PHYSICAL KNOWLEDGE AS SIDE INFORMATION FOR NEURAL FERMIONIC WAVEFUNCTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural wavefunctions provide a flexible approach to fermionic ground-state problems, but learning physically relevant correlations can remain difficult under finite optimization budgets. Existing methods often encode physical knowledge through architecture-specific modifications, making it hard to separate the value of the physical information from the way it is incorporated. We introduce a modular side-information architecture that makes fixed, Hamiltonian-derived physical computations available to a general fermionic neural wavefunction through trainable readers, while preserving its backbone architecture and determinant readout. The readers can act on particle features, pair communication, or orbitals, making the physical function, reader, and injection site explicit design choices. We evaluate this design on a three-dimensional attractive Fermi gas at infinite scattering length and a two-dimensional moiré electron system. In the Fermi gas, combining a two-body scattering-informed orbital correction with opposite-spin pair communication lowers the backbone energy by $1.82\,\hbar^2 k_F^2/m$ at 10,000 updates. With the reader and spectral modes fixed, the scattering relation lowers the energy relative to a matched constant relation. In the moiré system, four cross-attention readers using Hamiltonian-derived relations lower the energy by 0.12 to $0.19\,E_{\mathrm{h}}^*$ at the same budget; density modes also help when read in every block. Across the two systems, the gain depends on the physical functions, their readers, and the injection sites.

1 INTRODUCTION

Neural-network wavefunctions represent a fermionic many-body state using a determinant constructed from learned, permutation-equivariant features of all particle coordinates (Pfau et al., 2020; Hermann et al., 2020; von Glehn et al., 2023). Access to the full many-body configuration allows them to represent correlations among all particles. In practice, however, optimization budgets are finite, while physics already provides useful information about the structures the network needs to capture, including low-energy single-particle modes, the short-distance form of two-body scattering, pairing, and collective density fluctuations. Existing architectures incorporate such knowledge through Jastrow and backflow factors, pair orbitals, or conditioning (Drummond et al., 2004; López Ríos et al., 2006; Luo & Clark, 2019; Kim et al., 2024; Nazaryan & Fu, 2026). These modifications often change three aspects at once: the physical content, the trainable module that uses it, and the site at which it enters the network. When performance improves, it is therefore difficult to isolate the contribution of each aspect.

We ask when physical knowledge can improve a general neural wavefunction, and whether its usefulness depends on how and where it is introduced into the network. To study these questions separately, we treat physical knowledge as *side information* (Figure 1). At each Monte Carlo step, a fixed physical computation is evaluated from the Hamiltonian and the current configuration. A trainable reader then maps this information into the network at a chosen site, such as the particle features, pair messages, or the orbital matrix before the determinant. The reader is initialized to have no effect, so the augmented model starts from the original network and recovers it exactly when the reader is switched off. At fixed Hamiltonian, each physical resource is determined by the particle coordinates, so the side path does not provide additional information about the configuration. Instead, it provides a different representation of structure already present in the input, which may make that structure easier for the network to use. This separation lets us compare physical computations and their reading paths within a common backbone architecture.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We evaluate the framework on two systems with distinct correlation structure: a three-dimensional attractive Fermi gas dominated by short-range pair correlations, and a two-dimensional moiré electron system with more collective correlations. Our contributions are:

- We introduce a modular side-information architecture for neural fermionic wavefunctions, whose trainable readers connect prescribed physical computations to particle features, communication, and orbitals. The construction preserves exact antisymmetry and coordinate derivatives, while retaining the original model as a special case.
- In the Fermi gas at infinite scattering length, combining a scattering-informed orbital reader with opposite-spin pair communication lowers the coordinate-backbone energy by $1.82\,\hbar^2 k_F^2/m$ at a common 10,000-update budget and recovers about 30% of the gap to a specialized pairing ansatz. Matched controls support both gains; simpler pair relations retain the communication improvement.
- In the periodic moiré system, relation attention lowers the energy by up to $0.19\,E_{\mathrm{h}}^*$, or 4.6 combined MCSEs. Cross-attention readers give reductions of 0.12 to $0.19\,E_{\mathrm{h}}^*$ across the four readers in Table 2. At the same budget, reading location also matters: density modes help when read throughout the network, while the moiré one-body spectrum improves over equal-rank free-particle modes only at the orbital input.

Related work. Jastrow and backflow constructions, neural determinant wavefunctions, pair-based ansätze, attention-based periodic wavefunctions, and density-conditioned models all inject physical structure into variational states (Drummond et al., 2004; López Ríos et al., 2006; Luo & Clark, 2019; Bajdich et al., 2006; Pfau et al., 2020; Hermann et al., 2020; von Glehn et al., 2023; Li et al., 2022; Geier et al., 2025; Wilson et al., 2023; Pescia et al., 2024). Neural wavefunctions have been applied to superfluid Fermi gases (Lou et al., 2024; Kim et al., 2024) and to moiré systems (Luo et al., 2024), in each case with physics built into a dedicated architecture. Spatially weighted particle attention adds a trainable distance bias to attention logits (Gaggioli et al., 2026). Grouped-query attention, the internal mixture of experts (MoE), and Hamiltonian FiLM conditioning come from the QERNEL backbone (Nazaryan & Fu, 2026). We hold the backbone architecture fixed and vary only the physical computation, its reader, and its injection site.

2 METHOD

Figure 1 shows how physical readers augment the complete coordinate backbone. The gas model combines pair communication in panel (b) with orbital correction in panel (c). The moiré density model uses panel (d), and the moiré relation-attention readers (Section 2.6) follow the cross-attention of panel (b) without a pair state. These are separate models, each trained from scratch at its target Hamiltonian with backbone and reader parameters optimized jointly.

2.1 VARIATIONAL WAVEFUNCTION AND OPTIMIZATION

For a fixed particle number N, the wavefunction maps the configuration $R = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$ to a complex amplitude. The Hamiltonian specification H specifies the domain, boundary conditions, spin sector, and units. A periodic coordinate map $x(R; H)$ represents each position modulo the cell using sine and cosine pairs. As shown in Figure 1 (a), these inputs are embedded into width-64 particle features and processed by three grouped-query attention/MoE blocks followed by an orbital head. We use the QERNEL backbone (Nazaryan & Fu, 2026), whose internal MoE operates on particle tokens, while the physical readers enter at the locations indicated in the figure. The resulting wavefunction can be written as

$$\Psi_\Theta(R; H) = \mathcal{F}_\Theta\big(x(R; H); f_H(H), f_R(R), f_{RH}(R; H)\big).$$

The subscripts on f indicate whether a resource depends on the Hamiltonian, the configuration, or both. Each final particle feature is mapped to one row of a complex orbital matrix $\Phi(R; H)$, and the fermionic wavefunction is obtained from its determinant,

$$\Psi_\Theta(R; H) = \chi_H(R)\det[\Phi_{ij}(R; H)]_{i,j=1}^N,$$

where the symmetric factor χ_H is one except for the Jastrow comparison in Appendix F. If the preceding operations are permutation equivariant, exchanging two particles exchanges the corresponding

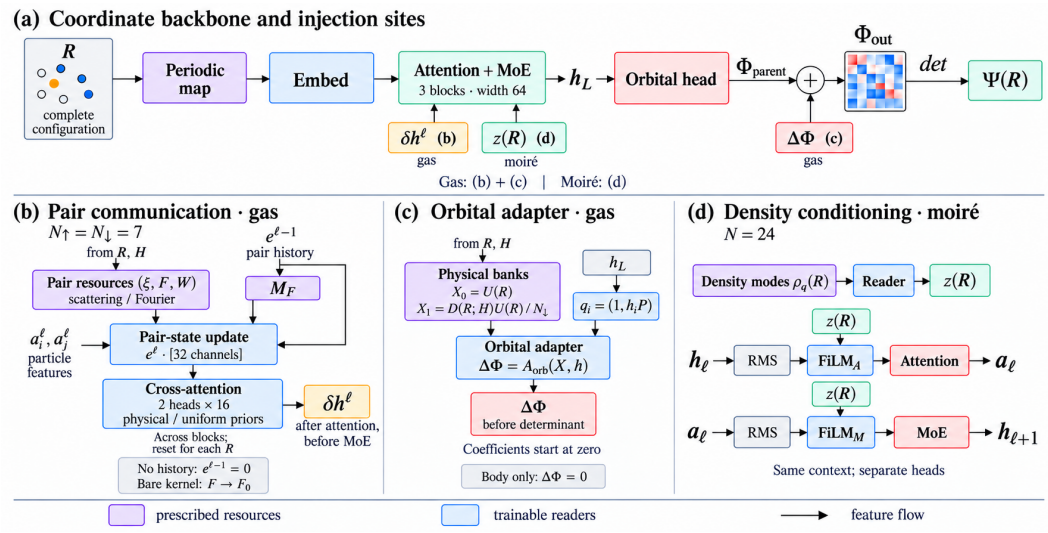


Figure 1: **Physical readers augment the coordinate backbone.** (a) The backbone maps the complete particle coordinates to a determinant wavefunction. The gas model combines (b) pair-state updates and cross-attention with (c) spectral and scattering orbital corrections. Pair relations and matching update the pair state, whose messages enter after backbone attention and before MoE; orbital corrections enter before the determinant. (d) The moiré density model shares one context across separate FiLM heads in every block. Gray boxes show separate gas controls; body-only adds the pair reader to the coordinate baseline. Pair states reset for each configuration, and all resource paths retain coordinate derivatives. Residual additions are omitted; Sections 2.4 and 2.5 give the full updates.

orbital rows and changes the sign of the determinant. The resources and readers are constructed to preserve this equivariance. A general exchange argument is given in Appendix A.

Variational Monte Carlo minimizes $E_H[\Psi] = \langle\Psi|H|\Psi\rangle/\langle\Psi|\Psi\rangle$ using configurations sampled from $|\Psi|^2$. Configuration-dependent resources are recomputed for each sampling proposal and remain part of the local-energy computation. Their defining parameters may be fixed, but their dependence on particle coordinates is retained and therefore contributes to the kinetic term. Further details are given in Appendix D.

2.2 A THREE-AXIS INTERFACE FOR PHYSICAL SIDE INFORMATION

A physical resource exposes a chosen computation through a map fixed with respect to Θ. At fixed H, complete coordinates already determine every configuration-derived resource, but making a physically structured computation explicit can change how the network represents and optimizes the same configuration. The interface has three components: the prescribed computation $\mathcal{P}$, trainable reader $\mathcal{R}$, and injection operator $\mathcal{U}$:

$$z_p(R) = \mathcal{R}_p[f_{\mathcal{P}}(R;H), h_p], \qquad h'_p = \mathcal{U}_p(h_p, z_p),$$

where h_p denotes the network state at site p, including particle or pair features as needed. Orbital and pair readers use this state; the density reader depends only on its resource. We call these resource-reading branches physical side experts, distinct from the backbone's internal MoE experts.

The site p records where a resource first changes the variational computation. We use five sites: p_{in} for the coordinate embedding, p_{body} for a particle token inside a backbone block, p_{pair} for a pair state or pair-to-particle message (written $p_{\text{pair}} \to p_{\text{body}}$ when its messages enter particle features), p_{orb} for the orbital matrix before the determinant, and p_{amp} for a symmetric log-amplitude factor. A later orbital readout can depend on modified body features, so the site label identifies the first intervention rather than the last place where its effect is visible.

A reader can combine resources, and one resource can serve several sites. Changing the resource, reader, or injection site defines a separate comparison; Table 4 records the combinations evaluated

here. Resource definitions remain fixed during training, while their values and coordinate derivatives are recomputed for each configuration. Each reader is initialized to reproduce its parent computation, so the augmented family contains the baseline (Proposition 2).

2.3 Interaction-aware orbital structure in the attractive Fermi gas

For the balanced $N = 14$ gas, the orbital adapter in Figure 1c combines low-energy spatial modes with a short-range scattering relation between opposite spins. The modes supply orbital variation, and the relation makes their use depend on nearby opposite-spin particles. The backbone retains the full many-body configuration.

The spin-resolved matrix $U(R)$ contains low-energy eigenfunctions of the periodic free-particle kinetic operator. For the Pöschl–Teller interaction, the zero-energy radial solution is proportional to $\tanh(\mu r)$ (Kim et al., 2024). Its logarithmic scattering profile is shifted and tapered to define $D(R; H)$, whose value and first two derivatives vanish at a prescribed healing distance. This confines the supplied relation to short separations while preserving the derivatives needed for kinetic energy; Appendix F specifies its periodic continuation.

We expose this prior through two orbital banks,

$$X_0 = U(R), \qquad X_1 = \frac{D(R; H)U(R)}{N_\downarrow}, \tag{1}$$

where X_1 makes the low-energy spatial modes sensitive to opposite-spin neighborhoods. The orbital reader uses $q_i = (1, h_i P)$ to form the correction $\Delta\Phi = \mathcal{A}_{\mathrm{orb}}(X, h)$:

$$\begin{aligned} \Delta\Phi_{i\alpha}(R) &= \sum_{s=0}^{1} \sum_{a,c} X_{s,ia}(R) q_{ic} C_{sca\alpha}, \\ \Phi_{\mathrm{out}}(R) &= \Phi_{\mathrm{parent}}\big(h(R)\big) + \Delta\Phi(R). \end{aligned} \tag{2}$$

Here h_i is the final particle feature, P is a fixed compression, a indexes spatial modes, and c indexes the components of q_i, including its constant component. Appendix F gives their dimensions. The coefficients C start at zero, recovering the parent computation. The correction enters the complex orbital matrix before the determinant. Replacing D by a constant opposite-spin mask holds the spectral bank, reader dimensions, and optimization protocol fixed while testing the scattering relation.

2.4 Pair-organized communication in the attractive gas

The pair reader in Figure 1b uses physical relations to organize communication inside the backbone. Its prescribed resources are local pair channels ξ, a matching matrix F, and a nonnegative attention envelope W, all evaluated from the current configuration and Hamiltonian.

Let F_{ij} combine closed-shell overlap and short-range scattering for opposite-spin particles. The matching action contrasts direct and exchanged assignments:

$$(\mathcal{M}_F e)_{ij} = \frac{\sum_{kl}(F_{ij}F_{kl} - F_{il}F_{kj})e_{kl}}{2(1 + \|F\|_{\mathrm{F}}^2)}. \tag{3}$$

Each coefficient is a two-by-two minor of F and vanishes when the pairs share a particle. The exchange relation organizes pair messages; the final determinant enforces fermionic antisymmetry.

In the gas, a_i^ℓ denotes particle i's input to attention in block ℓ, after normalization and any parent input modulation. Each opposite-spin pair carries a 32-channel state, initialized as $e_{ij}^0 = 0$ for every configuration and updated across blocks:

$$\begin{aligned} e_{ij}^\ell = \mathrm{sat}\Big[&\tfrac{1}{2} e_{ij}^{\ell-1} + L_\xi \xi_{ij} + L_s \tfrac{a_i^\ell + a_j^\ell}{2} \\ &+ L_d \tfrac{(a_i^\ell - a_j^\ell)^{\odot 2}}{2} + L_r \tfrac{\bar{e}_{i\cdot}^{\ell-1} + \bar{e}_{\cdot j}^{\ell-1}}{2} + L_m (\mathcal{M}_F e^{\ell-1})_{ij} \Big]. \end{aligned} \tag{4}$$

Here $\mathrm{sat}(t) = t/\sqrt{1 + t^2}$ acts elementwise, $\odot 2$ denotes elementwise squaring, and the barred terms average the other pairs sharing an endpoint. Two cross-attention heads, each of width 16, read the

pair state:

$$\alpha_{ij}^{\ell b} = \mathrm{softmax}_j\left[\frac{(Q_\ell a_i^\ell)_b^\top (K_\ell e_{ij}^\ell)_b}{\sqrt{16}} + \log p_{ij}^b\right],$$
$$\delta h_i^\ell = O_\ell\, \mathrm{concat}_{b=1,2}\left[\sum_j \alpha_{ij}^{\ell b}(V_\ell[e_{ij}^\ell, a_j^\ell])_b\right]. \tag{5}$$

The first head uses W with a uniform floor as its prior p^1; the second uses a uniform prior. Their message enters the attention residual before the MoE:

$$\bar{h}_\ell = h_\ell + \mathrm{Attn}_\ell(a^\ell) + \delta h^\ell,$$
$$h_{\ell+1} = \bar{h}_\ell + \mathrm{MoE}_\ell\big[\mathcal{T}_{\ell,M}(\bar{h}_\ell)\big], \tag{6}$$

Here a^ℓ and δh^ℓ collect particle rows, and $\bar{h}_\ell$ is the state after the attention residual. The map $\mathcal{T}_{\ell,M}$ applies the parent's normalization and input modulation before MoE, and the output maps O_ℓ start at zero. The final features enter q_i in the orbital adapter. Body-only attaches the same pair reader to the coordinate baseline, retaining its orbital head and setting $\Delta\Phi = 0$. The other controls retain the orbital-side parent. No-history resets the incoming pair state at each block. Bare-kernel replaces F by the closed-shell F_0, retaining ξ and W. Direct matching keeps only the direct assignment in Eq. (3). Fourier-pair replaces all three resources by calibrated periodic functions without scattering content. Pair state is never carried between Monte Carlo configurations. Appendix F.4 specifies these controls and the parent configurations.

