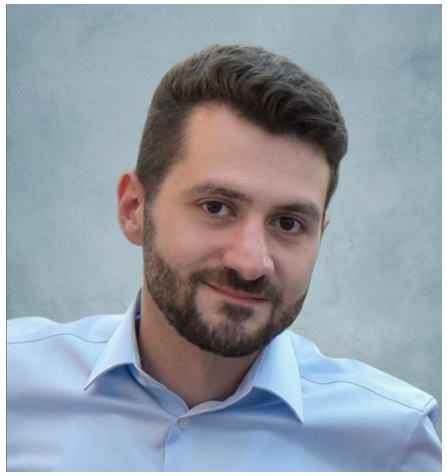
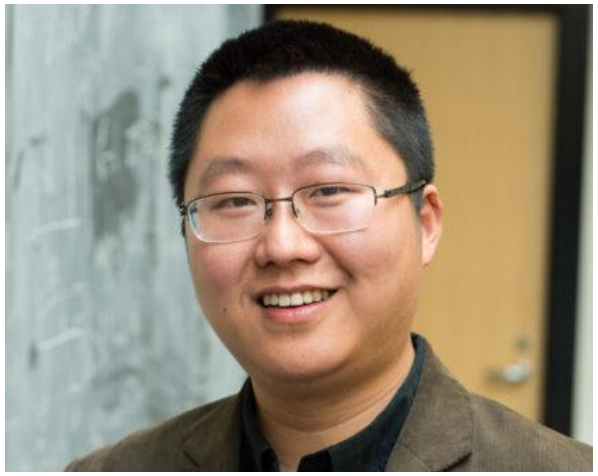
Is attention all you need to solve the correlated electron problem?



Khachatur Nazaryan

Timothy Zaklama



Liang Fu

Max Geier

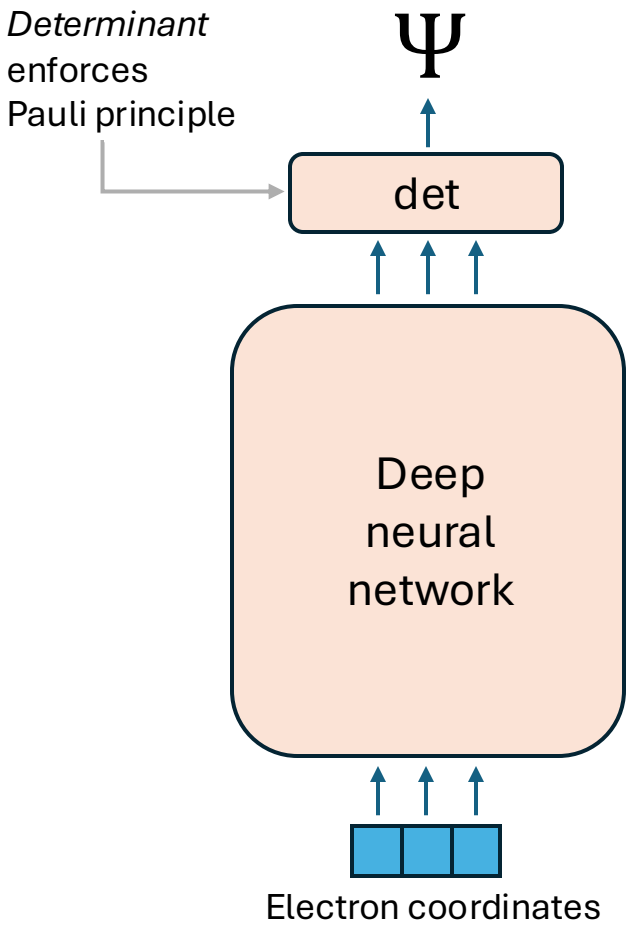
arXiv:2502.05383

subMIT, 07/01/2025

Philosophy of neural network wavefunctions

Many-parameter ansatz for **unbiased** approximation of *any* wavefunction,
but network structure should **reflect basic structure** of ground states for efficiency and accuracy.

➔ **Generative AI model:** The networks learns to represent the electron correlations.
➔ **Physics-informed**: Pauli principle enforced.



Minimal human bias

Electron correlations from generalized orbitals

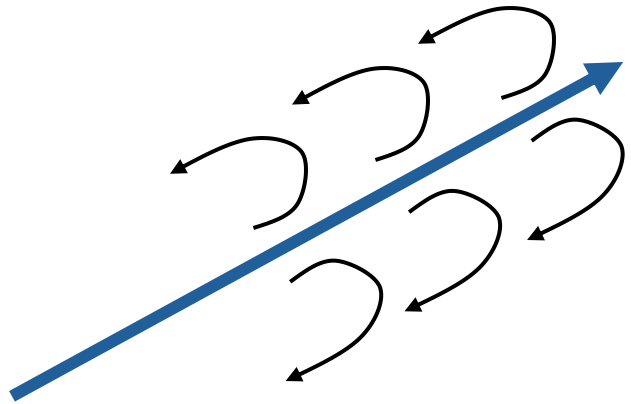
Describe how the state of each electron is affected by the state of all other electrons

$$\phi_j(\boldsymbol{r}_i) \rightarrow \phi_j\left(\boldsymbol{r}_i; \{\boldsymbol{r}_{/i}\}\right)$$

Slater determinant of *correlated orbitals* describes many-body wavefunction

$$\Psi(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N) = \det_{i,j} \phi_j\left(\boldsymbol{r}_i; \{\boldsymbol{r}_{/i}\}\right)$$

backflow

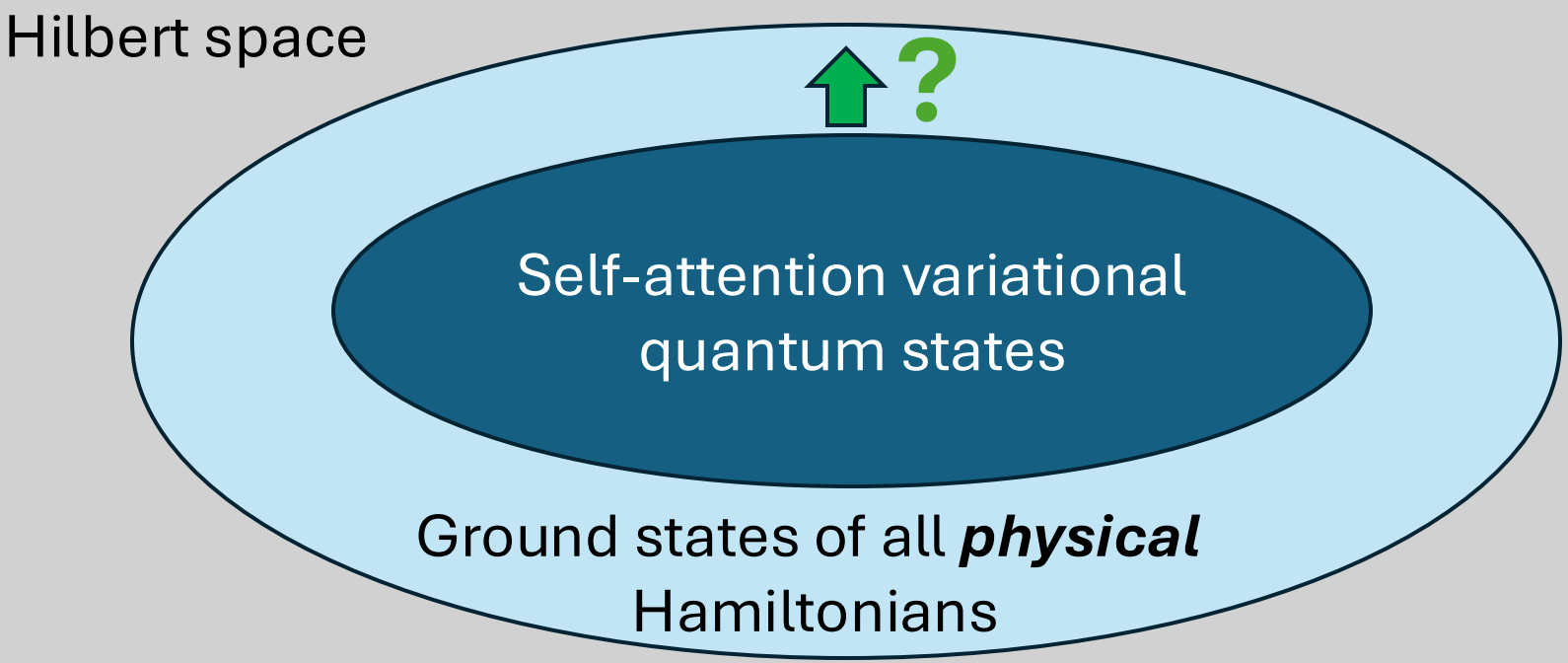


A particle moving in an electron liquid distorts the motion of all other particles along its way.

Feynman and Cohen (1956)

Is attention all you need to solve the correlated electron problem?

