

Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations

Akira Takahashi ^{*}, Yu Kumagai , Jun Miyamoto , Yasuhide Mochizuki , and Fumiyasu Oba *Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan*

(Received 29 July 2020; accepted 2 September 2020; published 9 October 2020)

Prediction models of both the electronic and ionic contributions to the static dielectric constants have been constructed using data from density functional perturbation theory calculations of approximately 1200 metal oxides via supervised machine learning. We developed two types of random forest regression models for oxides with the ground-state crystal structures: one model requires only compositional information and the other model also uses structural information. Although the training data included various atomic frameworks, the prediction models performed well even when only compositional information was used as feature descriptors. In prediction of the electronic contributions to the dielectric constants, the accuracies of the regression models with and without structural information were comparable, while the structural descriptors more clearly improved the prediction accuracy for the ionic contributions. We also analyzed the feature importance for prediction of the dielectric constants. The mean atomic mass and mass density were determined to be significant features in prediction of the electronic contributions without and with structural information, respectively. The standard deviation of the principal quantum number and mean neighbor distance variation were found to be important for the respective prediction models of the ionic contributions. The correlations between the dielectric constants and these features are discussed, along with the underlying physical mechanisms.

DOI: [10.1103/PhysRevMaterials.4.103801](https://doi.org/10.1103/PhysRevMaterials.4.103801)

I. INTRODUCTION

The dielectric response is one of the fundamental properties of materials, and it reflects their electronic and atomic structures. Furthermore, it is directly linked to important technological applications. Traditional capacitors use materials with superb dielectric properties that mostly originate from large ionic contributions. High- k dielectrics, which possess high static dielectric constants compared to silicon dioxide, sufficiently wide band gaps, and appropriate band alignment with semiconductors, are playing increasingly important roles as gate dielectrics and capacitors in semiconductor technology [1]. The electronic contributions to the dielectric constants, namely, the high-frequency or electronic dielectric constants, are related to the refractive indices, which are important quantities for designing optical applications, such as light extraction in light-emitting diodes [2], optical fibers [3], and antireflective coatings [4].

Among inorganic materials, metal oxides are good candidates for these applications because many metal oxides are stable under ambient conditions and easily synthesizable using various growth techniques. Moreover, some of them simultaneously possess wide band gaps and sufficiently high

dielectric constants. Typical examples include BaTiO₃ and SrTiO₃ for ceramic capacitor applications, and HfO₂ and related materials as gate dielectrics in field-effect transistors [5]. In addition, TiO₂ has been used as waveguides and resonators [6], Cr₂O₃ has been used in optical devices [7], and TiO₂-SiO₂ bilayers have been used as antireflective coatings [8,9].

In these applications, the static dielectric constant is an important parameter for the design and choice of not only the materials, but also the device structures. Theoretically, the static dielectric constants can be evaluated from first principles, often based on density functional perturbation theory (DFPT) [10]. Recent advances in the methodology, software, and hardware allow for high-throughput DFPT calculations, as well as density functional theory (DFT) calculations of structural and electronic properties [11–16]. To explore novel metal oxides and other inorganic materials with desirable dielectric properties and extract the overall physical and chemical trends, databases comprising several thousands of materials have been constructed. Examples include reports on both the electronic and ionic contributions to the static dielectric constants of 1762 oxides [17], 869 nonoxides [18], 1056 inorganic compounds [19], 2393 oxides [20], and 1521 semiconducting inorganic crystals [21], and the electronic static dielectric constants of 4040 materials, of which 3375 are oxides [22].

Machine learning (ML) techniques are promising to further accelerate the identification of novel dielectric oxides because they allow for the efficient exploration of target materials in a vast chemical space that cannot be fully covered by cost-intensive first-principles calculations. Such an approach, so-called *materials informatics*, also helps to uncover hidden

*takahashi.a.bb@m.titech.ac.jp

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

relationships among the physical, chemical, and structural properties, for example, the band gaps [23–25], melting points [26], and superconducting critical temperatures [27], have been related to the compositional and/or structural properties (e.g., composition, lattice parameter, and crystal structure).

For the electronic contribution to the static dielectric constants of inorganic materials, Morita *et al.* [28] recently reported ML prediction models using structure-dependent descriptors and discussed their relations to the preexisting Clausius-Mossotti (CM) and Penn models of the dielectric constants. Regarding the ionic contribution and the total static dielectric constants, to the best of our knowledge, reported regression models are limited to those for particular atomic frameworks. Pilania *et al.* [24] constructed prediction models of various material properties, including the dielectric constants, of one-dimensional chain structures. Noda *et al.* [29] reported a model for perovskite-type oxides using partial least-squares regression. Related to these studies, Umeda *et al.* [30] corrected the DFT calculation errors of the dielectric constants using ML, while Choudhary *et al.* [31] proposed a classification model for exploration of materials with dielectric constants higher than 10.

In this study, we constructed regression models to predict the electronic and ionic contributions to the static dielectric constants of oxides with the ground-state crystal structures, covering a wide range of atomic frameworks. High-throughput first-principles calculations based on DFT and DFPT were performed to generate the data for ML. To achieve accurate prediction by ML, the numerical errors of the training data should be removed as much as possible. In particular, dielectric constants calculated based on DFPT within the harmonic approximation are unreliable under certain conditions. Therefore, we selected the calculated data for ML in accordance with several criteria. We developed two ML models for each of the electronic and ionic contributions to the static dielectric constants: one relies on only compositional information, while the other uses both compositional and structural information as input descriptors. Such ML models will be useful for identifying superior dielectric materials in a vast chemical space. Moreover, we analyzed the importance of the descriptors to obtain physical insight into the dielectric constants.

This paper is organized as follows: In Sec. II, we describe the technical details of the high-throughput first-principles calculations of the dielectric constants and construction of the ML prediction models. In Sec. III, we discuss the trend of the dielectric constants, accuracy of the ML models, and importance of the descriptors. Finally, we summarize and conclude the work in Sec. IV.

II. METHODOLOGY

A. High-throughput first-principles calculations and database construction

The first-principles calculations based on DFT and DFPT were performed using the projector augmented-wave method [32] as implemented in the VASP code [33,34]. The Perdew-Burke-Ernzerhof functional tuned for solids (PBEsol) [35] was used for the exchange-correlation energy and potential,

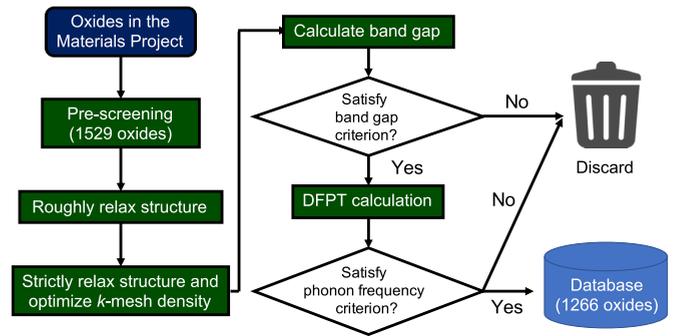


FIG. 1. Workflow for constructing a computational database of the dielectric constants of oxides.

and Hubbard U corrections were used to describe the localized nature of the d and f states [36,37]. The effective U parameters used in this work are given in Table S1 in the Supplemental Material [38].

