Transferable empirical pseudopotenials from machine learning

Rokyeon Kim[*] and Young-Woo Son [†]
Korea Institute for Advanced Study, Seoul 02455, Korea

Machine learning is used to generate empirical pseudopotentials that characterize the local screened interactions in the Kohn-Sham Hamiltonian. Our approach incorporates momentum-range-separated rotation-covariant descriptors to capture crystal symmetries as well as crucial directional information of bonds, thus realizing accurate descriptions of anisotropic solids. Trained empirical potentials are shown to be versatile and transferable such that the calculated energy bands and wave functions without cumbersome self-consistency reproduce conventional *ab initio* results even for semiconductors with defects, thus fostering faster and faithful data-driven material researches.

I. INTRODUCTION

First-principles calculations based on the density functional theory (DFT) [1,2] have become standard tools for studying the physical properties of materials [3–6]. Recently, applications of machine learning (ML) techniques to various computational methodologies based on the DFT has brought forth a new set of tools for investigating materials at the quantum scale [7–11]. Such novel approaches have given rise to a rapidly growing field, offering new insights and significant potential for prediction and analysis of materials properties [12–27].

One of popular applications in those developments has been the accelerated computations of physical quantities such as total energies and atomic forces [12–14,16–21,27]. By circumventing a part of various computationally demanding processes involved in DFT calculations, the ML techniques provide efficient ways for improving various simulation methods such as molecular dynamics. On the other hand, the integration of ML has also brought improvements to the exchange-correlation functionals within DFT [15,22–26,28], which are central to describing the many-electron effects in quantum systems.

Despite these strides, applications of ML to obtain precise quantum mechanical electronic structures for entire phase space of interests remain relatively unexplored [29–35]. Previous studies have utilized ML to study the electronic structures of one-dimensional [29], slab [30], and molecular [31–33] systems. We also note that the neural network was used to generate better transferable local pseudopotentials [36]. Only recently, a general ML framework to construct DFT Hamiltonian in the tight-binding approach has been developed [34,35]. Accurate quantum properties of solids such as energy bands and wave functions are central to design and discovery of new materials with desired properties.

However, the traditional DFT methods require a large amount of resources, partly because of the unavoidable self-consistent condition, posing significant challenges in data-intensive materials researches. Hence, there is a pressing need for a faster method that utilizes ML to accelerate electronic structure calculations without sacrificing the accuracy of first-principles methods.

Before the advent of *ab initio* methods based on the DFT, empirical pseudopotential method (EPM) [37] has been widely used as a fast and efficient method to calculate the electronic structure of materials because of its formal simplicity as well as less demanding computational resources. Despite its extensive use for various solids [37–43], the EPM has limitations such as inaccurate wave functions [44] and transferability issues of the obtained pseudopotential [41,42,45–47]. To improve EPM, Wang and Zunger proposed the local-density-derived EPM, which generates pseudopotentials by inverting the Kohn-Sham (KS) potential in DFT calculations [48]. However, the potentials obtained from this approach still suffer from a lack transferability to a wide range of materials, and the use of the spherical approximation results in errors in band structures, particularly for anisotropic crystal structures [49].

In this paper, we propose a neural network model to generate universal empirical pseudopotentials (EPs) encompassing the local screened interactions in solids. We demonstrate that our EPs are versatile and transferable, reproducing conventional first-principles results for energy bands and wave functions without self-consistency. To achieve this, we introduce a new set of rotation-covariant descriptors that capture the atomic and structural characteristics of target solids, enabling the transferability of learned EPs even to defective solids. Moreover, our model seamlessly integrates into existing computational packages and can be extended to use advanced *ab initio* methods to calculate optical and transport properties. We also demonstrate an extension to incorporate the nonlocal correlation effects without increasing computational complexity and resources. Overall, our method accurately reproduces converged KS Hamiltonians without

---
[*]rrykim@gmail.com
[†]hand@kias.re.kr

self-consistency, thus providing a reliable platform for data-intensive materials researches.

II. MACHINE LEARNING FRAMEWORK

In DFT using the local approximation for exchange-correlation functionals, the electronic structure of a solid is obtained by solving the KS equation [2],

$$\left[-\tfrac{1}{2}\nabla^2 + V_{\text{PS,nloc}} + V(\mathbf{r})\right]\psi_i = \epsilon_i \psi_i, \tag{1}$$

where ψ_i is the ith KS wave function with an energy eigenvalue of ϵ_i, $V_{\text{PS,nloc}}$ the nonlocal part of the pseudopotential, and $V(\mathbf{r})$ the local potential for all interactions experienced by a single electron. Specifically, $V(\mathbf{r})$ is the sum of three terms: the local pseudopotential, Hartree potential, and exchange-correlation potential. In the DFT approach, $V(\mathbf{r})$ for a given solid should be determined self-consistently. Instead of the conventional way of computing converged potentials, we are to employ ML techniques to generate atomic EPs of $v(\mathbf{r})$ for individual atoms, such that their summation equals the local potential of $V(\mathbf{r})$ in a crystal. (we use v and V to denote atomic and crystalline potentials, respectively). If the learned EPs can closely approximate the KS potentials, the KS Hamiltonian can be constructed only once, eliminating the self-consistency condition.

Specifically, let us consider a crystal structure denoted by $\mathcal{C}$, whose self-consistent local potential is $V_{\mathcal{C}}(\mathbf{r})$. We aim to find EPs of $v_{\mathcal{C}}^{\alpha}(\mathbf{r})$ of the αth atom in the crystal structure $\mathcal{C}$ that satisfies

$$V_{\mathcal{C}}(\mathbf{r}) = \sum_{\mathbf{R}} \sum_{\alpha} v_{\mathcal{C}}^{\alpha}(\mathbf{r} - \mathbf{R} - \boldsymbol{\tau}_{\alpha}), \tag{2}$$

where $\mathbf{R}$ is the Bravais lattice vector of the crystal, and $\boldsymbol{\tau}_{\alpha}$ is the position of the αth atom in the unit cell. In the momentum space, Eq. (2) can be written as

$$V_{\mathcal{C}}(\mathbf{G}_{\mathcal{C}}) = \sum_{\alpha} S_{\mathcal{C}}^{\alpha}(\mathbf{G}_{\mathcal{C}}) v_{\mathcal{C}}^{\alpha}(\mathbf{G}_{\mathcal{C}}), \tag{3}$$

where $S_{\mathcal{C}}^{\alpha}(\mathbf{G}_{\mathcal{C}}) = \frac{1}{\Omega_{\mathcal{C}}} e^{-i\mathbf{G}_{\mathcal{C}} \cdot \boldsymbol{\tau}_{\alpha}}$ is the structure factor of the αth atom, $\Omega_{\mathcal{C}}$ the unit cell volume of structure $\mathcal{C}$, and $\mathbf{G}_{\mathcal{C}}$ the reciprocal lattice vector of structure $\mathcal{C}$.