2.5 Collective density context in the periodic moiré system

For the fully polarized $N = 24$ moiré system, Figure 1d supplies collective density information to particle features. We compare using this global, permutation-invariant context at the embedding with reusing it throughout the backbone.

We provide normalized complex Fourier components at eight fixed long-wavelength vectors,

$$\rho_{\boldsymbol{q}}(R) = N^{-1/2}\sum_j e^{i\boldsymbol{q}\cdot\boldsymbol{r}_j}, \qquad f^\rho = \mathcal{N}_\rho\big((\rho_{\boldsymbol{q}})_{\boldsymbol{q}\in\mathcal{Q}}\big). \tag{7}$$

One $16 \to 32 \to 8$ reader maps f^ρ to a single context $z(R)$, shared across particles and blocks. Each block has separate FiLM heads before attention and before MoE:

$$\mathcal{U}_p(h_p, z_p) = \gamma_p(z_p) \odot h_p + \beta_p(z_p). \tag{8}$$

Each head broadcasts its scale and shift across particles, preserving equivariance while acting on position-dependent tokens. With layer-specific row-wise RMS normalization, the residual updates are

$$a_\ell = h_\ell + \mathrm{Attn}_\ell[\mathcal{U}_{\ell,A}(\mathrm{RMSNorm}_{\ell,A}(h_\ell), z(R))],$$
$$h_{\ell+1} = a_\ell + \mathrm{MoE}_\ell[\mathcal{U}_{\ell,M}(\mathrm{RMSNorm}_{\ell,M}(a_\ell), z(R))].$$

The front control applies the same six FiLM heads consecutively after embedding. Appendix B gives this placement and the internal MoE, whose router and four experts receive the same conditioned particle token. The one-body spectral comparison uses a separate reader defined there.

2.6 Relation attention in the periodic moiré system

To transfer the communication idea of Section 2.4, we keep its form but rebuild its relations from the moiré problem. The gas pair state has no counterpart here, because the fully polarized system has no opposite-spin channel. In every block, particle i attends over a set $\mathcal{T}_i$ of relation tokens τ_{it} and a null token with zero value,

$$\alpha_{it}^\ell = \mathrm{softmax}_{t\in\mathcal{T}_i\cup\{\varnothing\}}\left[\frac{(Q_\ell a_i^\ell)^\top K_\ell \tau_{it}}{\sqrt{d}}\right], \qquad a_i^\ell \leftarrow a_i^\ell + O_\ell \sum_{t\in\mathcal{T}_i} \alpha_{it}^\ell V_\ell \tau_{it}, \tag{9}$$

where a_i^ℓ is the normalized input to backbone attention and O_ℓ starts at zero. Two relation families define $\mathcal{T}_i$. *Pair relations* to the other electrons come from the one-body Bloch reference at $\lambda = 1$ with

smoothed occupation of 24 electrons: the normalized exchange overlap $|\gamma_{ij}|^2/(n_i n_j)$ of the occupied kernel γ, a ridge-conditioned overlap and orbital availability derived from the kernel over all electrons, and the reference density n_i, together with the partner feature a_j. Their control keeps the reader and replaces these relations with a lattice-geometry similarity, $\exp\{-\frac{1}{3}\sum_g[1-\cos \boldsymbol{g}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j)]\}$ over the primitive reciprocal star, with unit availability and density. *Cell relations* give each electron one token per moiré cell, carrying a soft cell assignment and relative phases, plus six tokens with the density relations $\sum_{j\neq i} e^{i\boldsymbol{q}\cdot(\boldsymbol{r}_i-\boldsymbol{r}_j)}$ at the shortest simulation-cell wavevectors. The occupation variant adds to each cell token the fluctuation of the cell occupation by the other electrons; the density variant sets that channel to zero. Both contrasts keep the reader and its initialization fixed (Appendix B.2).

3 Experiments

Setup. The gas has $N = 14$ fermions ($7\uparrow +7\downarrow$) in a cubic periodic box with a modified Pöschl–Teller attraction at infinite vacuum scattering length and $k_F r_e = 0.2$ (Appendix F). The moiré system has $N = 24$ fully spin-polarized electrons in a triangular moiré potential with Ewald-summed Coulomb interactions, at potential strength $\lambda = 1$ (Appendix C). Both use the width-64 backbone above. Every model is trained once, from the same untrained initialization and without pretraining, for 10,000 updates (stochastic reconfiguration for the gas, Adam for the moiré system); selected gas models continue to 20,000. We evaluate energies at frozen parameters with separate Monte Carlo chains. Error bars are conditional Monte Carlo standard errors (MCSEs) of each frozen state. For the gas, most checkpoints have a standard readout (40 draws per walker) and an independently initialized longer readout (256 draws), and differences in the main text are taken within one protocol (Appendix F). For scale, Table 1 also lists a separately designed Pfaffian–Jastrow–backflow ansatz (Kim et al., 2024), trained under the same protocol.

3.1 Attractive Fermi gas

Figure 2 compares the orbital and communication readers, Table 1 gives the gas energies, and Figure 3 shows that the two main gains hold over the last 4,000 updates.

Interaction-aware orbitals. Adding low-energy plane waves to the orbital reader lowers the energy by 0.474 ± 0.045 relative to the same reader without spectral modes (standard readout; Appendix F). Letting the correction depend on nearby opposite-spin particles helps further: replacing a constant opposite-spin relation by the healed scattering shape lowers the energy by 0.254 ± 0.043 (standard readout), mostly through the kinetic term (Table 19). Both models share the spectral modes and the reader, so this difference isolates the scattering shape.

Pair-organized communication. The pair reader lowers the energy of the orbital model by a further 0.549 ± 0.020 (independent readout) and keeps an advantage of 0.175 ± 0.019 at 20,000 updates (Table 12). Almost all of the reduction at 10,000 updates is kinetic, -0.542 ± 0.014 (Table 11). Attached directly to the coordinate backbone, the same reader lowers its energy by 0.816 ± 0.024 (independent readout), so the communication path also helps without the orbital reader. The orbital and communication gains overlap in part, and their factorial interaction is $+0.27 \pm 0.03$ on the independent readout.

Alternative realizations of the pair prior. Controls that keep the reader and site but change the pair content give energies close to those of the full reader in the independent readout: Fourier-pair functions without scattering content (-0.003 ± 0.017), a matching kernel without the scattering factor ($+0.003 \pm 0.017$), and direct matching without the exchanged assignment (-0.007 ± 0.017; Tables 15 and 16). Resetting the pair state at every block, which also disables the matching action and residual pair paths, is slightly lower still (-0.046 ± 0.016). In these runs, simpler relations retain the communication gain despite changes to the scattering template, exchange sign, and pair history. The three content controls keep the opposite-spin pair organization; the Fourier functions, for example, replace the scattering realization of the pair prior by a less specific one.

Reading the orbital banks inside the backbone. The orbital banks of Eq. (1), which help at the orbital site, raise the energy by 0.85 (independent readout) when a bilinear reader applies them in

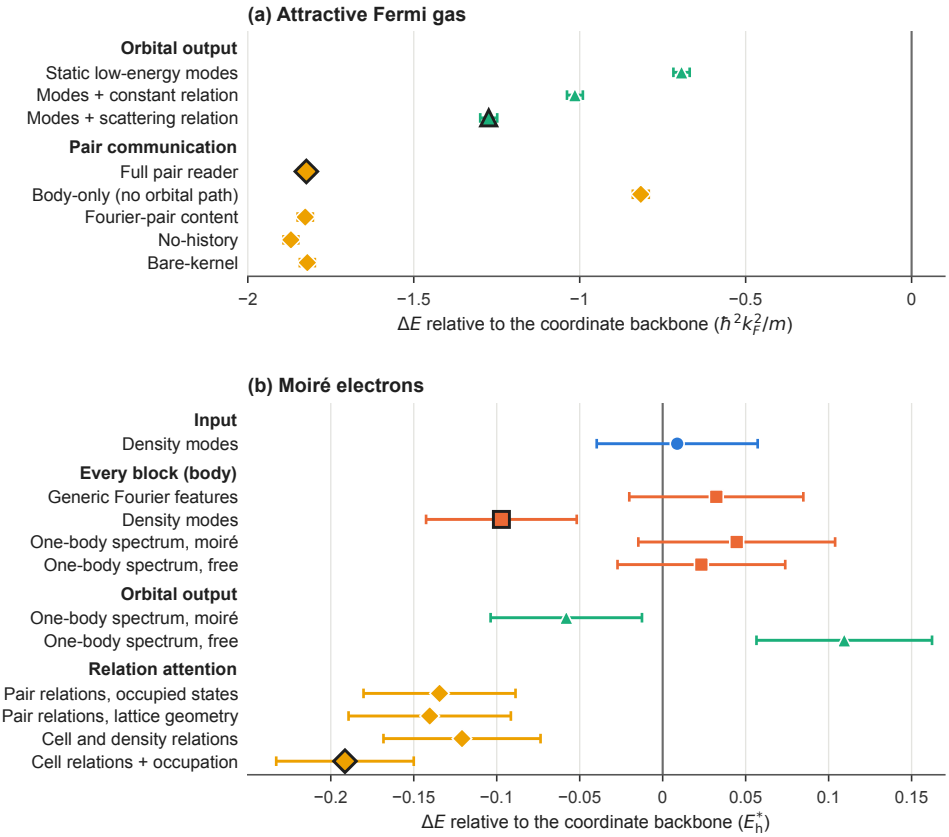


Figure 2: **Endpoint comparisons, grouped by first injection site.** Energy change at 10,000 updates relative to the coordinate backbone of each system; negative values are lower energies. Color and marker shape mark the injection site, and outlined markers identify the main gas gains and the highlighted moiré readers. Error bars combine the MCSEs of the model and its baseline in quadrature. (a) Gas, independent 256-draw readout. The scattering-relation point is the separately trained run marked † in Table 1, and the Pfaffian–Jastrow–backflow reference lies at -5.87 and is omitted. (b) Moiré system, readouts of Table 2, including the relation-attention readers.

every backbone block. Norm-preserving feature rotations driven by the same banks give no gain (Table 15). Adding further physical inputs to the orbital model, through a low-energy subspace reader, a short-range Jastrow factor, or an exchange-hole attention bias, changes its energy by less than 0.05 in each case (Table 11).

3.2 Periodic moiré system

Where to read collective information. Reading the eight density modes of Eq. (7) inside every block reaches $-7.4456 \pm 0.0290\, E_{\mathrm{h}}^{*}$ at 10,000 updates, $0.097 \pm 0.046\, E_{\mathrm{h}}^{*}$ below the coordinate backbone and $0.106 \pm 0.047\, E_{\mathrm{h}}^{*}$ below front reading of the same modes (paired errors, Table 2). Generic Fourier features of the same dimension, read the same way, stay above the backbone. Repeated use of this collective information lowers the endpoint energy relative to supplying it once at the input.

Transferring the communication idea. The relation-attention readers give the largest mean gains in Table 2. Relative to the coordinate backbone, the four readers of Section 2.6 lower the energy by 0.134 ± 0.046 (occupied-state pair relations), 0.140 ± 0.049 (lattice-geometry pair relations), 0.121 ± 0.047 (cell and density relations), and $0.191 \pm 0.041\, E_{\mathrm{h}}^{*}$ (cell relations with occupation), which is 2.6 to 4.6 combined MCSEs. The two pair-relation sets are different realizations of the same prior, both built from the moiré problem, and give comparable energies: occupied-state relations of

Table 1: Gas energies at 10,000 updates in $\hbar^2 k_F^2/m$, with MCSEs from one training run per model. The two columns read the same trained parameters with different frozen-state chains. The exception is the two scattering-relation rows, which are separate training runs of the same model and hyperparameters. The second (†) is the matched reference for the pair rows other than body-only. A dash marks a readout not reported.

Model	Physical content and first site	Standard, 40 draws	Independent, 256 draws
Coordinate backbone	None	10.5997 ± 0.0282	10.5625 ± 0.0192
Static orbital reader	Low-energy modes; p_{orb}	9.8270 ± 0.0338	9.8693 ± 0.0152
Orbital, constant relation	Modes and constant opposite-spin relation; p_{orb}	9.5716 ± 0.0278	9.5481 ± 0.0141
Orbital, scattering relation	Modes and healed scattering relation; p_{orb}	9.3174 ± 0.0328	—
Orbital, scattering relation†	Same configuration, separate run	9.2748 ± 0.0492	9.2884 ± 0.0168
Pair communication	Pair state and cross-attention on the orbital model; $p_{\mathrm{pair}} \to p_{\mathrm{body}}$	8.7166 ± 0.0270	8.7389 ± 0.0117
Body-only	Same pair reader on the coordinate backbone	9.7273 ± 0.0272	9.7466 ± 0.0143
Fourier-pair	Periodic pair functions without scattering content	—	8.7356 ± 0.0125
Bare-kernel	Scattering factor removed from F	8.7427 ± 0.0276	8.7419 ± 0.0118
No-history	Incoming pair state reset at every block	8.6566 ± 0.0338	8.6924 ± 0.0112
Pairing reference	Separate Pfaffian–Jastrow–backflow ansatz	4.6780 ± 0.0299	4.6921 ± 0.0118

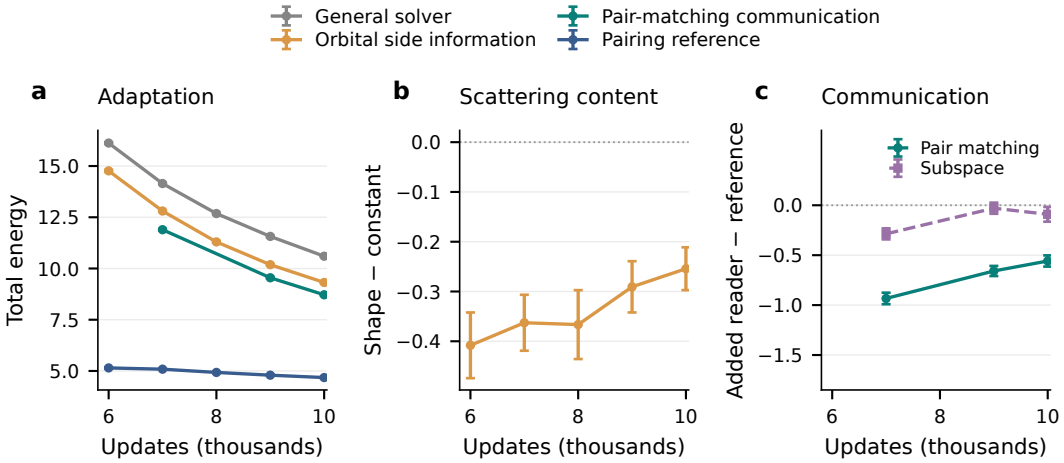


Figure 3: **Gas energies over the last 4,000 updates.** (a) Total energies from 6,000 to 10,000 updates. The pair-communication model keeps the spectral and scattering orbital path. The orbital curve is the original run; the matched orbital reference of the pair model is used in (c). (b) Scattering shape minus constant relation with the same orbital reader. (c) Communication readers relative to their matched orbital reference. Points are standard 40-draw frozen-checkpoint readouts; differences combine MCSEs in quadrature. Lines connect saved checkpoints without smoothing. Energies in $\hbar^2 k_F^2/m$.

the moiré one-body reference yield energies $0.006 \pm 0.045\, E_{\mathrm{h}}^*$ above lattice-geometry relations read by the same reader. Adding signed first-order response channels changes the energy by an unresolved $-0.045 \pm 0.040\, E_{\mathrm{h}}^*$ relative to the occupied-state reader. With the response signs removed, the reader shows no resolved gain over the coordinate backbone ($+0.035 \pm 0.075\, E_{\mathrm{h}}^*$; Appendix B.2). Adding the cell-occupation fluctuations lowers the energy by a further $0.071 \pm 0.039\, E_{\mathrm{h}}^*$, a trend of 1.8 MCSEs in one training run. It also moves the reduction toward the interaction energy, which falls by $0.163 \pm 0.027\, E_{\mathrm{h}}^*$ relative to the density variant while the external-potential term changes by an unresolved $+0.10 \pm 0.13\, E_{\mathrm{h}}^*$ (Table 6).

