npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01433-0

Convolutional network learning of self-consistent electron density via grid-projected atomic fingerprints

Ryong-Gyu Lee & Yong-Hoon Kim

The self-consistent field (SCF) generation of the three-dimensional (3D) electron density distribution (ρ) represents a fundamental aspect of density functional theory (DFT) and related first-principles calculations, and how one can shorten or bypass the SCF loop represents a critical question in electronic structure theory from both practical and fundamental standpoints. Herein, a machine learning strategy, DeepSCF, is presented in which the map between the SCF ρ and the initial guess density (ρ_0) constructed by the summation of neutral atomic densities is learned using 3D convolutional neural networks (CNNs). High accuracy and transferability of DeepSCF are achieved by first encoding ρ_0 on a 3D grid and then expanding the input features to include atomic fingerprints beyond ρ_0. The prediction of the residual density ($\delta\rho$) rather than ρ itself is targeted, and given that $\delta\rho$ is indicative of chemical bonding information, a dataset of small-sized organic molecules featuring diverse bonding characters is adopted. The fidelity of DeepSCF is finally enhanced by subjecting the atomic geometries of the dataset to random rotations and strains. The effectiveness of DeepSCF is demonstrated using a complex carbon nanotube-based DNA sequencer model. This work evidences that the nearsightedness in electronic structure can be optimally represented via the spatial locality in CNNs, offering insight into the success of various machine learning-based atomistic materials simulations.

The mean field or self-consistent field (SCF) method represents a ubiquitous computational approach to deal with complex scientific problems formulated as several coupled differential equations and appear in a wide range of contexts such as the Landau theory for phase transitions[1], Bogoliubov-de Gennes equations for superconductivity[2], Gummel's equations for semiconductor devices[3], and Kohn-Sham density functional theory (DFT) for ab initio electronic structure calculations[4,5], to name a few. In the case of DFT, which has become the standard computational tool for a wide range of science and engineering fields, the SCF solutions of the Kohn-Sham (KS) equations identify the three-dimensional (3D) ground-state electron density $\rho(\vec{r})$ and, in doing so, obtain the variationally minimized total energy. Despite its success, the applicability of DFT calculations is typically limited to a few hundred to thousand atoms due to the cubic scaling of the computational cost with respect to the number of atoms. To enable large-scale DFT calculations, advanced algorithms to accelerate SCF cycles have been devised[6] and in parallel several routes to reformulate DFT in the context of order-N and orbital-free DFT methods have also been explored[5,7].

More recently, machine learning (ML) techniques have emerged as promising alternatives, and much effort is presently being devoted in developing ML strategies that predict DFT secondary output such as total energy and atomic forces[8] or DFT primary output such as the exchange-correlation functional[9–11], wavefunctions[12], local density of states[13], density matrix[14], and electron density. Limiting ourselves to the electron density prediction, arguably the most straightforward and flexible route in the latter category and is true to the spirit of the Hohenberg–Kohn theorem, it can be broadly categorized into the models based on basis functions[15–18] and those based on grid representations[19–24]. The basis functions-based models generate $\rho(\vec{r})$ by expanding it in terms of usually atom-centered basis functions and the coefficients learned by fitting the dataset. On the other hand, grid-based models learn the map between input fingerprints and the $\rho(\vec{r})$ represented on a 3D real-space mesh. However, current methods still suffer from several shortcomings: The performance of the former strongly depends on the quality of basis functions and the models developed for a specific set of basis functions are less transferable to other basis

School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea.
✉e-mail: y.h.kim@kaist.ac.kr

