad (5)$$

In our implementation, we drop the prefactor λ_k in $g_k(\mathbf{x})$ to normalize the gradient norm, which can be equivalently understood as using the state-dependent learning rate η/λ_k for a base learning rate

η; this rescales the gradient of each state by a positive factor and keeps the fixed point. A similar state-dependent learning rate was recently studied for k-SVD by gradient descent with the LoRA principle [13].

3.3 Implementation details

In Sections 3.1 and 3.2, we derive a parametric gradient for finding the ordered eigenpairs as a fixed point out of iterative gradient-based updates. In practical optimization, there exist several issues that hinder effective training with a direct implementation of the gradient above: heavy-tails of the local energy and overlap statistics, the step-size sensitivity of the gradient-based eigenvalue updates, the conditioning of each network's objective, and unstable order of the states at the beginning of training. The first is a statistical issue, while the rest are related to optimization. In this section, we provide a solution to each problem without altering the fixed-point property of the desired solution.

3.3.1 Asymmetric clipping for local energy and overlap statistics

To implement the gradient in Eq. (5), we need the local energy statistics $\mathcal{T}\psi_k/\psi_k$ and local overlap statistics ψ_ℓ/ψ_k. Since both statistics may blow up around the nodes of ψ_k (i.e., where $\psi_k \approx 0$), it is a common practice in the literature [26, 38, 27, 37] to apply a clipping on these values. However, we empirically find that an asymmetric clipping strategy, clipping local energies while using raw local overlap statistics, leads to much more stable training; we attribute the effectiveness of the asymmetric clipping rule to differences in how the statistics behave near convergence, as explained below.

First, note that the local energy $\mathcal{T}\psi_k/\psi_k$ is constant at the fixed point, so the clipping is inactive there; Appendix F.1 gives the clipping rule. In contrast, the local overlap ψ_ℓ/ψ_k stays heavy-tailed at the fixed point; theoretically, we can show that, under the Born density p_k, the probability that its magnitude exceeds t decays as t^{-3} under regularity conditions; see Appendix F.2.1. This suggests that, for example, the clipping rule of Szabó et al. [37], which clips outside the median ± 10 median absolute deviations, may still unexpectedly perturb some walkers even around convergence, since the heavy tail places some walkers outside the clip window; see Figure 4c in Appendix F.2.2. The perturbation can bias the batch mean of the ratio, whose expectation $\mathsf{R}^I_{k\ell}$ vanishes at the fixed point.

We demonstrate empirically on Li that clipping the ratio substantially biases its batch mean, both early in training and near convergence. Early in training, when the overlap remains nonzero and the overlap term helps separate the networks, clipping shifts the batch mean away from a reference estimate obtained using long Metropolis chains targeting a mixture of p_k and p_ℓ, by an amount many times larger than its variation across batches; see Figure 4. At the final training step, near a fixed point at which the overlap should vanish, the clip still changes the walkers of many ratios and can leave their batch means spuriously nonzero; see Appendix F.2.2.

3.3.2 Analytical eigenvalue updates

A gradient update of the eigenvalue parameters $\lambda_{1:K}$ together with the networks needs a step size for λ, which may be hard to tune together with training the neural ansatzes. To simplify the updates, we instead propose to track the eigenvalues by an analytical update at each step. Since the objective $\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k})$ is strictly convex quadratic in λ_k, we solve its stationary equation $\partial_{\lambda_k}\mathcal{J}_{\mathsf{LoRA}} = 0$ to obtain an analytical update. Mirroring the sequential update of ψ_k, we hold the lower-state parameters $\{\lambda_j : j < k\}$ fixed and solve $\partial_{\lambda_k}\mathcal{J}_{\mathsf{LoRA}} = 0$ sequentially for $k = 1, \ldots, K$, yielding

$$\boldsymbol{\lambda}^\star = \left(\mathsf{I} + \mathsf{M}^{\mathsf{seq}} \odot (\mathsf{R}^I \odot (\mathsf{R}^I)^\intercal)\right)^{-1} \mathrm{diag}(\mathsf{R}^{\mathcal{T}}), \tag{6}$$

which can be computed in one cheap $O(K^2)$ triangular sweep per step; Appendix E.2 gives the step-by-step derivation. Because the system is lower-triangular, a heavy-tailed estimate at state m enters $\lambda^\star_j$ only for $j \geq m$. In practice, we plug EMAs of the estimates of R^I and $\mathrm{diag}(\mathsf{R}^{\mathcal{T}})$ into Eq. (6) to smooth the trajectory across steps. With the exact R^I and $\mathrm{diag}(\mathsf{R}^{\mathcal{T}})$ in place of these estimates, the solve returns $\lambda_k = \mu_k$ at the ordered eigenpairs, so the fixed point is unchanged.

3.3.3 Per-network operator shifts for curvature shaping

Another source of instability in the sequential nesting gradient of Eq. (5) is the imbalance between its energy and overlap terms. While the energy term distinguishes eigenstates through spectral gaps, the

overlap term is weighted by $\lambda_\ell \approx -E_\ell$, whose magnitude can be much larger than the spectral gaps. This creates steep directions toward lower eigenstates and relatively flat directions toward higher ones, leading to unstable convergence.

To mitigate this imbalance, we introduce a network-specific operator shift. Since network k is updated only through the k-th term of Eq. (4), it can use $\mathcal{T}_k := E_{\mathrm{ref},k} I - \mathcal{H}$, with $E_{\mathrm{ref},k}$ fixed during gradient computation. The shift adds a constant $-E_{\mathrm{ref},k}$ to the local energy, which vanishes upon centering, while changing each lower-state overlap coefficient from $\lambda_\ell \approx -E_\ell$ to $\lambda_\ell + E_{\mathrm{ref},k} \approx E_{\mathrm{ref},k} - E_\ell$ near orthogonality. Thus, choosing $E_{\mathrm{ref},k}$ near E_k reduces the relative weight of the overlap term without changing the centered energy contribution.

To see its effect on curvature, assume $\psi_\ell \propto \varphi_\ell$ for $\ell < k$ and expand $q_k := \frac{\psi_k}{\|\psi_k\|} = \sum_n a_{kn}\varphi_n + \chi_k$, where φ_n are the discrete eigenfunctions of $\mathcal{H}$ with energies E_n, and χ_k belongs to its essential spectral subspace. The effective objective underlying the k-th gradient, with the prefactor λ_k removed, is approximately $E_k + \sum_{\ell<k}(E_{\mathrm{ref},k} - E_k)\, a_{k\ell}^2 + \sum_{n>k}(E_n - E_k)\, a_{kn}^2 + \langle \chi_k \mid (\mathcal{H} - E_k)\chi_k \rangle$. Thus, the lower-state curvature decreases from $-E_k$ to $E_{\mathrm{ref},k} - E_k$, while the higher-state curvature remains unchanged.

In our implementation, we choose $E_{\mathrm{ref},k} = -\lambda_k + c\,\sigma_k$, where $-\lambda_k$ estimates E_k via Eq. (6), σ_k is the walker standard deviation of the clipped local energy, and $c = 5$. The margin $c\,\sigma_k$ helps ensure $E_{\mathrm{ref},k} > E_k$, so that the target state retains a positive eigenvalue under $\mathcal{T}_k$ and remains recoverable by LoRA. Near convergence, the lower-state curvature becomes approximately $c\,\sigma_k$ rather than $|E_k|$. Whenever $E_{\mathrm{ref},k} > E_k$, φ_k remains a minimizer, up to degeneracies. Appendix G.1 analyzes the curvature imbalance for carbon, and Section 4.4 demonstrates the effect of the shift.

3.3.4 Online state ordering for faster convergence

We find that online state reordering lets the networks learn the lowest K states without waiting for their energies to line up with their indices during training. We assign, at every step, the order-k objective to the network whose EMA energy ranks k-th within its spin sector (Section 4.1 and Appendix G.2). The ranking reads only the EMA energies, so it is the same under every shift $E_{\mathrm{ref},k}$ of Section 3.3.3. The assignment changes only which network carries which index, so the fixed point is unchanged.

3.4 Choice of parameterization and per-iteration efficiency

NestedLoRA-VMC needs K Laplacians per step, against K^2 for NES-VMC. Like the penalty methods, NestedLoRA-VMC applies $\mathcal{H}$ to each state only at that state's own walkers, $(\mathcal{H}\psi_k)(\mathbf{x}^k)$, while the local-energy matrix of NES-VMC needs $(\mathcal{H}\psi_k)(\mathbf{x}^i)$ at every pair (i,k). A Laplacian is the most expensive evaluation in a step: it sums $3n_{\mathsf{el}}$ second derivatives, so with the forward Laplacian [19] it costs about $3n_{\mathsf{el}}$ forward evaluations, $O({n_{\mathsf{el}}}^4)$ against $O({n_{\mathsf{el}}}^3)$ for one forward evaluation (Appendix I.1).

FermiNet [26] and the Psiformer [38] compute a feature of each electron with a *backbone* of permutation-equivariant layers, map these features to an *orbital matrix*, and sum the determinants of such matrices. Within one forward evaluation, the determinant of the orbital matrix sets the order ${n_{\mathsf{el}}}^3$, and on first-row atoms most of the cost is in the dense layers of the backbone, which act on a 256-dimensional feature of each electron. These costs, times the number of forward evaluations per step that Appendix I.2 counts by component, give the theoretical scaling in Table 1. The per-step wall time measured on Ne follows it for NES-VMC, the penalty class [37, Figure 1] and NestedLoRA-VMC (Appendix I.3).

NES-VMC uses one *joint* backbone with K heads, each of which maps the shared features of the electrons to the orbital matrices of one state. Whether this costs less than K disjoint networks depends on where $\mathcal{H}$ is applied. In NES-VMC every state needs its Laplacian at every walker set, and a joint backbone computes the backbone's part once per walker set for all K heads. In NestedLoRA-VMC and the penalty methods each state needs its Laplacian only at its own walkers, so the Laplacian costs the same under either parameterization, and a joint backbone saves only on the overlaps, which carry no factor $3n_{\mathsf{el}}$. We use K disjoint networks, following Szabó et al. [37].

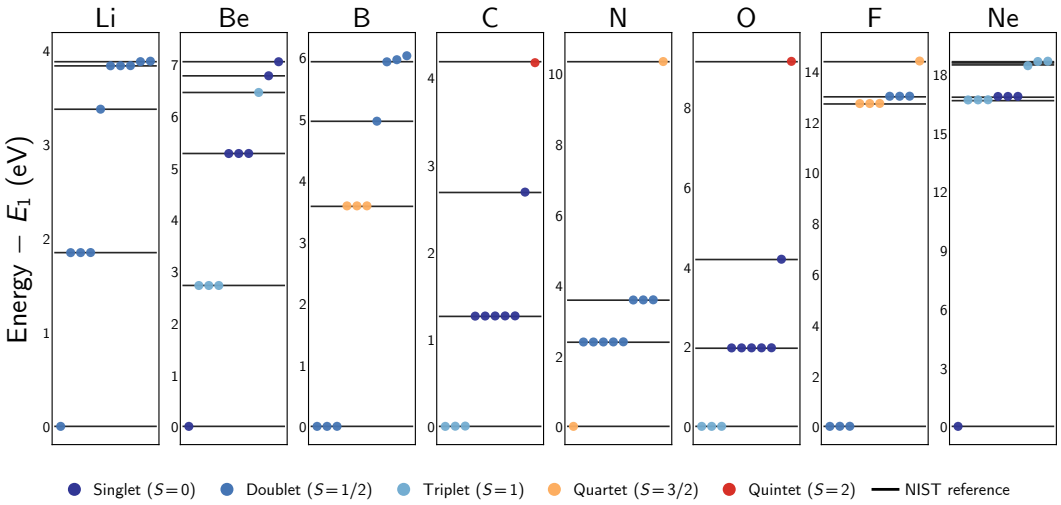


Figure 1: First-row K=10 energy-level diagram (Li–Ne). Per-atom recovered spectra (NestedLoRA-VMC) on top of NIST term centroids [32] (horizontal segments). Markers are colored by total spin S (per-network measured $\langle \hat{S}^2 \rangle$). Within each subplot, markers are placed left-to-right in order of energy, so states close in energy separate horizontally. For an atom with several seeds in Table 2, each marker is the mean over seeds of the energy and $\langle \hat{S}^2 \rangle$ of the state at that rank. Table 9 gives the term and configuration of each state.

4 Experiments

We evaluate NestedLoRA-VMC on the first-row atoms, Li through Ne, with the Psiformer architecture [38], for the lowest K=10 eigenstates of $\mathcal{H}$. Its excited-state energies are comparable in accuracy to those of NES-VMC [27] and the penalty-based method of Szabó et al. [37], against term centroids from the National Institute of Standards and Technology (NIST) (Section 4.2). On carbon, the per-network operator-shift of Section 3.3.3 brings the five networks of the degenerate $2p^2\ ^1D$ level close together in energy, whereas with $\mathcal{T} = -\mathcal{H}$ they stay spread over the level (Section 4.4).[5]

4.1 Setup

Ansatz and MCMC sampler. Every network is a Psiformer [38] and samples its own Born density with one Metropolis chain; Appendix H.2 gives the sampler settings.

Pretraining. For every atom we pretrain to a complete-active-space self-consistent-field (CASSCF) target with `aug-cc-pVTZ` as the underlying single-particle basis. The active space (N_e, N_o), N_e electrons in N_o orbitals, is chosen per atom to cover the K=10 states of its sectors (Table 6). The per-state CASSCF targets and the pretraining schedule are in Appendix H.5.

Spin sectors and spin penalty. To obtain pure spin states more accurately, we run NestedLoRA-VMC on $\mathcal{T} = -(\mathcal{H} + \hat{S}^2)$ so that each state has one total spin S. The numbers $n_\uparrow$ and $n_\downarrow$ of Section 2 fix $m_S = (n_\uparrow - n_\downarrow)/2$, so the *spin sector* of a given m_S holds every state with $S \geq |m_S|$ (cf. Pfau et al. [26], von Glehn et al. [38]). Added to $\mathcal{H}$, the spin penalty $\hat{S}^2$ raises each state by $S(S{+}1)$ and keeps the eigenfunctions, so the lowest states of a sector take its smallest spin $S = |m_S|$ (Appendix D.3.3). In training, each atom has one sector $m_S = S$ for each spin multiplicity $2S{+}1$ among its lowest K=10 states; the sectors, with fixed numbers of networks that sum to K=10, are trained in one run, and their states are merged by energy. Szabó et al. [37] use this penalty for every atom and fix m_S for each higher spin multiplicity, which gives the same sectors (Appendix H.4).

Reference data. A term ^{2S+1}L has total spin S and total orbital angular momentum L, with $L = 0, 1, 2$ written S, P, D; Table 6 gives the terms of each atom. We compare energies with term centroids

[5] Source code for NestedLoRA-VMC is available at `https://github.com/MinchanJeong/NestedLoRA-VMC`.

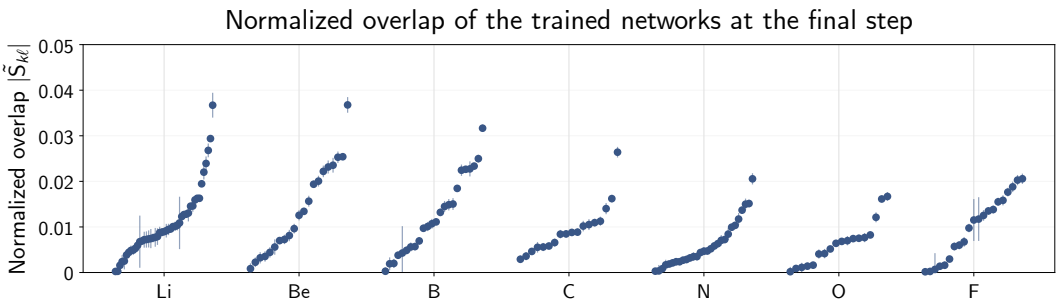


Figure 2: Normalized overlap $|\tilde{S}_{k\ell}|$ of every pair of networks within a spin sector at the final training step of the seed-1 run of each atom from Li to F, with the pairs of an atom ordered by $|\tilde{S}_{k\ell}|$. Each value comes from two long Metropolis chains (Appendix F.2.2), and Table 4 gives their convergence. Error bars are bootstrap standard errors (Appendix F.2.2).

from the NIST atomic-spectroscopy compilation of Sansonetti and Martin [32]. The centroid of a term is the mean energy of its fine-structure levels $J = |L - S|, \ldots, L + S$, each weighted by its degeneracy $2J+1$; spin–orbit coupling splits a term into these levels, and $\mathcal{H}$ gives them one energy. NIST gives each level as its energy above the ground level, so every comparison uses the excitation energy ΔE of each state above the ground-term centroid.

Training. Training runs are $100{,}000$ steps at an electron batch of $4{,}096$ with the Kronecker-factored approximate curvature (KFAC) optimizer [21], a natural-gradient approximation applied per state (Appendix H.3).

4.2 Excited-state energies on first-row atoms

The per-atom mean absolute error (MAE) of NestedLoRA-VMC in Table 2 is below chemical accuracy on every atom from Li to F and comparable to that of NES-VMC [27] and of the penalty-based method of Szabó et al. [37]. NES-VMC reaches its MAE with $200{,}000$ training steps per atom [27, Table S1], twice the $100{,}000$ of NestedLoRA-VMC and of Szabó et al. [37, SI Table S6].

Table 2: Per-atom mean absolute error (mHa) of $K{=}10$ excited-state energies against $(2J{+}1)$-weighted NIST term centroids [32]. An entry $m_{\pm s}$ is the mean m and standard deviation s over random seeds; the others are from one run.

Method	Li	Be	B	C	N	O	F	Ne
NES-VMC [27]	0.02	0.04	0.54	0.09	0.16	0.26	1.90	1.95
Szabó et al. [37] (Penalty-based)	1.06	0.20	0.63	0.59	2.95	1.43	1.40	2.86
NestedLoRA-VMC	$0.06_{\pm 0.01}$	$0.03_{\pm 0.01}$	$0.64_{\pm 0.23}$	$0.18_{\pm 0.01}$	0.28	0.27	0.65	$1.55_{\pm 0.56}$

Figure 1 shows the recovered $K{=}10$ spectra on each atom. All three methods are scored against the same NIST term centroids, with the NES-VMC and penalty errors computed from the per-state values published in their papers. Per-state energies of NestedLoRA-VMC are the mean of the per-walker local energy over the last $3{,}000$ training steps, evaluated with exact interparticle distances (Appendix J). The compute time available set the number of training seeds per atom: three for Li, Be and B, two for C and Ne, and one for N, O and F. Appendix J gives the scoring, the per-state errors of all eight atoms, and in Table 8 the 95% interval of each run's MAE.

4.3 Overlap of the trained networks

Exact eigenstates are mutually orthogonal, so we measure the normalized overlap of every pair of networks within a spin sector at the final training step as a surrogate for how well the networks learn the eigenstates, with the long chains of Appendix F.2.2 (Figure 2). The trained networks are close to orthogonal on every atom from Li to F.

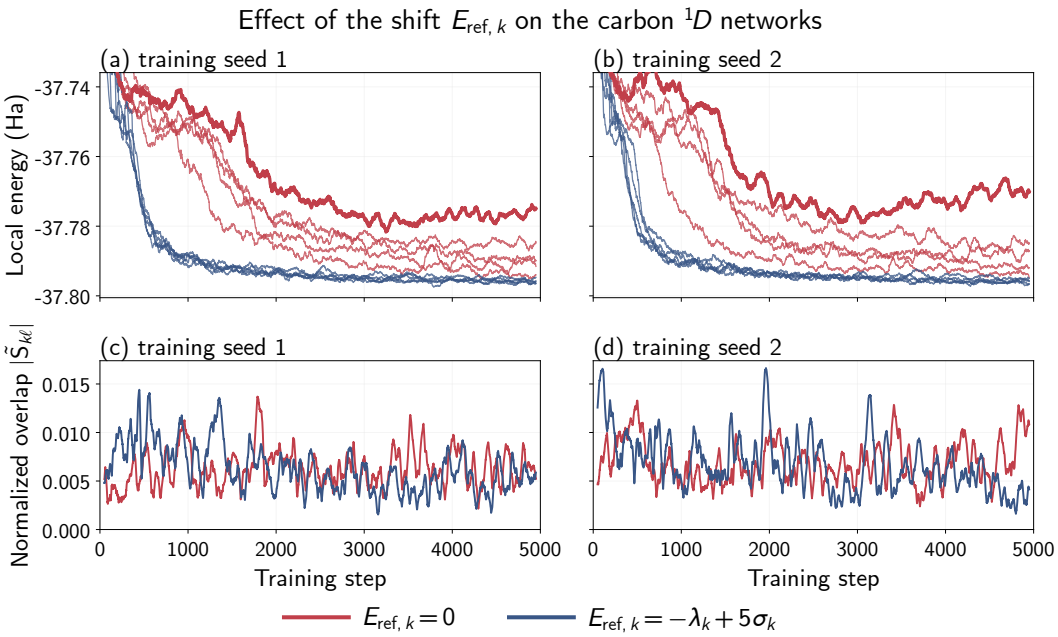


Figure 3: Carbon runs of Section 4.4, with $E_{\mathrm{ref},k} = 0$ or with the shift $E_{\mathrm{ref},k} = -\lambda_k + 5\sigma_k$, for training seed 1 (a, c) and seed 2 (b, d). (a, b) Batch mean of the clipped local energy of the five networks of the ^1D term; the thick line is the network furthest above the others with $E_{\mathrm{ref},k} = 0$. (c, d) Normalized overlap from the unclipped batch mean, averaged over the 10 pairs of these networks. Every curve is a mean over a sliding window of 100 steps, taken before the absolute value in (c, d).

4.4 On the effectiveness of the per-network operator-shift on carbon

On carbon we test whether the shift of Section 3.3.3 improves the energy convergence of the five networks of the $2p^2\ ^1D$ term, whose exact energies are equal. We train 5,000 steps from the pretrained start of the seed-1 carbon run of Table 2, with the setup of Section 4.1 and training seeds 1 and 2. For each seed one run uses $E_{\mathrm{ref},k} = 0$, that is $\mathcal{T} = -\mathcal{H}$, and one the shift $E_{\mathrm{ref},k} = -\lambda_k + 5\sigma_k$ (Figure 3). With $E_{\mathrm{ref},k} = 0$ the five networks stay spread over the term. With the shift their mean energies over the last 1,000 steps span less than a tenth as much.

5 Conclusion

We present NestedLoRA-VMC, a new approach to excited-state VMC based on nested low-rank approximation. Unlike penalty-based methods and NES-VMC, NestedLoRA-VMC exploits the structure of the spectral approximation problem to recover eigenstates in energy order while retaining per-network VMC optimization. This provides a principled alternative to existing approaches, with computational savings in Laplacian evaluations that become increasingly significant as the number of target states grows.

An important next step is to extend NestedLoRA-VMC to molecular systems and larger K. Its VMC-compatible formulation also opens opportunities to incorporate advanced optimization techniques, such as Wasserstein gradient flows [22] and improved discretizations of imaginary-time evolution [36].

Limitations. Empirical validation is limited to first-row atoms (Li–Ne) at K=10. For molecular systems, reference excitation energies at fixed nuclear geometries are best estimates whose uncertainty grows with system size, motivating evaluation criteria beyond agreement with reference values. Our multi-spin calculations also require preselecting spin sectors and allocating networks among them; automating this selection remains an open problem. Finally, NestedLoRA-VMC retains the $O({n_{\mathrm{el}}}^4)$ Laplacian cost of standard neural-network VMC, which remains a major bottleneck for larger molecules. Reducing this cost requires further mathematical and computational advances.