Transferring the orbital idea. We also tested the orbital approach from the gas using a low-energy one-body basis derived from the moiré Hamiltonian. At 10,000 updates, a residual reader built from the 25 lowest eigenmodes of the one-body moiré Hamiltonian yields an energy $0.167 \pm 0.049\, E_{\mathrm{h}}^*$ below the same reader built from free-particle modes of equal rank. The latter provide the low-energy basis used in the gas but omit the moiré potential. This matches the gas, where the scattering relation

Table 2: Moiré energies at 10,000 updates in effective Hartree, with MCSEs from one training run per model. The two spectral resources use the same rank and reader; each pair of relation-attention rows shares one reader.

Physical resource	First reading site	Energy
None (coordinate backbone)	—	-7.3484 ± 0.0349
Pair relations, occupied states	Relation attention, $p_{\mathrm{pair}} \to p_{\mathrm{body}}$	-7.4829 ± 0.0296
Pair relations, lattice geometry	Relation attention, $p_{\mathrm{pair}} \to p_{\mathrm{body}}$	-7.4888 ± 0.0342
Cell and density relations	Relation attention, p_{body}	-7.4693 ± 0.0319
Cell relations with occupation	Relation attention, p_{body}	$\mathbf{-7.5399 \pm 0.0223}$
Density modes	Front, p_{in}	-7.3397 ± 0.0336
Density modes	Every block, p_{body}	-7.4456 ± 0.0290
Generic Fourier features	Every block, p_{body}	-7.3161 ± 0.0391
One-body spectrum, moiré potential	Every block, p_{body}	-7.3038 ± 0.0479
One-body spectrum, free particle	Every block, p_{body}	-7.3251 ± 0.0365
One-body spectrum, moiré potential	Orbital input	-7.4065 ± 0.0293
One-body spectrum, free particle	Orbital input	-7.2390 ± 0.0397

lowered the energy relative to a constant relation at the same site. Relative to the backbone, the moiré modes give an unresolved $-0.058 \pm 0.046\, E_{\mathrm{h}}^{*}$ and the free modes end above it. When the spectra are read inside every block, their energy difference is unresolved ($+0.021 \pm 0.060\, E_{\mathrm{h}}^{*}$; Table 2, errors in quadrature). The physical-versus-free orbital contrast is unresolved at 20,000 updates (Appendix E.1).

4 Discussion and limitations

Treating physical knowledge as side information allows controlled comparisons of its content, reader, and injection site. At 10,000 updates, Hamiltonian-specific orbital functions gave lower energies than their generic counterparts in both systems. Several simpler relation sets retained the communication gains.

The method of injecting a physical prior through a side path transferred between the two systems, while the trained modules and descriptors were built anew for each. A prior in this sense is a kind of physical knowledge, such as a low-energy one-body basis at the orbitals or relations that organize communication. It plays the same role in each system, but its concrete form has to be derived again from each Hamiltonian. The gas pair state has no counterpart in the fully polarized moiré system and had to be rebuilt as relations of the moiré problem. The free-particle basis that serves the gas ended above the moiré backbone, whereas the basis derived from the moiré Hamiltonian had a lower mean energy. How the network reads a function also mattered: the same orbital functions raised the energy when read inside the backbone, and density modes helped the moiré model only when read in every block. For physics-informed neural wavefunctions, this suggests choosing the site and the reader together with the physical content, and testing each choice against matched controls. Readers with zero-initialized output maps preserve the baseline at initialization, allowing the same comparison design to be applied to new systems.

Limitations. Each model was trained once, and the error bars are conditional Monte Carlo errors that exclude training-seed variability. All models start from scratch and are compared at equal numbers of updates. Their energies remain far from converged ground-state values, so the results concern optimization within a budget rather than final accuracy, and they do not account for the added cost per update. Our communication controls exchange one realization of a prior for another; relations unrelated to either problem were not tested. In the moiré system, generic features also gain from in-block reading, so the placement benefit specific to density is unresolved ($-0.028 \pm 0.064\, E_{\mathrm{h}}^{*}$), and in an earlier, wider (width-96) model at 20,000 updates the density reader no longer helps ($+0.047 \pm 0.041\, E_{\mathrm{h}}^{*}$; Appendix E). We study one Hamiltonian point per system, do not test weight transfer across Hamiltonians, and do not claim a physical mechanism, such as pair condensation, behind the gains.

AI USE STATEMENT

The authors formulated the research questions and determined the experimental design. Generative AI tools, including ChatGPT, Codex, and Claude, assisted with method implementation and debugging, checks of mathematical formulations against code, translation, and the analysis and interpretation of results. They also provided feedback during methodological discussions and supported literature search, manuscript drafting and editing, and figure preparation. AI-assisted code and quantitative claims were checked against implementation tests and recorded experimental outputs, and references were checked against primary sources. The reported numerical results come from the computational experiments described in this paper. The authors evaluated the AI-assisted suggestions, made the scientific decisions, and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Appendices B–D specify the physical resources, periodic-electron Hamiltonian, optimization settings, coordinate derivatives, and uncertainty estimation. Appendix F gives the Fermi-gas Hamiltonian, reader equations, initialization, and sampling and optimization protocols. The supplementary source-data archive contains the numerical tables and plotting scripts underlying Figures 3–7, with file checksums and instructions for regenerating the plots. These materials distinguish training trajectories from frozen-state readouts and report conditional Monte Carlo errors.

REFERENCES

M. Bajdich, L. Mitas, G. Drobný, L. K. Wagner, and K. E. Schmidt. Pfaffian pairing wave functions in electronic-structure quantum Monte Carlo simulations. *Physical Review Letters*, 96(13):130201, 2006. doi: 10.1103/PhysRevLett.96.130201.

N. D. Drummond, M. D. Towler, and R. J. Needs. Jastrow correlation factor for atoms, molecules, and solids. *Physical Review B*, 70:235119, 2004. doi: 10.1103/physrevb.70.235119. URL https://doi.org/10.1103/physrevb.70.235119.

Filippo Gaggioli, Sam Azadi, and Liang Fu. Accurate self-attention wavefunctions at large scale, 2026. URL https://arxiv.org/abs/2607.08616. arXiv preprint arXiv:2607.08616.

Stefano Gandolfi. Quantum monte carlo study of strongly interacting fermi gases, 2013. URL https://arxiv.org/abs/1312.1612.

Max Geier, Khachatur Nazaryan, Timothy Zaklama, and Liang Fu. Self-attention neural network for solving correlated electron problems in solids. *Physical Review B*, 112(4):045119, July 2025. ISSN 2469-9950, 2469-9969. doi: 10.1103/qxc3-bkc7. URL https://doi.org/10.1103/qxc3-bkc7.

Jan Hermann, Zeno Schätzle, and Frank Noé. Deep-neural-network solution of the electronic Schrödinger equation. *Nature Chemistry*, 12:891–897, 2020. doi: 10.1038/s41557-020-0544-y.

Jane Kim, Gabriel Pescia, Bryce Fore, Jannes Nys, Giuseppe Carleo, Stefano Gandolfi, Morten Hjorth-Jensen, and Alessandro Lovato. Neural-network quantum states for ultra-cold fermi gases. *Communications Physics*, 7:148, 2024. doi: 10.1038/s42005-024-01613-w. URL https://www.nature.com/articles/s42005-024-01613-w.

Xiang Li, Zhe Li, and Ji Chen. Ab initio calculation of real solids via neural network ansatz. *Nature Communications*, 13:7895, 2022. doi: 10.1038/s41467-022-35627-1.

P. López Ríos, A. Ma, N. D. Drummond, M. D. Towler, and R. J. Needs. Inhomogeneous backflow transformations in quantum Monte Carlo calculations. *Physical Review E*, 74:066701, 2006. doi: 10.1103/PhysRevE.74.066701. URL https://doi.org/10.1103/PhysRevE.74.066701.

Wan Tong Lou, Halvard Sutterud, Gino Cassella, W. M. C. Foulkes, Johannes Knolle, David Pfau, and James S. Spencer. Neural wave functions for superfluids. *Physical Review X*, 14:021030, 2024. doi: 10.1103/PhysRevX.14.021030.

Di Luo and Bryan K. Clark. Backflow transformations via neural networks for quantum many-body wave functions. *Physical Review Letters*, 122(22):226401, 2019. doi: 10.1103/PhysRevLett.122.226401.

Di Luo, David D. Dai, and Liang Fu. Simulating moiré quantum matter with neural network, 2024. URL https://arxiv.org/abs/2406.17645. arXiv preprint arXiv:2406.17645.

Khachatur Nazaryan and Liang Fu. QERNEL: A scalable large electron model, 2026. URL https://arxiv.org/abs/2604.26018. arXiv preprint arXiv:2604.26018.

Gabriel Pescia, Jannes Nys, Jane Kim, Alessandro Lovato, and Giuseppe Carleo. Message-passing neural quantum states for the homogeneous electron gas. *Physical Review B*, 110(3):035108, 2024. doi: 10.1103/PhysRevB.110.035108.

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. *Physical Review Research*, 2(3):033429, 2020. doi: 10.1103/PhysRevResearch.2.033429.

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 (ICLR)*, 2023. doi: 10.48550/arXiv.2211.13672. URL https://openreview.net/forum?id=xveTeHVlF7j.

Max Wilson, Saverio Moroni, Markus Holzmann, Nicholas Gao, Filip Wudarski, Tejs Vegge, and Arghya Bhowmik. Neural network ansatz for periodic wave functions and the homogeneous electron gas. *Physical Review B*, 107(23):235139, 2023. doi: 10.1103/PhysRevB.107.235139.

A RESOURCE DEFINITIONS AND STRUCTURAL PROPERTIES

A.1 RESOURCE DEPENDENCE AND PAIR DESCRIPTORS

Table 3 groups resources by their functional dependence.

Table 3: Resource classes defined by functional dependence. The examples indicate the scope of the framework, including resources that have not been evaluated. Geometry-dependent mode choices belong to the joint class when the geometry varies with H.

Class	Dependence	Examples
Hamiltonian	$f_H(H)$	Couplings, boundary twists, prescribed sectors or problem-level index labels.
Configuration	$f_R(R)$	Density modes and pair geometry in a fixed coordinate convention.
Joint	$f_{RH}(R;H)$	Potential values at particles, interaction-dependent pair or scattering profiles.

Pair information can also be summarized at each particle through equivariant maps such as

$$f_{i,\alpha}^{\mathrm{pair}} = \frac{1}{N-1}\sum_{j\neq i} f_\alpha(\boldsymbol{r}_i - \boldsymbol{r}_j; H), \tag{10}$$

where prescribed f_α may describe smooth geometry or an interaction-dependent pair solution. An interaction-dependent solution belongs to the joint resource class. Equation (10) defines an admissible interface whose usefulness must be evaluated for each chosen pair profile.

A.2 EXCHANGE SYMMETRY

Proposition 1 (Exchange symmetry)**.** *Suppose the complete computation of Φ is permutation equivariant, including the resources, readers, injection maps, and backbone layers. For a permutation matrix Π and symmetric χ_H, $\Psi_\Theta(\Pi R; H) = \det(\Pi)\Psi_\Theta(R; H)$.*

Proof. Let Π permute electron rows, with the adapter satisfying $x(\Pi R; H) = \Pi x(R; H)$ by assumption. Global resources are invariant under this permutation, particle-resolved resources transform by Π, and pair-resolved resources transform by $f \mapsto \Pi f \Pi^\top$. Equivariant readers and pooling preserve the corresponding row transformation through each injection map. Shared row-wise maps, normalization, attention, residual addition, and the inherited MoE also preserve equivariance. Attention scores transform as $S \mapsto \Pi S \Pi^\top$, and row-wise softmax commutes with the permutation. Induction through the network gives $\Phi(\Pi R; H) = \Pi\Phi(R; H)$. Multiplication by the symmetric envelope and the determinant identity yield $\Psi(\Pi R; H) = \det(\Pi)\Psi(R; H)$. □

A.3 NEUTRAL RECOVERY

Proposition 2 (Neutral recovery)**.** *Let $\mathcal{V}_0$ and $\mathcal{V}_{\mathrm{side}}$ denote the baseline and augmented variational families for a fixed backbone and readout. Suppose every added resource path admits a parameter setting that reproduces the original computation pointwise. Then $\mathcal{V}_0 \subseteq \mathcal{V}_{\mathrm{side}}$ and, over admissible nonzero states, $\inf_{\Psi\in\mathcal{V}_{\mathrm{side}}} E_H[\Psi] \leq \inf_{\Psi\in\mathcal{V}_0} E_H[\Psi]$.*

Proof. Copy any baseline parameter setting into the augmented backbone and readout. Set each added resource path to its stipulated neutral parameter value. The two computations then agree pointwise and define the same wavefunction. Restriction to admissible nonzero states preserves the inclusion of the baseline family. Taking the energy infimum over the larger family gives the stated inequality. □

For the FiLM parameterization in Eq. (8), zeroing the scale and shift output heads, including their biases, gives $(\gamma, \beta) = (\mathbf{1}, \mathbf{0})$. The internal MoE then receives the unmodulated particle stream. Zero orbital coefficients C remove the additive orbital correction, and zero output maps O_ℓ remove the pair messages; these settings recover the respective parent computations. Setting a descriptor to zero need not do so. A design that supplies resources directly to a gate must also neutralize that path. Family inclusion leaves finite-budget optimization performance as an empirical question.

Table 4: Interface atlas for the completed comparisons reported in this manuscript. Each row specifies the physical prior, the trainable reader or operator, and the first injection site. The last column points to the quantitative comparison; rows marked as controls retain the other routes stated in that comparison. Exploratory catalog entries that were not completed or did not define a conclusion-changing comparison are not treated as evidence.

Physical prior $\mathcal{P}$	Reader / operator $\mathcal{R}, \mathcal{U}$	Site	Completed comparison
None (complete coordinates only)	Parent backbone and determinant	$p_{\mathrm{in}} \to p_{\mathrm{out}}$	Backbone references in Tables 16, 2, and 19.
Lowest-shell spectral modes U	Zero-initialized complex linear orbital readout UA	p_{orb}	Static spectral reader versus constant orbital reference; Appendix F.
Spectral modes and healed scattering bank (U, DU)	Dynamic linear map $\mathcal{A}_{\mathrm{orb}}$, Eq. (2)	p_{orb}	healed-scattering orbital route; scattering-shape versus constant-relation control in Table 19.
Vacuum short-range scattering profile D^{vac}	Relation-weighted attention logits and values	p_{body}	Physical versus constant body relation; Appendix F.
Compact pair channels ξ and antisymmetrized minors of F	Pair state and cross-attention, Eqs. (4) and (5)	$p_{\mathrm{pair}} \to p_{\mathrm{body}}$	Combined orbital-and-pair route / pair matching versus its orbital-side reference in Tables 11 and 16.
Calibrated periodic Fourier-pair functions	Same pair-minor and cross-attention maps as the combined route	$p_{\mathrm{pair}} \to p_{\mathrm{body}}$	Fourier-pair description versus the combined route; Table 15.
Same pair resources on coordinate baseline	Same pair reader; baseline orbital head with $\Delta\Phi = 0$	$p_{\mathrm{pair}} \to p_{\mathrm{body}}$	Body-only control; Table 16.
Same pair resources, pair history suppressed	Pair cross-attention reader with $e_{ij}^{\ell-1} = 0$ at each block	$p_{\mathrm{pair}} \to p_{\mathrm{body}}$	No-history control; Table 16.
Closed-shell kernel only G	Bare matching kernel $F_0 = G/s_0$ with the same reader	p_{pair}	Bare-kernel control; Table 16.
Closed-shell modes V and charge/spin projections	Ridge subspace fit with recurrent coefficients Z_t^{ℓ}	p_{body}	Low-energy subspace with matching and reference-subtracted matching; Tables 14 and 15.
Same-spin exchange-hole profile or opposite-spin D^{c}	Scalar attention-logit correction or symmetric Jastrow factor $\exp(J)$	p_{body} or p_{amp}	Exchange-hole and Jastrow comparisons; Tables 11 and 15.
Orbital banks (U, DU)	Bilinear body readout; norm-preserving feature rotation; their combination	p_{body}	Same banks read in alternative ways; Tables 14 and 15.
Multiple pair, orbital, and amplitude channels	Multi-site, restricted-orbital, gated particle-message, or body-and-amplitude readers	p_{body}, p_{orb}, or p_{amp}	Alternative complete readers; Tables 17 and 18.
Simulation-cell density modes $\rho_{\boldsymbol{q}}$	Shared MLP followed by FiLM heads	p_{in} or p_{body}	Front versus within-body placement and generic Fourier controls; Tables 2 and 9.
Simulation-cell pair harmonics $f_{\boldsymbol{q}}^{\mathrm{pair}} = (\lvert\rho_{\boldsymbol{q}}\rvert^2 - 1)/(N-1)$	The same matched projection and FiLM reader as the density comparison	p_{in} or p_{body}	Physical versus generic pair content and placement; Appendix E and Table 9.
One-body free or potential-dependent eigenmodes	Zero-initialized residual spectral reader	p_{body} or p_{orb}	Potential spectrum versus free spectrum and coordinate baseline; Appendix B.1.
Bloch-reference pair relations (exchange overlap, ridge-conditioned overlap, availability, density) or lattice-geometry similarity	Zero-initialized cross-attention with a null token, Eq. (9)	$p_{\mathrm{pair}} \to p_{\mathrm{body}}$	Occupied-state versus lattice-geometry relations; Table 6.
Soft moiré-cell geometry and simulation-cell density relations, with or without cell-occupation fluctuations	Zero-initialized grouped cross-attention with a null token, Eq. (9)	p_{body}	With versus without occupation fluctuations; Table 6.
Primitive-cell density power or complex density	Shared six-channel density reader	p_{body}	Historical width-96 context comparison; Table 10.
Local potential/force and collective density resources	Separate local body FiLM and global orbital FiLM, with a joint sum	p_{body} and p_{orb}	Local, collective, and joint historical controls; Appendix G.

