Highly Accurate Real-space Electron Densities with Neural Networks

Lixue Cheng,[1, a)] P. Bernát Szabó,[1, 2, a)] Zeno Schätzle,[1, 2, a)] Derk Kooi,[1] Jonas Köhler,[1] Klaas J. H. Giesbertz,[1] Frank Noé,[1, 2, b)] Jan Hermann,[1, c)] Paola Gori-Giorgi,[1, d)] and Adam Foster[1, e)]
[1)] *Microsoft Research, AI for Science*
[2)] *Freie Universität Berlin*

Variational ab-initio methods in quantum chemistry stand out among other methods in providing direct access to the wave function. This allows in principle straightforward extraction of any other observable of interest, besides the energy, but in practice this extraction is often technically difficult and computationally impractical. Here, we consider the electron density as a central observable in quantum chemistry and introduce a novel method to obtain accurate densities from real-space many-electron wave functions by representing the density with a neural network that captures known asymptotic properties and is trained from the wave function by score matching and noise-contrastive estimation. We use variational quantum Monte Carlo with deep-learning ansätze (deep QMC) to obtain highly accurate wave functions free of basis set errors, and from them, using our novel method, correspondingly accurate electron densities, which we demonstrate by calculating dipole moments, nuclear forces, contact densities, and other density-based properties.

I. INTRODUCTION

Variational ab-initio methods form the bedrock of quantum chemistry, be it the Hartree–Fock method and variational quantum Monte Carlo (VMC) in first quantization or various flavors of configuration interaction (CI) in second quantization.[1] They provide direct access to the wave function, unlike perturbation-based or projection-based methods. Access to the full wave function in theory allows the extraction of any observable of interest, over and above the electronic energy, but calculating such observables is often computationally involved. Consider for instance the extraction of real-space one- and two-electron densities from many-electron wave functions. With real-space wave functions, such as those yielded by VMC, one has to carefully design low-variance estimators for the density, and then sample those estimators individually at each point at which the density is to be determined, limiting resolution and further use.[2–5] With second-quantized wave functions, such as those yielded by CI, one is often severely limited by the accuracy of the basis set. The basis-set extrapolation techniques typically used for relative energies[6,7] are not available for electron densities, and Gaussian basis sets have built-in incorrect asymptotic behavior at the nuclei as well as far away from them. Non-variational methods (e.g. coupled cluster and Møller–Plesset) are popular to improve the accuracy of the total energy, but extracting observables from them are non-trivial as there is no unique approach to do so. For example, in coupled cluster we have the (cheap) unrelaxed density matrices and the inequivalent relaxed density via the costly λ-equations.[8]

In particular, the one-electron density $\rho(\mathbf{r})$ succinctly describes many electronic properties of matter, and it is a primary example of useful information that can be extracted from a many-electron wave function. It is the key object in density functional theory, forms the basis of bond classification,[9] and allows computation of other derived observables, such as electrostatic multipole moments, nuclear forces, or contact densities. Here, we address the existing issues in obtaining accurate one-electron densities and introduce a novel method for their extraction from real-space many-electron wave functions. The *neural electron real-space density* (NERD) approach retains the accuracy of the underlying wave-function method; once fitted, it provides the density everywhere in space at negligible computational cost, and is asymptotically correct by design both at the nuclei and far away from them. NERD represents the one-electron density as a neural network model with built-in physical constraints, and is fitted with a loss function composed of two terms on data consisting of spatial electron coordinates sampled from the target wave function. The first loss term uses score matching[10,11] to match gradients of the density model to gradients of the wave function. Score matching is a best-in-class algorithm for learning *local* features of the density, but it is known to struggle to model a density with multiple modes that are separated by large regions of low probability,[12] which we encounter in the case of molecular dissociation. To mitigate this, the second loss term uses noise-contrastive estimation,[13] which yields well-calibrated *global* features of the learned electron density, such as charges on dissociated molecular fragments.

NERD is only as accurate as the underlying wave functions and in this work we chose to couple it to wave functions from VMC with deep-learning ansätze (deep QMC). Deep QMC is a recent class of ab-initio electronic structure methods that model the many-electron wave function Ψ using a suitable neural network,[14–16] see Hermann *et al.*[17] for an in-depth review. Training this network relies only on the variational principle and requires no data as an input, as the electron configurations are generated by Monte Carlo sampling during training. This approach does not use a Gaussian basis, but instead learns a much more flexible neural network model of the

[a)] These authors contributed equally to this work.
[b)] Electronic mail: franknoe@microsoft.com
[c)] Electronic mail: jan.hermann@microsoft.com
[d)] Electronic mail: pgorigiorgi@microsoft.com
[e)] Electronic mail: adam.e.foster@microsoft.com

wave function, and as such offers one of the most accurate many-electron wave-function models in existence.

To analyze and verify the densities from NERD models, we first compare them on small molecules to densities from other methods such as full CI (FCI) in a large basis, showing that we capture the same general features but with improved asymptotics. We then show that for quantities involving *derivatives* of the density, our model is superior to Gaussian-based alternatives. We also demonstrate that a NERD model, once trained, delivers very accurate values for one-electron observables through numerical integration. Hence, with one single density model we have access not only to very accurate densities and density derivatives, but also to forces on the nuclei (via the Hellmann–Feynman theorem), spin-densities on the nuclei, dipole moments, etc. This is a crucial difference with respect to traditional QMC evaluation of these observables, which requires different dedicated estimators for each one of them.[2–5,18] Finally, we investigate the scalability and convergence rate of our density model using benzene as an example.

In summary, we present a method, dubbed NERD, to extract one-electron density from real-space many-electron wave functions at high fidelity, and show that it is possible to obtain highly accurate one-electron densities from deep-learning wave function ansätze. Figure 1 displays the schematic diagram of training NERDs.

II. PRELIMINARIES

A. Variational and deep quantum Monte Carlo

VMC is a method for approximately solving the Schrödinger equation by minimizing the energy expectation value of a parametrized ansatz. We focus on solving the time-independent, electronic Schrödinger equation of molecular systems in the Born–Oppenheimer approximation, defined by the position and charge of nuclei $I = 1, \dots, M$, denoted $\mathbf{R}_I, Z_I$.

A first-quantized, real-valued wave function ansatz parametrized by θ is a function $\Psi_\theta(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \in \mathbb{R}$, with $\mathbf{x}_i = (\mathbf{r}_i, \sigma_i)$ denoting the spatial-spin coordinates of electron i, that is constrained to be antisymmetric under the exchange of any two electrons. The parameters θ are trained variationally by minimizing the Rayleigh quotient

$$\theta' = \arg\min_\theta \frac{\langle \Psi_\theta | \hat{H} | \Psi_\theta \rangle}{\langle \Psi_\theta | \Psi_\theta \rangle}, \tag{1}$$

and the electron configurations $\mathbf{r}_i$ at which (1) and its derivatives are computed are generated via Monte Carlo sampling from the many-body density $|\Psi_\theta|^2$. After optimization, the ground state energy can be obtained as $E = \langle \Psi_{\theta'} | \hat{H} | \Psi_{\theta'} \rangle / \langle \Psi_{\theta'} | \Psi_{\theta'} \rangle$.

Traditional VMC uses ansätze built from Slater determinants of precomputed single-particle orbitals, and optimizes at most a few thousand wave-function parameters, typically in the form of a symmetric Jastrow factor. Deep QMC introduces wave functions parametrized by neural networks which may depend on millions of parameters and are trained with deep-learning techniques such as stochastic gradient descent. Neural network wave functions were first introduced by Carleo and Troyer.[19] In 2020, PauliNet and FermiNet were introduced as the first high-accuracy solutions for molecular ground states, using a first-quantized ab-initio neural network approach.[15,16] Various new network architectures and algorithmic improvements have since been made to further improve accuracy and efficiency.[20–28]

Most neural network wave functions for molecules employ generalized Slater determinants, that augment the single-particle orbitals of conventional Slater determinants with many-body correlation,

$$\Psi_{\boldsymbol{\theta}}(\mathbf{x}_{1:N}) = \textstyle\sum_p c_p \det[\mathbf{A}^p_\uparrow(\mathbf{x}_{1:N})] \det[\mathbf{A}^p_\downarrow(\mathbf{x}_{1:N})]\,, \tag{2}$$

$$A^p_{\sigma ik}(\mathbf{x}_{1:N}) = \phi^p_{\sigma k}(\mathbf{r}^\sigma_i, \{\mathbf{r}^\uparrow\}, \{\mathbf{r}^\downarrow\}) \times \varphi^p_{\sigma k}(\mathbf{r}^\sigma_i)\,, \tag{3}$$

where $\mathbf{x}_{1:N} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ and $\mathbf{r}_{1:N} = \{\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N\}$ denote the combination of all the spatial-spin coordinates and electron configurations, respectively. Here, electrons are grouped into (permutation invariant) sets of spin-up and spin-down electrons, $\{\mathbf{r}^\uparrow\}$, $\{\mathbf{r}^\downarrow\}$, $\sigma = \uparrow$ or $\downarrow$ to denote different groups of spins, $\phi^p_{\sigma k}$ are many-body spin orbitals, and $\varphi^p_{\sigma k}$ are single-particle spin envelopes that ensure the correct asymptotic behavior of the wave function with increasing distance from the nuclei. The ansatz may be a linear combination of multiple generalized Slater determinants, which are distinguished with the p index. The form of $\phi^p_{\sigma k}$ in (3) is closely related to the backflow transformation,[29,30] which introduces quasi-particles to obtain many-body orbital functions. The key observation motivating this choice of many-body orbitals is that the antisymmetry of Slater determinants constructed from many-body orbitals is preserved as long as the orbital functions are equivariant with the exchange of electrons.

For further details on VMC with neural network wave functions, see Hermann *et al.*[17] In all experiments of the present paper, the Psiformer wave-function architecture is used, the details of which are described in von Glehn, Spencer, and Pfau[28] .