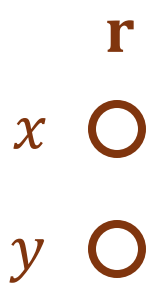
Employ **attention to learn correlations** between electrons.



How to construct a neural network variational wavefunction

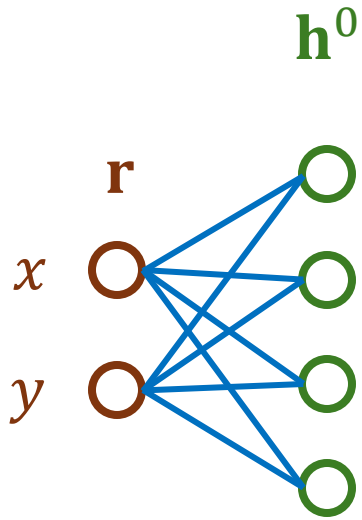
Representing a single-particle orbital $\phi: \mathbb{R}^2 \to \mathbb{C}$

Representing a single-particle orbital $\phi: \mathbb{R}^2 \to \mathbb{C}$



feature

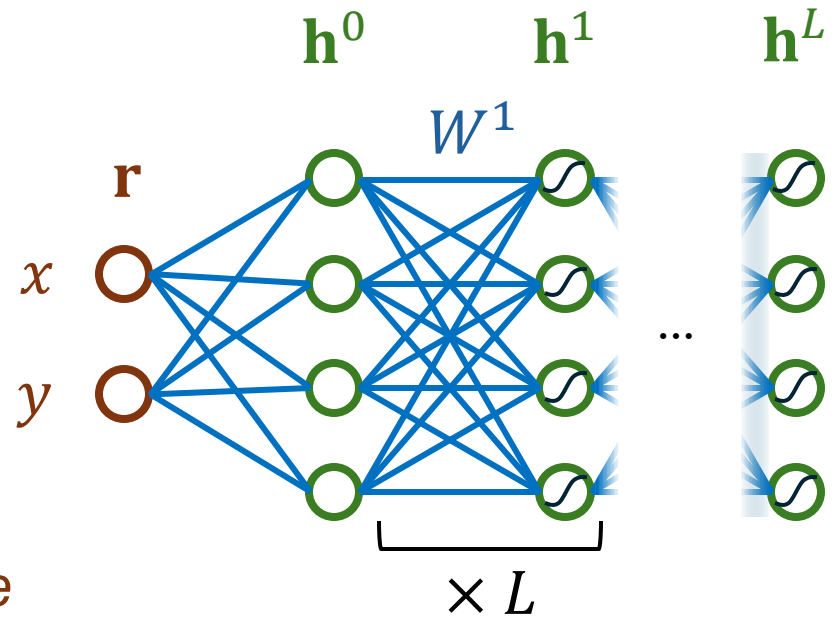
Representing a single-particle orbital $\phi: \mathbb{R}^2 \to \mathbb{C}$



feature

embed: $\mathbf{h}^0 = W^0\mathbf{r}$

Representing a single-particle orbital $\phi: \mathbb{R}^2 \to \mathbb{C}$

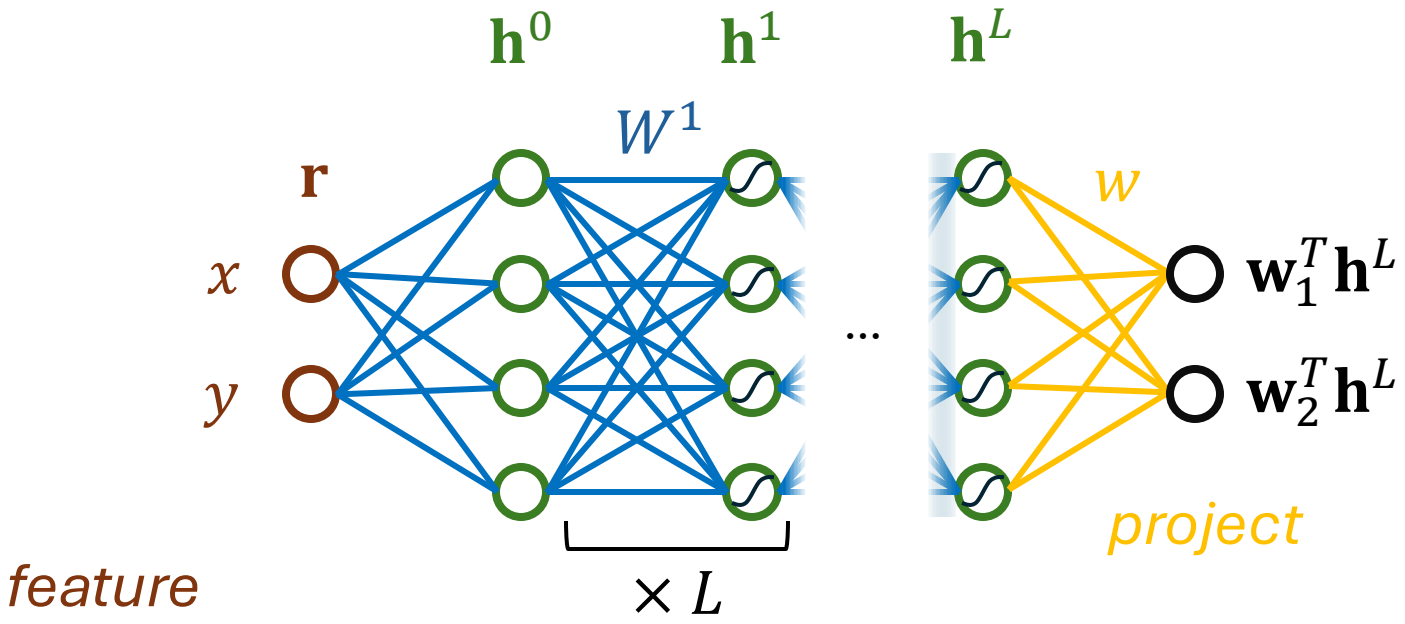


feature

embed: $\mathbf{h}^0 = W^0\mathbf{r}$

transform: $\mathbf{h}^{l+1} = \mathbf{h}^l + \tanh(W^l\mathbf{h}^l + \mathbf{b}^l)$

Representing a single-particle orbital $\phi: \mathbb{R}^2 \rightarrow \mathbb{C}$



$$\phi(\boldsymbol{r}) = \mathbf{w}_1^T\mathbf{h}^L + i\mathbf{w}_2^T\mathbf{h}^L$$

embed: $\mathbf{h}^0 = W^0\mathbf{r}$

transform: $\mathbf{h}^{l+1} = \mathbf{h}^l + \tanh(W^l\mathbf{h}^l + \mathbf{b}^l)$

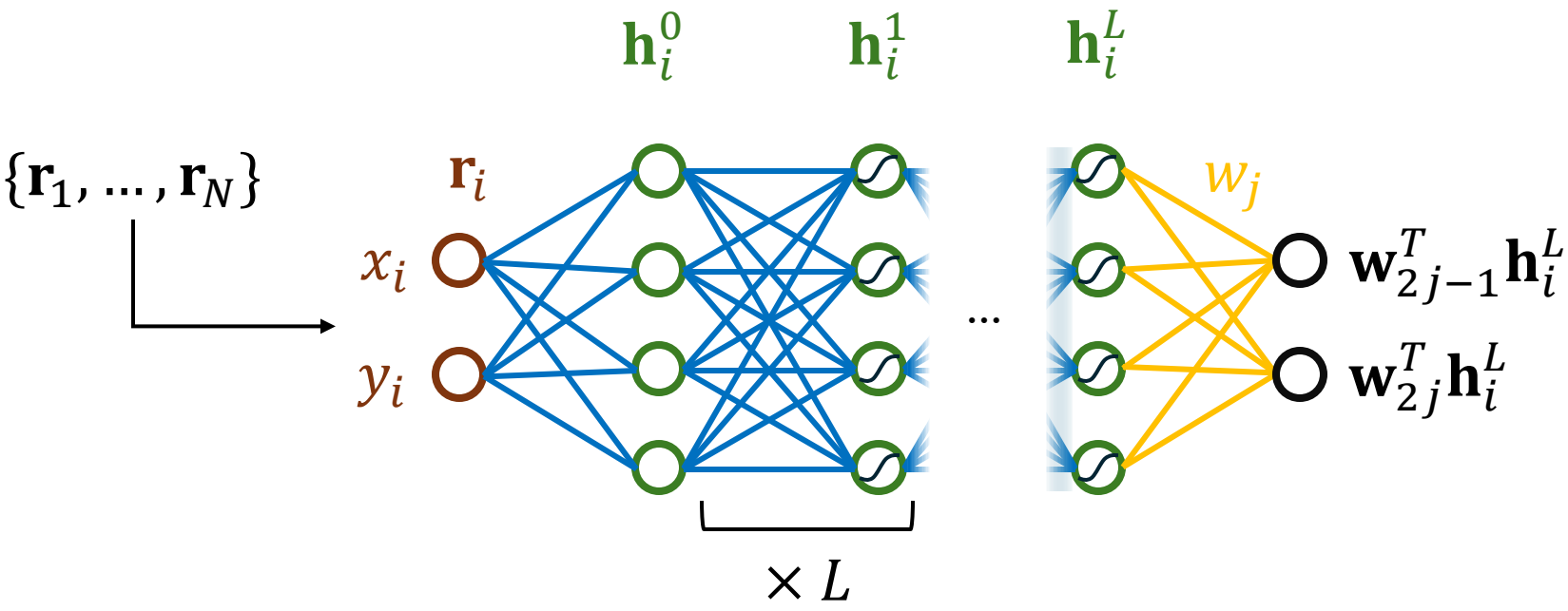
Universal approximation theorem

The family of neural networks lies dense in the function space.