We constructed the computational database for ML following the workflow shown in Fig. 1. We focused on oxides with the ground-state crystal structures (the lowest total energies among known polymorphs) and sufficiently large band gaps. Target oxides were selected from the Materials Project database [11] using the criteria of excess energy less than 0.1 meV/atom with respect to the convex hull and band gap larger than 0.3 eV. Note that the band gaps tend to be underestimated in the Materials Project because of the use of PBE(+ U) [39], and the degree of the underestimation is 50% or even more in the worst cases [16,40]. Mixed anion compounds were excluded. Further, because the accuracy of calculated dielectric constants is closely related to that of the ML model, we selected the candidate oxides for the ML data sets in accordance with the following criteria: (1) oxides with finite local magnetic moments, such as $3d$ transition metal oxides, were excluded because calculations of magnetic systems are less convincing when using the PBEsol functional even with Hubbard U corrections and (2) oxides with space group $P1$ and/or containing more than 40 atoms in the primitive unit cell were removed because their large degrees of freedom in the lattice parameters and internal atomic coordinates and/or high computational demand can make fine-structure optimization difficult. The total number of candidates for the DFT and DFPT calculations was 1529.

The computational results of the ionic static dielectric tensors strongly depend on the lattice parameters and internal atomic coordinates. Therefore, we performed the following two-step structure optimization procedure: (1) optimization with a rough condition using a coarse k -point mesh until the residual atomic forces and stresses were less than 0.2 eV/Å and 0.1 GPa, respectively, and (2) optimization with much tighter conditions of 0.1 meV/Å and 0.02 GPa. The first optimization is useful to roughly optimize the structures with reduced computational cost, particularly when the optimized structures are far from the initial structures, although such cases were rare in this study where the initial structures were taken from the Materials Project database. In the second optimization, the convergence of the structure and energy was also checked as a function of the k -point mesh density (see Fig. S1

in the Supplemental Material [38] for the detailed workflow). After determining the theoretical crystal structures with high accuracy, the ground-state structure of each composition was selected if more than two structures existed after the aforementioned screening of the Materials Project database. This procedure typically gave the same ground states as those in the Materials Project, despite the difference in the approximations and computational details used in the calculations. We then calculated the electronic band structures to obtain the band gaps. The phonon frequencies at the Γ points and dielectric tensors including local field effects were evaluated on the basis of DFPT using k -point meshes that were $2^3 = 8$ times denser than those determined from the structure and energy convergence. The electronic densities of states (DOSs) and complex dielectric functions were also calculated for selected materials using $3^3 = 27$ times denser k -point meshes than those used for the structure and energy convergence.

Although the workflow shown in Fig. 1 is seemingly simple, its automation is fairly complicated. To achieve automation, we used several open source packages. The SPGLIB [41] code was used to symmetrize the crystal structures and identify their space groups, and the SEEKPATH code [42] was used to determine the paths in the reciprocal spaces for the band-structure calculations. Furthermore, we generated VASP input files and controlled workflows using in-house programs, which rely on PYMATGEN [43], FIREWORKS [44], CUSTODIAN [43], and ATOMATE [45].

To construct a robust database for the above purpose, we discarded the following oxides. First, we ruled out oxides with band gaps smaller than 0.5 eV because small band gaps can cause large numerical errors in the electronic dielectric tensors. In addition, oxides with imaginary frequencies of the optical phonon modes at the Γ point were excluded because they are dynamically unstable and out of the scope of this study. We also excluded oxides with real optical phonon frequencies lower than 0.1 THz at the Γ point, which may contain large numerical errors in the ionic dielectric tensors. Additionally, the calculated frequencies of the acoustic phonon modes are slightly imaginary because of the residual atomic forces caused by the truncation error of structure optimization. Therefore, we discarded oxides exhibiting acoustic phonon modes with imaginary frequencies at the Γ point with absolute values larger than 0.3 THz as unconvincing computational results. Oxides whose calculations were aborted by unintentional VASP errors, those duplicated with others after structure optimization, and those for which we could not define the values of the descriptors listed in Table S3 in the Supplemental Material were also excluded [38].

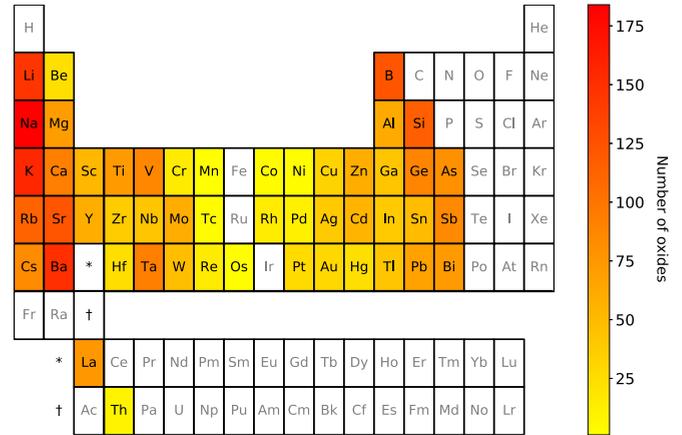


FIG. 2. Number of oxides containing the respective cation species in the data set.

As a result, we obtained the dielectric tensors of 1266 oxides covering a wide range of compositions and structures. The distribution of these oxides with respect to the constituent cations is shown in Fig. 2. Oxides with 50 types of cations remained in the database. The distribution bias of the data set is because of two reasons. First, the Materials Project database is mostly based on experimentally reported materials and can therefore be prone to contain materials that are cheap, nontoxic, easy to synthesize and handle, and/or stable under ambient conditions, although this is natural for a material database in some sense. The second reason pertains to our criteria to select materials: the elimination of magnetic systems removes most oxides of transition-metal elements and lanthanides with half-filled d - or f -electrons from the data set. More statistics of the computational database are given in Fig. S2 and Table S4 in the Supplemental Material [38].

