

ARTICLE OPEN



A deep learning framework to emulate density functional theory

Beatriz G. del Rio^{1,2}✉, Brandon Phan¹ and Rampi Ramprasad²✉

Density functional theory (DFT) has been a critical component of computational materials research and discovery for decades. However, the computational cost of solving the central Kohn–Sham equation remains a major obstacle for dynamical studies of complex phenomena at-scale. Here, we propose an end-to-end machine learning (ML) model that emulates the essence of DFT by mapping the atomic structure of the system to its electronic charge density, followed by the prediction of other properties such as density of states, potential energy, atomic forces, and stress tensor, by using the atomic structure and charge density as input. Our deep learning model successfully bypasses the explicit solution of the Kohn–Sham equation with orders of magnitude speedup (linear scaling with system size with a small prefactor), while maintaining chemical accuracy. We demonstrate the capability of this ML–DFT concept for an extensive database of organic molecules, polymer chains, and polymer crystals.

npj Computational Materials (2023)9:158; <https://doi.org/10.1038/s41524-023-01115-3>

INTRODUCTION

Density functional theory (DFT)^{1,2} has become one of the most valuable computational tools for the materials research community. It has guided the discovery of new catalysts^{3,4}, the design of materials for energy storage^{5–8}, and the exploration of material behavior under extreme conditions^{9–11}, among other applications. The success of DFT lies in the transformation of the cumbersome many-electron many-nuclear problem of quantum mechanics to an effective one-electron Kohn–Sham (KS) equation². Solving the KS equation for a material with a given atomic configuration provides information about the ground state electronic structure of the system in the form of one-electron wave functions (or charge density) and one-electron eigenvalues (or density of states). These quantities, i.e., either the wave functions and eigenvalues or the charge density and density of states, are the most essential and complete information of the material from which a host of properties can be computed, such as the potential energy, atomic forces, and stress tensor. DFT-based research has seen several advancements over the last several decades in the areas of theory, algorithms, and computational infrastructure, instrumental in the above-mentioned discoveries. Nevertheless, practical and routine DFT calculations of complex materials involving several thousands of atoms to probe phenomena that occur over timescales of the order of nanoseconds or longer remain inaccessible.

Over the last decade, machine learning (ML) based approaches are actively being considered in various ways to meet the length- and time-scale demands encountered during DFT computations. ML provides a powerful pathway to replace a cumbersome or expensive “input–output” problem with a cheap “surrogate” model. The accuracy and versatility of such models depend on the number and diversity of input–output examples the model has seen before and the internal architecture of such models. The past decade has seen several successful ML efforts applied to various material properties and application spaces^{12–26}.

The present contribution attempts to provide an efficient emulation of DFT by treating the KS equation itself as an

input–output problem. Of relevance to the present contribution are our own past work^{27,28}, in which the problem in question was addressed to a limited level, and the recent work by Brockherde et al.²⁹. The latter work attempts to bypass explicitly solving the KS equation using a plane-wave basis representation of the electron density. While the accuracy of the model was demonstrated for small molecules, it was not transferable to large systems. Since this first attempt, two main methodologies have been investigated to predict the charge density: grid-based schemes^{27,30–32} and atom-based representations in terms of basis functions^{33–37}. The main advantage of a grid-based approach is the high accuracy obtained and general applicability to localized and delocalized electron densities. However, no information about individual atomic charges can be retrieved, and the high computational cost hinders its applicability to large databases. On the other hand, predicting the charge density as atomic contributions in terms of a basis set significantly reduces the computational cost and provides information on the individual atomic charges at the cost of lower accuracy, especially in systems with a delocalized electron density. One important advantage of atom-based representations is the higher transferability to new and larger systems, which is essential for a successful deployment of the model. Nevertheless, challenges still remain with respect to achieving comparable model performance for larger systems relative to smaller systems used during the training phase. Although methods have been developed to varying degrees of success to predict either the electronic structure or basic atomic properties such as total potential energy, atomic forces, or stress tensor, there is yet no scheme that has successfully unified simultaneous prediction of both types of properties in a comprehensive KS–DFT emulation.

In this work, the KS equation is handled in an alternative manner using a deep learning scheme, both in terms of methodological advancements and applicability, which predicts the electron density first and then employs it as an additional descriptor of the material to further predict other electronic and atomic quantities. The electronic quantities predicted other than the charge density are the density of states, valence band

¹Departamento de Física Teórica, Atómica y Óptica, Universidad de Valladolid, Paseo de Belén 7, 47011 Valladolid, Spain. ²School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Drive NW, Atlanta, GA 30332, USA. ✉email: beatriz.gonzalez.rio@uva.es; rampi.ramprasad@mse.gatech.edu

maximum (VBM), conduction band minimum (CBM), and band gap (E_{gap}). The atomic or global quantities predicted are the total potential energy, atomic forces, and stress tensor, essential for molecular dynamics (MD) simulations. Our scheme also expands the flexibility and transferability of the model, allowing for training on molecules, 2D, and 3D systems within a large chemical space composed of carbon (C), hydrogen (H), nitrogen (N), and oxygen (O). Overall, the present contribution allows for a near-complete DFT emulation within a practical context, surpassing previous works^{27,33,34,37,38} in terms of methodological advancements as well as expanding the portfolio of predicted quantities.

Figure 1 shows several critical components which are part of our ML workflow. We created a reference, or “training,” data composed of molecules, polymer chains, and polymer crystal structures containing C, H, N, and O, and their corresponding properties computed using traditional DFT. Each reference atomic configuration is then represented using an ML-friendly atomic fingerprinting scheme. In this work, we used the atom-centered AGNI fingerprints^{39,40}, which represent the structural and atomic-level chemical environment of each atom in a machine-readable form such that it is translation, permutation, and rotation invariant. Because not all properties predicted are rotation invariant, i.e., electron density, atomic forces, and stress tensor, we define an intermediate internal reference system that allows us to easily transform any quantity to its value in the reference system of choice. To establish a direct (and nonlinear) mapping of the fingerprints (the “input”) to the spectrum of properties mentioned earlier (the “output”), we used deep neural networks.

Inspired by DFT, we give particular prominence to the electronic charge density and follow a two-step learning procedure. The first learning problem (step 1) involves predicting the electronic charge density given just the atomic configuration. Our protocol employs Gaussian-type orbitals (GTOs) as descriptors of the electronic charge density, but we do not use a predefined basis set; the model learns the most optimal basis from the data examples, thus expanding the flexibility of the model. Once the electronic charge density descriptors have been predicted, they are supplied as an auxiliary input (along with the atomic configuration fingerprints) to predict all other properties (listed as step 2 of Fig. 1). This strategy is consistent with the core concept underlying DFT (that the electronic charge density determines all properties of the system), and is aligned with the first rudimentary ML attempt almost a decade ago in which a variety of properties were predicted given just the electronic charge density⁴¹. Furthermore, in practice, this 2-step route also leads to more accurate and transferable results, as we will show below.