Acknowledgments and Disclosure of Funding

JJR and GWW were supported, in part, by the MIT-IBM Watson AI Lab under Agreement No. W1771646. MJ and SYY were partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST); No. RS-2024-00457882, National AI Research Lab Project). MJ was supported in part by Korea Institute for Advanced Study (KIAS) grant funded by the Korea government, and by the Center for Advanced Computation at KIAS.

Competing interests

All authors declare no competing interests.

References

[1] Zhijie Deng, Jiaxin Shi, and Jun Zhu. NeuralEF: Deconstructing kernels by deep neural networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proc. Int. Conf. Mach. Learn.*, volume 162 of *Proceedings of Machine Learning Research*, pages 4976–4992. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/deng22b.html`. Cited on p. 16.

[2] Andreas Dreuw, Jennifer L. Weisman, and Martin Head-Gordon. Long-range charge-transfer excited states in time-dependent density functional theory require non-local exchange. *J. Chem. Phys.*, 119(6): 2943–2946, 2003. doi: 10.1063/1.1590951. Cited on p. 16.

[3] M T Entwistle, Z Schätzle, P A Erdman, J Hermann, and F Noé. Electronic excited states in deep variational Monte Carlo. *Nat. Commun.*, 14(1):274, January 2023. ISSN 2041-1723. doi: 10.1038/s41467-022-35534-5. Cited on pp. 2, 3, 4, 16, and 17.

[4] William MC Foulkes, Lubos Mitas, RJ Needs, and Guna Rajagopal. Quantum Monte Carlo simulations of solids. *Rev. Mod. Phys.*, 73(1):33–83, 2001. Cited on p. 3.

[5] Nicholas Gao, Jonas Köhler, and Adam Foster. folx - forward laplacian for jax, 2025. URL `http://github.com/microsoft/folx`. Cited on p. 37.

[6] Nicholas Gao, Till Grutschus, Frank Noé, and Stephan Günnemann. Excited Pfaffians: Generalized neural wave functions across structure and state. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 33184–33213. PMLR, 2026. Cited on pp. 16 and 28.

[7] Weiguo Gao, Yingzhou Li, and Bichen Lu. Triangularized orthogonalization-free method for solving extreme eigenvalue problems. *J. Sci. Comput.*, 93(3):63, 2022. doi: 10.1007/s10915-022-02025-0. Cited on pp. 2 and 16.

[8] Douglas Hendry, Alessandro Sinibaldi, and Giuseppe Carleo. Grassmann variational Monte Carlo with neural wave functions. *Phys. Rev. B*, 114(13):134404, 2026. doi: 10.1103/l8mr-567g. Cited on p. 16.

[9] J. Hermann, P. B. Szabó, Z. Schätzle, M. Mezera, A. Foster, K. Maziarz, H. E. Sauceda, and M. Brockschmidt. DeepQMC, 2024. URL `https://github.com/deepqmc/deepqmc`. Cited on pp. 27, 29, 33, 34, 35, 37, and 41.

[10] Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.*, 12(10):891–897, October 2020. ISSN 1755-4330, 1755-4349. doi: 10.1038/s41557-020-0544-y. Cited on pp. 1 and 16.

[11] Jan Hermann, James Spencer, Kenny Choo, Antonio Mezzacapo, W M C Foulkes, David Pfau, Giuseppe Carleo, and Frank Noé. Ab initio quantum chemistry with neural-network wavefunctions. *Nat. Rev. Chem.*, 7(10):692–709, October 2023. ISSN 2397-3358,2397-3358. doi: 10.1038/s41570-023-00516-8. Cited on p. 16.

[12] W. Hunziker. On the spectra of Schrödinger multiparticle Hamiltonians. *Helvetica Physica Acta*, 39: 451–462, 1966. URL `https://www.osti.gov/biblio/4511685`. Cited on p. 19.

[13] Yassir Jedra and Devavrat Shah. k-SVD with gradient descent. *arXiv preprint arXiv:2502.00320*, 2025. URL https://arxiv.org/abs/2502.00320. Cited on pp. 6 and 16.

[14] Minchan Jeong, Jongha Jon Ryu, Se-Young Yun, and Gregory W. Wornell. Efficient parametric SVD of Koopman operator for stochastic dynamical systems. In *Adv. Neural Inf. Proc. Syst.*, 2025. URL https://openreview.net/forum?id=kL2pnzClyD. Cited on pp. 2 and 16.

[15] Adrien Kahn, Luca Gravina, and Filippo Vicentini. Variational subspace methods and application to improving variational Monte Carlo dynamics. *Quantum*, 10:2082, 2026. doi: 10.22331/q-2026-04-23-2082. Cited on p. 16.

[16] Tosio Kato. Fundamental properties of Hamiltonian operators of Schrödinger type. *Transactions of the American Mathematical Society*, 70(2):195–211, 1951. Cited on p. 19.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Repr.*, 2015. Cited on p. 16.

[18] Anna I. Krylov. Equation-of-motion coupled-cluster methods for open-shell and electronically excited species: The hitchhiker's guide to Fock space. *Annu. Rev. Phys. Chem.*, 59:433–462, 2008. doi: 10.1146/annurev.physchem.59.032607.093602. Cited on p. 16.

[19] Ruichen Li, Haotian Ye, Du Jiang, Xuelan Wen, Chuwei Wang, Zhe Li, Xiang Li, Di He, Ji Chen, Weiluo Ren, and Liwei Wang. A computational framework for neural network-based variational Monte Carlo with Forward Laplacian. *Nat. Mach. Intell.*, 6(2):209–219, February 2024. ISSN 2522-5839,2522-5839. doi: 10.1038/s42256-024-00794-x. Cited on pp. 7 and 16.

[20] Neepa T. Maitra, Fan Zhang, Robert J. Cave, and Kieron Burke. Double excitations within time-dependent density functional theory linear response. *J. Chem. Phys.*, 120(13):5932–5937, 2004. doi: 10.1063/1.1651060. Cited on p. 16.

[21] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/martens15.html. Cited on pp. 9, 16, and 35.

[22] Kirill Neklyudov, Jannes Nys, Luca Thiede, Juan Carrasquilla, Qiang Liu, Max Welling, and Alireza Makhzani. Wasserstein quantum Monte Carlo: a novel approach for solving the quantum many-body Schrödinger equation. In *Adv. Neural Inf. Proc. Syst.*, volume 36, pages 63461–63482, 2023. Cited on p. 10.

[23] Shivesh Pathak, Brian Busemeyer, João NB Rodrigues, and Lucas K Wagner. Excited states in variational Monte Carlo using a penalty method. *J. Chem. Phys.*, 154(3):034101, 2021. doi: 10.1063/5.0030949. Cited on pp. 2, 3, 16, and 17.

[24] Ruben Pauncz. *The Symmetric Group in Quantum Chemistry*. Mathematical Chemistry Series. CRC Press, Boca Raton, 1995. Cited on p. 18.

[25] David Pfau, Stig Petersen, Ashish Agarwal, David GT Barrett, and Kimberly L Stachenfeld. Spectral inference networks: Unifying deep and spectral learning. In *Int. Conf. Learn. Repr.*, 2019. Cited on p. 16.

[26] David Pfau, James S Spencer, Alexander 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(3):033429, September 2020. doi: 10.1103/PhysRevResearch.2.033429. Cited on pp. 1, 6, 7, 8, 16, 33, and 37.

[27] David Pfau, Simon Axelrod, Halvard Sutterud, Ingrid von Glehn, and James S Spencer. Accurate computation of quantum excited states with neural networks. *Science*, 385(6711):eadn0137, 2024. Cited on pp. 2, 3, 4, 6, 8, 9, 16, 17, 32, 36, 40, and 41.

[28] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics, Vol. IV: Analysis of Operators*. Academic Press, 1978. Cited on p. 19.

[29] Michael Reed and Barry Simon. *Methods of Modern Mathematical Physics, Vol. I: Functional Analysis*. Academic Press, revised and enlarged edition, 1980. Cited on p. 19.

[30] J. Jon Ryu, Xiangxiang Xu, Hasan Sabri Melihcan Erol, Yuheng Bu, Lizhong Zheng, and Gregory W. Wornell. Operator SVD with neural networks via nested low-rank approximation. In *Proc. Int. Conf. Mach. Learn.*, 2024. URL https://openreview.net/forum?id=qESG5HaaoJ. Cited on pp. 2, 4, 5, 16, and 22.

[31] J. Jon Ryu, Samuel Zhou, and Gregory W. Wornell. Revisiting orbital minimization method for neural operator decomposition. In *Adv. Neural Inf. Proc. Syst.*, 2025. Cited on pp. 5 and 16.

[32] Jean E. Sansonetti and William Clyde Martin. Handbook of basic atomic spectroscopic data. *Journal of Physical and Chemical Reference Data*, 34(4):1559–2259, 2005. doi: 10.1063/1.1800011. Cited on pp. 8, 9, 31, and 36.

[33] Zeno Schätzle, P Bernát Szabó, Matěj Mezera, Jan Hermann, and Frank Noé. DeepQMC: An open-source software suite for variational optimization of deep-learning molecular wave functions. *J. Chem. Phys.*, 159 (9):094108, 2023. doi: 10.1063/5.0157512. Cited on pp. 35 and 41.

[34] James S. Spencer, David Pfau, and FermiNet Contributors. FermiNet, 2020. URL `http://github.com/deepmind/ferminet`. Cited on pp. 32, 33, 35, 36, 37, and 38.

[35] John F. Stanton and Rodney J. Bartlett. The equation of motion coupled-cluster method. A systematic biorthogonal approach to molecular excitation energies, transition probabilities, and excited state properties. *J. Chem. Phys.*, 98(9):7029–7039, 1993. doi: 10.1063/1.464746. Cited on p. 16.

[36] Baiyu Su and Qiang Liu. Quadratic quantum variational Monte Carlo. In *Adv. Neural Inf. Proc. Syst.*, 2024. URL `https://openreview.net/forum?id=lDtABI541U`. Cited on p. 10.

[37] P Bernát Szabó, Zeno Schätzle, Michael T Entwistle, and Frank Noé. An improved penalty-based excited-state variational Monte Carlo approach with deep-learning ansatzes. *J. Chem. Theory Comput.*, 20(18): 7922–7935, August 2024. ISSN 1549-9618,1549-9626. doi: 10.1021/acs.jctc.4c00678. Cited on pp. 2, 3, 4, 6, 7, 8, 9, 16, 17, 32, 35, 36, 38, 40, and 41.

[38] Ingrid von Glehn, James S Spencer, and David Pfau. A self-attention ansatz for ab-initio quantum chemistry. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=xveTeHVlF7j`. Cited on pp. 1, 6, 7, 8, 16, 33, and 37.

[39] William A Wheeler, Kevin G Kleiner, and Lucas K Wagner. Ensemble variational Monte Carlo for optimization of correlated excited state wave functions. *Electron. Struct.*, 6(2):025001, 2024. Cited on pp. 2, 3, 16, and 17.

[40] G. M. Zhislin and E. L. Mandel'. Spectrum of molecular electron energy operators in spaces of functions of specified symmetry. *Theoretical and Mathematical Physics*, 1(2):227–232, 1969. doi: 10.1007/BF01028050. Cited on pp. 19 and 20.

[41] Grigorii M. Zhislin. A study of the spectrum of the Schrödinger operator for a system of several particles. *Trudy Moskovskogo Matematicheskogo Obshchestva*, 9:81–120, 1960. URL `http://mi.mathnet.ru/mmo97`. In Russian. Cited on p. 19.

Appendix

A Notation summary **15**

B Related work **16**

C Technical details for prior methods **16**
C.1 Natural-excited-states 16
C.2 Penalty-based methods 17

D Discussion on the mathematical formulation **18**
D.1 Mathematical preliminaries 18
D.2 The LoRA objective for an unbounded operator 20
D.3 Global minimizers of $\mathcal{J}_{\mathsf{LoRA}}$ 20
D.4 Fixed point of sequential nesting 23

E Deferred proofs **23**
E.1 Gradient expressions 24
E.2 Derivation of the analytic eigenvalue update in Eq. (6) 27

F Monte Carlo statistics **27**
F.1 Energy statistics 28
F.2 Overlap statistics 28

G Implementation details **31**
G.1 The per-network shifted operator 31
G.2 State ordering 32
G.3 Spin penalty 32

H Experimental setup details **32**
H.1 Architectures 33
H.2 Samplers 35
H.3 Natural-gradient descent with KFAC 35
H.4 Per-atom setup 35
H.5 Pretraining 35
H.6 NaN handling in training 36

I Per-step cost analysis **37**
I.1 Forward-Laplacian primitive (folx) 37
I.2 Cost decomposition: Hamiltonian coupling and ansatz parameterization 37
I.3 Empirical wall-time scaling on Ne 38

J Per-state errors across first-row atoms **38**

K Compute resources and asset licenses **41**

A Notation summary

Table 3 collects the symbols used throughout the body and appendix in one place. Each symbol is defined where it first appears.

Table 3: Notation summary.

Symbol	Meaning
Particle and walker variables	
n_{el}, n_{nu}	Number of electrons / nuclei in the molecule
K	Rank of the LoRA approximation; number of states, the ground state included
$\mathbf{r}_i$, $\mathbf{R}_I$	3D positions of the i-th electron, $i \in [n_{\mathsf{el}}]$, and of the I-th nucleus, $I \in [n_{\mathsf{nu}}]$
$\mathbf{x} = (\mathbf{r}_1, \dots, \mathbf{r}_{n_{\mathsf{el}}})$	Single-walker n_{el}-electron configuration
$\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^K)$	K electron configurations of NES-VMC, of which $\mathbf{x}^i$ is the i-th
Operators and spectrum	
$\mathscr{H}$	Hilbert space of wavefunctions antisymmetric under the exchange of two same-spin electrons
$\mathcal{H}$	Molecular electronic Hamiltonian
$\hat{S}^2$	Total spin-squared operator
$\mathcal{T} := -\mathcal{H}$ or $-(\mathcal{H} + \hat{S}^2)$	Negated Hamiltonian without and with spin penalty, target of HS approximation
$\mathcal{I}$ / I	Identity operator (script) / identity matrix (sans-serif)
$\sigma_{\mathrm{ess}}(\mathcal{T})$	Essential spectrum of $\mathcal{T}$
Wavefunction ansatzes and architecture parameters	
ψ_k	Per-state wavefunction ansatz, $k \in [K]$
f_k	Unnormalized rank-1 LoRA component
φ_k	Eigenfunction of $\mathcal{T}$
$\Psi(\mathbf{X}) \in \mathbb{C}^{K \times K}$	Matrix $[\psi_k(\mathbf{x}^i)]_{i,k=1}^K$
$\Phi(\mathbf{X}) := \det \Psi(\mathbf{X})$	NES-VMC total (determinantal) ansatz
θ	Neural-network parameters of the ansatz ψ_θ
Local energies and Gram-type matrices	
$\mathcal{E}_\theta(\mathbf{x})$, $\mathcal{E}_k(\mathbf{x})$	Local energy $(\mathcal{H}\psi_\theta)(\mathbf{x})/\psi_\theta(\mathbf{x})$ of the ansatz ψ_θ, and its per-state form $(\mathcal{H}\psi_k)(\mathbf{x})/\psi_k(\mathbf{x})$
$\mathsf{E}_{\mathsf{loc}}(\mathbf{X}) := \Psi(\mathbf{X})^{-1}(\mathcal{H}\Psi)(\mathbf{X})$	NES-VMC local energy matrix ($K \times K$)
$\mathsf{S}_{k\ell} := \langle \psi_k \mid \psi_\ell \rangle$	Overlap matrix ($K \times K$)
$\mathsf{H}_{k\ell} := \langle \psi_k \mid \mathcal{H}\psi_\ell \rangle$	Energy matrix ($K \times K$, NES-VMC)
Born densities, expectations, and statistics	
$p_k(\mathbf{x}) := \lvert\psi_k(\mathbf{x})\rvert^2 / \langle \psi_k \mid \psi_k \rangle$	Per-state Born density
$\mathbb{E}_{p_k}[\cdot]$	Expectation under the per-state Born density p_k
$\mathbb{E}_{\mathbf{X} \sim \lvert\Phi\rvert^2}[\cdot]$	Expectation under the joint NES-VMC sampling distribution
$\mathrm{Cov}_{p_k}(\cdot, \cdot)$	Covariance under p_k
σ_k	Walker standard deviation of the local energy of network k in a step
$E_{\mathrm{ref},k}$, c	Reference energy $-\lambda_k + c\,\sigma_k$ of network k, with margin c=5 (Section 3.3.3)
LoRA objective and nesting	
$\mathcal{L}_{\mathsf{LoRA}}(f_{1:K})$	HS rank-K LoRA objective in f variables
$\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:K}, \lambda_{1:K})$	Normalized LoRA objective in (ψ, λ) variables
λ_k	Per-state scale parameter ($= \mu_k$ at the nesting fixed point)
$\mathsf{R}^{\mathcal{O}}_{k\ell} := \langle \psi_k \mid \mathcal{O}\psi_\ell \rangle / \langle \psi_k \mid \psi_k \rangle$	Unified ratio functional, $\mathcal{O} \in \{\mathcal{T}, I\}$
$\tilde{\Delta}^{\mathcal{O}}_{k\ell}$	Score-coupled covariance (Appendix E.1.4)
$\mathsf{M}^{\mathsf{seq}}_{k\ell} := \mathbb{1}\{k > \ell\}$	Sequential nesting mask (strictly lower triangular)
$\odot$	Hadamard (entry-wise) product

B Related work

NestedLoRA-VMC is penalty-free like NES-VMC and needs K Laplacians per step like the penalty-based methods. It takes its objective from the neural-network eigensolver NestedLoRA and learns the states in energy order through sequential nesting (Section 3.2).

Ab initio methods for excited states. The accurate methods of ab initio quantum chemistry grow steeply in cost with the number n_{el} of electrons. Full configuration interaction expands ψ in Slater determinants and gives the ordered eigenstates exactly, but needs exponentially many determinants [26]. Coupled cluster with single, double and perturbative triple excitations, CCSD(T), costs $O({n_{\mathsf{el}}}^7)$ for the ground state [38], and equation-of-motion coupled cluster costs $O({n_{\mathsf{el}}}^6)$ or more for excited states [35, 18]. Time-dependent density functional theory is far cheaper, but its usual approximations fail for long-range charge-transfer excitations and miss double excitations [2, 20].

Neural-network VMC for the ground state. Neural-network VMC represents the wavefunction by a neural network and minimizes its energy with stochastic optimizers [17, 21], using samples of the Born density as in Section 2; Hermann et al. [11] review the field. On the ground states of small molecules, FermiNet [26], PauliNet [10] and Psiformer [38] reach chemical accuracy, an error below 1 kcal/mol or about 1.6 millihartree (mHa). With the forward Laplacian [19], each optimization step costs $O({n_{\mathsf{el}}}^4)$ for a fixed network size, against $O({n_{\mathsf{el}}}^7)$ for CCSD(T). NestedLoRA-VMC keeps this machinery for every state: each ψ_k is its own network and is sampled from its own Born density.

Excited-state VMC. Penalty-based methods [23, 3, 39, 37] train one ansatz per state on the sum of the energies plus pairwise overlap penalties, and the penalty weights set whether the minimum is the ordered eigenstates. Gao et al. [6] train one network for many excited states across molecular geometries on a penalty objective, and estimate the overlaps from samples pooled over all states.

NES-VMC [27] minimizes the energy of an extended system whose ansatz is the determinant of the K ansatzes evaluated on K configurations, and obtains the individual states by a diagonalization after training. The inverse of this $K \times K$ matrix enters its local energy, which therefore develops heavy tails when two ansatzes are nearly linearly dependent. Hendry et al. [8] write this objective as the energy of a K-dimensional subspace and derive a stochastic reconfiguration on the manifold of such subspaces, and Kahn et al. [15] derive the same formalism and apply it to time evolution. Section 2.2 describes the penalty-based methods and NES-VMC, and Appendix C states their objectives in full.

Neural-network eigensolvers. Outside VMC, several methods learn the leading eigenfunctions of an operator with neural networks and estimate inner products under a fixed data or proposal distribution. SpIN [25] maximizes a trace objective whitened by the covariance of the learned functions and orders them with a masked gradient. NeuralEF [1] solves a sequence of Rayleigh-quotient problems in which each function is penalized against the earlier ones. NestedLoRA [30] minimizes the error of a rank-K approximation of the operator, a polynomial in the inner products of the learned functions, and orders the eigenfunctions by nesting, which descends from the triangularized gradient of Gao et al. [7].

NestedLoRA-VMC brings the NestedLoRA objective to VMC: the reparameterization of Section 3.2 writes it as expectations under the Born densities of the ansatzes. NestedLoRA has been applied to low-dimensional partial differential equations and representation learning [30] and to the Koopman operators of stochastic dynamical systems [14]; Ryu et al. [31] apply nesting to the objective of the orbital minimization method, and Jedra and Shah [13] study k-SVD by gradient descent with the same low-rank-approximation principle.

C Technical details for prior methods

This section states the objectives of the prior excited-state methods that the main text compares with NestedLoRA-VMC, in the notation of this paper.

C.1 Natural-excited-states

We state the two forms of the natural-excited-states Rayleigh quotient in Eq. (2) and the expectation of the total local energy matrix. Let the matrix ansatz $\Psi(\mathbf{X}) \in \mathbb{C}^{K\times K}$ and the total ansatz $\Phi(\mathbf{X}) =$

$\det \Psi(\mathbf{X})$ be defined as in Section 2.2, and let the extended Hamiltonian be $\mathfrak{H} = \sum_{i=1}^{K} \mathfrak{H}_i$ with $\mathfrak{H}_i = I^{\otimes(i-1)} \otimes \mathcal{H} \otimes I^{\otimes(K-i)}$, which acts as $\mathcal{H}$ on the i-th configuration $\mathbf{x}^i$ and as the identity on the other $K-1$.

Theorem C.1. *Let* $\mathsf{S} := [\langle \psi_k \mid \psi_\ell \rangle]_{k,\ell=1}^{K}$ *be the overlap matrix and* $\mathsf{H} := [\langle \psi_k \mid \mathcal{H}\psi_\ell \rangle]_{k,\ell=1}^{K}$ *the energy matrix. Then*

$$\frac{\langle \Phi \mid \mathfrak{H}\Phi \rangle}{\langle \Phi \mid \Phi \rangle} = \operatorname{tr}(\mathsf{S}^{-1}\mathsf{H}).$$

Theorem C.2 ([27])**.**

$$\frac{\langle \Phi \mid \mathfrak{H}\Phi \rangle}{\langle \Phi \mid \Phi \rangle} = \mathbb{E}_{\mathbf{X}\sim|\Phi|^2}\left[\frac{\mathfrak{H}\Phi(\mathbf{X})}{\Phi(\mathbf{X})}\right] = \mathbb{E}_{\mathbf{X}\sim|\Phi|^2}[\operatorname{tr}(\mathsf{E}_{\mathsf{loc}}(\mathbf{X}))],$$

where the total local energy matrix *is* $\mathsf{E}_{\mathsf{loc}}(\mathbf{X}) := \Psi(\mathbf{X})^{-1}(\mathcal{H}\Psi)(\mathbf{X}) \in \mathbb{C}^{K\times K}$*, with* $(\mathcal{H}\Psi)(\mathbf{X}) := \big((\mathcal{H}\otimes\cdots\otimes\mathcal{H})\Psi\big)(\mathbf{X}) = [\,(\mathcal{H}\psi_k)(\mathbf{x}^i)\,]_{i,k=1}^{K}$.