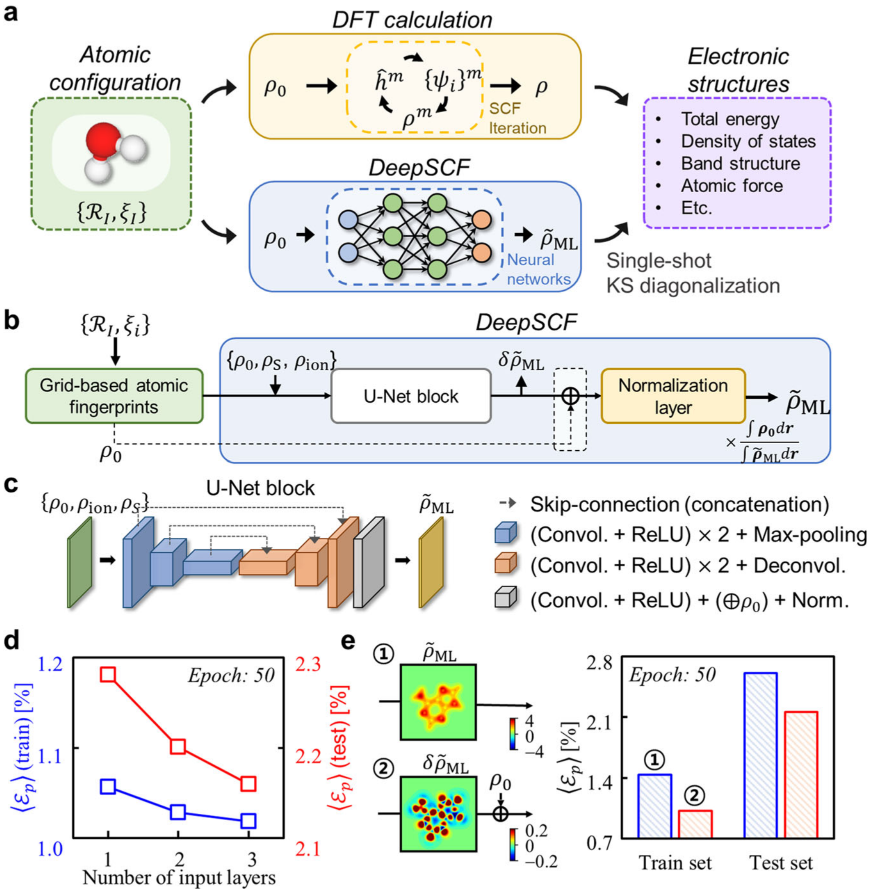

a
Atomic configuration
$\{\mathcal{R}_I, \xi_I\}$
DFT calculation
ρ_0
$\hat{h}^m$ $\{\psi_i\}^m$
ρ^m
ρ
SCF Iteration
DeepSCF
ρ_0
$\tilde{\rho}_{\mathrm{ML}}$
Neural networks
Electronic structures
Total energy
Density of states
Band structure
Atomic force
Etc.
Single-shot KS diagonalization
b
$\{\mathcal{R}_I, \xi_i\}$
Grid-based atomic fingerprints
DeepSCF
$\{\rho_0, \rho_S, \rho_{\mathrm{ion}}\}$
U-Net block
$\delta\tilde{\rho}_{\mathrm{ML}}$
Normalization layer
$\tilde{\rho}_{\mathrm{ML}}$
$\times \frac{\int \rho_0 dr}{\int \tilde{\rho}_{\mathrm{ML}} dr}$
ρ_0
c
U-Net block
$\{\rho_0, \rho_{\mathrm{ion}}, \rho_S\}$
$\tilde{\rho}_{\mathrm{ML}}$
Skip-connection (concatenation)
(Convol. + ReLU) × 2 + Max-pooling
(Convol. + ReLU) × 2 + Deconvol.
(Convol. + ReLU) + ($\oplus\rho_0$) + Norm.
d
Epoch: 50
$\langle \mathcal{E}_p \rangle$ (train) [%]
$\langle \mathcal{E}_p \rangle$ (test) [%]
1.2
1.1
1.0
2.3
2.2
2.1
1
2
3
Number of input layers
e
①
$\tilde{\rho}_{\mathrm{ML}}$
4
0
−4
②
$\delta\tilde{\rho}_{\mathrm{ML}}$
ρ_0
0.2
0
−0.2
$\langle \mathcal{E}_p \rangle$ [%]
2.8
2.1
1.4
0.7
Epoch: 50
①
②
Train set
Test set

Fig. 1 | Overview of the DeepSCF model. **a** The DFT calculation (orange box) and DeepSCF scheme (blue box) to predict the electronic structure of a material with atomic geometry ($\mathcal{R}_I$) and chemical composition (ξ_I). **b** The grid-projected atomic features extracted from the $\mathcal{R}$ and ξ are encoded into the initial layers of the U-Net block, and its output $\delta\tilde{\rho}_{\mathrm{ML}}$ and ρ_0 at the initial layer are summed (denoted by $\bigoplus$) and normalized by ρ_0, producing the final output ($\tilde{\rho}_{ML}$). **c** The neural networks architecture of the DeepSCF model is based on a U-Net block, which consists of encoder (blue box) and decoder (orange box), and finetuning (gray box) blocks (see Supplementary Note 1 for details). **d** The performance of input atomic features. The 1, 2, and 3 number of input layers respectively represent $\{\rho_0\}$, $\{\rho_0, \rho_{ion}\}$, and $\{\rho_0, \rho_{ion}, \rho_s\}$ features. **e** Comparison between the ρ-learning scheme ① and the $\delta\rho$-learning (DeepSCF) approach ②. For the former, the DeepSCF architecture was modified at the final finetuning block by removing the dotted box part in (**b**) (left panel). The performance of the two approaches estimated in by $\left\langle \mathcal{E}_p \right\rangle$ for both training and test datasets.

functions. For the latter group, the currently implemented approaches attempted to construct neural networks attached to each grid point, requiring substantial computational resources and quickly becoming impractical as the system size increases. Accordingly, modern state-of-the-art computer vision approaches, such as the convolutional neural network (CNN), could not be employed.

In this work, partly motivated by our earlier work on the multigrid acceleration of the SCF cycle within DFT[25,26], we develop the DeepSCF scheme that learns the map between atomic fingerprints encoded on a 3D mesh and the converged SCF density based on CNN model (Fig. 1). One central implication of the nearsightedness of electronic structure asserted by Walter Kohn is that local properties like chemical bonding and functional groups are transferable from one environment to another[27], and we hypothesized that this nearsightedness principle could be particularly well mapped to the spatial locality (or local connectivity) in CNNs[28]. This is to our knowledge the first successful demonstration of directly predicting $\rho(\vec{r})$ using the CNN algorithm, where advanced neural network architectures such as U-Net is available[29]. To realize this goal, we encode the structural and chemical signatures of target materials into 3D real-space grid by augmenting the summation of (modified) neutral atomic densities $\rho_0(\vec{r})$, typically used as the initial guess electron density, with additional grid-projected atomic fingerprints. Given that the density difference $\delta\rho$ between ρ_0 and ρ is information the SCF cycle generates and corresponds to the chemical bonding information (in practice, the correspondence is usually approximate due to the modification of neutral atomic densities), we devise a new U-Net-based architecture that includes a skip connection to learn the residual $\delta\rho$. We train our model by adopting a molecular database that includes diverse chemical bonding configurations and additionally enhancing the training dataset by randomly rotating and straining molecular geometries. We will confirm that the quality of our DeepSCF or non-SCF DFT calculations is comparable to that of fully SCF DFT counterparts and demonstrate its size-extensibility and transferability using crystalline polyethylene and graphene, and a large-scale carbon nanotube-based DNA sequencer model. The possibility of highly accurate ML-based prediction of ρ suggests the mechanistic origins of the outstanding performance of higher-level ML-based materials simulation methods such as neural network force fields.

Results

Key features of the DeepSCF model

In Fig. 1b, we show the schematics of the DeepSCF framework that learns the map between the summation of neutral atomic densities $\rho_0(\vec{r})$ and the SCF electron density $\rho(\vec{r})$ based on the CNN (For details, see Supplementary Note 1). We first project on a 3D mesh $\rho_0(\vec{r})$ and other atomic orbitals-related fingerprints, which will be detailed below. They encode the atomic geometry R and chemical composition ξ while effectively differentiating various chemical environments. Then, these input features are applied to the CNN model and converted into the final product $\tilde{\rho}_{\mathrm{ML}}$. As the CNN architecture, we employed the U-Net, which consists of encoding and decoding blocks composed of repeated sets of the convolution layers and rectified linear unit (ReLU) operations[29]. The encoding (decoding) blocks utilize the max-pooling (deconvolution) operations to decrease (increase) the resolution of hidden features. In addition, to integrate the output features from the encoding block into the decoding block, these blocks were linked by skip-connections (dotted arrows) (Fig. 1c; for details, see Supplementary Fig. 1 and Supplementary Table 1). As will be detailed later, rather than directly predicting $\tilde{\rho}_{\mathrm{ML}}$, we used the U-Net block to generate the residual $\delta\tilde{\rho}_{\mathrm{ML}}$. We added it to ρ_0 and normalized $\rho_0 + \delta\tilde{\rho}_{\mathrm{ML}}$ by the total number of electrons to prepare the final feature $\tilde{\rho}_{\mathrm{ML}}$. The electronic structure is then obtained by a single-shot KS diagonalization over the $\tilde{\rho}_{\mathrm{ML}}$ without the SCF loop.