RESULTS

Database

In this work, we focused on organic materials composed of four atoms: C, H, N, and O. We created a database containing 67 molecules, 178 polymer chains, and 55 polymer crystals composed of C-C single, double, and triple bonds, as well as aromatic rings. To provide the neural network with sufficient examples of configurational diversity, we procured random snapshots of each type of structure from DFT-based MD runs at high temperatures. For the molecules and polymer chains, the MD runs were performed at 300 K, and for the polymer crystals, the MD runs involved temperatures from 100 to 2500 K. In total, we used over 118,000 structures for the training and testing of ML-DFT. For each type of structure, we divided the selected configurations into training and test sets, following a 90:10 split. Additionally, the models were trained using an 80:20 split of the training set between training and validation. All performance results were computed using the independent test set. More details can be found in the Supplementary Information (SI). All the DFT reference

data calculations were performed using the Vienna Ab Initio Simulation Package (VASP)^{42,43}.

Fingerprinting

Within this work, we have employed two different types of fingerprints or descriptors: atomic (or structural) fingerprints and electron charge density descriptors. The atomic descriptors are the AGNI atomic fingerprints, which describe the structural and chemical environment of each atomic configuration. These previously developed AGNI atomic fingerprints have been used to create ML potentials and force fields^{44–46} for a variety of materials as well as in our previous work on predicting the grid-based electronic structure²⁷ and the atom-based density of states²⁸. The atomic fingerprints, computed for each atom, combine scalar, vector, and tensor-like expressions by summing over various Gaussian functions, resulting in translation, permutation, and rotation invariant descriptors.

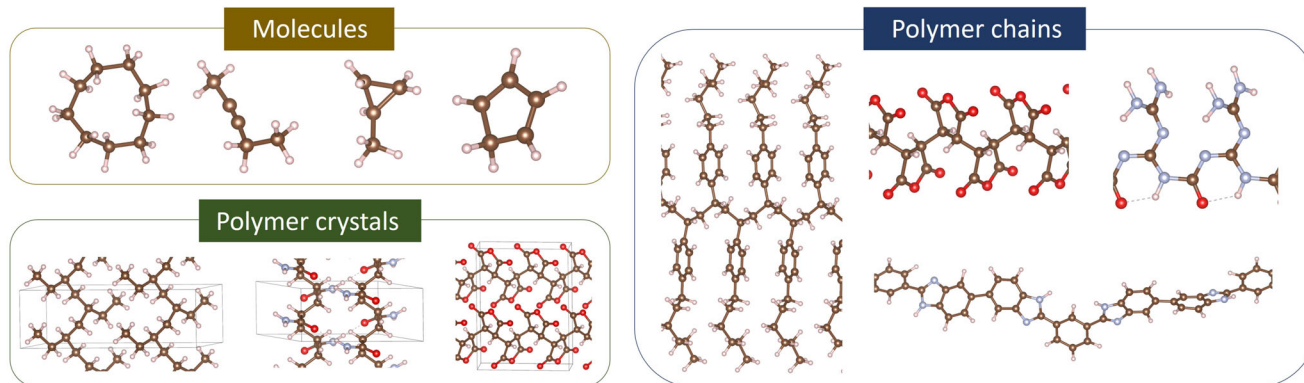
The predicted electronic charge density descriptors constitute the second type of fingerprints used in this work. Once all the configurations have been fingerprinted, the AGNI atomic fingerprints are used as input for the charge density model, which predicts the decomposition of the atomic charge density in terms of GTO basis functions. The set of optimal GTO basis functions is selected by the model in terms of the exponent of the Gaussian along with the constant multiplying the GTO; no initial decomposition of the reference charge density in terms of GTOs is performed. The error made during the training of the model is computed by projecting each set of atomic GTO basis functions onto the same grid points used for the reference DFT charge density. As the input atomic fingerprints are translation, permutation, and rotation invariant, the predicted constants and exponents of the GTO basis functions decomposition are within the internal reference system of the atom. Because the reference electron charge density is not rotation invariant, a transformation from each atom’s internal reference system to the global reference system of the electron charge density (the Cartesian system) is required before projecting the predicted charge density onto the grid points. The vectors for the transformation matrix of each atom are defined using the two nearest neighbors (independent of the element type): the first vector is the one pointing to the first nearest neighbor, and the second vector is defined as perpendicular to the plane containing the central atom and its two nearest neighbors, and the third vector is perpendicular to the first two vectors. All vectors are normalized to obtain an orthonormal reference system (more details in the SI). The resulting transformation matrix is used to convert from the orthonormal internal reference system of each atom onto the Cartesian reference system, allowing the transformation of any rotation-invariant value of a property (such as the decomposition onto GTOs) onto the Cartesian reference system. Because of the computational cost of projecting the GTOs onto grid points, we restricted the training to structures with up to 50 atoms per element. Once the model is trained, the predicted constants and exponents for each atomic fingerprint will be referred to as charge density descriptors. Unlike atomic fingerprints, which are determined by a set of predefined equations, these charge density descriptors are learned by the neural network and provide an electronic description of the system. More details on the charge density model can be found in the SI.

Charge density prediction

To study the performance of the charge density model, we computed the mean absolute percentage error (ϵ_{ρ}) for each configuration as

$$\epsilon_{\rho}(\%) = 100 \cdot \frac{\sum_j |\rho_{\text{DFT}}(r_j) - \rho_{\text{ML}}(r_j)|}{\sum_j \rho_{\text{DFT}}(r_j)} \quad (1)$$