Proposition C.1. *If* $\psi_{1:K}$ *are linearly independent, then* $\mathbb{E}_{\mathbf{X}\sim|\Phi|^2}[\mathsf{E}_{\mathsf{loc}}(\mathbf{X})] = \mathsf{S}^{-1}\mathsf{H}$*. Its trace gives* $\operatorname{tr}(\mathsf{S}^{-1}\mathsf{H}) = \mathbb{E}_{\mathbf{X}\sim|\Phi|^2}[\operatorname{tr}(\mathsf{E}_{\mathsf{loc}}(\mathbf{X}))]$*, the equality of the right-hand sides of Theorems C.1 and C.2.*

Proof. By Cramer's rule, $\Phi(\mathbf{X})\,[\mathsf{E}_{\mathsf{loc}}(\mathbf{X})]_{jk}$ is the determinant of $\Psi(\mathbf{X})$ with its j-th column replaced by $[(\mathcal{H}\psi_k)(\mathbf{x}^i)]_{i=1}^{K}$. The Andréief identity $\int \big(\det[f_b(\mathbf{x}^a)]_{a,b}\big)^{*} \det[g_b(\mathbf{x}^a)]_{a,b}\, d\mathbf{X} = K!\,\det[\langle f_a \mid g_b\rangle]_{a,b}$ then gives $\int \Phi^{*}(\mathbf{X})\,\Phi(\mathbf{X})\,[\mathsf{E}_{\mathsf{loc}}(\mathbf{X})]_{jk}\, d\mathbf{X} = K!\det \mathsf{S}^{(j,k)}$, where $\mathsf{S}^{(j,k)}$ is S with its j-th column replaced by the k-th column of H, and $\int |\Phi(\mathbf{X})|^2\, d\mathbf{X} = K!\det \mathsf{S}$, which is nonzero by linear independence. The ratio $\det \mathsf{S}^{(j,k)}/\det \mathsf{S}$ is $(\mathsf{S}^{-1}\mathsf{H})_{jk}$, again by Cramer's rule. □

C.2 Penalty-based methods

PauliNet penalty (Entwistle et al., 2023). Entwistle et al. [3] extended PauliNet to excited states with the divergent overlap penalty

$$\mathcal{L}_{\mathsf{penalty}}(\psi_{1:K}) = \sum_{k=1}^{K} \mathbb{E}_{p_k}[\mathcal{E}_k] + \alpha \sum_{i<j} \frac{1}{1-|\tilde{\mathsf{S}}_{ij}|}, \quad \tilde{\mathsf{S}}_{ij} := \frac{\langle \psi_i \mid \psi_j \rangle}{\|\psi_i\|\|\psi_j\|}.$$

The symmetrized overlap is estimated consistently by the product of expectations

$$\tilde{\mathsf{S}}_{ij} = \operatorname{sgn}\!\left(\mathbb{E}_{p_i}\!\left[\frac{\psi_j}{\psi_i}\right]\right) \times \sqrt{\mathbb{E}_{p_i}\!\left[\frac{\psi_j}{\psi_i}\right]\mathbb{E}_{p_j}\!\left[\frac{\psi_i}{\psi_j}\right]}.$$

Pathak–Wheeler form. Pathak et al. [23], Wheeler et al. [39] use a squared overlap that admits a direct unbiased estimator,

$$\mathcal{L}_{\mathsf{penalty}}(\psi_{1:K}) = \sum_{k=1}^{K} \mathbb{E}_{p_k}[\mathcal{E}_k] + \sum_{i<j} \alpha_{ij}\, |\tilde{\mathsf{S}}_{ij}|^2.$$

Wheeler et al. [39] established an ensemble variational principle for this family. Its objective carries strictly decreasing energy weights $w_1 > \cdots > w_K > 0$ on $\sum_i w_i \mathbb{E}_{p_i}[\mathcal{E}_i]$ and a single global penalty $\alpha \sum_{i<j} |\tilde{\mathsf{S}}_{ij}|^2$. The ordered bottom-$K$ eigenfunctions of $\mathcal{H}$ are then the global minimum once $\alpha > \alpha_c = \max_{i<j}\big[(E_j - E_i)\, w_i w_j/(w_i - w_j)\big]$, with E_i the eigenvalues the optimum identifies. The strictly decreasing weights break the rotational degeneracy of the bottom-K subspace: at equal weights $\alpha_c \to \infty$ and the minimum is the subspace, not its ordered basis. Because α_c scales with the spectral gaps $E_j - E_i$, the penalty schedule is coupled to the spectrum the optimization computes. The per-pair α_{ij} form displayed above is the practical variant; Szabó et al. [37] scale α_{ij} to running gap estimates (below).

Szabó et al. [37] adopted the Pathak–Wheeler form with three modifications. First, the orthogonality constraint is applied with respect to lower-lying states only: gradients of the overlap term flow only

into ψ_j for $j > i$, while ψ_i is detached from the computational graph (their Section 2.1.2). Second, an explicit per-state spin penalty $\beta \sum_i \langle \hat{S}^2 \rangle_i$ (their Eq. 7) selects the target spin sector at training time; Appendix G.3 gives the corresponding operator shift in our framework, and Appendix H.4 the spin sectors that their setting implies. Third, α_{ij} is set at every step to (their Eq. 17)

$$\alpha_{ij} = \tilde{\alpha} \cdot \max\big\{|\mathrm{ema}(\hat{E}_j - \hat{E}_i)|,\ \mathrm{ema}(\mathrm{std}_{p_i}(\mathcal{E}_i)),\ 0.001\big\}\, \mathbb{1}\{i < j\},$$

where $\hat{E}_i$ is the batch mean of $\mathcal{E}_i$, $\mathrm{std}_{p_i}(\mathcal{E}_i)$ is its batch standard deviation, and $\mathrm{ema}(\,\cdot\,)$ is an EMA over training steps. This reduces the user-tunable hyperparameter to a single global scale $\tilde{\alpha}$ (default $\tilde{\alpha} = 4$, with $\tilde{\alpha} = 2$ for Li, Be and $\tilde{\alpha} = 1$ for benzene per their SI Table S6).

D Discussion on the mathematical formulation

The LoRA objective of NestedLoRA-VMC is well defined for the molecular Hamiltonian, its global minimizers span the lowest K eigenstates, and the fixed point of sequential nesting gives these eigenstates in energy order. The first claim holds for any K trial functions in $\mathscr{H}$, because the objective of the unbounded $\mathcal{T}$ is the Hilbert–Schmidt objective of $\mathcal{T}$ projected onto their span, an operator of finite rank (Appendix D.2). The other two claims hold when the lowest K energies of $\mathcal{H}$ lie strictly below its essential spectrum, which we assume in Appendix D.1.4, and are shown in Appendices D.3 and D.4.

D.1 Mathematical preliminaries

The three claims are statements about $\mathcal{H}$ on the spatial space $\mathscr{H}$, and this subsection gives the mathematical facts about $\mathscr{H}$ and the spectrum of $\mathcal{H}$ that their proofs use.

D.1.1 Reduction to the spatial space $\mathscr{H}$

The eigenstates of the electrons with $n_\uparrow$ spins up correspond one-to-one to the eigenfunctions of $\mathcal{H}$ in $\mathscr{H}$, with the same energies, and their levels are those of total spin $S \geq |m_S|$, where $m_S = (n_\uparrow - n_\downarrow)/2$.

The space $\mathscr{H}$ is the subspace of $L^2(\mathbb{R}^{3n_{\mathsf{el}}})$ of functions of the electron positions $\mathbf{x} = (\mathbf{r}_1, \ldots, \mathbf{r}_{n_{\mathsf{el}}})$ that are antisymmetric under the exchange of two electrons of the same spin, where electrons $1, \ldots, n_\uparrow$ have spin up and the other $n_\downarrow = n_{\mathsf{el}} - n_\uparrow$ have spin down. The antisymmetrizer over the permutations within each spin is an orthogonal projection with range $\mathscr{H}$, so $\mathscr{H}$ is closed and is a Hilbert space with the inner product of $L^2(\mathbb{R}^{3n_{\mathsf{el}}})$. Our network architecture restricts the functions it can represent to $\mathscr{H}$ by construction (Appendix H.1).

An electronic state (*state* in this section) is a function of the positions and spins of the electrons, an element of $L^2(\mathbb{R}^{3n_{\mathsf{el}}}) \otimes (\mathbb{C}^2)^{\otimes n_{\mathsf{el}}}$ that changes sign when two electrons exchange both position and spin. The states with $n_\uparrow$ spins up correspond one-to-one to the functions in $\mathscr{H}$, and the correspondence is a linear isomorphism: it maps $\psi \in \mathscr{H}$ to the antisymmetrization Ψ, over all electrons, of ψ times the spin function with electrons $1, \ldots, n_\uparrow$ up. Since $\mathcal{H}$ acts on the positions alone and is symmetric in the electrons, it acts on states as $\mathcal{H} \otimes I$, which commutes with this antisymmetrization, so $(\mathcal{H} \otimes I)\Psi = E\Psi$ exactly when $\mathcal{H}\psi = E\psi$.

The total spin operators $\hat{S}^2$, with eigenvalues $S(S{+}1)$, and $\hat{S}_z$ act on the spins of the states, and the states with $n_\uparrow$ spins up form the eigenspace of $\hat{S}_z$ with eigenvalue m_S. Both commute with $\mathcal{H} \otimes I$, so each of its eigenspaces is spanned by multiplets of total spin S, each consisting of $2S + 1$ states with $\hat{S}_z$ eigenvalues $-S, \ldots, S$ [24, Sections 5.2–5.3]. A multiplet contains one state with $n_\uparrow$ spins up exactly when $S \geq |m_S|$, so the states with $n_\uparrow$ spins up form the orthogonal sum of their subspaces of total spin $S = |m_S|, |m_S| + 1, \ldots, n_{\mathsf{el}}/2$, each left invariant by $\mathcal{H} \otimes I$. Through the isomorphism, $\mathscr{H} = \bigoplus_{S \geq |m_S|} \mathscr{H}^{(S)}$, where $\mathscr{H}^{(S)}$ corresponds to total spin S and is left invariant by $\mathcal{H}$, and the eigenvalues of $\mathcal{H}$ in $\mathscr{H}$ are the energies of total spin $S \geq |m_S|$, with one eigenfunction per multiplet.

Example: lithium. The antisymmetry that defines $\mathscr{H}$ excludes functions with energy below the physical ground state. Consider lithium with m_S fixed at $1/2$, that is, $n_\uparrow = 2$ and $n_\downarrow = 1$. The spatial function $1s(\mathbf{r}_1)\, 1s(\mathbf{r}_2)\, 1s(\mathbf{r}_3)$ of three hydrogenic orbitals is symmetric in the three electrons

and, for a suitable orbital exponent, has an energy below the physical ground state; it lies outside $\mathscr{H}$, because it is symmetric under the exchange of the two spin-up electrons.

D.1.2 Self-adjoint operators and their spectrum

The min–max values of a self-adjoint operator $\mathcal{T}$ bounded above list its eigenvalues above the essential spectrum from the top, and the paragraphs below define these terms for a general $\mathcal{T}$.

Self-adjoint operators. A linear operator $\mathcal{T}$ on a dense domain $D(\mathcal{T})$ of a Hilbert space is *bounded* if $\|\mathcal{T}f\| \le C\|f\|$ for a constant C and all $f \in D(\mathcal{T})$. It is *symmetric* if $\langle f \mid \mathcal{T}g\rangle = \langle \mathcal{T}f \mid g\rangle$ for all $f, g \in D(\mathcal{T})$, and *self-adjoint* if, in addition, its adjoint has the same domain, $D(\mathcal{T}^\dagger) = D(\mathcal{T})$.

Spectrum. For a self-adjoint $\mathcal{T}$, the spectrum $\sigma(\mathcal{T})$ is the set of $z \in \mathbb{R}$ for which $\mathcal{T} - z$ has no bounded inverse on the Hilbert space. The discrete spectrum $\sigma_{\mathrm{disc}}(\mathcal{T})$ is the set of isolated eigenvalues of finite multiplicity, and the essential spectrum is $\sigma_{\mathrm{ess}}(\mathcal{T}) := \sigma(\mathcal{T}) \setminus \sigma_{\mathrm{disc}}(\mathcal{T})$.

Spectral theorem. By the spectral theorem in projection-valued-measure form [29, Theorem VIII.6], a self-adjoint $\mathcal{T}$ satisfies $\mathcal{T} = \int z\, dP_{\mathcal{T}}(z)$, where the spectral projection $P_{\mathcal{T}}(B)$ of $\mathcal{T}$ onto a Borel set $B \subset \mathbb{R}$ is an orthogonal projection. It vanishes when B is disjoint from $\sigma(\mathcal{T})$, and $P_{\mathcal{T}}(\{z\})$ projects onto the eigenspace of an eigenvalue z [29, pp. 236 and 264, Theorem VIII.5(e)]. For $\psi \in D(\mathcal{T})$ this gives $\langle \psi \mid \mathcal{T}\psi\rangle = \int z\, d\nu_\psi(z)$ with the spectral measure $\nu_\psi(B) := \|P_{\mathcal{T}}(B)\psi\|^2$.

Min–max values. For a self-adjoint $\mathcal{T}$ bounded above and $j \ge 1$, the min–max values

$$\mu_j := \sup_{\mathcal{V} \subset D(\mathcal{T}),\, \dim \mathcal{V} = j} \; \min_{\psi \in \mathcal{V},\, \|\psi\| = 1} \langle \psi \mid \mathcal{T}\psi\rangle \tag{7}$$

satisfy $\mu_1 \ge \mu_2 \ge \cdots$. By the min–max principle for $-\mathcal{T}$, which is bounded below [28, Theorem XIII.1], the μ_j greater than $\sup \sigma_{\mathrm{ess}}(\mathcal{T})$ are the eigenvalues of $\mathcal{T}$ greater than this value, in decreasing order and counted with multiplicity, and every other μ_j equals $\sup \sigma_{\mathrm{ess}}(\mathcal{T})$.

D.1.3 The molecular Hamiltonian on $\mathscr{H}$

The molecular Hamiltonian $\mathcal{H}$ is self-adjoint on $\mathscr{H}$, so it has a real spectrum.

The non-relativistic electronic Coulomb Hamiltonian (cf. Section 2)

$$\mathcal{H} = -\frac{1}{2}\sum_{i=1}^{n_{\mathsf{el}}} \nabla_{\mathbf{r}_i}^2 - \sum_{i=1}^{n_{\mathsf{el}}}\sum_{I=1}^{n_{\mathsf{nu}}} \frac{Z_I}{\|\mathbf{r}_i - \mathbf{R}_I\|} + \sum_{1 \le i < j \le n_{\mathsf{el}}} \frac{1}{\|\mathbf{r}_i - \mathbf{r}_j\|}$$

is unbounded but self-adjoint on $L^2(\mathbb{R}^{3n_{\mathsf{el}}})$ with the Sobolev space $H^2(\mathbb{R}^{3n_{\mathsf{el}}})$ as its dense domain, by the Kato–Rellich theorem [16], since the Coulomb potential satisfies the Kato smallness condition with respect to the kinetic operator.

The restriction of $\mathcal{H}$ to $D(\mathcal{H}) := H^2(\mathbb{R}^{3n_{\mathsf{el}}}) \cap \mathscr{H}$ is self-adjoint on $\mathscr{H}$, because $\mathcal{H}$ commutes with the exchange of two same-spin electrons and the antisymmetrizer maps $H^2(\mathbb{R}^{3n_{\mathsf{el}}})$ into itself, which also makes $D(\mathcal{H})$ dense in $\mathscr{H}$. We work with $\mathcal{T} = -\mathcal{H}$ on $D(\mathcal{T}) = D(\mathcal{H})$ throughout, and every function on which $\mathcal{T}$ acts is taken in $D(\mathcal{T})$ without further mention.

D.1.4 Spectrum and bound states

On $\mathscr{H}$, the spectrum of $\mathcal{H}$ consists of isolated eigenvalues of finite multiplicity below the ionization threshold Σ and the essential spectrum $\sigma_{\mathrm{ess}}(\mathcal{H}) = [\Sigma, \infty)$ above it. On each $\mathscr{H}^{(S)}$, the Hunziker–van Winter–Zhislin (HVZ) theorem [12, 41, 28] in its form for a fixed total spin [40] gives the essential spectrum $[\Sigma_S, \infty)$, where Σ_S is the lowest energy of the $(n_{\mathsf{el}}-1)$-electron system with total spin $S - \frac{1}{2}$ or $S + \frac{1}{2}$. Since $\mathscr{H}$ is the finite orthogonal sum of the $\mathscr{H}^{(S)}$, each left invariant by $\mathcal{H}$ (Appendix D.1.1), the essential spectrum on $\mathscr{H}$ is the union of these intervals, $[\Sigma, \infty)$ with $\Sigma = \min_S \Sigma_S$, and below Σ the discrete spectrum is the union of those on the $\mathscr{H}^{(S)}$. The threshold Σ is the lower of the ground-state energies of the two $(n_{\mathsf{el}}-1)$-electron systems obtained by removing one spin-up or one spin-down electron to infinity.

The eigenfunctions of $\mathcal{H}$ in $\mathscr{H}$ are the bound states, and they span only part of $\mathscr{H}$. The energies of the bound states below Σ form the discrete spectrum, and these states are localized near the nuclei. The essential spectrum $[\Sigma, \infty)$ contains the energies of scattering states, which have at least one electron at infinity and whose solutions of $\mathcal{H}\psi = E\psi$ lie outside $\mathscr{H}$. It can also contain eigenvalues: a bound state of total spin S with energy between Σ and Σ_S is discrete on $\mathscr{H}^{(S)}$ and embedded in the essential spectrum on $\mathscr{H}$. The orthogonal complement of the closed span of the bound states, the continuous spectral subspace, holds square-integrable states such as a wave packet of an electron moving away from the nuclei. The orthogonal complement of the bound states below Σ, the essential spectral subspace, contains the continuous spectral subspace and the bound states with energies from Σ upward.

We assume that $\mathcal{T} = -\mathcal{H}$ has at least K discrete eigenvalues separated from the essential spectrum:

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K > \sup \sigma_{\text{ess}}(\mathcal{T}). \tag{8}$$

Since $\sup \sigma_{\text{ess}}(\mathcal{T}) = -\Sigma$, this is the existence of at least K bound states below Σ, which holds for all systems in our experiments. The number of bound states below Σ depends on the system; neutral atoms and positive ions have infinitely many by Zhislin's theorem [40].

D.2 The LoRA objective for an unbounded operator

The LoRA objective of any trial functions $f_1, \ldots, f_K$ is the Hilbert–Schmidt objective of the projected operator $\mathcal{T}_{\mathcal{V}} := P_{\mathcal{V}}\, \mathcal{T}\, P_{\mathcal{V}}$, an operator of finite rank, where $P_{\mathcal{V}}$ is the orthogonal projection onto $\mathcal{V} := \operatorname{span}\{f_1, \ldots, f_K\}$; this holds also for the unbounded molecular $\mathcal{T}$.

Hilbert–Schmidt operators. A bounded operator on $\mathscr{H}$ is *Hilbert–Schmidt* (HS) if $\|\mathcal{T}\|_{\mathsf{HS}}^2 := \sum_n \|\mathcal{T} e_n\|^2 < \infty$ for an orthonormal basis (e_n) of $\mathscr{H}$, a sum that does not depend on the basis, and *finite-rank* if its range is finite-dimensional; a finite-rank operator is HS.

The body's identity $\mathcal{L}_{\mathsf{LoRA}}(f_{1:K}) = \|\mathcal{T} - \sum_k |f_k\rangle\langle f_k|\|_{\mathsf{HS}}^2 - \|\mathcal{T}\|_{\mathsf{HS}}^2$ assumes $\mathcal{T}$ to be HS, which fails for molecular $\mathcal{T}$. The objective itself is defined for any $f_1, \ldots, f_K$, since the matrix elements $\langle f_k \mid \mathcal{T} f_\ell \rangle$ are finite. Like the Rayleigh quotient, which evaluates $\mathcal{T}$ on its trial function alone, the objective depends on $\mathcal{T}$ only through these matrix elements, that is, through $\mathcal{T}$ on $\mathcal{V}$. The projected operator $\mathcal{T}_{\mathcal{V}}$ is finite-rank, hence HS, and has the same matrix elements, $\langle f_k \mid \mathcal{T} f_\ell \rangle = \langle f_k \mid \mathcal{T}_{\mathcal{V}} f_\ell \rangle$, so the identity holds exactly for $\mathcal{T}_{\mathcal{V}}$:

$$\mathcal{L}_{\mathsf{LoRA}}(f_{1:K}) = \left\| \sum_{k=1}^{K} |f_k\rangle\langle f_k| - \mathcal{T}_{\mathcal{V}} \right\|_{\mathsf{HS}}^2 - \|\mathcal{T}_{\mathcal{V}}\|_{\mathsf{HS}}^2. \tag{9}$$

Within a fixed $\mathcal{V}$, minimizing $\mathcal{L}_{\mathsf{LoRA}}$ is therefore the best positive-semidefinite (PSD) rank-K approximation of $\mathcal{T}_{\mathcal{V}}$, and the global minimization also moves $\mathcal{V}$.

D.3 Global minimizers of $\mathcal{J}_{\mathsf{LoRA}}$

Theorem D.1 (Global minimizers of $\mathcal{J}_{\mathsf{LoRA}}$)**.** *Let $\mathcal{T}$ be self-adjoint and bounded above on $\mathscr{H}$, with the min–max values μ_j of Eq.* (7)*, and assume*

$$\mu_K > \max\bigl(\sup \sigma_{\text{ess}}(\mathcal{T}),\, 0\bigr). \tag{10}$$

Then $\mu_1 \geq \cdots \geq \mu_K$ are the top K eigenvalues of $\mathcal{T}$, counted with multiplicity, and

$$\min_{\psi_{1:K} \neq 0,\ \lambda_{1:K} \in \mathbb{R}^K_{\geq 0}} \mathcal{J}_{\mathsf{LoRA}}(\psi_{1:K}, \lambda_{1:K}) = -\sum_{k=1}^{K} \mu_k^2.$$

Every minimizer satisfies

$$\sum_{k=1}^{K} \lambda_k \frac{|\psi_k\rangle\langle\psi_k|}{\langle \psi_k \mid \psi_k \rangle} = \sum_{k=1}^{K} \mu_k |\varphi_k\rangle\langle\varphi_k|$$

for some orthonormal $\varphi_{1:K}$ with $\mathcal{T}\varphi_k = \mu_k \varphi_k$. If moreover $\mu_K > \mu_{K+1}$, this operator is the same for every minimizer: it is the weighted spectral projector of $\mathcal{T}$ onto its top-K eigenspace.

The min–max formula defines μ_j for every self-adjoint $\mathcal{T}$ bounded above, including the operator with the spin penalty of Appendix D.3.3, so Eq. (10) can be stated for any such $\mathcal{T}$; the first claim is the min–max principle of Appendix D.1.2, since $\mu_K > \sup \sigma_{\text{ess}}(\mathcal{T})$. For $\mathcal{T} = -\mathcal{H}$, $\sup \sigma_{\text{ess}}(\mathcal{T}) = -\Sigma$ with the ionization threshold $\Sigma \leq 0$ of Appendix D.1.4, so Eq. (10) reduces to Eq. (8).

D.3.1 Two lemmas

The proof uses two lemmas, with $\mathcal{T}$ and μ_j as in Theorem D.1 and Eq. (10) assumed. The first is the lower bound, through the projected operator $\mathcal{T}_\mathcal{V}$ of Appendix D.2. For a finite-rank self-adjoint S, write S_+ for its positive part, and $(x)_+ := \max(x, 0)$.