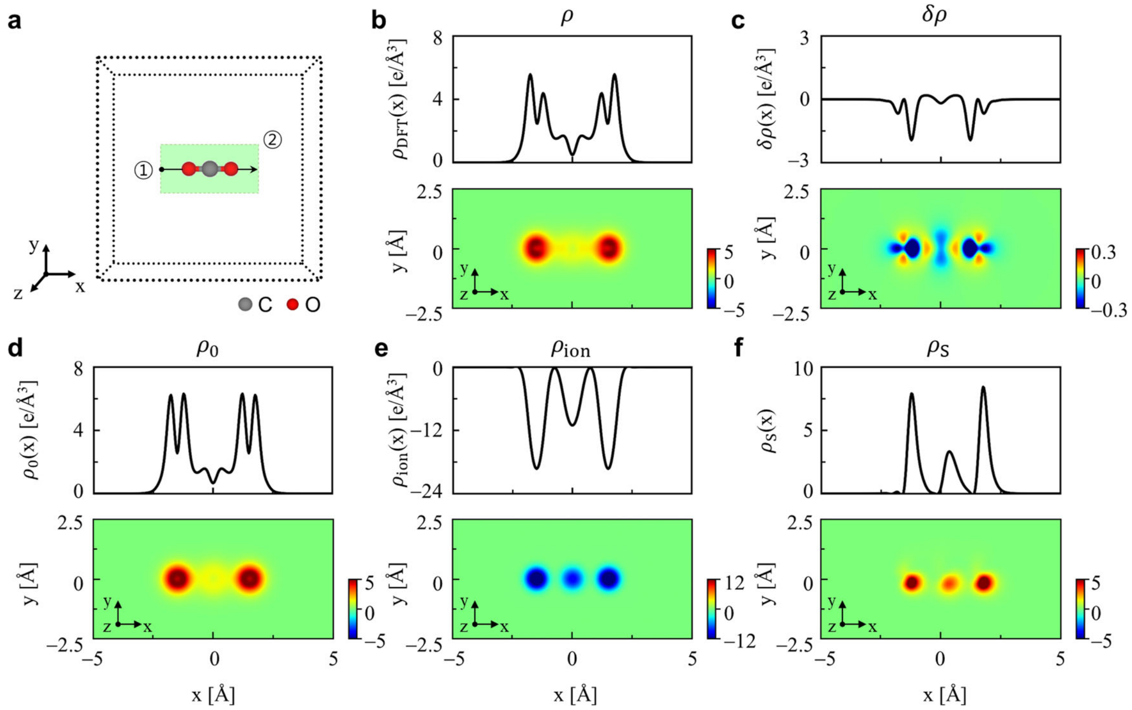


Fig. 2 | Input and output features of the carbon dioxide example. a For the carbon dioxide molecule, we visualize the one-dimensional and two-dimensional cross-sections (① and ②) of the spatial distributions of output **b** ρ and **c** $\delta\rho$ and input features **d** ρ_0, **e** ρ_{ion} and **f** ρ_s through the molecular bond axis.

In evaluating the prediction accuracy, we introduced the absolute percentage error in the predicted density

$$\mathcal{E}_p(\%) = 100 \times \frac{\int d\vec{\boldsymbol{r}} \left|\rho(\vec{\boldsymbol{r}}) - \widetilde{\rho}_{\mathrm{ML}}(\vec{\boldsymbol{r}})\right|}{\int d\vec{\boldsymbol{r}}\rho(\vec{\boldsymbol{r}})} \tag{1}$$

and used it together with the errors in the total energy, atomic forces, and highest occupied molecular orbital (HOMO), and lowest unoccupied molecular orbital (LUMO) values. In addition, since different systems i contain different numbers of electrons N_e^i, we estimated the average prediction accuracy for a dataset including several structures using the weighted formula, i.e., $\left\langle \mathcal{E}_p \right\rangle(\%) = 100 \times \frac{\sum_i^N N_e^i \mathcal{E}_p^i}{N_e}$, where the N_e is the number of electrons in the total dataset.

Input and output features

We next discuss in more detail the input and output features of DeepSCF, which we designed to capture the essential features of the $\rho_0(\vec{r})$ to $\rho(\vec{r})$ map (Fig. 2; see also Supplementary Fig. 2). To apply the CNN algorithm to first-principles electronic structure calculations, for theoretical and practical reasons, we propose to project $\rho_0(\vec{r})$ and other atomic signatures on a 3D real-space grid. Various well-tested atomistic quantities are already available in established DFT codes, and here we employed the SIESTA package based on the confined numerical atomic orbital basis set[30] and adopted in addition to (i) the summation of atomic electron densities $\rho_0(\vec{r}) = \sum_I \rho_I^{\mathrm{atom}}(\vec{r})$, where ρ_I^{atom} is the neutral atomic density of atom I, (ii) the diffuse ion charge density

$$\rho_{\mathrm{ion}}(\vec{r}) = -\frac{1}{4\pi}\sum_I \nabla^2 V_I^{local}(\vec{r}), \tag{2}$$

where the V_I^{local} is the local pseudopotential of atom I, and (iii) the atomic orbital overlap density

$$\rho_s(\vec{r}) = \sum_{\mu\nu} \phi_\nu^*(\vec{r})\phi_\mu(\vec{r}), \tag{3}$$

where the μ, ν are the labels for atomic basis orbitals, projected on a grid with a uniform grid spacing of 0.25 Å. In Fig. 1d, in predicting the dataset as will be detailed later, we indeed find that, with the increasing number of input fingerprints applied to the initial layers, the error $\left\langle \mathcal{E}_p \right\rangle$ of trained models decrease for both training and test datasets. This result shows that each input feature contributes to the capture of the final target, demonstrating their fidelity in encoding the 3D electronic structure.

In terms of the learning target of DeepSCF, as mentioned above, we selected rather than the density ρ itself the residual density $\delta\rho(\vec{r}) = \rho(\vec{r}) - \rho_0(\vec{r})$ as the target of ML prediction[9]. First of all, it directly corresponds to the information the SCF cycle acquires. Numerically, $\delta\rho(\vec{r})$ is more smoothly distributed on a spatial grid than $\rho(\vec{r})$ (compare, e.g., Fig. 2b, c), effectively reducing the requirement on grid resolution and thus enhancing the computational efficiency. From the theoretical viewpoint, $\delta\rho$ encodes the electron transfer between constituent atoms. In other words, while the correspondence typically becomes approximate due to the modifications of neutral atomic densities, $\delta\rho$ corresponds to the chemical bond information. To validate that the learning of $\delta\rho$ instead of ρ indeed allows more accurate prediction of various chemical environments, we prepared the ρ-based architecture ① and $\delta\rho$-based architecture ② of the DeepSCF model (Fig. 1e, left panel), in the latter of which the final feature becomes the summation of ρ_0 and $\delta\widetilde{\rho}_{\mathrm{ML}}$ (dotted arrow in Fig. 1b). Testing both models, we confirmed that the $\delta\rho$-learning architecture ② gives lower $\left\langle \mathcal{E}_p \right\rangle$ values than its ρ-learning counterpart ① for both training and test datasets (Fig. 1e right panel; see Supplementary Table 2 for detail).