a) Types of structures in ML-DFT database



b) Initial step: fingerprinting

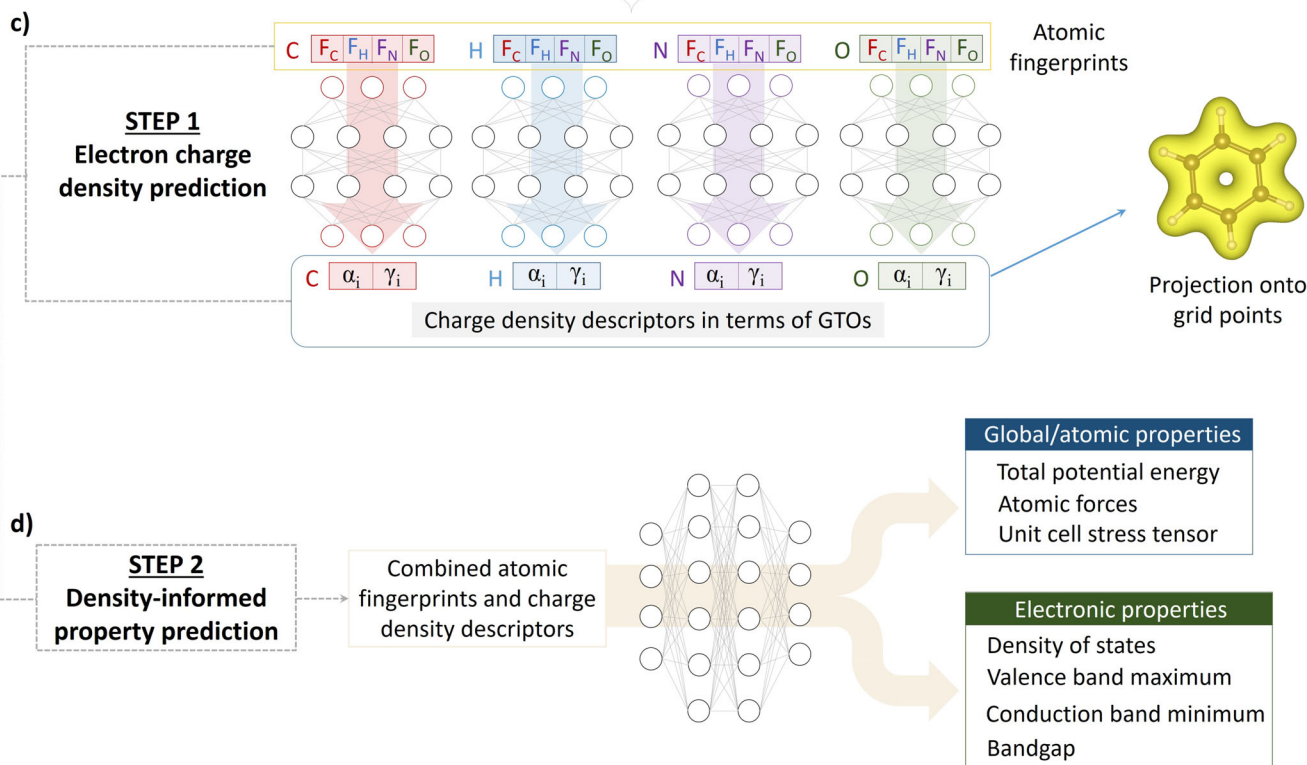
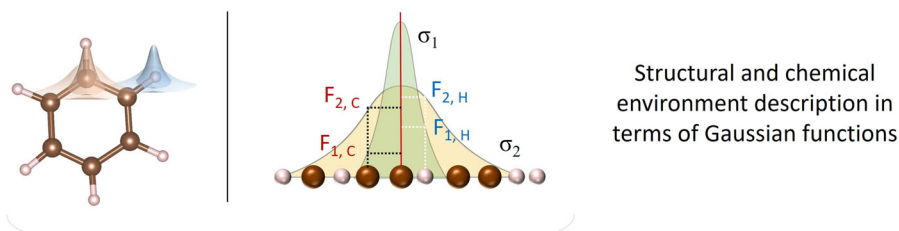


Fig. 1 ML-DFT database and two-step workflow. **a** The reference database contains DFT data from organic molecules, polymer chains, and polymer crystals. After creating the database, the atomic configurations are fingerprinted to describe the structural and chemical environment of each atom (**b**). Within step 1 (**c**), the resulting atomic fingerprints are used as the input layer to predict the electronic charge density in terms of various Gaussian-type orbitals (GTOs) descriptors. The projection of these GTOs onto grid points provides the charge density. In step 2 (**d**), the combined atomic fingerprints and charge density descriptors serve as input for the prediction of other DFT properties such as potential energy, atomic forces, stress tensor, density of states, valence band maximum, conduction band minimum, and the bandgap.

where $\rho_{\text{DFT}}(r_j)$ and $\rho_{\text{ML}}(r_j)$ are, respectively, the reference DFT charge density and ML-DFT charge density at grid point j , for the same configuration.

The accuracy of the charge density model can be observed in Fig. 2a, where ϵ_ρ for the test configurations of the training set ranges mainly from 1.0% to 3.0%, with a few cases extending up

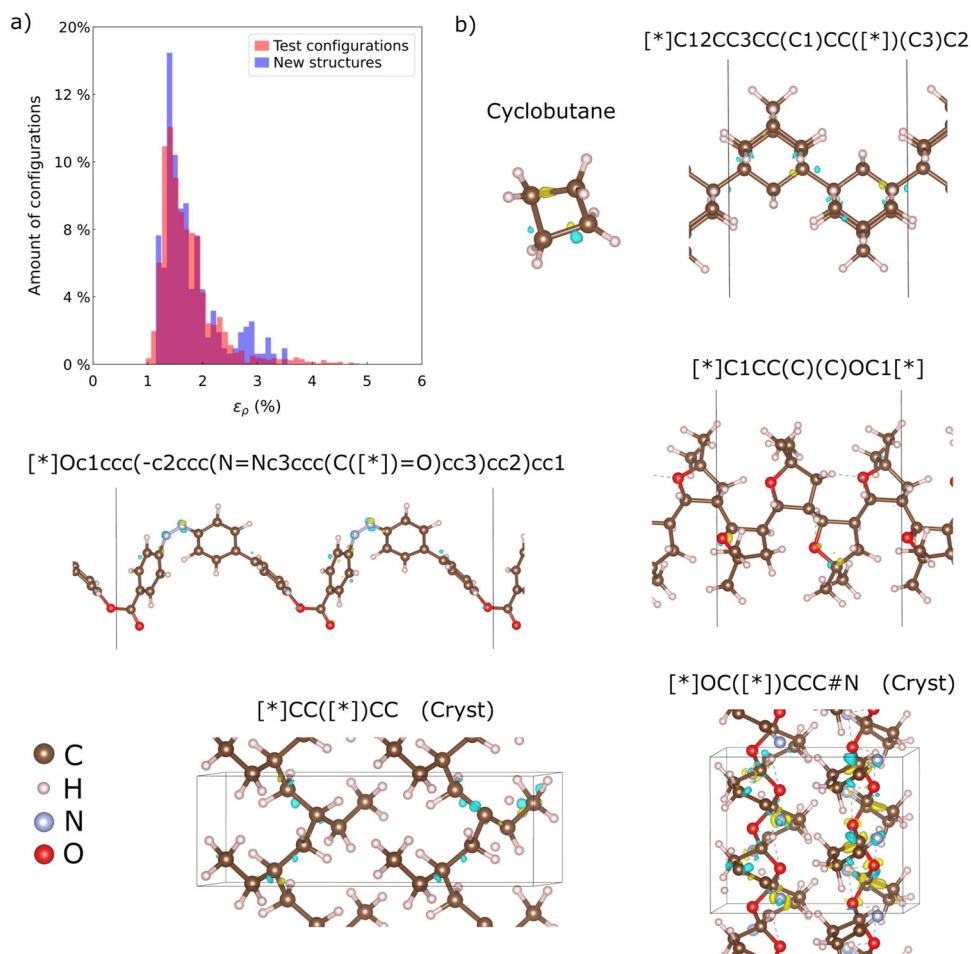


Fig. 2 Performance of ML-DFT charge density model. **a** Histogram of the mean absolute percentage error, ϵ_p , for the charge density on the test configurations of the training and for new larger structures. The amount of structures in each bar is indicated as a percentage with respect to the total in each set. **b** Charge density difference between DFT and ML-DFT for a molecule (cyclobutane), three polymer chains, and two crystalline polymers (Cryst). Cyan and yellow isosurfaces refer to an error of ± 0.005 e bohr⁻³.

to 5.0%. Most notably, the performance on the new structures (more than 50 atoms per element) not included during training is very similar to those structures used for training. An overall performance value can be calculated as the mean $\bar{\epsilon}_p$ computed by summing ϵ_p for all configurations and normalizing by the number of electrons in each configuration. For the test configurations, $\bar{\epsilon}_p$ is 1.75%, increasing slightly to 1.97% for the new structures. Figure 2b presents the valence charge density difference between the reference DFT and predicted ML-DFT for various atomic structures: the cyclobutane molecule, three different polymer chains, and two polymer crystals. The cyan and yellow isosurfaces refer to an error of ± 0.005 e bohr⁻³. The isosurfaces occupy a very small volume due to the high accuracy of the predicted ML-DFT charge density.