Lemma D.1 (Lower bound)**.** *Assume Eq.* (10)*. For $f_1, \dots, f_K$, let*

- $\mathcal{V} = \operatorname{span}\{f_{1:K}\}$ *and* $d = \dim \mathcal{V}$*,*
- $A_\mathcal{V} = \sum_k |f_k\rangle\langle f_k|$*, and*
- $\omega_1 \geq \cdots \geq \omega_d$ *the eigenvalues of $\mathcal{T}_\mathcal{V}$ on $\mathcal{V}$.*

Then

$$\mathcal{L}_{\mathsf{LoRA}}(f_{1:K}) \geq -\sum_{j=1}^{d} (\omega_j)_+^2 \geq -\sum_{k=1}^{K} \mu_k^2.$$

The first inequality is an equality only if $A_\mathcal{V} = (\mathcal{T}_\mathcal{V})_+$. The second is an equality only if $d = K$ and $\omega_j = \mu_j$ for every j.

Proof. Consider

$$\begin{aligned}
\mathcal{L}_{\mathsf{LoRA}}(f_{1:K}) &\overset{(a)}{=} \|\mathcal{T}_\mathcal{V} - A_\mathcal{V}\|_{\mathsf{HS}}^2 - \|\mathcal{T}_\mathcal{V}\|_{\mathsf{HS}}^2 \\
&\overset{(b)}{\geq} \min_{A \succeq 0,\, \operatorname{rank}(A) \leq K} \|\mathcal{T}_\mathcal{V} - A\|_{\mathsf{HS}}^2 - \|\mathcal{T}_\mathcal{V}\|_{\mathsf{HS}}^2 \\
&\overset{(c)}{=} \|\mathcal{T}_\mathcal{V} - (\mathcal{T}_\mathcal{V})_+\|_{\mathsf{HS}}^2 - \|\mathcal{T}_\mathcal{V}\|_{\mathsf{HS}}^2 \overset{(d)}{=} -\|(\mathcal{T}_\mathcal{V})_+\|_{\mathsf{HS}}^2 \overset{(e)}{=} -\sum_{j=1}^{d} (\omega_j)_+^2 \\
&\overset{(f)}{\geq} -\sum_{k=1}^{K} \mu_k^2.
\end{aligned}$$

The steps hold for the following reasons.

- (a) is Eq. (9).
- (b) holds since $A_\mathcal{V}$ is PSD of rank $\leq K$.
- (c) holds since $\mathcal{T}_\mathcal{V}$ has rank $\leq K$, so the rank constraint is inactive and the best PSD approximant is $(\mathcal{T}_\mathcal{V})_+$ (a PSD A cannot match the negative eigenvalues of $\mathcal{T}_\mathcal{V}$).
- (d) and (e) are direct calculations using $\mathcal{T}_\mathcal{V} - (\mathcal{T}_\mathcal{V})_+ = (\mathcal{T}_\mathcal{V})_- \perp (\mathcal{T}_\mathcal{V})_+$.
- (f) is due to the Courant–Fischer min–max principle, i.e., $\omega_j \leq \mu_j$, and monotonicity under $(\cdot)_+^2$ since $\mu_j > 0$.

The nearest PSD operator in Hilbert–Schmidt norm is unique, so (b) is tight only if $A_\mathcal{V} = (\mathcal{T}_\mathcal{V})_+$. Since $(\omega_j)_+^2 \leq \mu_j^2$ for each $j \leq d$ and $\mu_j > 0$ for every $j \leq K$, (f) is tight only if $d = K$ and $\omega_j = \mu_j$ for every j. □

Equality in Lemma D.1 fixes the eigenvalues of $\mathcal{T}_\mathcal{V}$ on $\mathcal{V}$, and the second lemma is the induction step of the proof of the theorem: given eigenfunctions $\varphi_1, \dots, \varphi_{j-1}$, it turns a vector of the subspace W_j of Lemma D.2 that attains μ_j into an eigenfunction of $\mathcal{T}$.

Lemma D.2 (Orthogonal complement)**.** *Let $1 \leq j \leq K$, let $\varphi_1, \dots, \varphi_{j-1}$ be orthonormal with $\mathcal{T}\varphi_i = \mu_i \varphi_i$, and let W_j be the orthogonal complement of their span. Then $\langle \psi \mid \mathcal{T}\psi \rangle \leq \mu_j$ for every unit $\psi \in W_j \cap D(\mathcal{T})$, and a unit $\varphi \in W_j \cap D(\mathcal{T})$ with $\langle \varphi \mid \mathcal{T}\varphi \rangle = \mu_j$ satisfies $\mathcal{T}\varphi = \mu_j \varphi$.*

Proof. Let $Q = P_\mathcal{T}\big((\mu_j, \infty)\big)$; we show that $Q\psi = 0$ for $\psi \in W_j$. Eq. (10) and $j \leq K$ give $\mu_j \geq \mu_K > \sup \sigma_{\mathrm{ess}}(\mathcal{T})$, so every point of the spectrum of $\mathcal{T}$ in (μ_j, ∞) lies outside $\sigma_{\mathrm{ess}}(\mathcal{T})$ and

is an isolated eigenvalue of finite multiplicity. By the min–max principle of Appendix D.1.2, an eigenvalue $z > \mu_j$ of multiplicity m appears m times in the list $\mu_1 \geq \mu_2 \geq \cdots$, at indices $i < j$. The m vectors φ_i with these indices are orthonormal and satisfy $\mathcal{T}\varphi_i = z\varphi_i$ by the hypothesis of the lemma, so they span the m-dimensional eigenspace of z. Since $P_{\mathcal{T}}$ vanishes on the gaps between these eigenvalues, Q is the sum of their spectral projections, so its range is the sum of their eigenspaces, which lies in the span of $\varphi_1, \ldots, \varphi_{j-1}$ and is orthogonal to W_j. For a unit $\psi \in W_j \cap D(\mathcal{T})$ with spectral measure ν_ψ under $\mathcal{T}$, consider

$$\langle\psi|(\mu_j - \mathcal{T})\psi\rangle \overset{(a)}{=} \int (\mu_j - z)\, d\nu_\psi(z) \overset{(b)}{=} \int_{(-\infty,\mu_j]} (\mu_j - z)\, d\nu_\psi(z) \overset{(c)}{\geq} 0.$$

The steps hold for the following reasons.

- (a) is the spectral theorem, with $\nu_\psi(\mathbb{R}) = \|\psi\|^2 = 1$.
- (b) holds since $\nu_\psi\big((\mu_j, \infty)\big) = \|Q\psi\|^2 = 0$, as $\psi \in W_j$ is orthogonal to the range of Q.
- (c) holds since $\mu_j - z \geq 0$ for $z \leq \mu_j$.

Equality in (c) holds exactly when ν_ψ is concentrated at μ_j, that is, when $\mathcal{T}\psi = \mu_j\psi$. □

D.3.2 Proof of Theorem D.1

Proof of Theorem D.1. Write $f_k = \sqrt{\lambda_k}\,\psi_k/\|\psi_k\|$, so that $\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:K}, \lambda_{1:K}) = \mathcal{L}_{\mathsf{LoRA}}(f_{1:K})$ by the definition of $\mathcal{J}_{\mathsf{LoRA}}$ in Section 3.2. The operator of the theorem is then $\sum_k \lambda_k \frac{|\psi_k\rangle\langle\psi_k|}{\langle\psi_k \mid \psi_k\rangle} = \sum_k |f_k\rangle\langle f_k| = A_{\mathcal{V}}$, with $A_{\mathcal{V}}$ as in Lemma D.1.

Minimum. Lemma D.1 gives $\mathcal{L}_{\mathsf{LoRA}}(f_{1:K}) \geq -\sum_k \mu_k^2$, and $f_k = \sqrt{\mu_k}\,\varphi_k$ for orthonormal eigenfunctions $\mathcal{T}\varphi_k = \mu_k\varphi_k$, given by $\psi_k = \varphi_k$ and $\lambda_k = \mu_k$, attains it, so a minimizer exists.

Form of every minimizer. Let $(\psi_{1:K}, \lambda_{1:K})$ be a minimizer, so that its $f_{1:K}$ attains the minimum and both inequalities of Lemma D.1 are tight: $\dim \mathcal{V} = K$, $\mathcal{T}_{\mathcal{V}}$ has the eigenvalues $\mu_1, \ldots, \mu_K$ on $\mathcal{V}$, and $A_{\mathcal{V}} = (\mathcal{T}_{\mathcal{V}})_+$. We show that $A_{\mathcal{V}} = \mathcal{T}_{\mathcal{V}}$. Let $\varphi_{1:K}$ be an orthonormal basis of $\mathcal{V}$ with $\mathcal{T}_{\mathcal{V}}\varphi_j = \mu_j\varphi_j$; we show $\mathcal{T}\varphi_j = \mu_j\varphi_j$ by induction on j. If this holds for $\varphi_{1:j-1}$, then φ_j is a unit vector in W_j with $\langle\varphi_j \mid \mathcal{T}\varphi_j\rangle = \langle\varphi_j \mid \mathcal{T}_{\mathcal{V}}\varphi_j\rangle = \mu_j$, so $\mathcal{T}\varphi_j = \mu_j\varphi_j$ by Lemma D.2. Since $\mu_k \geq \mu_K > 0$, $\mathcal{T}_{\mathcal{V}} = \sum_k \mu_k|\varphi_k\rangle\langle\varphi_k|$ is PSD, and $A_{\mathcal{V}} = (\mathcal{T}_{\mathcal{V}})_+ = \mathcal{T}_{\mathcal{V}}$.

Uniqueness. If $\mu_K > \mu_{K+1}$, at most K orthonormal eigenfunctions have eigenvalues $\geq \mu_K$: for $K+1$ of them, $\langle\psi \mid \mathcal{T}\psi\rangle$ of a unit ψ in their span is an average of their eigenvalues, hence $\geq \mu_K$, so the definition of μ_{K+1} gives $\mu_{K+1} \geq \mu_K$. $\mathcal{V}$ is then the top-K eigenspace, and $A_{\mathcal{V}}$ is the same for every minimizer. If $\mu_K = \mu_{K+1}$, $\mathcal{V}$ contains the eigenfunctions with eigenvalues above μ_K and a subspace of $\ker(\mathcal{T} - \mu_K)$ of the remaining dimension; the minimum and the eigenvalues of $A_{\mathcal{V}}$ do not depend on this choice, while $A_{\mathcal{V}}$ itself does. □

D.3.3 Remarks

Nonnegative scale parameters. The proof uses $\lambda_k \geq 0$ in step (b) of Lemma D.1, where $A_{\mathcal{V}} = \sum_k |f_k\rangle\langle f_k|$ is PSD. Over $\lambda_k < 0$, $\mathcal{J}_{\mathsf{LoRA}}$ is unbounded below, since $\mathcal{T} = -\mathcal{H}$ is unbounded below.

Spin penalty. The spin penalty of Section 4.1 replaces $\mathcal{H}$ by $\mathcal{H} + \hat{S}^2$; on $\mathscr{H}$ the total-spin operator $\hat{S}^2$ is bounded and commutes with $\mathcal{H}$, so $\mathcal{T} = -(\mathcal{H} + \hat{S}^2)$ is self-adjoint and bounded above, its eigenfunctions are those of $\mathcal{H}$ with each level raised by $S(S+1)$, and the theorem and Corollary D.1 apply to it.

Relation to Theorem C.5 of Ryu et al. [30]. Theorem D.1 holds for every self-adjoint $\mathcal{T}$ bounded above under Eq. (10), because it applies Eckart–Young to the finite-rank $\mathcal{T}_{\mathcal{V}}$ (Lemma D.1) and finds the eigenfunctions through the min–max values (Lemma D.2). Theorem C.5 of Ryu et al. [30] assumes an orthonormal basis of eigenfunctions of $\mathcal{T}$. The essential spectrum $(-\infty, -\Sigma]$ of $-\mathcal{H}$ contains the continuous spectrum of the scattering states, whose solutions of the eigenvalue equation lie outside $\mathscr{H}$. The eigenfunctions of $-\mathcal{H}$ therefore span a proper subspace of $\mathscr{H}$ (Appendix D.1.4). A trial function f_k has, in general, a component outside this subspace that an expansion of f_k in

eigenfunctions misses. In exchange, Theorem D.1 requires $\mu_K > 0$, while Theorem C.5 also covers an operator with only $r < K$ positive eigenvalues.

Scale parameters at a minimizer. With $f_k = \sqrt{\lambda_k}\,\psi_k/\|\psi_k\|$ as in the proof, the objective depends on $f_{1:K}$ only through $A_{\mathcal{V}} = \sum_k |f_k\rangle\langle f_k|$. For a minimizer with $A_{\mathcal{V}} = \sum_j \mu_j|\varphi_j\rangle\langle\varphi_j|$, the f_k are exactly $f_k = \sum_{j=1}^{K} U_{jk}\sqrt{\mu_j}\,\varphi_j$ with U a $K \times K$ unitary, because two families of K vectors with the same rank-K operator $\sum_k |f_k\rangle\langle f_k|$ differ by a unitary. Then $\lambda_k = \|f_k\|^2 = \sum_j |U_{jk}|^2\mu_j$. The joint minimum fixes the eigenvalues of $A_{\mathcal{V}}$ but not each λ_k, which equals μ_k only if U does not mix distinct eigenvalues. Corollary D.1 in Appendix D.4 shows that sequential nesting removes this freedom.

D.4 Fixed point of sequential nesting

This section treats sequential nesting once more, through its objective. The fixed point of sequential nesting is the ordered eigenpairs, up to a rotation within each degenerate level (Corollary D.1), while the joint minimum of Theorem D.1 fixes the states only up to a rotation among all K of them. The sequential nesting gradient derived in Appendix E.1.4 can be read as K optimizations, one per network: network k minimizes the order-k objective $\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k})$ over its own arguments, with the networks below it held fixed, and the corollary characterizes the fixed point of these K optimizations.

If networks $1, \dots, k-1$ hold the top $k-1$ eigenpairs, the order-k objective minimized over λ_k depends on ψ_k only through $q(\hat{\psi}_k) = \langle\hat{\psi}_k \mid \mathcal{T}\hat{\psi}_k\rangle - \sum_{\ell<k}\mu_\ell|\langle\hat{\psi}_\ell \mid \hat{\psi}_k\rangle|^2$ with $\hat{\psi}_k = \psi_k/\|\psi_k\|$. This q is largest exactly at the eigenfunctions of eigenvalue μ_k orthogonal to the lower networks (Lemma D.2). Induction on k then gives each network one eigenfunction and $\lambda_k = \mu_k$.

Corollary D.1 (Fixed point of sequential nesting)**.** *Assume Eq.* (10)*, and let* $(\psi_{1:K}, \lambda_{1:K})$ *be such that, for each* $k = 1, \dots, K$*,* (ψ_k, λ_k) *minimizes the order-*k *objective* $\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k})$ *over its own arguments with* $(\psi_{1:k-1}, \lambda_{1:k-1})$ *held fixed. Then for every* k*,* ψ_k *is an eigenfunction of* $\mathcal{T}$ *with eigenvalue* μ_k*, orthogonal to* $\psi_{1:k-1}$*, and* $\lambda_k = \mu_k$*. The statement also holds at* $\mu_K = \mu_{K+1}$*, and within a degenerate level* ψ_k *may be any such eigenfunction.*

Proof. By induction on k. Suppose that $\hat{\psi}_\ell := \psi_\ell/\|\psi_\ell\|$ for $\ell < k$ are orthonormal eigenfunctions of eigenvalue μ_ℓ and that $\lambda_\ell = \mu_\ell$, which is vacuous for $k = 1$, and let P be the orthogonal projection onto their span. With $f_\ell = \sqrt{\lambda_\ell}\,\hat{\psi}_\ell$, the order-$k$ objective is $-2\sum_{\ell\le k}\langle f_\ell|\mathcal{T}f_\ell\rangle + \sum_{\ell,m\le k}|\langle f_\ell|f_m\rangle|^2$, and its terms with f_k are $-2\langle f_k|\mathcal{T}f_k\rangle$, the cross terms $2\sum_{\ell<k}|\langle f_\ell|f_k\rangle|^2$ and the diagonal term $\|f_k\|^4 = \lambda_k^2$. Their sum is $-2\lambda_k\, q(\hat{\psi}_k) + \lambda_k^2$ with

$$q(\hat{\psi}) := \langle\hat{\psi}|\mathcal{T}\hat{\psi}\rangle - \sum_{\ell<k}\mu_\ell|\langle\hat{\psi}_\ell|\hat{\psi}\rangle|^2 = \langle(1-P)\hat{\psi}|\mathcal{T}(1-P)\hat{\psi}\rangle,$$

where the last equality holds because $\mathcal{T}$ maps the span of $\hat{\psi}_{1:k-1}$ into itself. Over $\lambda_k \ge 0$ the minimum is $-q(\hat{\psi}_k)_+^2$, at $\lambda_k = q(\hat{\psi}_k)_+$, so $\hat{\psi}_k$ maximizes q over unit vectors. For $q(\hat{\psi}_k) > 0$ this λ_k is the analytic update of Eq. (6) (Appendix E.2) with $\lambda_\ell = \mu_\ell$ for $\ell < k$, so the corollary also covers λ_k set by that update. By Lemma D.2, $q(\hat{\psi}) \le \mu_k\|(1-P)\hat{\psi}\|^2 \le \mu_k$, and since $\mu_k \ge \mu_K > 0$ both bounds are attained exactly when $P\hat{\psi} = 0$ and $\mathcal{T}\hat{\psi} = \mu_k\hat{\psi}$; such a $\hat{\psi}$ exists because $\mu_{1:K}$ are counted with multiplicity. Hence $\hat{\psi}_k$ is an eigenfunction of eigenvalue μ_k orthogonal to $\hat{\psi}_{1:k-1}$, and $\lambda_k = \mu_k$. □

E Deferred proofs

This section derives the gradient expressions and the eigenvalue update that the main text states without proof.

E.1 Gradient expressions

E.1.1 Proof of Eq. (1) (VMC gradient)

The gradient expression of the VMC Rayleigh quotient can be immediately derived from the following lemma.

Lemma E.1. *For a symmetric operator O,*

$$\frac{\nabla_\theta \langle \psi_\theta \mid O\psi_\theta \rangle}{\langle \psi_\theta \mid \psi_\theta \rangle} = 2\,\mathrm{Re}\Big(\mathbb{E}_{p_\theta(\mathbf{x})} \Big[\frac{O\psi_\theta(\mathbf{x})}{\psi_\theta(\mathbf{x})} \nabla_\theta \log \psi_\theta^*(\mathbf{x}) \Big] \Big).$$

Proof. The inner product between a vector-valued function and a scalar function denotes the coordinate-wise operation. By the chain rule, the commutativity of O with ∇_θ, and the symmetry of O,

$$\nabla_\theta \langle \psi_\theta \mid O\psi_\theta \rangle = 2\,\mathrm{Re}\langle \nabla_\theta \psi_\theta \mid O\psi_\theta \rangle = 2\|\psi_\theta\|^2\,\mathrm{Re}\,\mathbb{E}_{p_\theta(\mathbf{x})}\Big[\frac{O\psi_\theta(\mathbf{x})}{\psi_\theta(\mathbf{x})}\, \nabla_\theta \log \psi_\theta^*(\mathbf{x}) \Big],$$

where the second equality multiplies and divides by $|\psi_\theta(\mathbf{x})|^2/\|\psi_\theta\|^2 = p_\theta(\mathbf{x})$. Dividing both sides by $\langle \psi_\theta \mid \psi_\theta \rangle = \|\psi_\theta\|^2$ gives the claim. □

Proof of Eq. (1). Consider

$$\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \frac{\nabla_\theta \langle \psi_\theta \mid \mathcal{H}\psi_\theta \rangle}{\langle \psi_\theta \mid \psi_\theta \rangle} - \frac{\langle \psi_\theta \mid \mathcal{H}\psi_\theta \rangle \nabla_\theta \langle \psi_\theta \mid \psi_\theta \rangle}{\langle \psi_\theta \mid \psi_\theta \rangle^2} \\
&= \frac{\nabla_\theta \langle \psi_\theta \mid \mathcal{H}\psi_\theta \rangle}{\langle \psi_\theta \mid \psi_\theta \rangle} - \mathcal{L}(\theta) \frac{\nabla_\theta \langle \psi_\theta \mid \psi_\theta \rangle}{\langle \psi_\theta \mid \psi_\theta \rangle} \\
&\overset{(a)}{=} 2\,\mathrm{Re}\Big(\mathbb{E}_{p_\theta(\mathbf{x})} \Big[\frac{\mathcal{H}\psi_\theta(\mathbf{x})}{\psi_\theta(\mathbf{x})} \nabla_\theta \log \psi_\theta^*(\mathbf{x}) \Big] - \mathcal{L}(\theta) \mathbb{E}_{p_\theta(\mathbf{x})} \Big[\nabla_\theta \log \psi_\theta^*(\mathbf{x}) \Big] \Big) \\
&= 2\,\mathrm{Re}\Big(\mathbb{E}_{p_\theta(\mathbf{x})} \Big[(\mathcal{E}_\theta(\mathbf{x}) - \mathbb{E}_{p_\theta(\mathbf{x})}[\mathcal{E}_\theta(\mathbf{x})]) \nabla_\theta \log \psi_\theta^*(\mathbf{x}) \Big] \Big).
\end{aligned}$$

Here, (a) follows from Lemma E.1. When $\psi_\theta(\mathbf{x}) \in \mathbb{R}$ for any $\mathbf{x}$, note that

$$\nabla_\theta \log \psi_\theta^*(\mathbf{x}) = \nabla_\theta(\log|\psi_\theta(\mathbf{x})| + \log \mathrm{sgn}(\psi_\theta(\mathbf{x}))) = \nabla_\theta \log|\psi_\theta(\mathbf{x})| + \nabla_\theta \log \mathrm{sgn}(\psi_\theta(\mathbf{x})).$$

Here, taking $\log$ as the complex logarithm with its branch cut along the negative imaginary axis, $\nabla_\theta \log \mathrm{sgn}(\psi_\theta(\mathbf{x}))$ is well defined and equals 0 for almost every $\mathbf{x}$, since $\mathrm{sgn}(\psi_\theta(\mathbf{x})) \in \{-1, 1\}$ is locally constant in θ wherever $\psi_\theta(\mathbf{x}) \neq 0$. Thus we can conclude $\nabla_\theta \log \psi_\theta^*(\mathbf{x}) = \nabla_\theta \log|\psi_\theta(\mathbf{x})|$, and numerically we ignore this sign term and differentiate $\log|\psi_\theta|$ alone. This gives Eq. (1) with $\nabla_\theta \log|\psi_\theta|$ in place of $\nabla_\theta \log \psi_\theta^*$ for a real-valued ansatz. □

E.1.2 Derivation of the LoRA gradient

Lemma E.2. *Define*

$$\Gamma^O_{k\ell} := \frac{\nabla_\theta \langle \psi_k \mid O\psi_\ell \rangle}{\langle \psi_k \mid \psi_k \rangle}.$$

Then, we have

$$\nabla_\theta \mathsf{R}^O_{k\ell} = \Gamma^O_{k\ell} - \mathsf{R}^O_{k\ell} \Gamma^I_{kk}. \tag{11}$$

Proof. The quotient rule on $\mathsf{R}^O_{k\ell} = \langle \psi_k \mid O\psi_\ell \rangle / \langle \psi_k \mid \psi_k \rangle$ gives $\nabla_\theta \mathsf{R}^O_{k\ell} = \Gamma^O_{k\ell} - \mathsf{R}^O_{k\ell}\, \Gamma^I_{kk}$. □

Lemma E.3. *Define*

$$\Delta^O_{k\ell} := \frac{\langle \nabla_\theta \psi_k \mid O\psi_\ell \rangle}{\langle \psi_k \mid \psi_k \rangle} = \mathbb{E}_{p_k(\mathbf{x})} \Big[\frac{O\psi_\ell(\mathbf{x})}{\psi_k(\mathbf{x})} \nabla_\theta \log \psi_k^*(\mathbf{x}) \Big].$$

Then, for a symmetric operator O, we have

$$\Gamma^{O}_{k\ell} = \Delta^{O}_{k\ell} + \frac{\|\psi_\ell\|^2}{\|\psi_k\|^2}(\Delta^{O}_{\ell k})^*$$

In particular, when $\ell = k$, we have $\Delta^{I}_{kk} = \gamma_k$ and

$$\Gamma^{I}_{kk} = \gamma_k + \gamma_k^*,$$

where we define

$$\gamma_k := \mathbb{E}_{p_k}\big[\nabla_\theta \log \psi_k^*\big].$$

Proof. The product rule gives $\nabla_\theta\langle\psi_k \mid O\psi_\ell\rangle = \langle\nabla_\theta\psi_k \mid O\psi_\ell\rangle + \langle\psi_k \mid O\nabla_\theta\psi_\ell\rangle$, and since O is symmetric the second term is $\langle\nabla_\theta\psi_\ell \mid O\psi_k\rangle^*$; dividing by $\langle\psi_k \mid \psi_k\rangle$ gives the identity. Setting $\ell = k$ and $O = I$ in the definition of $\Delta^{O}_{k\ell}$ gives $\Delta^{I}_{kk} = \gamma_k$. □

The gradient identity used to assemble Eq. (4) follows from the lemma below.