For varying crystal structures, the EP of $v_{\mathcal{C}}^{\alpha}(\mathbf{G}_{\mathcal{C}})$ should depend on the local environment around the αth atom because the local symmetry, atomic coordination, and bond characteristics for each atom alter substantially. To account for these factors, we introduce a descriptor of $\mathbf{d}_{\mathcal{C}}^{\alpha}$, whose specific form will be discussed later. By incorporating this descriptor into the EP, we express the potential as functions of $\mathbf{d}_{\mathcal{C}}^{\alpha}$ and $\mathbf{G}_{\mathcal{C}}$, respectively. So, resulting EPs can be written as $v[\mathbf{d}_{\mathcal{C}}^{\alpha}](\mathbf{G}_{\mathcal{C}})$. This is the universal EP, which must be transferable across different systems due to its ability to capture the complex local environments through the descriptor of $\mathbf{d}_{\mathcal{C}}^{\alpha}$.

With the introduction of the descriptor, we can apply ML to generate the EPs. Since the KS potentials from DFT calculations are to be learned, we employ a ML model, as depicted in Fig. 1, similar to the high-dimensional neural network introduced by Behler and Parrinello [12]. Our model employs a weight-sharing neural network (denoted as EP NN in Fig. 1) to represent the EP of $v[\mathbf{d}](\mathbf{G})$ with inputs of $\mathbf{d}$ and $\mathbf{G}$ (dropping the indices α and $\mathcal{C}$ for simplicity). We optimize the ML model
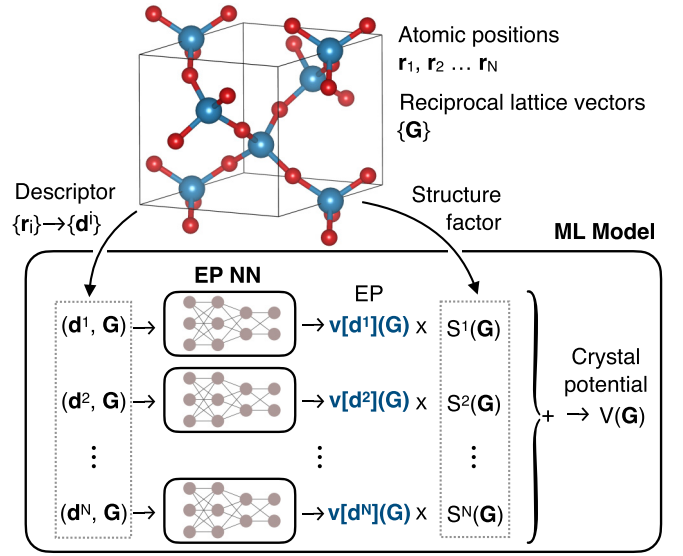

FIG. 1. The structure of a ML model for EPs. From a crystal structure having N atoms, we calculate N descriptors of $\mathbf{d}^i$ representing the local environment at each atom. These descriptors, along with momentum $\mathbf{G}$, are fed into the EP neural network (EP NN) as input. The resulting $v[\mathbf{d}^i](\mathbf{G})$ are multiplied by the structure factors of $S^i(\mathbf{G})$, then summed over all atoms to yield the crystal potential of $V(\mathbf{G})$.

by minimizing the loss function for various crystal structures $\mathcal{C}$. The loss function compares the DFT crystalline potential of $V_{\mathcal{C},\text{DFT}}(\mathbf{G}_{\mathcal{C}})$ with the corresponding ML-predicted $V_{\mathcal{C},\text{ML}}(\mathbf{G}_{\mathcal{C}})$ for a trained structure $\mathcal{C}$:

$$\text{Loss} = \frac{1}{N_D} \sum_{\{\mathcal{C}\}} \sum_{\mathbf{G}_{\mathcal{C}}} |V_{\mathcal{C},\text{DFT}}(\mathbf{G}_{\mathcal{C}}) - V_{\mathcal{C},\text{ML}}(\mathbf{G}_{\mathcal{C}})|^2, \tag{4}$$

where N_D is the number of the training data and $\{\mathcal{C}\}$ indicates the summation for all training structures. Here, $V_{\mathcal{C},\text{ML}}(\mathbf{G}_{\mathcal{C}})$ is calculated by

$$V_{\mathcal{C},\text{ML}}(\mathbf{G}_{\mathcal{C}}) = \sum_{\alpha} S_{\mathcal{C}}^{\alpha}(\mathbf{G}_{\mathcal{C}}) v[\mathbf{d}_{\mathcal{C}}^{\alpha}](\mathbf{G}_{\mathcal{C}}), \tag{5}$$

where $S_{\mathcal{C}}^{\alpha}(\mathbf{G}_{\mathcal{C}})$ can be readily calculated using the reciprocal lattice vectors and atomic positions. After the ML procedure, we have the EP of $v[\mathbf{d}](\mathbf{G})$ for general $\mathbf{d}$ and $\mathbf{G}$, which allow us to predict band structures for various new crystal structures.

To construct $V_{\mathcal{C},\text{ML}}$ reflecting the crystal symmetry of a given structure $\mathcal{C}$, we introduced a momentum-dependent rotation-covariant descriptor that is modified from the atom-density representation [50]. In the representation [50], the sum of Gaussian functions with a variance of σ^2 is assigned as a density of ρ^Z for each atomic species Z inside a cutoff radius of r_{cut} such as $\rho^Z(\mathbf{r}) = \sum_{\alpha \in Z} \exp[-\frac{(\mathbf{r}-\mathbf{r}_{\alpha})^2}{2\sigma^2}]$. Here the origin sets to the point where the local environment is evaluated. With this density, we use two different types of descriptors as follows.

The first type is the density coefficient (DC) descriptor, denoted as $\mathbf{c}$, of which an element of c_{lmn}^Z is defined as $c_{lmn}^Z = \int d\mathbf{r} \, g_n^*(r) Y_{lm}^*(\hat{\mathbf{r}}) \rho^Z(\mathbf{r})$, where $g_n(r)$ is a set of radial basis functions, and $Y_{lm}(\hat{\mathbf{r}})$ spherical harmonics. For practical applications, we limit the expansion using suitable $n_{\max}$

TABLE I. Unit cells of SiO$_2$ used for the generation of machine learning data.