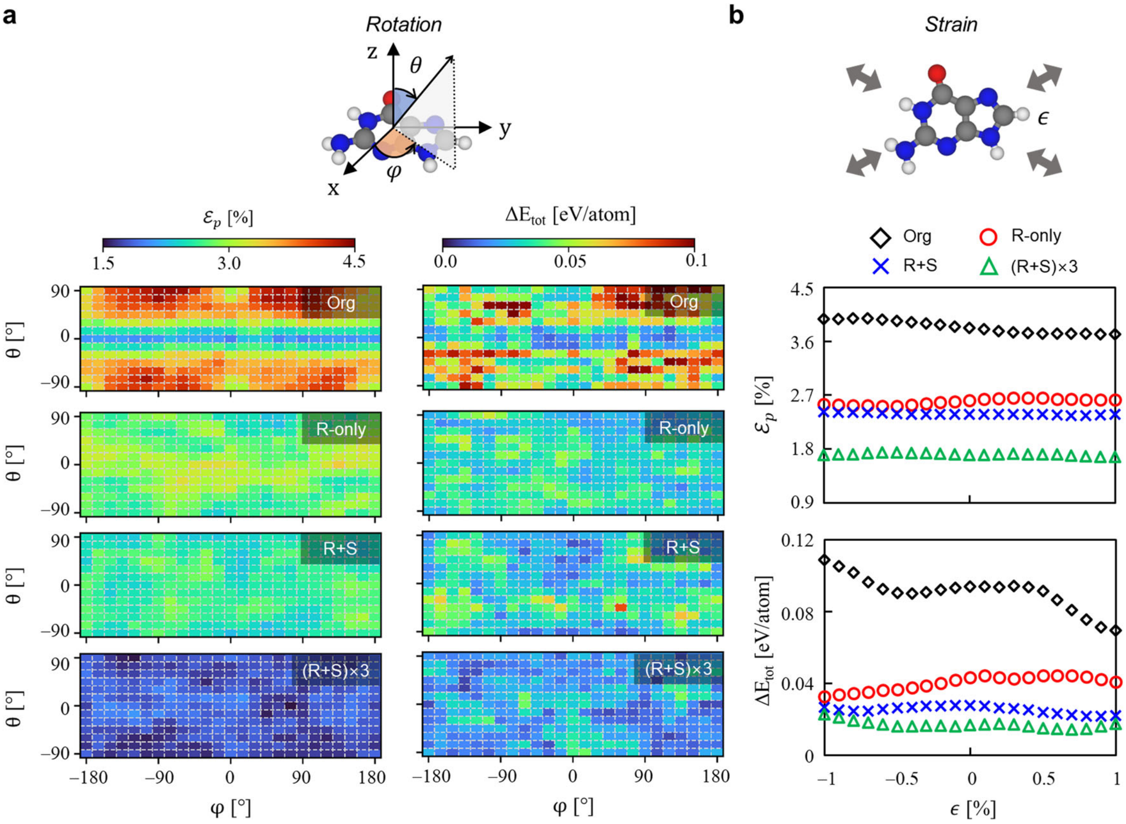


Fig. 3 | Dataset enhancement. a The $\mathcal{E}_p$ distributions (left panels) and corresponding total energy error ΔE_{tot} distributions (right panels) for rotated guanine structures are shown together with the definition of rotation directions (top panel). From top to bottom: results from the Org, R-only, R + S, and (R + S) × 3 models. **b** The $\mathcal{E}_p$ (middle panel) and ΔE_{tot} distributions (bottom panel) obtained by applying the strain ϵ ranging from −1 to +1% to the guanine structures of $\theta = -150°$ and $\varphi = -90°$ (top panel). The black diamonds, red circles, blue crosses and green triangles correspond to the results obtained from the Org, R-only, R + S, and (R + S) × 3 models, respectively.

Dataset enhancement

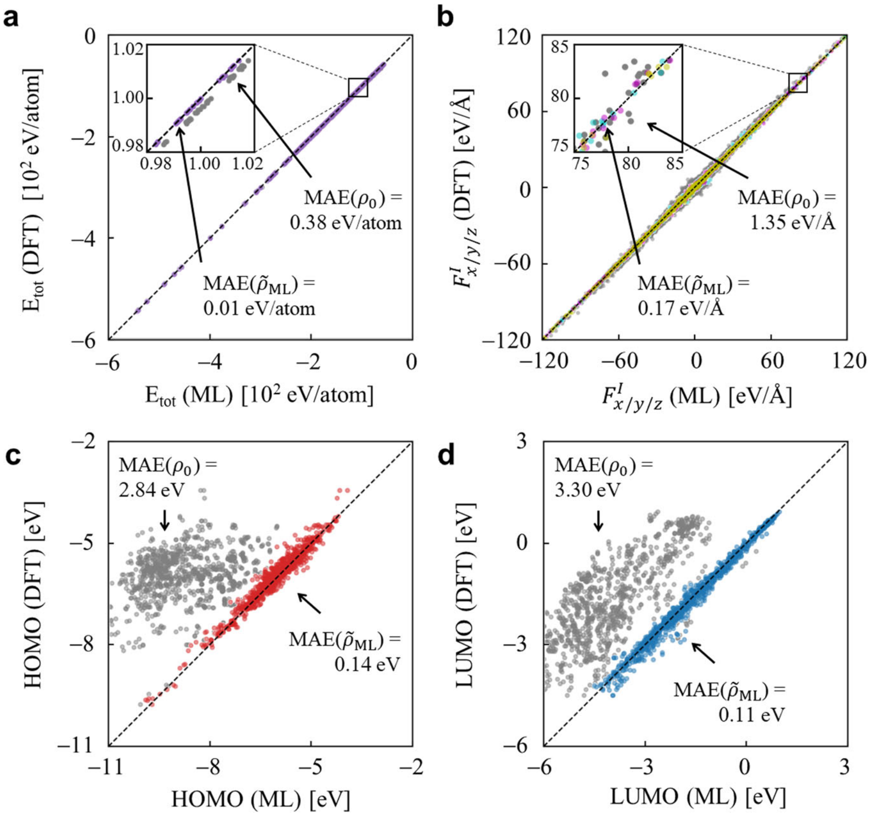
Given that we target to predict $\delta\rho$ that approximately corresponds to the chemical bond information, another essential ingredient of the DeepSCF model is the preparation of a proper dataset that can cover diverse chemical bonding cases. For this purpose, we used the TABS dataset which consists of molecules including various chemical species[31] (see "Methods" for details). However, while the CNN is equivariant to the translation of input features (Supplementary Fig. 3a), it is inherently not equivariant to the rotation of input features. Indeed, upon considering differently oriented guanine structures, we found that the model derived from the original dataset that consists of the molecular geometries lying on the *xy*-plane results in much larger $\mathcal{E}_p$ for the molecules rotated from the *xy*-plane (Supplementary Fig. 3b). Moreover, the original dataset only contains the optimized structures, from which we found that the derived ML models become inefficient in predicting the conditions beyond the equilibrium geometries (Supplementary Fig. 3c).