Lemma E.4. *Define*

$$\tilde{\Delta}^{O}_{k\ell} := \mathrm{Cov}_{p_k}\Big(\frac{O\psi_\ell}{\psi_k}, \nabla_\theta \log \psi_k^*\Big) = \Delta^{O}_{k\ell} - \mathsf{R}^{O}_{k\ell}\gamma_k. \tag{12}$$

Then,

$$\nabla_\theta \mathsf{R}^{\mathcal{T}}_{kk} = \tilde{\Delta}^{\mathcal{T}}_{kk} + (\tilde{\Delta}^{\mathcal{T}}_{kk})^* = 2\,\mathrm{Re}(\tilde{\Delta}^{\mathcal{T}}_{kk}).$$

For a symmetric operator O, we have

$$\begin{aligned}\nabla_\theta \mathsf{R}^{O}_{k\ell}\mathsf{R}^{I}_{\ell k} &= \mathsf{R}^{O}_{k\ell}\tilde{\Delta}^{I}_{\ell k} + (\mathsf{R}^{O}_{\ell k}\tilde{\Delta}^{I}_{k\ell})^* + \mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{O}_{k\ell} + (\mathsf{R}^{I}_{k\ell}\tilde{\Delta}^{O}_{\ell k})^* \\ &= 2\,\mathrm{Re}\Big\{\mathsf{R}^{O}_{k\ell}\tilde{\Delta}^{I}_{\ell k} + \mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{O}_{k\ell}\Big\} \quad \textit{if } O = I.\end{aligned}$$

Proof. Continuing from Eq. (11),

$$\begin{aligned}\nabla_\theta \mathsf{R}^{O}_{k\ell} &= \Gamma^{O}_{k\ell} - \mathsf{R}^{O}_{k\ell}\Gamma^{I}_{kk} \\ &= \Delta^{O}_{k\ell} + \frac{\|\psi_\ell\|^2}{\|\psi_k\|^2}(\Delta^{O}_{\ell k})^* - \mathsf{R}^{O}_{k\ell}(\gamma_k + \gamma_k^*) \\ &= (\Delta^{O}_{k\ell} - \mathsf{R}^{O}_{k\ell}\gamma_k) + \frac{\|\psi_\ell\|^2}{\|\psi_k\|^2}(\Delta^{O}_{\ell k} - \mathsf{R}^{O}_{\ell k}\gamma_\ell)^* + \frac{\|\psi_\ell\|^2}{\|\psi_k\|^2}(\mathsf{R}^{O}_{\ell k}\gamma_\ell)^* - \mathsf{R}^{O}_{k\ell}\gamma_k^* \\ &= \tilde{\Delta}^{O}_{k\ell} + \frac{\|\psi_\ell\|^2}{\|\psi_k\|^2}(\tilde{\Delta}^{O}_{\ell k})^* + \frac{\|\psi_\ell\|^2}{\|\psi_k\|^2}(\mathsf{R}^{O}_{\ell k}\gamma_\ell)^* - \mathsf{R}^{O}_{k\ell}\gamma_k^*.\end{aligned}$$

This leads to:

$$\begin{aligned}\mathsf{R}^{O}_{k\ell}\nabla_\theta \mathsf{R}^{I}_{\ell k} &= \mathsf{R}^{O}_{k\ell}\tilde{\Delta}^{I}_{\ell k} + (\mathsf{R}^{O}_{\ell k}\tilde{\Delta}^{I}_{k\ell})^* + (\mathsf{R}^{O}_{\ell k}\mathsf{R}^{I}_{k\ell}\gamma_k)^* - \mathsf{R}^{O}_{k\ell}\mathsf{R}^{I}_{\ell k}\gamma_\ell^*, \\ \mathsf{R}^{I}_{\ell k}\nabla_\theta \mathsf{R}^{O}_{k\ell} &= \mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{O}_{k\ell} + (\mathsf{R}^{I}_{k\ell}\tilde{\Delta}^{O}_{\ell k})^* + (\mathsf{R}^{I}_{k\ell}\mathsf{R}^{O}_{\ell k}\gamma_\ell)^* - \mathsf{R}^{I}_{\ell k}\mathsf{R}^{O}_{k\ell}\gamma_k^*.\end{aligned}$$

Summing them up, we get

$$\begin{aligned}\nabla_\theta \mathsf{R}^{O}_{k\ell}\mathsf{R}^{I}_{\ell k} &= \mathsf{R}^{O}_{k\ell}\nabla_\theta \mathsf{R}^{I}_{\ell k} + \mathsf{R}^{I}_{\ell k}\nabla_\theta \mathsf{R}^{O}_{k\ell} \\ &= \Big(\mathsf{R}^{O}_{k\ell}\tilde{\Delta}^{I}_{\ell k} + (\mathsf{R}^{O}_{\ell k}\tilde{\Delta}^{I}_{k\ell})^* + \mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{O}_{k\ell} + (\mathsf{R}^{I}_{k\ell}\tilde{\Delta}^{O}_{\ell k})^*\Big) \\ &\quad + \Big((\mathsf{R}^{O}_{\ell k}\mathsf{R}^{I}_{k\ell})^* - \mathsf{R}^{I}_{\ell k}\mathsf{R}^{O}_{k\ell}\Big)(\gamma_k^* + \gamma_\ell^*) \\ &= \mathsf{R}^{O}_{k\ell}\tilde{\Delta}^{I}_{\ell k} + (\mathsf{R}^{O}_{\ell k}\tilde{\Delta}^{I}_{k\ell})^* + \mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{O}_{k\ell} + (\mathsf{R}^{I}_{k\ell}\tilde{\Delta}^{O}_{\ell k})^*.\end{aligned}$$

This concludes the proof. □

E.1.3 Final gradient expression for implementation

Recall that $\tilde{\Delta}^{\mathcal{O}}$ is defined in Eq. (12). For simplicity, we assume real ansatzes without loss of generality. Let θ_k denote the parameters of network k; with ∇_{θ_k} in place of ∇_θ in Eq. (12), $\tilde{\Delta}^{\mathcal{O}}_{ij} = 0$ for $i \neq k$.

Theorem E.1 (LoRA gradient)**.**

$$\begin{cases} \nabla_{\theta_k} \mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k}) & = -4\lambda_k \tilde{\Delta}^{\mathcal{T}}_{kk} + 4 \sum_{\ell=1}^{k} \lambda_k \lambda_\ell \mathsf{R}^I_{\ell k} \tilde{\Delta}^I_{k\ell}, \\ \partial_{\lambda_k} \mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k}) & = -2\mathsf{R}^{\mathcal{T}}_{kk} + 2 \sum_{\ell=1}^{k} \lambda_\ell \mathsf{R}^I_{\ell k} \mathsf{R}^I_{k\ell}. \end{cases}$$

Sequential nesting applies this gradient for each k.

Surrogate objective for autograd. To compute the centered gradient of Eq. (5) with a generic autograd, we differentiate a surrogate objective whose derivative gives this gradient. The surrogate, whose autograd derivative is twice the centered gradient, is

$$\breve{\mathcal{L}}_{\mathsf{LoRA}}(\psi_{1:K}, \lambda_{1:K}) := 2 \sum_{k=1}^{K} \mathbb{E}_{p_k}\big[\mathsf{sg}(g_k - \mathbb{E}_{p_k}[g_k])\, \log|\psi_k|\big], \tag{13}$$

where sg is the stop-gradient operator. The surrogate depends on θ only through $\log|\psi_k|$, in which it is linear, so a generic autograd differentiates it directly. In g_k, $\mathsf{R}^I_{\ell k}$ is estimated by its batch mean.

E.1.4 Sequential nesting identity

We state the lemma that gives the overlap term of Eq. (4) in terms of the covariance

$$\tilde{\Delta}^{\mathcal{O}}_{k\ell} := \mathrm{Cov}_{p_k}\big(\mathcal{O}\psi_\ell/\psi_k,\, \nabla_\theta \log \psi_k^*\big) \qquad \text{for } \mathcal{O} \in \{\mathcal{T}, I\}.$$

Lemma E.5 (Sequential nesting, LoRA case)**.** *The gradient contribution from the second term of $\mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k})$, without its weights $\lambda_i\lambda_j$, summed over k is*

$$\sum_{k=1}^{K} \big\langle \nabla_\theta \psi_k \;\big|\; \partial_{\psi_k} \textstyle\sum_{i,j=1}^{k} \mathsf{R}^I_{ij}\mathsf{R}^I_{ji} \big\rangle = 4 \textstyle\sum_{k,\ell=1}^{K} \mathsf{M}^{\mathsf{seq}}_{k\ell}\, \mathsf{R}^I_{\ell k}\, \tilde{\Delta}^I_{k\ell}.$$

Proof. For $\ell \neq k$, the quotient rule gives

$$\frac{\partial \mathsf{R}^{\mathcal{O}}_{k\ell}}{\partial \psi_k(\mathbf{x})} = \frac{\psi_k(\mathbf{x})}{\|\psi_k\|^2} \Big(\frac{\mathcal{O}\psi_\ell(\mathbf{x})}{\psi_k(\mathbf{x})} - 2\mathsf{R}^{\mathcal{O}}_{k\ell} \Big).$$

Therefore, using $\nabla_\theta \psi_k/\psi_k = \nabla_\theta \log|\psi_k|$ on the support of p_k,

$$\begin{aligned} \langle \nabla_\theta \psi_k \mid \partial_{\psi_k} \mathsf{R}^{\mathcal{O}}_{k\ell} \rangle &= \int \nabla_\theta \psi_k(\mathbf{x})\, \frac{\partial \mathsf{R}^{\mathcal{O}}_{k\ell}}{\partial \psi_k(\mathbf{x})}\, \mathrm{d}\mathbf{x} = \mathbb{E}_{p_k(\mathbf{x})}\Big[\Big(\frac{\mathcal{O}\psi_\ell(\mathbf{x})}{\psi_k(\mathbf{x})} - 2\mathsf{R}^{\mathcal{O}}_{k\ell} \Big) \frac{\nabla_\theta \psi_k(\mathbf{x})}{\psi_k(\mathbf{x})} \Big] \\ &= \Delta^{\mathcal{O}}_{k\ell} - 2\gamma_k \mathsf{R}^{\mathcal{O}}_{k\ell} = \tilde{\Delta}^{\mathcal{O}}_{k\ell} - \gamma_k \mathsf{R}^{\mathcal{O}}_{k\ell}. \end{aligned}$$

For the second factor $\mathsf{R}^I_{\ell k}$, for $\ell \neq k$, $\langle \nabla_\theta \psi_k \mid \partial_{\psi_k} \mathsf{R}^I_{\ell k} \rangle = (\|\psi_k\|^2/\|\psi_\ell\|^2)\, \Delta^I_{k\ell}$. Combining both, applying the identities $(\|\psi_k\|^2/\|\psi_\ell\|^2)\mathsf{R}^{\mathcal{O}}_{k\ell} = \mathsf{R}^{\mathcal{O}}_{\ell k}$ and $\mathsf{R}^{\mathcal{O}}_{k\ell}\mathsf{R}^I_{\ell k} = \mathsf{R}^{\mathcal{O}}_{\ell k}\mathsf{R}^I_{k\ell}$, and using $\Delta - \gamma R = \tilde{\Delta}$, when $\ell \neq k$:

$$\begin{aligned} \langle \nabla_\theta \psi_k \mid \partial_{\psi_k} \mathsf{R}^{\mathcal{O}}_{k\ell}\mathsf{R}^I_{\ell k} \rangle &= \mathsf{R}^I_{\ell k}\tilde{\Delta}^{\mathcal{O}}_{k\ell} - \gamma_k \mathsf{R}^{\mathcal{O}}_{k\ell}\mathsf{R}^I_{\ell k} + \mathsf{R}^{\mathcal{O}}_{k\ell}\, (\|\psi_k\|^2/\|\psi_\ell\|^2)\, \Delta^I_{k\ell} \\ &= \mathsf{R}^I_{\ell k}\tilde{\Delta}^{\mathcal{O}}_{k\ell} + \mathsf{R}^{\mathcal{O}}_{\ell k}(\Delta^I_{k\ell} - \gamma_k \mathsf{R}^I_{k\ell}) = \mathsf{R}^{\mathcal{O}}_{\ell k}\tilde{\Delta}^I_{k\ell} + \mathsf{R}^I_{\ell k}\tilde{\Delta}^{\mathcal{O}}_{k\ell}. \end{aligned}$$

In $\sum_{i,j=1}^{k} \mathsf{R}^{\mathcal{O}}_{ij}\mathsf{R}^{I}_{ji}$, each $\ell < k$ appears in the two terms $(i,j) = (k,\ell)$ and (ℓ,k), which are equal by the second identity. The term (k,k) is $\mathsf{R}^{\mathcal{O}}_{kk}$, since $\mathsf{R}^{I}_{kk} = 1$, and the quotient rule gives $\langle \nabla_\theta \psi_k \mid \partial_{\psi_k} \mathsf{R}^{\mathcal{O}}_{kk} \rangle = 2\tilde{\Delta}^{\mathcal{O}}_{kk}$, which is the $\ell = k$ summand below since $\tilde{\Delta}^{I}_{kk} = 0$. Hence,

$$\left\langle \nabla_\theta \psi_k \,\middle|\, \partial_{\psi_k} \sum_{i,j=1}^{k} \mathsf{R}^{\mathcal{O}}_{ij}\mathsf{R}^{I}_{ji} \right\rangle = 2\sum_{\ell=1}^{k} (\mathsf{R}^{\mathcal{O}}_{\ell k}\tilde{\Delta}^{I}_{k\ell} + \mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{\mathcal{O}}_{k\ell}).$$

At $\mathcal{O} = I$ the bracket is $2\mathsf{R}^{I}_{\ell k}\tilde{\Delta}^{I}_{k\ell}$ and vanishes at $\ell = k$, so summing over k gives the lemma. □

E.2 Derivation of the analytic eigenvalue update in Eq. (6)

We derive the closed form of Eq. (6) step by step. For each $k \in [K]$, the partial derivative of the LoRA objective with respect to λ_k at the sequentially nested level is

$$\frac{\partial}{\partial \lambda_k} \mathcal{J}_{\mathsf{LoRA}}(\psi_{1:k}, \lambda_{1:k}) = -2\mathsf{R}^{\mathcal{T}}_{kk} + 2\sum_{\ell=1}^{k} \lambda_\ell \mathsf{R}^{I}_{k\ell}\mathsf{R}^{I}_{\ell k}.$$

Setting this to zero,

$$\lambda^\star_k = \mathsf{R}^{\mathcal{T}}_{kk} - \sum_{\ell=1}^{k-1} \lambda^\star_\ell \mathsf{R}^{I}_{k\ell}\mathsf{R}^{I}_{\ell k}.$$

Solving sequentially in k yields

$$\begin{aligned}
\lambda^\star_1 &= \mathsf{R}^{\mathcal{T}}_{11}, \\
\lambda^\star_2 &= \mathsf{R}^{\mathcal{T}}_{22} - \lambda^\star_1 \mathsf{R}^{I}_{21}\mathsf{R}^{I}_{12}, \\
&\vdots \\
\lambda^\star_K &= \mathsf{R}^{\mathcal{T}}_{KK} - \sum_{\ell=1}^{K-1} \lambda^\star_\ell \mathsf{R}^{I}_{K\ell}\mathsf{R}^{I}_{\ell K}.
\end{aligned}$$

In matrix-vector form,

$$\begin{aligned}
\boldsymbol{\lambda}^\star &= \operatorname{diag}(\mathsf{R}^{\mathcal{T}}) - \begin{bmatrix} 0 \\ \lambda^\star_1 \mathsf{R}^{I}_{21}\mathsf{R}^{I}_{12} \\ \vdots \\ \lambda^\star_1 \mathsf{R}^{I}_{K,1}\mathsf{R}^{I}_{1,K} + \ldots + \lambda^\star_{K-1}\mathsf{R}^{I}_{K,K-1}\mathsf{R}^{I}_{K-1,K} \end{bmatrix} \\
&= \operatorname{diag}(\mathsf{R}^{\mathcal{T}}) - \big(\mathsf{M}^{\mathsf{seq}} \odot (\mathsf{R}^{I} \odot (\mathsf{R}^{I})^{\intercal})\big)\boldsymbol{\lambda}^\star,
\end{aligned}$$

which rearranges to the closed form quoted in the main text:

$$\boldsymbol{\lambda}^\star = \big(\mathsf{I} + \mathsf{M}^{\mathsf{seq}} \odot (\mathsf{R}^{I} \odot (\mathsf{R}^{I})^{\intercal})\big)^{-1} \operatorname{diag}(\mathsf{R}^{\mathcal{T}}).$$

In our implementation, $\mathsf{R}^{\mathcal{T}}$ and R^{I} in this solve are estimated by exponential moving averages (EMAs) of their per-step batch means, which smooths $\boldsymbol{\lambda}^\star$ across steps. The EMA is warm-started: at step 1 it takes the first batch statistic itself rather than blending against an arbitrary prior $\hat{x}_0$, so the smoothing starts at step 2. The gradient coefficient g_k of Eq. (5) uses the per-step batch means.

F Monte Carlo statistics

This section gives the estimates of energies and overlaps that Section 3.3 takes from the walkers.

Averages over steps. An exponential moving average (EMA) over training steps t of a batch statistic x_t is $m_t = \alpha_t x_t + (1-\alpha_t)\, m_{t-1}$. The inputs $\mathsf{R}^{\mathcal{T}}$ and R^{I} of the eigenvalue solve (Appendix E.2) and the logarithm of the norm-ratio estimate (Appendix F.2) are EMAs with $\alpha_t = 0.01$, started at $m_1 = x_1$. The energies that rank the networks (Appendix G.2) are EMAs with the schedule $\alpha_t = \max(10^{-3}, 1/(2 + t/10))$ of DeepQMC [9], which falls from $1/2$ to 10^{-3} within the first $\sim 10^4$ steps.

F.1 Energy statistics

Clipping. The local energy in $\tilde{\Delta}^{T}_{kk}$, $x = \mathcal{E}_k(\mathbf{x})$ plus $\hat{S}^2_{\mathrm{loc}}(\mathbf{x})$ when the spin penalty is on (Appendix G.3), is clipped to the window $[m - w\,d,\ m + w\,d]$ around its batch median m, with $w = 10$ and $d := \mathbb{E}[|x - m|]$. The clipped values enter the energy term of g_k and the energy input of the eigenvalue solver.

F.2 Overlap statistics

The ratio $\psi_\ell(\mathbf{x})/\psi_k(\mathbf{x})$ in $\tilde{\Delta}^{I}_{k\ell}$ enters training unclipped, because a clip of this heavy-tailed ratio biases the overlap estimate (Appendices F.2.1 and F.2.2). The coefficient $\mathsf{R}^{I}_{\ell k}$ of g_k in Eq. (5) is the unclipped batch mean of the reverse ratio ψ_k/ψ_ℓ on the walkers of ψ_ℓ, and the eigenvalue solve of Eq. (6) takes the EMA of the same unclipped batch means as R^{I}. Before these means are formed, each ratio is divided by the square root of an estimate of the norm ratio $\langle\psi_\ell \mid \psi_\ell\rangle/\langle\psi_k \mid \psi_k\rangle = \mathbb{E}_{p_k}[(\psi_\ell/\psi_k)^2]$, obtained by exponentiating an EMA (Appendix F) of the logarithm of its batch mean. This division makes the batch mean of a ratio an estimate of the normalized overlap $\tilde{\mathsf{S}}_{k\ell}$. The two ratios of a pair are rescaled by estimates from different chains, and the overlap term of g_k multiplies one rescaled ratio by the batch mean of the other, so this term is multiplied by a factor that is one at the exact norm ratios; the estimate could be replaced by the iterative bridge sampling of Gao et al. [6], or the rescaling removed.

F.2.1 Heavy-tail behavior of the overlap-term ratio at nodal surfaces

The ratio $\psi_\ell(\mathbf{x})/\psi_k(\mathbf{x})$ at $\mathbf{x} \sim p_k$ has a heavy tail near the nodes of ψ_k, and it is the per-sample integrand of the overlap term of g_k in Eq. (5). At a node of ψ_k, $|\psi_k|$ vanishes while ψ_ℓ is generically nonzero, so the ratio diverges for a walker close to a node of the denominator, while its mean $\mathsf{R}^{I}_{k\ell}$ stays finite.

Closed-form example: Hydrogen $2p_z$ vs $2p_x$. The simplest non-trivial pair already shows the heavy tail: take $\psi_k = \psi_{2p_z}$ as the denominator and $\psi_\ell = \psi_{2p_x}$ as the numerator on Hydrogen. Both have the same radial factor, which cancels, so the overlap ratio reduces to the angular ratio

$$R(\mathbf{x}) \;=\; \frac{\psi_{2p_x}(\mathbf{x})}{\psi_{2p_z}(\mathbf{x})} \;=\; \tan\theta \cdot \cos\phi,$$

sampled at $\mathbf{x} \sim |\psi_{2p_z}|^2$. Setting $u = \cos\theta$, the squared density factorizes as $\frac{3}{2}u^2\,\mathrm{d}u \cdot \frac{1}{2\pi}\,\mathrm{d}\phi$, so the polar and azimuthal factors are independent. The polar tail is $\mathbb{P}(|\tan\theta| > w) = (1 + w^2)^{-3/2}$ exactly. Integrating over ϕ and differentiating gives the closed-form density

$$f_R(r) \;=\; \frac{2}{\pi}\,\frac{1}{(1 + r^2)^2}, \qquad r \in \mathbb{R}, \tag{14}$$

which (after the rescaling $x = \sqrt{3}\,r$) is exactly Student's t distribution with $\nu = 3$ degrees of freedom. The tail decays as $\mathbb{P}(|R| > t) \sim 4/(3\pi t^3)$ as $t \to \infty$, and consequently $\mathbb{E}[|R|^p] < \infty$ if and only if $p < 3$: the per-sample overlap estimator has a finite mean and a finite variance, but its third absolute moment diverges logarithmically.

General nodal sets.