Materials Project ID	Number of atoms	Space group	Materials Project ID	Number of atoms	Space group
mp-6922	9	$P6_222$ (No. 180)	mp-10851	9	$P6_422$ (No. 181)
mp-6930	9	$P3_221$ (No. 154)	mp-10948	12	$Pbcn$ (No. 60)
mp-6945	12	$P4_12_12$ (No. 92)	mp-546794	12	$I\bar{4}2d$ (No. 122)
mp-6947	6	$P4_2/mnm$ (No. 136)	mp-555235	24	Cc (No. 9)
mp-7000	9	$P3_121$ (No. 152)	mp-555891	12	$P2_1$ (No. 4)
mp-7087	12	$P6_3/mmc$ (No. 194)	mp-556961	24	$P2_12_12_1$ (No. 19)
mp-7648	24	$C222_1$ (No. 20)	mp-559091	12	$P6_322$ (No. 182)
mp-7905	12	$Ibam$ (No. 72)	mp-640917	9	$P3_221$ (No. 154)
mp-8059	24	$P2_13$ (No. 198)	mp-669426	24	$P2_1/c$ (No. 14)
mp-8352	24	$Fd\bar{3}m$ (No. 227)	mp-972808	12	$P2_1$ (No. 4)
mp-9258	12	$Pa\bar{3}$ (No. 205)	mp-1071820	12	$Ibam$ (No. 72)

and $l_{\max}$ values. The second type is the smooth overlap of atomic positions (SOAP) descriptor [50], denoted as **p**, whose elements are defined as $p_{nn'l}^{ZZ'} = \pi\sqrt{\frac{8}{2l+1}}\sum_m (c_{nlm}^Z)^* c_{n'lm}^{Z'}$. With these definitions, the descriptor **d** can be either **c** or **p** depending on the purpose of ML as discussed below. We note that only the SOAP descriptor seems to be widely used because its rotation-invariant nature may be suitable to represent scalar quantities of solids such as energy. However, it may not be adequate for describing the potential because it cannot reflect directional information. We also note that, while the descriptor **p** was originally designed for the kernel methods [50], it can be used as an input for a neural network [51].

We use the DC descriptor for the EPs because, under a general rotation operation of $\mathcal{R}$, $v[\mathbf{d}](\mathbf{G})$ should satisfy $v[\mathbf{d}](\mathbf{G}) = v[\mathcal{R}\mathbf{d}](\mathcal{R}\mathbf{G})$, i.e., both **d** and **G** must be rotated simultaneously to yield the same v. The DC descriptor $\mathbf{c}_\mathcal{C}^\alpha$ is rotation-covariant such that it is rotated by the well-defined rules using the Wigner D matrices [50]. Therefore we can use $\mathbf{c}_\mathcal{C}^\alpha$ as a main descriptor for learning EPs that can contain all local information for a crystal structure $\mathcal{C}$. On the other hand, with the SOAP descriptor $\mathbf{p}_\mathcal{C}^\alpha$, the EP is rotation-invariant, only depending on a magnitude $(G_\mathcal{C})$ of $\mathbf{G}_\mathcal{C}$ such that the resulting EP can be written as $v[\mathbf{p}_\mathcal{C}^\alpha](G_\mathcal{C})$. In this case, instead

of Eq. (5), we use the ML potential written as

$$\bar{V}_{\mathcal{C},\mathrm{ML}}(G_\mathcal{C}) = \sum_\alpha \bar{S}_\mathcal{C}^\alpha(G_\mathcal{C}) v[\mathbf{p}_\mathcal{C}^\alpha](G_\mathcal{C}), \qquad (6)$$

where $\bar{S}_\mathcal{C}^\alpha(G_\mathcal{C})$ is the spherically averaged quantity of $S_\mathcal{C}^\alpha(\mathbf{G}_\mathcal{C})$. Then, in the loss function, we compare $\bar{V}_{\mathcal{C},\mathrm{ML}}(G_\mathcal{C})$ with the spherical averaged $\bar{V}_{\mathcal{C},\mathrm{DFT}}(G_\mathcal{C})$. To ensure both efficiency and accuracy of our ML model, we adopt a hybrid approach using two distinct ML models that treats the directional dependency of the EP differently depending on the momentum values. For large G, the detailed directional information is found to be less relevant so that the spherical symmetric form suffices to construct EPs (see Fig. 2). Therefore the descriptor **p** is adopted to represent the potential for large G. For small G,
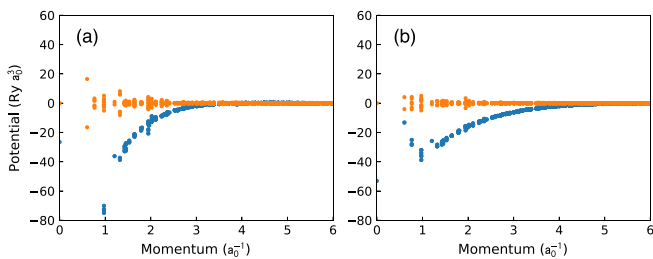

FIG. 2. Shape of EPs for (a) Si and (b) O in α-quartz SiO$_2$. Blue (orange) dots represent real (imaginary) parts of the potentials. Here, we directly inverted Eq. (3) without a ML procedure to obtain the potentials [52]. The vertical distribution of data points indicates multiple momentum vectors for each momentum value, showing the directional dependence of the potentials. For large momentum values, real values align along a line and imaginary values are suppressed, which can be also observed in other SiO$_2$ structures.
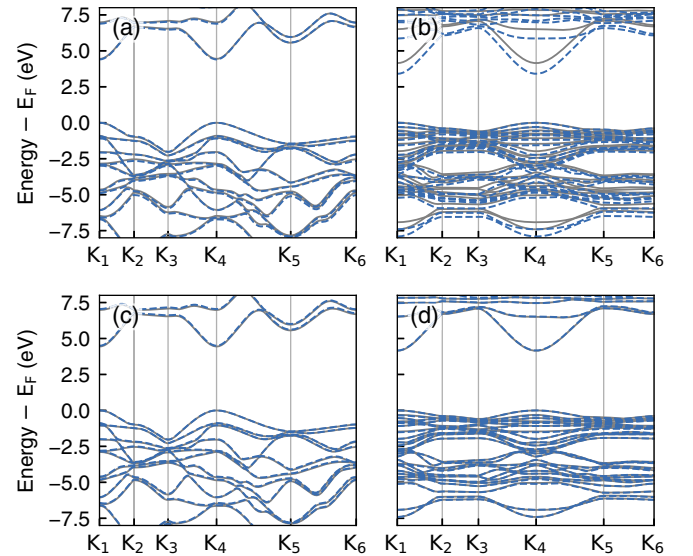