B. Theoretical background of dielectric constant calculation

Among the methodologies for evaluating the electronic contribution to the dielectric tensors, we used the DFPT approach including local field effects, as described above. In addition, we also performed calculations of the complex dielectric functions at the independent-particle level to discuss the electronic contribution from the viewpoint of the character of the electronic states. For the non-spin-polarized systems considered in this study, the imaginary part of the electronic macroscopic dielectric constant as a function of frequency ω can be expressed as [46]

$$\varepsilon_{\alpha\beta}^{(2)}(\omega) = \frac{4\pi e^2}{\Omega} \lim_{q \rightarrow 0} \frac{1}{q^2} \sum_{c,v} 2 \int \langle u_{c\mathbf{k}+\mathbf{e}_\alpha q} | u_{v\mathbf{k}} \rangle \langle u_{c\mathbf{k}+\mathbf{e}_\beta q} | u_{v\mathbf{k}} \rangle^* \delta(\epsilon_{c\mathbf{k}} - \epsilon_{v\mathbf{k}} - \omega) d\mathbf{k}, \quad (1)$$

where $u_{v\mathbf{k}}(u_{c\mathbf{k}})$ and $\epsilon_{v\mathbf{k}}(\epsilon_{c\mathbf{k}})$ are the cell-periodic part of a valence (conduction) band wavefunction and its eigenvalue, respectively, α and β ($=x, y, z$) are the Cartesian directions, \mathbf{e}_α and \mathbf{e}_β are unit vectors along the α and β directions, e

is the charge of an electron, and Ω is the volume of the primitive unit cell. The summations with respect to c and v run over the conduction and valence band indices, respectively. The electronic contribution to a static macroscopic dielectric

tensor, namely, an electronic static dielectric tensor, is then obtained by the Kramers-Kronig transformation

$$\varepsilon_{\alpha\beta}^{(1)}(\omega)\Big|_{\omega=0} = 1 + \frac{2}{\pi} P \int_0^{\infty} \frac{\varepsilon_{\alpha\beta}^{(2)}(\omega')}{\omega'} d\omega', \quad (2)$$

where P is the principal value. These equations indicate that the electronic static dielectric tensor is mainly determined by three factors. The first factor is the joint DOS composed of the valence and conduction band DOSs in Eq. (1), which is defined as

$$2 \int \delta(\varepsilon_{c\mathbf{k}} - \varepsilon_{v\mathbf{k}} - \omega) d\mathbf{k}. \quad (3)$$

The second factor is $\lim_{q \rightarrow 0} \frac{1}{q^2} \langle u_{c\mathbf{k}+\mathbf{e}_{\alpha}q} | u_{v\mathbf{k}} \rangle \langle u_{c\mathbf{k}+\mathbf{e}_{\beta}q} | u_{v\mathbf{k}} \rangle^*$ in Eq. (1), which is proportional to $\langle u_{c\mathbf{k}} | i\nabla_{\alpha} - \mathbf{k}_{\alpha} | u_{v\mathbf{k}} \rangle \langle u_{c\mathbf{k}} | i\nabla_{\beta} - \mathbf{k}_{\beta} | u_{v\mathbf{k}} \rangle^*$ [46,47]. When considering the situation where $\mathbf{k} = \mathbf{0}$ and $\alpha = \beta$, it becomes $|\langle u_{c\mathbf{k}} | i\nabla_{\alpha} | u_{v\mathbf{k}} \rangle|^2$, indicating how much the valence band states transition to the conduction band states under an electric field. The third factor is the band gap and widths of the valence and conduction bands because the frequency is the denominator of the integrand in Eq. (2) [22].

The ionic contribution to a static dielectric tensor is given by

$$\varepsilon_{\text{ion},\alpha\beta} = 4\pi \sum_m \frac{f_{m,\alpha\beta}^2}{\omega_m^2}, \quad (4)$$

where ω_m is the frequency of optical phonon mode m . $f_{m,\alpha\beta}$ is the oscillator strength tensor, which is expressed as

$$f_{m,\alpha\beta} = \sqrt{\frac{1}{\Omega} \left(\sum_{\kappa\alpha'} Z_{\kappa,\alpha\alpha'} U_m^*(\kappa\alpha') \right) \left(\sum_{\kappa\beta'} Z_{\kappa,\beta\beta'} U_m^*(\kappa\beta') \right)}, \quad (5)$$

where $Z_{\kappa,\alpha\beta}$ is the Born effective charge (BEC) tensor of ion κ , and $U_m^*(\kappa\alpha')$ is the atomic displacement expressed by an eigenvector of phonon mode m . The summations run over the indices of ion κ in the primitive unit cell and directions α' and β' ($=x, y, z$) in the space. The BEC tensor is defined as

$$Z_{\kappa,\alpha\beta} = \Omega \frac{\partial P_{\beta}}{\partial \tau_{\kappa,\alpha}}, \quad (6)$$

where $\tau_{\kappa,\alpha}$ is the atomic perturbation displacement acting on ion κ along direction α and P_{β} is the macroscopic polarization along direction β .

These equations indicate that the average BECs are approximately quadratically correlated with the ionic dielectric tensors. Regarding the frequencies, lower-lying optical phonon modes have a greater effect on the ionic dielectric tensors, and the inverse of their frequency is quadratically related to the ionic dielectric tensors.

In this study, for simplicity, we discuss the spherically averaged electronic (ε_{el}) and ionic (ε_{ion}) static dielectric constants. From Eq. (4), ε_{ion} is very sensitive to the phonon frequency when the lower-lying optical phonon modes have very low frequencies. Although the aforementioned procedure excluded such oxides from the data set to some or a large extent, the calculated ε_{ion} may still possess large numerical error when ε_{ion} is large. We therefore used the logarithms of the dielectric

constants as the objective variables of ML to suppress the effect of such a numerical error.

C. Machine learning

To construct ML models, information about the materials needs to be converted to descriptors representing the features of the materials. We used 81 compositional and 553 structural descriptors extracted using the PYMATGEN [43] and MATMINER [40] packages (see Table S3 in the Supplemental Material [38]). Two models were constructed for each of ε_{el} and ε_{ion} : one that uses only compositional descriptors (ML without SD) and the other that uses structural descriptors (SD) in addition to the compositional descriptors (ML with SD). Although the first model can be used for the prediction of dielectric constants even when the crystal structure for the target material is unknown and only its chemical formula is given, it should be less accurate than the second because of less representational ability. However, there are one-to-one correspondences between the compositions and ground-state crystal structures and, therefore, the accuracy of the first should be systematically improved by increasing the size of the training data set.

We used a random forest (RF) regression method [48], one of the supervised ML techniques, as implemented in the SCIKIT-LEARN code [49]. An RF model consists of many decision trees constructed from randomly sampled subtraining data sets. The target value is predicted by averaging the output values of the decision trees. The RF technique allows the importance of the descriptors to be evaluated from the average of the variance reductions of the decision trees.

To prevent overfitting caused by a large number of descriptors, we pruned the descriptors with the lowest 20% importance values during one training iteration [50]. The best ML model during the recursive process was determined from the cross-validation (CV) score, the procedure of which is shown in Fig. 3. Although several ML models were obtained by performing the recursive process several times from the beginning, we report the results with the best CV scores.