A.4 Regularity

The full state must satisfy its boundary conditions and have a finite, nonzero norm within the Hamiltonian's quadratic-form domain. Pointwise local-energy estimation requires the corresponding derivative and integrability conditions. Smooth periodic resources and readers preserve periodicity, but finite variance near nodes or Coulomb coalescences requires separate analysis. Resources retained in the wavefunction continue to impose an inductive bias even without reference supervision.

B MOIRÉ RESOURCES AND REPORTED IMPLEMENTATION

The fully polarized periodic model uses $N = 24$, hidden width 64, and three grouped-query attention/MoE blocks. Each attention layer has eight query heads and two key/value heads. The MoE evaluates three routed experts and one shared expert, followed ultimately by a complex single-determinant readout. All principal comparisons retain this backbone width and block architecture. Historical width-96 runs are identified separately in Appendices E and G and are not evidence for the fixed width-64 backbone.

Denote the FiLM-conditioned particle token entering the MoE in Figure 1d by $u_{\ell i}(R)$. The inherited dense MoE is

$$\mathrm{MoE}_\ell(u_\ell)_i = E_{\ell,\mathrm{sh}}(u_{\ell i}) + \sum_{k=1}^{3} g_{\ell ik}(R) E_{\ell k}(u_{\ell i}),$$
$$g_{\ell i}(R) = \mathrm{softmax}(W_\ell u_{\ell i}(R) + b_\ell)\,.$$

The router and all four experts receive the same conditioned particle token. All experts are evaluated, and softmax weights combine the three routed outputs before the shared output is added without gating. Parameters are shared across particles, while inputs and routing weights depend on i and R. These internal experts have no assigned physical resources or spatial partitions; the density reader supplies one context to their common input stream.

The periodic coordinate adapter uses sine and cosine pairs at the wavevectors

$$\{\boldsymbol{b}_1, \boldsymbol{b}_2, -\boldsymbol{b}_1 - \boldsymbol{b}_2, 6\boldsymbol{b}_1, 6\boldsymbol{b}_2, -6(\boldsymbol{b}_1 + \boldsymbol{b}_2)\}.$$

The first two reciprocal basis vectors retain complete positions modulo the simulation cell. The density reader uses the following ordered coefficients in that basis:

$$(-1,0),(0,-1),(0,1),(1,0),(-2,0),(0,-2),(0,2),(2,0).$$

The eight complex modes are packed into sixteen real channels. Opposite modes obey $\rho_{-\boldsymbol{q}} = \rho_{\boldsymbol{q}}^*$ and therefore carry dependent information. The selected list does not form an isotropic reciprocal-space shell. For $\boldsymbol{\rho} = (\rho_{\boldsymbol{q}})$ and the fixed scale $\rho_{\mathrm{ref}} = 0.204976296284$, the normalization is

$$\mathcal{N}_\rho(\boldsymbol{\rho}) = \mathrm{pack}_{\mathbb{R}}\left(\frac{\boldsymbol{\rho}}{\sqrt{\rho_{\mathrm{ref}}^2 + \frac{1}{8}\sum_{a=1}^{8}|\rho_a|^2}}\right).$$

Normalization is evaluated for each configuration and contributes its full spatial derivatives.

The nonlinear reader has dimensions $16 \to 32 \to 8$. It computes $z(R)$ once per wavefunction evaluation; the repeated $z(R)$ boxes in Figure 1d denote this same context. Six separately parameterized FiLM heads supply scales and shifts at the backbone injection sites,

$$\gamma_p(z) = \mathbf{1} + 0.1\,\mathrm{sat}(\mathrm{Linear}_{p,\gamma}(z))\,,$$
$$\beta_p(z) = 0.1\,\mathrm{sat}(\mathrm{Linear}_{p,\beta}(z))\,.$$

Here each Linear is a learned affine map, initialized at zero to give neutral scales and shifts. The side variant applies these heads before attention and before the MoE in each block. The front variant applies all six transformations consecutively after the embedding, with no RMS normalization inserted between them. The reader and head dimensions match across variants, while their interaction with the backbone changes with placement. The concatenation control uses a different linear input interface, with total parameter counts listed in Table 8.

Appendix F specifies the gas implementation, including its orbital and relation-attention readers.

B.1 ONE-BODY SPECTRAL KERNELS IN THE MOIRÉ SYSTEM

The spectral comparison uses the one-body operator $h_1 = \frac{1}{2}(-\mathrm{i}\nabla - \boldsymbol{A})^2 + V_M(\boldsymbol{r})$ at $\lambda = 1$, with the potential and sign convention in Appendix C. A plane-wave Galerkin solve supplies the lowest $r = 25$ modes u_a. This is the first rank at least 24 whose window ends outside degeneracies for both h_1 and the free operator $h_0 = \frac{1}{2}(-\mathrm{i}\nabla - \boldsymbol{A})^2$. The window is fixed before neural optimization. Reciprocal-index cutoffs 30 and 36 agree within 9.82×10^{-12} in eigenvalues, 1.16×10^{-6} in evaluated kernels, 4.14×10^{-6} in first derivatives, and 2.81×10^{-5} in Laplacians on the reference configurations.

Writing $u_a(\boldsymbol{r}) = \sum_{\boldsymbol{G}} C_{\boldsymbol{G}a} e^{\mathrm{i}\boldsymbol{G}\cdot\boldsymbol{r}}$ with $C^\dagger C = I$, the resource and its residual reader are

$$K_{ij} = \frac{1}{r}\sum_{a=1}^{r} u_a(\boldsymbol{r}_i) u_a(\boldsymbol{r}_j)^*,$$
$$\Delta h_i = \frac{1}{N}\sum_j \left[\mathrm{Re}\, K_{ij}\, h_j W_R + \mathrm{Im}\, K_{ij}\, h_j W_I\right].$$

The kernel is invariant under unitary rotations of the retained modes. Its uniform spatial mean diagonal is one; no configuration normalization removes its local amplitude. The two real 64×64 reader matrices start at zero. Body reading reuses them at the six normalized attention/MoE inputs; orbital-input reading applies them once after the final normalization. Both variants add 8,192 parameters and retain the original initial wavefunction. All coordinate derivatives are included. The free control uses the same rank, basis, and reader, while retaining its own low-kinetic-energy spatial structure. Table 5 lists the endpoints.

Table 5: Moiré spectral comparisons at 10,000 updates in E_{h}^*, with conditional MCSEs from 32 draws of 64 walkers.

Model	Energy
Coordinate baseline	-7.34842 ± 0.03495
Physical spectrum, body	-7.30381 ± 0.04785
Free spectrum, body	-7.32510 ± 0.03651
Physical spectrum, orbital input	-7.40651 ± 0.02933
Free spectrum, orbital input	-7.23904 ± 0.03969

B.2 Relation-attention readers in the Moiré system

Both relation readers of Section 2.6 act in each of the three blocks on the normalized attention input a^ℓ, before backbone attention; the backbone attention and MoE are unchanged. Queries are $Q_\ell a_i^\ell$. The null token has a trainable logit per block and zero value. The output maps O_ℓ and their biases start at zero, so both readers recover the coordinate backbone at initialization (Proposition 2). Resource definitions are fixed before training and are recomputed with all coordinate derivatives for every configuration.

Pair relations. The reference is the lowest one-body band of h_1 (Appendix B.1) on the 6×6 supercell momentum grid, 36 Bloch functions ϕ_a with smooth occupations f_a of fixed trace $\sum_a f_a = 24$ and smearing parameter 0.036. With $\gamma_{ij} = \sum_a f_a \phi_a(\boldsymbol{r}_i)\phi_a(\boldsymbol{r}_j)^*$, $n_i = \gamma_{ii}$, and $P = (\gamma + \eta I)^{-1}$ for $\eta = 0.5$, the relations are

$$x_{ij} = \frac{|\gamma_{ij}|^2}{n_i n_j}, \qquad c_{ij} = \frac{|P_{ij}|^2}{P_{ii}P_{jj}}, \qquad s_i = \frac{P_{ii}^{-1} - \eta}{n_i}.$$

Here x is the normalized exchange overlap of the occupied kernel, s_i is a ridge-regularized Schur complement that measures how much occupied orbital space remains at $\boldsymbol{r}_i$ given all electrons, and c is the corresponding conditional overlap. The edge vector is $(x_{ij}, c_{ij}, s_i, s_j, b_{ij}, j_{ij}^x, j_{ij}^y, n_i, n_j)$. The occupied-state and lattice-geometry readers set $b = j^x = j^y = 0$. Two response variants of the occupied-state reader fill these channels from the static first-order response of the same reference to plane-wave perturbations at the six shortest simulation-cell wavevectors $\boldsymbol{q}$. Here $\delta n_i(\boldsymbol{q})$ and $\delta \boldsymbol{j}_i(\boldsymbol{q})$ denote the induced density and current at $\boldsymbol{r}_i$. We set $b_{ij} = \mathrm{Re}\sum_{\boldsymbol{q}} \delta n_i(\boldsymbol{q}) e^{-\mathrm{i}\boldsymbol{q}\cdot\boldsymbol{r}_j}$ and define the current channels $j_{ij}^{x,y}$ in the same way from $\delta \boldsymbol{j}_i$, with each channel divided by a fixed scale. The signed variant uses these values directly. The evenized variant replaces each value v by $\sqrt{v^2 + \delta^2} - \delta$ with $\delta = 10^{-3}$, which removes its sign. The lattice-geometry control uses the similarity in Section 2.6 for both x_{ij} and c_{ij} (squared in the second slot), with $s_i = n_i = 1$. In each block, keys are linear maps of (a_j, e_{ij}) together with three environment channels per endpoint: the row sums of (b, j^x, j^y). Values are linear maps of (a_j, e_{ij}) with a gate $1 + 0.1\tanh(\cdot)$ computed from the response channels. Self-attention to $j = i$ is masked. Query, key, and value widths are 32, and all four pair readers have the same shape and 27,747 added parameters.

Cell relations. Let $\boldsymbol{u}$ be fractional simulation-cell coordinates and $\boldsymbol{c}_a$ the 36 moiré-cell centers. The soft assignment

$$w_{ia} = \mathrm{softmax}_a\Big\{-\kappa \sum_{\boldsymbol{t}} \big[1 - \cos 2\pi \boldsymbol{t} \cdot (\boldsymbol{u}_i - \boldsymbol{c}_a)\big]\Big\}, \qquad \boldsymbol{t} \in \{(1,0),(0,1),(1,1)\},$$

uses a fixed κ for which the envelope near a center is Gaussian with width $a_M/3$. Each cell token carries the six cosines and sines of these phases, w_{ia}, type indicators, and the self-excluded occupation fluctuation

$$\chi_{ia} = \frac{n_a^{(-i)} - \bar{n}}{\sqrt{\bar{n}}}, \qquad n_a^{(-i)} = \sum_{j \neq i} w_{ja}, \qquad \bar{n} = \frac{N-1}{36},$$

which the density variant sets to zero. Six further tokens carry the direction and relative length of $\boldsymbol{q} \in \{\pm \boldsymbol{b}_1, \pm \boldsymbol{b}_2, \pm(\boldsymbol{b}_1 + \boldsymbol{b}_2)\}$ and the real and imaginary parts of $(N-1)^{-1/2} \sum_{j \neq i} e^{\mathrm{i}\boldsymbol{q}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)}$. All 42 tokens have 12 channels. Keys and values are read from the tokens alone, with four query heads, two key/value heads of width 8, and 13,827 added parameters.

Endpoints. All six models use training seed 20260731, the Adam protocol of Appendix D, and the readout protocol of Eq. (11). Table 6 gives the energy components. Relative to the backbone, the pair readers give -0.13447 ± 0.04580 (occupied states) and $-0.14039 \pm 0.04890\, E_{\mathrm{h}}^*$ (lattice geometry), and the cell readers give -0.12092 ± 0.04733 (density) and $-0.19144 \pm 0.04144\, E_{\mathrm{h}}^*$ (occupation). The signed and evenized response variants give -0.17900 ± 0.04391 and $+0.03485 \pm 0.07550\, E_{\mathrm{h}}^*$. For the pair reader, occupied-state minus lattice-geometry relations is $+0.00592 \pm 0.04523\, E_{\mathrm{h}}^*$, signed response minus the occupied-state reader is $-0.04453 \pm 0.03978\, E_{\mathrm{h}}^*$, evenized response minus the occupied-state reader is $+0.16932 \pm 0.07317\, E_{\mathrm{h}}^*$, signed minus evenized response is $-0.21385 \pm 0.07201\, E_{\mathrm{h}}^*$. For the cell reader, occupation minus density is $-0.07052 \pm 0.03891\, E_{\mathrm{h}}^*$, with interaction, kinetic, and external-potential changes of -0.16292 ± 0.02669, -0.01010 ± 0.15029, and $+0.10251 \pm 0.13412\, E_{\mathrm{h}}^*$. The evenized state has about twice the readout MCSE of the other readers. Errors combine marginal MCSEs in quadrature, because paired lineage has not been verified for these runs. The components share samples within each readout, so their errors are not independent of the total-energy error.

Table 6: Moiré relation-attention endpoints at 10,000 updates in E_{h}^*, with conditional MCSEs from 32 draws of 64 walkers and one training run per model. Components are the kinetic, Ewald interaction, and moiré-potential terms.

Model	Energy	Kinetic	Interaction	Potential
Coordinate backbone	-7.34842 ± 0.03495	5.77660 ± 0.08543	-6.54207 ± 0.02287	-6.58295 ± 0.07732
Pair, occupied states	-7.48289 ± 0.02960	5.86409 ± 0.10495	-6.66105 ± 0.01801	-6.68592 ± 0.09070
Pair, lattice geometry	-7.48881 ± 0.03421	5.82769 ± 0.10617	-6.62203 ± 0.02781	-6.69447 ± 0.09751
Pair, signed response	-7.52742 ± 0.02658	6.17855 ± 0.09898	-6.66527 ± 0.02141	-7.04070 ± 0.08299
Pair, evenized response	-7.31357 ± 0.06692	5.88665 ± 0.09983	-6.60560 ± 0.02141	-6.59462 ± 0.07642
Cell, density	-7.46934 ± 0.03191	5.96263 ± 0.11326	-6.59623 ± 0.02203	-6.83574 ± 0.09879
Cell, occupation	-7.53986 ± 0.02226	5.95252 ± 0.09878	-6.75915 ± 0.01507	-6.73324 ± 0.09071

C Reported Hamiltonian and units

C.1 Periodic moiré benchmark

In effective atomic units, the periodic family is

$$H(\lambda) = \frac{1}{2} \sum_i (-\mathrm{i}\nabla_i - \boldsymbol{A})^2 + \lambda \sum_i V_M(\boldsymbol{r}_i) + W_{\mathrm{Ew}}(R),$$

$$V_M(\boldsymbol{r}) = -2V_0 \sum_{a=1}^{3} \cos(\boldsymbol{g}_a \cdot \boldsymbol{r} + \varphi).$$

The finite-potential setting uses $\lambda = 1$, and the zero-potential setting uses $\lambda = 0$. Both have $N = 24$, full spin polarization, and $\boldsymbol{A} = (0.5, 0)$ in units of $(a_{\mathrm{B}}^*)^{-1}$. The triangular supercell contains 6×6 primitive cells at filling $2/3$. Table 7 lists the physical parameters used in these calculations.