FIG. 3. Band structures of two representative SiO$_2$ structures, C1 and C2 (see text). Gray solid and blue dashed lines represent the DFT and EPM results, respectively. $\mathrm{K}_i(i = 1,\ldots,6)$ are reduced coordinates in reciprocal space, representing (0, 0, 0), (0.5, 0, 0), (0.5, 0.5, 0), (0, 0, 0), (0.5, 0.5, 0.5), (0.5, 0, 0) in their corresponding crystal structures, respectively. E_F denotes the Fermi energy. The full-spherical approximation results for (a) C1 and (b) C2. The hybrid model results for (c) C1 and (d) C2.
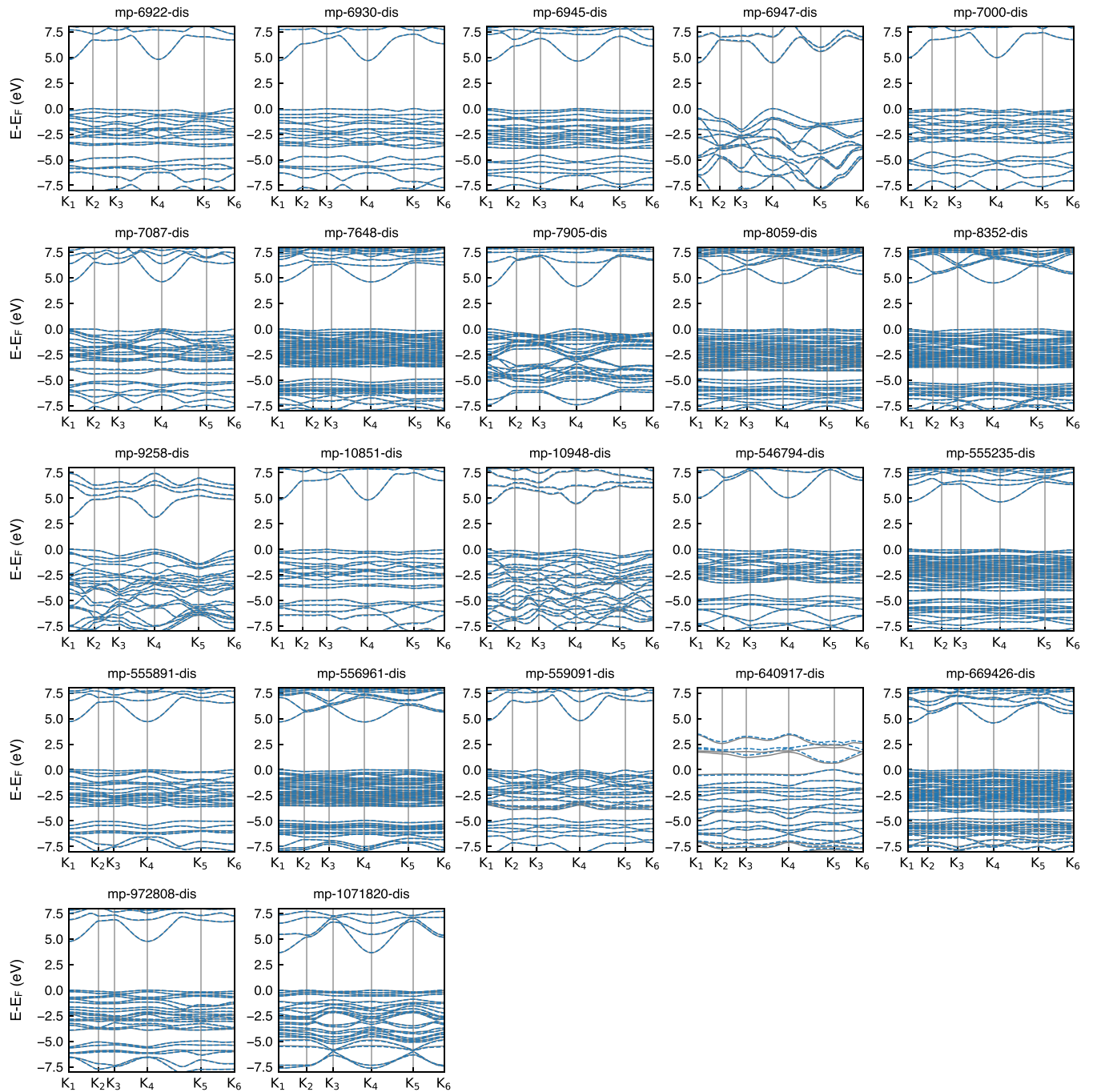
FIG. 4. Band structures for SiO_2 with Material Project IDs ranging from mp-6922 to mp-1071820 (see Table I), where the length of the lattice vectors and internal coordinates for each structure are randomly disturbed within 5% (denoted as "-dis" in labels). K_i $(i = 1, \ldots, 6)$ are reduced coordinates in reciprocal space, representing $(0, 0, 0)$, $(0.5, 0, 0)$, $(0.5, 0.5, 0)$, $(0, 0, 0)$, $(0.5, 0.5, 0.5)$, $(0.5, 0, 0)$ in their corresponding crystal structures, respectively. Solid gray and dashed blue lines represent the DFT and the EPM results, respectively.

however, the directional information becomes crucial so that the descriptor **c** is used. Formally, in the hybrid ML process, the form $v[\mathbf{p}](G)$ is used for $G > G_{cut}$, while $v[\mathbf{c}](\mathbf{G})$ for $G \leqslant G_{cut}$, where $G_{cut} = 3\,a_0^{-1}$ (a_0 is the Bohr radius).

III. APPLICATIONS TO SiO₂

Having established the universal EPM, we demonstrate a few examples showing the versatility of our ML model. We

choose silicon dioxide (SiO_2) or silica as an example owing to its diverse crystal structures and wide-ranging applications [53]. To generate data for the ML model, we prepared 22 stable polymorphs of SiO_2 from the Materials Project [54] as listed in Table I. For each sample, we constructed $p \times q \times r$ supercells (p, q, and r are integers) limiting the maximum number of atoms in the supercell to 24. We then randomly perturbed the length of the supercell lattice vectors and the internal coordinates of the atoms within 5% to increase the

diversity of the generated data. For 10% of the learning samples, we create arbitrary defective structures by randomly removing atoms in the unit cell. In total, we prepared 10,224 inputs for DFT calculations. For each input $\mathcal{C}$, we calculated $V_{\mathcal{C},\mathrm{DFT}}$, $S_{\mathcal{C}}^{\alpha}$, $\mathbf{c}_{\mathcal{C}}^{\alpha}$, and $\mathbf{p}_{\mathcal{C}}^{\alpha}$, which were used to train the ML model. We implemented our universal EPMs in the QUANTUM ESPRESSO package [55], which we also used for performing reference DFT calculations. Norm-conserving LDA pseudopotentials from the PSEUDODOJO [56] were utilized. The k-point mesh was set with approximately 1,000 k points per reciprocal atom and the cutoff energy for the plane-wave basis set was 84 Ry.

We used an in-house modified version of the DSCRIBE package [57] to calculate descriptors. Gaussian functions with $\sigma = 1$ Å were employed for the density. For the basis functions, we used Gaussian type orbitals and real spherical harmonics as implemented in the package. The DC descriptor $\mathbf{c}$ was calculated with $r_{\mathrm{cut}} = 10$ Å, $n_{\mathrm{max}} = 7$, $l_{\mathrm{max}} = 7$, resulting in a total of 896 features. For the SOAP descriptor $\mathbf{p}$, we used $r_{\mathrm{cut}} = 10$ Å, $n_{\mathrm{max}} = 6$, $l_{\mathrm{max}} = 6$, and applied a weighting function of the form $c(1 + 2(\frac{r}{r_0})^3 - 3(\frac{r}{r_0})^2)^m$ with $c = 1$, $m = 2$, $r_0 = 10$ Å [58], leading to a total of 546 features.