III. RESULTS AND DISCUSSION

A. Validation of the DFPT calculations

First, we compared the theoretical dielectric constants of ten representative oxides with the experimental values (Table I).

The calculated ε_{el} values are in good agreement with the experimental values, although they are slightly overestimated, partly because of underestimation of the band gaps by PBEsol(+ U) (see Sec. II B). Such tendencies have also been found for prototypical binary oxides [55,56] and semiconductors [16]. The calculated ε_{ion} values deviate more from the experimental values, in particular, ε_{ion} of SrTiO₃ is only 13% of the experimental value at room temperature.

The difference of the ε_{ion} value of SrTiO₃ from the experimental value is mainly attributed to three reasons: (1) the crystal structure difference; (2) ignoring of the quantum nuclear effect; and (3) the approximation of the exchange-correlation functional. For (1), SrTiO₃ is cubic at room temperature, while our computational ground-state structure model corresponds to the low-temperature tetragonal phase.

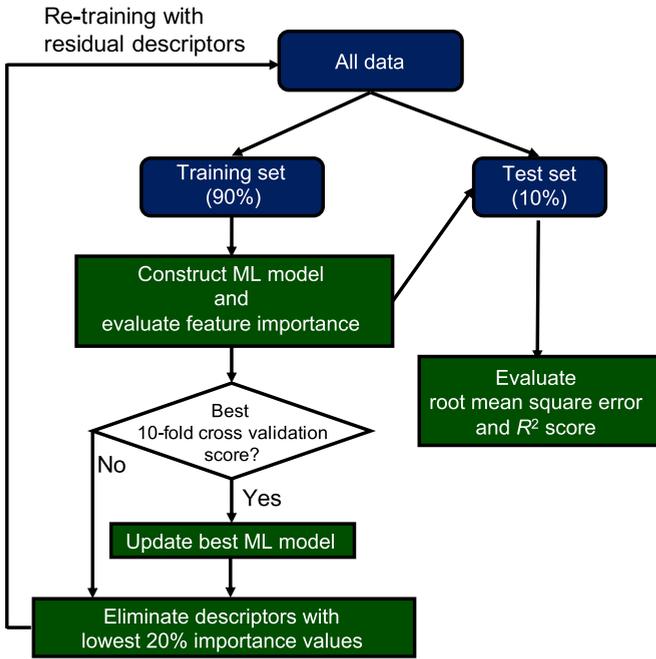


FIG. 3. Schematic diagram of the recursive feature elimination process.

Regarding (2), we ignored the quantum paraelectric nature of SrTiO₃ [57]. Indeed, the dielectric constant of SrTiO₃ becomes much larger at low temperature (approximately 18,000 at 1.4 K [58]), which makes the discrepancy even larger. Regarding (3), using the semilocal PBEsol exchange-correlation functional with Hubbard U corrections to the Ti-3d states can cause some error. Petousis *et al.* [59] pointed out that there is a large discrepancy between the ϵ_{ion} value calculated using PBEsol and that calculated using the local density approximation. We also found such functional and Hubbard U parameter dependencies for SrTiO₃ (see Table S5 in the Supplemental Material [38]). This functional issue also affects the ϵ_{ion} values of other materials. Other effects, such as ignoring the temperature effect in the calculations (e.g.,

electron-phonon and phonon-phonon coupling), and imperfections in the samples in the experiments (e.g., point defects and grain boundaries) also make a direct comparison between theory and experiment difficult. Excluding SrTiO₃, the root-mean-squared percentage errors (RMSPE) between the computational and experimental values are 10.2% and 21.8% for ϵ_{el} and ϵ_{ion} , respectively.

The distributions of ϵ_{el} and ϵ_{ion} with respect to the band gap are shown in Figs. 4(a) and 4(b), respectively. First, and most importantly, ϵ_{ion} is generally about an order of magnitude greater than ϵ_{el} . Therefore, focusing on ϵ_{ion} is essential for discovering high- k materials, as has been previously believed. The largest ϵ_{el} value at a given band gap tends to exhibit an inverse correlation with the band gap, as shown in Fig. 4(a), which is consistent with the previously reported tendency [18,19,22]. This correlation can be understood from Eq. (2), which will be discussed in Sec. III B. Here, we used the minimum band gaps for both the direct- and indirect-type band structures in Fig. 4. The threshold energies of the imaginary part of the dielectric function [ϵ'' in Eq. (1)] correspond to the direct band gaps, except for the cases where electronic transitions over the direct gaps are symmetrically forbidden [16,60]. Therefore, the direct gaps should be more relevant to the electronic dielectric constants, particularly for materials with indirect-type band structures, as pointed out by Naccarato *et al.* [22]. However, in an analysis of the statistics, using the minimum or direct gaps does not substantially change the correlation between the band gaps and the dielectric constants (see Fig. S3 in the Supplemental Material [38]), and we use the minimum gaps for simplicity.

To take a closer look at the spread of ϵ_{el} for materials with similar band gaps, we calculated the complex dielectric functions of PtO₂ and Cs₂O. Both oxides have indirect-type band structures, as shown along with their dielectric functions and DOSs in Fig. S4 in the Supplemental Material [38]. The first shows the largest ϵ_{el} (12.5) with minimum and direct band gaps of 0.95 and 1.30 eV, while the second has a low ϵ_{el} (4.6) and 0.93-eV minimum and 1.50-eV direct band gaps. In ϵ'' of PtO₂, a sharp peak is located slightly above the photon energy corresponding to the direct band-gap value. This spectrum

TABLE I. Comparison of the dielectric constants of ten oxides from the DFPT calculations and the reported experimental values at room temperature [51–54]. ϵ_{el} and ϵ_{ion} are the electronic and ionic static dielectric constants, respectively. Some of the experimental ϵ_{el} values were estimated as the squares of the refractive indices. The experimental ϵ_{ion} values were obtained by subtracting ϵ_{el} from the total static dielectric constants.

Compound	Space group (DFT)	ϵ_{el} (DFPT)	ϵ_{el} (experiment)	ϵ_{ion} (DFPT)	ϵ_{ion} (experiment)
SrTiO ₃	$I4/mcm$	6.1	5.3 ^a [51]	38.0	~ 300 ^a [51]
Al ₂ O ₃	$R\bar{3}c$	3.3	3.1 [51]	6.9	7.0 [51]
MgO	$Fm\bar{3}m$	3.2	3.0 [51]	6.8	6.7 [51]
ZnO	$P6_3mc$	4.3	3.7 [52]	4.4	4.7 [52]
SrO	$Fm\bar{3}m$	3.9	3.5 [52]	15.6	11.0 [52]
BeO	$P6_3mc$	3.1	2.9 [51]	3.9	4.2 [51]
CaO	$Fm\bar{3}m$	3.9	3.3 [52]	12.1	8.7 [52]
Mg ₂ SiO ₄	$Pnma$	2.9	2.7 [53]	4.2	4.5 [53,54]
LiTaO ₃	$R3c$	4.9	4.6 [51]	29.8	41.8 [51]
LiGaO ₂	$Pna2_1$	3.3	3.1 [51]	3.7	4.1 [51]

^aValue for the cubic ($Pm\bar{3}m$) phase.