Various other charge-density dependent properties can be calculated from the predicted charge density and charge density descriptors, such as the partial atomic charges, the reduced density gradient for the analysis of non-covalent interactions, and the dipole moment. Examples and comparisons with DFT results are included in the SI. In addition, to further extend the applicability of our method to situations requiring the full electron density, ML-DFT also provides the core electron density by mapping it to the DFT reference using 1s orbitals, with an accuracy of around $\bar{\epsilon}_{p,core}(\%) = 5 \cdot 10^{-5}$. When the full electron density is used, the error of the predicted total electron density is reduced by 27% overall, with $\bar{\epsilon}_{p,full} = 1.28\%$ on test configurations and $\bar{\epsilon}_{p,full} = 1.44\%$ on new structures. More details in the SI.

Important previous ML work to predict the charge density-based atomic contributions required an initial decomposition onto predefined basis sets, introducing an additional error³⁴. The ML model was trained to predict the components of the decomposition for each atom, resulting in good predictions of the charge density for cases within the training space but leading to a lack of transferability to new cases. The authors used the full charge density instead of only the chemically active valence charge density used in our work. This difference results in their work presenting lower errors in the charge density prediction, as the high-valued core charge density is easily predicted and effectively lowers the percentage errors. Moreover, the requirement for already available basis sets to decompose the charge density can become a hindrance to the applicability of this method to some elements. Comparison with more recent work predicting the valence electron density using an atom-centered approach^{35,36}, shows our method has better accuracy. Another important comparison with recent work is the high error cases within the training space, which in this study extend up to a maximum of 5.02%, whereas in³⁷ they reach up to 11%. Overall, our ML-DFT charge density surpasses previous methods in terms of accuracy and/or methodology, simplifying the protocol and, in the process, extending the applicability and transferability.

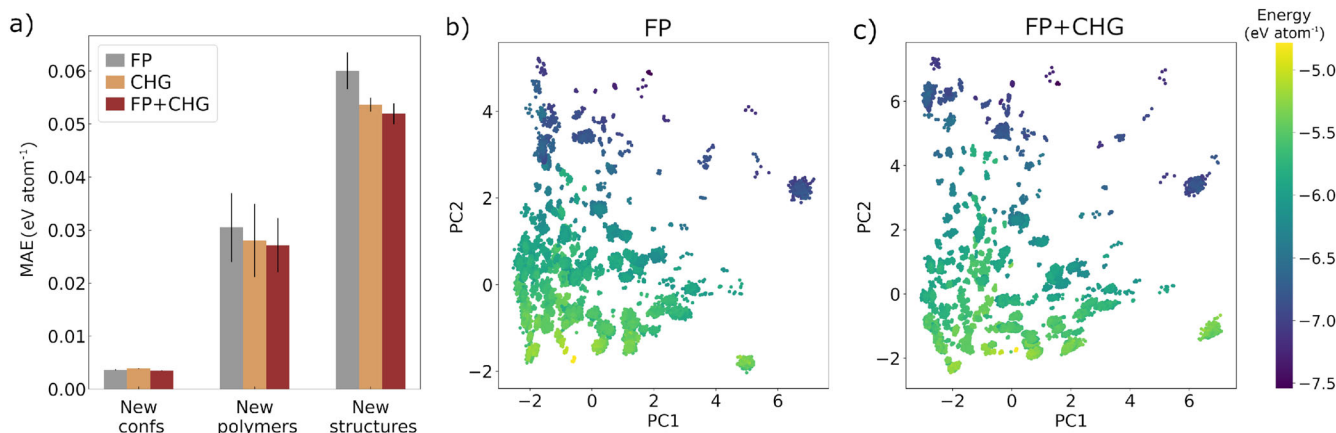


Fig. 3 Effect of using charge density descriptors along with atomic fingerprints. **a** Histogram of MAE from the fivefold cross-validation for the potential energy prediction with three different input descriptors: (1) the atomic fingerprints (FP); (2) the charge density descriptors (CHG); (3) the atomic fingerprints along with the charge density descriptors (FP + CHG). Plots of the two main principal components (PC) from only using **b** the atomic fingerprints and using **c** both the atomic fingerprints and the charge density descriptors. The points are colored with respect to the total potential energy.

Total potential energy, atomic forces, and stress tensor

DFT posits that the ground state charge density has a one-to-one mapping with the ground state potential energy. Similarly, in our ML-DFT emulator, once the atomic charge density descriptors are predicted, they can be used as input (along with the AGNI atomic fingerprints) to predict the potential energy, atomic forces, and stress tensor. To evaluate the improvement in the accuracy and transferability of the potential energy model by including the charge density descriptors in the input layer, we considered three different options: using the atomic fingerprints only, the charge density descriptors only, and the atomic fingerprints and charge density descriptors together. We only used the polymer chains for training and left the polymer crystals to test the transferability to new structures. We performed fivefold cross-validation and evaluated the performance on test configurations for the polymer chains used during training (a.k.a. new configurations), for new polymer chains (a.k.a. new polymers), as well as for polymer crystals (a.k.a. new structures).

Figure 3a shows the histogram with the mean and standard deviation of the mean absolute error (MAE) value of the total potential energy per atom of each type of descriptor on the test configurations. The combined atomic fingerprints and charge density descriptors not only improve the accuracy and transferability but also reduce the deviation of the predictions, resulting in more robust models. As can be seen in the principal component analysis plots in Fig. 3b, c, the addition of the charge density descriptors results in a better separation of the structures with different potential energies, thus improving the prediction capabilities of the model.

In our approach, the potential energy, atomic forces, and stress tensor components are each predicted directly (without using one to derive the others) by employing the same transformation matrix from the charge density model to transform the atomic forces and stress tensor components into the Cartesian reference system. This allows a significant reduction of errors in the atomic forces and stress tensor components while also improving the transferability to new structures. More details on the model and a quantitative test showing the improved performance can be found in the SI.