B. Electron density

The electron spin-density is the one-body marginal of $|\Psi|^2$ normalized to the number of electrons N,

$$\rho(\mathbf{x}) = N \int \mathrm{d}\mathbf{x}_{2 \dots N} |\Psi(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N)|^2. \tag{4}$$

There are two known key properties of the electron density.

a. Kato's cusp condition Kato's cusp condition[31] states that for each nucleus, the electron density has

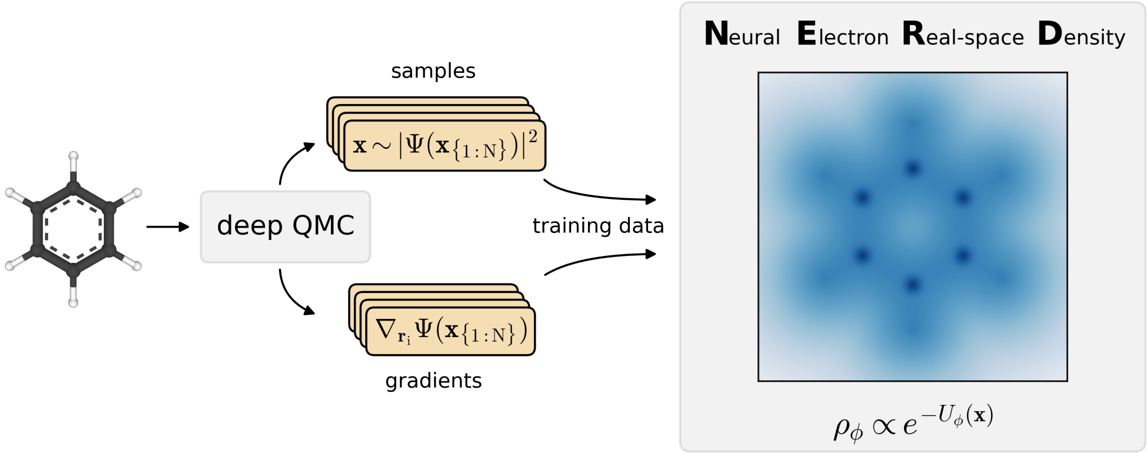
Neural Electron Real-space Density
samples
$\mathbf{x} \sim |\Psi(\mathbf{x}_{\{1:\mathrm{N}\}})|^2$
deep QMC
training data
$\nabla_{\mathbf{r}_\mathrm{i}}\Psi(\mathbf{x}_{\{1:\mathrm{N}\}})$
gradients
$\rho_\phi \propto e^{-U_\phi(\mathbf{x})}$

FIG. 1: Schematic diagram for training of NERDs. For an input molecular geometry a highly accurate wave function is obtained with deep QMC. Samples from the wave function and associated gradients serve as training data in a combined score matching and noise contrastive estimation approach to optimizing the NERD.

a non-differentiable 'cusp' at the point $\mathbf{R}_I$ such that[32]

$$\frac{1}{4\pi}\lim_{\epsilon\downarrow 0}\int_{\mathbf{r}\in S^2}\mathbf{r}\cdot\nabla\log\rho(\mathbf{R}_I+\epsilon\mathbf{r},\sigma) = -2Z_I. \tag{5}$$

b. Exponential decay and tail convergence Away from the nuclei, ρ should be differentiable[33,34] and the tail of the distribution decays exponentially. Specifically,

$$\log\rho(\mathbf{r},\sigma) \leq -2\kappa_\sigma\|\mathbf{r}\| + 2\beta_\sigma\log(1+\|\mathbf{r}\|) + \text{const.}, \tag{6}$$

where $\kappa_\sigma = \sqrt{2\mathrm{IP}(\sigma)}$ with $\mathrm{IP}(\sigma) > 0$ as the ionization potential and $\beta_\sigma = \left(\sum_I Z_I - N + 1\right)/\kappa_\sigma - 1$.[35,36] Unlike with Kato's cusp condition, $\mathrm{IP}(\sigma)$ requires calculation to obtain, so we treat this constraint as only guiding the functional form of the model.

C. Machine-learning energy based models

In machine learning, energy[37] based models (EBMs) are a highly flexible class of probability density models that focus on learning an approximation to the logarithm of the unnormalized density of a target distribution. An EBM with parameters ϕ over a variable $\mathbf{x}$ has a probability density

$$p_\phi(\mathbf{x}) = \frac{e^{-U_\phi(\mathbf{x})}}{\int e^{-U_\phi(\mathbf{x}')}d\mathbf{x}'}, \tag{7}$$

where U_ϕ is a real-valued function (e.g. a neural network). Due to the intractability of the normalizing constant in Eq. (7), naïve maximum likelihood training of EBMs is not possible. Numerous training approaches have been proposed,[38] including contrastive divergence,[39] score matching,[10,11] and noise-contrastive estimation.[13]

III. ELECTRON DENSITY ESTIMATION

A. Designing an EBM for the electron density

We propose an EBM approach to the problem of electron density estimation. EBMs are a promising model class for this problem because they do not place any direct constraints on the form of the density model, meaning that we can incorporate scientific priors as we see fit to match known asymptotic properties of the electron density.

Specifically, we propose an additive form for the unnormalised log-density

$$U_\phi(\mathbf{x}) = E_\phi(\mathbf{x}) + M_\phi(\mathbf{x}) + C(\mathbf{r}), \tag{8}$$

where E_ϕ is an envelope, M_ϕ is a simple multi-layer perceptron (MLP), and C controls the cusps around the nuclei.

For the envelopes, E_ϕ, we first define the smoothed distances $s_I = \sqrt{1+\|\mathbf{r}-\mathbf{R}_I\|^2}$. To capture the long distance exponential tails and to anchor the density around the nuclei, we use a sum of exponentially decaying envelopes

$$E_\phi(\mathbf{x}) = -\log\left(\sum_I \pi_I(\sigma)e^{-\zeta_I(\sigma)s_I}\right), \tag{9}$$

where π, ζ are learned parameters (we learn separate values for different spin channels).

Secondly, M_ϕ is a neural network. The input features to the network are defined in a manner inspired by the Psiformer, as a concatenation of four-dimensional featurelets, one for each nucleus

$$\text{input} = \text{concat}_I\begin{pmatrix}\log(1+s_I)\\ (\mathbf{r}-\mathbf{R}_I)\log(1+s_I)/s_I\end{pmatrix}. \tag{10}$$

The network then consists of a four-layer MLP with skip connections, SiLU activations[40] and with each hidden layer having size 64. There are two output heads for the up- and down-spin channels.

Finally, the cusp term is

$$C(\mathbf{r}) = 2\sqrt{\pi}\sum_I \operatorname{erf}\left(\tfrac{1}{2}Z_I\|\mathbf{r}-\mathbf{R}_I\|\right). \tag{11}$$

We can verify that this term will satisfy Kato's cusp condition in Eq. (5) (see Supporting Information Sec. S2). Since E_ϕ and M_ϕ are smooth everywhere, this also means that the model as a whole satisfies Kato's cusp condition.

B. Score matching

The general principle behind score matching is that, if we can train our model so that $\nabla U_\phi = -\nabla \log \rho$ everywhere, then $e^{-U_\phi} = \rho$ up to a normalizing constant. To achieve this, we could consider minimizing the loss

$$\mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})}\left[\|-\nabla_\mathbf{r}\log\rho(\mathbf{x}) - \nabla_\mathbf{r}U_\phi(\mathbf{x})\|^2\right], \tag{12}$$

however, this is impractical because it requires us to already know $\nabla \log \rho$. Instead, we can show that, starting from definition in Eq. (4), we can minimize a loss involving gradients of the *wave function*

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\|-2\nabla_{\mathbf{r}_i}\log|\Psi(\mathbf{x}_{1:N})| - \nabla_\mathbf{r}U_\phi(\mathbf{x}_i)\|^2\right], \tag{13}$$

where data is sampled from $\mathbf{x}_{1:N} \sim |\Psi(\mathbf{x}_{1:N})|^2$ by running Markov Chain Monte Carlo on the deep QMC wave function in the same manner as in QMC training. A full derivation is presented in Supporting Information Sec. S1B; the connection to force matching is discussed in Supporting Information Sec. S1C. Since we have access to the wave function, we do *not* require a noising–denoising approach and can therefore target the exact electron density.

Score matching is a best-in-class algorithm for learning *local* features of the density, however, it is known to struggle to model a distribution with multiple modes that are separated by a large region of low probability.[12] Because score matching is based only on the gradient of the log-density, there is only a very weak signal to correctly calibrate the relative masses of several far-separated modes.

C. Noise contrastive estimation

To correct the relative masses of separated modes, we seek a loss function that directly trains the values of the log-density. Noise contrastive estimation[13] (NCE) is one such approach. In NCE, we set up a synthetic classification problem to distinguish samples from the true distribution, ρ, and from a noise distribution, p_n, which has a known density. In our case, we sample the joint distribution $y, \mathbf{x} \sim p(y=1)\rho(\mathbf{x}) + p(y=0)p_n(\mathbf{x})$ where $p(y=1)$ is a free parameter that determines the fraction of true and noise samples. Samples from ρ are obtained by sampling $|\Psi|^2$. We then set up a 'classifier' model to predict y given $\mathbf{x}$ with probability

$$p_\phi(y=1\mid\mathbf{x}) = \frac{e^{-U_\phi(\mathbf{x})}}{e^{-U_\phi(\mathbf{x})}+\nu p_n(\mathbf{x})}, \tag{14}$$

where $\nu = p(y=1)/p(y=0)$, and then minimize the negative log-likelihood loss on our observed data

$$\mathbb{E}_{y,\mathbf{x}\sim p(y=1)\rho(\mathbf{x})+p(y=0)p_n(\mathbf{x})}[-\log p_\phi(y\mid\mathbf{x})]. \tag{15}$$

When the model is sufficiently powerful, the trained classifier will match the Bayes-optimal classifier, with

$$p(y=1\mid\mathbf{x}) = \frac{\rho(\mathbf{x})}{\rho(\mathbf{x})+\nu p_n(\mathbf{x})}, \tag{16}$$

requiring $e^{-U_\phi} = \rho$.

The success of NCE is known to depend heavily on the choice of noise distribution.[41] For example, p_n should be at least as heavy-tailed as the data. In the quantum chemistry context, we use the known exponential tails of the density, and the fact that electron density should cluster near nuclei, to define a suitable noise distribution. Our noise distribution, $p_n(\mathbf{x})$, is a weighted mixture of exponential-tailed distributions centered on atoms,

$$p_n(\mathbf{x}) \propto \sum_I Z_I e^{-\|\mathbf{r}-\mathbf{R}_I\|}. \tag{17}$$

Combining the score matching loss (Eq. (13)) with the NCE loss (Eq. (15)) gives a training mechanism to accurately capture both local and global features of the target density. Our overall loss function used during training is therefore

$$\ell = \ell_{\mathrm{SM}} + \lambda\ell_{\mathrm{NCE}}, \tag{18}$$

where we use $\lambda = 1$ unless otherwise stated.

IV. RESULTS

Detailed experimental settings for the QMC calculations and the NERD trainings are listed in Tables S1 and S2 (Supporting Information Sec. S9), respectively.

A. Small atoms

As the first set of experiments, we assess the quality of the deep QMC wave functions and subsequently that of the NERD models, on a set of small atoms ranging from He to Ne. First, the accuracy of the many-body wave functions is validated by comparing deep QMC ionization potentials (IPs) with reference experimental values.[42] The QMC ionization potential values are computed as $E_{\mathrm{IP}} = E_{\mathrm{cation}} - E_{\mathrm{atom}}$, where E_{atom} and E_{cation} are the energy expectation values of the trained Psiformer ansätze for the atom and cation, respectively. The computed Psiformer ionization potentials agree with experimental results extremely well, with percentage errors ranging from 0.029% to 0.37%, indicating the high accuracy of the corresponding wave functions.

TABLE I: Ionization potentials (IPs) of first row atoms from experimental results[42] and deep QMC calculations. All energy values are in the unit of eV.

Atom	Exp. IP	deep QMC IP	Error%
He	24.587	24.594	0.029%
Li	5.3917	5.3873	−0.082%
Be	9.3227	9.3576	0.374%
B	8.2980	8.3047	0.081%
C	11.260	11.238	−0.195%
N	14.534	14.563	0.200%
O	13.618	13.656	0.279%
F	17.423	17.438	0.086%
Ne	21.565	21.608	0.199%

TABLE II: Contact densities (densities at nuclei) from Slater-Jastrow VMC with a dedicated zero-variance-zero-bias estimator,[5] CCSD/aug-cc-pVQZ wave functions, and our NERD model. All the values are in atomic units (a.u.).

