Solving the Many-Electron Schrödinger Equation with Deep Neural Networks

W.M.C. Foulkes[1] David Pfau[2] James S. Spencer[2]
Alexander G. de G. Matthews[2]

[1]Department of Physics, Imperial College London
[2]DeepMind, 6 Pancras Square, London

Workshop on Ab Initio Simulations of Correlated Fermions
Kiel University, Germany
8^{th}–9^{th} July 2020

Outline

Why?

To get chemistry right at room temperature, we need total energies to better than

Chemical Accuracy

1 kcal/mole $\approx$ 43 meV/particle $\approx$ 1.6 mHartree/particle

Why?

To get chemistry right at room temperature, we need total energies to better than

Chemical Accuracy

1 kcal/mole $\approx$ 43 meV/particle $\approx$ 1.6 mHartree/particle

- At room temperature, $k_B T \approx 25$ meV.

Why?

To get chemistry right at room temperature, we need total energies to better than

Chemical Accuracy

1 kcal/mole $\approx$ 43 meV/particle $\approx$ 1.6 mHartree/particle

- At room temperature, $k_B T \approx 25$ meV.
- Energy differences between competing structures can be < 10 meV.

Why?

To get chemistry right at room temperature, we need total energies to better than

Chemical Accuracy

1 kcal/mole $\approx$ 43 meV/particle $\approx$ 1.6 mHartree/particle

- At room temperature, $k_B T \approx 25$ meV.
- Energy differences between competing structures can be < 10 meV.
- Energy differences between different magnetic structures and different correlated states can be much smaller again.

Our Grand Unified Theory

The Many-Electron Schrödinger Hamiltonian

$$\sum_i -\frac{1}{2}\nabla_i^2 - \sum_i \sum_I \frac{Z_I e^2}{|\boldsymbol{r}_i - \boldsymbol{d}_I|} + \sum_{i>j} \frac{e^2}{|\boldsymbol{r}_i - \boldsymbol{r}_j|} + \sum_{I>J} \frac{Z_I Z_J e^2}{|\boldsymbol{d}_I - \boldsymbol{d}_J|}$$

The many-electron Schrödinger equation is NP-hard, so the best we can realistically hope for are approximate results.

Energy and Length Scales

Energy scales

Total electronic	$> 10^2$ eV
Chemical bond	few eV
Chemical reaction	10^{-1} eV
Correlations and magnetism	$< 10^{-2}$ eV

Energy and Length Scales

Energy scales

Total electronic	$> 10^2$ eV
Chemical bond	few eV
Chemical reaction	10^{-1} eV
Correlations and magnetism	$< 10^{-2}$ eV

Length scale for accuracy of 0.1 eV

$$E/E_F = 10^{-2}$$
$$\lambda/\lambda_F \approx 10$$
$$\lambda \sim 20\,\text{Å}$$

Density Functional Theory

can sometimes achieve chemical accuracy — if you choose the right functional. But how can you tell in advance?

Perturbation Theory

works well for weakly-correlated materials. But what about strong correlations and non-Fermi-liquid ground states?

Guessing the Wavefunction

was the method responsible for most of our few successes in understanding strongly-correlated non-Fermi-liquid systems:

- BCS
- FQHE
- Bethe Ansatz
- ...

Parameterized wavefunction guessing

- Hartree-Fock
- Configuration interation
- Slater-Jastrow
- Coupled cluster
- RVB/Pfaffian/geminal
- DMRG, matrix product states, tensor product states
- RPA
- ...

Outline

Variational Quantum Monte Carlo

- Guess $\Psi_T(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) = \Psi_T(\boldsymbol{R})$.
- Evaluate

$$E_T[\Psi_T] = \int \Psi_T^*(\boldsymbol{R}) \hat{H} \Psi_T(\boldsymbol{R})\, d\boldsymbol{R}$$
$$= \int \left(\frac{\hat{H}\Psi_T(\boldsymbol{R})}{\Psi_T(\boldsymbol{R})} \right) |\Psi_T(\boldsymbol{R})|^2\, d\boldsymbol{R}$$

  using Monte Carlo integration.
- Adjust Ψ_T to minimise $E_T[\Psi_T]$.

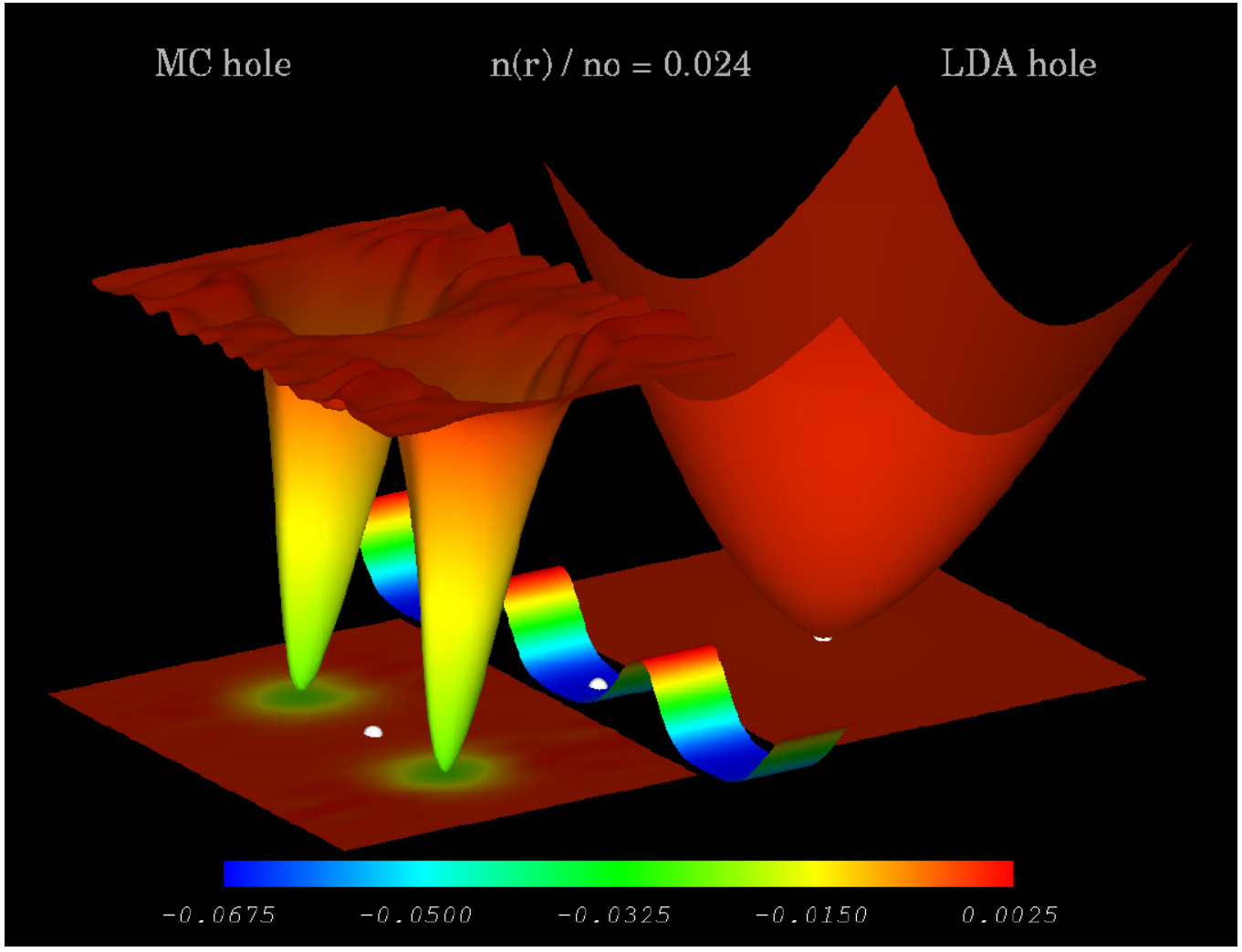
The Slater-Jastrow Trial Wavefunction

remains the most popular guess for large three-dimensional systems:

$$\Psi_T(\boldsymbol{R}) = \exp\left(-\sum_{i>j} u(r_{ij})\right) D$$

$$D = \begin{vmatrix} \psi_1(\boldsymbol{r}_1) & \psi_1(\boldsymbol{r}_2) & \dots & \psi_1(\boldsymbol{r}_N) \\ \psi_2(\boldsymbol{r}_1) & \psi_2(\boldsymbol{r}_2) & \dots & \psi_2(\boldsymbol{r}_N) \\ . & . & \dots & . \\ . & . & \dots & . \\ \psi_N(\boldsymbol{r}_1) & \psi_N(\boldsymbol{r}_2) & \dots & \psi_N(\boldsymbol{r}_N) \end{vmatrix}$$

It works surprisingly well in weakly-correlated Fermi-liquid-like solids. Systems of thousands of electrons can be studied.



Exchange-correlation hole in sine-wave jellium

Limitations

- VMC with a SJ trial function is not normally capable of chemical accuracy, even in weakly correlated systems. (DMC is much better.)
- How to extend the SJ form when it doesn't work?
  - Linear combinations of determinants are often used, but this approach does not scale well.
  - No help when Fermi-liquid physics is very wrong.

Observation 1

Artificial neural networks are flexible and efficient function approximators.

Observation 1

Artificial neural networks are flexible and efficient function approximators.

Neural QMC

Perhaps we can represent $\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$ as a deep neural network?