Proposition F.1. *Assume, for some $M, \eta > 0$, that*

- *(i) ψ_k is real and continuously differentiable,*
- *(ii) $|\psi_\ell| \le M$,*
- *(iii) $\nabla\psi_k \neq 0$ wherever $|\psi_k| < \eta$, and*
- *(iv) for $u, v \in \mathbb{R}$ the function $\Gamma(u, v) := \int_{\{\psi_k = u\}} \mathbf{1}\{|\psi_\ell| > |v|\}\,|\nabla\psi_k|^{-1}\,\mathrm{d}\sigma$ satisfies $\Gamma(u, v) \to \Gamma(0, v)$ as $u \to 0$ for almost every v, and $\Gamma(u, v) \le \bar{\Gamma}(v)$ for $|u| < \eta$ with $\int v^2\,\bar{\Gamma}(v)\,\mathrm{d}v < \infty$.*

Then $R = \psi_\ell/\psi_k$ at $\mathbf{x} \sim p_k$ satisfies

$$\lim_{t\to\infty} t^3\mathbb{P}(|R| > t) \;=\; \frac{2}{3\langle\psi_k \mid \psi_k\rangle}\int_{\{\psi_k = 0\}} \frac{|\psi_\ell|^3}{|\nabla\psi_k|}\,\mathrm{d}\sigma. \tag{15}$$

Proof. For $t > M/\eta$ the event $|R| > t$, that is $|\psi_k| < |\psi_\ell|/t$, lies in $\{|\psi_k| < \eta\}$, where the coarea formula gives $\mathbb{P}(|R| > t) = \langle \psi_k \mid \psi_k \rangle^{-1} \int_{-\eta}^{\eta} u^2\, \Gamma(u, tu)\, \mathrm{d}u$. The substitution $u = v/t$ gives $t^3 \mathbb{P}(|R| > t) = \langle \psi_k \mid \psi_k \rangle^{-1} \int_{-\eta t}^{\eta t} v^2\, \Gamma(v/t, v)\, \mathrm{d}v$, which tends to $\langle \psi_k \mid \psi_k \rangle^{-1} \int v^2\, \Gamma(0, v)\, \mathrm{d}v$ by dominated convergence. Exchanging the two integrals, each point of $\{\psi_k = 0\}$ contributes $\int_{|v|<|\psi_\ell|} v^2\, \mathrm{d}v = \frac{2}{3}|\psi_\ell|^3$, which gives Eq. (15). □

For the pair of Eq. (14), with $\psi_k = \psi_{2p_z} = z\, e^{-r/2}$ and $\psi_\ell = \psi_{2p_x} = x\, e^{-r/2}$, the nodal set is the plane $z = 0$, the integral in Eq. (15) is 64 and $\langle \psi_k \mid \psi_k \rangle = 32\pi$, so the limit is the constant $4/(3\pi)$ found above. The exponent 3 comes from assumptions (i) and (iii), under which ψ_k changes sign linearly across the hypersurface $\{\psi_k = 0\}$. A complex ψ_k, which vanishes on a set of codimension two, gives t^{-4} instead. A crossing of two nodal surfaces, where $\nabla\psi_k = 0$, gives $t^{-3} \log t$. Assumption (iv) excludes a heavier tail from the region where one electron is far from the nuclei. No other property of the two functions enters, so the t^{-3} tail applies to molecular wavefunctions that satisfy these conditions. Under the same conditions, the square $(\psi_\ell/\psi_k)^2$ in the norm-ratio estimate and, where $\nabla_\theta \psi_k \neq 0$ on the node, the product of the ratio with the score in Eq. (5) have a finite mean and an infinite variance.

Distributions at convergence. The local energy and the ratio differ in how their distributions behave as the networks converge. At an eigenstate the local energy is constant, $\mathcal{E}_k(\mathbf{x}) = E_k$ for every $\mathbf{x}$, so its distribution under p_k shrinks to a point, every sample falls inside the clip window, and the energy clip does not act. At the eigenfunctions the ratio $R = \varphi_\ell/\varphi_k$ keeps the tail $\mathbb{P}(|R| > t) \sim t^{-3}$ of Proposition F.1 near every node of φ_k.

F.2.2 Measurements on trained networks

On trained networks, a clip biases the batch means of the overlap statistics by much more than it reduces their spread (Section 3.3.1). The networks trained with the unclipped ratio are close to orthogonal (Figure 2).

Long chains. Two long Metropolis chains for each pair of networks ψ_k and ψ_ℓ, run on saved checkpoints, give the reference value of $\mathbb{E}_{p_k}[\psi_\ell/\psi_k]$ and the normalized overlap of the pair. The chains sample the density proportional to $\rho = |\psi_k|^2 + e^{-a}|\psi_\ell|^2$, so they cross the nodes of ψ_k more often than the chain of ψ_k. With $\hat{\mathbb{E}}_\rho[f]$ the mean of f over their walkers, we estimate

$$\mathbb{E}_{p_k}\left[\frac{\psi_\ell}{\psi_k}\right] = \frac{\langle \psi_k \mid \psi_\ell \rangle}{\|\psi_k\|^2} \approx \frac{\hat{\mathbb{E}}_\rho[\psi_k \psi_\ell/\rho]}{\hat{\mathbb{E}}_\rho[|\psi_k|^2/\rho]},$$

$$\frac{\langle \psi_k \mid \psi_\ell \rangle}{\|\psi_k\|\,\|\psi_\ell\|} \approx \frac{\hat{\mathbb{E}}_\rho[\psi_k \psi_\ell/\rho]}{\sqrt{\hat{\mathbb{E}}_\rho[|\psi_k|^2/\rho]\,\hat{\mathbb{E}}_\rho[|\psi_\ell|^2/\rho]}}.$$

Since the walkers follow ρ/Z_ρ with $Z_\rho = \int \rho$, $\hat{\mathbb{E}}_\rho[f/\rho]$ tends to $Z_\rho^{-1} \int f$, so Z_ρ cancels in each ratio and every a gives the same limit. The constant a sets only the share of walkers on $|\psi_k|^2$, $\|\psi_k\|^2/(\|\psi_k\|^2 + e^{-a}\|\psi_\ell\|^2)$. Optimally, e^a is the norm ratio $\|\psi_\ell\|^2/\|\psi_k\|^2$ estimated from shorter chains, which makes this share about $1/2$.

Bias of the clip (Figure 4a& 4b). Clipping R at a fixed number of median absolute deviations (MADs) strongly biases its batch mean; the bias remains at a width of 50 MADs and is gone only at 100 MADs (Figure 4b). Figure 4b shows that the bias of the clip is many times the spread across batches, even though the clip lowers this spread by removing the tail of R. The clipped batch mean is thus a precise estimate of a shifted value. For the reverse ratio ψ_5/ψ_1 on the chain of ψ_1, the unclipped batch mean lies near the reference. The ratio clip lowers it to half of it or less.

We provisionally apply the ratio clip of the DeepQMC code [9] at 10 MADs to $R = \psi_1/\psi_5$ at step 1,000 of a Li run. The fraction of walkers beyond s MADs in Figure 4a is the fraction a ratio clip of width s changes. The walkers of ψ_5 come from its *training chain* and from *restarted chains*, which cross the nodes of ψ_5 more often. The two long chains of this pair use $a = 0$, which gives the same limit, with 4,096 walkers each and 45,000 steps; their walkers, resampled to p_5 and then moved by 300 Metropolis steps on p_5, start the restarted chains. Without a clip, the batch mean agrees with the

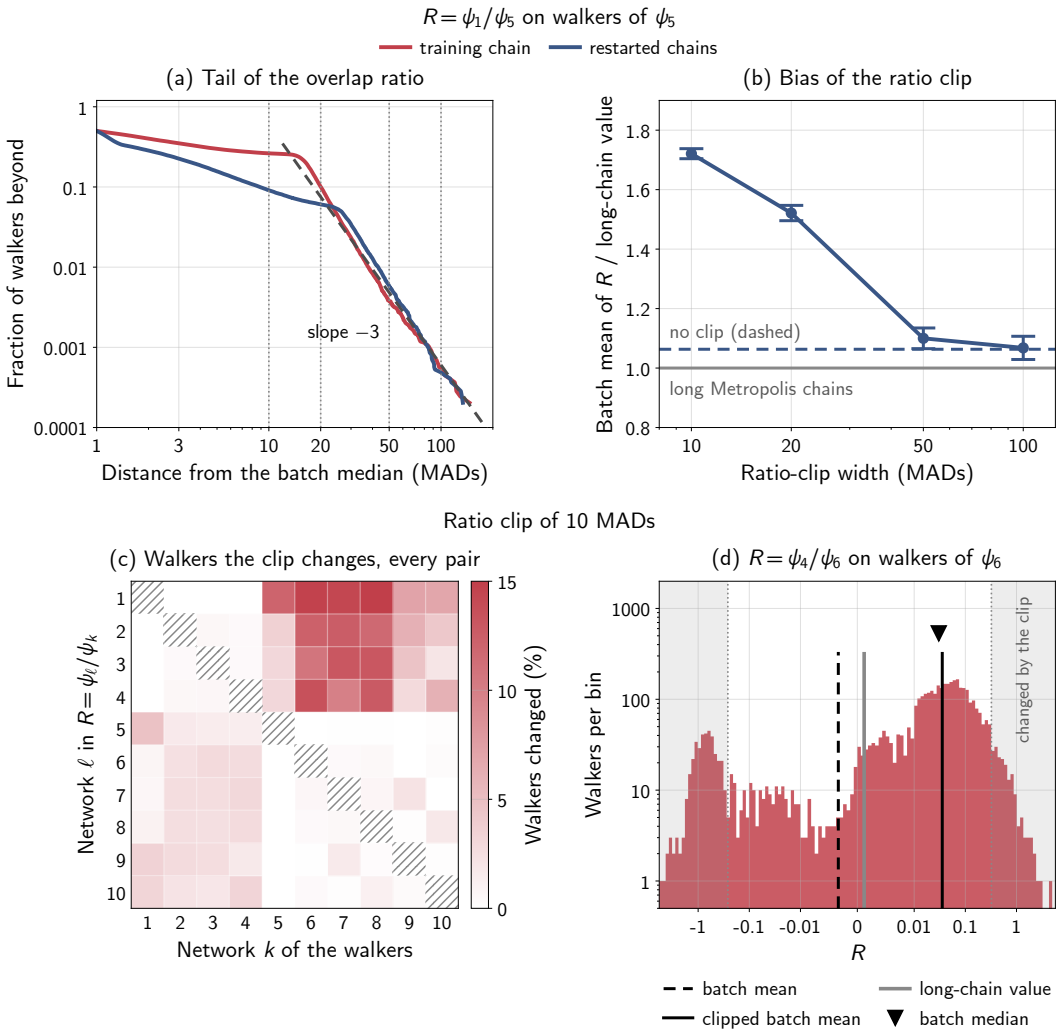


Figure 4: Overlap ratio $R = \psi_\ell/\psi_k$ of networks k and ℓ of the Li ansatz, on the walkers of ψ_k. Panels (a) and (b) are at training step 1,000 and number the networks by ascending energy at that step; panels (c) and (d) are at step 100,000 and number them as in Table 9. Dotted lines in (a) mark the clip widths of (b); error bars in (b) are standard errors over six batches; the axis of R in (d) is logarithmic beyond ± 0.01. Panel (b) uses the restarted chains (Appendix F.2.2).

long-chain reference within its spread across batches on the restarted chains and has the opposite sign on the training chain, so Figure 4b shows the restarted chains only.

The ratio at the end of training (Figure 4c& 4d). The ratio $R = \psi_\ell/\psi_k$ keeps a heavy, power-law tail to the end of training, as the analysis of Appendix F.2.1 supports. At step 100,000 of the seed-1 Li run, we measure the tail of each of the 90 ratios by the fraction of the 4,096 training walkers of ψ_k at which R lies more than s MADs from its median. Proposition F.1 gives this fraction an asymptotic log-log slope of -3 in s; the slopes fitted to the 90 ratios have median -2.5 and quartiles -3.0 and -1.9.

At the end of training, where the overlaps should vanish, the ratio clip at 10 MADs still changes the walkers of many ratios (Figure 4c) and biases their batch means. The clip window is centered on the batch median, so the clip removes more walkers on the heavier side of the tail and moves the batch mean toward the median (Figure 4d). For $R = \psi_4/\psi_6$, the ratio whose batch mean the clip moves farthest relative to its long-chain value, the clip moves the batch mean from -0.0033 to $+0.0354$, close to the median $+0.0302$, while the long chains give $+0.0012 \pm 0.0006$.

Overlaps of the trained networks. The near-orthogonality of the trained networks in Figure 2 of Section 4.3 is measured on long chains, with two chains for each pair that repeat the same measurement as a check on each other. The figure covers the pairs within a spin sector at the final step of the seed-1 runs of Li to F, since pairs in different sectors are orthogonal by spin symmetry. Each chain has $2{,}048$ walkers and runs $5{,}000$ steps of burn-in and up to 10^5 steps. The standard error of each estimate is computed by resampling the walkers, each with its whole trajectory, which accounts for the correlation along a stationary chain.

Table 4 classes each pair by whether its two long chains agree and are long enough: *converged* when they agree within three combined standard errors, every integrated autocorrelation time τ_{int} is resolved, and each chain is at least $50\,\tau_{\mathrm{int}}$ long; *short* when only the last condition fails; *differ* when the chains differ by three combined standard errors or more; and *unresolved* when a τ_{int} window reaches the end of its chain. A chain shorter than $50\,\tau_{\mathrm{int}}$ was extended once, except on F.

Table 4: Pairs of networks of each atom in Figure 2 by the convergence of their two long chains, and the largest $|\tilde{\mathsf{S}}_{k\ell}|$ over all pairs of the atom.

Atom	Pairs	Converged	Short	Differ	Unresolved	Largest $\lvert\tilde{\mathsf{S}}_{k\ell}\rvert$
Li	45	29	14	0	2	0.037
Be	21	21	0	0	0	0.037
B	24	18	3	2	1	0.032
C	18	18	0	0	0	0.026
N	29	28	0	1	0	0.021
O	18	18	0	0	0	0.017
F	21	9	12	0	0	0.021

G Implementation details

Every choice in this section keeps the eigenfunctions of $\mathcal{T} = -\mathcal{H}$ and changes how training reaches them. Appendix F gives the Monte Carlo estimates of energies and overlaps.

G.1 The per-network shifted operator

The weight from the analytic solve. The overlap-term weights λ_ℓ of g_k under the shifted operator $\mathcal{T}_k = E_{\mathrm{ref},k}\, I - \mathcal{H}$ follow from the analytic solve of Eq. (6), and one extra triangular solve gives them for all K networks. The shift adds $E_{\mathrm{ref},k}$ to each $\mathsf{R}^{\mathcal{T}}_{\ell\ell}$ and leaves R^{I} unchanged, and the solve is linear in $\mathrm{diag}(\mathsf{R}^{\mathcal{T}})$, so the new weights, the eigenvalue parameters of $\mathcal{T}_k$, are

$$\boldsymbol{\lambda}^{(k)} = \boldsymbol{\lambda}^{\star} + E_{\mathrm{ref},k}\,\mathbf{u}, \qquad \mathbf{u} := \left(\mathsf{I} + \mathsf{M}^{\mathsf{seq}} \odot (\mathsf{R}^{I} \odot (\mathsf{R}^{I})^{\intercal})\right)^{-1}\mathbf{1}, \tag{16}$$

with $\mathbf{1}$ the all-ones vector. Network k uses $\lambda_\ell^{(k)}$ in place of λ_ℓ in $g_k(\mathbf{x})$. When the overlaps between networks vanish, $\mathbf{u} = \mathbf{1}$. With the prefactor λ_k dropped, the shift adds the constant $-E_{\mathrm{ref},k}$ to the energy term of $g_k(\mathbf{x})$, which cancels in the centered gradient, so only the weights of the overlap term change.

Controlled experiment on carbon. The shift lowers the curvature of a carbon ^1D network toward the lower levels from about 724 times the smaller curvature outside the level to at most 17 times it, which stabilizes the energy convergence (Section 4.4). For a network of the ^1D level other than the first in order, the smaller curvature outside the level is the ^1D–^1S gap of $52.2\,\mathrm{mHa}$ [32]. The first has no lower network in its spin sector, so the shift does not change its curvatures. With $\mathcal{T} = -\mathcal{H}$, that is $E_{\mathrm{ref},k} = 0$, the curvature toward the lower levels is $|E_k| \approx 37.8\,\mathrm{Ha}$, so the ratio κ of this curvature to the gap is about 724; with $E_{\mathrm{ref},k} = -\lambda_k + c\,\sigma_k$ at $c = 5$ it is $c\,\sigma_k \lesssim 0.87\,\mathrm{Ha}$ on the seed-1 carbon run of Table 2 after step $15{,}000$, above the gap, so $\kappa \lesssim 17$.

Weight floor. At an exact eigenstate, $-\lambda_k = E_k$ and $\sigma_k = 0$, so $E_{\mathrm{ref},k} = E_k$ and the lower-state curvature vanishes. A trained network is an approximate eigenstate, so in practice σ_k stays positive. In theory, if $E_\ell = E_k$ for some $\ell < k$, the weight $\lambda_\ell^{(k)} = E_{\mathrm{ref},k} - E_\ell$ is $c\,\sigma_k = 0$, and a rotation of ψ_k toward ψ_ℓ within their level keeps $\sigma_k = 0$, so the objective leaves their overlap free. This case

arises only at an exact eigenstate, outside the practical regime, and as a guard the implementation replaces every weight $\lambda_\ell^{(k)}$ of Eq. (16) by $\max(\lambda_\ell^{(k)}, 1\,\mathrm{mHa})$.

G.2 State ordering

Within each spin sector we permute the order indices of the networks at every step, ranking the networks by the EMA (Appendix F) of each network's batch mean of $\mathcal{E}_k + \hat{S}^2_{\mathrm{loc}}$, taken before clipping (Appendix G.3). The EMAs of $\mathsf{R}^{\mathcal{T}}$, R^{I} and the norm ratios are kept per network, and the eigenvalue solve permutes $\mathsf{R}^{\mathcal{T}}$ and R^{I} into the step's order. At the fixed point of Corollary D.1 the local energies are constant and the networks are already in energy order, so the assignment leaves the fixed point unchanged; within a degenerate level the order is free, as the corollary allows. After step $10{,}000$, in our runs of Li to F, the assignment changes between networks that end in different levels in three runs: on one Li seed the networks of the $3p\ {}^2P^\circ$ and $3d\ {}^2D$ levels swap near step $17{,}000$, and on two B seeds two networks swap within the $3p\ {}^2P^\circ$ level until one of them descends to the ${}^2D$ level.

G.3 Spin penalty

The spin penalty replaces the operator $-\mathcal{H}$ of the Hilbert–Schmidt objective by $-(\mathcal{H} + \hat{S}^2)$, which keeps the eigenstates and raises each level of total spin S by $S(S{+}1)$. The molecular Hamiltonian commutes with $\hat{S}^2$ for spin-independent (Coulomb) interactions, so $\mathcal{H}$ and $\mathcal{H}{+}\hat{S}^2$ share an eigenbasis.

In NestedLoRA-VMC, g_k in Eq. (5) replaces $\mathcal{T}\psi_k/\psi_k$ with the augmented local quantity:

$$\frac{(\mathcal{H} + \hat{S}^2)\psi_k(\mathbf{x})}{\psi_k(\mathbf{x})} = \mathcal{E}_k(\mathbf{x}) + \hat{S}^2_{\mathrm{loc}}(\mathbf{x}).$$

The local spin $\hat{S}^2_{\mathrm{loc}}$ is evaluated by the Löwdin swap-ratio formula on the same MCMC sample $\mathbf{x}$ (Eq. 44 of Pfau et al. [27]). For $n_\uparrow \geq n_\downarrow$ up- and down-spin electrons, the Löwdin formula is:

$$\hat{S}^2_{\mathrm{loc}}(\mathbf{x}) = m_S(m_S{+}1) + n_\downarrow - \sum_{i\in\uparrow,\, j\in\downarrow} \frac{\psi_k(\tau_{ij}\mathbf{x})}{\psi_k(\mathbf{x})},$$

where τ_{ij} exchanges the positions of up-spin electron i and down-spin electron j.

Because both local quantities are evaluated on identical walkers, the per-walker sum $\mathcal{E}_k + \hat{S}^2_{\mathrm{loc}}$ is the local energy of $\mathcal{H} + \hat{S}^2$, and we clip this sum as one quantity with the window of Appendix F.1 (as the NES-VMC implementation does [34]). Clipping $\mathcal{E}_k$ alone and subtracting a separate batch-mean $\langle \hat{S}^2 \rangle$ would leave the heavy tails of $\hat{S}^2_{\mathrm{loc}}$ unclipped.

We add $\hat{S}^2$ to the operator in the run of every atom, in one m_S sector per spin multiplicity (Appendix H.4), so the penalty enters only the K local quantities g_k, one for each network. Szabó et al. [37] instead add the penalty to the loss as a separate term (Appendix C.2), and NES-VMC [27] adds a multiple of $\hat{S}^2$ to the Hamiltonian for the singlet states of cyclopentadienone and tetrazine, where the penalty enters all K^2 entries of its local-energy matrix.

H Experimental setup details

This section gives the settings of the runs reported in Section 4.1. Each network ψ_k is a Psiformer (Appendix H.1), and we discuss its computational stability in float32 (Appendix H.1.1). Each network samples its own walkers by Metropolis steps (Appendix H.2) and is optimized by natural gradient descent with KFAC (Appendix H.3). The per-atom settings give the spin sectors, the terms each sector holds and the CASSCF active space (Appendix H.4), and pretraining fits each network to one root of the CASSCF (Appendix H.5). With the envelope shift, the local energies, $\log \psi_k$ and updates of the reported runs stayed finite, and non-finite values that could later occur are handled by dropping the affected walkers and undoing the affected steps (Appendix H.6).

Table 5: Settings of each network ψ_k.

Input of electron i	$\mathbf{r}_i - \mathbf{R}_I$ and $\lVert\mathbf{r}_i - \mathbf{R}_I\rVert$ for every nucleus I, each scaled by $\log(1 + \lVert\mathbf{r}_i - \mathbf{R}_I\rVert)/\lVert\mathbf{r}_i - \mathbf{R}_I\rVert$, and the spin $\sigma_i = \pm 1$
Embedding	linear map without bias to dimension 256
Layers	4, each a self-attention over all electrons with 4 heads followed by a two-layer $\tanh$ MLP of width 256, both with residual sums
Orbitals	linear map without bias from the embedding of electron i to f_{dij}
Determinants	$n_{\mathrm{det}} = 16$, each over all n_{el} electrons, summed with unit weights
Envelopes	π^{σ}_{djI} and ζ^{σ}_{djI} initialized to 1
Jastrow factor	$J = -\sum_{i<j} c_{ij}\,\alpha^2/(\alpha + \lVert\mathbf{r}_i - \mathbf{r}_j\rVert)$ with $c_{ij} = 1/4$ for equal and $1/2$ for opposite spins, and one trained α for each case, initialized to 1

H.1 Architectures

All experiments use Psiformer [38] in the DeepQMC framework [9]; each per-state ansatz ψ_k in our NestedLoRA-VMC method is an independent Psiformer instance. As in FermiNet [26], each network has two stages: a permutation-equivariant backbone computes one feature vector per electron, and orbitals obtained from these features by a linear map enter determinants, which make ψ_k antisymmetric.

Permutation-equivariant backbones. Each layer of the Psiformer or FermiNet backbone updates the feature of an electron from its own feature and from the unordered set of the features of all electrons, so permuting the electrons permutes the features in the same way. In the Psiformer each layer is a self-attention over all electrons followed by an MLP [38], and in FermiNet each layer appends to the feature of an electron the means of the one- and two-electron features over the electrons of each spin [26].

Our Psiformer instance has four self-attention layers, four heads and MLP hidden dimension 256 (Table 5). Relative to the reference implementation [34], the DeepQMC Psiformer omits the layer normalization after each residual sum.