Table 7: Periodic Hamiltonian parameters. Moiré energies are total energies in E_{h}^{*}. The Fermi-gas calculations use the separate energy unit specified in Appendix F.

Quantity	Value
m^{*}/m_{e}, ϵ_{r}	0.35, 10
a_{B}^{*}, E_{h}^{*}	1.5119348883 nm, 95.2398518610 meV
a_{M}, V_{0}, φ	$5.3117366774\,a_{\mathrm{B}}^{*}$, $0.1574970950\,E_{\mathrm{h}}^{*}$, $\pi/4$
$\boldsymbol{L}_{1}/a_{\mathrm{B}}^{*}$	$(31.8704200644, 0)$
$\boldsymbol{L}_{2}/a_{\mathrm{B}}^{*}$	$(15.9352100322, 27.6005934051)$
$\boldsymbol{b}_{1}a_{\mathrm{B}}^{*}$, $\boldsymbol{b}_{2}a_{\mathrm{B}}^{*}$	$(0.1971478661, -0.1138233736)$, $(0, 0.2276467471)$
$\boldsymbol{g}_{1}, \boldsymbol{g}_{2}, \boldsymbol{g}_{3}$	$6\boldsymbol{b}_{1}$, $6\boldsymbol{b}_{2}$, $-6(\boldsymbol{b}_{1}+\boldsymbol{b}_{2})$

The two-dimensional Ewald convention includes the neutralizing background and self term. For cell area Ω and $\eta_{\mathrm{Ew}} = \sqrt{\Omega/\pi}$, the interaction is

$$\begin{aligned} v_{\mathrm{Ew}}(\boldsymbol{r}) &= \sum_{\boldsymbol{L}} \frac{\operatorname{erfc}(|\boldsymbol{r}+\boldsymbol{L}|/(2\eta_{\mathrm{Ew}}))}{|\boldsymbol{r}+\boldsymbol{L}|} \\ &\quad + \frac{2\pi}{\Omega} \sum_{\boldsymbol{G}\neq 0} \frac{\operatorname{erfc}(\eta_{\mathrm{Ew}}|\boldsymbol{G}|)}{|\boldsymbol{G}|} \cos(\boldsymbol{G}\cdot\boldsymbol{r}) - \frac{4\pi\eta_{\mathrm{Ew}}}{\Omega\sqrt{\pi}}, \\ W_{\mathrm{Ew}}(R) &= \sum_{i<j} v_{\mathrm{Ew}}(\boldsymbol{r}_{i}-\boldsymbol{r}_{j}) + \frac{N}{2}\xi_{\mathrm{Ew}}, \\ \xi_{\mathrm{Ew}} &= \sum_{\boldsymbol{L}\neq 0} \frac{\operatorname{erfc}(|\boldsymbol{L}|/(2\eta_{\mathrm{Ew}}))}{|\boldsymbol{L}|} + \frac{2\pi}{\Omega} \sum_{\boldsymbol{G}\neq 0} \frac{\operatorname{erfc}(\eta_{\mathrm{Ew}}|\boldsymbol{G}|)}{|\boldsymbol{G}|} \\ &\quad - \frac{4\pi\eta_{\mathrm{Ew}}}{\Omega\sqrt{\pi}} - \frac{1}{\eta_{\mathrm{Ew}}\sqrt{\pi}}. \end{aligned}$$

The real- and reciprocal-space sums use integer coefficients $-5, \ldots, 5$ in each direction, with the exclusions shown above. Any change to the Coulomb multiplier applies to all of W_{Ew}, including its constant terms. Reference comparisons require matching boundary and Ewald conventions, with converged sum truncations.

The wavefunction obeys ordinary periodic boundary conditions in the fixed gauge. The constant $\boldsymbol{A}$ has zero curl and represents a boundary holonomy with phases $-\boldsymbol{L}_{a}\cdot\boldsymbol{A}$ modulo 2π. It gives no nonzero magnetic field and does not establish magnetic topology. Changing only the external potential leaves this boundary condition unchanged.

D Differentiation, optimization, and statistical protocol

D.1 Spatial derivatives and variational gradients

For a constant vector potential and $\hbar = m = 1$, the local energy away from nodes is

$$E_{\mathrm{L}} = -\frac{1}{2}\sum_{i}\left[\nabla_{i}^{2}\log\Psi + \nabla_{i}\log\Psi\cdot\nabla_{i}\log\Psi\right] + \mathrm{i}\boldsymbol{A}\cdot\sum_{i}\nabla_{i}\log\Psi + \frac{N|\boldsymbol{A}|^{2}}{2} + V(R).$$

The gradient product contains no complex conjugation. Every configuration-dependent resource enters these derivatives through the chain rule below.

Locally write $\log\Psi(R) = \mathcal{L}(R, f(R))$, where $\mathcal{L}$ contains the explicit coordinate processing at fixed parameters. Index real resource channels by a and coordinate components by ν. Subscripts on $\mathcal{L}$ denote partial derivatives with coordinates and resources treated as independent arguments. The first derivative is

$$\partial_{\nu}\log\Psi = \mathcal{L}_{\nu} + \sum_{a}\mathcal{L}_{a}\partial_{\nu}f_{a},$$

and

$$\partial_\nu^2 \log \Psi = \mathcal{L}_{\nu\nu} + 2 \sum_a \mathcal{L}_{\nu a} \partial_\nu f_a + \sum_{a,b} \mathcal{L}_{ab} (\partial_\nu f_a)(\partial_\nu f_b) + \sum_a \mathcal{L}_a \partial_\nu^2 f_a.$$

Resource normalization couples the channels, and its derivatives enter both expressions. Pure Hamiltonian resources have zero coordinate derivatives at fixed H. Omitting derivatives through a configuration-dependent resource would generally change the local-energy expression for the evaluated wavefunction. Real and imaginary parts require consistent treatment near branch choices for $\log \det \Phi$. These identities apply locally away from the nodes of the wavefunction.

For real parameters Θ_a and $O_a = \partial_{\Theta_a} \log \Psi$, the unclipped variational gradient is

$$\partial_{\Theta_a} E_H = 2 \operatorname{Re} \mathbb{E}_{|\Psi|^2} \left[(E_{\mathrm{L}} - E_H) O_a^* \right],$$

under the usual differentiation and integrability assumptions. The moiré training protocol uses Adam with clipping of complex local-energy deviations around a component-wise complex median. The threshold is five times the median deviation magnitude, with raw values retained when that median is zero. Clipping is an optimization heuristic and changes the estimator used for parameter updates. Final energies and observables are always evaluated from raw, unclipped samples.

D.2 Moiré readout protocol and conditional uncertainty

Training uses 128 persistent walkers per Hamiltonian. At each readout, we freeze the parameters and clone a 64-walker readout bank from the training state together with its random-number state. The readout therefore leaves the training walker bank and its state unchanged. We perform 512 single-particle moves per walker, then record 32 draws separated by 16 moves. All reported checkpoint energies are computed from unclipped local-energy values.

For a scalar readout y_{tc} with $T = 32$ draws and $C = 64$ chains, define the overall, chain, and temporal-block means as

$$\overline{y} = \frac{1}{TC} \sum_{t,c} y_{tc}, \qquad a_c = \frac{1}{T} \sum_t y_{tc},$$

$$b_k = \frac{1}{4C} \sum_{t=4k-3}^{4k} \sum_c y_{tc}, \quad k = 1, \ldots, 8.$$

The error rule is

$$\epsilon_y = \max \left\{ \frac{\mathrm{sd}(a_1, \ldots, a_{64})}{\sqrt{64}}, \frac{\mathrm{sd}(b_1, \ldots, b_8)}{\sqrt{8}} \right\}, \tag{11}$$

where each standard deviation is the sample standard deviation. For the mode average $N_q^{-1} \sum_{\boldsymbol{q}} S(\boldsymbol{q})$, we first average over modes for each (t, c) sample. We then apply the error estimator to these averages, preserving dependence between signed modes.

For contrasts with verified common random-number lineage and a matching readout protocol, we apply Eq. (11) to the samplewise difference $y_{tc}^{(1)} - y_{tc}^{(2)}$. Common random numbers can induce covariance even when the models accept different configurations. For contrasts without verified pairing, we combine the marginal errors in quadrature. This convention need not be conservative when the covariance is unknown. The primitive-cell density contrasts have verified paired lineage, while gas contrasts use quadrature under the protocol in Appendix F.

The readout also records the second-half mean minus the first-half mean. This drift diagnostic and the maximum-error rule do not establish equilibration. The reported errors remain conditional sampling estimates and exclude training-seed variability and exact ground-state error.

Extended moiré readouts continue the same cloned chains to $T = 256$ draws and reproduce the original 32-draw prefix. Equation (11) then uses eight temporal blocks of 32 draws and all 64 chain means. The last 224 draws provide a consistency check and overlap the full estimate. The gas diagnostic records first-half minus second-half drift, opposite to the moiré convention.

D.3 PHYSICAL OBSERVABLES

With the $N^{-1/2}$ normalization in Eq. (7), we report eight structure factors and their mean,

$$S(\boldsymbol{q}) = \mathbb{E}|\rho_{\boldsymbol{q}}|^2, \qquad \overline{S} = \frac{1}{8}\sum_{\boldsymbol{q}\in\mathcal{Q}} S(\boldsymbol{q}).$$

The mode average is formed for each sample before estimating uncertainty, so signed modes are not treated as independent replicates. In the fixed gauge, the total current is defined by

$$\boldsymbol{J} = \mathbb{E}\left[\sum_i (\mathrm{Im}\,\nabla_i \log\Psi - \boldsymbol{A})\right].$$

This convention reports total current, with no division by the particle number N. The full complex local-energy variance $\mathbb{E}|E_{\mathrm{L}} - \mathbb{E}E_{\mathrm{L}}|^2$ also differs from the variance of the real part alone.

E DENSITY COMPARISONS AND A HISTORICAL WIDER MODEL

Table 8 documents the fixed-backbone density comparison summarized in Section 3. Table 10 retains a historical run with a wider backbone. All runs use training seed 20260731. They evaluate shared density conditioning, with the fixed-backbone implementation shown in Figure 1d. Error bars are conditional MCSEs at frozen checkpoints and exclude variability across independent training runs.

Table 8: Fixed width-64 backbone: density comparisons at 10,000 updates. Total energy is in E_{h}^*.

Model	Parameters	Total energy ± MCSE
Coordinate-only MoE	235097	-7.34842 ± 0.03495
Density concat	236121	-7.30530 ± 0.03558
Generic concat	236121	-6.95627 ± 0.08311
Density front	242817	-7.33967 ± 0.03362
Generic front	242817	-7.23789 ± 0.03955
Density side	242817	-7.44560 ± 0.02896
Generic side	242817	-7.31610 ± 0.03910

At width 64, density side minus baseline is $-0.09718 \pm 0.04636\,E_{\mathrm{h}}^*$ and side minus front is $-0.10593 \pm 0.04672\,E_{\mathrm{h}}^*$ under the same update budget, with paired errors. Combining the marginal MCSEs in quadrature instead gives $0.0454\,E_{\mathrm{h}}^*$ for the first difference. Matching reader parameters leaves different interactions with normalization and nonlinear layers, so front and side need not have equal expressivity. The generic Fourier controls also differ in bandwidth and derivative scale.

The global pair harmonic used in the comparison satisfies

$$f_{\boldsymbol{q}}^{\mathrm{pair}} = \frac{2}{N(N-1)}\sum_{i<j}\cos[\boldsymbol{q}\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)] = \frac{|\rho_{\boldsymbol{q}}|^2 - 1}{N-1}.$$

Overlapping pair and density modes therefore encode related quantities through this identity. Their different mode lists prevent attribution of performance differences to nonlinearity or phase alone. Equation (10) defines a separate interface for particle-resolved resources.

Table 9 compares reading locations along the fixed width-64 trajectories. Density conditioning inside the backbone retains a lower endpoint, although its advantage over front conditioning narrows relative to 4,000 updates. Pair harmonics favor different locations at intermediate checkpoints and show no resolved endpoint preference. At 10,000 updates, the content-by-location interaction, defined as the physical side-minus-front contrast minus its generic counterpart, is $-0.02772 \pm 0.06360\,E_{\mathrm{h}}^*$ for density and $+0.03397 \pm 0.06964\,E_{\mathrm{h}}^*$ for pair harmonics. Neither resolves a physical-content-specific advantage of side placement over the corresponding generic placement effect.

Side minus baseline is $+0.04673 \pm 0.04145\,E_{\mathrm{h}}^*$ in the wider model. This endpoint gives no resolved advantage for density conditioning at the larger budget. The comparison concerns shared-stream conditioning and provides no test of resource assignment to individual experts.

Table 9: Dependence of placement comparisons on optimization budget, with the fixed width-64 backbone. Entries are paired side-minus-front total energies in E_{h}^{*}, with conditional MCSEs. Negative values favor repeated conditioning within the backbone.

Updates	Density modes	Pair harmonics
1,000	-0.05839 ± 0.10678	$+0.12911 \pm 0.08405$
2,000	-0.24998 ± 0.07967	-0.17430 ± 0.07655
4,000	-0.33479 ± 0.06463	$+0.23070 \pm 0.07494$
6,000	-0.04728 ± 0.06210	$+0.05132 \pm 0.05539$
8,000	-0.15934 ± 0.05689	$+0.03904 \pm 0.06405$
10,000	-0.10593 ± 0.04672	$+0.02113 \pm 0.04932$

Table 10: Historical width-96 comparison at 20,000 updates, separate from the fixed-backbone results. The side model no longer has a lower mean energy in this comparison. Residual training and readout drift prevents either endpoint from establishing the variational limit.

Historical model	Total energy $\pm$ MCSE (E_{h}^{*})
Coordinate-only MoE	-7.75599 ± 0.02518
Density side	-7.70926 ± 0.02925

E.1 Hamiltonian-specific one-body spectral control

The one-body spectral comparison in Section 3 is a frozen-state readout at $\lambda = 1$ and $\boldsymbol{A} = (0.5, 0)$, against the coordinate baseline trained at that single Hamiltonian. Its endpoint energies are -7.34842 ± 0.03495 for the baseline and -7.40651 ± 0.02933 for the physical spectrum at the orbital input, a difference of $-0.05808 \pm 0.04563\, E_{\mathrm{h}}^{*}$ with the MCSEs combined in quadrature. We therefore report it as unresolved. Against the free-particle spectrum with the same rank, basis, reader, and 8,192 added parameters (Appendix B.1), the physical spectrum is lower by $0.16747 \pm 0.04936\, E_{\mathrm{h}}^{*}$ at the orbital input, whereas inside the blocks the difference is $+0.02129 \pm 0.06019\, E_{\mathrm{h}}^{*}$. The free spectrum at the orbital input ends $0.10938 \pm 0.05289\, E_{\mathrm{h}}^{*}$ above the baseline, so the orbital-input contrast combines a small gain from the physical modes with a loss from the free modes. At 20,000 updates, the supplementary spectral readouts give physical minus free spectrum at the orbital input as $-0.02487 \pm 0.03809\, E_{\mathrm{h}}^{*}$, and physical spectrum minus backbone is $+0.03814 \pm 0.03358\, E_{\mathrm{h}}^{*}$; both are unresolved. All values come from one training run per model.