Observation 2

- Nothing requires the orbitals in a Slater determinant

$$D = \begin{vmatrix} \psi_1(\boldsymbol{r}_1) & \psi_1(\boldsymbol{r}_2) & \dots & \psi_1(\boldsymbol{r}_N) \\ \psi_2(\boldsymbol{r}_1) & \psi_2(\boldsymbol{r}_2) & \dots & \psi_2(\boldsymbol{r}_N) \\ . & . & \dots & . \\ . & . & \dots & . \\ \psi_N(\boldsymbol{r}_1) & \psi_N(\boldsymbol{r}_2) & \dots & \psi_N(\boldsymbol{r}_N) \end{vmatrix}$$

  to be functions of the coordinates of a single electron.
- The only requirement is that exchanging any two input variables, $\boldsymbol{r}_i$ and $\boldsymbol{r}_j$, exchanges two columns.

- This allows us to replace the single-electron orbitals $\psi_i(\boldsymbol{r}_j)$ by multi-electron functions

$$\psi_i(\boldsymbol{r}_j; \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{j-1}, \boldsymbol{r}_{j+1}, \ldots, \boldsymbol{r}_N) = \psi_i(\boldsymbol{r}_j; \{\boldsymbol{r}_{/j}\})$$

- D remains antisymmetric as long as $\psi_i(\boldsymbol{r}_j; \{\boldsymbol{r}_{/j}\})$ is invariant under any change in the order of the arguments after $\boldsymbol{r}_j$:

$$D = \begin{vmatrix} \psi_1(\boldsymbol{r}_1, \{\boldsymbol{r}_{/1}\}) & \psi_1(\boldsymbol{r}_2, \{\boldsymbol{r}_{/2}\}) & \ldots & \psi_1(\boldsymbol{r}_N, \{\boldsymbol{r}_{/N}\}) \\ \psi_2(\boldsymbol{r}_1, \{\boldsymbol{r}_{/1}\}) & \psi_2(\boldsymbol{r}_2, \{\boldsymbol{r}_{/2}\}) & \ldots & \psi_2(\boldsymbol{r}_N, \{\boldsymbol{r}_{/N}\}) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \psi_N(\boldsymbol{r}_1, \{\boldsymbol{r}_{/1}\}) & \psi_N(\boldsymbol{r}_2, \{\boldsymbol{r}_{/2}\}) & \ldots & \psi_N(\boldsymbol{r}_N, \{\boldsymbol{r}_{/N}\}) \end{vmatrix}$$

- This allows us to replace the single-electron orbitals $\psi_i(\boldsymbol{r}_j)$ by multi-electron functions

$$\psi_i(\boldsymbol{r}_j; \boldsymbol{r}_1, \ldots, \boldsymbol{r}_{j-1}, \boldsymbol{r}_{j+1}, \ldots, \boldsymbol{r}_N) = \psi_i(\boldsymbol{r}_j; \{\boldsymbol{r}_{/j}\})$$

- D remains antisymmetric as long as $\psi_i(\boldsymbol{r}_j; \{\boldsymbol{r}_{/j}\})$ is invariant under any change in the order of the arguments after $\boldsymbol{r}_j$:

$$D = \begin{vmatrix} \psi_1(\boldsymbol{r}_1, \{\boldsymbol{r}_{/1}\}) & \psi_1(\boldsymbol{r}_2, \{\boldsymbol{r}_{/2}\}) & \ldots & \psi_1(\boldsymbol{r}_N, \{\boldsymbol{r}_{/N}\}) \\ \psi_2(\boldsymbol{r}_1, \{\boldsymbol{r}_{/1}\}) & \psi_2(\boldsymbol{r}_2, \{\boldsymbol{r}_{/2}\}) & \ldots & \psi_2(\boldsymbol{r}_N, \{\boldsymbol{r}_{/N}\}) \\ . & . & \ldots & . \\ . & . & \ldots & . \\ \psi_N(\boldsymbol{r}_1, \{\boldsymbol{r}_{/1}\}) & \psi_N(\boldsymbol{r}_2, \{\boldsymbol{r}_{/2}\}) & \ldots & \psi_N(\boldsymbol{r}_N, \{\boldsymbol{r}_{/N}\}) \end{vmatrix}$$

Fermi Net

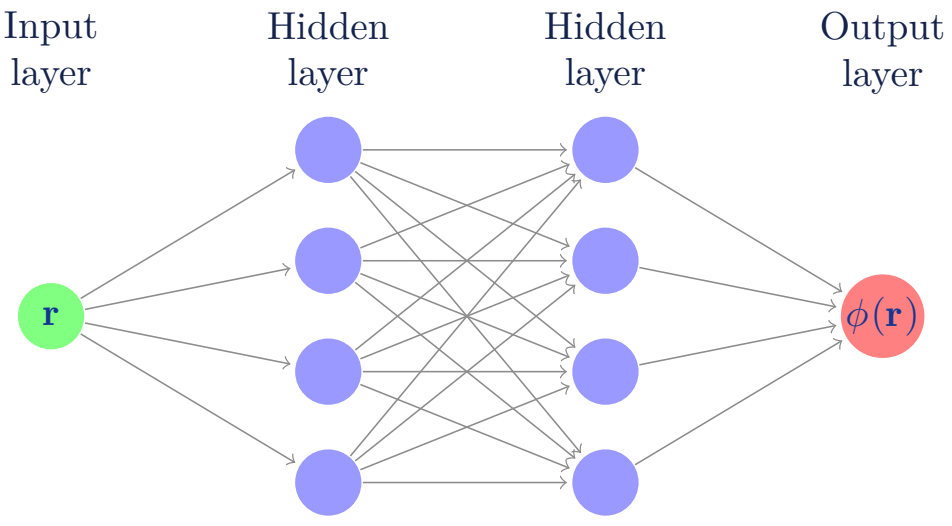
The construction of these permutation-equivariant functions with a neural network is the main innovation of Fermi Net

Outline

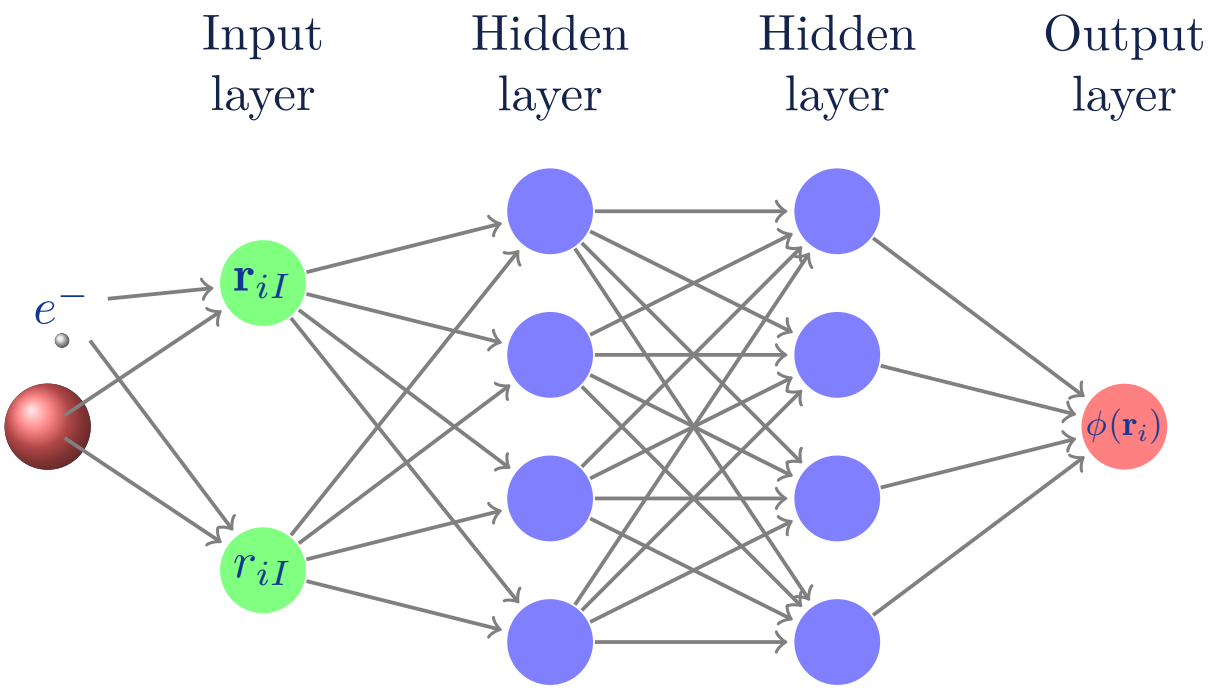
Neural Network Wavefunction

$$\boldsymbol{h}^{\ell+1} = f(\boldsymbol{W}^{\ell}\boldsymbol{h}^{\ell} + \boldsymbol{b}^{\ell})$$