Orbitals and determinants. The second stage completes the antisymmetry: orbitals computed by a linear map from the feature of electron i fill row i of each determinant $\det A_d$, so the exchange of two electrons of the same spin swaps two rows of every determinant and changes the sign of ψ_k, and every network therefore lies in $\mathscr{H}$. Figure 5 shows the detailed structure.

Each network is $\psi = e^J \sum_{d=1}^{n_{\mathrm{det}}} \det A_d$ with a Jastrow factor e^J symmetric in the electrons and Slater matrices

$$A_d = \begin{pmatrix} \phi_{d1}(\mathbf{r}_1) & \cdots & \phi_{dn_{\mathsf{el}}}(\mathbf{r}_1) \\ \vdots & & \vdots \\ \phi_{d1}(\mathbf{r}_{n_{\mathsf{el}}}) & \cdots & \phi_{dn_{\mathsf{el}}}(\mathbf{r}_{n_{\mathsf{el}}}) \end{pmatrix}, \qquad \phi_{dj}(\mathbf{r}_i) = f_{dij} \sum_{I=1}^{n_{\mathsf{nu}}} \pi^{\sigma_i}_{djI} \exp(-e_{dijI}),$$

with exponents $e_{dijI} = |\zeta^{\sigma_i}_{djI}|\,\lVert\mathbf{r}_i - \mathbf{R}_I\rVert$. Row $i \in [n_{\mathsf{el}}]$ holds the orbitals $j \in [n_{\mathsf{el}}]$ at electron i. The index $d \in [n_{\mathrm{det}}]$ runs over the determinants and $I \in [n_{\mathsf{nu}}]$ over the nuclei. Each orbital is the network output f_{dij} times the envelope $\sum_I \pi^{\sigma_i}_{djI} \exp(-e_{dijI})$. The output f_{dij} depends on all electrons. The envelope parameters π^{σ}_{djI} and ζ^{σ}_{djI} are trained for each spin σ and taken at the spin σ_i of electron i.

H.1.1 Computational stability

Below we discuss the issues that can arise in float32 from the envelopes and from the derivative of the electron–nucleus distance, together with their solutions.

Envelope underflow. We keep $\log|\psi|$ and its derivatives finite in float32 by shifting the exponents in each row of the Slater matrix A_d above, a shift that leaves them exact. If a row of the shifted matrix $\tilde{A}_d$ still has largest absolute entry below the smallest normal float32 value 1.2×10^{-38}, its precision is lost and we set $\det A_d$ and its derivative to zero. Each entry of A_d is the network output f_{dij} times an envelope that decays exponentially in the distance of electron i from the nuclei, so without the

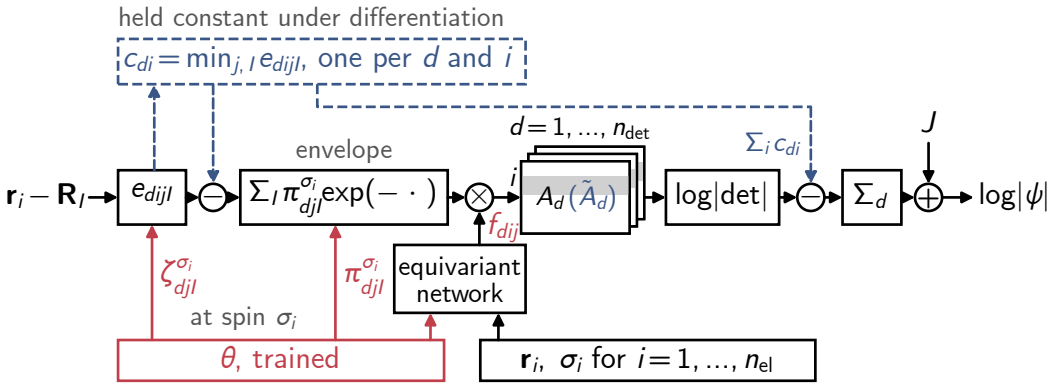


Figure 5: Computation of $\log|\psi|$ with the envelope shift, from the electron positions and spins and the trained parameters θ. The constant c_{di}, held fixed under differentiation, is subtracted from the exponents in row i of A_d and its sum over i from $\log|\det \tilde{A}_d|$.

shift the entries of an electron far from every nucleus are stored as zero and the derivative of $\log|\psi|$ becomes non-finite.

Consider an electron i far from every nucleus, so that every exponent e_{dijI} is large. Each term $\exp(-e_{dijI})$ falls below 1.2×10^{-38} once $e_{dijI} \gtrsim 87$ and is stored as zero once $e_{dijI} \gtrsim 103$, so row i is zero or subnormal in some or all A_d. The derivative of $\log|\det A_d|$ is the transposed inverse $(A_d^{-1})^\top$, and the inverse of a matrix with a zero or subnormal row is non-finite. A determinant that contributes nothing to ψ still passes this value on, because its zero weight multiplies a non-finite derivative.

We subtract the smallest exponent in row i of A_d, $c_{di} = \min_{j,I} e_{dijI}$, from every exponent e_{dijI} of that row (Figure 5). This multiplies row i by $e^{c_{di}}$ and gives the shifted matrix $\tilde{A}_d$, whose largest exponential in row i is $\exp(0) = 1$. The shift is the subtract-max trick of the softmax applied to the logits $-e_{dijI}$, whose largest value is $-c_{di}$, and it leaves $\log|\psi|$ and its derivatives exact. In the softmax the trick prevents overflow of large logits; here every logit $-e_{dijI}$ is at most zero, and for a distant electron the largest one, $-c_{di}$, is large and negative, so the subtraction raises every entry of row i and prevents underflow.

A determinant is linear in each row, so $\log|\det A_d| = \log|\det \tilde{A}_d| - \sum_i c_{di}$ for every value of c_{di}. We compute $\log|\det \tilde{A}_d|$ from rows stored at full precision and subtract $\sum_i c_{di}$ before the determinants are summed. Since the identity holds for every c_{di}, we hold c_{di} constant under differentiation. A second subtract-max, over the determinants instead of the entries of a row, sums the determinants:

$$x_d = \log|\det \tilde{A}_d| - \sum_{i=1}^{n_{\text{el}}} c_{di}, \quad M = \max_d x_d, \quad \log|\psi| = J + M + \log\Big|\sum_{d=1}^{n_{\text{det}}} \operatorname{sgn}(\det \tilde{A}_d)\, e^{x_d - M}\Big|.$$

The first shift keeps the entries of $\tilde{A}_d$ from underflowing, and the second keeps the terms e^{x_d} of the sum, lowered by $\sum_i c_{di}$, from underflowing.

Safe norm of the electron–nucleus distance. The reported runs were trained with the safe norm $\rho = \sqrt{r^2 + a^2}$ of the DeepQMC code [9], where a^2 is the float32 machine epsilon ($a = 3.45 \times 10^{-4}$ bohr), in place of each electron–nucleus distance $r = \|\mathbf{r}_i - \mathbf{R}_I\|$ in the exponents of the envelopes and in the input features. This helps with the undefined derivative $(\mathbf{r}_i - \mathbf{R}_I)/r$ of r where r vanishes, since that of ρ is finite everywhere.

While this choice may be numerically stable, it also induces rare jumps of the local energy, and the reported energies evaluate the network and the local energy with exact interparticle distances (Appendix J), which removes these jumps. The potential $-Z/r$ takes the exact distance r, so within about a of a nucleus the safe norm breaks the cancellation between the kinetic energy and the nuclear Coulomb term. For an envelope $e^{-\zeta\rho}$, the term $-\partial_r\psi/(r\psi)$ of the kinetic energy equals ζ/ρ and

is at most ζ/a, whereas for $e^{-\zeta r}$ it equals ζ/r and cancels $-Z/r$ at the nuclear cusp when $\zeta = Z$. A walker with an electron within about a of a nucleus thus has a local energy far below its value with exact distances, which makes the unclipped batch mean of the local energy jump at the step that samples it.

We conjecture that training with the exact distance r removes these jumps and gives more accurate energies, since the envelopes then keep this cancellation at every distance from a nucleus. We also suspect that the safe norm is a broader precaution than the computation needs, because the FermiNet code [34] keeps the exact electron–nucleus distance and masks only the difference of each electron with itself, and the same mask appears applicable here.

H.2 Samplers

The training chains follow the Metropolis sampler of DeepQMC v1.2.0 [9, 33]: each network ψ_k has its own chain of $4{,}096$ walkers, which samples p_k (Section 4.1). A proposal moves all electrons of a walker by a Gaussian step of width τ and is accepted with probability $\min\{1, |\psi_k(\mathbf{x}')|^2/|\psi_k(\mathbf{x})|^2\}$. Each training step takes 30 Metropolis steps, and after each Metropolis step τ is multiplied by $\max(a_{\mathrm{acc}}, 0.05)/0.57$, where a_{acc} is the acceptance rate over the walkers of the chain; τ starts at 1 bohr. The chains start from the walkers at the end of pretraining and are equilibrated under the pretrained networks before training.

H.3 Natural-gradient descent with KFAC

Optimization uses per-state natural gradient descent: each per-state gradient is left-multiplied by the inverse Fisher information F_k^{-1}, with

$$F_k := \mathbb{E}_{p_k}[\nabla_\theta \log p_k(\mathbf{x})\, \nabla_\theta \log p_k(\mathbf{x})^\intercal].$$

The parameter update direction is

$$\Delta\theta = 2 \sum_{k=1}^{K} F_k^{-1}\, \mathbb{E}_{p_k(\mathbf{x})}\big[(g_k(\mathbf{x}) - \mathbb{E}_{p_k}[g_k])\, \nabla_\theta \log |\psi_k|\big]. \tag{17}$$

In practice each F_k is approximated by KFAC [21], applied per state. The factorization by state holds because each network samples its own Born density (Section 3.4): the walkers $\mathbf{x} \sim p_k$ enter only F_k, so the KFAC blocks decouple across k.

Hyperparameters. The optimizer uses an inverse-time learning-rate schedule $\eta_n = 0.05/(1 + n/5 \times 10^4)$. KFAC-specific settings: norm constraint 10^{-3}, constant damping 10^{-3}, Fisher factors computed exactly from the registered output $\log|\psi_k|$ rather than from sampled targets, inverse update period 1, no burn-in. The norm constraint is applied to each spin sector separately: the update of the networks of a sector is multiplied by $\min\{1, (10^{-3}/(\eta_n^2 G))^{1/2}\}$, where G is the sum over the networks of that sector of the inner product of each network's preconditioned gradient with its gradient.

H.4 Per-atom setup

Sector decision. Every atom runs NestedLoRA-VMC on $\mathcal{T} = -(\mathcal{H} + \hat{S}^2)$ in one m_S sector per spin multiplicity, and the states of all sectors are merged by energy (Section 4.1). The augmented Hamiltonian shifts each state by $S(S{+}1)$ Ha. This shift differs by at least 2 Ha between two values of S in one sector, far above the excitation energies in the $K{=}10$ window (Table 9), so a single-sector run with the spin penalty would lift every state with $S > |m_S|$ out of the $K{=}10$ window. One m_S sector per spin multiplicity is also the recipe of Szabó et al. [37] (their Section 2.1.3), who fix m_S to reach each higher spin multiplicity and keep the spin penalty on.

H.5 Pretraining

Per-state CASSCF targets. Network k of a sector is pretrained to the leading determinants of the k-th root of the sector's state-averaged CASSCF.

Table 6: Per-atom physical configuration. Each row lists the terms one spin sector $m_S = S$ (Section 4.1) holds, in NIST term-centroid order, with the number of networks assigned to each term in parentheses. A fully covered term gets its $2L+1$ orbital sub-states; a term the $K=10$ window cuts keeps fewer (e.g. Li's ${}^2D(2)$ keeps 2 of its 5). The CAS (N_e, N_o) is chosen per atom by counting the orbitals of the $K=10$ configurations (Li, C, Ne) or by enlarging the valence shell by $\{3s, 3p\}$ (Be, B, N, O, F).

atom	m_S	$K=10$ block	CAS (N_e, N_o)
Li	$1/2$	${}^2S(1)$, ${}^2P^\circ(3)$, ${}^2S(1)$, ${}^2P^\circ(3)$, ${}^2D(2)$	$(1, 13)$
Be	0	${}^1S(1)$, ${}^1P^\circ(3)$, ${}^1S(1)$, ${}^1D(1)$	$(2, 8)$
	1	${}^3P^\circ(3)$, ${}^3S(1)$	
B	$1/2$	${}^2P^\circ(3)$, ${}^2S(1)$, ${}^2D(3)$	$(3, 8)$
	$3/2$	${}^4P(3)$	
C	0	${}^1D(5)$, ${}^1S(1)$	$(4, 4)$
	1	${}^3P(3)$	
	2	${}^5S^\circ(1)$	
N	$1/2$	${}^2D^\circ(5)$, ${}^2P^\circ(3)$	$(5, 8)$
	$3/2$	${}^4S^\circ(1)$, ${}^4P(1)$	
O	0	${}^1D(5)$, ${}^1S(1)$	$(6, 8)$
	1	${}^3P(3)$	
	2	${}^5S^\circ(1)$	
F	$1/2$	${}^2P^\circ(3)$, ${}^2P(3)$	$(7, 8)$
	$3/2$	${}^4P(3)$, ${}^4P^\circ(1)$	
Ne[†]	0	${}^1S(1)$, ${}^2[1/2]^\circ(3)$	$(8, 8)$
	1	${}^2[3/2]^\circ(3)$, ${}^2[1/2](1)$, ${}^2[5/2](1)$, ${}^2[3/2](1)$	

[†]NIST names each level by its total angular momentum J, which spin–orbit coupling conserves, and by the basis state with the largest weight in that level [32]. For Li–F this basis state is an LS term, and for the excited levels of Ne it is a $J_1 l$-coupled state ${}^{2S_2+1}[K]^\pi_J$: the core's angular momentum J_1 and the outer electron's orbital angular momentum l give K, and the leading 2 is the multiplicity of the outer electron's spin.

Pretraining sampler. As in FermiNet's excited-state pretraining [34], each state's pretraining walkers come from a chain on the density of its CASSCF reference that continues from one iteration to the next, whereas the penalty baseline [37] restarts its chains at every iteration.

Optimization. The pretrain loss is the mean-squared error between the network orbital matrices and the per-state CASSCF target. We optimize with Lamb at learning rate 3×10^{-4} (β_1=0.9, β_2=0.999) for 10,000 steps in all reported runs.

H.6 NaN handling in training

A non-finite $\log \psi_k$, local energy or cross-state ratio at a single walker makes every batch mean it enters non-finite. A non-finite update then leaves the parameters non-finite at every later step. Although the reported runs kept $\log \psi_k$, the local energy and the parameters finite with the envelope shift of Appendix H.1.1, we adopt two safeguards from the NES-VMC stability handler of Pfau et al. [27, Section S8]. (i) Per-walker exclusion: any walker with non-finite $\log \psi_k$, local energy or cross-state ratio is dropped from every batch reduction in g_k, including the analytic-eigenvalue solver inputs and the EMAs. (ii) Step undo with a failure counter: a step that leaves the parameters, the optimizer state or the statistics fed back to the next step non-finite is undone, as FermiNet does [34]. A step with a non-finite walker or an undone update increments a counter of consecutive failures, and at 100 the run restarts from its last checkpoint. The other two safeguards of that handler address the joint matrix of NES-VMC and larger systems. Extra MCMC on zero entries $\psi_i(\mathbf{x}^j)$ treats a sparse joint matrix with nonzero $\det \Psi$. In NestedLoRA-VMC each network k samples its walkers from $|\psi_k|^2$, so ψ_k is nonzero at every one of them. The median-of-minibatches energy estimator was introduced for systems with > 25 electrons, beyond the first-row atoms studied here.

Table 7: Cost of one training step by component. $\mathcal{C}_{\mathrm{bb}}$ is one forward evaluation of the backbone, $\mathcal{C}_{\mathrm{post}}$ that of one state's orbitals, envelopes and determinants, and $\mathcal{C}_{\mathrm{eval}} = \mathcal{C}_{\mathrm{bb}} + \mathcal{C}_{\mathrm{post}}$; a training step takes n_{MH} Metropolis steps. The last column gives the leading order in K and n_{el}. Bold: the Laplacian, the largest single evaluation.

	Component	Joint backbone, K heads	K disjoint networks	Order
NES-VMC	Laplacian	$\boldsymbol{3n_{\mathsf{el}}(K\mathcal{C}_{\mathrm{bb}} + K^2\mathcal{C}_{\mathrm{post}})}$	$\boldsymbol{3n_{\mathsf{el}}K^2\mathcal{C}_{\mathrm{eval}}}$	$K^2{n_{\mathsf{el}}}^4$
	Sampling	$n_{\mathrm{MH}}(K\mathcal{C}_{\mathrm{bb}} + K^2\mathcal{C}_{\mathrm{post}})$	$n_{\mathrm{MH}}K^2\mathcal{C}_{\mathrm{eval}}$	$n_{\mathrm{MH}}K^2{n_{\mathsf{el}}}^3$
	Inverse of Ψ	K^3	K^3	K^3
NestedLoRA-VMC, penalty	Laplacian	$\boldsymbol{3n_{\mathsf{el}}K\mathcal{C}_{\mathrm{eval}}}$	$\boldsymbol{3n_{\mathsf{el}}K\mathcal{C}_{\mathrm{eval}}}$	$K{n_{\mathsf{el}}}^4$
	Sampling	$n_{\mathrm{MH}}K\mathcal{C}_{\mathrm{eval}}$	$n_{\mathrm{MH}}K\mathcal{C}_{\mathrm{eval}}$	$n_{\mathrm{MH}}K{n_{\mathsf{el}}}^3$
	Overlaps	$K^2\mathcal{C}_{\mathrm{post}}$	$K^2\mathcal{C}_{\mathrm{eval}}$	$K^2{n_{\mathsf{el}}}^3$

I Per-step cost analysis

NestedLoRA-VMC evaluates K Laplacians per training step and NES-VMC evaluates K^2. This section counts the cost of one step for both methods and measures the wall time of NestedLoRA-VMC on Ne as K grows.

One forward evaluation of a network costs $\mathcal{C}_{\mathrm{eval}} = \mathcal{C}_{\mathrm{bb}} + \mathcal{C}_{\mathrm{post}}$, the sum over the two stages of Appendix H.1, with $\mathcal{C}_{\mathrm{bb}} \sim {n_{\mathsf{el}}}^2$ and $\mathcal{C}_{\mathrm{post}} \sim {n_{\mathsf{el}}}^3$ at fixed depth and width. The backbone of a network is its permutation-equivariant layers, of order ${n_{\mathsf{el}}}^2$ from the self-attention of Psiformer or the symmetric features of FermiNet. The orbitals, envelopes and determinants of one state have the order ${n_{\mathsf{el}}}^3$ from the determinants. A Laplacian costs $\mathcal{C}_{\mathrm{lap}} \sim 3n_{\mathsf{el}}\,\mathcal{C}_{\mathrm{eval}} \sim {n_{\mathsf{el}}}^4$ (Appendix I.1). A training step takes $n_{\mathrm{MH}} = 30$ Metropolis steps (Appendix H.2).

I.1 Forward-Laplacian primitive (folx)

We compute every Laplacian $\Delta f = \sum_{k=1}^{3n_{\mathsf{el}}} \partial^2 f / \partial x_k^2$ with the forward Laplacian of `folx` [5], at about half the cost of the standard Hessian-diagonal loop. DeepQMC [9] and FermiNet [34] by default run $3n_{\mathsf{el}}$ forward-mode Jacobian–vector products through the gradient $\nabla_{\mathbf{x}} f$, each about one forward and one backward pass, $\sim 2 \cdot 3n_{\mathsf{el}}\,\mathcal{C}_{\mathrm{eval}}$ in total. The forward Laplacian instead propagates $(h, \nabla h, \Delta h)$ with $\nabla h \in \mathbb{R}^{3n_{\mathsf{el}} \times d_l}$ through the network in one pass, and the Jacobian products $(3n_{\mathsf{el}} \times d_l) \cdot (d_l \times d_{l+1})$ of its layers sum to $\mathcal{C}_{\mathrm{lap}} \sim 3n_{\mathsf{el}}\,\mathcal{C}_{\mathrm{eval}}$.

The sparse mode of `folx` shrinks the Jacobian to $(s \times d_l)$ in layers whose outputs depend on at most s input coordinates, which holds with $s = 3$ only in the leading per-electron embedding and MLP. From the first self-attention of Psiformer [38] or the first mean over electrons of FermiNet [26] on, the features of an electron depend on all $3n_{\mathsf{el}}$ coordinates, so the Laplacian cost stays at $\sim 3n_{\mathsf{el}}\,\mathcal{C}_{\mathrm{eval}}$.

I.2 Cost decomposition: Hamiltonian coupling and ansatz parameterization

Table 7 counts every component of a step under the two ansatz parameterizations, a joint backbone with K heads and K disjoint networks.

Hamiltonian coupling: off-diagonal Laplacian count. The K^2 Laplacians of NES-VMC come from its local-energy matrix $\mathsf{E}_{\mathsf{loc}}(\mathbf{X}) = \Psi(\mathbf{X})^{-1}(\mathcal{H}\Psi)(\mathbf{X})$, which contains $(\mathcal{H}\psi_k)(\mathbf{x}^i)$ at every state-walker pair (i, k). In NestedLoRA-VMC and the penalty methods the cross term is the overlap $\langle \psi_k \mid \psi_\ell \rangle$, a forward-only evaluation, so only the K diagonal local energies $\mathcal{H}\psi_k/\psi_k$ need a Laplacian, a factor K fewer than NES-VMC. The inverse Ψ^{-1} and the product $\Psi^{-1}(\mathcal{H}\Psi)$ add $O(K^3)$ per step to NES-VMC, below its Laplacian cost $K^2{n_{\mathsf{el}}}^4$ for $n_{\mathsf{el}} \gtrsim K$.

Ansatz parameterization and walker-target factorization. The parameterization changes the overlap term of NestedLoRA-VMC, and the Laplacian and sampling terms of NES-VMC. The walker target $|\psi_1|^2 \cdots |\psi_K|^2$ of NestedLoRA-VMC and the penalty methods factorizes into K Markov chains, so sampling and Laplacians evaluate each network at its own walkers under either parameterization. The overlaps evaluate every network at every walker set, at $K^2\,\mathcal{C}_{\mathrm{post}}$ with a joint backbone, whose K heads share the backbone features of a walker set, and at $K^2\,\mathcal{C}_{\mathrm{eval}}$ with the K

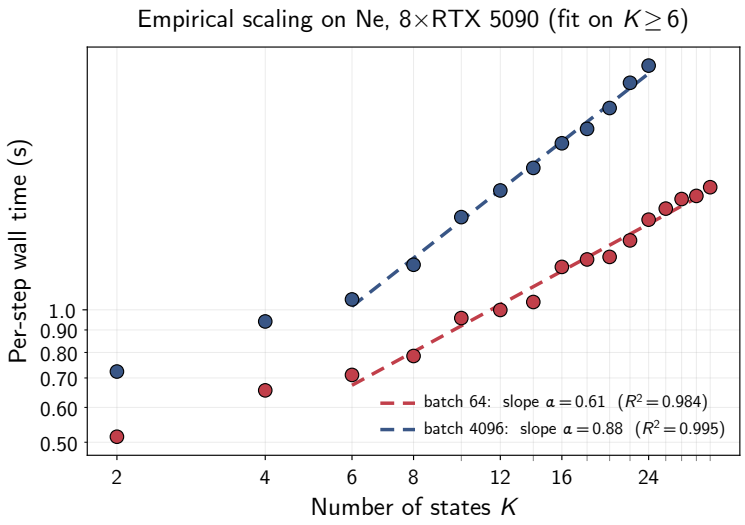


Figure 6: Empirical per-step wall time of NestedLoRA-VMC's main optimization loop on Ne, $8\times$RTX 5090, FP32, at batches of $4{,}096$ and 64. Dashed lines are least-squares power-law fits. The spin penalty is off.

disjoint networks that NestedLoRA-VMC uses, following Szabó et al. [37]. The target $|\det \Psi|^2$ of NES-VMC couples the K walker sets, so its sampling and Laplacians evaluate the full matrix $\Psi(\mathbf{X})$, at $K\,\mathcal{C}_{\mathrm{bb}} + K^2\,\mathcal{C}_{\mathrm{post}}$ with a joint backbone, as in the NES-VMC reference implementation [34], and at $K^2\,\mathcal{C}_{\mathrm{eval}}$ with disjoint networks.