For the EP NN (see Fig. 1), we employed fully connected neural networks with the following structures. For the input $(\mathbf{c}, \mathbf{G})$ comprising the DC descriptor $\mathbf{c}$ and the vector $\mathbf{G}$, we employed three hidden layers, each consisting of 1024 neurons with rectified linear unit (ReLU) activation function. The output layer was composed of two neurons, representing the real and imaginary parts of the potential $v[\mathbf{c}](\mathbf{G})$, respectively. Similarly, for the input $(\mathbf{p}, G)$ comprising the SOAP descriptor $\mathbf{p}$ and the scalar G, we utilized three hidden layers, each consisting of 512 neurons with ReLU activation function. The output layer consisted of a single neuron representing the potential $v[\mathbf{p}](G)$. TENSORFLOW [59] with the Keras API [60] was utilized to implement the entire ML process.

Figure 3 shows the band structures of two SiO_2 structures denoted as C1 and C2, which were created by randomly perturbing the unit cell vectors and internal coordinates of stishovite (space group $P4_2/mnm$) and fibrous (space group $Ibam$) polymorphs, respectively. We compare the band structures obtained using the full-spherical approximation and the hybrid model. In the full-spherical approximation, we assumes the form $v[\mathbf{p}](G)$ for all momentum values as opposed to the hybrid model. Figures 3(a) and 3(b) show the band structures using the full-spherical approximation for C1 and C2, respectively. For C1, the full-spherical approximation yields satisfactory agreement with its DFT band dispersions, but for C2, we observe discrepancies in the band structures because of its anisotropic crystal structure. Figures 3(c) and 3(d) show the energy bands of C1 and C2, respectively, using the hybrid model. The hybrid model significantly improves the accuracy of the EPM predictions of bands, being identical to those from DFT. These results emphasize the importance of incorporating directional information in the EPs to accurately describe the band structure, especially for anisotropic materials.

For the crystal structures listed in Table I, the EPM band structures calculated from the hybrid model are displayed in Fig. 4. Here, we adjusted each unit cell's lattice vectors and internal coordinates randomly perturbed by up to 5% to test
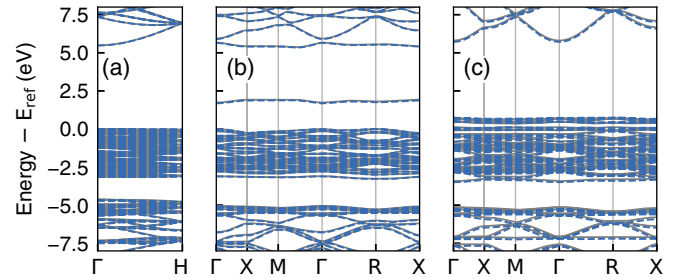

FIG. 5. (a) The band structure of a crystal with 72 atoms in a unit cell (see details in text). The band structures of defective cells for (b) oxygen vacancy and (c) silicon vacancy. Solid gray and dashed blue lines represent the DFT and the EPM results, respectively. In (a), $E_{\mathrm{ref}} = E_F$. In (b), we set E_{ref} at the band energy of the $(N-2)$th electrons, while in (c) $(N+4)$th electrons, where N is the number of electrons in each crystal structure.

the transferability and predictive capability of our EPM. In comparing the band structures, we observe a good agreement between the EPM and DFT results. An exception is noted for mp-640917 in the conduction bands, attributed to its highly anisotropic nature.

In Fig. 5(a), we present the band structure of a crystal having 72 atoms in its unit cell (Materials Project ID of mp-6930 with a space group of $P3_221$), which exhibits excellent agreement between the EPM and DFT results. It is worth noting that our hybrid ML model has been trained on a data set containing no more than 24 atoms in unit cell, and yet it accurately predicts the energy bands of a crystal with the significantly larger cell. This finding highlights the potential of our method in reliably predicting the electronic properties of materials of larger cell size, even with trained EPs from smaller cell sizes.

As shown in all bands from EPM in Figs. 3–5, we confirm that our hybrid EPM correctly predicts the band degeneracies at all high-symmetry points, indicating that the EP has learned the required crystalline symmetry for the potentials. Although our model does not include the symmetry components unlike equivariant neural network methods [35,61–64], we were able to learn the correct symmetry by designing a suitable rotation-covariant descriptor, generating data, and developing a ML model.

Figures 5(b) and 5(c) present the energy bands of SiO_2 with oxygen and silicon vacancies, respectively. We confirm that the defective structures for our test run do not belong to the learning set. Our EPM accurately predicts the band structures and defect energy levels if compared with the DFT results. Our study highlights the capability of EPM to accurately predict the defect band structure with computational efficiency compared to DFT.

As a final example, we use both band energies and wave functions to compute the dielectric function. Figures 6(a) and 6(b) display the imaginary part of the noninteracting dielectric function of $\epsilon_{xx}^{(2)}(\omega)$ for a pristine SiO_2 and one with an oxygen vacancy, respectively, demonstrating the excellent agreement between matrix elements obtained from the DFT and EPM results.
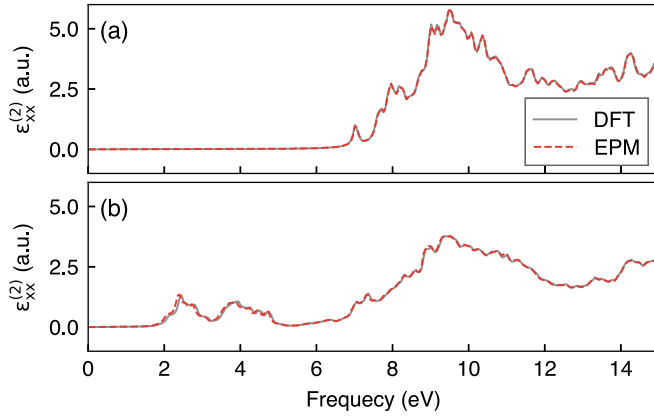
FIG. 6. Frequency dependent noninteracting dielectric functions calculated for (a) a pristine SiO_2 structure (b) and the SiO_2 structure with an oxygen vacancy.
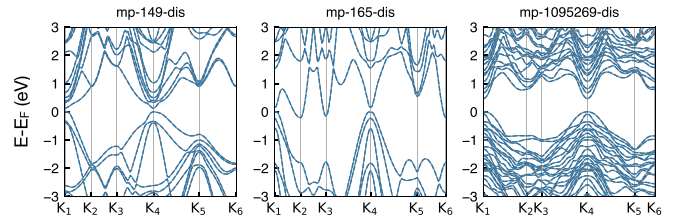

FIG. 7. Band structures for Si with Material Project IDs mp-149, mp-165, and mp-1095269 (see Table II), where the length of the lattice vectors and internal coordinates for each structure are randomly disturbed within 5% (denoted as "-dis" in labels). K_i ($i = 1, \ldots, 6$) are reduced coordinates in reciprocal space, representing (0, 0, 0), (0.5, 0, 0), (0.5, 0.5, 0), (0, 0, 0), (0.5, 0.5, 0.5), (0.5, 0, 0) in their corresponding crystal structures, respectively. Solid gray and dashed blue lines represent the DFT and the EPM results, respectively.