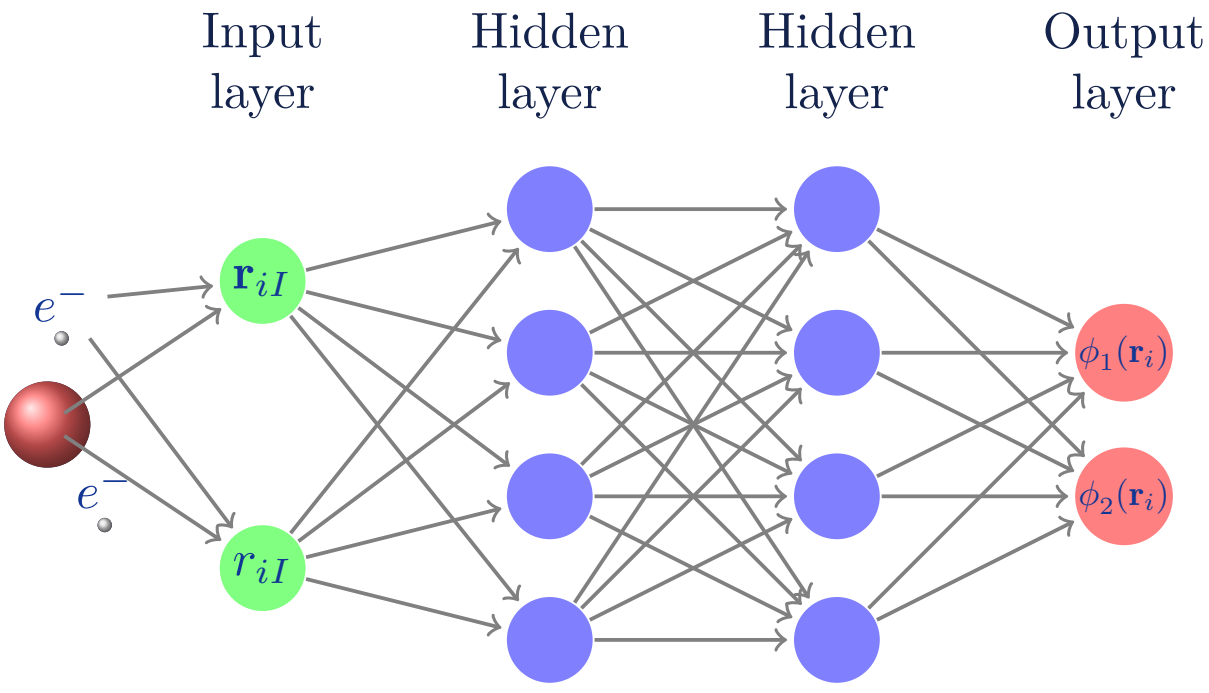
Learn network parameters using automatic differentiation and back propagation.



Neural Network Wavefunction

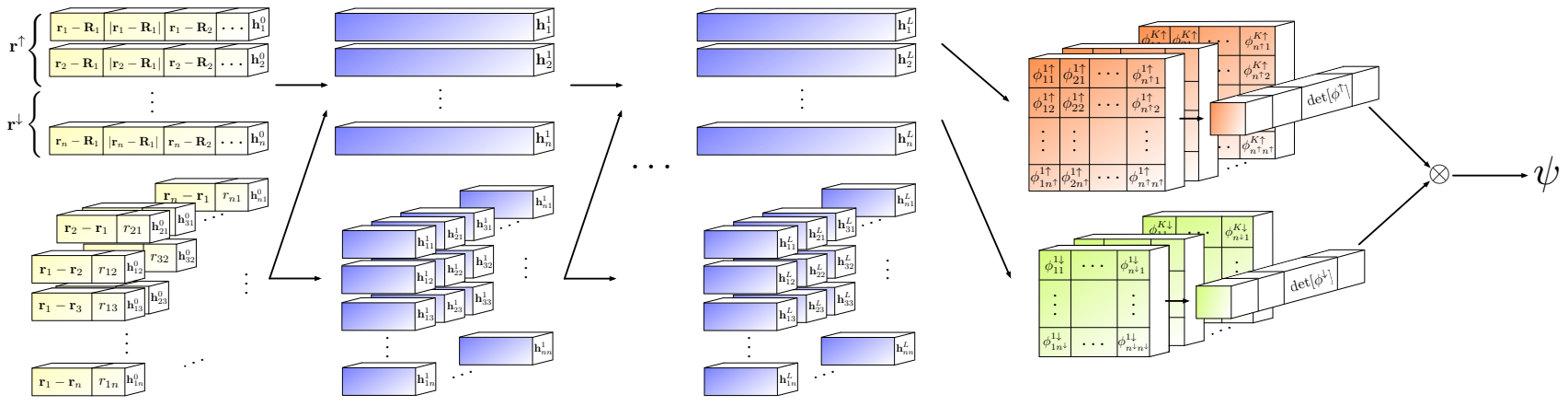


One Slater Determinant



$$\Psi(\boldsymbol{r}_1, \boldsymbol{r}_2) = \begin{vmatrix} \phi_1(\boldsymbol{r}_1) & \phi_1(\boldsymbol{r}_2) \\ \phi_2(\boldsymbol{r}_1) & \phi_2(\boldsymbol{r}_2) \end{vmatrix}$$

Fermi Net Wavefunction



- $\boldsymbol{h}_i^0$ is a vector with elements $\boldsymbol{r}_i - \boldsymbol{d}_I$ and $|\boldsymbol{r}_i - \boldsymbol{d}_I|$ (for all ions I).
  - The elements of $\boldsymbol{h}_i^0$ depend on $\boldsymbol{r}_i$ only.
  - Including $|\boldsymbol{r}_i - \boldsymbol{d}_I|$ helps the network learn the electron-nuclear cusps.

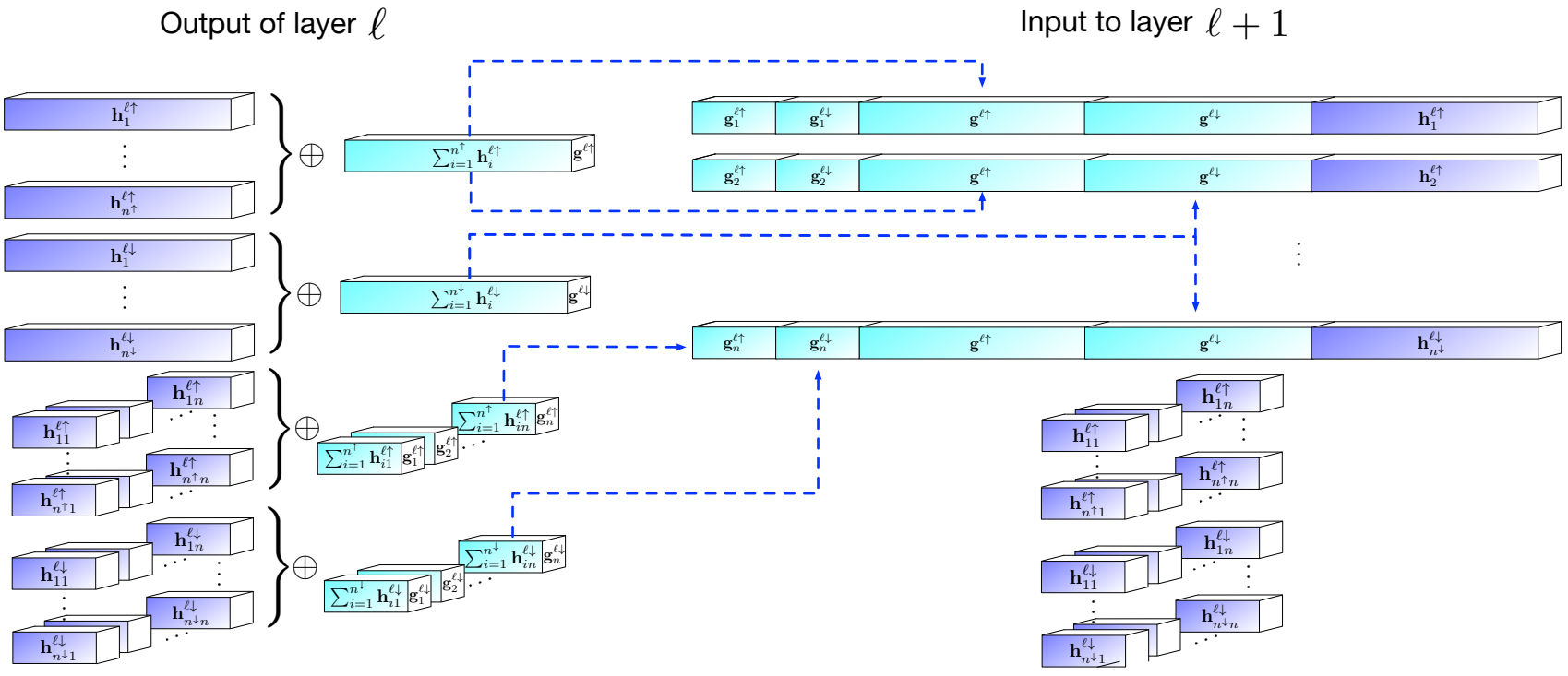
- $\boldsymbol{h}_i^0$ is a vector with elements $\boldsymbol{r}_i - \boldsymbol{d}_I$ and $|\boldsymbol{r}_i - \boldsymbol{d}_I|$ (for all ions I).
  - The elements of $\boldsymbol{h}_i^0$ depend on $\boldsymbol{r}_i$ only.
  - Including $|\boldsymbol{r}_i - \boldsymbol{d}_I|$ helps the network learn the electron-nuclear cusps.
- As data propagates through the network

$$\boldsymbol{h}_i^0 \to \boldsymbol{h}_i^1 \to \ldots \to \boldsymbol{h}_i^L$$

information about the positions of other electrons is mixed in such that

$$\boldsymbol{h}_i^L = \boldsymbol{h}_i^L(\boldsymbol{r}_i, \{\boldsymbol{r}_{/i}\})$$