Hornik, Stinchcombe, and White, *Neural networks* **2** (5) 359-366 (1989)

Neural network Hartree-Fock

Neural network Hartree-Fock



$$\phi_j(\boldsymbol{r}_i) = \mathbf{w}_{2j-1}^T \mathbf{h}_i^L + i\mathbf{w}_{2j}^T \mathbf{h}_i^L$$

Slater determinant

$$\Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \det_{ij} \phi_j(\boldsymbol{r}_i)$$

Universal approximation theorem

Because single-particle orbitals are densely approximated, so are Slater determinants.

Capturing correlations

$$\Psi(\mathbf{r}_1, \ldots, \mathbf{r}_N) = \sum_k \omega_k \det\left(\Phi_j^k\left(\mathbf{r}_i; \{\mathbf{r}_{/i}\}\right)\right)$$

Slater determinants of generalized orbitals $\Phi_j(\mathbf{r}_i; \{\mathbf{r}_{/i}\})$
with **permutation equivariant** dependence on remaining electrons
$\Phi_j(\mathbf{r}_i; \{\ldots, \boldsymbol{r}_l, .., \boldsymbol{r}_k, \ldots\} = \Phi_j(\mathbf{r}_i; \{\ldots, \boldsymbol{r}_k, \ldots, \boldsymbol{r}_l, \ldots\})$

Any antisymmetric function can be written as a Slater determinant
of generalized orbitals*

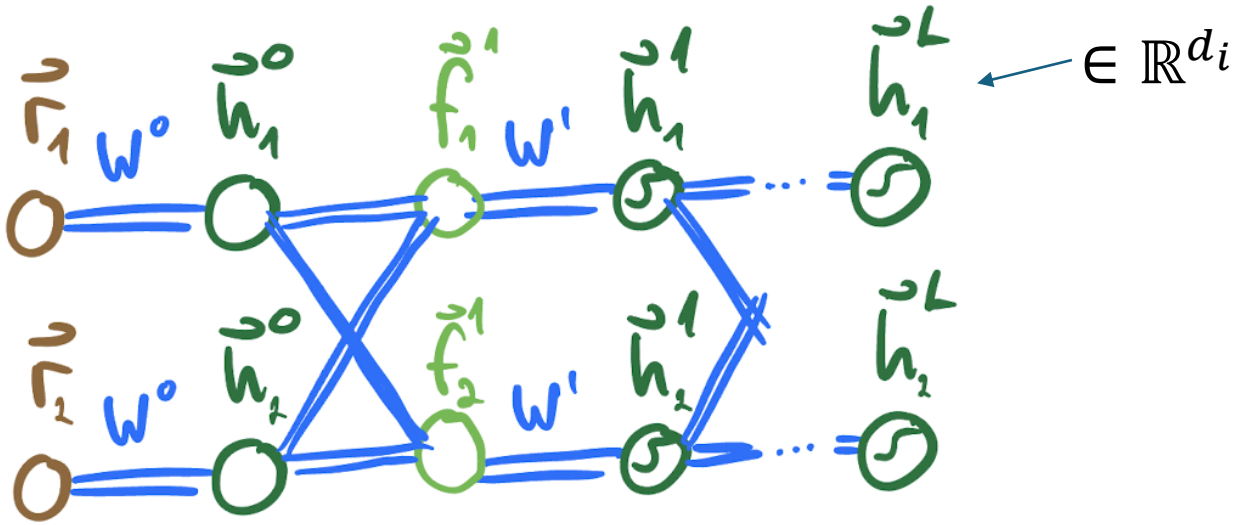
Pfau et al., Phys Rev. Research 2, 033429 (2020)

Permutation equivariant functions

Extend on Hartree-Fock:

Which functions $\mathbf{f}: \mathbb{R}^{d_i \times N} \to \mathbb{R}^{d_i}$ are permutation equivariant in $N-1$ arguments $\{\boldsymbol{h}_{/i}\}$, nonlinear, and depend explicitly on $\boldsymbol{h}_i$?

$$\mathbf{f}(\boldsymbol{h}_i; \dots, \boldsymbol{h}_l, \dots, \boldsymbol{h}_k, \dots, \boldsymbol{h}_N) = \mathbf{f}(\boldsymbol{h}_i; \dots, \boldsymbol{h}_k, \dots, \boldsymbol{h}_l, \dots, \boldsymbol{h}_N)$$

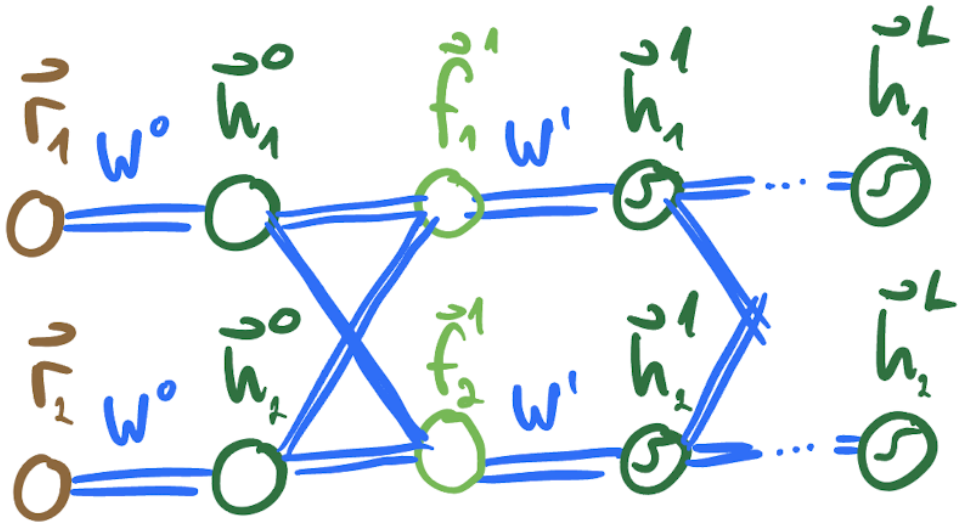


Permutation equivariant functions

Extend on Hartree-Fock:

Which functions $\mathbf{f}: \mathbb{R}^{d_i \times N} \to \mathbb{R}^{d_i}$ are permutation equivariant in $N-1$ arguments $\{\boldsymbol{h}_{/i}\}$, nonlinear, and depend explicitly on $\boldsymbol{h}_i$?