Atom	VMC ZVZB [5]	CCSD/aug-cc-pVQZ	NERD
He	—	3.3830	3.6503
Li	13.847	13.467	13.914
Be	35.540	34.616	35.625
B	72.242	70.423	72.517
C	128.16	124.96	128.76
N	207.02	201.99	207.66
O	312.75	305.80	314.24
F	450.64	439.79	452.64
Ne	622.50	608.61	624.49

NERD models are then trained for each atom using the verified Psiformer wave function. As a first quality indicator of NERD, we show that it yields accurate estimates for the electron densities at the position of the nuclei, also known as *contact density*. Contact density is linked to various experimental techniques for probing the local chemical environment of a nucleus, such as Mößbauer isomer shift in Mößbauer spectroscopy[43] and the Fermi-contact contribution to the hyperfine coupling in electron paramagnetic resonance spectroscopy.[44] In spite of its significance, obtaining accurate contact densities continues to be a challenging task in electronic structure theory.[4,5] Densities computed using theories involving Gaussian basis functions have notoriously large errors due to the incorrect cusp behavior of Gaussians. Traditionally, accurate contact densities are also difficult to extract from real-space QMC wave functions, prompting the development of several dedicated estimators.[4,5] Here we highlight that NERD yields accurate densities everywhere in one shot, including at the nuclei. In fact, as shown in Table II, the obtained NERD values are close to the contact densities obtained in Ref. 5 from a Slater-Jastrow VMC wave function with a dedicated zero-variance zero-bias (ZVZB) estimator. For comparison, we also report CCSD contact densities computed in-house with the relatively large basis set of aug-cc-pVQZ, which, as expected from the lack of nuclear cusp, consistently yields much lower values.

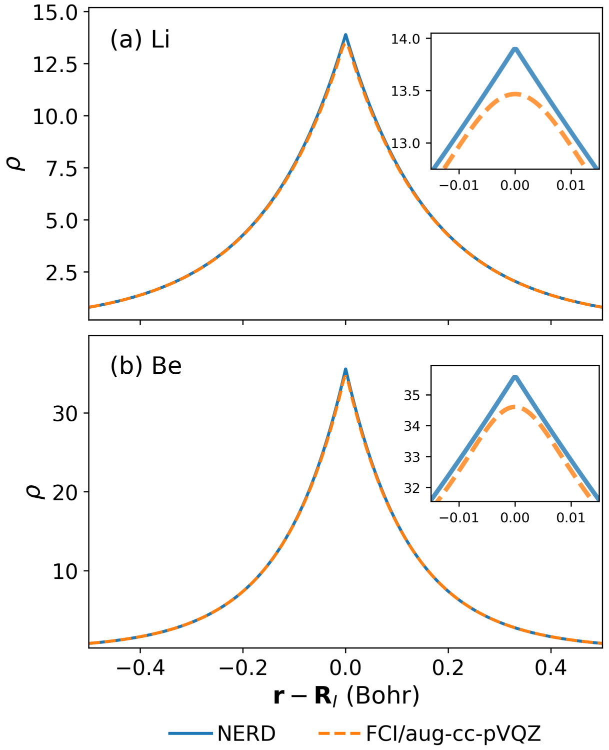

FIG. 2: Comparison between densities from the neural network model and FCI/aug-cc-pVQZ for (a) Li atom and (b) Be atom. Both nuclei are located at the origin. Inset plots are displayed to better visualize the cusps around nuclei.

Next, three key features of the obtained densities are observed at different radial distance scales. Without specification, all the displayed density are plotted along the x-axis in real-space (evaluated by fixing the y and z coordinates). First, the behaviour of NERD and FCI/aug-cc-pVQZ densities are compared close to the nucleus. Following Eq. (5), it is known that electronic densities should approach exponentials around the nuclear cusp, with a non-differentiable point at the nucleus. As is clear from the inset plots of Fig. 2, densities from NERD models for the Li and Be atoms have exactly this behaviour, while FCI/aug-cc-pVQZ densities exhibit smooth local maxima instead. This deficiency of the FCI/aug-cc-pVQZ densities is again a direct consequence of the atom-centered Gaussian basis functions, which are differentiable at the nucleus. On the other hand, the carefully crafted nuclear cusp term of Eq. (11) ensures the correct behaviour of the NERDs. In Supporting Information Sec. S7B, we additionally discuss the accuracy of spin-up and spin-down NERDs for Li and Be in Fig. S2.

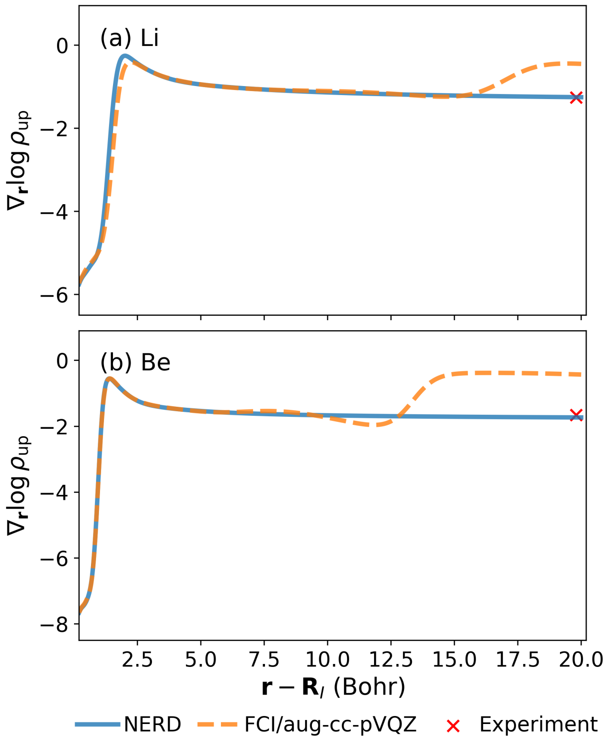

FIG. 3: Comparison between radial derivatives of log of spin-up densities from the neural network model, FCI/aug-cc-pVQZ and converting literature ionization potential values via Eq. (6) for (a) Li atom and (b) Be atom. Both nuclei are located at the origin.

Second, the correct exponential decay of the NERD model is highlighted in Fig. 3, showing $\nabla_{\mathbf{r}} \log \rho(r)$ as a function of r for the Li and Be atoms. According to Eq. (6), $\nabla_{\mathbf{r}} \log \rho(r)$ should approach the constant $-2\kappa_\sigma = -2\sqrt{2I_{\mathrm{P}}(\sigma)}$ when $r \to \infty$, marked with red crosses on Fig. 3. Comparing NERDs with those obtained from FCI/aug-cc-pVQZ calculations, it is clear that the former converges to the correct asymptotic value, while the latter does not. The incorrect behaviour of the FCI/aug-cc-pVQZ density is rooted in its use of Gaussian basis functions, with which the correct exponential decay cannot be expressed. In contrast, our NERD models are constructed using exponential envelopes, which enables them to correctly describe densities far away from nuclei.

Third, the reliability of the NERD models is investigated at intermediate distances from the nucleus. Figure 4 compares our NERD models with highly accurate references obtained with specialized FCI-based methods[45] for the Be and Ne atoms. At this intermediate range where we expect both methods to work well, QMC densities agree well with literature values, validating their soundness.

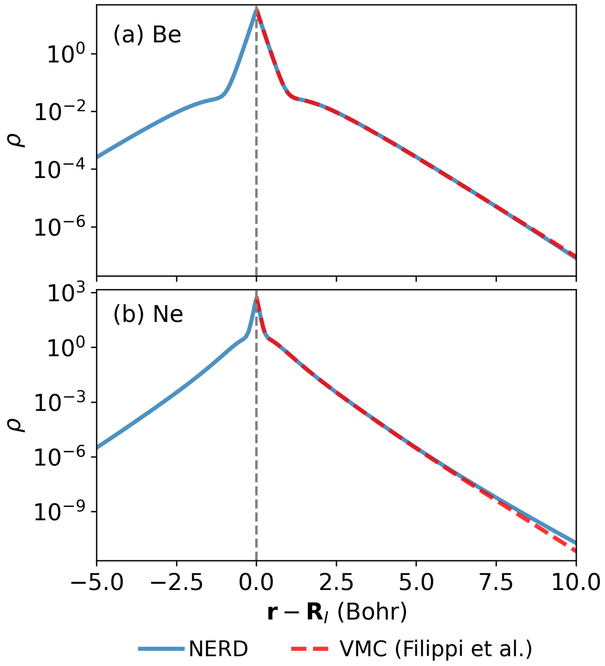

FIG. 4: Comparison between the total densities from the neural network model and literature for (a) Be atom and (b) Ne atom. Both nuclei are located at the origin and indicated by the gray dashed lines. The y-axis is plotted on a log scale. The literature densities are computed using VMC and obtained from Filippi et al.[45]

B. LiH dissociation

Next, the dissociation of the LiH molecule is investigated, by extracting QMC densities for a total of twelve molecular geometries, with bond lengths varying between 1.8 and 10 Bohr. Cuts from these densities along the bond axis are shown on the top panel of Fig. 5, using darker shades of blue for longer bond lengths, with the Li atom centered at the origin, and the H atom shifting gradually to the right. From this plot, one can immediately verify that the magnitudes of the density peaks around the two atoms remain constant across all bond lengths. This behavior is the result of including the NCE loss term during density fitting, ensuring the correct estimation of the relative magnitudes of separated density modes. The FCI/aug-cc-pVTZ density of LiH at its equilibrium geometry is shown on the top panel of Fig. 5 with the orange dashed line. It is in excellent agreement with the density extracted from QMC in the high density region around the nuclei, and only deviates from it appreciably in areas of low density, farther from the nuclei. This

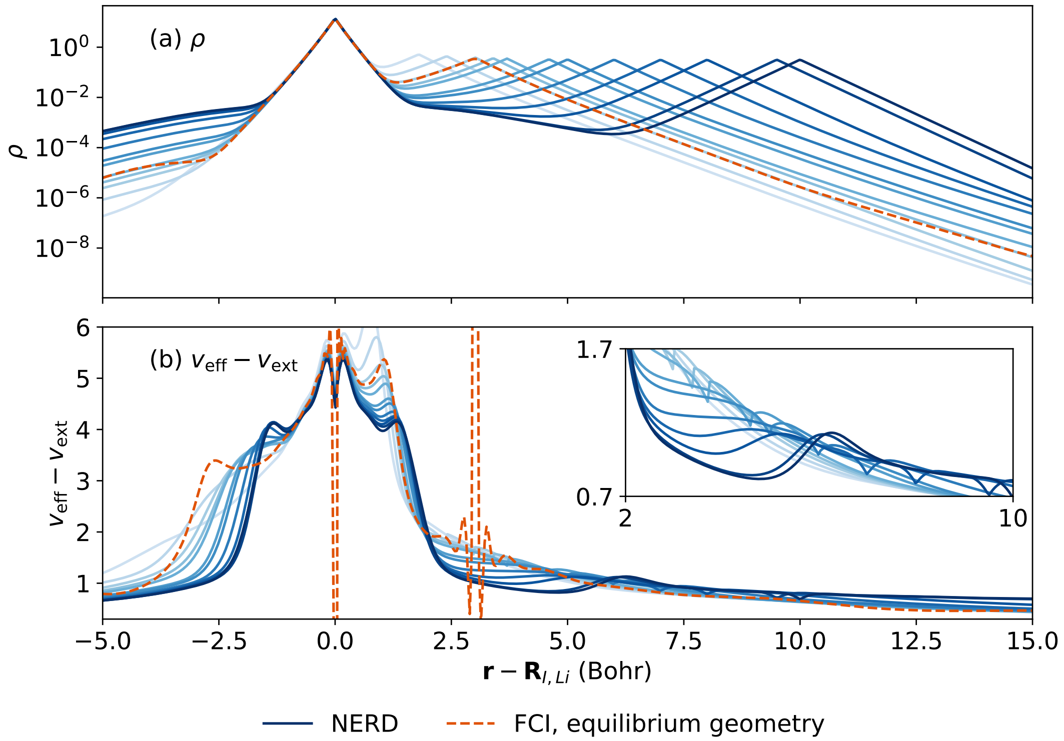

FIG. 5: (a) NERD and FCI/aug-cc-pVTZ densities along the dissociation of the LiH molecule, (b) effective potentials (see eq. (19)) extracted from the densities. The densities in (a) are plotted on a logarithm scale for y-axis. Darker shades of blue correspond to longer bond lengths, while the FCI/aug-cc-pVTZ calculation was carried out at equilibrium bond length only.

mismatch between the tails of the two densities can again be attributed to the use of a Gaussian basis set in the FCI calculation, with which the correct, exponentially decaying tails cannot be expressed.