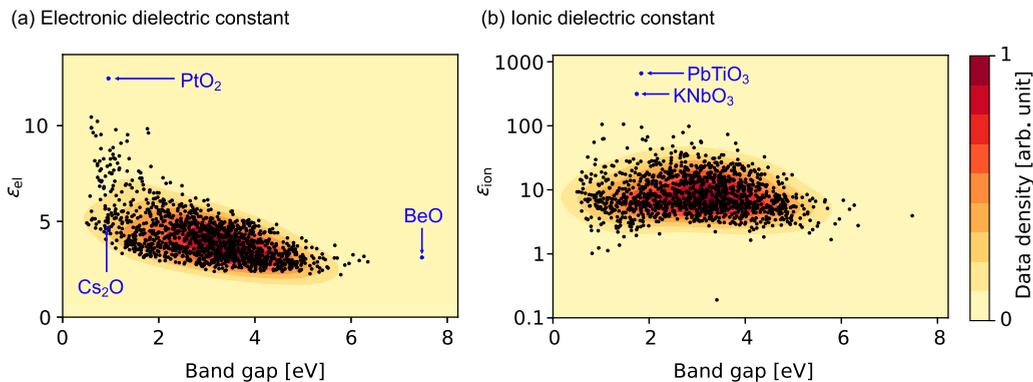


FIG. 4. Distributions of the (a) electronic and (b) ionic static dielectric constants with respect to the band gap. The data density was estimated by kernel density estimation.

shape would lead to the large ϵ_{el} value given Eq. (2). Such a peak is ascribed to the characteristics that both the valence and conduction bands are composed mainly of the covalent bonding between the Pt-5*d* and O-2*p* states. Conversely, the valence band of Cs₂O is mainly composed of the O-2*p* states and the conduction band is composed of the Cs-6*s*, 6*p*, and 5*d* states, which is associated with its strong ionic nature. The DOS near the CBM is much smaller than that of PtO₂. Further, because the volumetric density of the oxide ions of Cs₂O (0.010 O²⁻/Å³) is lower than that of PtO₂ (0.064 O²⁻/Å³), the DOS originating from the O-2*p* states is also lower. These factors lead to the lower joint DOS expressed by Eq. (3). Actually, $\epsilon^{(2)}$ of Cs₂O gradually increases when moving from the threshold and the height of the main peak is much lower than that of PtO₂. The low ϵ_{el} value can be ascribed to the spectrum shape of $\epsilon^{(2)}$.

In contrast to the case of ϵ_{el} , ϵ_{ion} does not show such an obvious correlation with the band gap [Fig. 4(b)]. This tendency appears to be consistent with Eq. (4), which does not explicitly include the band gap. However, because the BEC has a relationship with the change of orbital hybridization [61], the band gap could have some correlation with ϵ_{ion} , which will be discussed in Sec. III C. In our data set, the tetragonal (*P4mm*) phase of PbTiO₃ has an extremely large ϵ_{ion} value (661.4). We calculated its phonon band structure (Fig. S5 in the Supplemental Material [38]). As anticipated, the lowest-lying optical mode shows a very low frequency at the Γ point. Furthermore, the BEC of Ti (+6.7) is much larger than the formal charge of +4. However, three things need to be kept in mind. First, as discussed above, the errors of the lattice parameters and internal atomic coordinates may change the frequency by a few terahertz or more, which makes the accuracy of the predicted large ϵ_{ion} uncertain. Second, imaginary frequencies appear at the *R* and *A* points, indicating that PbTiO₃ is dynamically unstable under our computational conditions. We were not able to exclude such cases in the database construction procedure, where the phonon modes were only investigated at the Γ point. Finally, the calculated ground-state structure and ϵ_{ion} are sensitive to the effective U (U_{eff}) parameter. Investigating the ground-state structure of PbTiO₃ by imaginary phonon mode analysis [62,63], we found that the most stable structure with U_{eff} (Ti) = 3 eV is rhombohedral (*R3c*), which is not available in the Materials Project database and inconsistent with the experimentally

observed tetragonal structure at low temperature [64]. However, when U_{eff} (Ti) is set to 0 eV, the space group of the theoretical ground-state structure of PbTiO₃ is *P4mm* and ϵ_{ion} decreases to 103.9.

Yim *et al.* reported that cubic BeO in the rocksalt structure shows both a large band gap and a high dielectric constant, significantly deviating from the observed distribution of the dielectric constant with respect to the band gap [17]. However, it is not included in our data set because the ground state of BeO is not the rocksalt phase but the wurtzite phase, the latter of which shows a much lower static dielectric constant (7.1) than the former (274 in Ref. [17]).

B. Prediction models of the electronic dielectric constants

We first investigated the performance of the ML models of ϵ_{el} . We estimated the prediction errors of the ML models using the root-mean-squared errors (RMSE) and coefficients of determination (R^2) of the test set. The ML without SD achieves $R^2 = 0.87$, while ML with SD is slightly better with $R^2 = 0.89$ (Fig. 5). The RMSPEs of the ML without and with SD before taking the logarithm are 9.3% and 8.5% for the test set, respectively. These results indicate that the compositional information well describes ϵ_{el} for the oxides considered in this study.

Because ML regression is generally interpolative, its prediction accuracy is linked to the abundance of similar training data. From the opposite point of view, the presence of outliers in the test set, typical examples of which are given in Table II, can be ascribed to the deviation of the chemical composition and/or structure type from the main ones. To visualize this, we classified our data set into 641 structure types using the structure matcher implemented in PYMATGEN [43] and counted their frequencies in the training data. We found that the outliers tend to comprise rare cation species in our data set, irrespective of the structure type. Examples are LaRhO₃, BaHgO₂, and Re₂PbO₈ with Rh, Hg, and Re, respectively (Fig. 2).