After confirming the advantage of employing the fingerprints and the charge density descriptors together, we trained the energy, forces, and stress tensor model on the entire training set of molecules, chains, and crystals. Figure 4 shows the performance of the model on the test configurations for the atomic potential energy per atom (Fig. 4a), the stress tensor components (Fig. 4b),

and the atomic forces (Fig. 4c). Both the potential energy and stress tensor components are predicted with great accuracy, with an MAE of 3.3 meV atom⁻¹ for the potential energy, and a mean root-mean-squared error (RMSE) of 6.42 kB for the diagonal stress components. However, the predicted atomic forces present a mean RMSE of 0.759 eV Å⁻¹, with the C atomic forces presenting larger deviations from the reference DFT values than the other elements studied. This deviation is mainly observed in the polymer crystals, as observed in Fig. 4d; from the histogram of the error in the predicted forces, most of the errors are contained within ± 1 eV Å⁻¹. There are very few instances with errors larger than ± 3 eV Å⁻¹, with the highest error obtained at ~ 11 eV Å⁻¹.

Some possible reasons behind these high errors in the atomic forces could be attributed to the insufficient sampling of highly disordered structures present in the crystal polymers, as well as the inability to capture non-local effects, such as the long-range van der Waals dispersion forces, using local atomic fingerprints. Similar results with large errors in the atomic forces have been reported in other previous studies on machine-learned atomic potentials and force fields for pure carbon structures^{47–49} and carbon-containing structures⁵⁰, where various methods of database optimization are used, such as active learning. These methods could be used with ML-DFT to improve the performance of the atomic forces but are out of the scope of this study.

The density of states predictions

As previously mentioned, the solution of the KS equation in DFT describes the electronic structure of the system in the form of the charge density and the DOS. This last property is essential for computing multiple electronic properties of the system, such as the VBM, CBM, and E_{gap} . Using a similar approach as previously described for the potential energy, the ML-DFT DOS predictor also employs as input both atomic fingerprints and charge density descriptors. Following previous work from our group^{27,28}, the reference DOS curve is previously shifted with respect to the reference energy of vacuum and discretized every 0.1 eV from -33 eV to 1 eV. Due to this constraint, we only trained the model using vacuum-containing structures: molecules and polymer chains. To achieve the highest accuracy in the VBM and CBM, and consequently the bandgap, their DFT reference values are obtained directly from the (shifted) eigenvalues and not from the smeared DOS. Due to the intensive nature of the VBM and CBM, the entire DOS/VBM-CBM model first predicts the smeared DOS curve as the sum of each atomic contribution, normalizes the total

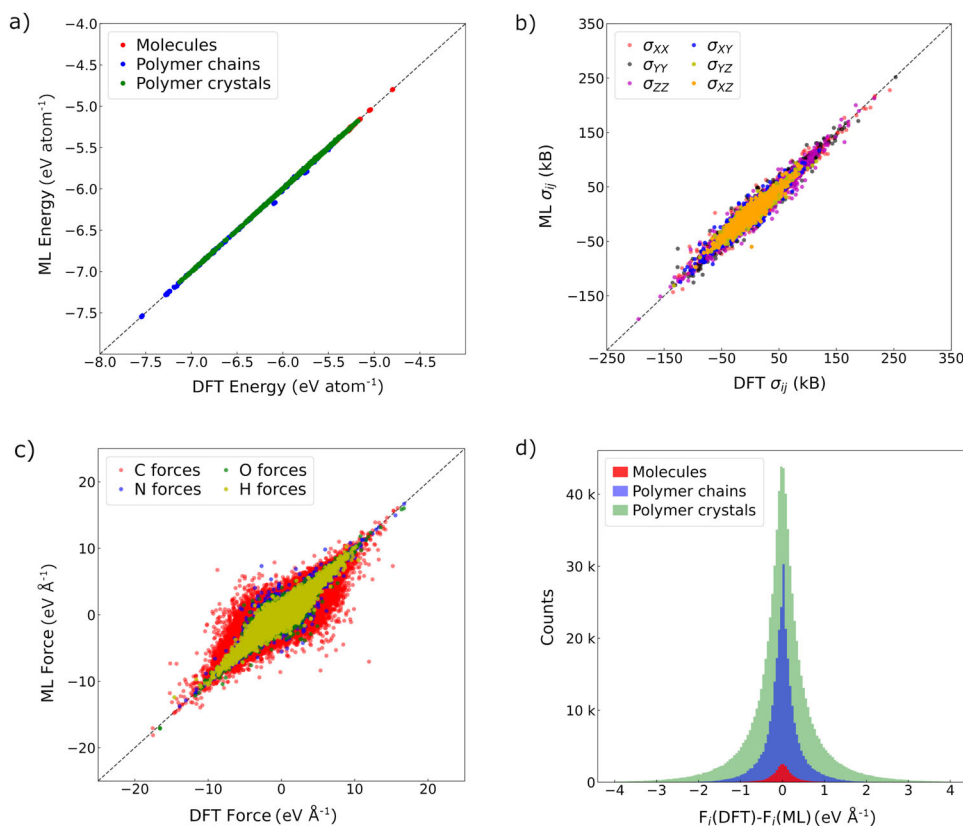


Fig. 4 Performance of the energy, stress tensor components, and atomic forces from the ML-DFT model for the test configurations of the entire database of molecules, polymer chains, and polymer crystals. **a** Parity plot of the potential energy per atom. **b** Parity plot of the six different components of the stress tensor. **c** Parity plot of the atomic forces for each type of element. **d** Histogram of the error between the reference atomic forces ($F_i(\text{DFT})$) and the predicted atomic forces ($F_i(\text{ML})$).

DOS with respect to the number of valence electrons, and feeds it to a second sub-neural network which predicts the VBM and CBM (additional details in the SI). Figure 5a–f displays six different DOS for test configurations of different molecules and polymer chains. As can be observed, the test cases possess a large variety of DOS curves. Nevertheless, the ML-DFT DOS model can accurately predict them. Also, the accuracy of the predicted DOS is high enough to evaluate differences due to the atomic movement (see details in SI).

As with electron density prediction, previous works on predicting the DOS are divided between a grid-based scheme^{27,32} and an atom-centered approach^{28,51–53}. Focusing on the more recent studies, Ellis et al.³² predict the DOS through the LDOS for liquid and solid Al using a grid-based approach. The DOS obtained from the predicted LDOS shows very good accuracy, but the DOS of Al presents a smooth shape with very small variations, even from solid to liquid. However, the computational cost of using a grid-based approach significantly hinders the use of the method for large databases. Kong et al.⁵³ use a significantly different and much more advanced method to represent the atomic structure based on graph neural networks along with an encoder–decoder technique. The method learns the similarity between crystalline structures, which is then translated into the DOS. One advantage of this technique is probably the lower cost when applied to diverse chemistries. However, it seems focused on crystalline structures, and its application to slightly unrelaxed structures is dependent upon the atomic coordinates being sufficiently similar to any of the DFT-relaxed structures used during training.

Figure 5a–f also includes the DFT and ML-DFT VBM and CBM. Due to the use of the DFT eigenvalues for VBM and CBM along with the smeared DOS as a reference, the locations of DFT and ML-

DFT VBM and CBM are not at zero-valued DOS. As can be observed from the parity plots between the DFT reference and the predicted ML-DFT values of VBM, CBM, and E_{gap} , in Fig. 5g–i, the ML-DFT values are in agreement with the reference DFT eigenvalues within MAEs of 0.069, 0.051, and 0.08 eV, respectively.