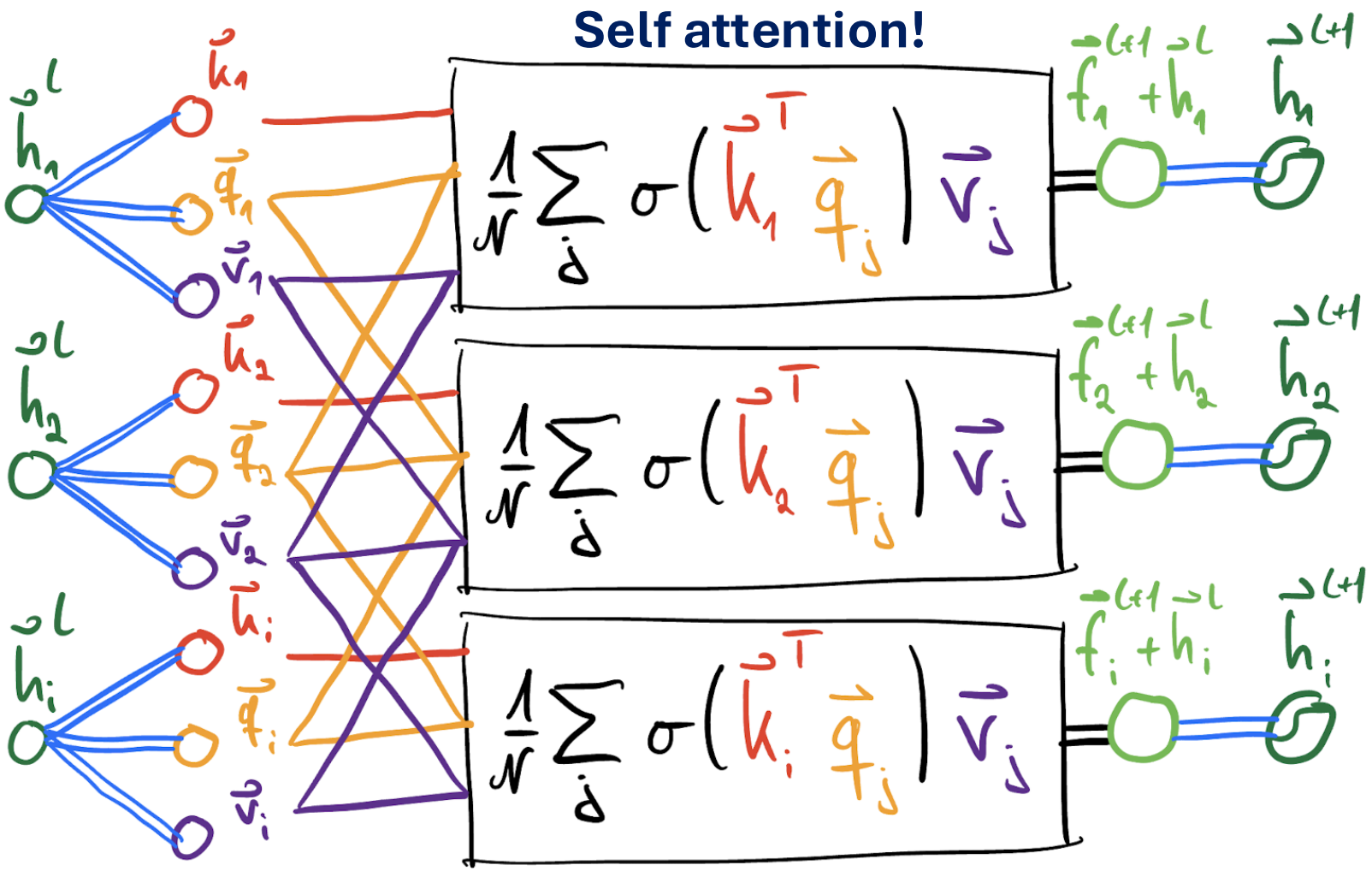
$$\mathbf{f}(\boldsymbol{h}_i; \dots, \boldsymbol{h}_l, \dots, \boldsymbol{h}_k, \dots, \boldsymbol{h}_N) = \mathbf{f}(\boldsymbol{h}_i; \dots, \boldsymbol{h}_k, \dots, \boldsymbol{h}_l, \dots, \boldsymbol{h}_N)$$



$$\mathbf{f}_i(\boldsymbol{h}_i; \{\boldsymbol{h}_{/i}\}) = \sum_j \sigma\left(\boldsymbol{h}_i^T W_{\mathrm{k}}^T W_{\mathrm{q}} \boldsymbol{h}_j\right) W_{\mathrm{v}} \boldsymbol{h}_j$$

σ... non-linear function

Capturing correlations with self-attention

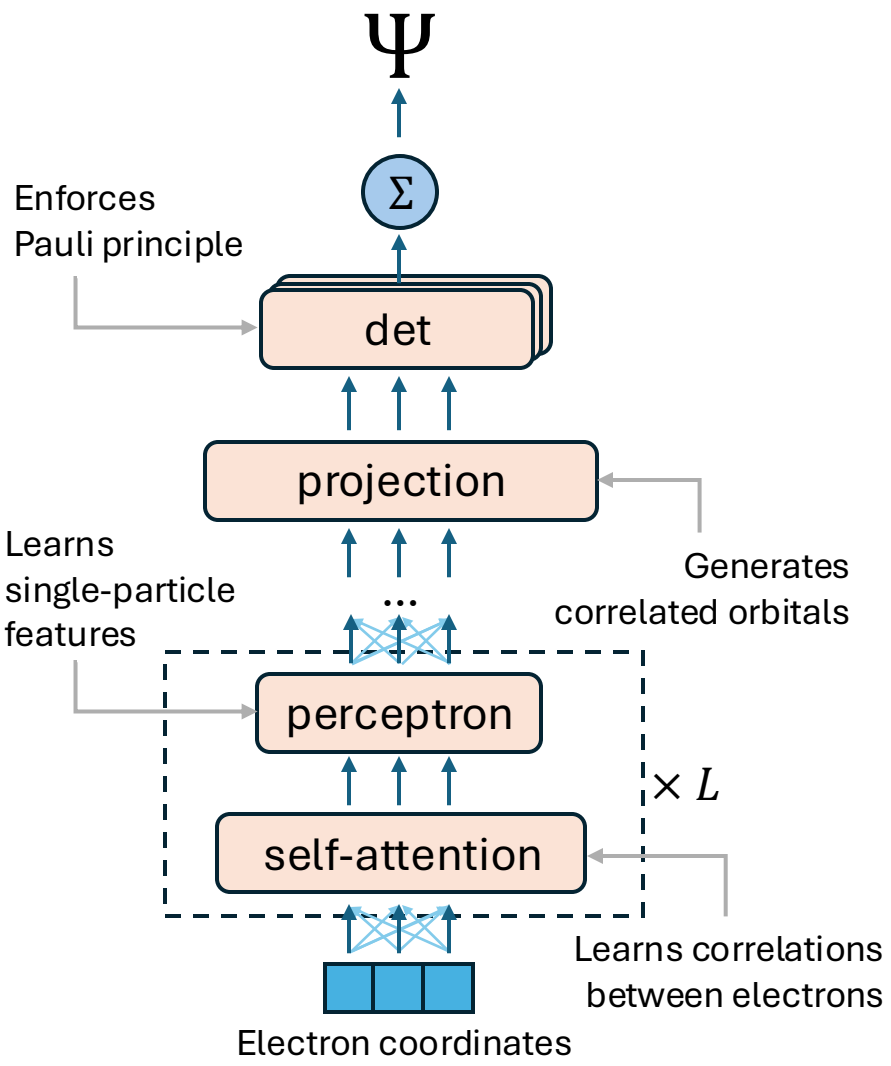


$$\vec{k}_i = W_k^l \vec{h}_i^l$$
$$\vec{q}_i^l = W_q^l \vec{h}_i^l$$
$$\vec{v}_i^l = W_v^l \vec{h}_i^l$$

$$\sigma(x) = \exp(x)$$

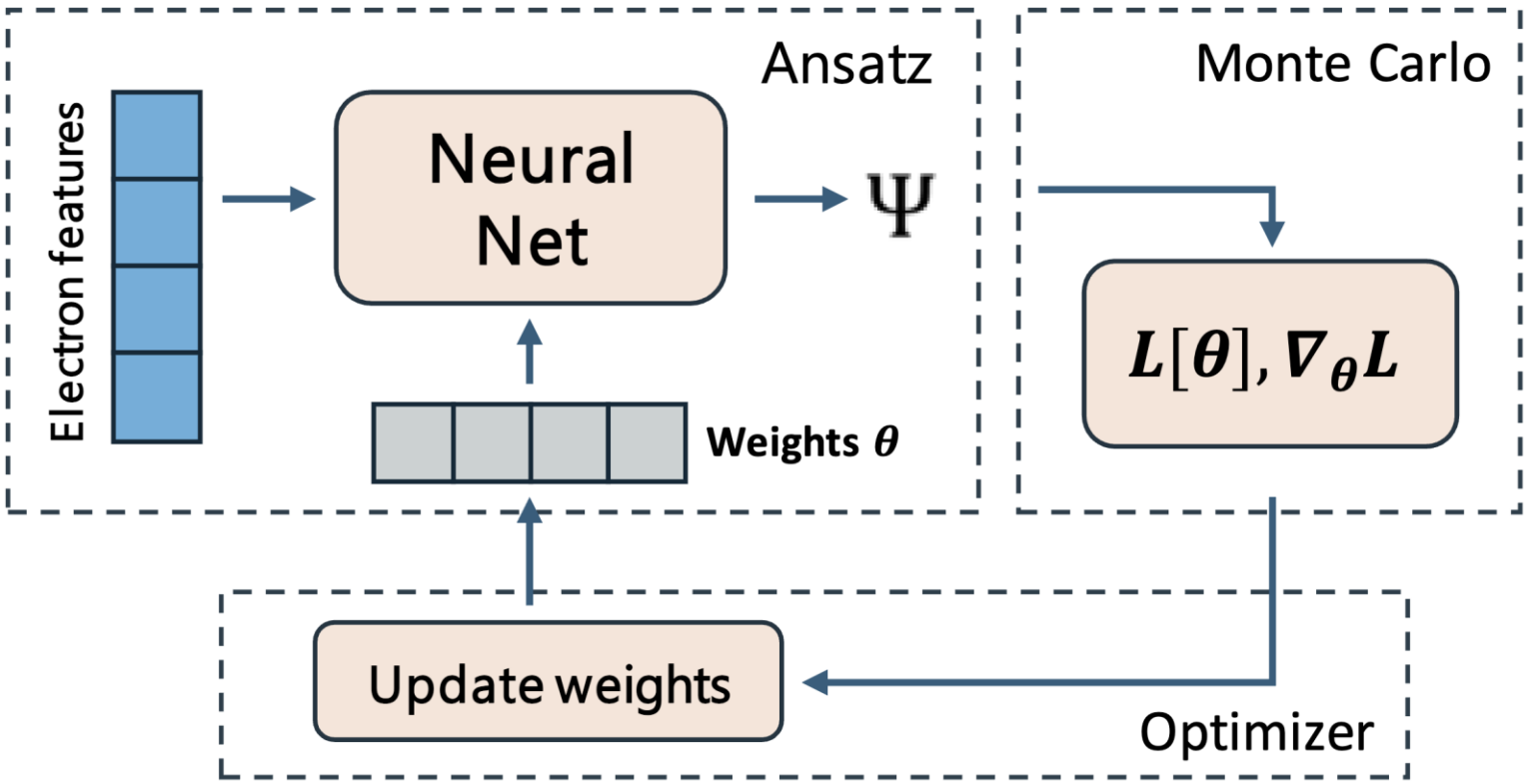
von Glehn et al., arXiv:2211.13672 (2022)

Self-attention wavefunction ansatz



Optimizing the variational wavefunction with variational Monte Carlo

Variational Monte Carlo



How to sample an unnormalized probability density $P(\mathbf{R})$

Metropolis algorithm.