To validate the accuracies of the ML models and analyze the control factors of ϵ_{el} , we will now discuss the most important descriptors of each ML model. For the ML model without SD, the mean value of the atomic masses (MAM), which depends on the cation species in the oxides, was found to be the most important descriptor [Fig. 6(a)]. The correlation

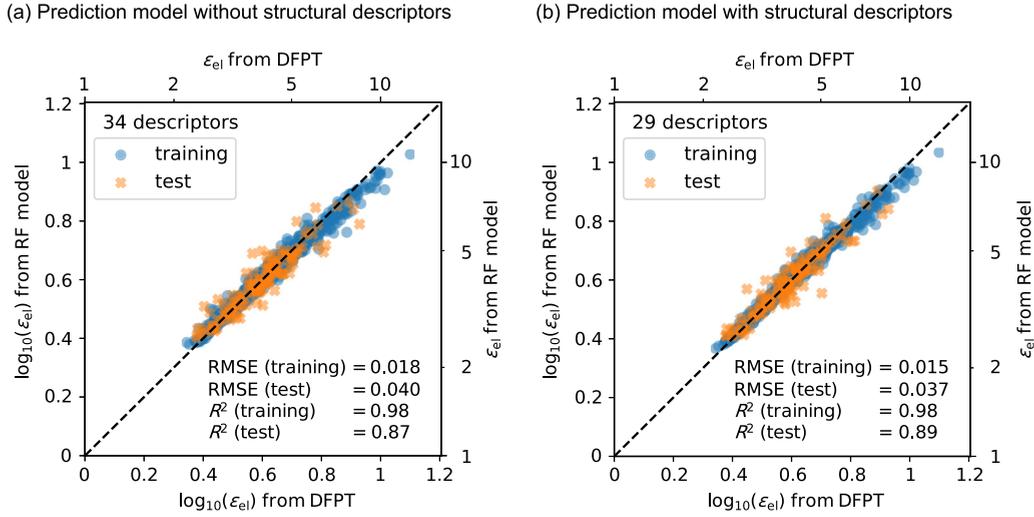


FIG. 5. Parity plots of the electronic dielectric constants ϵ_{el} between the reference DFPT values and ML predicted values (a) without and (b) with structural descriptors. The blue circles and orange crosses represent the training data (used to construct the model) and test data (not used to construct the model), respectively.

coefficient (r) between the MAM and ϵ_{el} is 0.6 [Fig. S6(a) in the Supplemental Material [38]].

This positive correlation can be interpreted in two ways. First, heavy cations with large ionic radii have long bond lengths with oxide ions, weakening the Madelung potential and thereby lowering the band gap in a simple picture of the ionic crystal. As mentioned above, the reciprocal relationship between the band gap and ϵ_{el} has been demonstrated both theoretically and empirically in the present and previous studies [17,22,28,70–73]. In fact, the MAM has a negative correlation with the band gap ($r = -0.5$) in our database [Fig. S7(a) in the Supplemental Material [38]]. Thus, a larger MAM would tend to give higher ϵ_{el} .

Additionally, the positive correlation between the MAM and ϵ_{el} can be interpreted on the basis of the CM model (see Sec. IX in the Supplemental Material [38] for details). From

the CM relation, ϵ_{el} can be expressed as

$$\epsilon_{el} = \frac{3}{1 - \frac{4\pi}{3} \sum_j N_j \alpha_j} - 2, \quad (7)$$

where N_j and α_j are the number of ion j per unit volume and electronic polarizability of ion j . The large electronic polarizability of ions (or atoms) increases ϵ_{el} in Eq. (7). Heavy cations tend to possess large atomic polarizabilities because the energies of their frontier orbitals are close and these orbitals easily hybridize [74]. In fact, there is a robust positive correlation between the electronic polarizabilities of gaseous cations [75] and the atomic masses [Fig. S10(b) in the Supplemental Material [38]]. Note that the CM model is based on α_j , i.e., hybridization of the atomic orbitals within each isolated ion under an electric field. However, in reality, hybridization

TABLE II. Five oxides with the largest prediction errors in the ML models of the electronic dielectric constants ϵ_{el} (a) without and (b) with structural descriptors.

Compound	Space group	Number of training data entries with the same structure type	ϵ_{el} (DFPT)	ϵ_{el} (ML)	$\log_{10}(\epsilon_{el} \text{ (ML)}) - \log_{10}(\epsilon_{el} \text{ (DFPT)})$	Percentage error
(a)						
LaRhO ₃	<i>Pnma</i>	11	8.4	6.1	-0.14	-26%
BaHgO ₂	<i>R$\bar{3}m$</i>	6	3.7	4.9	0.12	+33%
SrPbO ₃	<i>Pnma</i>	11	6.4	4.9	-0.11	-23%
KZnB ₃ O ₆	<i>P$\bar{1}$</i>	1	2.5	3.2	0.11	+28%
Re ₂ PbO ₈	<i>P31m</i>	1	4.0	5.0	0.10	+26%
(b)						
Li ₃ CuO ₃	<i>P4₂/mnm</i>	3	5.0	3.6	-0.15	-29%
Na ₂ WO ₄	<i>Fd$\bar{3}m$</i>	11	2.8	3.7	0.12	+31%
Re ₂ PbO ₈	<i>P31m</i>	1	4.0	5.0	0.10	+25%
GaCuO ₂	<i>R$\bar{3}m$</i>	6	5.2	6.4	0.09	+24%
LaRhO ₃	<i>Pnma</i>	11	8.4	6.9	-0.09	-18%

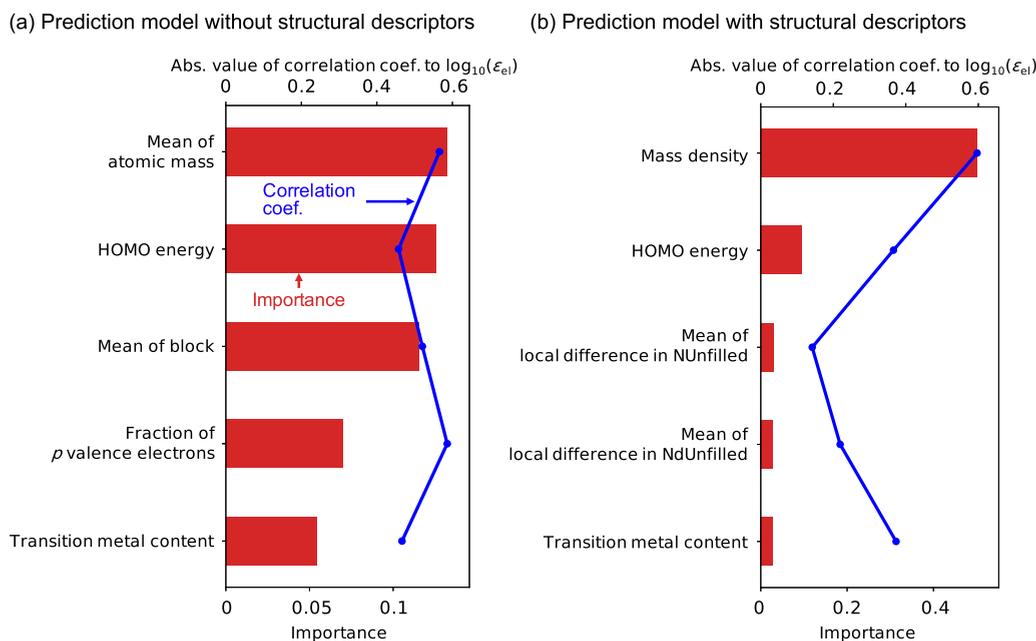


FIG. 6. Importance values of the five most significant descriptors of the prediction models of the electronic dielectric constants ϵ_{el} (a) without and (b) with structural descriptors. The red bars and blue lines represent the importance values of RF regression and the absolute values of the correlation coefficients, respectively. The details of the highest occupied molecular orbital (HOMO) energy, fraction of *p* valence electrons, mean of the local difference in NUnfilled, mean of the local difference in NdUnfilled, and transition metal content can be found in Refs. [65], [66,67], [68,69], [68,69], and [67], respectively.