To further assess the quality of the NERDs, they are used to compute the so called effective potential

$$v_{\text{eff}}(\mathbf{x}) = \frac{1}{8}\|\nabla_{\mathbf{r}} \log \rho(\mathbf{x})\|^2 + \frac{1}{4}\nabla^2_{\mathbf{r}} \log \rho(\mathbf{x}). \qquad (19)$$

This potential is an important quantity appearing in the effective Schrödinger equation for the square root of the density.[46] More importantly for the present work it exhibits some well known, yet traditionally hard-to-capture features during the dissociation of LiH, the proper description of which constitutes an exceedingly high benchmark of accuracy for NERD models.[47,48] Among these features is the formation of a step with a slight overshoot between the two atoms as the bond length increases.

The bottom panel of Fig. 5 depicts the effective potential extracted from QMC densities at various bond lengths, with the contribution of the external potential (Coulomb potential of nuclei) removed. The exact numerical recipe for accurately extracting the $v_{\text{eff}} - v_{\text{ext}}$ potential from our NERD models is described in Supporting Information Sec. S3. Due to the carefully selected functional form of our model, the obtained effective potential precisely cancels the diverging external potential at the nuclear cusps, resulting in the bounded, continuous blue curves on the bottom panel of Fig. 5. This is in stark contrast to the effective potential derived from the FCI/aug-cc-pVTZ density at the equilibrium bond length, shown in orange. Due to the Gaussian basis set, there are subtle errors in FCI density close to the nuclei, therefore the derived effective potential cannot exactly cancel the external potential, leading to oscillations and divergences in the dashed orange curve at 0 and 3 Bohr.

The inset plot in the bottom panel of Fig. 5 highlights the region between the two atoms and shows the gradual formation of the expected step feature as the bond length increases. The correct description of this elusive feature is made possible by the precisely reproduced exponential decays of the two density modes, and would therefore be extremely challenging using Gaussian basis sets.

We additionally verify the accuracy of NERDs by computing the dipole moments, $\boldsymbol{\mu}$, from densities and Psiformer wave functions as follows:

$$\boldsymbol{\mu} = \int \mathbf{r}\rho(\mathbf{x})\, d\mathbf{x}, \qquad (20)$$

Figure 6 displays the computed dipole moments for the same LiH systems assessed in Fig. 5. Evaluations from NERD models, evaluations from Psiformer wave functions,

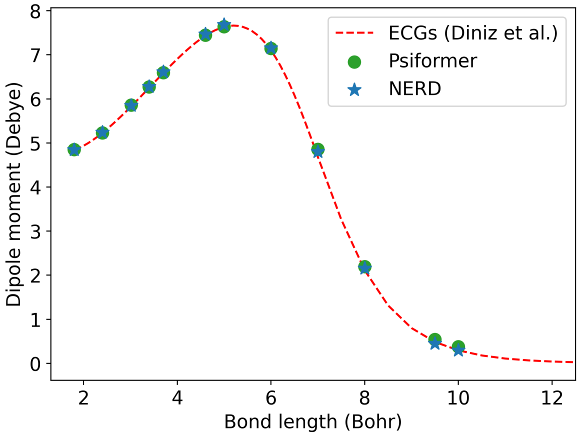

FIG. 6: Dipole moments along the dissociation of the LiH molecule. Results evaluated from NERDs are compared with the direct evaluation of the underlying Psiformer wave functions and accurate reference calculations of Diniz *et al.*[49] based on explicitly correlated Gaussians (ECGs) with shifted centers.

and the highly-accurate calculations with all-particle explicitly correlated Gaussian functions with shifted centers (ECGs) from Ref. 49, have excellent agreement along the entire dissociation curve. This indicates the ability of NERD models to accurately yield other density-based properties.

C. H$_4$

Deep QMC is known to produce highly accurate results on systems with strong multi-reference character,[17] such as the H$_4$ molecule in a square configuration.[50,51] To highlight the advantages of QMC densities for these systems, Fig. 7 displays planar cuts of H$_4$ densities obtained from (a) a Psiformer wave function with our NERD model, and (b) a CCSD computation in the aug-cc-pVQZ basis. It is clear that single-reference CCSD yields an incorrect, asymmetric density, as a result of randomly picking one of the two equally important Slater-determinants as its reference determinant. On the other hand, the density obtained from QMC is perfectly symmetric, verifying the ability of deep QMC to model multi-reference systems and validating the correctness of NERD model.

D. Force estimation on diatomic systems

Here we demonstrate that our highly accurate score matching densities can be used to obtain quantitatively correct Hellmann–Feynman forces.[53,54] To that end, we evaluate energies and Hellmann–Feynman forces for a range of bond lengths of various homonuclear diatomic molecules around their equilibrium geometries. Specifically, we show that (a) accurate forces can be computed efficiently from the density, (b) these forces match the Monte Carlo estimates from the wave function used for density extraction, and (c) bias correction schemes accounting for deficiencies in the wave function are not required.

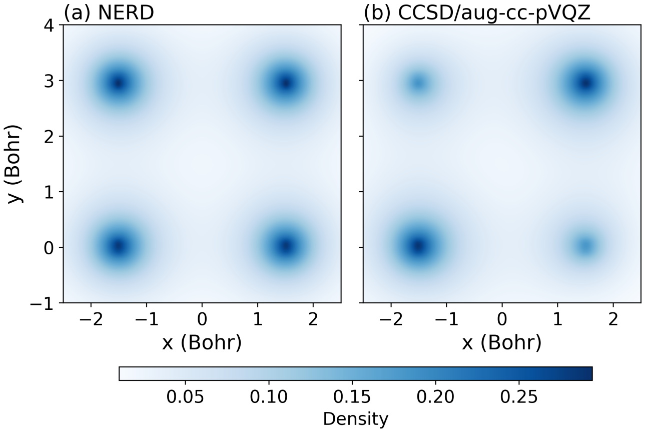

FIG. 7: Comparison of spin-up densities computed from (a) NERD model using Psiformer wave function and (b) CCSD/aug-cc-pVQZ on a square H$_4$ molecule with bond lengths of 3.0236 Bohr

Similar to other observables, the Hellmann–Feynman forces from the NERD models are obtained by numerically evaluating the expectation value integrals on a grid (Eq. S42 in Supporting Information). While the three-dimensional electron density allows for such treatment, integrals of the $3N$ dimensional wave function have to be estimated through Monte Carlo integration. Due to the prohibitively high variance of the bare Monte Carlo estimator for nuclear forces, we compare the results with the improved zero-variance Assaraf-Caffarel (AC-ZV) estimator[18] (see Supporting Information Sec. S6). We also test the zero-variance zero-bias Assaraf–Caffarel (AC-ZVZB) estimator (see Supporting Information Sec. S6), which corrects for a potential systematic error due to discrepancies of the wave function from the true ground state, at the cost of additional local energy evaluations. Additionally, we compute finite difference forces based on the energy expectation values of our Psiformer wave functions. This serves as an additional validation of the quality of the wave function in one dimension, but is not a viable strategy on higher dimensional potential energy surfaces. To benchmark the forces against reference results obtained with QMC,[52] we chose the three systems investigated by Qian *et al.*[52] i.e. H$_2$, Li$_2$, and N$_2$.

For all three tested systems, the results presented in Fig. 8 show that the energies obtained with deep QMC are very close to the reference energies and both the forces extracted from the NERD model as well as the Monte Carlo estimates from the wave function are in good agreement with the reference. Firstly, we find that the integration of the Hellmann–Feynman force on the grid (see Sec. S5 in Supporting Information) gives a good, numerically stable

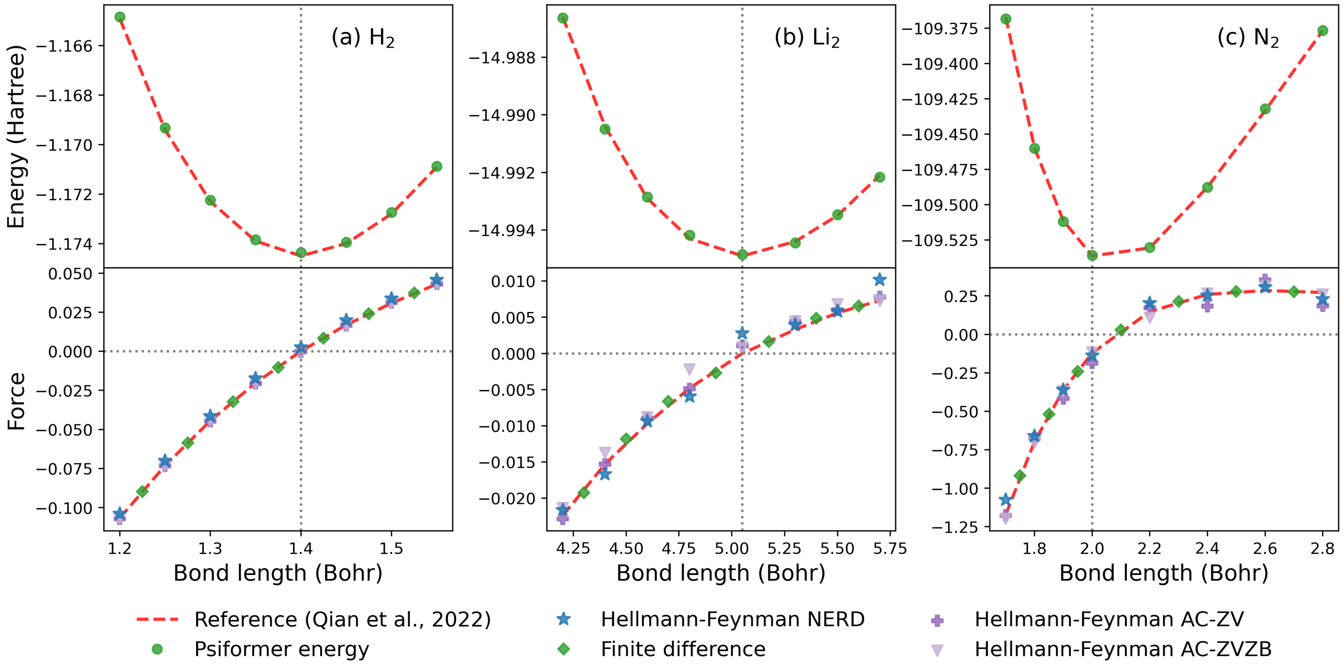

FIG. 8: Estimating the inter-atomic force for three diatomic systems. The top panels in (a)-(c) show the QMC energies computed from Psiformer using the settings described in Table S1 in Supporting Information as well as reference energies of Qian *et al.*[52] obtained with FermiNet using 100k training steps for H_2, and 200 k training steps for Li_2 and N_2, respectively. In the lower panels, we compare (1) reference forces using zero-variance Assaraf-Caffarel estimator (AC-ZV) estimator evaluated on FermiNet[52] (2) integration of the Hellmann-Feynman force from the density (at 40k density fitting steps), (3) finite difference from QMC energies (4) forces using AC-ZV estimator evaluated on Psiformer (10k evaluation steps), and (5) forces using AC-ZVZB estimator evaluated on Psiformer (10k evaluation steps).

estimate of the force at low computational costs. Comparing the forces obtained from integrating the density with the direct extraction from the underlying wave function shows excellent agreement, further validating the accuracy of the density fitting procedure. We attribute the remaining differences to noise in the density fitting and the stochastic evaluation of the Hellmann–Feynman forces with Monte Carlo. In fact, investing compute in fitting the density and obtaining forces from the NERD model can be considered a workable alternative to the direct force estimation. The results moreover indicate that the high accuracy of deep QMC wave functions removes the need for bias correction, due to a generally good agreement of the uncorrected forces from both the density as well as the wave function and the reference. To further investigate the potential benefits of bias correction we evaluate the AC-ZVZB estimator on the wave functions. We find that in our experimental setting the bias correction of the AC-ZVZB estimator may not be worth the high cost of additional local energy evaluations.