IV. APPLICATIONS TO SILICON

In this section, we detail calculations conducted on a few polymorphs of silicon. While we've crafted separate datasets and an ML model for Si, it's worth noting that they could be merged with those of SiO_2 to create a unified ML model applicable to both SiO_2 and Si. To generate data for the ML model, we prepared three stable polymorphs of Si as listed in Table II. The unit cells were retrieved from the Materials Project [54]. For each sample, we constructed $p \times q \times r$ supercells (p, q, and r are integers) limiting the maximum number of atoms in the supercell to 24. We then randomly perturbed the length of the supercell lattice vectors and the internal coordinates of the atoms within 5% to increase the diversity of the generated data. In total, we prepared 2496 inputs for DFT calculations.

We employed the same parameters as those used in the SiO_2 to compute descriptors. As a result of reducing the number of atomic species, **c** contains 448 features, and **p** contains 147 features. For the EP NN, we utilized three hidden layers, each containing 512 neurons, for the input of (**c**, **G**). Similarly, for the input of (**p**, G), we employed three hidden layers, each comprising 128 neurons.

The band structures corresponding to the unit cells listed in Table II are illustrated in Fig. 7, where we used the hybrid model for the EPM. To assess our EPM, we introduced random variations to the lattice vectors and internal coordinates, with perturbations up to 5%. For all three distinct structures, there is an excellent agreement between the DFT and EPM results, showing the reliability and accuracy of our approach.

V. DISCUSSION AND SUMMARY

Before concluding, we provide a couple of remarks on our universal EPM. Since the present ML model constructs a converged KS Hamiltonian in momentum space, a high energy cutoff is not necessary to converge density matrix. The only limiting factor for the cutoff comes from the kinetic energy. We demonstrate such an efficiency by reducing the cutoff for energy bands calculations of silica polymorphs. As shown in Fig. 8, in case of 40% reduction of the cutoff for conventional DFT calculations, the almost unaltered bands can be obtained using only about 4% of computational resources used for DFT. Table III compares the time needed for band structure calculations using DFT and EPM for three different systems. For the unit cell of α-quartz containing nine atoms, the self-consistent DFT calculation using a $5 \times 5 \times 4\,k$ mesh for electron density took 107 seconds, and the subsequent band structure calculation for 300 k points required 226 seconds. In contrast, at the same energy cutoff of 84 Ry, the EPM computed the band structure in 47 seconds. Further reducing the energy cutoff to 50 Ry shortened this time to
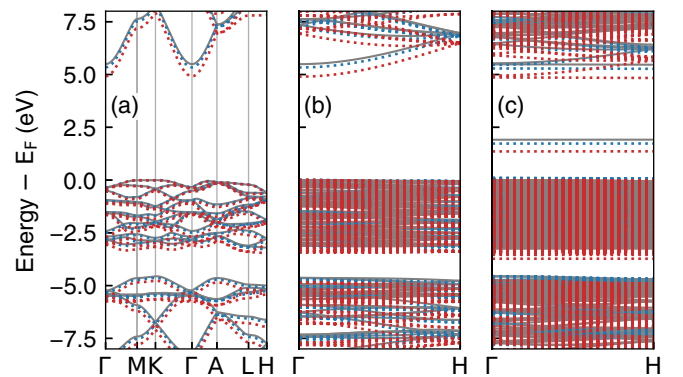
TABLE II. Unit cells of Si used for the generation of machine learning data.

Materials Project ID	Number of atoms	Space group
mp-149	8	$Fd\bar{3}m$ (No. 227)
mp-165	4	$P6_3/mmc$ (No. 194)
mp-1095269	24	$Cmcm$ (No. 63)

FIG. 8. Band structures obtained with different energy cutoff values for (a) α-quartz SiO_2 (9 atoms), (b) $2 \times 2 \times 2$ supercell (72 atoms), and (c) $3 \times 3 \times 3$ supercell with an oxygen vacancy (242 atoms). For each figure, the solid gray lines represent the DFT results obtained using an energy cutoff of 84 Ry, while the dotted blue (red) lines correspond to the EPM results obtained using an energy cutoff of 50 (40) Ry.

TABLE III. Comparison of computation times for band structures.

Number of atoms	Number of CPU[a] nodes	k-mesh (density)	Number of k-points (band)	time (second)			
				DFT density	DFT band	EPM band (84 Ry)	EPM band (50 Ry)
9[b]	1	$5 \times 5 \times 4$	300	107	226	47	26
72[c]	2	$3 \times 3 \times 2$	30	2,294	4,080	807	276
242[d]	2	$2 \times 2 \times 2$	20	31,500	85,740	14,340	6,900

[a]Intel(R) Xeon(R) CPU E5-2690 v2 @ 2.800GHz (10 cores).

[b]α-quartz SiO_2.

[c]$2 \times 2 \times 2$ supercell.

[d]$3 \times 3 \times 3$ supercell with an oxygen vacancy.

26 seconds. When expanding to a larger 72-atom supercell using, the self-consistent DFT calculation for density using a $3 \times 3 \times 2$ k-mesh took 2294 seconds. The band structure calculation for 30 k points were completed in 4,080 seconds. However, using the EPM and energy cutoffs of 84 Ry and 50 Ry, the band structure computation times were reduced to 807 and 276 seconds, respectively. We also note that, for a Hamiltonian matrix from our EPM, a desired set of energies and wave functions within a target energy window can be extracted [65,66]. These merits would lead to a significant reduction of resources in making computation-driven database for specific materials properties.

We also tested a possible extension to include nonlocal correlations within DFT. Recent studies [67–70] have shown that the DFT with self-consistent intersite Hubbard interactions can yield excellent many-body quasiparticle energy gaps of semiconductors. Using the local charge density from our EPM, we performed one-shot calculations for the nonlocal interactions of Si and SiO_2 and obtained their energy bands shown in Fig. 9, agreeing with results from the aforementioned methods very well.

We recognize certain limitations inherent in our framework. First, while our model demonstrates effectiveness in predicting electronic properties such as band structures, its capability to accurately predict total energies is not yet fully developed. The current accuracy of our ML model, approximately 99.9%, does not meet the criteria required for reliable total energy predictions. Second, the training of our model has been focused on periodic systems of a single chemical composition, which restricts its direct inference to surfaces or systems with varied chemical compositions. Retraining with an expanded and more diverse dataset is required for the reliable predictions for these systems. The refinement of our model to improve total energy predictions and the development of a ML architecture and dataset to enhance extrapolation capabilities present opportunities for future research.