of the orbitals of different ions should be important because the valence and conduction bands of most metal oxides mainly consist of O-2*p* and cation orbitals, respectively. We compared ϵ_{el} predicted from the DFPT calculations to those from the CM model with the electronic polarizabilities of gaseous ions calculated by Pauling [75] but the prediction was rather inaccurate and the dielectric constants were negative in some cases. (RMSE = 127.4 and $R^2 = -2.4 \times 10^4$ for nonlogarithm values). This is probably because of the effects in solids, such as hybridization of the orbitals of different ions, not being taken into account in the CM model.

Regarding ML with SD, the mass density of the crystal was found to be the most important descriptor [Fig. 6(b)]. It has a more robust positive correlation ($r = 0.7$) with ϵ_{el} than the MAM [Fig. S6(b) in the Supplemental Material [38]]. This dependency can be partly explained by the same logic as the MAM, as indicated from the fact that the mass density has a positive correlation with the MAM ($r = 0.7$) [Fig. S7(b) in the Supplemental Material [38]]. Furthermore, the mass density also indicates how close the ions (or atoms) are packed, and the improvement of the prediction accuracy can be ascribed to such information. In fact, a higher number density of ions leads to a higher number density of electrons and DOS per volume, thereby increasing ϵ_{el} . The CM model also indicates that larger density of atoms N_j leads to larger ϵ_{el} .

Pantelides [76] also showed theoretically that ϵ_{el} can be expressed as functions of the interatomic spacing and cation-dependent parameters. Braun *et al.* [77] reported that the effective refractive index of SiO₂ can be defined as a function of the porosity, while Malyi *et al.* [78] demonstrated that ϵ_{el} of the polymorphs of SiO₂ can be represented as

the reciprocal of their volumes by DFT calculations. Thus, our derived significant descriptors, i.e., the MAM and mass density, are compatible with the previously proposed physical and chemical pictures.

C. Prediction models of the ionic dielectric constants

We will now discuss the ML models of ϵ_{ion} . The prediction accuracy of the ML without SD is worse in terms of the R^2 values for the test set ($R^2 = 0.65$) [Fig. 7(a)] than that for ϵ_{el} ($R^2 = 0.87$) [Fig. 5(a)]. The structural descriptors clearly improve the R^2 value for ϵ_{ion} from 0.65 to 0.73 [Fig. 7(b)], unlike the case of ϵ_{el} . However, the RMSPEs for the test set in the ML without and with SD are comparable (48.3% and 48.0%, respectively). The top five outliers in the test set for each of the ML models are listed in Table III. As in the case of ϵ_{el} , the behavior of these outliers is partly ascribed to bias of the database. Outliers tend to occur when their structure types are rare in the training data. This tendency is contrary to that of ϵ_{el} , whose prediction errors can be large when the constituent cation species are rare in the database even when the structure type is common in the training data (Table II). The percentage errors are the largest for Au₂O₃ for both the models without and with SD (−323% and −371%, respectively). These large errors of Au₂O₃ significantly affect the RMSPEs for all of the data. When Au₂O₃ is excluded, the RMSPEs without and with SD decrease to 39.0% and 35.0%, respectively, but these values are still worse than those for ϵ_{el} . The prediction accuracy of ϵ_{ion} is lower, probably because (1) the numerical errors of ϵ_{ion} by DFPT calculations are larger than those of ϵ_{el} , (2) ϵ_{ion} is much more sensitive to the structure (e.g., a

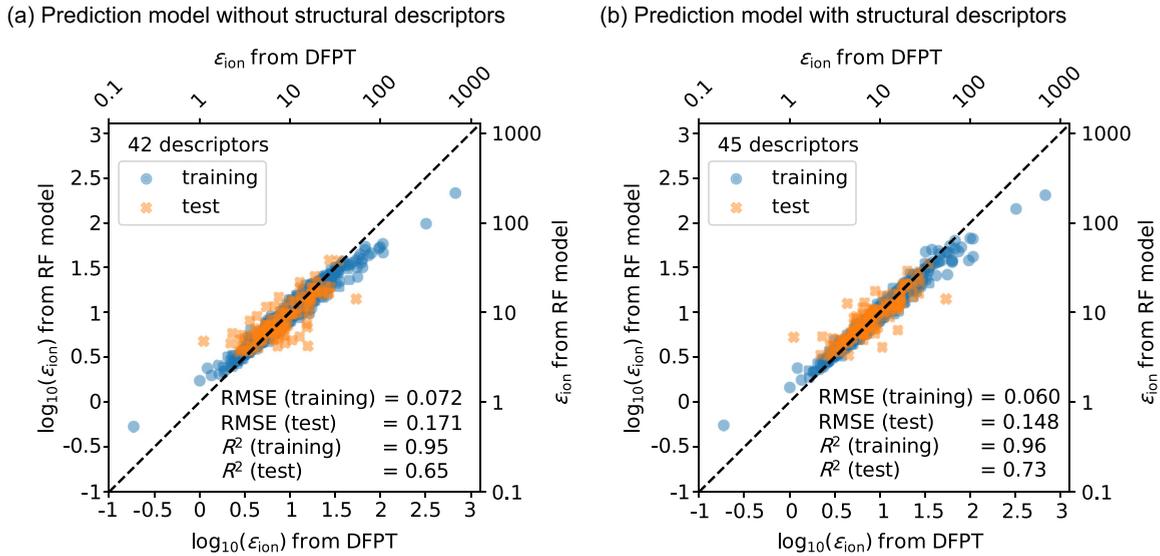


FIG. 7. Parity plots of the ionic dielectric constants ϵ_{ion} between the reference DFPT values and ML predicted values (a) without and (b) with structural descriptors. The blue circles and orange crosses represent the training data (used to construct the model) and test data (not used to construct the model), respectively.

subtle difference in the bond lengths and lattice parameters can greatly affect the phonon frequencies) and more SDs are required to improve the prediction, and (3) ϵ_{ion} has a much wider range of values than ϵ_{cl} . Nevertheless, these regression models with the aforementioned level of prediction accuracy will be useful for the initial screening of high- k oxides, even when only compositional information is available.