In conclusion, this experiment shows that our density models provide high quality forces, exemplifying the utility of density extraction and validating the accuracy of the density fitting procedure. Remarkably, fitting the density on samples from the wave function and evaluating integrals on grids regularizes the high variance of the bare force estimator and gives results in good agreement with the more advanced estimators introduced by Assaraf and Caffarel. Furthermore, a single NERD model suffices to evaluate many observables at a negligible cost compared to the direct estimation of each of them from the wave function by the means of Monte Carlo integration.

E. Benzene

We finally investigate the scalability and convergence properties of our method on the benzene molecule (with 42 electrons) at its equilibrium geometry. To assess the convergence of our model, we trained five NERD models with different seeds against the same Psiformer checkpoint. At various stages of density training, we computed several metrics and the variance between seeds. To check of the rotation invariance of the density (Fig. 9a), we compute the total variation (TV) distance[55] between ρ and 60°-rotated ρ, which should be zero as the ground state density is invariant under the D_{6h} point group. TV distance is a commonly used metric in machine learning to quantify the differences between two probability distributions and its definition is in Supporting Information Eq. S41. We also check on the spin-symmetry of the density (Fig. 9b) by computing the TV distance between spin-up and spin-

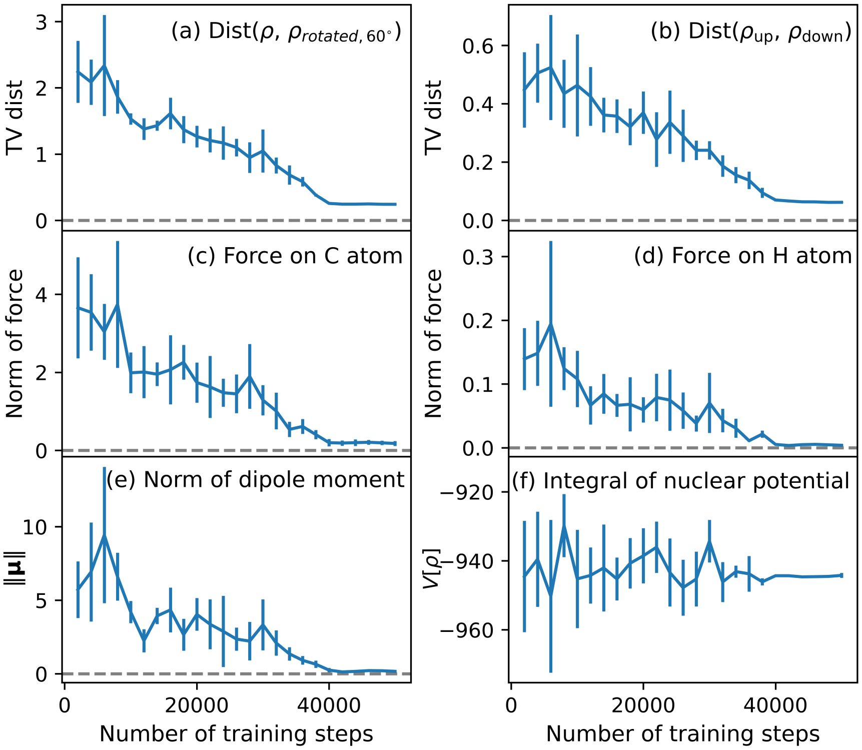

FIG. 9: Evaluating the convergence of our NERD model using various metrics on a single Psiformer wave function on the equilibrium benzene. (a) TV distance between ρ and 60°-rotated ρ (b) TV distance between densities of two spins, (c) Hellmann-Feynman force on C, (d) Hellmann-Feynman force on H, (e) Norm of molecular dipole moment, and (f) Integral of nuclear potential. Along the training processes, all the values are computed by averaging results from 5 independent trainings and the error bars are 2 standard deviations of the corresponding means. The gray dashed lines represent the ground truth values. Note that the model learning rate is decayed during training, see Sec. S9 in Supporting Information

down densities, $\mathrm{TV}(\rho_{\mathrm{up}}, \rho_{\mathrm{down}})$. This should also be zero for the ground state. Another metric to check is Hellmann-Feynman force[53,54] on one certain atom, which could be computed by Eq. S42 in Supporting Information. Figure 9c and 9d display the Hellmann-Feynman force on a carbon and a hydrogen atom, respectively, along the training processes. Both values converge to nearly 0 as expected because the molecule is at its experimental equilibrium geometry. The dipole moment (Eq. 20) should also be zero due to the overall symmetry of the system as shown in Fig. 9e. Finally, we compute the nuclear potential by Eq. S43 in Supporting Information to assess the reduction in variance during training (Fig. 9f). For all the 4 metrics we assessed, as training progresses, the variances between different seeds decrease close to 0 and the values converge close to their theoretical values after 40 000 training steps. This investigation suggests that our NERD model is accurate and stable with a reasonable convergence speed even for larger systems.

V. CONCLUSION

In this study, we propose an approach, NERD, to extract electronic densities from wave functions computed using real-space quantum chemistry methods via score matching and noise contrastive estimation techniques. We first validate the quality of the deep QMC wave functions from Psiformer by showing the agreements between the ionization potential computed from Psiformer energies and experimental values for small atoms. Then, the significant features of the densities, including Kato's cusp condition, tail convergences, and reliability of the densities across the entire space, are also systematically checked using Li, Be and Ne, as examples. We also investigate the performances of our NERDs and dipole moments computed from NERD models along the dissociation curves of LiH. The curvature of FCI densities is incorrect around nuclei while our NERDs provide the correct solutions. The quality of our NERD is additionally verified by the resulting physical solutions for two symmetric systems, i.e., H_4. The NERD models are also applied to compute Hellmann-Feynman forces accurately along the H_2, Li_2, and N_2 dissociation curves. Finally, we show the scalability of NERD by evaluating densities and density-related properties of benzene. As a general method, NERD approach paves a way to obtain high-quality densities from any wave function computed by real-space electronic structure theory.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENT

We thank Professor C. J. Umrigar for providing us the VMC densities of Be and Ne atoms and Prof. Carlos F. Bunge for providing us the FCI densities with 8s up to k basis functions of Li. We also appreciate the useful discussions from Dr. Rianne van den Berg.

[1] L. Piela, *Ideas of Quantum Chemistry*, 2nd ed. (Elsevier, 2014).

[2] R. Assaraf, M. Caffarel, and A. Scemama, "Improved Monte Carlo estimators for the one-body density," Phys. Rev. E **75**, 035701 (2007).

[3] D. Varsano, M. Barborini, and L. Guidoni, "Kohn-Sham orbitals and potentials from quantum Monte Carlo molecular densities," J. Chem. Phys. **140**, 054102 (2014).

[4] P. Håkansson and M. Mella, "Efficient and robust quantum Monte Carlo estimate of the total and spin electron densities at nuclei," J. Chem. Phys. **129**, 124101 (2008).

[5] M. C. Per, I. K. Snook, and S. P. Russo, "Zero-variance zero-bias quantum Monte Carlo estimators for the electron density at a nucleus," J. Chem. Phys. **135**, 134112 (2011).

[6] D. G. Truhlar, "Basis-set extrapolation," Chem. Phys. Lett. **294**, 45–48 (1998).

[7] A. Karton, N. Sylvetsky, and J. M. L. Martin, "W4-17: A diverse and high-confidence dataset of atomization energies for benchmarking high-level electronic structure methods," J. Comput. Chem. **38**, 2063–2075 (2017).

[8] S. A. Kucharski and R. J. Bartlett, "Noniterative energy corrections through fifth-order to the coupled cluster singles and doubles method," J. Chem. Phys. **108**, 5243–5254 (1998).

[9] B. Silvi and A. Savin, "Classification of chemical bonds based on topological analysis of electron localization functions," Nature **371**, 683–686 (1994).

[10] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching," J. Mach. Learn. Res. **6** (2005).

[11] P. Vincent, "A connection between score matching and denoising autoencoders," Neural Comput. **23**, 1661–1674 (2011).

[12] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, Vol. 32, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019).

[13] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (JMLR Workshop and Conference Proceedings, 2010) pp. 297–304.

[14] J. Han, L. Zhang, and W. E, "Solving many-electron schrödinger equation using deep neural networks," J. Comput. Phys. **399**, 108929 (2019).

[15] J. Hermann, Z. Schätzle, and F. Noé, "Deep-neural-network solution of the electronic schrödinger equation," Nat. Chem. **12**, 891–897 (2020).

[16] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes, "Ab initio solution of the many-electron schrödinger equation with deep neural networks," Phys. Rev. Res. **2**, 033429 (2020).

[17] J. Hermann, J. Spencer, K. Choo, A. Mezzacapo, W. M. C. Foulkes, D. Pfau, G. Carleo, and F. Noé, "Ab initio quantum chemistry with neural-network wavefunctions," Nat. Rev. Chem. **7**, 692–709 (2023).

[18] R. Assaraf and M. Caffarel, "Zero-variance zero-bias principle for observables in quantum monte carlo: Application to forces," J. Chem. Phys. **119**, 10536–10552 (2003).

[19] G. Carleo and M. Troyer, "Solving the quantum many-body problem with artificial neural networks," Science **355**, 602–606 (2017).

[20] J. S. Spencer, D. Pfau, A. Botev, and W. M. C. Foulkes, "Better, faster fermionic neural networks," (2020), arXiv:2011.07125 [physics.comp-ph].

[21] Z. Schätzle, P. B. Szabó, M. Mezera, J. Hermann, and F. Noé, "DeepQMC: An open-source software suite for variational optimization of deep-learning molecular wave functions," J. Chem. Phys. **159**, 094108 (2023).

[22] N. Gao and S. Günnemann, "Neural pfaffians: Solving many many-electron schrödinger equations," arXiv preprint arXiv:2405.14762 (2024).

[23] R. Li, H. Ye, D. Jiang, X. Wen, C. Wang, Z. Li, X. Li, D. He, J. Chen, W. Ren, *et al.*, "A computational framework for neural network-based variational monte carlo with forward laplacian," Nat. Mach. Intell. **6**, 209–219 (2024).

[24] L. Gerard, M. Scherbela, P. Marquetand, and P. Grohs, "Gold-standard solutions to the schrödinger equation using deep learning: How much physics do we need?" in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 10282–10294.

[25] J. Lin, G. Goldshlager, and L. Lin, "Explicitly antisymmetrized neural network layers for variational monte carlo simulation," J. Comput. Phys. **474**, 111765 (2023).