As a final note, in Fig. 6, we compare the computational cost of DFT with our ML-DFT approach for various structures of different sizes. Both types of calculations (DFT and ML-DFT) were performed in serial mode on one core of a Ryzen 9 5900X node. The total CPU time cost of DFT is significantly higher and has a cubic dependence on the system size. However, the total time for the electronic structure prediction with our ML-DFT is orders of magnitude lower than DFT with a linear dependence on system size. The ML-DFT model depends on the number of element types in the system: the red squares represent cases with only carbon and hydrogen; the red star is for a polymer chain with three elements (carbon, hydrogen, and oxygen); and the red cross represents the case of a polymer crystal with all four elements.

DISCUSSION

This work represents an important step toward a physically-informed ML-based DFT emulator, which successfully, accurately, and simultaneously reproduces many of the outputs of the KS equation. Following the essence of DFT, material properties are determined by the descriptors of the structure and the predicted charge density, resulting in increased accuracy with respect to traditional ML potentials for a fraction of the computational cost of traditional DFT.

To represent the charge density with physically-informed descriptors, we employed GTOs, which fully adapt to each

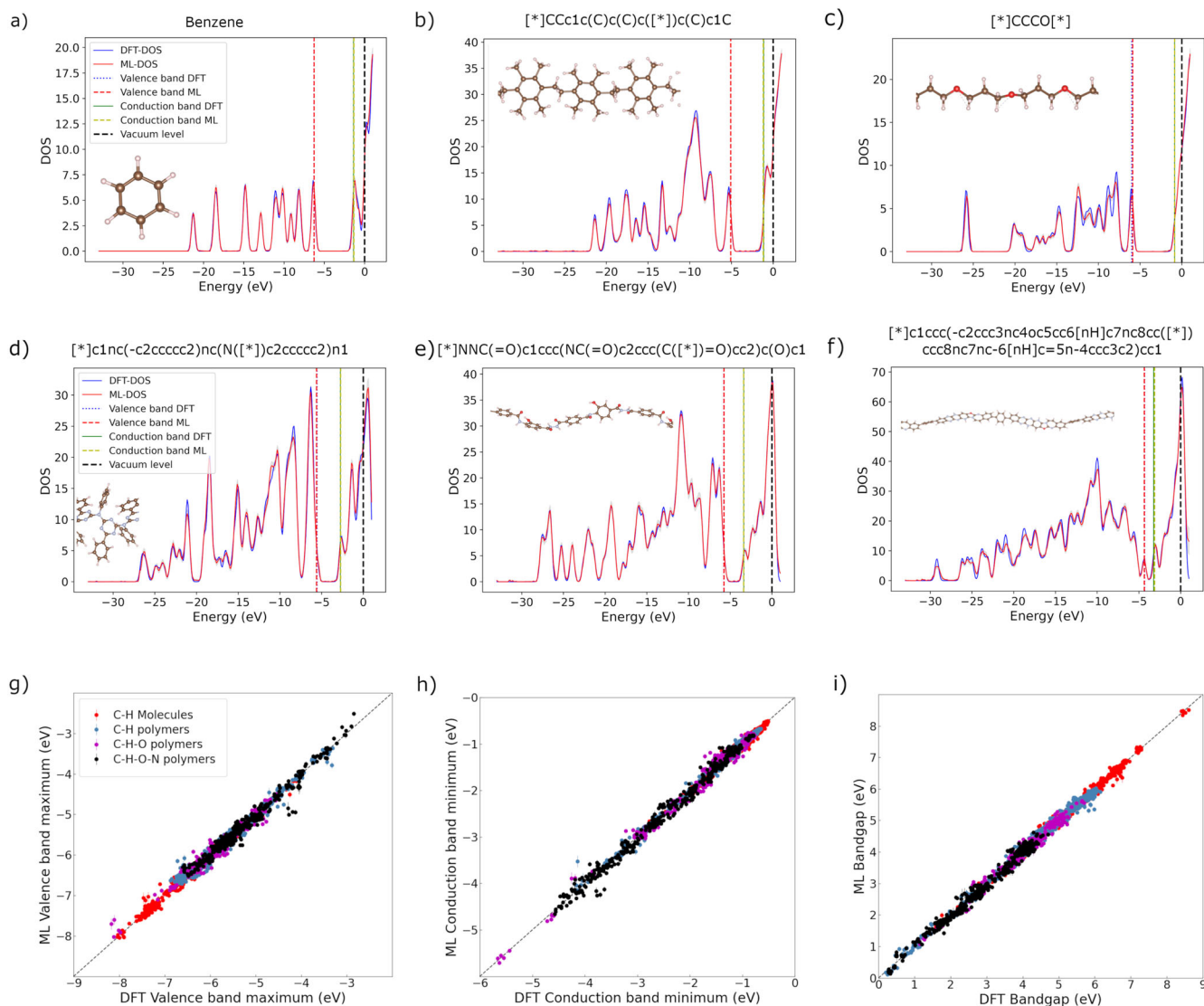


Fig. 5 Performance of the ML-DFT DOS model. a–f DFT DOS (blue) and ML-DFT DOS (red) for test configurations of six different molecules and polymer chains. The DFT and ML-DFT VBM/CBM predictions are included. The vertical dashed dark green line indicates the vacuum energy used as the global energy reference. The gray shadow indicates the standard deviation in the predicted DOS curves due to the dropout layers employed in the ML-DFT DOS model. g–i Parity plots of the ML-DFT VBM, CBM, and resulting E_{gap} for the test configurations. Vertical lines indicate the standard deviation.

individual atom and can be used for any element type. The resulting descriptors contain physical information about the system in two ways. First, as descriptors of the electronic structure of the system, which can be employed to further predict density-dependent properties. We demonstrate this density-enhanced protocol for the atomic properties of energy, atomic forces, and stress tensors, obtaining better performance and more robust models than by using only atomic structure descriptors. Second, the descriptors can be used to directly calculate the partial atomic charges, obtaining high accuracy when compared to the leading methods in the field. We expect further applications of these physically-informed descriptors of the charge density for the prediction of other electron density-related properties such as the dipole moment (initial tests in SI) or polarizability.

Continued future improvements in the performance and robustness of the ML-DFT approach will allow for a wide range of applications: from electronic structure prediction to structure search and optimization, while integration with MD codes will allow for emulations of ab initio MD simulations. Large-scale

dynamical simulations of disordered systems such as liquids or glasses may be performed, which are challenging for traditional DFT. Generalization of this methodology to more delocalized electron densities, such as in metals, may require modifications to the type of basis functions used in the decomposition of the charge density and will be explored in the future (see test on liquid Li in the SI). We also envision further modifications to increase the accuracy and transferability by shifting our present fingerprinting from predefined equations to allow deep learning architectures to search for the best atomic descriptors.