To address these problems, we considered the possibilities of randomly rotating and straining the ground-state geometries and additionally increasing the dataset size. To systematically demonstrate the dataset enhancement effects, in addition to the original the *xy*-plane-lying ground-state geometry dataset (Org), we prepared the comparison models that were trained on the rotated (R-only) structures and the simultaneously rotated and −2% ~ +2% strained (R + S) structures. In addition, we tripled the size of the R + S dataset by creating additional two copies of molecules with different rotations and strains [(R + S) × 3]. To test the performance of different models, we considered a guanine molecule that was not included in the TABS dataset. First, rotating the ground-state geometry guanine, we obtained for the Org, R-only, R + S, and (R + S) × 3 models the average $\mathcal{E}_p$ values of 3.18%, 2.82%, 2.59%, and 1.79%, respectively (Fig. 3a, left panels). For the total energy, we again observed the trend of systematic improvements by obtaining the error ΔE_{tot} values of 0.05 eV/atom, 0.03 eV/atom, 0.03 eV/atom, and 0.02 eV/atom for the Org, R-only, R + S, and (R + S) × 3 models, respectively (Fig. 3a, right panels). It should also be emphasized that, unlike the Org model (Fig. 3a, top panel), the enhanced R-only, R + S, and (R + S) × 3 models exhibit uniform $\mathcal{E}_p$ and ΔE_{tot} distributions over all rotation directions, representing their capability to equivariantly predict differently orientated molecules. Similarly, by straining a guanine molecule with the strain ϵ ranging from −1% to +1%, we again observed that the Org, R-only, R + S, and (R + S) × 3 models provide the average $\mathcal{E}_p$ values of 3.37%, 2.71%, 2.42%, and 1.80%, respectively (Fig. 3b), with the corresponding ΔE_{tot} values of 0.09 eV/atom, 0.04 eV/atom, 0.03 eV/atom, and 0.02 eV/atom. We thus concluded that within DeepSCF the rotation-equivariant and highly accurate prediction properties are obtained by applying random rotations and strains to the Org dataset and then increasing the dataset size. Accordingly, we will below exclusively adopt the (R + S) × 3 model (see Supplementary Fig. 4 and Supplementary Table 3 for details).

Model performance

For the test dataset, we confirmed that the $\widetilde{\rho}_{\mathrm{ML}}$ fits ρ well with $\langle \mathcal{E}_p \rangle$ of 1.67%, corresponding to 0.99 of the coefficient of determination (R^2) score. As a reference, the $\langle \mathcal{E}_p \rangle$ value of the initial condition ρ_0 was 13.83%, quantifying the drastic improvement in prediction accuracy achieved within DeepSCF.

Fig. 4 | Model performance. The posterior **a** total energy (purple), **b** F_x^I (cyan), F_y^I (magenta), and F_z^I (yellow), **c** HOMO (red), and **d** LUMO (blue) derived from the DeepSCF calculations are compared with corresponding DFT calculation results. Gray circles are the prediction results from ρ_0.



In addition, we obtained the mean-absolute-errors (MAEs) of 0.01 eV/atom, 0.17 eV/Å, 0.14 eV, and 0.11 eV, for the predicted total energy (E_{tot}), the average of atomic forces on the atoms I along the x-, y-, and z-directions (F_x^I, F_y^I, and F_z^I), the lowest unoccupied molecular orbital (LUMO), and the highest occupied molecular orbital (HOMO) levels, respectively. Again, as references, the adoptation of ρ_0 resulted in the prediction MAEs of E_{tot}, average atomic forces, LUMO, and HOMO at the levels of 0.38 eV/atom, 1.35 eV/Å, 2.84 eV, and 3.30 eV, respectively (Fig. 4a–d), quantifying the excellent accuracy of the DeepSCF approach. In Supplementary Fig. 5, we discuss the specific cases of pyridinium chloride (C_5H_6NCl) and bromopyruvic acid ($C_3H_3BrO_3$) molecules included in the test dataset.

Next, we consider the computational efficiency of the DeepSCF approach. In DeepSCF calculations, the computational time for predicting $\widetilde{\rho}_{\text{ML}}$ (blue) requires only about 0.6% of the total computational time, making the computational speed-up approximately corresponds to the number of SCF steps. In Supplementary Fig. 6, we first summarized the computational times of full-SCF and DeepSCF DFT calculations for the TABS test dataset. The average computational times were 258.6 sec. and 22.5 sec. for the full-SCF DFT (black) and DeepSCF DFT calculations (red), respectively, representing an approximately 1150% speed-up. In addition, we applied the DeepSCF approach to random-sequence stretched single-stranded DNA models with the size ranging from 128 atoms (4 nucleobases) to 2096 atoms (64 nucleobases) (see Supplementary Fig. 7). Here, as the system size and complexity increases, we observed the number of SCF steps increases from 11 to 131 and the full-SCF DFT computational time concomitantly increases from 120 sec. to 727131 sec. On the other hand, the computational time within the DeepSCF approach increased from 11 sec. to 5551 sec., representing on the average 5610% speed-up in the computational time.

Next, we consider the size-extensibility and transferability, which would be essential in predicting unseen materials beyond the molecular training dataset. As a critical advantage of adopting a CNN architecture, the DeepSCF model can handle any size of input feature, and the size-extensibility should be naturally extended to crystalline systems by applying periodic boundary conditions. In Fig. 5, we present the DeepSCF calculation results obtained for butane, polyethylene, and graphene, which are not present in the TABS dataset. Note that, while butane has a finite structure like other molecular systems in the training dataset, polyethylene and graphene are crystalline systems that are periodically extended along the one-dimensional (1D) and two-dimensional (2D) directions, respectively (Fig. 5a). Nonetheless, the accuracies of the DeepSCF-predicted $\widetilde{\rho}_{\text{ML}}$ as well as the $\widetilde{\rho}_{\text{ML}}$ -derived density of states (DOS) obtained for periodic polyethylene and graphene comparable to that for isolated butane (Fig. 5b–d).