In summary, we present a new empirical computational method using neural network as accurate as *ab initio* methods based on DFT. Our method combines the merits of EPM with ML approaches by introducing a new hybrid descriptor reflecting local symmetries of materials reliably even when they have defects. We demonstrate the versatility and efficacy of our approach by computing energy bands of polymorphs of silica and silicon as well as their defective structures. Since transferable empirical pseudopotentials in our study can replace all local Hartree, atomic, and exchange-correlation potentials in KS Hamiltonians without self-consistency, results can be applied to all post-processing computational tools within existing first-principles calculations packages. Moreover, as demonstrated in defective SiO_2, our methodology holds promise for nonideal structures. So, we also expect that our methods can be used to generate transferable empirical pseudopotential for molecules, clusters, and disordered systems. We anticipate that our universal EPM will play important roles in constructing accurate computation-driven materials database.
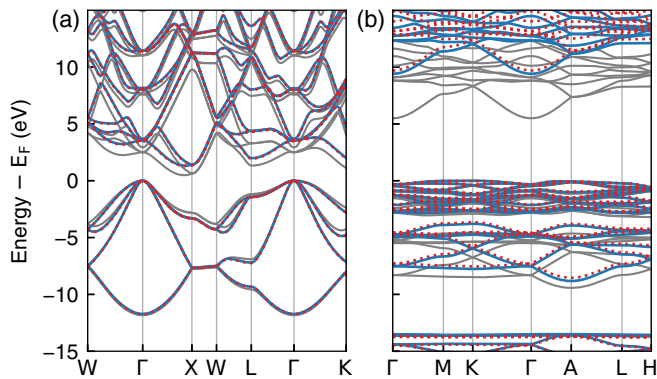

FIG. 9. Band structures of (a) Si and (b) SiO_2 incorporating the effects of the on-site interaction U and the intersite interaction V. In each figure, the solid gray lines represent LDA results, the solid blue lines denote self-consistent LDA$+U+V$ result, and the dotted red lines show the one-shot $+U+V$ results performed on top of the EPM results.

[1] P. Hohenberg and W. Kohn, Inhomogeneous electron gas, Phys. Rev. **136**, B864 (1964).

[2] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, Phys. Rev. **140**, A1133 (1965).

[3] R. O. Jones and O. Gunnarsson, The density functional formalism, its applications and prospects, Rev. Mod. Phys. **61**, 689 (1989).

[4] R. O. Jones, Density functional theory: Its origins, rise to prominence, and future, Rev. Mod. Phys. **87**, 897 (2015).

[5] S. G. Louie, Y.-H. Chan, F. H. da Jornada, Z. Li, and D. Y. Qiu, Discovering and understanding materials through computation, Nat. Mater. **20**, 728 (2021).

[6] N. Marzari, A. Ferretti, and C. Wolverton, Electronic-structure methods for materials design, Nat. Mater. **20**, 736 (2021).

[7] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature (London) **521**, 436 (2015).

[8] M. I. Jordan and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects, Science **349**, 255 (2015).

[9] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[10] F. Giustino, J. H. Lee, F. Trier, M. Bibes, S. M. Winter, R. Valentí, Y.-W. Son, L. Taillefer, C. Heil, A. I. Figueroa, B. Plaçais, Q. Wu, O. V. Yazyev, E. P. A. M. Bakkers, J. Nygård, P. Forn-Díaz, S. D. Franceschi, J. W. McIver, L. E. F. F. Torres, T. Low *et al.*, The 2021 quantum materials roadmap, J. Phys. Mater. **3**, 042006 (2020).

[11] H. J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit *et al.*, Roadmap on machine learning in electronic structure, Electron. Struct. **4**, 023004 (2022).

[12] J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[13] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. **104**, 136403 (2010).

[14] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, J. Comput. Phys. **285**, 316 (2015).

[15] L. Li, J. C. Snyder, I. M. Pelaschier, J. Huang, U.-N. Niranjan, P. Duncan, M. Rupp, K.-R. Müller, and K. Burke, Understanding machine-learned density functionals, Int. J. Quantum Chem. **116**, 819 (2016).

[16] V. Botu and R. Ramprasad, Learning scheme to predict atomic forces and accelerate materials simulations, Phys. Rev. B **92**, 094306 (2015).

[17] N. Artrith and A. Urban, An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO_2, Comput. Mater. Sci. **114**, 135 (2016).

[18] A. Khorshidi and A. A. Peterson, *Amp:* A modular approach to machine learning in atomistic simulations, Comput. Phys. Commun. **207**, 310 (2016).

[19] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K. R. Müller, Machine learning of accurate energy-conserving molecular force fields, Sci. Adv. E **3**, e1603015 (2017).

[20] L. Zhang, J. Han, H. Wang, R. Car, and W. E, Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics, Phys. Rev. Lett. **120**, 143001 (2018).

[21] H. Wang, L. Zhang, J. Han, and W. E, DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics, Comput. Phys. Commun. **228**, 178 (2018).

[22] K. Ryczko, D. A. Strubbe, and I. Tamblyn, Deep learning and density-functional theory, Phys. Rev. A **100**, 022512 (2019).

[23] Y. Chen, L. Zhang, H. Wang, and W. E, DeePKS: A comprehensive data-driven approach toward chemically accurate density functional theory, J. Chem. Theory Comput. **17**, 170 (2021).

[24] R. Nagai, R. Akashi, and O. Sugino, Completing density functional theory by machine learning hidden messages from molecules, npj Comput. Mater. **6**, 43 (2020).

[25] S. Dick and M. Fernandez-Serra, Machine learning accurate exchange and correlation functionals of the electronic density, Nat. Commun. **11**, 3509 (2020).

[26] J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, and A. J. Cohen, Pushing the frontiers of density functionals by solving the fractional electron problem, Science **374**, 1385 (2021).

[27] S. Kang, W. Jeong, C. Hong, S. Hwang, Y. Yoon, and S. Han, Accelerated identification of equilibrium structures of multicomponent inorganic crystals using machine learning potentials, npj Comput. Mater. **8**, 108 (2022).

[28] J. C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, and K. Burke, Finding density functionals with machine learning, Phys. Rev. Lett. **108**, 253002 (2012).

[29] G. Hegde and R. C. Bowen, Machine-learned approximations to density functional theory Hamiltonians, Sci. Rep. **7**, 42669 (2017).

[30] A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen, and R. Ramprasad, Solving the electronic structure problem with machine learning, npj Comput. Mater. **5**, 22 (2019).

[31] K. T. Schütt, M. Gastegger, A. Tkatchenko, K. R. Müller, and R. J. Maurer, Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions, Nat. Commun. **10**, 5024 (2019).

[32] M. Tsubaki and T. Mizoguchi, Quantum deep field: Data-driven wave function, electron density generation, and atomization energy prediction and extrapolation with machine learning, Phys. Rev. Lett. **125**, 206401 (2020).