METHODS

DFT details

All the reference data calculations were performed using DFT-MD simulations using the Vienna Ab Initio Simulation Package (VASP)^{42,43}. The exchange–correlation function was modeled using the Perdew–Burke–Ernzerhof approximation⁵⁴, and the ion–electron interaction was modeled using projector-

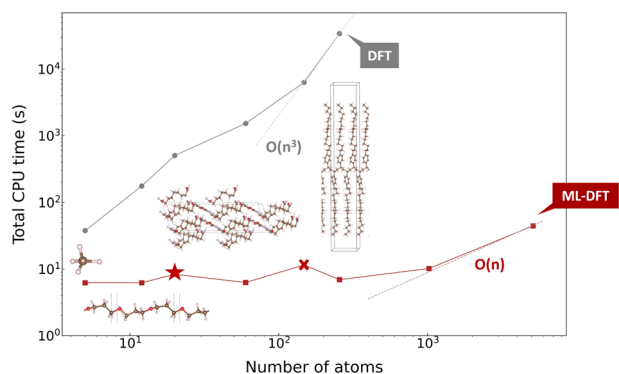


Fig. 6 Total CPU time of DFT versus ML-DFT for electronic structure predictions. DFT shows an initial high computational cost and a cubic dependence on the system size. On the other hand, ML-DFT is orders of magnitude faster than DFT and linearly dependent on the system size. Red squares: structures with carbon and hydrogen. Red star: structure with three elements. Red cross: structure with all four elements.

augmented wave (PAW) potentials⁵⁵. We employed a Monkhorst–Pack grid⁵⁶ with a density of 0.025 \AA^{-1} to sample the Brillouin zone. A plane wave basis set with a kinetic energy cutoff of 500 eV was used. The chosen kinetic energy cutoff and k-point sampling converged the total energy to less than 1 meV per atom. Tkatchenko and Scheffler vdW corrections were included⁵⁷. Gaussian smearing of 0.1 eV was used. The MD simulations were performed in the NVT ensemble, with a time step of 1 fs for the molecules and polymer chains at 300 K. For the polymer crystals, due to the high temperatures reached, we used a timestep of 0.5 fs. All structures were thermalized for 500-time steps at their initial temperature (300 or 100 K), and the snapshots were taken from the subsequent simulations spanning 1 ps for the molecules and polymer chains and 5 ps for the crystal polymers.

AGNI fingerprints

For a given atom i , three different types of AGNI fingerprints are defined (scalar, vector, and tensor)^{27,28}, expressed as the sum over the number of Gaussian functions (k) of width σ_k ,

$$S_{k,i} = c_k \sum_{j=1}^N \exp\left(\frac{-R_{ij}^2}{2\sigma_k^2}\right) f_c(R_{ij}) \quad (2)$$

$$V_{k,i}^a = c_k \sum_{j=1}^N \frac{r_{ij}^a}{R_{ij}} \exp\left(\frac{-R_{ij}^2}{2\sigma_k^2}\right) f_c(R_{ij}) \quad (3)$$

$$T_{k,i}^{\alpha\beta} = c_k \sum_{j=1}^N \frac{r_{ij}^\alpha r_{ij}^\beta}{R_{ij}^2} \exp\left(\frac{-R_{ij}^2}{2\sigma_k^2}\right) f_c(R_{ij}) \quad (4)$$

where c_k is the normalization constant defined as $\left(\frac{1}{\sqrt{2\pi}\sigma_k}\right)^3$, R_{ij} the distance between atom j and the center atom i , and $f_c(R_{ij})$ a cutoff function defined as $0.5 \left[\cos\left(\frac{\pi R_{ij}}{d_c}\right) + 1\right]$ for $R_{ij} \leq d_c$ and equal to 0 for $R_{ij} > d_c$. a and β represent the x , y , or z components of the radial vector between atoms i and j . In this work, we employed 18 different Gaussian widths on a logarithmic scale (base 10) from 0.5 to 6.0 Å, with a cutoff distance of $d_c = 5 \text{ \AA}$. While S_k is rotation invariant, V_k^a and $T_k^{\alpha\beta}$ are not, but can be combined into four rotation invariant expressions²⁷, which are employed as the fingerprints.

ML-DFT architecture

We used Keras⁵⁸ with the TensorFlow backend to implement the ML-DFT. The charge density and DOS models employ fully

connected layers and are trained using a mini-batch of 30, while the model for the potential energy, atomic force, and stress tensor uses a mini-batch of 100. All models use random sampling along with Adam optimizer with a learning rate of 0.0001 and momentum vectors $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The mean-squared error was employed as the objective function during all training. More details about the specific architectures of each model can be found in the SI.

DATA AVAILABILITY

All DFT data can be found at khazana.gatech.edu.

CODE AVAILABILITY

To promote its applicability within the community, we provide access to our ML-DFT emulator package with the presently trained models and tutorials (github.com/Ramprasad-Group). Additionally, a Google Colab notebook is included in the package to clone and utilize the package online for predictions.

Received: 18 December 2022; Accepted: 16 August 2023;

Published online: 29 August 2023

REFERENCES

- Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136(3B)**, B864 (1964).
- Kohn, W. & Sham, L. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–8 (1965).
- Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. B. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **5**, 909–913 (2006).
- Nørskov, J. K., Abild-Pedersen, F., Studt, F. & Bligaard, T. Density functional theory in surface chemistry and catalysis. *Proc. Natl Acad. Sci. USA* **108**, 937–943 (2011).
- Ceder, G. et al. Identification of cathode materials for lithium batteries guided by first-principles calculations. *Nature* **392**, 694–696 (1998).
- Schlapbach, L. & Züttel, A. Hydrogen-storage materials for mobile applications. *Nature* **414**, 353–358 (2001).
- Kang, K., Meng, Y. S., Bréger, J., Grey, C. P. & Ceder, G. Electrodes with High Power and High Capacity for Rechargeable Lithium Batteries. *Science* **311**, 977–980 (2006).
- Sharma, V. et al. Rational design of all organic polymer dielectrics. *Nat. Commun.* **5**, 4845 (2014).
- Christensen, N. E. & Novikov, D. L. Predicted superconductive properties of lithium under pressure. *Phys. Rev. Lett.* **86**, 1861–1864 (2001).
- Kolmogorov, A. et al. New superconducting and semiconducting Fe-B compounds predicted with an Ab initio evolutionary search. *Phys. Rev. Lett.* **105**, 217003 (2010).
- Wu, C. et al. Rational design of all-organic flexible high-temperature polymer dielectrics. *Matter* **5**, 2615–2623 (2022).
- Ma, X., Li, Z., Achenie, L. E. K. & Xin, H. Machine-learning-augmented chemisorption model for CO₂ electroreduction catalyst screening. *J. Phys. Chem. Lett.* **6**, 3528–33 (2015).
- Li, Z., Wang, S., Chin, W., Achenie, L. & Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **5**, 24131 (2017).
- O'Connor, N. J., Jonayat, A. S. M., Janik, M. J. & Senftle, T. P. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat. Catal.* **1**, 531539 (2018).
- García-Muelas, R. & López, N. Statistical learning goes beyond the D-band model providing the thermochemistry of adsorbates on transition metals. *Nat. Commun.* **10**, 4687 (2019).
- Toyao, T. et al. Machine learning for catalysis informatics: recent applications and prospects. *ACS Catal.* **10**, 2260–2297 (2020).
- Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**, 3762–3767 (2010).
- Davies, D. W. et al. Computer-aided design of metal chalcogenide semiconductors: from chemical composition to crystal structure. *Chem. Sci.* **9**, 1022–1030 (2018).
- Ye, W., Chen, C., Wang, Z., Chu, I.-H. & Ong, S. P. Deep neural networks for accurate predictions of crystal stability. *Nat. Commun.* **9**, 3800 (2018).