Permutation Equivariant Coupling



Neural QMC

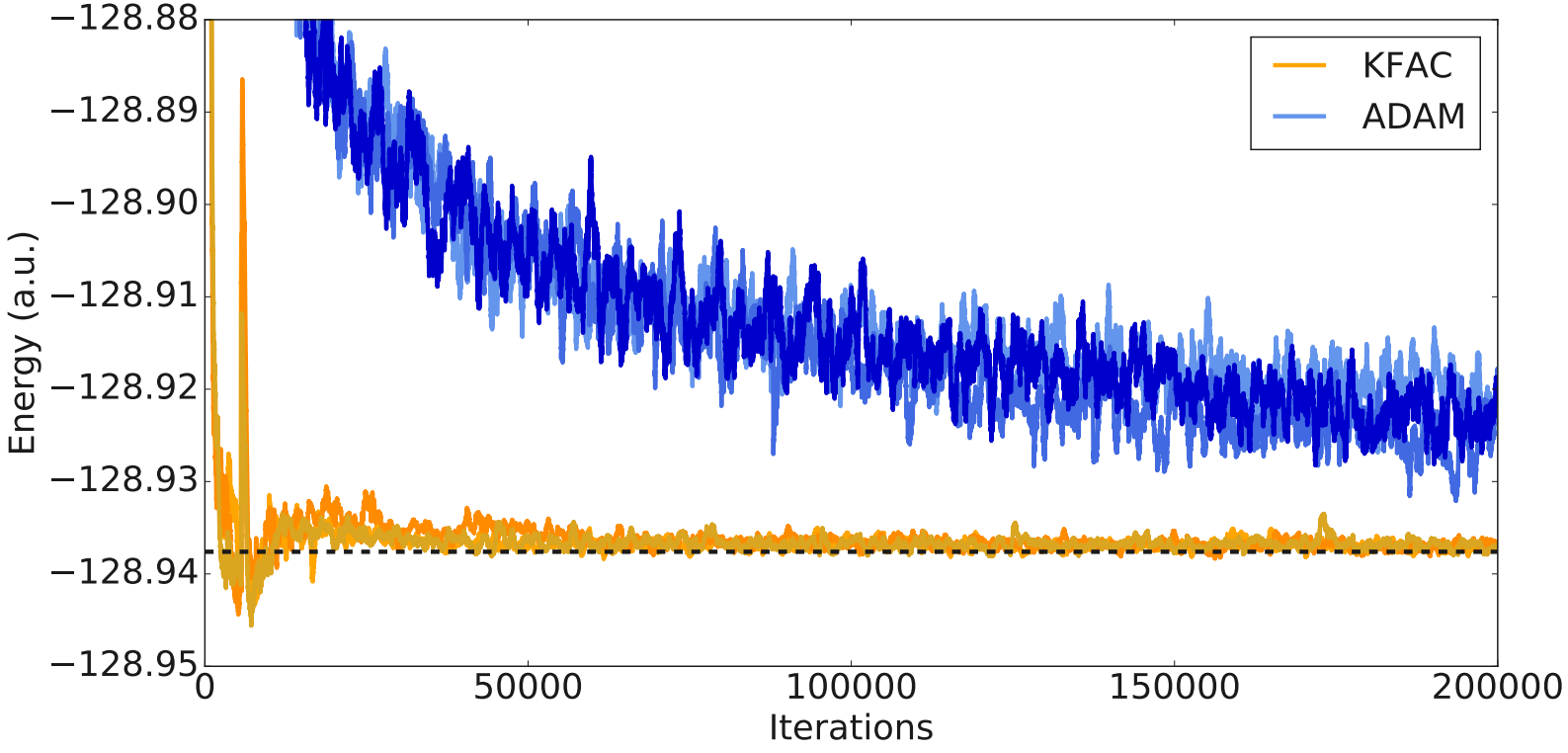
- Sample points $\boldsymbol{R} = (\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N)$ from current $|\Psi_T(\boldsymbol{R})|^2$ using the Metropolis algorithm.
- Use values of Ψ_T and its derivatives at the sampled points to estimate

$$E_T[\Psi_T] = \int \left(\frac{\hat{H}\Psi_T(\boldsymbol{R})}{\Psi_T(\boldsymbol{R})} \right) |\Psi_T(\boldsymbol{R})|^2 \, d\boldsymbol{R}$$

  and its first derivatives.
- Adjust network parameters to lower $E_T[\Psi_T]$.

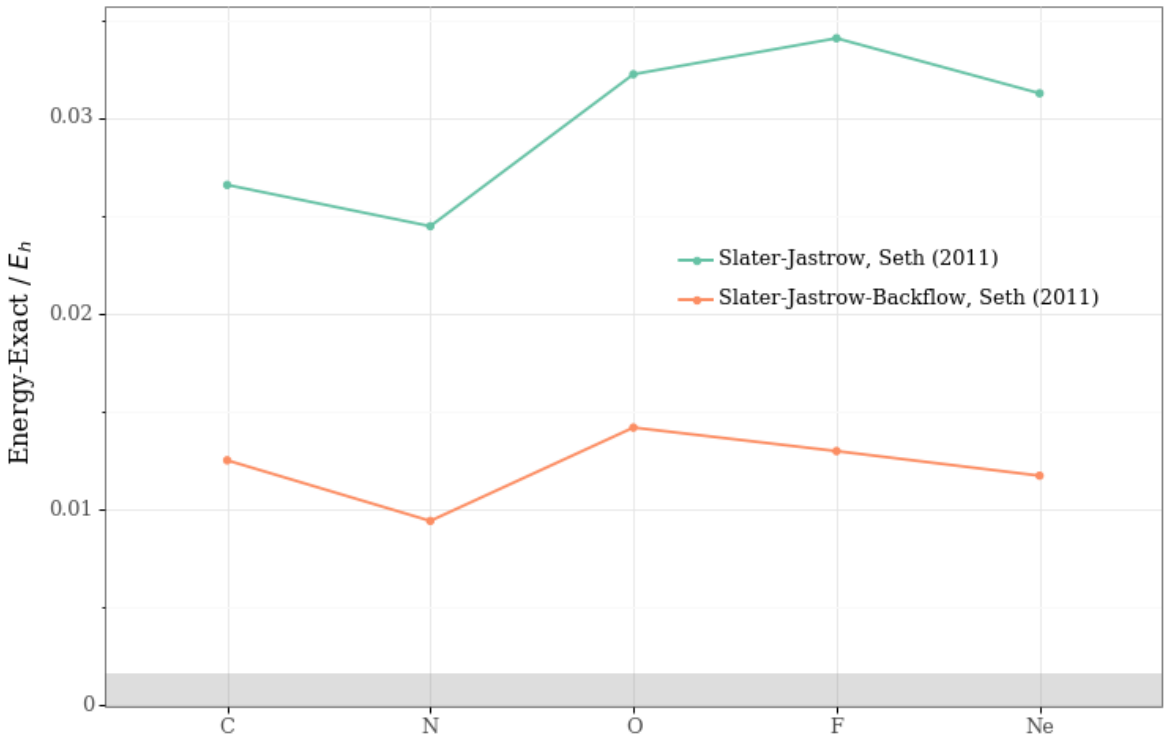
Kronecker-Factored Approximate Curvature

(approximates Natural Gradient Descent/Stochastic Reconfiguration)



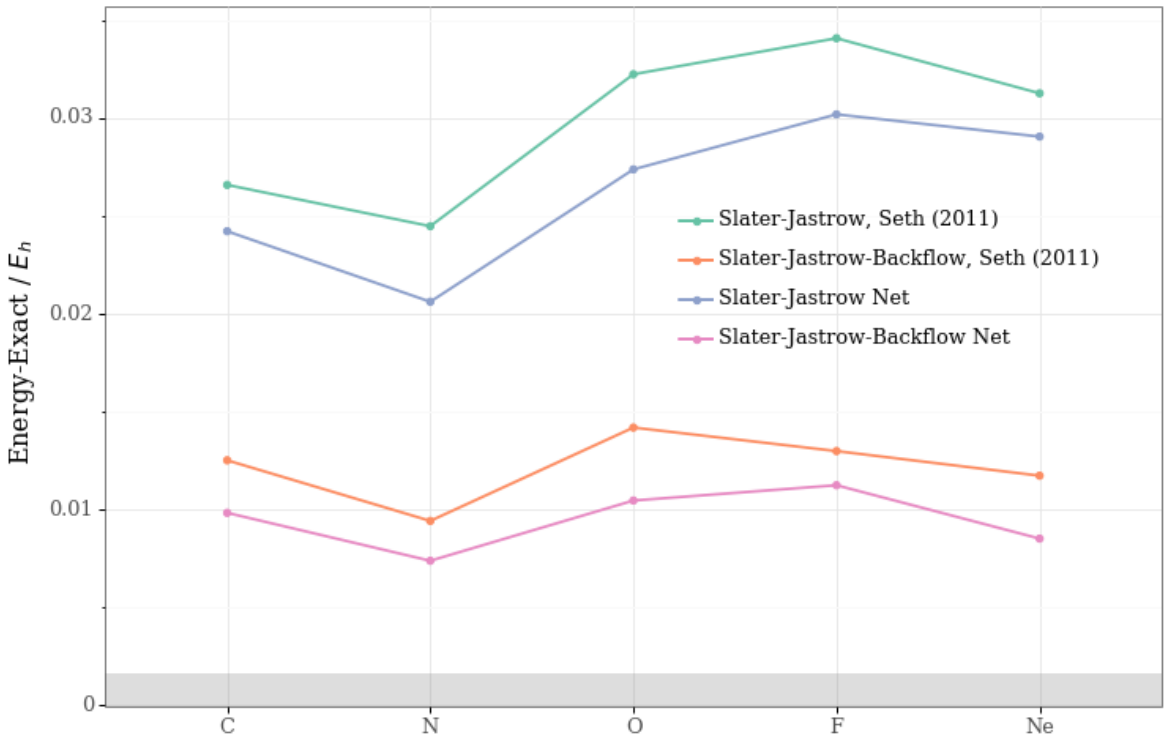
Outline

Single Determinant



Backflow captures a large fraction of the remaining correlation energy.