F Fermi gas, physical functions, and additional comparisons

F.1 Hamiltonian and solver protocol

The gas contains $N = 14$ particles, with $N_{\uparrow} = N_{\downarrow} = 7$, in a cubic periodic box at zero twist. We set $\hbar = m = k_F = v_0 = 1$, $\mu = 10$, and $L = (3\pi^2 N)^{1/3} = 7.456179338422642$. The target Hamiltonian for all gas comparisons is

$$H = -\frac{1}{2}\sum_i \nabla_i^2 - 2v_0\mu^2 \sum_{i\in\uparrow, j\in\downarrow} \operatorname{sech}^2(\mu r_{ij}^{\mathrm{MIC}}), \tag{12}$$

where r^{MIC} is the minimum-image Euclidean distance. The sum includes 49 opposite-spin pairs, with infinite vacuum scattering length and $k_F r_e = 0.2$ (Kim et al., 2024). We report total energies in $\hbar^2 k_F^2/m$, rather than effective Hartree or energy per particle. The Hamiltonian uses the minimum-image distance. The orbital resource uses the chord distance below; the communication resource uses a compact minimum-image continuation defined in Appendix F.4.

The general solver uses width 64, three attention/MoE blocks, eight query heads, two key/value heads, and one complex determinant. Every augmentation retains its 233,861 parameters and optimizes all backbone and reader weights. The parent orbital reader has 263,025 total parameters, the static spectral variant has 263,417, and each dynamic shape variant has 273,665. Each body-plus-shape model has 274,049 parameters, including its additional relation reader. Added readout coefficients

start at zero, while the backbone retains its original initialization rule. The runs use neither plane-wave initialization nor reference pretraining before target-system optimization.

Each run uses seed 20260731 and 512 persistent walkers. The original general, orbital, and specialized-reference runs distribute these walkers over two RTX 4090 GPUs. The subsequent communication comparisons use one device, with the same global walker count, initialization, Hamiltonian, and optimization rule. Their reference is the existing single-device spectral-and-scattering trajectory. The overall gap comparison uses the original general and specialized-reference trajectories; the device layouts are not independent training replicates. We optimize directly at the target interaction using stochastic reconfiguration (SR) with learning rate 10^{-4}. SR uses a diagonal moving scale and initial damping 10^{-3}. Conjugate-gradient solves use tolerance 10^{-8} and at most 200 iterations, with failed solves increasing damping by a factor of ten.

The specialized reference implements the Pfaffian–Jastrow–backflow construction of Kim et al. (2024) with two message-passing layers and 8,530 parameters. Its adaptive damping reached 10, while the general and shape trajectories retained 10^{-3}. The optimization rule is shared, although the realized damping histories differ between models. This direct-optimization comparison does not reproduce the original paper's transfer-learning schedule or converged results.

A closed-shell free Slater state provides an additional physical scale under the same Hamiltonian. Its kinetic energy is 4.26067391 and its interaction expectation is -0.24434610, giving total energy 4.01632781. The kinetic term follows from the occupied plane waves. The interaction term is obtained by integrating the potential over independent, uniform opposite-spin positions. Both the general solver and the trained pairing reference remain above this simple trial state. Consequently, the reported reductions measure finite-budget learning within the prescribed initialization and training protocol, rather than competitive ground-state accuracy or established pairing order.

F.2 Spectral and scattering providers

For a cubic periodic box, $e^{i\boldsymbol{k}\cdot\boldsymbol{r}}$ is an eigenfunction of $-\nabla^2/2$ with eigenvalue $|\boldsymbol{k}|^2/2$, where $\boldsymbol{k} = 2\pi\boldsymbol{n}/L$. Selecting small $|\boldsymbol{n}|$ therefore supplies functions with low one-body kinetic energy. The box geometry fixes the selected modes, whose values are evaluated at the current particle positions. They are a prescribed basis for orbital corrections, rather than orbitals obtained by optimizing an auxiliary mean-field problem.

Let $\mathcal{Q}_+$ contain one representative from each pair $\{\boldsymbol{n}, -\boldsymbol{n}\}$, with $\boldsymbol{n} \in \{-1, 0, 1\}^3$ and $0 < |\boldsymbol{n}|^2 \le 2$. The nine representatives define the spatial basis

$$u(\boldsymbol{r}) = \left(1, \{\sqrt{2}\cos(2\pi\boldsymbol{n}\cdot\boldsymbol{r}/L), \sqrt{2}\sin(2\pi\boldsymbol{n}\cdot\boldsymbol{r}/L)\}_{\boldsymbol{n}\in\mathcal{Q}_+}\right).$$

Applying spin masks to this 19-component basis gives $U \in \mathbb{R}^{N\times 38}$.

The short-range provider is motivated by two-body scattering in the Pöschl–Teller potential of Eq. (12) (Kim et al., 2024). In the equal-mass relative coordinate, the zero-energy radial equation at $v_0 = 1$ is

$$\left[-\frac{d^2}{dr^2} - 2\mu^2\,\mathrm{sech}^2(\mu r)\right] u_0(r) = 0, \qquad u_0(r) = \tanh(\mu r).$$

Thus $f_0(r) = u_0(r)/(k_F r)$ is a dimensionless s-wave shape, regular at the origin and proportional to $1/r$ beyond the potential range. This vacuum solution supplies local interaction knowledge. To evaluate it smoothly on the periodic configuration domain, the provider replaces r by a chord distance and takes the logarithm:

$$(r_{ij}^{\mathrm{chord}})^2 = (L/\pi)^2 \sum_d \sin^2[\pi(r_{id} - r_{jd})/L], \quad t_{ij} = \mu^2 (r_{ij}^{\mathrm{chord}})^2,$$

$$g(t) = \log\left[\frac{\mu}{k_F}\frac{\tanh\sqrt{t}}{\sqrt{t}}\right].$$

The value at the origin is obtained by analytic continuation. With opposite-spin mask $m_{ij} = \mathbf{1}_{\sigma_i\neq\sigma_j}$ and cutoff $t_c = (\mu L/N^{1/3})^2$, the physical relation is

$$D_{ij} = m_{ij} w(t_{ij}/t_c)\,[g(t_{ij}) - g(t_c)],$$
$$w(s) = \begin{cases} (1-s)^3, & s < 1, \\ 0, & s \ge 1. \end{cases} \tag{13}$$

The taper makes D and its first two derivatives vanish at the healing boundary. It follows the physical idea of matching short-range correlations to an uncorrelated value at a finite distance (Gandolfi, 2013), but its distance and polynomial form are prescribed here. The periodic and healed descriptor is consequently not an exact two-body solution in the simulation box, an optimized pair orbital, or a measure of pair condensation. The matched control sets $D_{ij} = m_{ij}$ while retaining U, q_i, C, and all parent readers. Both variants retain the same physical spectrum and trainable reader. Changes in function rank and coordinate derivatives are part of the tested change in relation content.

In Eq. (2), the fixed compression $P \in \mathbb{R}^{64\times 4}$ contains orthonormal discrete-cosine columns,

$$P_{jc} = \sqrt{2/64}\cos[\pi(j+1/2)c/64], \qquad j = 0,\dots,63, \quad c = 1,\dots,4.$$

The resulting q_i has five components, and $C \in \mathbb{C}^{2\times 5\times 38\times 14}$ adds 10,640 real parameters. The compression P is a computational choice with no assigned physical meaning. Features h_i are taken after the parent's final orbital FiLM and before its orbital projection. Both function banks and their products with q_i retain all spatial derivatives.

The static spectral reader uses seven complex plane waves per spin from the lowest shell, giving $U_7 \in \mathbb{C}^{14\times 14}$. It adds $U_7 A$ to the parent orbital matrix, with $A \in \mathbb{C}^{14\times 14}$ initialized at zero. At 10,000 updates, its energy was 9.82701 ± 0.03379, compared with 10.30149 ± 0.02978 for the parent constant orbital reader. The added spectral computation changed the energy by -0.47448 ± 0.04504. Comparing this reader with the dynamic shape model changes both the mode library and the reader, leaving their individual contributions unresolved.

F.3 Body relations and composition

The body reader uses the vacuum relation $D^{\mathrm{vac}}_{ij} = m_{ij} g(t_{ij})$. In each head of the three attention layers, learned relation vectors a^K and a^V modify the logits and values. With attention queries $\boldsymbol{q}_i$, keys $\boldsymbol{k}_i$, and values $\boldsymbol{v}_i$, the reader is

$$s_{ij} = T(\boldsymbol{q}_i^\top \boldsymbol{k}_j/\sqrt{8}) + b_{ij} + D^{\mathrm{vac}}_{ij}\boldsymbol{q}_i^\top a^K/\sqrt{8},$$
$$y_i = \sum_j \mathrm{softmax}_j(s_{ij})(\boldsymbol{v}_j + D^{\mathrm{vac}}_{ij} a^V).$$

Here T and b denote the parent's logit transformation and bias, and each backbone head has width 8. The key term is added outside T, preserving the inherited output projection and residual order. The 384 added real parameters are all initialized at zero. The constant body control replaces D^{vac} with m and retains the healed orbital provider. In both models, the updated backbone features determine the orbital mixing coefficients. The two readers therefore interact through the shared feature computation.

F.4 Pair-matching communication and low-energy subspace reading

The combined route retains the spectral-and-scattering orbital parent and adds the pair reader in each of its three backbone blocks, as shown in Figure 1b,c. The body-only control instead adds that reader to the coordinate baseline. Both recompute the prescribed functions at every configuration and retain all coordinate derivatives, including those of kernel normalization. The orbital parent also includes its inherited constant-input conditioning and orbital-reader parameters; body-only omits that parent construction, so it is not obtained solely by zeroing C in an otherwise identical full model.

Closed-shell and compact scattering functions. Let $k_0 = 2\pi/L$ and define the seven real closed-shell modes

$$v(\boldsymbol{r}) = \frac{1}{\sqrt{7}}\big(1, \{\sqrt{2}\cos(k_0 r_d)\}_{d=x,y,z}, \{\sqrt{2}\sin(k_0 r_d)\}_{d=x,y,z}\big).$$

Their overlap is $G_{ij} = v(\boldsymbol{r}_i)^\top v(\boldsymbol{r}_j) = [1 + 2\sum_d \cos k_0(r_{id} - r_{jd})]/7$. The second-harmonic kernel $G^{(2)}$ replaces k_0 by $2k_0$.

For this reader, the scattering function uses a compact minimum-image continuation, distinct from the orbital provider in Eq. (13). Set $s = (r^{\mathrm{MIC}}_{ij})^2$, $r_h = k_F^{-1}$, $r_c = L/N^{1/3}$, and

$$g_{\mathrm{c}}(s) = \log\frac{\tanh(\mu\sqrt{s})}{\mu\sqrt{s}}, \qquad d_0 = -g_{\mathrm{c}}(r_c^2),$$
$$D^{\mathrm{c}}_{ij} = w_{\mathrm{c}}(s)[g_{\mathrm{c}}(s) + d_0].$$

The value at $s = 0$ is defined by analytic continuation. The taper equals one for $s \le r_h^2$ and zero for $s \ge r_c^2$. Between these limits,

$$w_c(s) = 1 - I_q(7, 7), \qquad q = \frac{s - r_h^2}{r_c^2 - r_h^2},$$

where I_q is the regularized incomplete beta function. This gives a C^6 cutoff before the minimum-image seam. The additional radial channels are

$$T_{ij} = -2s\,\partial_s D^c_{ij}, \qquad I_{ij} = w_c(s)\,\mathrm{sech}^2(\mu\sqrt{s}), \qquad W_{ij} = w_c(s)e^{-k_F^2 s/2}.$$

They respectively describe the radial response of the log-scattering shape, the normalized attractive-potential shape, and a smooth local envelope. Each opposite-spin pair receives the six channels $\xi_{ij} = (D^c_{ij}/d_0, T_{ij}/4, I_{ij}, W_{ij}, G_{ij}, G^{(2)}_{ij})$. The matching matrix in Eq. (3) is

$$F_{ij} = G_{ij}\exp(D^c_{ij})/s_0, \qquad s_0 = 1.9128373948645523,$$

with the scale fixed before training. Its two-pair minors have the exchange algebra of an antisymmetrized product of two pair functions. They organize communication; the final fourteen-particle antisymmetry is supplied by the retained determinant.

Pair features and particle messages. Equations (4) and (5) define the pair update and readout, with $i, k \in \uparrow$ and $j, l \in \downarrow$. The pair features have 32 channels. Each barred term averages the other six entries in the same row or column, excluding (i, j). The five maps L are shared across layers; the query, key, value, and output maps are layer-specific. The two width-16 heads use the priors

$$p^1_{ij} = \frac{1/7 + W_{ij}}{1 + \sum_{j'} W_{ij'}}, \qquad p^2_{ij} = \frac{1}{7}.$$

The same maps serve both spin sectors, with transposed pair indices for the reverse direction. Equation (6) adds the pair message to the attention residual before the updated particle stream reaches the MoE. Output maps O_ℓ start at zero; other projections use a separate Xavier initialization, preserving the parent's initial wavefunction. The reader adds 30,912 parameters, for 304,577 in the combined model. Pair state persists across layers of one wavefunction evaluation and is reset for every new configuration, including sampling proposals and local-energy evaluations.

Body-only retains the pair update in Eq. (4) and the cross-attention map in Eq. (5), with the coordinate baseline as parent and no orbital adapter. No-history retains the orbital parent but sets the incoming $e^{\ell-1}_{ij}$ to zero at every block. This removes the old-state residual, row/column means, and matching action together; their associated maps are inactive, while current pair channels, particle features, and cross-attention remain active. The bare-kernel control retains the same state and reader but substitutes

$$(F_0)_{ij} = G_{ij}/s_0 \quad \text{for} \quad F_{ij} = G_{ij}\exp(D^c_{ij})/s_0,$$

and evaluates the normalization in Eq. (3) with the selected kernel. It retains the local channels ξ, prior envelope W, and orbital parent. The antisymmetrized two-pair subtraction is unchanged. The Fourier-pair control instead replaces (ξ, F, W) together, while retaining the same reader and orbital parent.

For opposite-spin pairs, set $\boldsymbol{\theta}_{ij} = 2\pi(\boldsymbol{r}_i - \boldsymbol{r}_j)/L$. For each shell $s = 1, \dots, 6$, let $\mathcal{Q}_s \subset \mathbb{Z}^3$ contain one representative from each $\{\boldsymbol{q}, -\boldsymbol{q}\}$ with $|\boldsymbol{q}|^2 = s$. The Fourier-pair resource uses the even, cubic-invariant shell functions

$$\zeta_s(\boldsymbol{\theta}) = \sqrt{\frac{2}{|\mathcal{Q}_s|}} \sum_{\boldsymbol{q} \in \mathcal{Q}_s} \cos(\boldsymbol{q} \cdot \boldsymbol{\theta}), \qquad s = 1, \dots, 6.$$

For channel indices $c = 0, 1, 2, 4, 5$, set $\widetilde{\xi}_{ij,c} = \mu_c + \sigma_c \zeta_{s(c)}(\boldsymbol{\theta}_{ij})$, using shells $s(c) = 1, 2, 3, 4, 5$, respectively. The remaining channel is the positive prior envelope,

$$\widetilde{\xi}_{ij,3} = \widetilde{W}_{ij} = A_W \exp\left[\kappa \sum_{d=1}^{3} \cos\theta_{ij,d}\right],$$

$$\widetilde{F}_{ij} = \mu_F + \frac{\sigma_F}{\sqrt{6}} \sum_{s=1}^{6} \zeta_s(\boldsymbol{\theta}_{ij}).$$

Uniform-geometry calibration fixes the means and standard deviations before training, without energies or trained configurations. For envelope mean μ_W and variance σ_W^2, $A_W = \mu_W / I_0(\kappa)^3$ and $I_0(2\kappa)^3/I_0(\kappa)^6 - 1 = \sigma_W^2/\mu_W^2$, where I_0 is the modified Bessel function. This matches individual resource means and variances, while leaving spatial correlations and coordinate-derivative statistics unmatched.

The pair update uses $\widetilde{\xi}$, the matching action is Eq. (3) with $F = \widetilde{F}$, and the first attention head uses $\widetilde{W}$ with the same uniform floor. The pair reader, injection site, and healed-scattering orbital parent are retained, as are all derivatives through the replacement resources.

Reading a low-energy subspace across layers. The subspace reader uses the same seven modes in $V_{ia} = v_a(\boldsymbol{r}_i)$. Define charge and spin channels $V_{\mathrm{c}} = V$ and $V_{\mathrm{s}} = \mathrm{diag}(\sigma_i)V$, where $\sigma_i = \pm 1$. It fits the current particle features to these channels using

$$Z^{\ell}_{t,\mathrm{fit}} = (V^\top V + \tfrac{1}{2} I)^{-1} V_t^\top a^{\ell},$$
$$Z^{\ell}_{t} = \tfrac{1}{2} Z^{\ell-1}_{t} + \tfrac{1}{2} Z^{\ell}_{t,\mathrm{fit}}, \qquad t \in \{\mathrm{c}, \mathrm{s}\}.$$

The first layer uses its fit directly. The reader returns $\sum_t V_t Z_t^{\ell} B_{\ell t}$ at the same site as the pair messages, with zero-initialized 64×64 matrices $B_{\ell t}$. It adds 24,576 parameters. These coefficients organize hidden features in a prescribed low-energy basis; they are not measured occupations of the interacting state. The combined reader computes both branches from the same a^{ℓ} and adds their messages, with independent states across layers.