20. Yasina, A. S. & Musho, T. D. A machine learning approach for increased throughput of density functional theory substitutional alloy studies. *Comput. Mater. Sci.* **181**, 109726 (2020).
21. Zhang, Z., Li, M., Flores, K. & Mishra, R. Machine learning formation enthalpies of intermetallics. *J. Appl. Phys.* **128**, 105103 (2020).
22. Xiong, S. et al. A combined machine learning and density functional theory study of binary Ti-Nb and Ti-Zr alloys: Stability and Young's modulus. *Comput. Mater. Sci.* **184**, 109830 (2020).
23. Kauffman, K. et al. Discovery of high-entropy ceramics via machine learning. *npj Comput. Mater.* **6**, 42 (2020).
24. Kaufmann, K. & Vecchio, K. Searching for high entropy alloys: a machine learning approach. *Acta Mater.* **198**, 178–222 (2020).
25. Chen, L. et al. Polymer informatics: current status and critical next steps. *Mater. Sci. Eng. R. Rep.* **144**, 100595 (2021).
26. Tran, H. D. et al. Machine-learning predictions of polymer properties with polymer genome. *J. Appl. Phys.* **128**, 171104 (2020).
27. Chandrasekaran, A. et al. Solving the electronic structure problem with machine learning. *npj Comput. Mater.* **5**, 22 (2019).
28. G. del Rio, B., Kuenneth, C., Tran, H. D. & Ramprasad, R. An efficient deep learning scheme to predict the electronic structure of materials and molecules: the example of graphene-derived allotropes. *J. Phys. Chem. A* **124**, 9496–9502 (2020).
29. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
30. 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).
31. Kamal, D., Chandrasekaran, A., Batra, R. & Ramprasad, R. A charge density prediction model for hydrocarbons using deep neural networks. *Mach. Learn.: Sci. Technol.* **1**, 025003 (2020).
32. Ellis, J. et al. Accelerating finite-temperature Kohn-Sham density functional theory with deep neural networks. *Phys. Rev. B* **104**, 035120 (2021).
33. Grisafi, A. et al. Transferable machine-learning model of the electron density. *ACS Cent. Sci.* **5**, 57–64 (2019).
34. Fabrizio, A., Grisafi, A., Meyer, B., Ceriotti, M. & Corminboeuf, C. Electron density learning of non-covalent systems. *Chem. Sci.* **10**, 9424–9432 (2019).
35. Cuevas-Zuñivira, B. & Pacios, L. F. Analytical model of electron density and its machine learning inference. *J. Chem. Inf. Model.* **60**, 3831–3842 (2020).
36. Cuevas-Zuñivira, B. & Pacios, L. F. Machine learning of analytical electron density in large molecules through message-passing. *J. Chem. Inf. Model.* **61**, 2658–2666 (2021).
37. Jørgensen, P. & Bhowmik, A. Equivariant graph neural networks for fast electron density estimation of molecules, liquids, and solids. *npj Comput. Mater.* **8**, 183 (2022).
38. Kamal, D., Chandrasekaran, A., Batra, R. & Ramprasad, R. A charge density prediction model for hydrocarbons using deep neural networks. *Mach. Learn.* **166**, 025003 (2020).
39. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B* **92**, 1–5 (2015).
40. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C* **121**, 511–522 (2017).
41. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
42. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–86 (1996).
43. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *J. Comput. Mater. Sci.* **6**, 15–50 (1996).
44. Botu, V. & Ramprasad, R. Learning scheme to predict atomic forces and accelerate materials simulations. *Phys. Rev. B* **92**, 094306 (2015).
45. Botu, V., Batra, R., Chapman, J. & Ramprasad, R. Machine learning force fields: construction, validation, and outlook. *J. Phys. Chem. C* **121**, 511–522 (2017).
46. Huan, T. et al. A universal strategy for the creation of machine learning-based atomistic force fields. *npj Comput. Mater.* **3**, 37 (2017).
47. Rowe, P., Deringer, V., Gasparotto, P., Csányi, G. & Michaelides, A. An accurate and transferable machine learning potential for carbon. *J. Chem. Phys.* **153**, 034702 (2020).
48. Shaaidu, Y. et al. A systematic approach to generating accurate neural network potentials: the case of carbon. *npj Comput. Mater.* **7**, 52 (2021).
49. Wang, J. et al. A deep learning interatomic potential developed for atomistic simulation of carbon materials. *Carbon* **186**, 1–8 (2022).
50. Yoo, P. et al. Neural network reactive force field for c, h, n, and o systems. *npj Comput. Mater.* **7**, 9 (2021).
51. Umeno, Y. & Kubo, A. Prediction of electronic structure in atomistic model using artificial neural network. *Comput. Mater. Sci.* **168**, 164–171 (2019).
52. Ben Mahmoud, C., Anelli, A., Csányi, G. & Ceriotti, M. Learning the electronic density of states in condensed matter. *Phys. Rev. B* **102**, 235130 (2020).
53. Kong, S. et al. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* **13**, 949 (2022).
54. Perdew, J., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–8 (1996).
55. Kresse, G. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–75 (1999).
56. Monkhorst, H. & Pack, J. Special points for Brillouin-Zone integrations. *Phys. Rev. B* **13**, 5188–92 (1976).
57. Tkatchenko, A. & Scheffler, M. Accurate molecular Van Der Waals interactions from ground-state electron density and free-atom reference data. *Phys. Rev. Lett.* **102**, 073005 (2009).
58. Chollet, F. et al. *Keras*. <https://keras.io> (2015).

ACKNOWLEDGEMENTS

This work is partially funded by the National Science Foundation under Award Numbers 1900017 and 1941029 and partially by the Office of Naval Research under Award Number N00014-18-1-2113. We thank Christopher Kuenneth and Huan Doan Tran for their useful discussions and Lihua Chen for proofreading the paper.

AUTHOR CONTRIBUTIONS

R.R. and B.G.R. conceptualized the ML-DFT framework. B.G.R. developed the ML-DFT framework, designed and trained the networks, and evaluated the results. B.G.R. and B.P. created the database. R.R. and B.G.R. 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/s41524-023-01115-3>.

Correspondence and requests for materials should be addressed to Beatriz G. del Rio or Rampi Ramprasad.

Reprints and permission 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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023