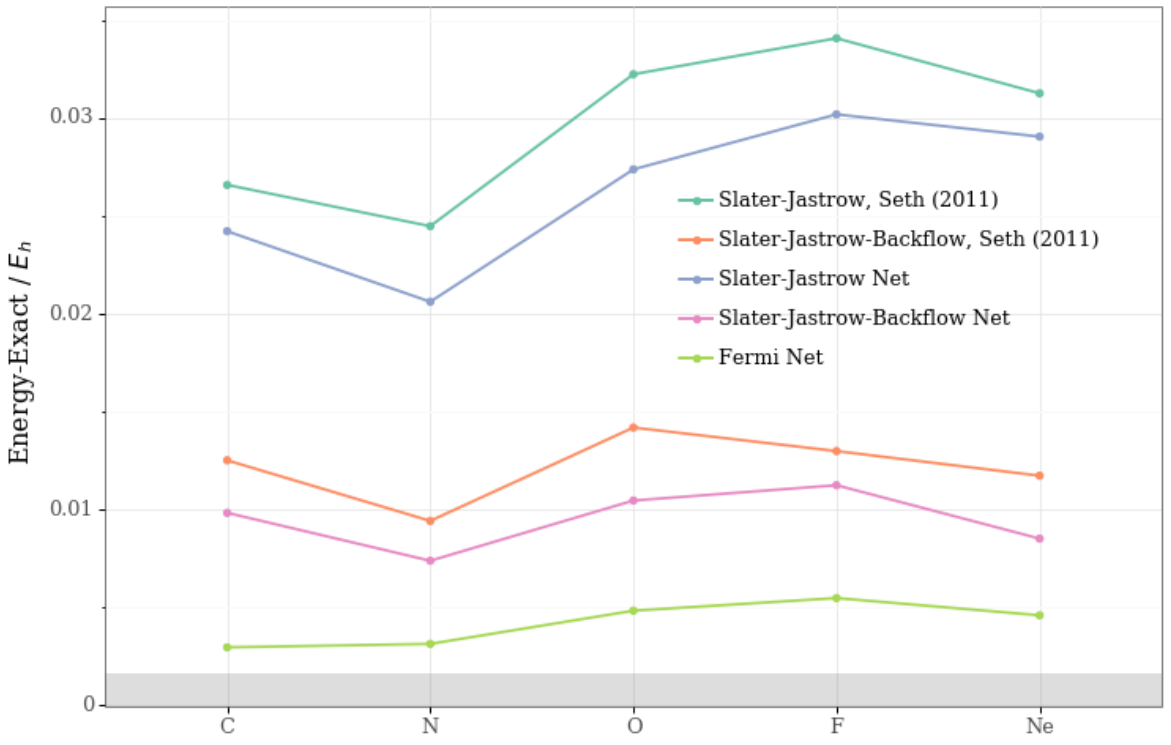
Single Determinant



Backflow captures a large fraction of the remaining correlation energy.

Neural networks offer more flexible functional forms.

Single Determinant

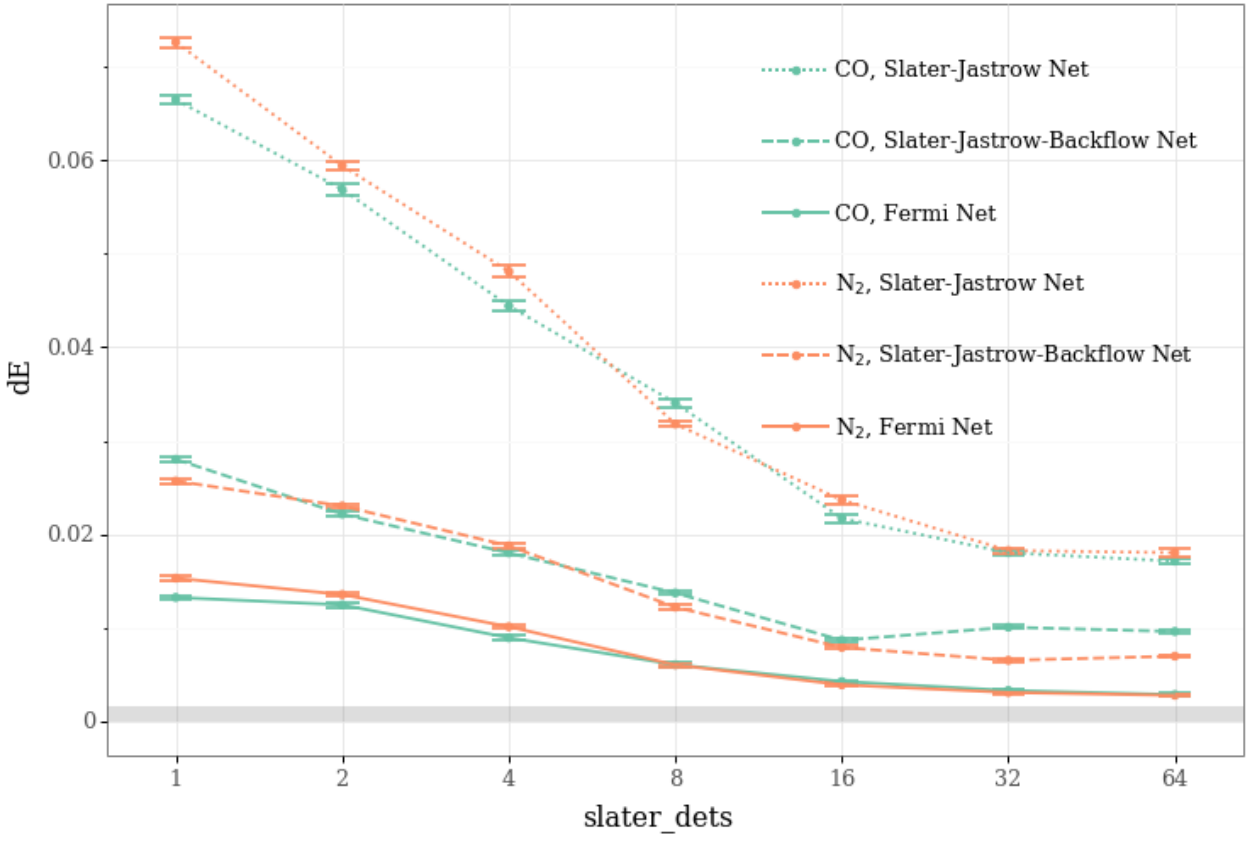


Backflow captures a large fraction of the remaining correlation energy.

Neural networks offer more flexible functional forms.

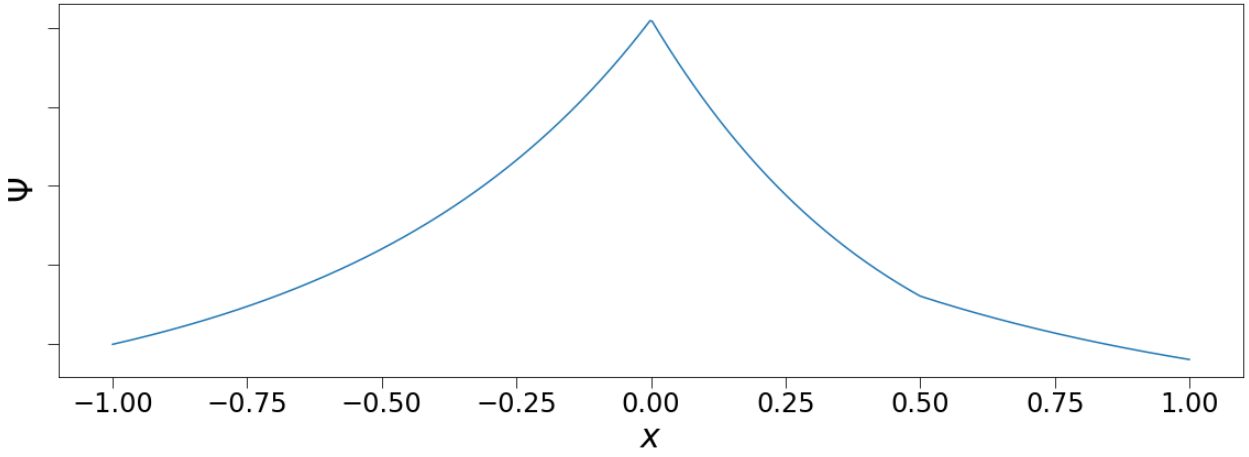
Fermi Net substantially reduces the error.