[33] O. Unke, M. Bogojeski, M. Gastegger, M. Geiger, T. Smidt, and K.-R. Müller, SE(3)-equivariant prediction of molecular wavefunctions and electronic densities, in *Advances in Neural Information Processing Systems*, edited by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Curran Associates, Inc., 2021), Vol. 34, p. 14434.

[34] H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan, and Y. Xu, Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation, Nat. Comput. Sci. **2**, 367 (2022).

[35] X. Gong, H. Li, N. Zou, R. Xu, W. Duan, and Y. Xu, General framework for E(3)-equivariant neural network representation

of density functional theory Hamiltonian, Nat. Commun. **14**, 2848 (2023).

[36] J. Woo, H. Kim, and W. Y. Kim, Neural network-based pseudopotential: development of a transferable local pseudopotential, Phys. Chem. Chem. Phys. **24**, 20094 (2022).

[37] J. C. Phillips, Energy-band interpolation scheme based on a pseudopotential, Phys. Rev. **112**, 685 (1958).

[38] D. Brust, J. C. Phillips, and F. Bassani, Critical points and ultraviolet reflectivity of semiconductors, Phys. Rev. Lett. **9**, 94 (1962).

[39] D. Brust, M. L. Cohen, and J. C. Phillips, Reflectance and photoemission from Si, Phys. Rev. Lett. **9**, 389 (1962).

[40] M. L. Cohen and J. C. Phillips, Spectral analysis of photoemissive yields in Si, Ge, GaAs, GaSb, InAs, and InSb, Phys. Rev. **139**, A912 (1965).

[41] M. L. Cohen and T. K. Bergstresser, Band structures and pseudopotential form factors for fourteen semiconductors of the diamond and zinc-blende structures, Phys. Rev. **141**, 789 (1966).

[42] J. R. Chelikowsky and M. L. Cohen, Nonlocal pseudopotential calculations for the electronic structure of eleven diamond and zinc-blende semiconductors, Phys. Rev. B **14**, 556 (1976).

[43] M. L. Cohen and J. R. Chelikowsky, *Electronic Structure and Optical Properties of Semiconductors* (Springer-Verlag, Berlin, 1988).

[44] Y. W. Yang and P. Coppens, On the experimental electron distribution in silicon, Solid State Commun. **15**, 1555 (1974).

[45] C.-Y. Yeh, S. B. Zhang, and A. Zunger, Confinement, surface, and chemisorption effects on the optical properties of si quantum wires, Phys. Rev. B **50**, 14405 (1994).

[46] K. A. Mäder and A. Zunger, Empirical atomic pseudopotentials for AlAs/GaAs superlattices, alloys, and nanostructures, Phys. Rev. B **50**, 17393 (1994).

[47] A. Zunger and L.-W. Wang, Theory of silicon nanostructures, Appl. Surf. Sci. **102**, 350 (1996).

[48] L.-W. Wang and A. Zunger, Local-density-derived semiempirical pseudopotentials, Phys. Rev. B **51**, 17398 (1995).

[49] H. Fu and A. Zunger, Local-density-derived semiempirical nonlocal pseudopotentials for InP with applications to large quantum dots, Phys. Rev. B **55**, 1642 (1997).

[50] A. P. Bartók, R. Kondor, and G. Csányi, On representing chemical environments, Phys. Rev. B **87**, 184115 (2013).

[51] E. Kocer, J. K. Mason, and H. Erturk, A novel approach to describe chemical environments in high-dimensional neural network potentials, J. Chem. Phys. **150**, 154102 (2019).

[52] For the system in which we aim to compute potentials, we generated N_{str} distinct crystal configurations by making minor adjustments to the internal coordinates. Subsequently, we formulated the matrix equation

$$\begin{pmatrix} V_1(\mathbf{G}) \\ \vdots \\ V_{N_{str}}(\mathbf{G}) \end{pmatrix} = \begin{pmatrix} S_1^{Si}(\mathbf{G}) & S_1^{O}(\mathbf{G}) \\ \vdots & \vdots \\ S_{N_{str}}^{Si}(\mathbf{G}) & S_{N_{str}}^{O}(\mathbf{G}) \end{pmatrix} \begin{pmatrix} v^{Si}(\mathbf{G}) \\ v^{O}(\mathbf{G}) \end{pmatrix}$$

and inverted it to obtain potentials for Si and O.

[53] *Silica Physical Behavior, Geochemistry and Materials Applications*, edited by P. J. Heaney, C. T. Prewitt, and G. V. Gibbs, Mineralogical Society of America Reviews in Mineralogy Vol. 29 (Washington, D.C., 1994).

[54] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and

K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[55] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos *et al.*, QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys.: Condens. Matter **21**, 395502 (2009).

[56] M. J. van Setten, M. Giantomassi, E. Bousquet, M. J. Verstraete, D. R. Hamann, X. Gonze, and G.-M. Rignanese, The PSEUDODojo: Training and grading a 85 element optimized norm-conserving pseudopotential table, Comput. Phys. Commun. **226**, 39 (2018).

[57] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, DScribe: Library of descriptors for machine learning in materials science, Comput. Phys. Commun. **247**, 106949 (2020).

[58] M. A. Caro, Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials, Phys. Rev. B **100**, 024112 (2019).

[59] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, tensorflow.org (2015).

[60] F. Chollet *et al.*, Keras, https://keras.io (2015).

[61] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, arXiv:1802.08219.

[62] B. Anderson, T. S. Hy, and R. Kondor, Cormorant: Covariant molecular neural networks, in *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), Vol. 32, p. 14537.

[63] F. Fuchs, D. Worrall, V. Fischer, and M. Welling, SE(3)-transformers: 3D roto-translation equivariant attention networks, in *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Curran Associates, Inc., 2020), Vol. 33, p. 1970.

[64] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun. **13**, 2453 (2022).

[65] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods* (SIAM, Philadelphia, 1988).

[66] E. Polizzi, Density-matrix-based algorithm for solving eigenvalue problems, Phys. Rev. B **79**, 115112 (2009).

[67] V. L. Campo Jr. and M. Cococcioni, Extended DFT + U + V method with on-site and inter-site electronic interactions, J. Phys.: Condens. Matter **22**, 055602 (2010).

[68] S.-H. Lee and Y.-W. Son, First-principles approach with a pseudohybrid density functional for extended hubbard interactions, Phys. Rev. Res. **2**, 043410 (2020).

[69] N. Tancogne-Dejean and A. Rubio, Parameter-free hybridlike functional based on an extended Hubbard model: DFT $+ U + V$, Phys. Rev. B **102**, 155117 (2020).

[70] W. Yang, S.-H. Jhi, S.-H. Lee, and Y.-W. Son, *Ab initio* study of lattice dynamics of group IV semiconductors using pseudohybrid functionals for extended Hubbard interactions, Phys. Rev. B **104**, 104313 (2021).