Finally, to demonstrate the capability of handling complex, large-scale systems that involve a wide range of chemical and bonding environments, we considered the 1644-atom nano-carbon electrode-based DNA sequencer model in which a single-stranded DNA (ssDNA) composed of 30 randomly distributed nucleobases is positioned between two capped metallic (6,6) carbon nanotube (CNT) electrodes (Fig. 6a and see "Methods" for details)[32–34]. We obtain the $\widetilde{\rho}_{ML}$ prediction error $\mathscr{E}_p$ of 2.10%, which is comparable to that of 1.67% for the training dataset (Fig. 6b). Analyzing the source of errors, we predict that the performance could be further improved by adding molecules that contain phosphorus elements (Fig. 6b, red arrow) and/or weak dispersion interactions (Fig. 6b, blue arrow) to our dataset. Performing non-SCF calculation based on this $\widetilde{\rho}_{\text{ML}}$, we also evaluated the DOS and atomic forces. We find that the DOS calculated from $\widetilde{\rho}_{ML}$ are comparable to DFT data (Fig. 6c). Moreover, the $\widetilde{\rho}_{\text{ML}}$-derived atomic forces are also in excellent agreement with DFT results with the MAE of 1.35 eV/Å, 1.07 eV/Å, and 1.26 eV/Å for F_x^I, F_y^I, and F_z^I, respectively (Fig. 6d).

Discussion

To summarize, we introduced the DeepSCF scheme, which leverages the 3D U-Net CNN architecture to predict the map between 3D grid-projected atomic fingerprints and the SCF residual density δρ. We achieved the high accuracy and transferability of the DeepSCF model by expanding the 3D mesh-encoded input features to include atomistic fingerprints beyond ρ_0, targeting the learning of δρ rather than ρ itself, adopting a dataset of small-sized organic molecules featuring diverse bonding characters, and subjecting

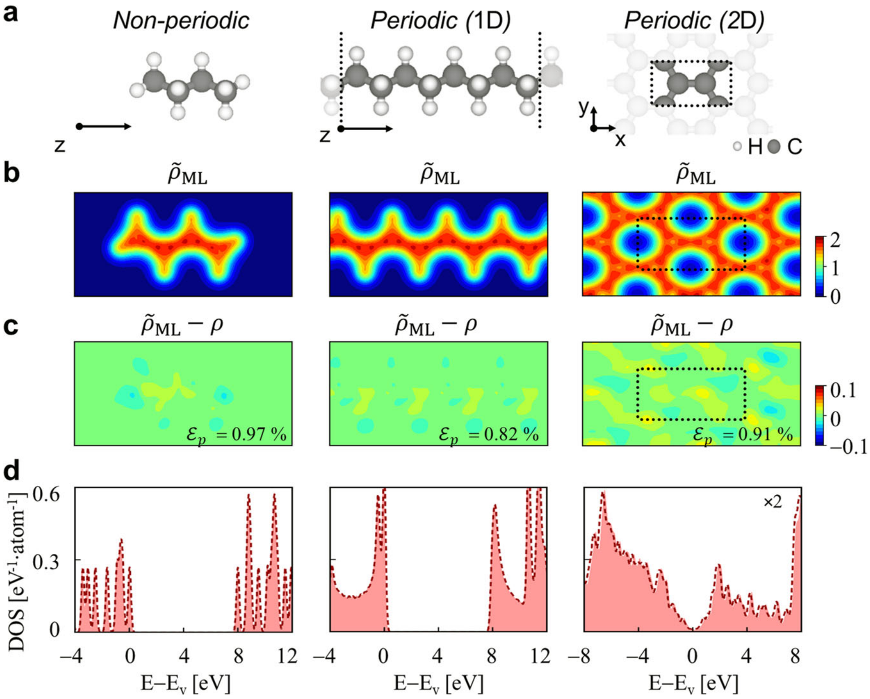


Fig. 5 | Isolated and periodic structures. **a** The butane (left), polyethylene (center) and graphene (right) structures. Two-dimensional cross-sections of **b** $\tilde{\rho}_{\mathrm{ML}}$ and **c** $\tilde{\rho}_{\mathrm{ML}} - \rho$ through the center of each structure along the y–z and x–y planes. The color scale is in units of e/Å^3. **d** The comparison between DOS obtained from DeepSCF (dashed lines) and full DFT calculations (shaded curves).

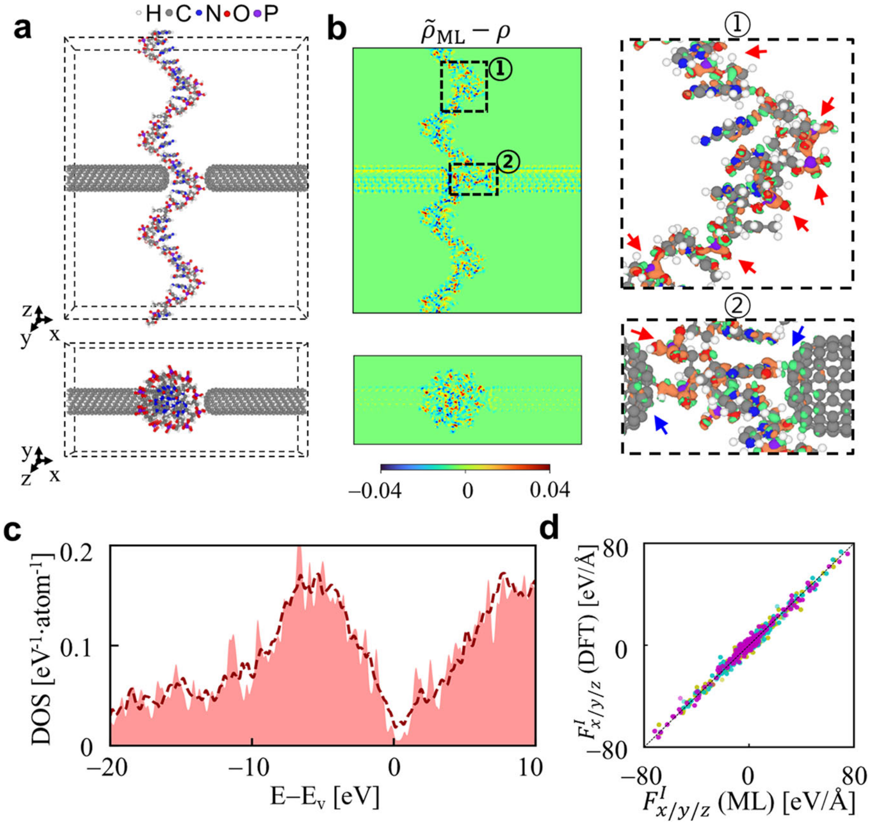


Fig. 6 | The CNT-based DNA sequencer model. **a** The DNA sequencer model was prepared by sandwiching a ssDNA between two CNT electrodes, periodically extended along the z and x directions, respectively. **b** The plane-averaged $\tilde{\rho}_{\mathrm{ML}} - \rho$ (left panel), where the color scale is in units of e/Å^3. On the right panels, for the ① and ② regions, the 3D contour plots of $\tilde{\rho}_{\mathrm{ML}} - \rho$ overlaid over the atomic structures are shown. Here, red and blue arrows highlight the positions of phosphorous atoms and the contacts between the ssDNA and CNT electrodes, respectively. The isosurface level for 3D contour plots is 0.03 e/Å^3, where the red and green colors indicate positive and negative values. **c** The comparison between DOS obtained from DeepSCF (dashed line) and the full DFT calculation (shaded curve). **d** The comparison between atomic forces on each atom I, F^I_x (cyan), F^I_y (magenta), and F^I_x (yellow), obtained from DeepSCF and the full DFT calculation.