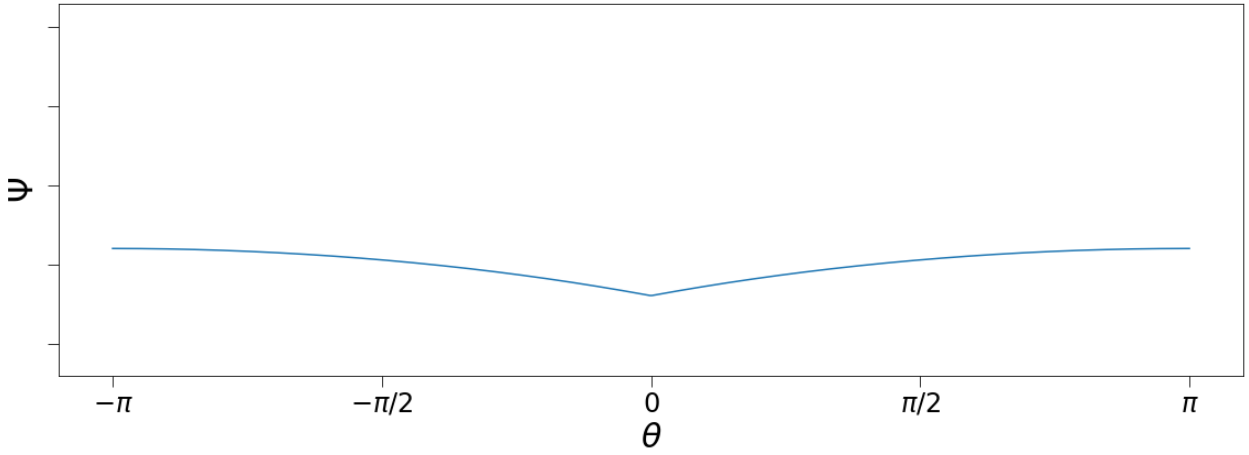
Adding Determinants

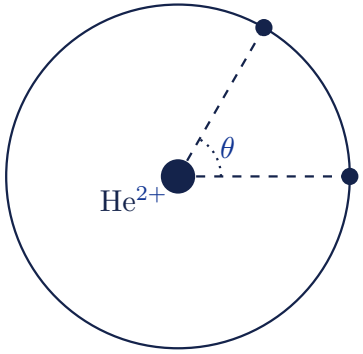


- Fermi Net converges quite rapidly with the number of determinants.
- Substantially easier to optimise than Slater-Jastrow and Slater-Jastrow-Backflow networks.

Learning the Cusps

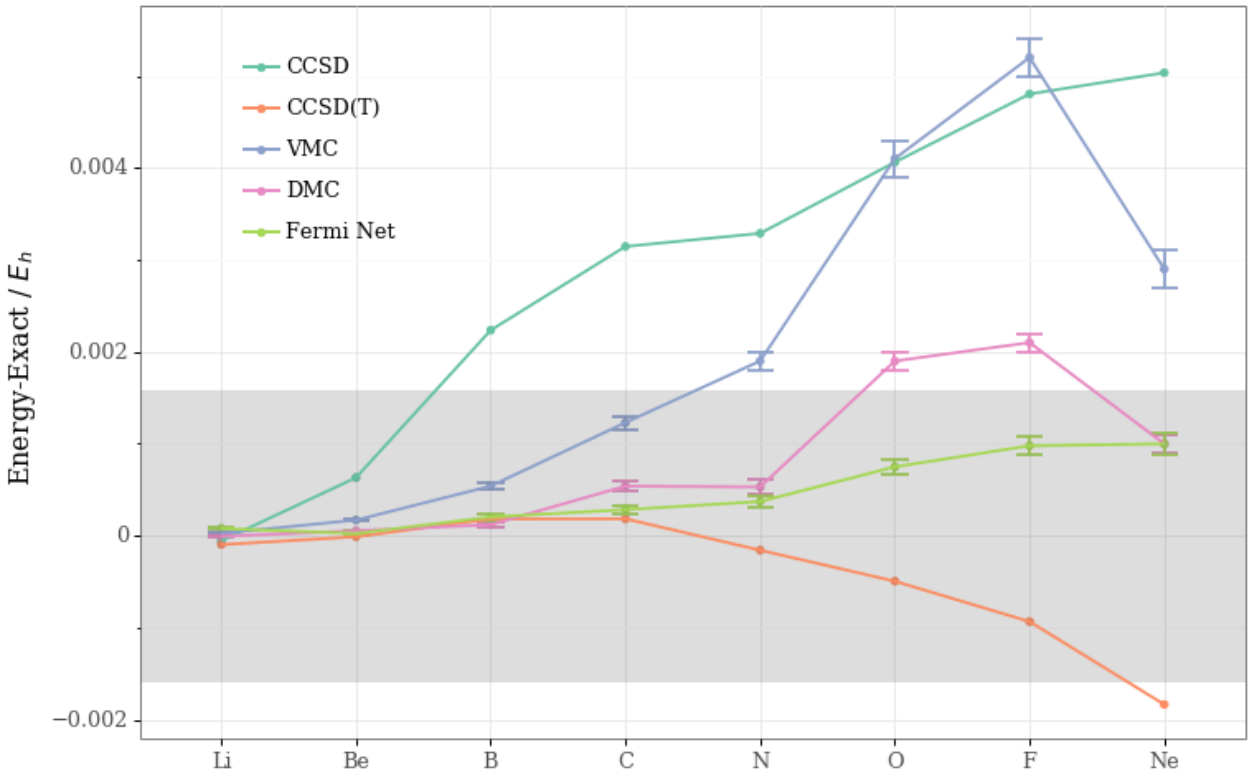






$$\left.\frac{\partial \ln|\Psi|}{\partial r_1}\right|_{r_1=0} = -1.9979(4), \qquad \left.\frac{\partial \ln|\Psi|}{\partial r_{12}}\right|_{r_{12}=0} = 0.4934(1).$$

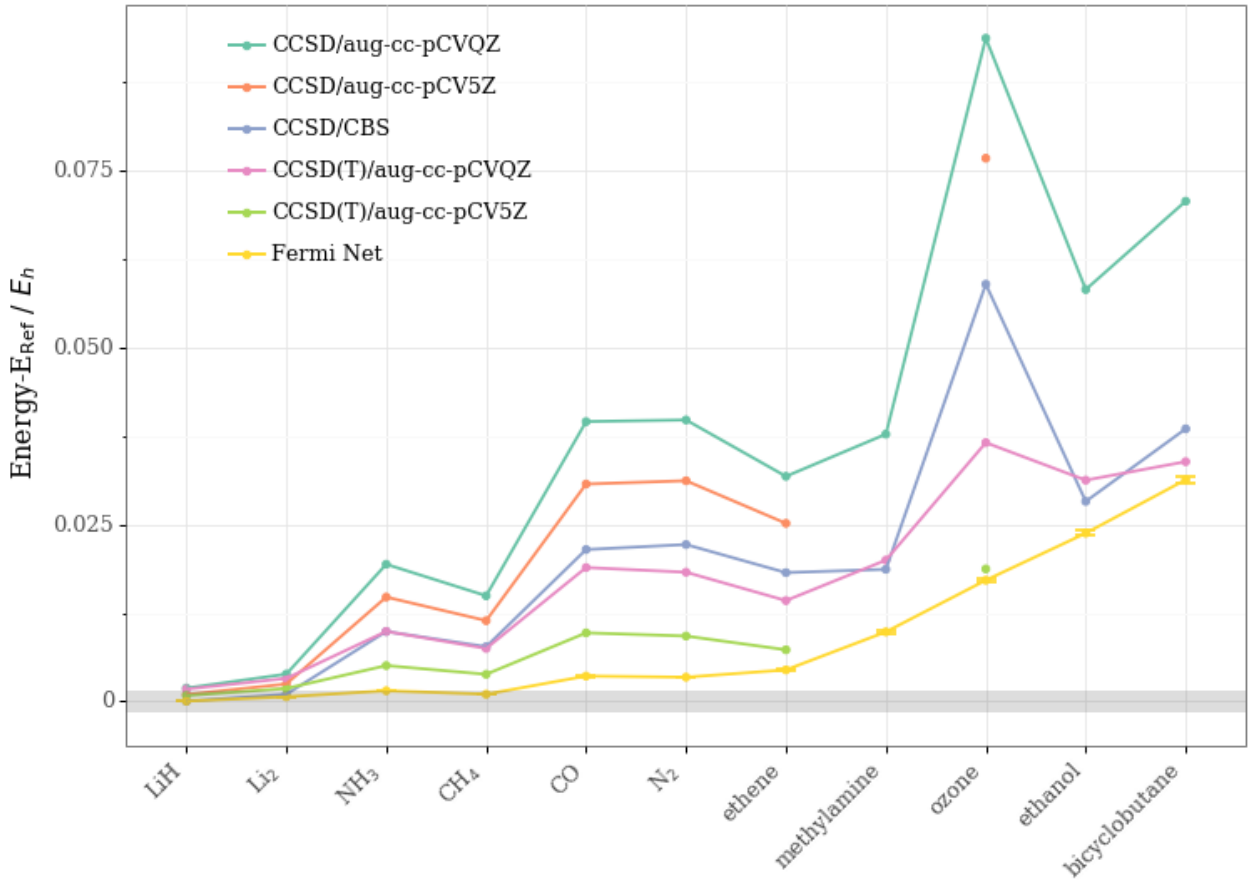
Atoms



- Fermi Net: 16 determinants.
- VMC, DMC: 50–100 configuration state functions. SJB wavefunction.
- Fermi Net outperforms conventional VMC and DMC.
- Consistently captures 99.7% of correlation energy.