Exchange-hole and Jastrow comparisons. A simpler same-spin resource follows the closed-shell exchange-hole shape. For distinct same-spin particles, define

$$w_{ij} = \tfrac{1}{2} + \tfrac{1}{2}\tfrac{7}{6}(1 - G_{ij}^2), \qquad b_{ij} = \log w_{ij} - \tfrac{1}{6} \sum_{k \neq i,\, \sigma_k = \sigma_i} \log w_{ik}.$$

A learned scalar per layer adds b to the original attention logits, with zero initial coefficients and no change to opposite-spin logits. The Jastrow comparison multiplies the wavefunction by $\exp(J)$, where $J = 1.25 \tanh(a_J) \sum_{i\uparrow, j\downarrow} D^{\mathrm{c}}_{ij}$ and a_J starts at zero. The full kinetic energy includes its cross terms with the neural determinant. We evaluate this correction alone and together with pair-matching communication.

F.5 Communication endpoints and controlled alternatives

Table 11 compares the completed readers using independently initialized 256-draw readouts at the 10,000-update checkpoint. The pair-matching gain is primarily kinetic, whereas the Jastrow factor trades lower interaction energy against higher kinetic energy. The subspace reader's reduction relative to the same reference narrowed from 3.99% of the reference energy at 4,000 updates (standard readout) to 0.04950 ± 0.02110, or 0.53%, at the independent 10,000-update readout.

Table 11: Completed communication and correlation readers at 10,000 updates. Each addition retains the common spectral-and-scattering orbital reference. Entries are total energies and components in $\hbar^2 k_F^2/m$, with conditional MCSEs from 256-draw readouts of one training run.

Added reader	Total energy	Kinetic energy	Interaction energy
None	9.2884 ± 0.0168	9.5434 ± 0.0093	-0.2551 ± 0.0134
Pair matching	8.7389 ± 0.0117	9.0018 ± 0.0104	-0.2629 ± 0.0104
Low-energy subspace	9.2389 ± 0.0127	9.4887 ± 0.0076	-0.2498 ± 0.0114
Short-range Jastrow	9.2509 ± 0.0197	9.5893 ± 0.0113	-0.3384 ± 0.0127
Exchange-hole logit bias	9.2971 ± 0.0139	9.5652 ± 0.0075	-0.2681 ± 0.0114

Longer training and parameter counts. Table 12 extends the gas comparison. At 16,000 updates, the standard readouts give 7.25345 ± 0.03899 for the coordinate backbone, 5.65402 ± 0.02559 for pair matching, and 4.21746 ± 0.02501 for the Pfaffian reference. Their gap recovery is 52.68%. At 20,000 updates, both standard and independently initialized readouts retain the pair-matching

Table 12: Gas comparison at 20,000 updates. Total energies and components use independently initialized 256-draw, 512-walker readouts, in $\hbar^2 k_F^2/m$. Errors are conditional MCSEs. These evaluate one training trajectory per model.

Model	Total energy	Kinetic energy	Interaction energy
Coordinate backbone	5.94447 ± 0.01219	6.17656 ± 0.00325	-0.23209 ± 0.01122
Orbital reference	4.73352 ± 0.01282	5.23691 ± 0.00417	-0.50339 ± 0.01447
Pair matching	4.55813 ± 0.01422	5.14899 ± 0.00475	-0.59087 ± 0.01739

advantage over the orbital reference; the independent difference is -0.17539 ± 0.01915. The simple Slater reference has total energy 4.01632781 under the same Hamiltonian.

For the 20,000-update pair and orbital states, independent total-energy half-chain differences are 0.09 and 0.31 reported MCSEs. The pair state's kinetic half-chain difference is -3.20 MCSEs, and its local-energy standard deviation is 4.758, versus 4.415 for the orbital reference. Thus, the lower total energy does not entail lower local-energy variance. The standard pair readout has a total-energy half-chain difference of -2.23 MCSEs; its mean 4.54382 ± 0.03691 agrees with the independent estimate.

Table 13 gives model sizes. The matching operator is evaluated by contractions without storing its full four-index tensor. All backbone and reader parameters are optimized jointly.

Table 13: Trainable parameter counts for the communication comparisons.

Added reader	Parameters
None	273,665
Pair matching	304,577
Low-energy subspace	298,241
Short-range Jastrow	273,666
Exchange-hole logit bias	273,668

Which part of the matching kernel helps? The direct-matching control retains the pair features, cross-attention reader, parameters, and disjoint-pair support, replacing the operator by

$$(\mathcal{M}_F^{\mathrm{dir}} e)_{ij} = \frac{F_{ij} \sum_{k \neq i,\, l \neq j} F_{kl} e_{kl}}{2(1 + \|F\|_{\mathrm{F}}^2)}.$$

It removes the exchanged assignment while continuing to exclude pairs sharing an endpoint. Both readers improve on their common reference at 4,000 updates (Table 14) and at 10,000 updates (Table 15). At the latter budget, direct minus antisymmetrized matching is -0.00663 ± 0.01726, leaving the exchange-specific contribution unresolved.

A reference-subtracted variant subtracts the numerator of Eq. (3) evaluated with $F_0 = G/s_0$ from the numerator evaluated with F. It uses the common denominator $2(1 + \|F\|_{\mathrm{F}}^2 + \|F_0\|_{\mathrm{F}}^2)$. This removes the closed-shell contribution when the scattering shape vanishes, but also changes normalization. Both subspace combinations ended above pair matching alone at 10,000 updates (Table 15).

A distinct matrix-response reader transports opposite-spin features through $\mathcal{B}_\eta(F) = 2\eta F(F^\top F + \eta^2 I)^{-1}$, with $\eta = 0.5$ and one zero-initialized output map per layer, shared between spin sectors. Its singular-value response is $2\eta s/(s^2 + \eta^2)$. This bounded response emphasizes an intermediate range of pair-matrix singular values; it is not a many-body pairing projector. At 7,000 updates it reached 13.23852 ± 0.07526, compared with 12.82643 ± 0.04091 for the orbital-side reference.

The completed 10,000-update controls are listed in Table 15. Pair matching plus Jastrow reached 8.71952 ± 0.01258, versus 8.73889 ± 0.01169 for pair matching alone. Both subspace combinations ended above pair matching alone, despite their lower intermediate energies. The Fourier pair reader replaces the added body descriptors, matching function, and attention prior with calibrated periodic functions, retaining opposite-spin pair organization, the direct-minus-exchange construction, the cross-attention reader, the trainable reader, and the scattering orbital path. It is a less system-specific physical description rather than a removal of physical priors. Its agreement with the original reader

Table 14: Communication and function-reading comparisons at a common intermediate budget of 4,000 updates. All rows retain the spectral-and-scattering orbital reference. Values use standard 40-draw readouts and conditional MCSEs in $\hbar^2 k_F^2/m$. Completed 10,000-update estimates appear in Table 15.

Added reader	Total energy $\pm$ MCSE
None	21.8246 ± 0.0765
Pair matching	20.6395 ± 0.1043
Low-energy subspace + pair matching	20.3592 ± 0.0674
Low-energy subspace + reference-subtracted matching	20.5331 ± 0.0761
Direct matching on disjoint pairs	20.7623 ± 0.1373
Regularized pair-matrix response	22.5570 ± 0.0821
Physical feature rotations	21.9676 ± 0.0781
Bilinear function reading	24.4646 ± 0.1028
Rotations + bilinear reading	24.6215 ± 0.0828

leaves the incremental value of the scattering-specific body refinement unresolved. Resetting its pair features between layers gave 8.76154 ± 0.01380, a difference of $+0.02599 \pm 0.01860$ from the persistent Fourier pair reader; this reset also deactivates the associated recurrent and collection paths.

Table 15: Completed content, matching, and function-reading controls at 10,000 updates, with independent 256-draw MCSEs. Every row retains the scattering orbital path. Energies use $\hbar^2 k_F^2/m$.

Added reader or control	Total energy $\pm$ MCSE
Pair matching	8.73889 ± 0.01169
Pair reader with Fourier pair description	8.73555 ± 0.01247
Fourier pair description with layerwise reset	8.76154 ± 0.01380
Direct matching on disjoint pairs	8.73226 ± 0.01269
Low-energy subspace + pair matching	8.93830 ± 0.01178
Low-energy subspace + reference-subtracted matching	8.90846 ± 0.01262
Pair matching + Jastrow	8.71952 ± 0.01258
Regularized pair-matrix response	9.60497 ± 0.01349
Physical feature rotations	9.33037 ± 0.01512
Bilinear function reading	10.13844 ± 0.01383
Rotations + bilinear reading	10.15437 ± 0.01451

Table 16: Fourteen-particle gas controls completed at 10,000 updates. Standard and independent columns are separate frozen-state readout protocols; the latter reinitializes only the readout chain. Errors are conditional MCSEs in $\hbar^2 k_F^2/m$.

Model or control	Standard, 40 draws	Independent, 256 draws
Coordinate backbone	10.59967 ± 0.02824	10.56251 ± 0.01925
Body-only	9.72730 ± 0.02723	9.74662 ± 0.01426
Matched healed-scattering orbital	9.27482 ± 0.04915	9.28837 ± 0.01683
Combined orbital-and-pair route	8.71660 ± 0.02695	8.73889 ± 0.01169
No pair history	8.65659 ± 0.03384	8.69240 ± 0.01117
Bare pair kernel	8.74272 ± 0.02762	8.74186 ± 0.01176

The body-only control adds the physical pair resources and cross-attention reader to the coordinate baseline, independently of the orbital-side parent construction. The no-history control retains the healed-scattering orbital route and the pair reader but replaces the incoming pair state by zero at every block; particle features still propagate, whereas the history, matching-action, old-state residual, and associated collection paths are inactive by construction. It is therefore a bundled intervention rather than an isolated memory ablation. The bare-kernel control retains the combined orbital-and-pair routes and replaces only $F = G\exp(D^{\mathrm{c}})/s_0$ by $F_0 = G/s_0$, with the corresponding denominator; its local pair channels and attention prior remain. The latter control retains the physical resources, whereas no-history also reduces the number of active parameters. Each trajectory used the

same 10,000-update protocol and one training seed (30,000 updates across the three controls); the independent column changes only the frozen-state readout chain.

Reading the same function banks in different ways. The bilinear body reader uses the two banks X_0 and X_1 of Eq. (1). In each layer it adds

$$\delta h^{\ell}_{iw} = \sum_{s,c,a} X_{s,ia}(1, a^{\ell}_i P)_c C^{\ell}_{scaw}$$

to the attention output before the residual and MoE, with real, zero-initialized coefficients and the same fixed compression P as the orbital reader. It adds 72,960 parameters. This makes the original physical functions available at every layer without a separate nonlinear resource compression.

The rotation reader instead uses these banks to control attention inputs. Let (u_p, v_p) be eight orthogonal feature planes drawn from the first sixteen nonconstant discrete-cosine columns in the 64-dimensional feature space. The angles are $\theta_i = [X_{0,i}, X_{1,i}]A_\ell$, with A_ℓ initialized at zero. For $a = a^{\ell}_i$, $A_p = a^{\top}u_p$, and $B_p = a^{\top}v_p$, the rotated attention input is

$$\widetilde{a} = a + \sum_{p=1}^{8} [(\cos\theta_p - 1)(A_p u_p + B_p v_p) + \sin\theta_p (A_p v_p - B_p u_p)] .$$

This preserves the feature norm pointwise and adds 1,824 parameters. The combined reader applies both operations from the same unmodified a^{ℓ}_i. The higher intermediate energies of bilinear reading and its combination (Table 14 and Figure 4b) show that functions effective at the orbital output need not help when read throughout the backbone. This compares different readers of the same physical banks rather than a placement ablation.

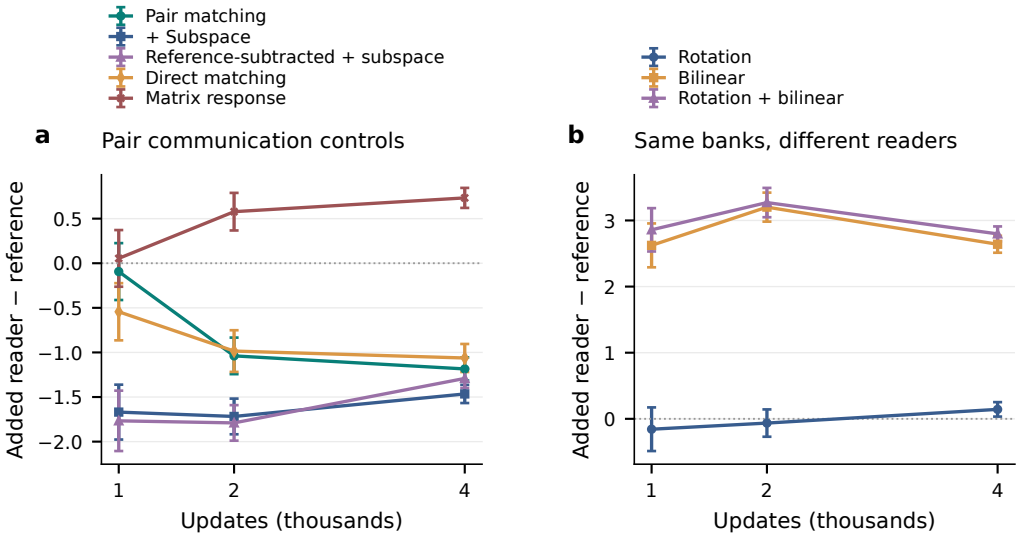


Figure 4: **Different uses of physical information at a shared intermediate budget.** (a) Pair-matching controls and combinations. (b) Rotation and bilinear readers of the same orbital function banks. Curves show all standard frozen-checkpoint differences up to 4,000 updates, relative to the common spectral-and-scattering orbital reference. Negative values favor the added reader. Error bars combine conditional MCSEs in quadrature; all trajectories use one training seed. Energies are in $\hbar^2 k_F^2/m$.

Readout identity and physical diagnostics. Standard 40-draw and independent 256-draw measurements evaluate the same trained checkpoint. For pair matching they give 8.71660 ± 0.02695

and 8.73889 ± 0.01169, respectively; for the single-device orbital-side reference they give 9.27482 ± 0.04915 and 9.28837 ± 0.01683. With standard readouts for all three states, pair communication closes 31.80% of the gap between the coordinate backbone and the pairing reference. The incremental improvement over the orbital reference, 0.54948 ± 0.02049, instead compares the two independent readouts in Table 11.

The independent pair-matching readout has total-energy and kinetic-energy half-chain differences of 2.46 and 2.71 reported MCSEs. Its local-energy standard deviation is 4.106, versus 4.237 for the reference, while its largest absolute local-energy sample is 174.91, versus 162.39. Close-pair event frequencies are 0.416% and 0.402%; the shortest-distance radial bin remains noisy, and the reference readout lacks the matching structure-factor measurements. The energy evidence therefore supports improved variational learning and a kinetic reduction, without establishing a change in pair condensation or uniformly improved sampling stability.

F.6 Additional ways to read and combine physical resources

The four additional gas comparisons retain the same grouped-query attention/MoE backbone, complete coordinate input, and 512-walker stochastic-reconfiguration protocol. We call the full combination a *multi-site reader*. Its providers include a smooth short-range shape, its radial response, an interaction kernel, local pair environments, a closed-shell orbital matrix, and filtered low-energy modes. Learned particle queries read pair features through two physically weighted attention heads in each body block. Separate readers add a closed-shell anchor and low-rank relation-weighted orbital corrections, together with static and environment-dependent log-amplitude corrections. All added contributions are zero at initialization. The backbone and readers are then optimized jointly from this common initial state.

The *restricted-orbital* variant retains these providers and all body, anchor, and amplitude paths. It removes only the radial-response and interaction channels from the added low-rank orbital correction. The shape channel and inherited scattering-shape orbital reader remain. Thus, this comparison tests two selected orbital uses, not the removal of physical orbital information altogether.

