# Solving the Many-Electron Schrödinger Equation

## With a Transformer Architecture

# Outline

# Setup: distances, spins, coordinates

- System: $N$ electrons, nuclei with charge $Z_I$ at fixed $\mathbf{R}_I$.
- Electron coordinates $\mathbf{r}_i \in \mathbb{R}^3$, spin $\sigma_i \in \{\uparrow, \downarrow\}$.
- Atomic units (distances in Bohr, energies in Hartree).
- Distances: $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, $r_{iI} = |\mathbf{r}_i - \mathbf{R}_I|$.

# The equation (TISE Hamiltonian)

- Kinetic energy: $T = -\frac{1}{2}\sum_{i=1}^{N} \nabla_i^2$.
- Electron-nuclear attraction: $V_{en} = -\sum_{i,I} \frac{Z_I}{r_{iI}}$.
- Electron-electron repulsion: $V_{ee} = \sum_{i<j} \frac{1}{r_{ij}}$.
- Hamiltonian: $H = T + V_{en} + V_{ee}$.
- Solve the time-independent Schrödinger equation: $H\,\Psi = E\,\Psi$.

# Fermi–Dirac statistics & antisymmetry

- Electrons are indistinguishable fermions.
- Exchanging two electrons flips the wavefunction's sign: $\Psi(\dots i, j \dots) = -\Psi(\dots j, i \dots)$.
- Pauli exclusion: no two electrons can occupy the same state.
- Use a Slater determinant to enforce an antisymmetric $\Psi$.

# Exponential decay of wavefunction

- Bound-state wavefunctions decay exponentially as $r \to \infty$.
- Example: hydrogen ground state $\psi(r) \sim e^{-r}$ (in a.u.).
- Ansätze include an exponential envelope so that $\Psi \to 0$ as $r \to \infty$.

# Kato cusp conditions

- Coulomb potentials cause a sharp cusp in $\Psi$ when particles coalesce.
- Electron–nucleus cusp: $\left.\frac{\partial \Psi}{\partial r_{iI}}\right|_{r_{iI}=0} = -Z_I\,\Psi(0).$
- Electron–electron cusp: $\left.\frac{\partial \Psi}{\partial r_{ij}}\right|_{r_{ij}=0} = \frac{1}{2}\,\Psi(0).$

# The Psiformer Ansatz (overview)

- Neural network trial wavefunction ("Wavefunction Transformer").
- Uses self-attention layers to model electron correlations.
- Permutation equivariant (independent of electron ordering).
- Greatly improves accuracy, especially for larger systems.
- Solves Schrödinger equation from first principles (no external data).

# Dissecting the Ansatz

- Slater–Jastrow form: $\Psi = \left( \sum_d c_d \det\left[\phi_k^d(\mathbf{r}_i)\right] \right) \exp(J)$.
- Slater determinant part ensures the correct antisymmetric exchange.
- One-electron orbitals $\phi_k^d$ are learned functions of all electron coordinates.
- Jastrow factor $J$ enforces electron-electron cusp conditions.
- Additional envelope functions ensure exponential decay at long range.

# Jastrow factor form (parallel/antiparallel spins)

- Jastrow factor: $J = \sum_{i<j} u_{\sigma_i,\sigma_j}(r_{ij})$.
- Use separate $u(r)$ for same-spin vs opposite-spin pairs.
- Example choice: $u_{\uparrow\downarrow}(r) = \frac{\alpha\, r}{1+\beta\, r}$ (two parameters).
- Satisfies cusp: $u(r) \approx \frac{1}{2} r$ as $r \to 0$.
- Jastrow adds correlation beyond antisymmetry (lowers the energy).

# Loss: Rayleigh quotient

- Variational principle uses the energy expectation as loss.
- $E[\Psi] = \dfrac{\langle \Psi \mid H \mid \Psi \rangle}{\langle \Psi \mid \Psi \rangle} \geq E_0$ (Rayleigh quotient).
- Minimizing $E[\Psi]$ drives the ansatz toward the ground state.
- $E_0$ is the true ground energy (minimum possible).

# Metropolis–Hastings sampling

- Markov Chain Monte Carlo algorithm to sample $|\Psi(\mathbf{r})|^2$.
- Propose random moves for electron coordinates.
- Accept move with probability $A = \min\left(1,\ \frac{|\Psi_{\text{new}}|^2}{|\Psi_{\text{old}}|^2}\right)$.
- After thermalization, obtained samples follow the target distribution $|\Psi|^2$.

# Variational Monte Carlo loop

- Initialize neural network parameters (trial $\Psi$).
- Sample electron configurations via Metropolis–Hastings.
- Evaluate energy and gradient from the sampled configurations.
- Update parameters to lower the energy (gradient descent).
- Iterate until convergence to the minimum energy (ground state).

# Optimizer: KFAC (natural gradient)

- Kronecker-Factored Approximate Curvature (KFAC) optimizer.
- Approximates the natural gradient using a factored curvature matrix.
- Speeds up training and improves stability for large networks.
- Used to efficiently optimize FermiNet and Psiformer wavefunctions.

# FermiNet baseline

- FermiNet: first deep-neural-network wavefunction ansatz (Pfau et al., 2020).
- Architecture: multiple dense layers with electron-wise feature streams.
- Outputs single-electron orbitals feeding into a Slater determinant.
- Achieved high accuracy on small molecules (near chemical accuracy).
- Faced scaling limits: accuracy and efficiency drop for larger systems.

# Psiformer vs FermiNet: Architecture & Attention

- FermiNet layers mix electron features via fixed functions; Psiformer uses self-attention.
- Self-attention: each electron attends to all others (learns interactions).
- Both ansatzes enforce antisymmetry via Slater determinants.
- Psiformer captures correlations more effectively with fewer parameters.
- Attention mechanism is permutation-invariant and scales to complex interactions.

# Computational power & scaling

- Solving many-electron systems is computationally intensive.
- Computational cost grows steeply with electron number $N$.
- Wavefunction evaluation involves expensive operations (e.g. determinants, $O(N^3)$).
- Requires significant computing resources (GPUs/TPUs, parallelization).

# Numerical stability

- Exponential and cusp factors can cause numerical overflow/underflow if not handled carefully.
- Ensuring stable Monte Carlo estimates (variance reduction techniques).
- Feature scaling and proper initialization improve training stability.
- Jastrow factor helps manage large energy fluctuations (by satisfying cusps).

# Scaling with electron count

- Wavefunction resides in a $3N$-dimensional configuration space.
- Monte Carlo integration becomes harder as $N$ increases (high dimensionality).
- More electrons often require deeper or wider neural networks.
- Ongoing research into architectures that scale better or transfer to larger $N$.

# Thanks

Thank you for your attention!