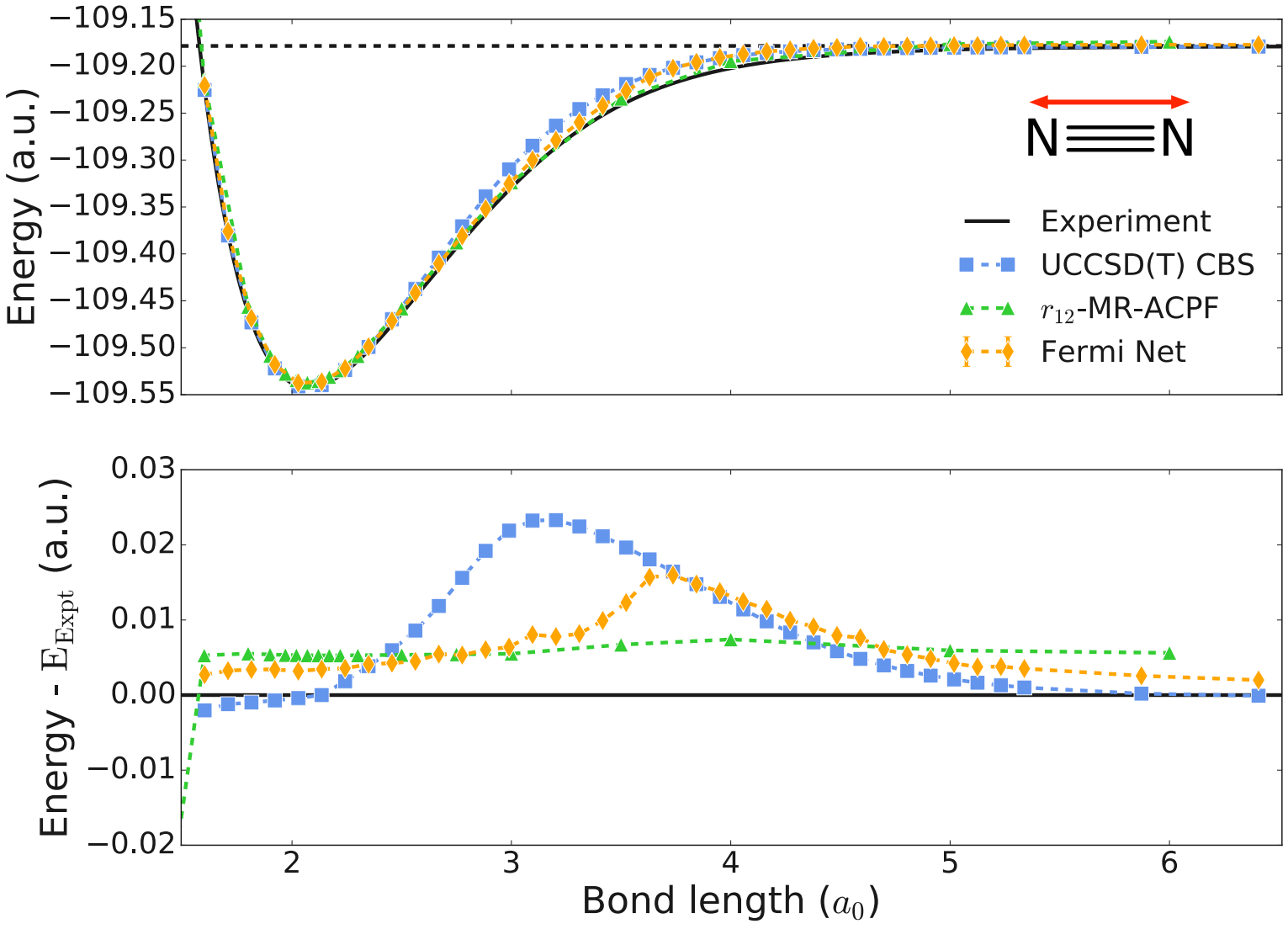
VMC, DMC: P. Seth, P.L. Ríos and R.J. Needs, J. Chem. Phys. **134**, 084105 (2011).
Exact: Chakravorty *et al.*, Phys. Rev. B **47**, 3649 (1993).

Molecules

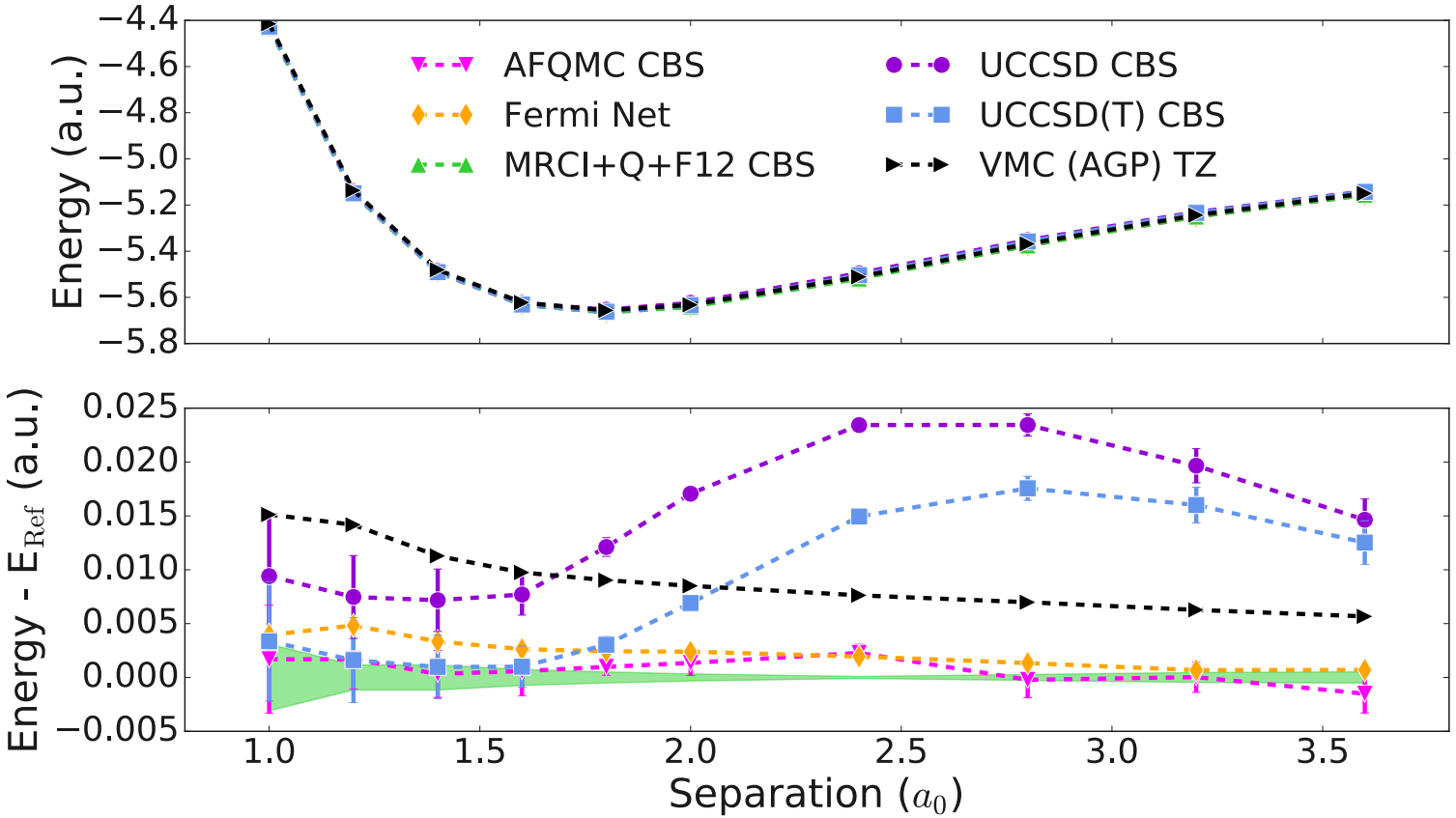


- Fermi Net outperforms CCSD(T) in QZ, 5Z basis sets.
- Accuracy degrades (smoothly) as number of electrons increases:
  - Fixed network configuration?
  - Optimisation?

N_2


−109.15
−109.20
−109.25
−109.30
−109.35
−109.40
−109.45
−109.50
−109.55
Energy (a.u.)
N≡N
Experiment
UCCSD(T) CBS
r_{12}-MR-ACPF
Fermi Net
0.03
0.02
0.01
0.00
−0.01
−0.02
Energy - E_{Expt} (a.u.)
2
3
4
5
6
Bond length (a_0)

H_{10}


AFQMC CBS
Fermi Net
MRCI+Q+F12 CBS
UCCSD CBS
UCCSD(T) CBS
VMC (AGP) TZ
Energy (a.u.)
−4.4
−4.6
−4.8
−5.0
−5.2
−5.4
−5.6
−5.8
Energy - E$_{\mathrm{Ref}}$ (a.u.)
0.025
0.020
0.015
0.010
0.005
0.000
−0.005
1.0
1.5
2.0
2.5
3.0
3.5
Separation (a_0)

Outline

Positives

- Fermi Net is *way* better than any other form of VMC applicable to real molecules with more than a couple of atoms.
- Fermi Net rivals coupled cluster for equilibrium geometries and outperforms it for molecules with a strong multi-determinant nature.
- The Fermi Net wavefunction can serve as a trial function for DMC and other projector methods.
- We have only just begin. Coupled cluster and SJ VMC have a fifty-year start.