[26] J. Kim, G. Pescia, B. Fore, J. Nys, G. Carleo, S. Gandolfi, M. Hjorth-Jensen, and A. Lovato, "Neural-network quantum states for ultra-cold fermi gases," Commun. Phys. **7**, 148 (2024).

[27] X. Li, C. Fan, W. Ren, and J. Chen, "Fermionic neural network with effective core potential," Phys. Rev. Res. **4**, 013021 (2022).

[28] I. von Glehn, J. S. Spencer, and D. Pfau, "A self-attention ansatz for ab-initio quantum chemistry," arXiv preprint arXiv:2211.13672 (2022).

[29] R. P. Feynman and M. Cohen, "Energy spectrum of the excitations in liquid helium," Phys. Rev. **102**, 1189–1204 (1956).

[30] D. M. Ceperley, "Fermion nodes," J. Stat. Phys. **63**, 1237–1267 (1991).

[31] T. Kato, "On the eigenfunctions of many-particle systems in quantum mechanics," Commun. Pure Appl. Math. **10**, 151–177 (1957).

[32] We express these conditions in terms of $\nabla \log \rho$ as it gives the cleanest comparison with the density models that we introduce in Section III.

[33] M. Fournais, Sørenand Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and T. Østergaard Sørensen, "The electron density is smooth away from the nuclei," Commun. Math. Phys. **228**, 401–415 (2002).

[34] M. Fournais, Sørenand Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and T. Østergaard Sørensen, "Analyticity of the density of electronic wavefunctions," Ark. Mat. **42**, 87–106 (2004).

[35] M. Hoffmann-Ostenhof and T. Hoffmann-Ostenhof, ""Schrödinger inequalities" and asymptotic behavior of the electron density of atoms and molecules," Phys. Rev. A **16**, 1782–1785 (1977).

[36] R. Ahlrichs, M. Hoffmann-Ostenhof, T. Hoffmann-Ostenhof, and J. D. Morgan, "Bounds on the decay of electron densities with

screening," Phys. Rev. A **23**, 2106–2117 (1981).

[37] The term 'energy' in EBMs is not connected to the energy in quantum chemistry that arises as an eigenvalue of the Hamiltonian. The 'energy' in our setting is equivalent to $-\log \rho$ and thus we use a notation based on $U_\phi \approx -\log \rho + C$.

[38] Y. Song and D. P. Kingma, "How to train your energy-based models," arXiv preprint arXiv:2101.03288 (2021).

[39] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," Neural Comput. **14**, 1771–1800 (2002).

[40] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," arXiv preprint arXiv:1606.08415 (2016).

[41] O. Chehab, A. Gramfort, and A. Hyvärinen, "The optimal noise in noise-contrastive learning is not what you think," in *Uncertainty in Artificial Intelligence* (PMLR, 2022) pp. 307–316.

[42] NIST, "Atomic Spectra Database - Ionization Energies Form — physics.nist.gov," `https://physics.nist.gov/PhysRefData/ASD/ionEnergy.html`, [Accessed 15-05-2024].

[43] M. Filatov, W. Zou, and D. Cremer, "Analytic calculation of contact densities and mössbauer isomer shifts using the normalized elimination of the small-component formalism," J. Chem. Theory Comput. **8**, 875–882 (2012).

[44] E. D. Hedegard, J. Kongsted, and S. P. Sauer, "Validating and analyzing epr hyperfine coupling constants with density functional theory," J. Chem. Theory Comput. **9**, 2380–2388 (2013).

[45] C. Filippi, X. Gonze, and C. Umrigar, "Generalized gradient approximations to density functional theory: comparison with exact results," arXiv preprint cond-mat/9607046 (1996).

[46] G. Hunter, "Conditional probability amplitudes in wave mechanics," Int. J. Quantum Chem. **9**, 237–242 (1975).

[47] O. V. Gritsenko and E. J. Baerends, "Effect of molecular dissociation on the exchange-correlation kohn-sham potential," Phys. Rev. A **54**, 1957 (1996).

[48] E. J. Baerends and O. V. Gritsenko, "A quantum chemical view of density functional theory," J. Phys. Chem. A **101**, 5383–5403 (1997).

[49] L. G. Diniz, N. Kirnosov, A. Alijah, J. R. Mohallem, and L. Adamowicz, "Accurate dipole moment curve and non-adiabatic effects on the high resolution spectroscopic properties of the lih molecule," J. Mol. Spectrosc. **322**, 22–28 (2016).

[50] M. Motta, D. M. Ceperley, G. K.-L. Chan, J. A. Gomez, E. Gull, S. Guo, C. A. Jiménez-Hoyos, T. N. Lan, J. Li, F. Ma, *et al.*, "Towards the solution of the many-electron problem in real materials: Equation of state of the hydrogen chain with state-of-the-art many-body methods," Phys. Rev. X **7**, 031059 (2017).

[51] N. C. Rubin, R. Babbush, and J. McClean, "Application of fermionic marginal constraints to hybrid quantum algorithms," New J. Phys. **20**, 053020 (2018).

[52] Y. Qian, W. Fu, W. Ren, and J. Chen, "Interatomic force from neural network based variational quantum monte carlo," J. Chem. Phys. **157** (2022), https://doi.org/10.1063/5.0112344.

[53] R. P. Feynman, "Forces in molecules," Phys. Rev. **56**, 340 (1939).

[54] P. Politzer and J. S. Murray, "The Hellmann-Feynman theorem: a perspective," J. Mol. Model. **24**, 1–7 (2018).

[55] A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer, New York, 2009).

Supporting Information for Highly Accurate Real-space Electron Densities with Neural Networks

Lixue Cheng,[1, a)] P. Bernát Szabó,[1, 2, a)] Zeno Schätzle,[1, 2, a)] Derk Kooi,[1] Jonas Köhler,[1] Klaas Giesbertz,[1] Frank Noé,[1, 2, b)] Jan Hermann,[1, c)] Paola Gori-Giorgi,[1, d)] and Adam Foster[1, e)]

[1)] *Microsoft Research, AI for Science*

[2)] *Freie Universität Berlin*

S1. SCORE MATCHING THEORY

A. Preliminary mathematical results

1. Bias-variance decomposition

Let $p(\mathbf{x})$ be a probability density function on a finite dimensional vector space $\mathcal{X}$ with Euclidean norm denoted $\|\cdot\|$. Define $\boldsymbol{\mu} = \int \mathbf{x}p(\mathbf{x})\,d\mathbf{x}$.

a. Claim For any $\mathbf{s} \in \mathcal{X}$,

$$\int p(\mathbf{x})\|\mathbf{s}-\mathbf{x}\|^2\,d\mathbf{x} = \|\mathbf{s}-\boldsymbol{\mu}\|^2 + \int p(\mathbf{x})\|\boldsymbol{\mu}-\mathbf{x}\|^2\,d\mathbf{x}. \tag{S1}$$

Proof. First, we have

$$\|\mathbf{s}-\mathbf{x}\|^2 = \|\mathbf{s}-\boldsymbol{\mu}\|^2 + 2(\mathbf{s}-\boldsymbol{\mu})\cdot(\boldsymbol{\mu}-\mathbf{x}) + \|\boldsymbol{\mu}-\mathbf{x}\|^2. \tag{S2}$$

Then,

$$\int p(\mathbf{x})(\mathbf{s}-\boldsymbol{\mu})\cdot(\boldsymbol{\mu}-\mathbf{x})\,d\mathbf{x} = (\mathbf{s}-\boldsymbol{\mu})\cdot\int p(\mathbf{x})(\boldsymbol{\mu}-\mathbf{x})\,d\mathbf{x} \tag{S3}$$

$$= (\mathbf{s}-\boldsymbol{\mu})\cdot\left(\boldsymbol{\mu} - \int \mathbf{x}p(\mathbf{x})\,d\mathbf{x}\right) \tag{S4}$$

$$= 0. \tag{S5}$$

This completes the proof. □

b. Corollary The mean value minimises the mean squared error

$$\boldsymbol{\mu} = \min_{\mathbf{s}} \int p(\mathbf{x})\|\mathbf{s}-\mathbf{x}\|^2\,d\mathbf{x}. \tag{S6}$$

2. Stein's Identity

Let $p(\mathbf{x}_2,\ldots,\mathbf{x}_N \mid \phi)$ be any family of probability densities parametrized by ϕ.

a. Claim

$$\int p(\mathbf{x}_2\ldots,\mathbf{x}_N \mid \phi)\,\nabla_\phi \log p(\mathbf{x}_2\ldots,\mathbf{x}_N \mid \phi)\,d\mathbf{x}_{2:N} = 0. \tag{S7}$$

[a)] These authors contributed equally to this work.

[b)] Electronic mail: franknoe@microsoft.com

[c)] Electronic mail: jan.hermann@microsoft.com

[d)] Electronic mail: pgorigiorgi@microsoft.com

[e)] Electronic mail: adam.e.foster@microsoft.com

Proof.

$$
\begin{aligned}
0 &= \nabla_\phi \int p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \phi)\, d\mathbf{x}_{2:N} && \text{(S8)}\\
&= \int \nabla_\phi p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \phi)\, d\mathbf{x}_{2:N} && \text{(S9)}\\
&= \int p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \phi)\, \nabla_\phi \log p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \phi)\, d\mathbf{x}_{2:N}. && \text{(S10)}
\end{aligned}
$$

□

B. Main result

We write $p(\mathbf{x}_1, \dots, \mathbf{x}_N) = |\Psi(\mathbf{x}_1, \dots, \mathbf{x}_N)|^2$ for the joint probability density of the N electrons. Then the electron density

$$
\rho(\mathbf{x}) = N \int p(\mathbf{x}, \mathbf{x}_2, \dots, \mathbf{x}_N)\, d\mathbf{x}_{2:N} \tag{S11}
$$

is the marginal probability density multiplied by N. Also recall that $\mathbf{x} = (\mathbf{r}, \sigma)$ where $\mathbf{r} \in \mathbb{R}$ is the spatial co-ordinate and $\sigma \in \{\uparrow, \downarrow\}$ is the spin. The *score function* is $\nabla_\mathbf{r} \log \rho(\mathbf{x})$, the gradient of the density with respect to position.

We now prove that Eq. (13) is a correct loss function to learn ρ in two claims.

a. Claim

$$
\nabla_\mathbf{r} \log \rho = \underset{s:\mathbb{R}^3\times\{\uparrow,\downarrow\}\to\mathbb{R}^3}{\arg\min} \int \|\nabla_{\mathbf{r}_1} \log p(\mathbf{x}_1, \dots, \mathbf{x}_N) - s(\mathbf{x}_1)\|^2\, p(\mathbf{x}_1, \dots, \mathbf{x}_N)\, d\mathbf{x}_{1:N}. \tag{S12}
$$

where the minimization is over functions s.