(i) Start walker at $\mathbf{R}$.

(ii) Propose move to $\mathbf{R}'$ with probability $T(\mathbf{R} \leftarrow \mathbf{R}')$.

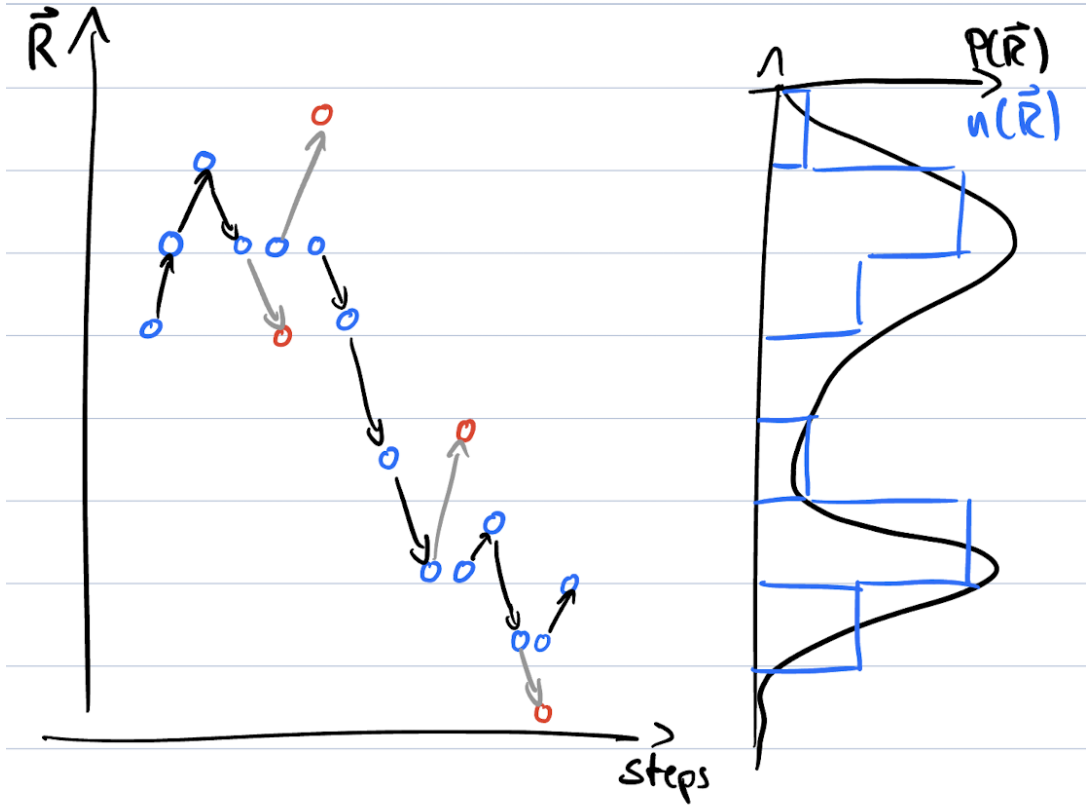
(iii) Accept with probability

$$A(\mathbf{R} \leftarrow \mathbf{R}') = \mathrm{Min}\left(1, \frac{T(\mathbf{R} \leftarrow \mathbf{R}')P(\mathbf{R}')}{T(\mathbf{R}' \leftarrow \mathbf{R})P(\mathbf{R})}\right)$$

Record $\mathbf{R}'$ if accepted, else record $\mathbf{R}$.

(iv) Repeat at (ii).

=> Walker density $n(\mathbf{R})$ proportional to $P(\mathbf{R})$.



Estimating energy

For local Hamiltonian $\hat{H}_{\mathbf{R}\mathbf{R}'} = \hat{H}_{\mathbf{R}}\delta_{\mathbf{R}\mathbf{R}'}$

$$\langle \hat{H} \rangle = \frac{1}{\int d\mathbf{R}|\Psi(\mathbf{R})|^2} \int d\mathbf{R}\Psi^*(\mathbf{R})\hat{H}_{\mathbf{R}}\Psi(\mathbf{R})$$

$$= \frac{1}{\int d\mathbf{R}|\Psi(\mathbf{R})|^2} \int d\mathbf{R}|\Psi(\mathbf{R})|^2 \left[\Psi^{-1}(\mathbf{R})\hat{H}_{\mathbf{R}}\Psi(\mathbf{R})\right]$$

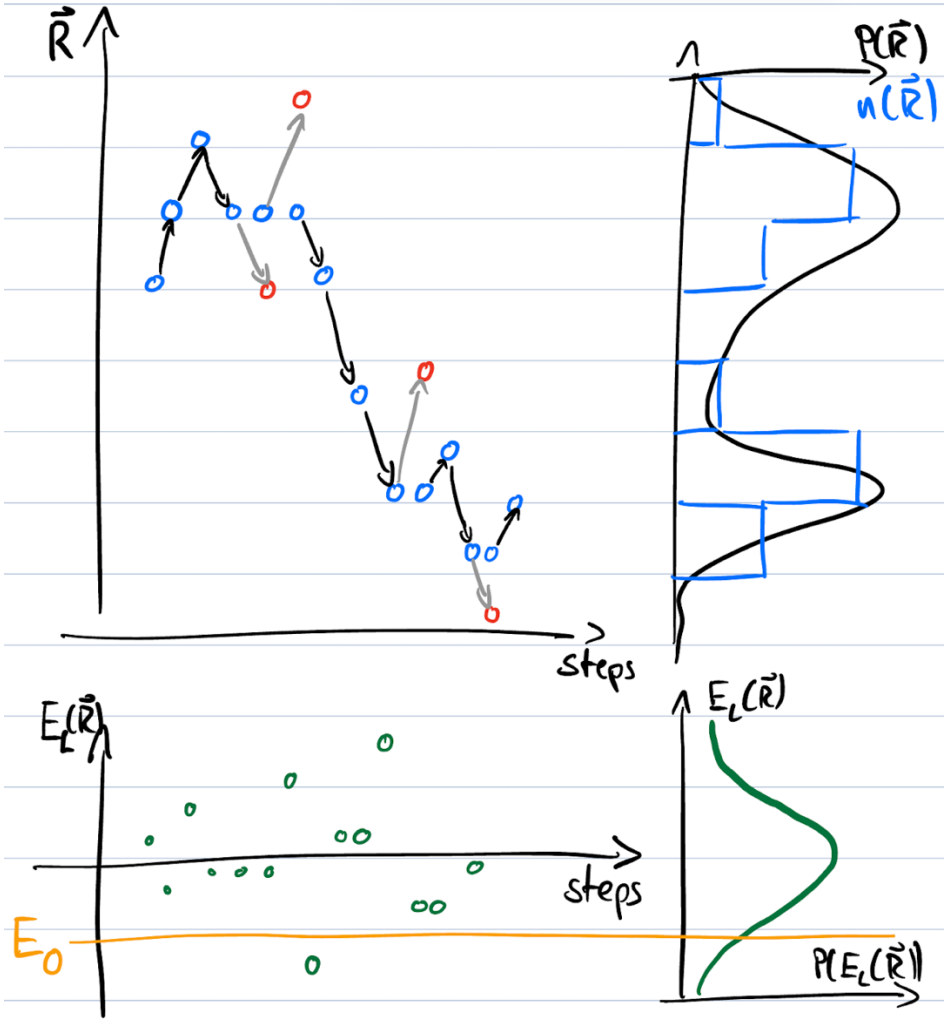
$$\langle \hat{H} \rangle \approx \frac{1}{M}\sum_{\mathbf{R}} E_L(\mathbf{R}) \quad \text{with} \quad E_L(\mathbf{R}) = \Psi^{-1}(\mathbf{R})\hat{H}_{\mathbf{R}}\Psi(\mathbf{R})$$

Properties:

(i) $P(E_L(\mathbf{R}))$ approaches normal distribution

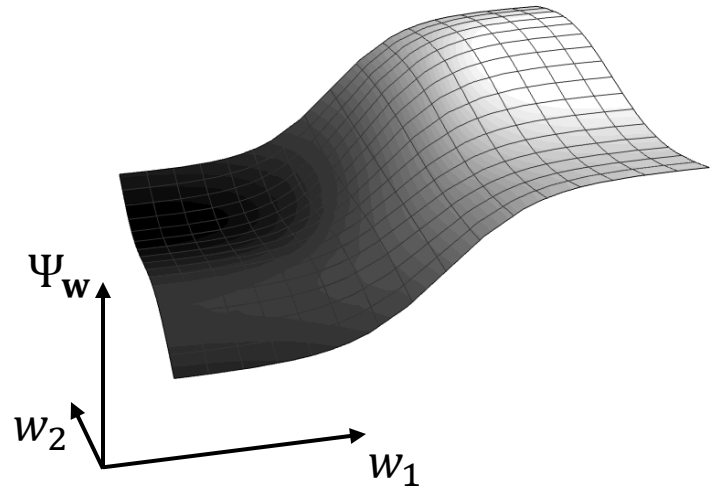
(ii) In eigenstate $\hat{H}\Psi = E_0\Psi$: $E_L(\mathbf{R}) = E_0$, $\mathrm{var}(E_L) = 0$

(iii) $\lim_{M\to\infty} \frac{1}{M}\sum_{\mathbf{R}} E_L(\mathbf{R})$ larger than ground state energy



Natural gradient descent

The **steepest descent direction** $\mathrm{d}\mathbf{w}$ of cost function $\mathcal{L}(\mathbf{w})$ minimizes $\mathcal{L}\,(\mathbf{w}+\mathrm{d}\mathbf{w})$ for fixed infinitesimal distance $\|d\mathbf{w}\|^2 = \epsilon$ where $\|\ldots\|$ is a distance measure of distributions $\Psi_{\mathbf{w}}$.



Distance measure on Hilbert space:

$$\|d\mathbf{w}\|^2 = 1 - |\langle\Psi_{\mathbf{w}+d\mathbf{w}}|\Psi_{\mathbf{w}}\rangle|^2 = \sum_{ij} g_{ij}(\mathbf{w})dw_i dw_j + \mathcal{O}(d\mathbf{w}^4)$$

Steepest descent direction:

$$d\mathbf{w} = \frac{1}{2\lambda}\, g^{-1}(\mathbf{w})\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w})$$

Equivalent to *imaginary time-evolution*.

In practice:

(i) Approximate curvature $g^{-1}(\mathbf{w})$

(ii) Coarse-grain curvature

$$g(\mathbf{w}) \to g(\mathbf{w}) + \delta\mathbb{I}$$

(iii) Use only $|\Psi_{\mathbf{w}}|$ for curvature estimate

(iv) *Stochastic* gradient estimates

Results

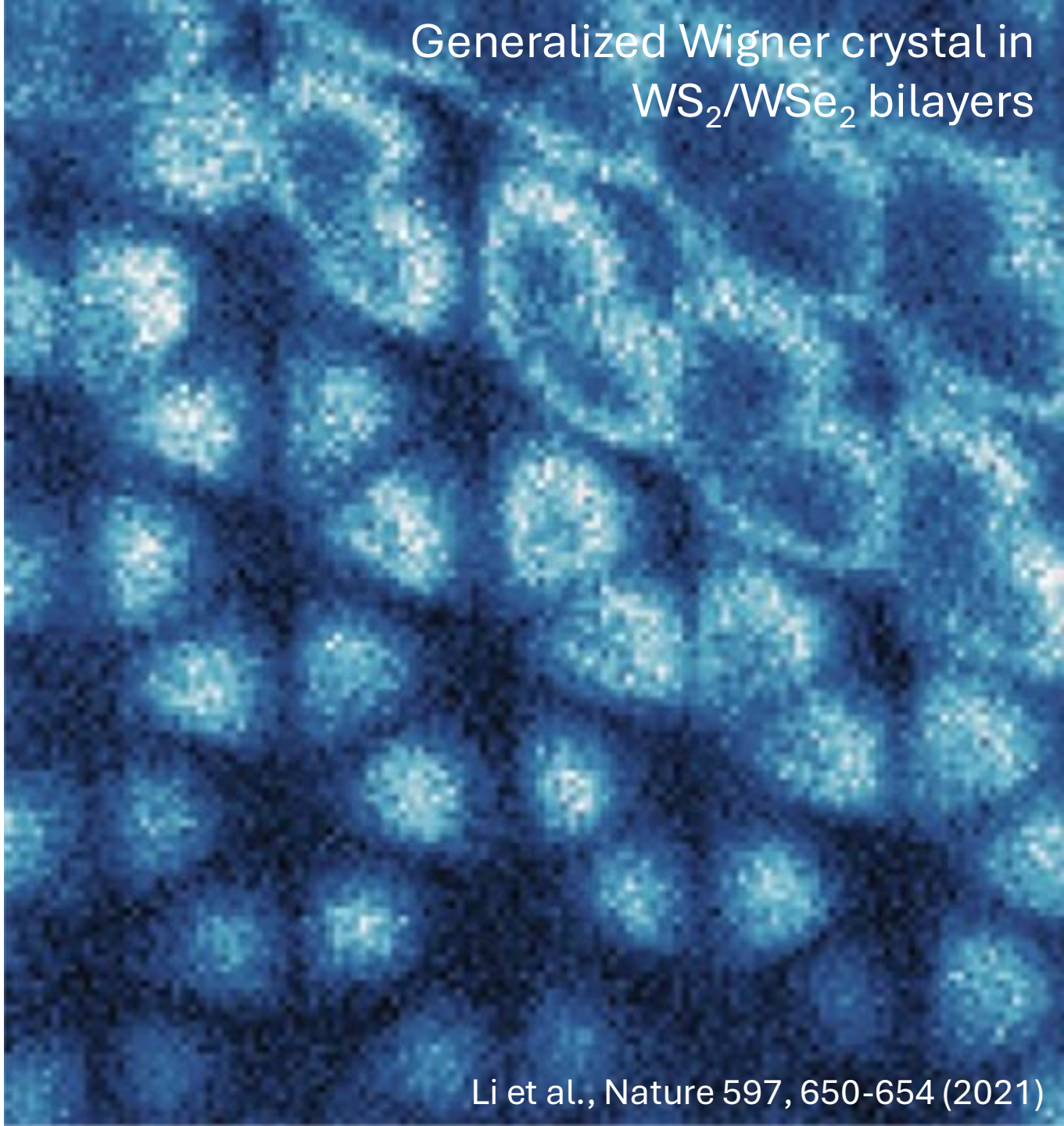
Two-dimensional Coulomb gas with periodic potential

$$\hat{H} = -\frac{1}{2}\sum_i \nabla_{\mathbf{r}_i}^2 + \sum_{i<j} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} + V_0 \sum_j \cos \mathbf{G}_j \mathbf{r}_i$$