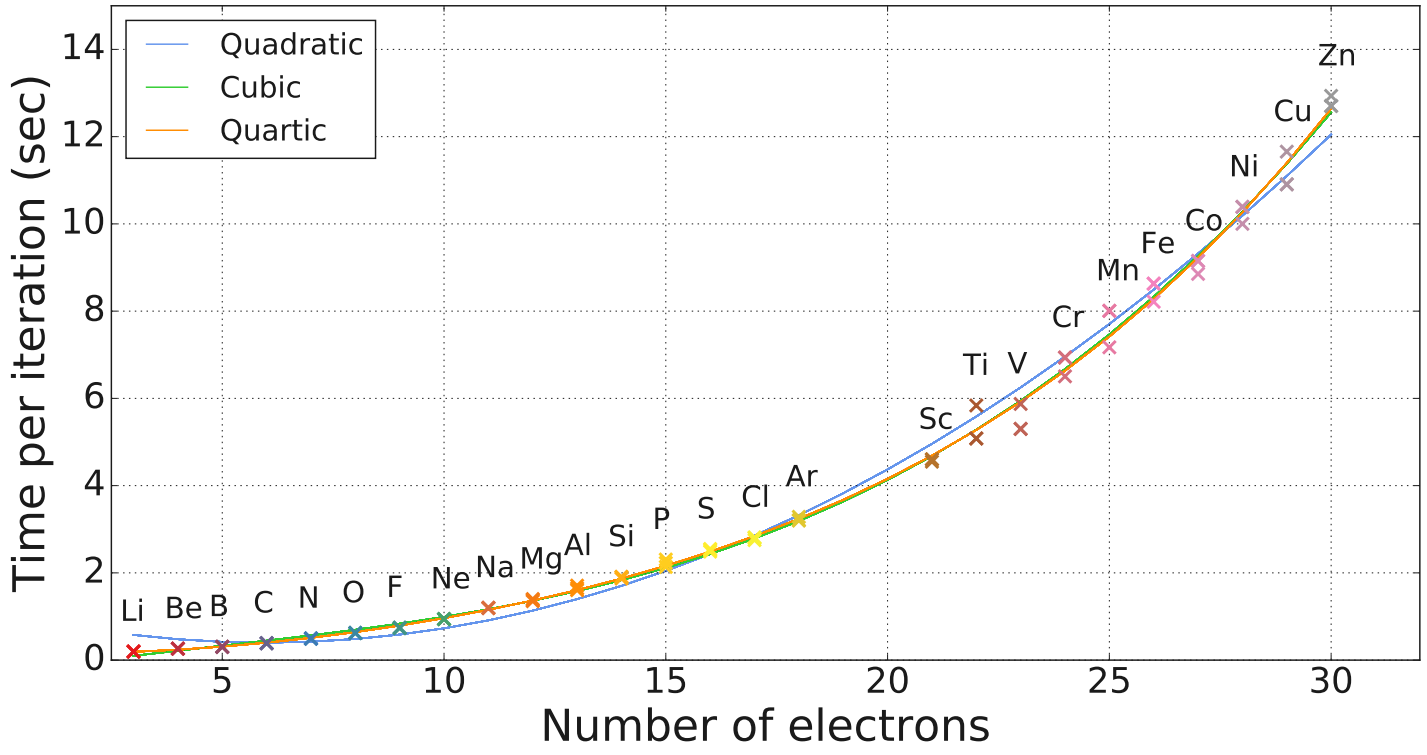
Questions

- How general is the Fermi Net wavefunction?
- Can we really get away with only a few determinants?
- Limits to accuracy?
- Solids?
- Size extensivity? (SJ VMC is extensive)
- Scaling? (SJ VMC scales very well)

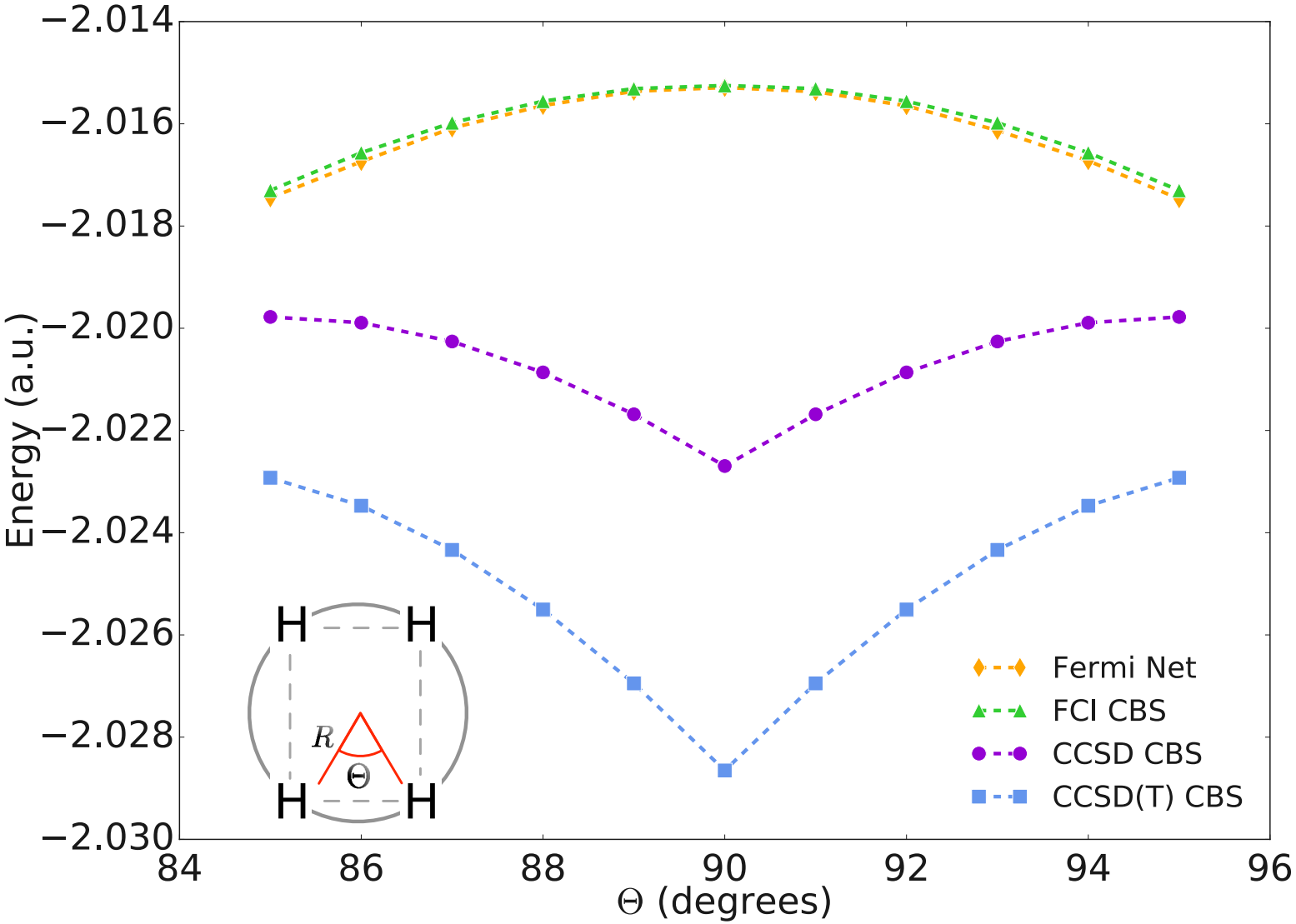
Range of Ideas and Approaches

	Systems	**Parameterisation**	**Optimisation**	**Accuracy**
NNB: Luo, Clark, PRL **122**, 226401 (2019)	Hubbard	backflow	first order	beats conventional backflow
PauliNet: Hermann *et al.*, arXiv:1909.08423	molecules continuum	Jastrow + backflow	ADAM	Boron: 97.3% c.e. H_{10}: 90–97% c.e.
NQS: Choo *et al.*, Nat. Commun. **11**, 2360 (2020)	molecules basis	map to spin system; approximate as RBM	stochastic reconfiguration	$<$ 1mH relative to FCI in STO-3G
Fermi Net: Pfau *et al.*, arXiv:1909.02487	molecules continuum	Everything	KFAC	Boron: 99.8% c.e. H_{10}: 98.5–99.3% c.e.

Scaling


Quadratic
Cubic
Quartic
Time per iteration (sec)
Number of electrons
0
2
4
6
8
10
12
14
5
10
15
20
25
30
Li
Be
B
C
N
O
F
Ne
Na
Mg
Al
Si
P
S
Cl
Ar
Sc
Ti
V
Cr
Mn
Fe
Co
Ni
Cu
Zn

H_4


−2.014
−2.016
−2.018
−2.020
−2.022
−2.024
−2.026
−2.028
−2.030
Energy (a.u.)
84
86
88
90
92
94
96
Θ (degrees)
H
H
H
H
R
Θ
Fermi Net
FCI CBS
CCSD CBS
CCSD(T) CBS