the dataset geometries to random strains and rotations. We then demonstrated the size-extensibility and transferability of the DeepSCF model using molecules, periodic polyethylene and graphene, and a large-scale CNT-based DNA sequencer model. This development not only offers practical guidelines for applying CNNs and related computer vision ML methods to accelerate demanding DFT calculations but also establishes the correspondence between the nearsightedness in electronic structure and the spatial locality in CNNs. We thus provided insights into the mechanistic underpinnings of the success of various higher-level ML-based atomistic materials simulation strategies such as neural network forcefields. Promising follow-up research directions include, e.g., the synergistic use of DeepSCF method components in accelerating ab initio molecular dynamics simulations.

Methods

Dataset

To train the DeepSCF model, we adopted the TABS database[31], which consists of 1641 molecules with 24 different functional group categories. The molecules all contain one or more C atoms and may have one or more H, N, O, F, S, Cl, and Br atoms. For the molecules within the database, we first performed DFT geometry optimizations and created a ground-state geometry dataset. Based on the ground-state dataset, we then generated additional datasets for R-only, R + S, (R + S) × 2 and (R + S) × 3 models by randomly rotating the atomic geometries, and additionally applying the straining (R + S) and also augmenting the dataset size ((R + S) × 2 and (R + S) × 3 model). For all molecular cases, we adopted the cubic unit cell with lattice parameters $a = b = c = 20$ Å. Each dataset was divided into 80% for the training set and 20% for the test set. To examine the model performance, we additionally considered guanine, butane, polyethylene and graphene structures. For guanine, we again adopted the cubic unit cell with $a = b = c = 20$ Å. For butane and polyethylene, we adopted a tetragonal unit cell with $a = b = 20$ Å and $c = 15$ Å. The polyethylene structure was periodically connected along the c-axis direction. For the graphene structure periodically repeated along the a-b plane, we adopted a tetragonal unit cell with lattice parameters $a = 30$ Å, $b = 32$ Å, and $c = 15$ Å. For the CNT-based DNA sequencer model, we positioned a ssDNA is positioned between two capped metallic (6,6) CNT electrodes. The ssDNA structure consists of the backbone and 30 nucleobases: adenine (A), cytosine (C), guanine (G), thymine (T). As the nucleobase sequence, we adopted a randomly generated configuration (ATTAGCCGAT) and repeated it three times. The final ssDNA structure contains 984 atoms with C, H, N, O, and P elements. Each of capped (6,6) CNT electrodes contains 330 C atoms. For the CNT-based DNA sequencer model, we adopted a tetragonal unit cell with lattice parameters $a = 80$ Å, $b = 32$ Å, and $c = 96$ Å.

DFT calculations

For all above datasets, density functional theory (DFT) calculations were carried out using the SIESTA package[30] within the generalized gradient approximation (GGA)[35]. Troullier-Martins-type norm-conserving pseudopotentials[36] were employed with the numerical atomic orbital basis sets of double-ζ-plus-polarization quality. The Monkhorst–Pack $\vec{k}$-point grid of 1 × 1 × 15, 2 × 2 × 1, and 2 × 1 × 2 was sampled in the Brillouin zone for the polyethylene, graphene supercell, and nano-carbon-based DNA sequencer models, respectively. The atomic geometries of TABS database which were optimized using B3LYP functional[37–39] were reoptimized within GGA-level until the Hellmann–Feynman ionic forces acting on each atom were below 0.02 eV/Å. The posterior electronic structures were obtained from the non-self-consistent field calculations within SIESTA package. DFT calculations for the DeepSCF model development were performed on an Intel Xeon® E5620 CPU with 4 cores and 32GB of RAM. The total computational time of performing the DFT calculations for the Org, R-only, R + S, (R + S) × 2, and (R + S) × 3 molecular datasets were 421301.6, 433168.9, 433168.9, 804043.9, 1280305.9 s, respectively. For the CNT-based DNA sequencer model, DFT calculations was performed on an Intel Xeon® Gold 6226 R CPU with 16 cores and 192GB of RAM. Total computational times for the full-SCF DFT calculation and DeepSCF calculation (including $\widetilde{\rho}_{\mathrm{ML}}$ prediction step) were 60580.9 and 2873.6 s, respectively.

Model training

We implemented the DeepSCF model within Pytorch framework[40]. Then, we trained the model to training set using the ADAM optimizer[41] by minimizing the mean-square error loss function

$$MSE(\widetilde{\rho}_{\mathrm{ML}}) = \frac{1}{N}\sum_{i=1}^{N}\left(\rho^{i}_{\mathrm{DFT}} - \widetilde{\rho}^{i}_{\mathrm{ML}}\right)^{2}, \tag{4}$$

where N is the total number of models. To avoid the overfitting problem, we randomly divided each dataset into 80% of training set and 20% of test set. To determine the model architecture, we chose the hyperparameters (kernel size and the number of channels in first layers) which show the smallest test errors for the Org model trained over 50 epochs (see Supplementary Table 1 for detail). Using these optimized hyperparameters, the models (the Org, R-only, R + S, and (R + S) × 3) were trained over 100 epochs. The average computational time for training the Org, R-only, R + S, (R + S) × 2, and (R + S) × 3 molecular datasets per single epoch are 818.2, 817.7, 818.3, 1638.0, 2535.1 s, respectively, corresponding to the 0.6 seconds of average training time per molecule. All training and evaluation for DeepSCF were performed on an NVIDIA GeForce GTX 3090 GPU with 24GB of RAM.

Data availability

The TABS database can be found at https://doi.org/10.1016/j.comptc.2014.05.010. The atomic geometries of molecular datasets generated based on the TABS database and tested example materials (butane, polyphenylene, graphene, and nano-carbon-based DNA sequencer) are provided at https://doi.org/10.5281/zenodo.10073810. These atomic geometries can be used to generate the input and target features of DeepSCF using the SIESTA package which is available at https://github.com/siesta-project. The input and target feature used for the Org model at https://doi.org/10.5281/zenodo.10075554.

Code availability

The Python source code of the DeepSCF package and utility codes to post-process the input and output features are available at Github.