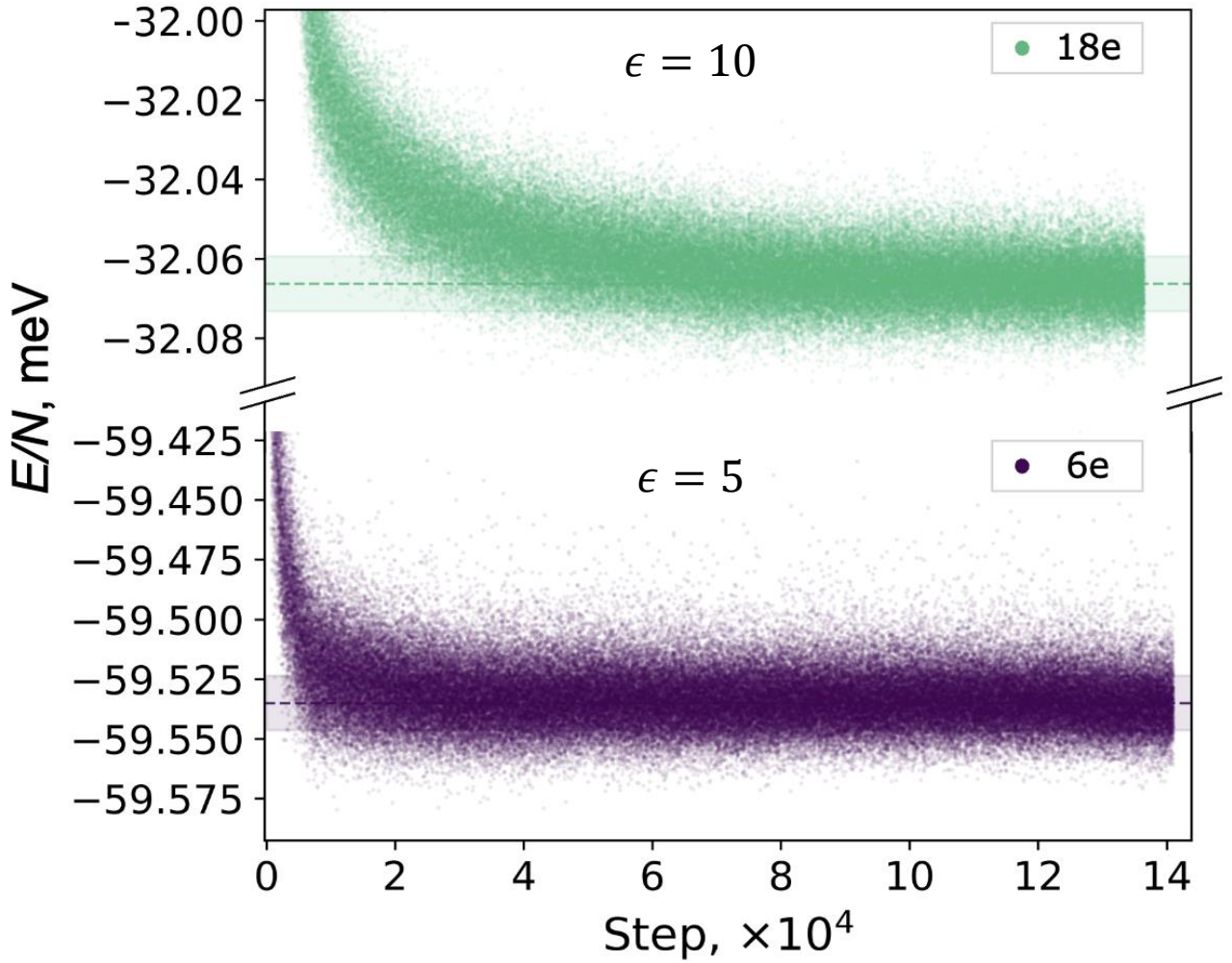
Filling fraction: $\nu = \frac{2}{3}$

Spin-polarized

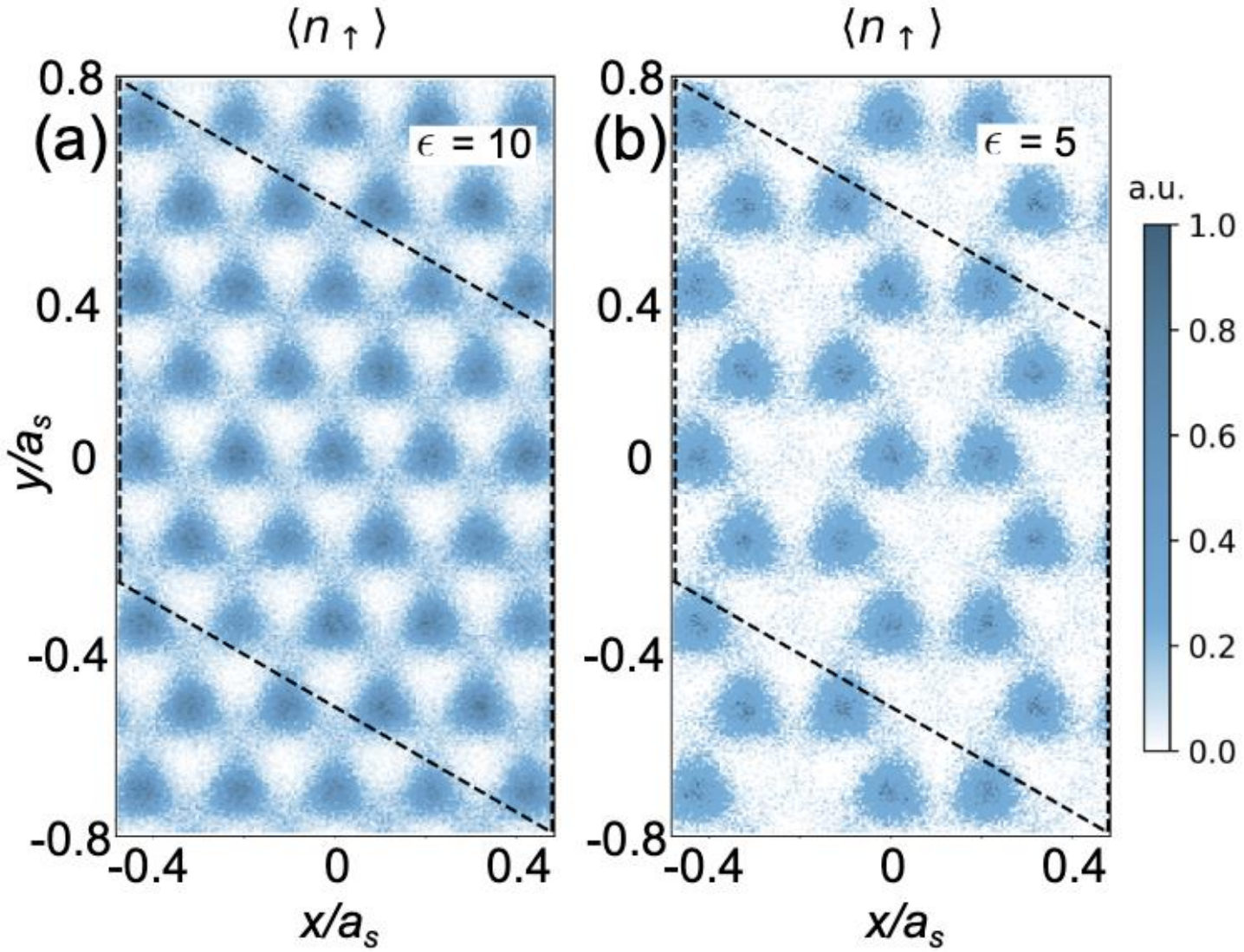
Geier, Nazaryan, Zaklama, and Fu (2025)