Cost of the spin penalty. For the first-row atoms, with $n_{\mathrm{el}} \leq 10$, the Löwdin sum of Appendix G.3 costs less than the K Laplacians of a step and much less than the $K(K-1)$ Laplacians that NestedLoRA-VMC saves against NES-VMC. Each of its $n_\uparrow n_\downarrow$ terms is one forward evaluation of ψ_k, because every orbital of the Psiformer depends on all electron positions. Evaluated at every walker, it adds $Kn_\uparrow n_\downarrow \leq K{n_{\mathrm{el}}}^2/4$ forward evaluations per walker and step to the forward counts of Table 7, of order $K{n_{\mathrm{el}}}^5$ against the order $K{n_{\mathrm{el}}}^4$ of the Laplacians.

In larger systems the spin term can therefore cost more than the Laplacians, but a small random subsample of the walkers suffices to estimate it, because the eigenvalues $S(S+1)$ of neighboring spin sectors differ by at least 2, far above the excitation energies. Evaluating the spin term on a random eighth of the walkers keeps its cost below that of the Laplacians up to about a hundred electrons, $n_{\mathrm{el}} < 96$. On Ne, training with the spin term on a random eighth of the walkers reached the same accuracy as training with it on every walker.

I.3 Empirical wall-time scaling on Ne

We confirm the scaling in K claimed in Section 3.4 by sweeping K on Ne. Measurements run on a single host with $8\times$ RTX 5090 (FP32, JAX 0.9.2). For each value of K we run 100 main-loop warm-up steps followed by 100 measurement steps. The sweep places all K networks in one spin sector and runs the training step with the spin penalty off. With two or more sectors, overlaps are evaluated only between networks of the same sector, and the spin penalty adds the cost given in Appendix I.2. Figure 6 reports two batch sizes. At the training batch of $4{,}096$ the walkers no longer fit in memory beyond $K{=}24$. The exponent fitted at this batch in Figure 6 has standard error 0.02 (95% interval ± 0.04) and describes the measured range $K{=}6$ to 24 on this host.

J Per-state errors across first-row atoms

This section reports the per-state errors behind the per-atom mean absolute error (MAE, in mHa) of Table 2, for NestedLoRA-VMC, NES-VMC, and Szabó et al. [37] against $(2J{+}1)$-weighted NIST

term centroids on each of the eight first-row atoms at $K{=}10$. Table 9 gives the per-state breakdown of one run per atom, whose MAE is in Table 8.

Per-state stderr and MAE interval. Each per-state energy E_k is the unweighted mean of the per-state batch mean of the local energy over training steps $97{,}000$ to $99{,}999$, and its stderr is the standard deviation of the means of 10 contiguous blocks of 300 steps, divided by $\sqrt{10}$. For Ne seed 1, read at step $99{,}000$, the mean runs over steps $97{,}000$ to $98{,}999$ and the stderr uses its last 6 blocks of 300 steps. The local energy is unclipped and is evaluated on each step's walkers with exact interparticle distances (Appendix H.1). Blocks of 100 steps give the same stderr: over the states of the runs in Table 2, the ratio of the 10-block stderr to the 30-block stderr has median 1.00, within the precision of a standard deviation estimated from 10 blocks.

The interval of the MAE comes from a block bootstrap, because the sampling error of the anchor E_1 enters every per-state error $\varepsilon_k = (E_k - E_1) - t_k$, with t_k the target excitation energy. Each of $4{,}000$ resamples draws the blocks with replacement, the same blocks for every state, and recomputes the MAE from the resampled per-state means: it sorts them, subtracts the lowest and scores the excitations against the targets t_k. The 95% interval runs from the 2.5th to the 97.5th percentile of the resampled MAE. Table 8 gives the interval of every run in Table 2; blocks of 100 and of 300 steps give nearly the same interval. The interval measures the sampling noise of one run, and on B it is narrower than the spread across the three seeds.

Table 8: MAE (mHa) of each run entering Table 2, with its 95% block-bootstrap interval for blocks of 100 and of 300 training steps over the last $3{,}000$ steps ($2{,}000$ for Ne seed 1). Steps is the number of training steps the run had completed when it was read.

Atom	Seed	Steps	MAE	Blocks of 100	Blocks of 300
Li	1	100,000	0.046	[0.034, 0.073]	[0.036, 0.070]
Li	2	100,000	0.066	[0.049, 0.093]	[0.051, 0.086]
Li	3	100,000	0.059	[0.046, 0.083]	[0.046, 0.086]
Be	1	100,000	0.019	[0.015, 0.042]	[0.015, 0.042]
Be	2	100,000	0.035	[0.028, 0.054]	[0.025, 0.056]
Be	3	100,000	0.036	[0.022, 0.064]	[0.026, 0.059]
B	1	100,000	0.498	[0.489, 0.533]	[0.487, 0.531]
B	2	100,000	0.521	[0.507, 0.558]	[0.505, 0.561]
B	3	100,000	0.915	[0.901, 0.954]	[0.897, 0.962]
C	1	100,000	0.177	[0.147, 0.207]	[0.150, 0.201]
C	2	100,000	0.193	[0.148, 0.243]	[0.139, 0.260]
N	1	100,000	0.276	[0.213, 0.338]	[0.214, 0.337]
O	1	100,000	0.265	[0.234, 0.368]	[0.243, 0.353]
F	1	100,000	0.647	[0.555, 0.781]	[0.561, 0.780]
Ne	1	99,000	1.944	[1.746, 2.139]	[1.665, 2.118]
Ne	2	100,000	1.153	[1.051, 1.280]	[1.068, 1.253]

B: tenth state. On B, the tenth states of NestedLoRA-VMC and of NES-VMC lie near the $2s^2 3p$ $^2P^\circ$ level ($\Delta E_{\mathrm{NIST}} = 0.22145$ Ha), not at a third ^2D substate. The $k{=}10$ row of Table 9 scores both against the ^2D centroid.

Table 9: Per-state K=10 results across first-row atoms (Li–Ne), one run per atom: seed 1 (Ne: seed 2), read at the step given in Table 8. Within an atom, the rows list the target terms by ascending ΔE_{NIST}, each repeated by its number of networks in Table 6, against the run's ten lowest states in ascending energy. E is the per-state mean defined above, with its stderr in parentheses. Errors in mHa against the NIST term centroids of Section 4.1; baseline errors from Pfau et al. [27, Table S2] and Szabó et al. [37, SI Table S1]. Config. omits the closed $1s^2$ shell (Ne: $1s^22s^2$); each column is rounded from unrounded values.

				This work			NIST		Baselines (mHa)	
System	k	Config.	Term	E (Ha)	$\langle\hat{S}^2\rangle$	ΔE (Ha)	ΔE_{NIST} (Ha)	Err (mHa)	NES-VMC [27]	Szabó [37]
Li	1	2s	^2S	-7.47805(1)	0.750	0.00000	0.00000	+0.00	—	—
Li	2	2p	$^2P^\circ$	-7.41015(2)	0.750	0.06790(2)	0.06791	+0.00	+0.00	+0.05
Li	3	2p	$^2P^\circ$	-7.41014(2)	0.750	0.06791(2)	0.06791	+0.00	+0.01	+0.07
Li	4	2p	$^2P^\circ$	-7.41010(1)	0.750	0.06795(2)	0.06791	+0.04	+0.01	+0.14
Li	5	3s	^2S	-7.35408(3)	0.750	0.12397(3)	0.12396	+0.01	+0.01	+0.22
Li	6	3p	$^2P^\circ$	-7.33709(1)	0.750	0.14096(2)	0.14091	+0.05	+0.01	+0.57
Li	7	3p	$^2P^\circ$	-7.33708(1)	0.750	0.14097(2)	0.14091	+0.06	+0.02	+2.12
Li	8	3p	$^2P^\circ$	-7.33704(2)	0.750	0.14101(2)	0.14091	+0.10	+0.03	+2.25
Li	9	3d	^2D	-7.33552(2)	0.750	0.14253(2)	0.14254	-0.01	+0.02	+1.45
Li	10	3d	^2D	-7.33538(2)	0.750	0.14267(2)	0.14254	+0.14	+0.03	+2.61
Be	1	$2s^2$	^1S	-14.66737(2)	0.000	0.00000	0.00000	+0.00	—	—
Be	2	2s2p	$^3P^\circ$	-14.56723(2)	2.000	0.10013(2)	0.10015	-0.01	-0.05	-0.04
Be	3	2s2p	$^3P^\circ$	-14.56722(1)	2.000	0.10015(2)	0.10015	+0.00	-0.04	+0.01
Be	4	2s2p	$^3P^\circ$	-14.56721(1)	2.000	0.10016(2)	0.10015	+0.01	-0.04	+0.02
Be	5	2s2p	$^1P^\circ$	-14.47344(1)	0.000	0.19393(2)	0.19394	-0.01	-0.05	-0.09
Be	6	2s2p	$^1P^\circ$	-14.47343(2)	0.000	0.19393(2)	0.19394	-0.01	-0.03	+0.07
Be	7	2s2p	$^1P^\circ$	-14.47342(2)	0.000	0.19395(2)	0.19394	+0.01	-0.03	+0.07
Be	8	2s3s	^3S	-14.43002(2)	2.000	0.23735(3)	0.23730	+0.05	+0.03	+0.07
Be	9	2s3s	^1S	-14.41819(3)	0.000	0.24918(3)	0.24913	+0.05	+0.10	+1.27
Be	10	$2p^2$	^1D	-14.40821(2)	0.000	0.25916(3)	0.25917	-0.01	-0.00	+0.14
B	1	$2s^22p$	$^2P^\circ$	-24.65381(2)	0.750	0.00000	0.00000	+0.00	—	—
B	2	$2s^22p$	$^2P^\circ$	-24.65381(2)	0.750	0.00000(3)	0.00000	+0.00	+0.02	+0.20
B	3	$2s^22p$	$^2P^\circ$	-24.65379(2)	0.750	0.00001(2)	0.00000	+0.01	+0.02	+0.20
B	4	$2s2p^2$	^4P	-24.52201(2)	3.750	0.13179(3)	0.13153	+0.27	+0.33	-0.47
B	5	$2s2p^2$	^4P	-24.52201(1)	3.750	0.13180(2)	0.13153	+0.27	+0.33	-0.42
B	6	$2s2p^2$	^4P	-24.52200(2)	3.750	0.13180(2)	0.13153	+0.27	+0.34	-0.33
B	7	$2s^23s$	^2S	-24.47139(3)	0.750	0.18242(4)	0.18239	+0.03	+0.10	-0.09
B	8	$2s2p^2$	^2D	-24.43588(3)	0.750	0.21793(4)	0.21801	-0.08	+0.07	+0.89
B	9	$2s2p^2$	^2D	-24.43584(4)	0.750	0.21797(4)	0.21801	-0.03	+0.07	+1.39
B	10	$2s2p^2$	^2D	-24.43229(3)	0.750	0.22151(3)	0.21801	+3.51	+3.62	+1.79
C	1	$2s^22p^2$	^3P	-37.84489(2)	2.000	0.00000	0.00000	+0.00	—	—
C	2	$2s^22p^2$	^3P	-37.84482(3)	2.000	0.00007(2)	0.00000	+0.07	+0.04	+0.00
C	3	$2s^22p^2$	^3P	-37.84472(4)	2.000	0.00016(4)	0.00000	+0.16	+0.06	+0.34
C	4	$2s^22p^2$	^1D	-37.79844(4)	0.000	0.04645(5)	0.04631	+0.14	+0.03	+0.19
C	5	$2s^22p^2$	^1D	-37.79843(2)	0.000	0.04645(2)	0.04631	+0.15	+0.03	+0.54
C	6	$2s^22p^2$	^1D	-37.79841(3)	0.000	0.04648(3)	0.04631	+0.17	+0.04	+0.79
C	7	$2s^22p^2$	^1D	-37.79839(2)	0.000	0.04650(3)	0.04631	+0.19	+0.04	+0.89
C	8	$2s^22p^2$	^1D	-37.79837(4)	0.000	0.04652(5)	0.04631	+0.21	+0.11	+1.05
C	9	$2s^22p^2$	^1S	-37.74631(3)	0.000	0.09857(3)	0.09850	+0.07	+0.04	+1.00
C	10	$2s2p^3$	$^5S^\circ$	-37.69174(1)	6.000	0.15315(2)	0.15357	-0.42	-0.40	-0.52
N	1	$2s^22p^3$	$^4S^\circ$	-54.58905(3)	3.750	0.00000	0.00000	+0.00	—	—
N	2	$2s^22p^3$	$^2D^\circ$	-54.50120(4)	0.750	0.08784(5)	0.08761	+0.23	+0.00	+1.89
N	3	$2s^22p^3$	$^2D^\circ$	-54.50120(5)	0.750	0.08785(6)	0.08761	+0.24	+0.03	+2.19
N	4	$2s^22p^3$	$^2D^\circ$	-54.50117(4)	0.750	0.08787(4)	0.08761	+0.27	+0.05	+2.79
N	5	$2s^22p^3$	$^2D^\circ$	-54.50113(5)	0.750	0.08792(6)	0.08761	+0.31	+0.05	+3.09
N	6	$2s^22p^3$	$^2D^\circ$	-54.50103(6)	0.750	0.08802(8)	0.08761	+0.41	+0.08	+3.19
N	7	$2s^22p^3$	$^2P^\circ$	-54.45746(6)	0.750	0.13159(7)	0.13140	+0.19	-0.00	+3.40
N	8	$2s^22p^3$	$^2P^\circ$	-54.45737(3)	0.750	0.13168(5)	0.13140	+0.28	+0.02	+4.30
N	9	$2s^22p^3$	$^2P^\circ$	-54.45726(6)	0.750	0.13178(6)	0.13140	+0.38	+0.02	+5.10
N	10	$2s^22p^2(^3P)3s$	^4P	-54.20916(4)	3.750	0.37988(5)	0.37971	+0.18	+1.16	+0.59
O	1	$2s^22p^4$	^3P	-75.06706(4)	2.000	0.00000	0.00000	+0.00	—	—

continued on next page

Table 9 – continued from previous page

System	k	Config.	Term	This work E (Ha)	This work $\langle\hat{S}^2\rangle$	This work ΔE (Ha)	NIST ΔE_{NIST} (Ha)	NIST Err (mHa)	Baselines (mHa) NES-VMC [27]	Baselines (mHa) Szabó [37]
O	2	$2s^2 2p^4$	^3P	-75.06705(3)	2.000	0.00001(5)	0.00000	+0.01	+0.00	+0.59
O	3	$2s^2 2p^4$	^3P	-75.06703(3)	2.000	0.00003(7)	0.00000	+0.03	+0.06	+0.79
O	4	$2s^2 2p^4$	^1D	-74.99483(6)	0.000	0.07223(7)	0.07194	+0.29	+0.02	-2.85
O	5	$2s^2 2p^4$	^1D	-74.99479(13)	0.000	0.07227(15)	0.07194	+0.32	+0.14	+1.35
O	6	$2s^2 2p^4$	^1D	-74.99477(8)	0.000	0.07229(9)	0.07194	+0.35	+0.14	+1.65
O	7	$2s^2 2p^4$	^1D	-74.99476(5)	0.000	0.07230(5)	0.07194	+0.36	+0.14	+1.65
O	8	$2s^2 2p^4$	^1D	-74.99470(5)	0.000	0.07236(6)	0.07194	+0.42	+0.20	+2.25
O	9	$2s^2 2p^4$	^1S	-74.91316(5)	0.000	0.15390(6)	0.15362	+0.28	+0.07	+1.68
O	10	$2s^2 2p^3(^4S^\circ)3s$	$^5S^\circ$	-74.73097(2)	6.000	0.33609(3)	0.33576	+0.33	+1.61	+0.06
F	1	$2s^2 2p^5$	$^2P^\circ$	-99.73354(8)	0.750	0.00000	0.00000	+0.00	—	—
F	2	$2s^2 2p^5$	$^2P^\circ$	-99.73343(8)	0.750	0.00012(15)	0.00000	+0.12	+0.16	+1.30
F	3	$2s^2 2p^5$	$^2P^\circ$	-99.73342(7)	0.750	0.00013(9)	0.00000	+0.13	+0.21	+1.70
F	4	$2s^2 2p^4(^3P)3s$	^4P	-99.26623(9)	3.750	0.46731(13)	0.46673	+0.59	+1.20	+0.17
F	5	$2s^2 2p^4(^3P)3s$	^4P	-99.26618(5)	3.750	0.46736(9)	0.46673	+0.63	+1.30	+1.17
F	6	$2s^2 2p^4(^3P)3s$	^4P	-99.26602(7)	3.750	0.46752(11)	0.46673	+0.79	+1.42	+1.27
F	7	$2s^2 2p^4(^3P)3s$	^2P	-99.25567(11)	0.750	0.47788(14)	0.47707	+0.81	+1.04	+1.93
F	8	$2s^2 2p^4(^3P)3s$	^2P	-99.25564(7)	0.750	0.47790(10)	0.47707	+0.83	+1.04	+1.93
F	9	$2s^2 2p^4(^3P)3s$	^2P	-99.25554(5)	0.750	0.47800(9)	0.47707	+0.93	+1.06	+2.23
F	10	$2s^2 2p^4(^3P)3p$	$^4P^\circ$	-99.20465(7)	3.750	0.52890(5)	0.52790	+0.99	+9.68	-0.90
Ne	1	$2p^6$	^1S	-128.93691(5)	0.000	0.00000	0.00000	+0.00	—	—
Ne	2	$2p^5(^2P^\circ_{3/2})3s$	$^2[3/2]^\circ$	-128.32451(11)	2.000	0.61240(15)	0.61145	+0.95	+1.40	+2.95
Ne	3	$2p^5(^2P^\circ_{3/2})3s$	$^2[3/2]^\circ$	-128.32438(8)	2.000	0.61253(11)	0.61145	+1.08	+1.40	+4.35
Ne	4	$2p^5(^2P^\circ_{3/2})3s$	$^2[3/2]^\circ$	-128.32409(7)	2.000	0.61283(7)	0.61145	+1.37	+1.40	+5.05
Ne	5	$2p^5(^2P^\circ_{1/2})3s$	$^2[1/2]^\circ$	-128.31793(8)	0.000	0.61898(11)	0.61794	+1.05	+1.56	-0.44
Ne	6	$2p^5(^2P^\circ_{1/2})3s$	$^2[1/2]^\circ$	-128.31782(8)	0.000	0.61909(11)	0.61794	+1.16	+1.56	-0.14
Ne	7	$2p^5(^2P^\circ_{1/2})3s$	$^2[1/2]^\circ$	-128.31771(11)	0.000	0.61920(12)	0.61794	+1.27	+1.71	+0.76
Ne	8	$2p^5(^2P^\circ_{3/2})3p$	$^2[1/2]$	-128.26008(10)	2.000	0.67684(12)	0.67854	-1.71	+2.14	+5.86
Ne	9	$2p^5(^2P^\circ_{3/2})3p$	$^2[5/2]$	-128.25296(12)	2.000	0.68396(12)	0.68220	+1.75	+3.95	+4.20
Ne	10	$2p^5(^2P^\circ_{1/2})3p$	$^2[3/2]$	-128.25240(12)	2.000	0.68451(14)	0.68456	-0.05	+2.42	+2.04

K Compute resources and asset licenses

Compute resources. Each run of Table 2 trains for $100{,}000$ KFAC steps at an electron batch of $4{,}096$ on one to four NVIDIA V100 (SXM2, 32 GB), two NVIDIA A40, or four NVIDIA H100 NVL. A training step took 2.8 to 9.3 s, depending on the atom and the cards, and these runs took over $3{,}600$ GPU-hours, estimated as the per-step time times the number of steps and cards. CASSCF pretraining of Ne took 47 minutes on four H100 NVL and 140 minutes on two A40. Preliminary work used about $2{,}880$ A6000-hours and $2{,}880$ A100-hours. The wall-time sweep of Appendix I.3 ran on one host with $8\times$ RTX 5090.

Asset licenses. This work uses the following open-source software, used under their respective licenses:

- **DeepQMC** v1.2.0 [9, 33] (MIT, `https://github.com/deepqmc/deepqmc`): base code, forked and modified for this work. The modifications independent of the LoRA objective are the pretraining sampler (Appendix H.5), the exact distances in the reported energies (Appendix H.1), and the handling of envelope underflow and non-finite steps (Appendix H.6).
- **JAX** v0.9.2 (Apache 2.0).
- **KFAC-JAX** v0.0.8 (Apache 2.0): second-order optimizer.
- **folx** v0.2.19 (MIT): forward-Laplacian primitive.

No new datasets are introduced. Source code for NestedLoRA-VMC is available at `https://github.com/MinchanJeong/NestedLoRA-VMC` under the MIT license.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the contributions; Sections 3 (theory) and 4 (first-row $K{=}10$ empirics) substantiate them.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Limitations section in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Theorem D.1 is stated with its assumptions and proved in Appendix D.3, and Corollary D.1 in Appendix D.4; Lemma E.5 is proved in Appendix E.1.4, the analytic eigenvalue update is derived in Appendix E.2, the LoRA θ-gradient in Appendix E.1.2, and Proposition F.1 is proved in Appendix F.2.1.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 fixes the operating recipe (ansatz, sampler, pretraining, spin sector, schedule); Psiformer architecture, per-atom setup, pretraining, spin penalty, state ordering, and NaN handling are detailed in Appendices H.1, H.4, H.5, G.3, G.2, H.6. Code is available at `https://github.com/MinchanJeong/NestedLoRA-VMC`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code, training scripts, and reproduction instructions are available at `https://github.com/MinchanJeong/NestedLoRA-VMC` (Appendix K).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 reports the training settings (B=4096, 100,000 KFAC steps, K=10) and Table 6 the spin sectors of each atom; the clipping width w=10 is in Appendix F.1, the optimizer hyperparameters in Appendix H, and the other implementation details in Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 2 gives the mean and standard deviation over random seeds for each atom trained with more than one seed. Within one run, each per-state energy in Table 9 carries its 10-block stderr over the last 3,000 training steps, and each MAE has a 95% block-bootstrap interval, given for blocks of 100 and of 300 steps in Table 8; the methodology is in Appendix J.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Per-step times and GPU-hours (reported runs, pretraining, preliminary work) are in Appendix K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the Code of Ethics; the work conforms to it. It is a numerical method for the electronic Schrödinger equation, with no human subjects, crowdsourcing, or dual-use artifacts.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: Foundational research on a numerical method for quantum-chemistry electronic structure; no direct path to negative societal applications.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: No pretrained generative models, scraped datasets, or high-misuse-risk artifacts are released. The released code is a variational Monte Carlo library and does not pose dual-use concerns.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix K lists the software we use (DeepQMC, JAX, KFAC-JAX, folx) with version and license; the NIST reference data and the FermiNet reference implementation are cited at point of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code for NestedLoRA-VMC is released at `https://github.com/MinchanJeong/NestedLoRA-VMC`; documentation and license in Appendix K.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human-subjects research; IRB not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing and editing; no core-methodology contribution.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.