Proof. For a given $\mathbf{x}_1$, the minimization problem is solvable and the minimizing value is the conditional expectation. This follows from the corollary of the bias-variance decomposition. Then,

$$
\begin{aligned}
s(\mathbf{x}_1) &= \int p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \mathbf{x}_1)\, \nabla_{\mathbf{r}_1} \log p(\mathbf{x}_1, \dots, \mathbf{x}_N)\, d\mathbf{x}_{2:N} && \text{(S13)}\\
&= \int p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \mathbf{x}_1)\, \nabla_{\mathbf{r}_1} \left(\log \rho(\mathbf{x}_1) + \log p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \mathbf{x}_1)\right)\, d\mathbf{x}_{2:N} && \text{(S14)}\\
&= \nabla_{\mathbf{r}_1} \log \rho(\mathbf{x}_1) + \int p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \mathbf{x}_1)\, \nabla_{\mathbf{r}_1} \log p(\mathbf{x}_2 \dots, \mathbf{x}_N \mid \mathbf{x}_1)\, d\mathbf{x}_{2:N} && \text{(S15)}\\
&= \nabla_{\mathbf{r}_1} \log \rho(\mathbf{x}_1), && \text{(S16)}
\end{aligned}
$$

using Stein's Identity in the last step (taking $\mathbf{x}_1$ as our ϕ). □

b. Corollary We now use the fact that $p(\mathbf{x}_1, \dots, \mathbf{x}_N)$ is exchangeable, i.e. for any permutation σ, $p(\mathbf{x}_1, \dots, \mathbf{x}_N) = p(\mathbf{x}_{\sigma(1)}, \dots, \mathbf{x}_{\sigma(N)})$. This follows from the antisymmetry of the wave function. Then

$$
\begin{aligned}
&\int \frac{1}{N} \sum_{i=1}^{N} \left(\|\nabla_{\mathbf{r}_i} \log p(\mathbf{x}_1, \dots, \mathbf{x}_n) - s(\mathbf{x}_i)\|^2\right) p(\mathbf{x}_1, \dots, \mathbf{x}_n)\, d\mathbf{x}_{1:N} && \text{(S17)}\\
&= \frac{1}{N} \sum_{i=1}^{N} \int \|\nabla_{\mathbf{r}_i} \log p(\mathbf{x}_1, \dots, \mathbf{x}_n) - s(\mathbf{x}_i)\|^2\, p(\mathbf{x}_1, \dots, \mathbf{x}_n)\, d\mathbf{x}_{1:N} && \text{(S18)}\\
&= \frac{1}{N} \sum_{i=1}^{N} \int \left\|\nabla_{\mathbf{r}_i} \log p(\mathbf{x}_i, \mathbf{x}_{\setminus i}) - s(\mathbf{x}_i)\right\|^2 p(\mathbf{x}_i, \mathbf{x}_{\setminus i})\, d\mathbf{x}_{1:N} && \text{(S19)}\\
&= \int \|\nabla_{\mathbf{r}_1} \log p(\mathbf{x}_1, \mathbf{x}_{2:N}) - s(\mathbf{x}_1)\|^2\, p(\mathbf{x}_1, \mathbf{x}_{2:N})\, d\mathbf{x}_{1:N}. && \text{(S20)}
\end{aligned}
$$

1. Unbiasedness of the training gradient estimator

Now assume $s(\cdot;\boldsymbol{\theta})$ is parameterized and we aim to find the minimizer of (S12) via stochastic gradient descent. Denote

$$\ell(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x}_1,\mathbf{x}_{2:N})}\left[\|\nabla_{\mathbf{r}_1}\log p(\mathbf{x}_1,\mathbf{x}_{2:N}) - s(\mathbf{x}_1;\boldsymbol{\theta})\|^2\right] \tag{S21}$$

$$\hat{\ell}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x}_1)}\left[\|\nabla_{\mathbf{r}_1}\log p(\mathbf{x}_1) - s(\mathbf{x}_1;\boldsymbol{\theta})\|^2\right]. \tag{S22}$$

a. Claim

$$\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\hat{\ell}(\boldsymbol{\theta}) \tag{S23}$$

Proof. Analyzing the residual gives

$$\ell(\boldsymbol{\theta}) - \hat{\ell}(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{x}_1,\mathbf{x}_{2:N})}\left[\|\nabla_{\mathbf{r}_1}\log p(\mathbf{x}_1,\mathbf{x}_{2:N})\|^2 - \|\nabla_{\mathbf{x}_1}\log p(\mathbf{x}_1)\|^2 - 2\boldsymbol{\delta}(\mathbf{x}_1,\mathbf{x}_{2:N})^\top s(\mathbf{x}_1;\boldsymbol{\theta})\right] \tag{S24}$$

with

$$\boldsymbol{\delta}(\mathbf{x}_1,\mathbf{x}_{2:N}) := \nabla_{\mathbf{r}_1}\log p(\mathbf{x}_1,\mathbf{x}_{2:N}) - \nabla_{\mathbf{r}_1}\log p(\mathbf{x}_1). \tag{S25}$$

Using eqs. (S13) - (S16) we get

$$\mathbb{E}_{p(\mathbf{x}_1,\mathbf{x}_{2:N})}\left[\boldsymbol{\delta}(\mathbf{x}_1,\mathbf{x}_{2:N})\right] = \mathbf{0} \tag{S26}$$

and as such

$$\nabla_{\boldsymbol{\theta}}\left(\ell(\boldsymbol{\theta}) - \hat{\ell}(\boldsymbol{\theta})\right) = \nabla_{\boldsymbol{\theta}}\mathbb{E}_{p(\mathbf{x}_1,\mathbf{x}_{2:N})}\left[\|\nabla_{\mathbf{x}_1}\log p(\mathbf{x}_1,\mathbf{x}_{2:N})\|_2^2 - \|\nabla_{\mathbf{x}_1}\log p(\mathbf{x}_1)\|_2^2\right] = \mathbf{0}. \tag{S27}$$

□

C. Connection to force matching

We note a strong connection between our approach to electron density estimation and the force matching approach to coarse-graining.? Force matching considers a more general coarse-graining operator that is a linear map π from the original co-ordinates $\mathbf{x}_1,\ldots,\mathbf{x}_N$ of a system to a smaller set of co-ordinates that describe certain aspects of the system. Our approach can be seen as a special case of coarse-graining where we use the operator

$$\mathbf{x}_1,\ldots,\mathbf{x}_N \mapsto \mathbf{x}_1. \tag{S28}$$

S2. KATO'S CUSP CONDITION

We use the shorthand $\mathbf{r}_I = \mathbf{r} - \mathbf{R}_I$ and $r_I = \|\mathbf{r}_I\|$. The cusp term in the model is given by

$$C(\mathbf{r}) = \sum_I 2\sqrt{\pi}\mathrm{erf}(\tfrac{1}{2}Z_I r_I) = \sum_I 4\int_0^{\frac{1}{2}Z_I r_I} e^{-t^2}dt. \tag{S29}$$

The gradient of this term is then

$$\nabla_{\mathbf{r}}C(\mathbf{r}) = \sum_I 2Z_I e^{-\frac{1}{4}Z_I^2 r_I^2}\hat{\mathbf{r}_I}. \tag{S30}$$

The Laplacian can be computed using standard formulae for spherical co-ordinates

$$\begin{aligned}\nabla_{\mathbf{r}}^2 C(\mathbf{r}) &= \sum_I \frac{1}{r_I^2}\frac{\partial}{\partial r_I}\left(r_I^2\frac{\partial C}{\partial r_I}\right) &\text{(S31)}\\ &= \sum_I \frac{1}{r_I^2}\frac{\partial}{\partial r_I}\left(2Z_I r_I^2 e^{-\frac{1}{4}Z_I^2 r_I^2}\right) &\text{(S32)}\\ &= \sum_I \frac{1}{r_I^2}\left(4Z_I r_I - Z_I^3 r_I^3\right)e^{-\frac{1}{4}Z_I^2 r_I^2} &\text{(S33)}\\ &= \sum_I \left(\frac{4Z_I}{r_I} - Z_I^3 r_I\right)e^{-\frac{1}{4}Z_I^2 r_I^2}, &\text{(S34)}\end{aligned}$$

furthermore, the asymptotic behaviour for small r_I is

$$= \sum_I \left(\frac{4Z_I}{r_I} - Z_I^3 r_I - Z_I^3 r_I + O(r_I^3) \right) \tag{S35}$$

$$= \sum_I \left(\frac{4Z_I}{r_I} - 2Z_I^3 r_I + O(r_I^3) \right). \tag{S36}$$

S3. EFFECTIVE POTENTIAL COMPUTATION

For the LiH experiment in Sec. IVB, we compute the following quantity

$$v_{\text{eff}}(\mathbf{x}) - v_{\text{ext}}(\mathbf{x}) = \frac{1}{8}\|\nabla_{\mathbf{r}} \log \rho(\mathbf{x})\|^2 + \frac{1}{4}\nabla_{\mathbf{r}}^2 \log \rho(\mathbf{x}) + \sum_I \frac{Z_I}{r_I}. \tag{S37}$$

Using the additive form of our density $U_\phi(\mathbf{x}) = E_\phi(\mathbf{x}) + M_\phi(\mathbf{x}) + C(\mathbf{r})$, gives

$$v_{\text{eff}}(\mathbf{x}) - v_{\text{ext}}(\mathbf{x}) = \frac{1}{8}\| - \nabla_{\mathbf{r}} U_\phi(\mathbf{x})\|^2 - \frac{1}{4}\nabla_{\mathbf{r}}^2 \left(E_\phi(\mathbf{x}) + M_\phi(\mathbf{x}) + C(\mathbf{r})\right) + \sum_I \frac{Z_I}{r_I}. \tag{S38}$$

From this equation, we use the fact that $\| - \nabla_{\mathbf{r}} U_\phi\|^2$, $\nabla_{\mathbf{r}}^2 E_\phi$, and $\nabla_{\mathbf{r}}^2 M_\phi$ are bounded by design. Cancellation of the poles in the external potential can therefore only come from $\nabla_{\mathbf{r}}^2 C$. We expand this Laplacian using the analysis from above, Eq. (S34), to give

$$\begin{aligned} v_{\text{eff}}[\rho](\mathbf{x}) - v_{\text{ext}}[\rho](\mathbf{x}) = & \frac{1}{8}\| - \nabla_{\mathbf{r}} U_\phi(\mathbf{x})\|^2 - \frac{1}{4}\nabla_{\mathbf{r}}^2 \left(E_\phi(\mathbf{x}) + M_\phi(\mathbf{x})\right) \\ & + \sum_I \left(\frac{-Z_I}{r_I}(e^{-\frac{1}{4}Z_I^2 r_I^2} - 1) + \tfrac{1}{4} Z_I^3 r_I e^{-\frac{1}{4}Z_I^2 r_I^2} \right). \end{aligned} \tag{S39}$$

Note that

$$\frac{-Z_I}{r_I}(e^{-\frac{1}{4}Z_I^2 r_I^2} - 1) = \frac{1}{4} Z_I^3 r_I + O(r_I^3) \tag{S40}$$

no longer explodes as $r_I \to 0$. For numerical stability, we explicitly replace $\frac{-Z_I}{r_I}(e^{-\frac{1}{4}Z_I^2 r_I^2} - 1)$ with the first term of its series expansion for small r_I.

S4. TOTAL VARIATION DISTANCE

The total variation (TV) distance is a metric on the space of probability distributions[?] that can be expressed as

$$\mathrm{TV}(\rho, \rho') = \frac{1}{2}\int |\rho(\mathbf{r}) - \rho'(\mathbf{r})| d\mathbf{r}. \tag{S41}$$

We extend this definition to electron densities (keeping their normalization to N, rather than 1). Whilst it is generally not possible to find "ground truth" electron densities, the TV distance is helpful to compare to references and to check expected symmetries of the density.

S5. ESTIMATION OF OBSERVABLES FROM ELECTRON DENSITY

We can also validate our NERD models by estimating certain one-body observable quantities. In every case, the observable quantity involves an expectation over the density, $\int \cdot\ \rho(\mathbf{x})\, d\mathbf{x}$. We compute these integrals using Lebedev–Laikov grids.[?]