The *gated particle-message* reader uses two physical kernels: the healed scattering relation and an opposite-spin cosine kernel at the first reciprocal shell. Each kernel pools partner features before a trainable linear map. A softmax gate, computed from particle features and kernel summaries, weights the resulting messages together with a null branch. These added messages act alongside the backbone's existing computation. They combine resource-specific messages through a learned gate.

Writing K^c_{ij} for the two fixed kernels and $\bar{h}^c_i = \sum_j K^c_{ij} h_j / \sum_j |K^c_{ij}|$, this reader has the explicit form

$$m_i = \sum_{c \in \{0,1,\varnothing\}} \pi_{ic} W_c \bar{h}^c_i, \qquad \bar{h}^{\varnothing}_i = h_i,$$
$$\pi_i = \mathrm{softmax}\big[g_\phi\big(h_i, \bar{h}^0_i, \bar{h}^1_i\big)\big] .$$

with m_i added at the body site. The null branch makes the reader neutral at initialization, while the gate learns how to combine the two physical relations.

The *body-and-amplitude* reader uses nine pair channels, including bounded scattering shape and radial response, interaction and envelope functions, and periodic pair-center harmonics. A bounded projection supplies body residuals, while learned particle features set the coefficients of a log-amplitude correction. This variant disables the inherited scattering-shape orbital correction as well as its own direct orbital tangent. It therefore compares a complete alternative use of the resources, rather than a single-factor orbital ablation.

If $\xi_{ij} \in \mathbb{R}^9$ denotes these channels, its two outputs can be written as

$$m_i = B_\ell\big(\mathrm{Pool}_{j \neq i}[\xi_{ij}, h^\ell_j]\big) ,$$
$$\Delta J_\ell(R) = \sum_i c_\ell(h^\ell_i)^\top \mathrm{Pool}_{j \neq i}[\xi_{ij}] ,$$

where B_ℓ is the bounded body projection and $c_\ell(h^\ell_i)$ is a learned feature-dependent coefficient. The first term enters p_{body}; ΔJ_ℓ is the corresponding increment to the symmetric log-amplitude factor at p_{amp}. The inherited healed-scattering orbital increment is disabled in this comparison.

All four trajectories use seed 20260731 and one training run. The standard readouts contain 40 draws for each of 512 walkers at six checkpoints. Seven supplementary early checkpoints per trajectory use eight draws. Each final checkpoint also has an independently initialized 256-draw readout. Table 17 reports the longer estimates. Table 18 retains both endpoint evaluations. Independent readout initialization does not provide an independent training replicate.

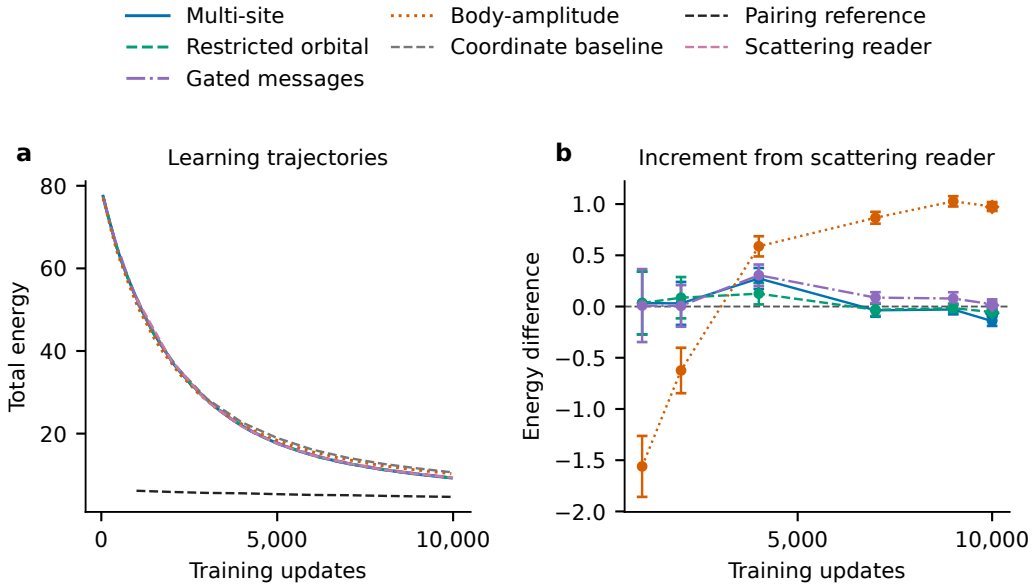


Figure 5: **Learning trajectories for alternative uses of physical resources.** (a) Curves show 100-step training means for all four readers. Dashed reference curves connect frozen-state readouts. (b) Energy differences from the scattering-shape reader at common checkpoints. Circles use 40-draw readouts. Diamonds use independent 256-draw endpoint readouts minus the standard 40-draw readout of the scattering-shape reader. Bars are combined conditional MCSEs from one training seed per trajectory. All energies use $\hbar^2 k_F^2/m$.

Table 17: Alternative physical readers in the Fermi gas at 10,000 updates. Energies are total values in $\hbar^2 k_F^2/m$. The terms K and V denote the kinetic and interaction contributions. Uncertainties are conditional MCSEs from the independent 256-draw endpoint readouts.

Reader	Total energy	K	V
Multi-site reader	9.2486 ± 0.0134	9.5438 ± 0.0098	-0.2952 ± 0.0132
Restricted orbital reader	9.2623 ± 0.0153	9.5858 ± 0.0077	-0.3235 ± 0.0171
Gated particle messages	9.3298 ± 0.0125	9.5946 ± 0.0078	-0.2648 ± 0.0119
Body-and-amplitude reader	10.2897 ± 0.0135	10.4777 ± 0.0075	-0.1881 ± 0.0113

Table 18: Two frozen-state evaluations of each additional reader at 10,000 updates. Values are total energy $\pm$ conditional MCSE, in $\hbar^2 k_F^2/m$. The two estimates evaluate the same trained parameters.

Reader	Standard, 40 draws	Independent, 256 draws
Multi-site reader	9.17496 ± 0.03358	9.24857 ± 0.01344
Restricted orbital reader	9.26321 ± 0.07040	9.26232 ± 0.01534
Gated particle messages	9.33976 ± 0.03398	9.32978 ± 0.01255
Body-and-amplitude reader	10.29384 ± 0.02903	10.28966 ± 0.01346

The multi-site and restricted-orbital endpoints differ by -0.01375 ± 0.02039, which does not resolve a benefit from the two added orbital channels. Relative to the standard readout of the scattering-shape reader, their independent-readout kinetic differences are -0.01727 ± 0.03535 and $+0.02477 \pm$

0.03481. The corresponding interaction differences are -0.05152 ± 0.02984 and -0.07981 ± 0.03173. These components describe the observed tradeoff without establishing separate kinetic improvements. The body-and-amplitude reader had an early advantage at 2,000 updates, -0.62424 ± 0.22170, but ended higher by 0.97230 ± 0.03549.

F.7 Which added paths affect the trained wavefunction?

We evaluate branch responses on 512 configurations from the first independent endpoint draw, holding all learned parameters fixed. For a branch g, define

$$\delta J_g(R) = \log|\Psi(R)| - \log|\Psi_{-g}(R)|, \qquad s_g = \sqrt{\langle(\delta J_g - \langle \delta J_g \rangle)^2\rangle}.$$

Centering removes a configuration-independent normalization shift. In the multi-site model, all added branches give $s_g = 0.05419$, compared with 0.05418 for the static amplitude correction, 0.002004 for the anchor, 0.0001559 for the body reader, 0.00001769 for the environment-dependent amplitude, and 0.000002622 for the added orbital corrections. Evaluation on a common configuration bank gives the same qualitative ordering. These responses do not sum into energy contributions and do not measure the performance of a retrained or resampled ablated state. They show that the added paths have very different effects on the final log amplitude despite being available throughout training.

Table 19: Gas endpoint decomposition at 10,000 updates. Entries give means and conditional MCSEs from raw samples at frozen states.

Variant	Total energy	Kinetic energy	Interaction energy
Constant relation	9.57159 ± 0.02775	9.76529 ± 0.02029	-0.19369 ± 0.01910
Scattering shape	9.31737 ± 0.03285	9.56104 ± 0.03396	-0.24367 ± 0.02675
Body + shape	9.31580 ± 0.03119	9.58312 ± 0.01890	-0.26732 ± 0.02842
Constant body + shape	9.34165 ± 0.05880	9.54470 ± 0.03465	-0.20305 ± 0.03388

The body-content difference was -0.02584 ± 0.06656, leaving its benefit unresolved. A separate relation-weighted message outside attention reached 10.23992 ± 0.03834, compared with 10.27548 ± 0.03152 for its constant-relation control. Neither comparison resolved a benefit from the physical relation at this budget. These tests leave open the performance of other body readers and the cause of the remaining energy gap, including any role of FiLM saturation. Figure 6 shows the kinetic and interaction changes of these readers.

F.8 Gas readout and uncertainty

Each frozen-state readout continues a cloned checkpoint bank for 100 burn-in moves, followed by 40 draws separated by 20 moves. Combining both ranks gives 40 draws for each of 512 walkers. The MCSE is the maximum of the chain-mean error and temporal errors from a blocking ladder with 40, 20, 10, and 5 blocks. At each level, the error is the sample standard deviation of block means divided by the square root of the block count. An unpaired final block is omitted only when constructing the next ladder level. The reported mean energy retains every draw from the readout.

Readouts use raw complex local energies and report their real-part mean as total energy. The gas SR protocol retains raw optimizer records without substituting a separately clipped energy series for these measurements. Contrasts combine marginal MCSEs in quadrature, with samplewise pairing unverified. Diagnostics include the imaginary mean, raw variance and tails, and temporal half-drift. All reported uncertainties are conditional on the saved state and sampling history; they are not uncertainty estimates across independent training runs.

G Historical density context and extended moiré readouts

The historical environment comparison uses the Hamiltonian in Appendix C at $\lambda = 0$ and 2. Both settings use width 96 and lie outside the fixed width-64 backbone comparisons in the main text. Their baseline and side models share a complete-coordinate backbone, and the resources use three

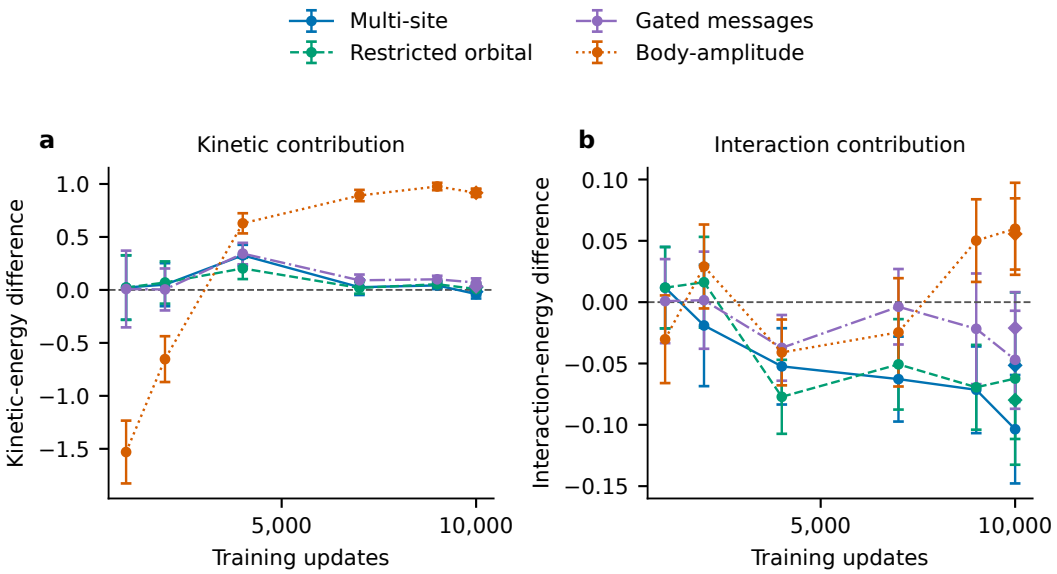


Figure 6: **Kinetic and interaction changes under alternative physical readers.** Contributions are shown relative to the scattering-shape reader at common checkpoints. Points and error bars use the same conditional MCSE convention as Figure 5. Component differences describe the trained states and are not causal branch attributions.

primitive-cell vectors $(\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3)$. These results do not establish an environment-dependent gain for the width-64 model. The complex-valued channels and their common normalization are

$$f_a^{\mathrm{complex}} = \rho_{\boldsymbol{g}_a}, \qquad f_a^{\mathrm{power}} = (|\rho_{\boldsymbol{g}_a}|^2 - 1)\sqrt{\frac{N}{2(N-1)}}(1+\mathrm{i}),$$

$$f^{\mathrm{norm}} = \mathrm{pack}_{\mathbb{R}}\left[\frac{f}{\sqrt{\rho_{\mathrm{ref}}^2 + \frac{1}{3}\sum_a |f_a|^2}}\right],$$

with the fixed ρ_{ref} from Appendix B. The normalized resource f^{norm} occupies six real channels in either case, although power contains three independent quantities and complex density contains six. Their second moments match under the independent uniform-particle reference used to set the scale. Their functional content and spatial derivatives differ under the common normalization. The reader dimensions are $6 \to 32 \to 8$, and both models have 529,313 trainable parameters. This interface uses primitive-cell modes, distinct from the eight simulation-cell modes in Appendix B. Energies use the 256-draw frozen-state readout described in Appendix D.

Table 20: Historical width-96 primitive-cell density comparison: paired side-minus-baseline total-energy differences in E_{h}^*. Every row uses the 12,000-update checkpoint and the 256-draw frozen-state readout.

Resource	$\lambda = 0$	$\lambda = 2$
Density power	$+0.07622 \pm 0.00938$	-0.21001 ± 0.01785
Complex density	$+0.01071 \pm 0.00811$	-0.09789 ± 0.01931

In this wider model, density power lowers the energy in the stronger potential and raises it without the external potential (Figure 7). This observation is restricted to the recorded architecture and budget.

In the fixed width-64 low-wavevector density setting, extended readouts compare the side model at 10,000 updates with the baseline at 12,000 updates. Their energy difference is $-0.02116 \pm$

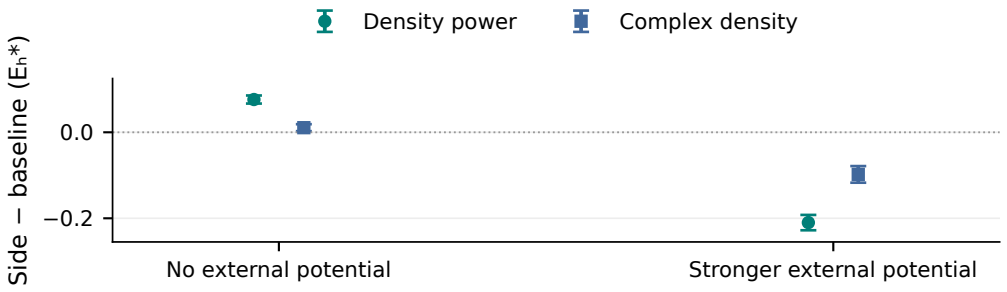


Figure 7: **Historical environment comparison with a width-96 backbone.** Side-minus-baseline total energy at 12,000 updates for potential multipliers $\lambda = 0$ and 2. Bars show paired conditional MCSEs from 256 draws of 64 continued chains per frozen state. Both resource forms use the same primitive-cell wavevectors and six-channel interface. This comparison is separate from the fixed width-64 backbone and simulation-cell density resource in the main text.

$0.02225\,E_{\mathrm{h}}^{*}$ at these different training budgets. In the historical width-96 model at 30,000 updates, side minus baseline is $+0.04588 \pm 0.01117\,E_{\mathrm{h}}^{*}$. The extended readouts refine the uncertainties at these checkpoints. The earlier 10,000- and 20,000-update results remain separate records for their respective training budgets.

A historical width-96 four-way moiré comparison combines local external-potential and force features with collective density-interference features. At 12,000 updates, baseline, local, collective, and joint energies were -15.90753 ± 0.03168, -15.36833 ± 0.05238, -15.67353 ± 0.04244, and $-15.73189 \pm 0.05836\,E_{\mathrm{h}}^{*}$, respectively. The interaction contrast was $E_{AB} - E_A - E_B + E_0 = -0.59757 \pm 0.09462\,E_{\mathrm{h}}^{*}$, while the joint model remained above baseline by $0.17564 \pm 0.06640\,E_{\mathrm{h}}^{*}$. The negative interaction contrast therefore did not translate into an improvement over the baseline.