Received: 30 May 2024; Accepted: 6 October 2024;
Published online: 24 October 2024

References

1. Kadanoff, L. P. More is the same; phase transitions and mean field theories. *J. Stat. Phys.* **137**, 777–797 (2009).
2. De Gennes, P. G. *Superconductivity of Metals And Alloys*. (Springer, 2000).
3. Gummel, H. K. A self-consistent iterative scheme for one-dimensional steady state transistor calculations. *IEEE Trans. Electron Devices* **11**, 455–465 (1964).
4. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
5. Martin, R. M. *Electronic Structure: Basic Theory and Practical Methods*. 2nd edn (Cambridge University Press, 2020).
6. Woods, N. D., Payne, M. C. & Hasnip, P. J. Computing the self-consistent field in kohn-sham density functional theory. *J. Phys. Condens. Matter* **31**, 453001 (2019).
7. Mi, W., Luo, K., Trickey, S. B. & Pavanello, M. Orbital-free density functional theory: an attractive electronic structure method for large-scale first-principles simulations. *Chem. Rev.* **123**, 12039–12104 (2023).
8. Mills, K. et al. Extensive deep neural networks for transferring small scale learning to large scale systems. *Chem. Sci.* **10**, 4129–4140 (2019).
9. Dick, S. & Fernandez-Serra, M. Machine learning accurate exchange and correlation functionals of the electronic density. *Nat. Commun.* **11**, 3509 (2020).
10. Nagai, R., Akashi, R. & Sugino, O. Completing density functional theory by machine learning hidden messages from molecules. *npj Comput. Mater.* **6**, 43 (2020).
11. Pederson, R., Kalita, B. & Burke, K. Machine learning and density functional theory. *Nat. Rev. Phys.* **4**, 357–358 (2022).
12. Schutt, K. T., Gastegger, M., Tkatchenko, A., Muller, K. R. & Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
13. Fiedler, L. et al. Predicting electronic structures at any length scale with machine learning. *Npj Comput. Mater.* **9**, 115 (2023).
14. Hazra, S., Patil, U. & Sanvito, S. Predicting the one-particle density matrix with machine learning. *J. Chem. Theory Comput.* **20**, 4569–4578 (2024).

15. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
16. Lewis, A. M., Grisafi, A., Ceriotti, M. & Rossi, M. Learning electron densities in the condensed phase. *J. Chem. Theory Comput.* **17**, 7203–7214 (2021).
17. Briling, K. R., Fabrizio, A. & Corminboeuf, C. Impact of quantum-chemical metrics on the machine learning prediction of electron density. *J. Chem. Phys.* **155**, 024107 (2021).
18. Rackers, J. A., Tecot, L., Geiger, M. & Smidt, T. E. A recipe for cracking the quantum scaling limit with machine learned electron densities. *Mach. Learn.* **4**, 015027 (2023).
19. Alred, J. M., Bets, K. V., Xie, Y. & Yakobson, B. I. Machine learning electron density in sulfur crosslinked carbon nanotubes. *Compos. Sci. Technol.* **166**, 3–9 (2018).
20. Chandrasekaran, A. et al. Solving the electronic structure problem with machine learning. *Npj Comput. Mater.* **5**, 22 (2019).
21. Fowler, A. T., Pickard, C. J. & Elliott, J. A. Managing uncertainty in data-derived densities to accelerate density functional theory. *J. Phys. Mater.* **2**, 034001 (2019).
22. Gong, S. et al. Predicting charge density distribution of materials using a local-environment-based graph convolutional network. *Phys. Rev. B* **100**, 184103 (2019).
23. Jørgensen, P. B. & Bhowmik, A. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *Npj Comput. Mater.* **8**, 1–10 (2022).
24. Focassio, B., Domina, M., Patil, U., Fazzio, A. & Sanvito, S. Linear jacobi-legendre expansion of the charge density for machine learning-accelerated electronic structure calculations. *npj Comput. Mater.* **9**, 87 (2023).
25. Kim, Y.-H., Lee, I.-H. & Martin, R. M. Object-oriented construction of a multigrid electronic-structure code with fortran 90. *Comput. Phys. Commun.* **131**, 10–25 (2000).
26. Lee, I.-H., Kim, Y.-H. & Martin, R. M. One-way multigrid method in electronic-structure calculations. *Phys. Rev. B* **61**, 4397–4400 (2000).
27. Prodan, E. & Kohn, W. Nearsightedness of electronic matter. *Proc. Natl Acad. Sci. USA* **102**, 11635–11638 (2005).
28. Alzubaidi, L. et al. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *J. Big Data* **8**, 53 (2021).
29. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*. 234–241 (Springer, 2015).
30. Soler, J. M. et al. The siesta method for ab initio order-*n* materials simulation. *J. Phys. Condens.* **14**, 2745–2779 (2002).
31. Blair, S. A. & Thakkar, A. J. Tabs: a database of molecular structures. *Comput. Theor. Chem.* **1043**, 13–16 (2014).
32. Kim, H. S. & Kim, Y.-H. Recent progress in atomistic simulation of electrical current DNA sequencing. *Biosens. Bioelectron.* **69**, 186–198 (2015).
33. Kim, H. S., Lee, S. J. & Kim, Y.-H. Distinct mechanisms of DNA sensing based on n-doped carbon nanotubes with enhanced conductance and chemical selectivity. *Small* **10**, 774–781 (2014).
34. Jung, S. W., Kim, H. S., Cho, A. E. & Kim, Y.-H. Nitrogen doping of carbon nanoelectrodes for enhanced control of DNA translocation dynamics. *ACS Appl. Mater. Interfaces* **10**, 18227–18236 (2018).
35. Perdew, J. P. Accurate density functional for the energy: real-space cutoff of the gradient expansion for the exchange hole. *Phys. Rev. Lett.* **55**, 1665–1668 (1985).
36. Troullier, N. & Martins, J. L. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B* **43**, 1993–2006 (1991).
37. Kassel, L. S. The limiting high temperature rotational partition function of nonrigid molecules I. General theory. II. Ch4, c2h6, c3h8, ch(ch3)3, c(ch3)4 and ch3(ch2)2ch3. III. Benzene and its eleven methyl derivatives. *J. Chem. Phys.* **4**, 276–282 (1936).
38. Lee, C., Yang, W. & Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
39. Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
40. Paszke, A. et al. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
41. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *The 3rd International Conference on Learning Representations* (ICLR, 2015).

Acknowledgements

This work was supported by the National Research Foundation of Korea (2023R1A2C2003816 and RS-2023-00253716). Computational resources were provided by KISTI Supercomputing Center (KSC-2023-CRE-0476).

Author contributions

Y.-H.K. formulated and oversaw the project. R.G.L. developed the computational framework and carried out calculations. Y.-H.K. and R.G.L. analyzed the computational results and co-wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01433-0.

Correspondence and requests for materials should be addressed to Yong-Hoon Kim.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.