Convergence

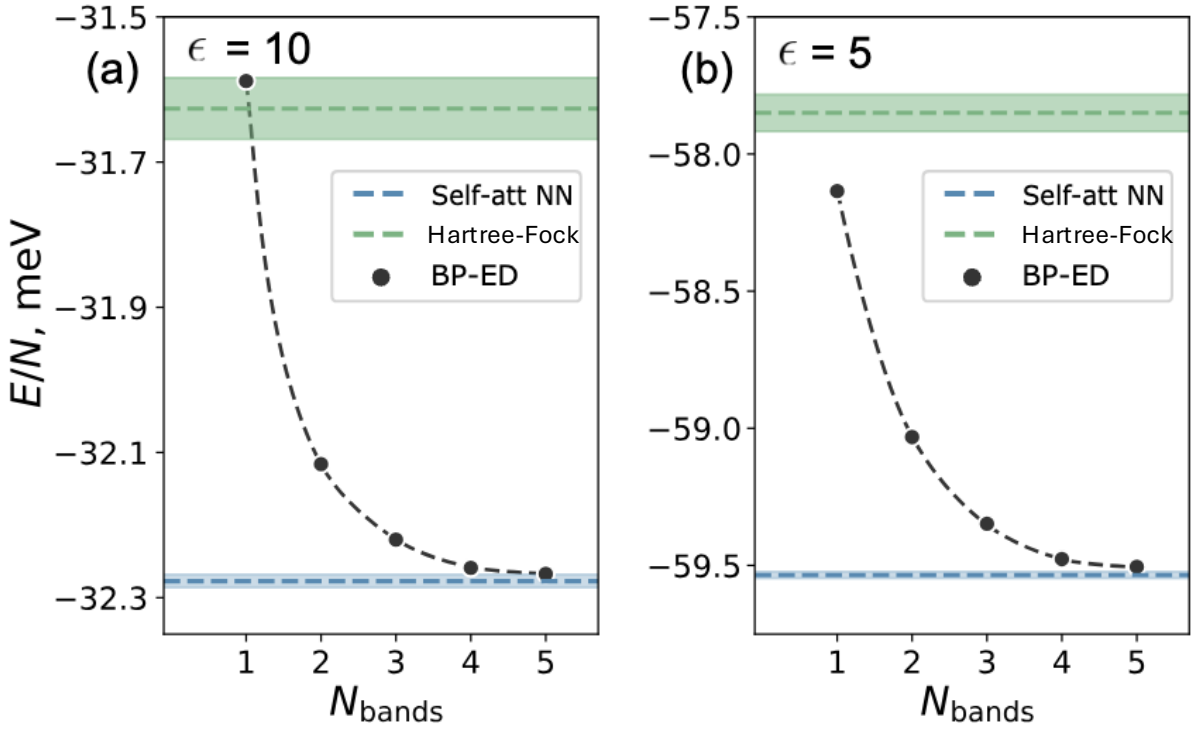


Fermi liquid to generalized Wigner crystal transition



Benchmark with band-projected exact diagonalization

9 lattice sites

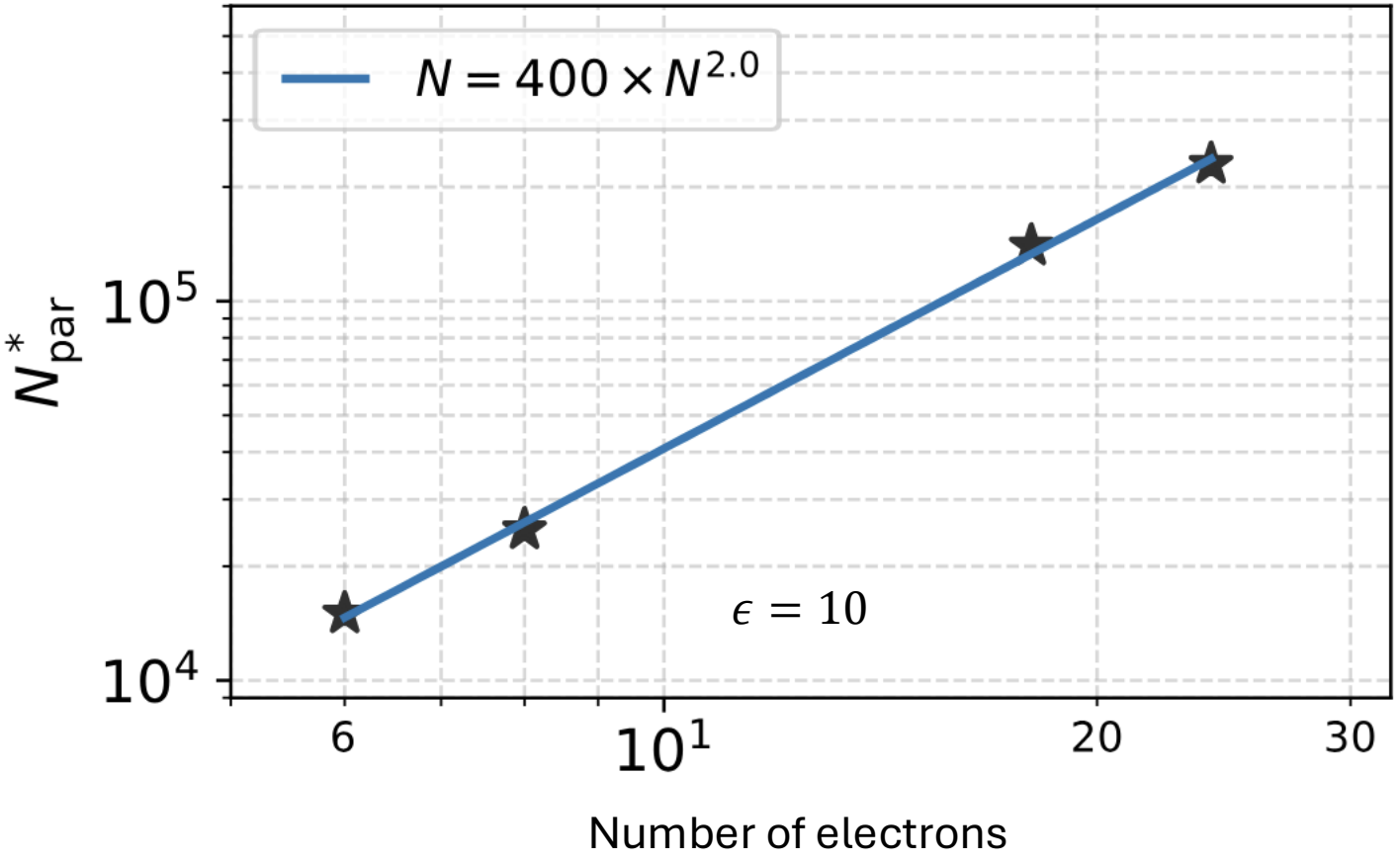


27 lattice sites

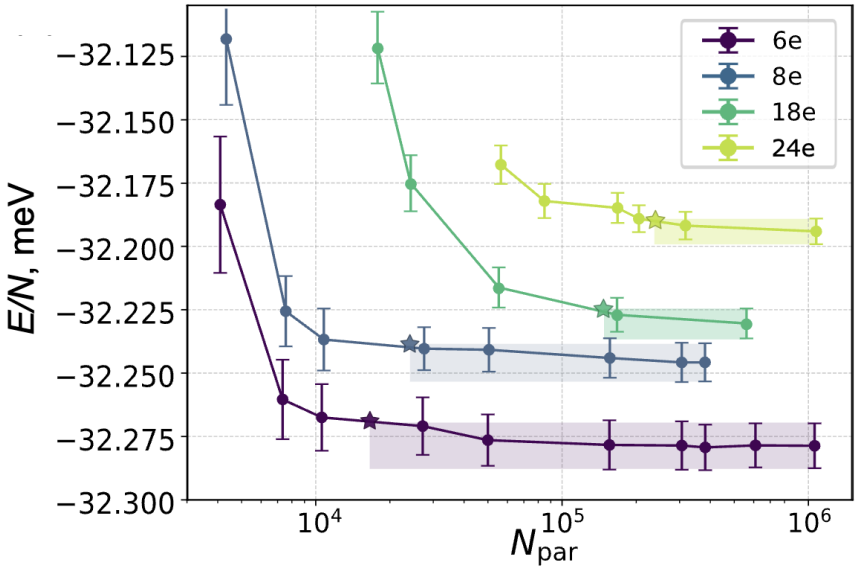
# sites	ϵ	self-attention NN	Hartree-Fock	BP-ED
27	10	**-32.070(7)**	-31.35(2)	-31.32443
27	5	**-59.127(9)**	-58.01(3)	-57.80848

Scaling law

Number of parameters required until convergence



Parameter convergence defined from variance during optimization

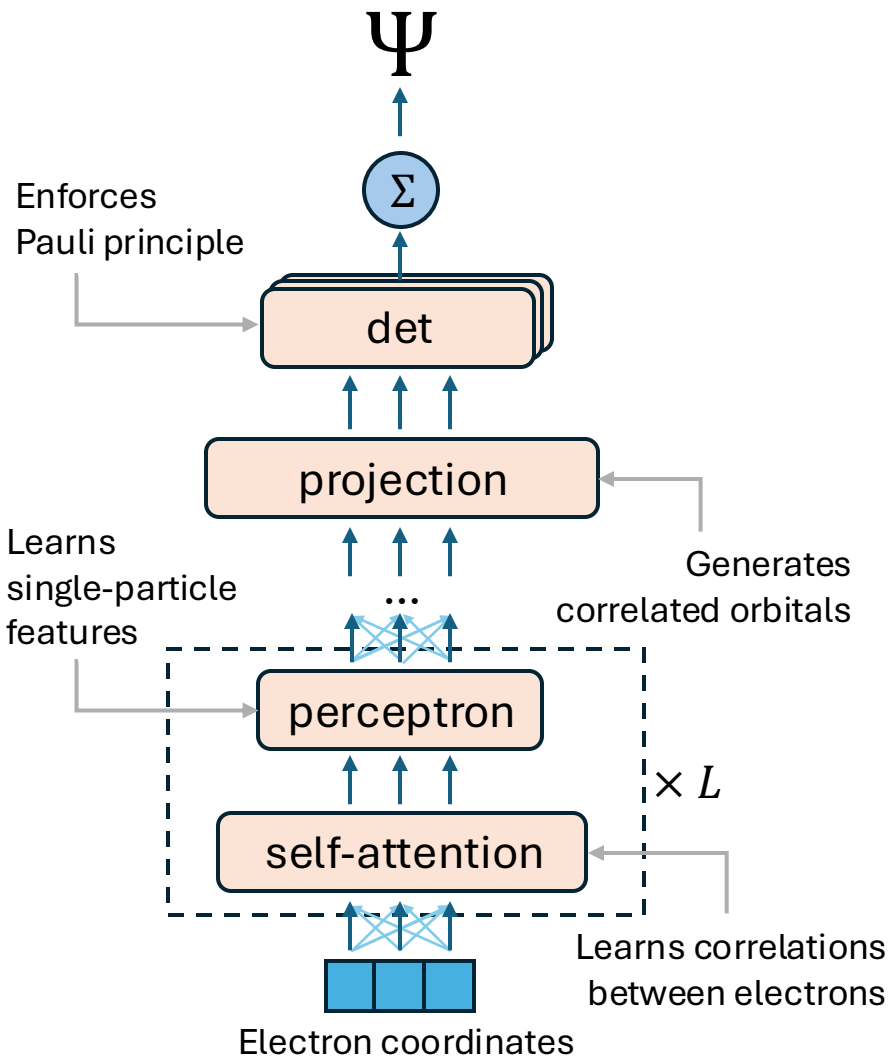


Is attention all you need to solve the correlated electron problem?

Many-parameter neural network provides an ansatz **without human bias**,
where **correlations are described by self-attention**,
achieving **quantitative accuracy.**

Spontaneous symmetry breaking [1] ✓

Topological order [2] ✓



[1] M. Geier, K. Nazaryan, T. Zaklama, and L. Fu, arXiv:2502.05383 (2025)
[2] Y. Teng, D. Dai, and L. Fu, arXiv:2412.00618 (2024)