By the Hellmann-Feynman Theorem,[? ?] the *force* on nucleus I is given by

$$\mathbf{F}_I = \sum_{J \neq I} \frac{Z_I Z_J (\mathbf{R}_I - \mathbf{R}_J)}{\|\mathbf{R}_I - \mathbf{R}_J\|^3} + \int \frac{Z_I (\mathbf{r} - \mathbf{R}_I)}{\|\mathbf{r} - \mathbf{R}_I\|^3} \rho(\mathbf{x})\, d\mathbf{x}\,. \tag{S42}$$

This formula holds when ρ corresponds to a variationally optimal wave function. We compare the approach of integrating ρ obtained from our NERD model with the estimators that directly operate on samples from the deep QMC wave function.

Second, we can compute the expectation value of the *nuclear potential*

$$V[\rho] = -\int \left(\sum_I \frac{Z_I}{\|\mathbf{r} - \mathbf{R}_I\|} \right) \rho(\mathbf{x})\, d\mathbf{x}; \tag{S43}$$

this is the only component of the Hamiltonian that can be directly calculated from the density, and is most sensitive to density quality near nuclei.

Third, *quadrupole moment*, Q, are defined as

$$Q_{ij} = \int \left(3r_i r_j - \|\mathbf{r}\|^2 \delta_{ij}\right) \rho(\mathbf{x})\, d\mathbf{x}, \tag{S44}$$

and help describe the electrostatic properties of a charge density.

S6. ASSARAF–CAFFAREL FORCE ESTIMATOR FOR THE WAVE FUNCTION

Forces in quantum Monte Carlo can be evaluated by applying the Hellmann-Feynman theorem, turning the energy derivative into an expectation value of the "classical" Coulomb force over samples from the wave function

$$\mathbf{F}_I = \sum_{J \neq I} \frac{Z_I Z_J (\mathbf{R}_I - \mathbf{R}_J)}{\|\mathbf{R}_I - \mathbf{R}_J\|^3} + \left\langle \frac{Z_I(\mathbf{r} - \mathbf{R}_I)}{\|\mathbf{r} - \mathbf{R}_I\|^3} \right\rangle_{\mathbf{r}\sim\Psi^2}. \tag{S45}$$

This estimator of the Hellman-Feynman force has the desired expectation value, but exhibits infinite variance due to the divergences of the force at the nuclei, severely limiting its practical applicability. To improve the efficiency, Assaraf and Caffarel proposed a force estimator that has the same mean but finite variance

$$\mathbf{F}_I^{ZV} = \mathbf{F}_I + \left\langle \frac{(H - E_{\text{loc}})\tilde{\Psi}}{\Psi} \right\rangle_{\mathbf{r}\sim\Psi^2}, \tag{S46}$$

where E_{loc} is the local energy of the wave function and $\tilde{\Psi}$ is an approximation of the wave function derivative.[?] The "minimal" version of the zero variance estimator cancels the divergences in the Coulomb force by choosing

$$\tilde{\Psi}_{\min} = Q\Psi \quad \text{with} \quad Q_{I,j} = Z_I \sum_i \frac{(\mathbf{r}_i - \mathbf{R}_I)_j}{\|\mathbf{r}_i - \mathbf{R}_I\|}, \tag{S47}$$

where I indexes atoms and j indexes the three spacial dimensions. Within this approximation the expectation value takes the form

$$\mathbf{F}_I^{ZV} = \left\langle -\nabla_{\mathbf{r}} Q \frac{\nabla_{\mathbf{r}}\Psi}{\Psi} \right\rangle_{\mathbf{r}\sim\Psi^2}. \tag{S48}$$

The estimator can be further extended with a term to correct a potential bias in the force estimate due to a discrepancy of the wave function from the true ground state

$$\mathbf{F}_I^{ZVZB} = \mathbf{F}_I^{ZV} + \left\langle 2(E_{\text{loc}} - E)\frac{\tilde{\Psi}}{\Psi} \right\rangle_{\mathbf{r}\sim\Psi^2}, \tag{S49}$$

where E is the energy expectation value of the wave function. Note that the $\mathbf{F}_I^{ZVZB}$ estimator requires the explicit evaluation of local energies, adding a significant computational overhead over the $\mathbf{F}_I^{ZV}$ estimator (see **?** for more details).

S7. ADDITIONAL EXPERIMENTS

A. Kato's cusp condition using $v_{\text{eff}} - v_{\text{ext}}$ vs r

In Fig. S1, the $v_{\text{eff}} - v_{\text{ext}}$ values are plotted as a function of $r = |\mathbf{r} - \mathbf{R}_I|$ for the small atoms tested in this study. For all atoms, $v_{\text{eff}} - v_{\text{ext}}$ values approach constants when $r \to 0$ as expected (Eq. S39 and S40). This further verifies that the neural network densities have the correct behavior near the nucleus and satisfy an exact Kato's cusp condition.

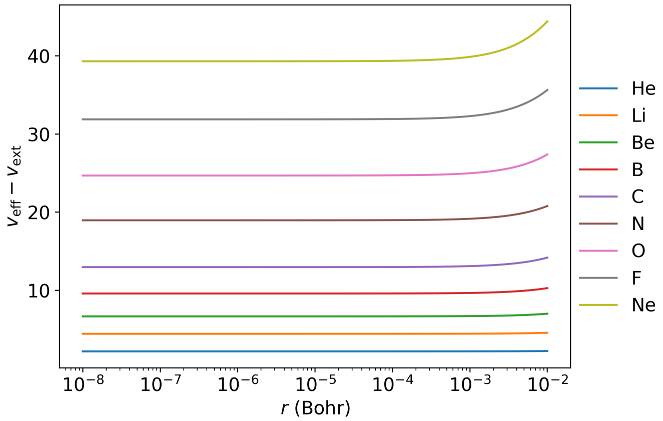

FIG. S1: $v_{\mathrm{eff}} - v_{\mathrm{ext}}$ vs r for 9 small atoms examined in this study. The x-axis is plotted on a log-scale to visualize the distances to nuclei

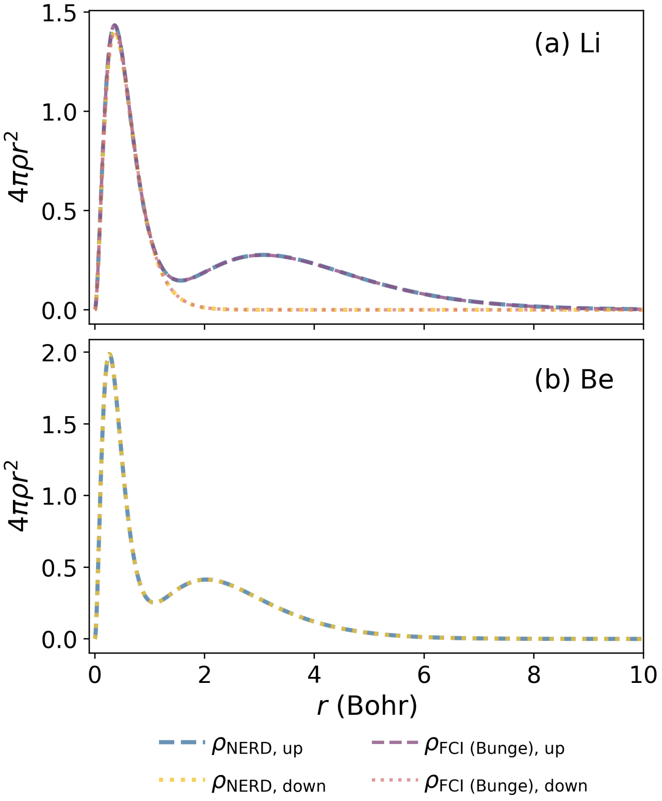

FIG. S2: Comparison of the ρ_{up} and ρ_{down} of (a) Li and (b) Be. To better visualize the differences between spin-up and spin-down densities, all the values are plotted as $4\pi\rho r^2$. The FCI spin density values for Li are computed with 8s up to k basis functions, and are kindly provided by Dr. Carlos Bunge.

B. Spin-up and spin-down density comparisons for Li and Be

We note that NERD models could provide accurate total densities and also spin densities. Figure S2 displays comparisons of spin NERDs for Li and Be using $4\pi\rho r^2$ values, where $r = |\mathbf{r} - \mathbf{R}_I|$ the corresponding FCI spin densities for Li are also compared.

C. Quadrupole moment evaluations for benzene

Figure S3 plots an additional experiment (similar to Fig. 9) on the convergence of quadrupole moments with training steps for benzene.

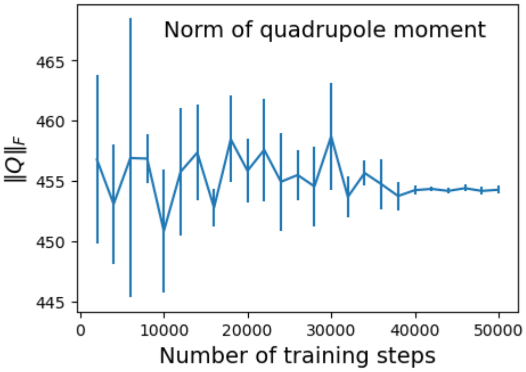

FIG. S3: Norm of the quadrupole moment for equilibrium benzene structure

S8. QUANTUM CHEMISTRY CALCULATIONS

All the densities from second quantized quantum chemistry methods in this study are obtained by PySCF 2.4.0[?] with the default settings. The calculations include FCI/aug-cc-pVQZ densities of Li and Be atoms (Figs. 2 and 3), FCI/aug-cc-pVTZ densities of the equilibrium structure of LiH with a bond length of 3.015 Bohr (Fig. 5), and CCSD/aug-cc-pVQZ densities of small atoms in Table 2 and H_4 square with bond lengths of 3.0236 Bohr (Fig. 6). The literature VMC densities of Be and Ne atoms are obtained from **?**

S9. EXPERIMENTAL SETTINGS FOR PSIFORMER AND NERD MODEL TRAINING

The following tables summarize the experimental settings of the Psiformer (Table S1) and the NERD models (Table S2) reported in this study. For all experiments, optimization during variational wave function training was done using KFAC[?] with learning rate 5×10^{-2} with geometric decay schedule, damping factor 10^{-3}, and norm constraint 10^{-3}. Supervised pretraining to Hartree-Fock used the Adam optimizer with learning rate 10^{-3}. Density training also used the Adam optimizer.

TABLE S1: Experimental settings of Psifomer models for different systems.

Setting	Atoms & ions	LiH dissociation	H_4	Benzene	H_2 & Li_2	N_2
Electron batch size	4096	2048	4096	4096	2048	4096
MCMC sampler	MALA[a]	MALA	MALA	MHA [b]	MHA	MALA
Pretraining steps	20000	20000	500	20000	5000	5000
Training steps	max(50k, 10k $\cdot n_{e^-}$)	100k	100k	200k	100k	200k
Energy evaluation Steps	10k	—	—	—	10k	10k
Force evaluation Steps	—	—	—	—	10k	10k
Dipole moment evaluation Steps	—	10k	—	—	—	—

[a] Metropolis adjusted Langevin algorithm. [b] Metropolis–Hastings algorithm.

TABLE S2: Experimental settings of NERD models for different systems.

Setting	Atoms	LiH dissociation	H_4	Benzene	H_2, Li_2 and N_2
Electron batch size	4096	2048	2048	4096	4096
MCMC sampler	MHA	MHA	MHA	MHA	MHA
Steps	max(25k, 5k $\cdot\, n_{e-}$)	25k	25k	50k	40k
NCE weight, λ	1	1	1	1	1
Learning rate	Cosine decay schedule: init value=0.01, decay steps=0.8 $\cdot$ steps, α=6e-5, b1=0.95, and b2=0.999				