The importance of the feature descriptors was also investigated for ϵ_{ion} . Regarding ML without SD, the standard deviation of the row (SDR) in the periodic table, i.e., the principal quantum numbers of the valence electrons of the constituent elements n_v , was found to be the most significant descriptor for ϵ_{ion} [Fig. 8(a)]. The correlation coefficient between the SDR and ϵ_{ion} is 0.5 [Fig. S8(a) in the Supplemental Material [38]]. Based on Eqs. (4) to (6), low optical phonon frequencies and large BECs increase ϵ_{ion} . The

SDR is determined by the n_v of the cations because the oxide ion is the only anion in our data set of oxides. Because oxygen has $2p$ valence electrons and neither hydrogen nor helium is included in our database, the SDR increases when the n_v of the cations increases. Cations with larger n_v tend to show lower phonon frequencies because of the (1) heavy atomic masses and (2) large atomic radii, which would decrease the Madelung energies and force constants. In fact, as the SDR increases, the lowest frequency of the optical mode tends to decrease ($r = -0.3$) [Fig. S8(c) in the Supplemental Material [38]]. Further, a larger SDR also increases the average of the BEC ($r = 0.5$) [Fig. S8(e) in the Supplemental Material [38]]. The BEC contains contributions from the static charge and transfer of charge induced by atomic displacement [61]. The static charges tend to be larger as the n_v of the cations increases within some groups in the periodic table because

TABLE III. Five oxides with the largest prediction errors in the ML models of the ionic dielectric constants ϵ_{ion} (a) without and (b) with structural descriptors.

Compound	Space group	Number of training data entries with the same structure type	ϵ_{ion} (DFPT)	ϵ_{ion} (ML)	$\log_{10}[\epsilon_{\text{ion}} (\text{ML})] - \log_{10}[\epsilon_{\text{ion}} (\text{DFPT})]$	Percentage error
(a)						
Au_2O_3	<i>Fdd2</i>	1	1.1	4.8	0.63	+323%
$\text{La}_3\text{HfGa}_5\text{O}_{14}$	<i>P321</i>	2	53.4	14.0	-0.58	-74%
SnB_4O_7	<i>Pmn2_1</i>	2	15.7	4.2	-0.57	-73%
Ag_2HgO_2	<i>P4_32_12</i>	1	2.3	5.8	0.40	+151%
Na_2BiO_3	<i>C2/c</i>	7	12.8	5.3	-0.38	-58%
(b)						
Au_2O_3	<i>Fdd2</i>	1	1.1	5.3	0.67	+371%
$\text{La}_3\text{HfGa}_5\text{O}_{14}$	<i>P321</i>	2	53.4	14.1	-0.58	-74%
$\text{Na}_{14}\text{Cd}_2\text{O}_9$	<i>P3</i>	1	4.4	12.5	0.46	+186%
Li_3CuO_3	<i>P4_2/mnm</i>	3	10.6	4.1	-0.41	-61%
SnB_4O_7	<i>Pmn2_1</i>	2	15.7	6.3	-0.40	-60%

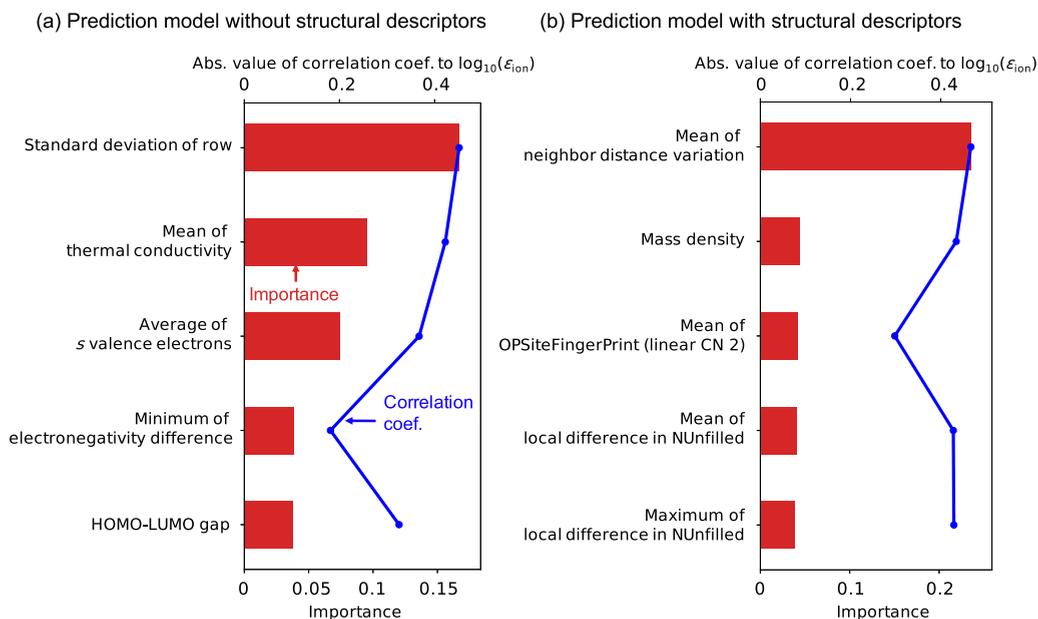


FIG. 8. Importance values of the five most significant descriptors of the prediction models of the ionic dielectric constants ϵ_{ion} (a) without and (b) with structural descriptors. The red bars and blue lines represent the importance values of RF regression and the absolute values of the correlation coefficients, respectively. The details of the average s valence electrons, minimum electronegativity difference, HOMO-lowest unoccupied molecular orbital (LUMO) gap, mean of the neighbor distance variation, mean of OPSiteFingerPrint, mean of the local difference in NUnfilled, and maximum of the local difference in NUnfilled can be found in Refs. [66,67], [67], [65], [68], [79], [68,69], and [68,69], respectively.

of lower electronegativity, leading to large electron transfer to oxide ions. Note that some cations have multiple possible oxidation states, such as Sn^{2+} and Sn^{4+} , and such a variety of oxidation states cannot be described by only n_v . Another possible reason is that the dynamic charge contribution to the BEC is related to the change of orbital hybridization [61], and thus the band gap can also affect the BEC. Larger n_v of cations would decrease the band gap, as mentioned above. In fact, there is a negative correlation between the SDR and the band gap ($r = -0.3$) (Fig. S9 in the Supplemental Material [38]).

The most important descriptor of the ML with SD is the mean of the neighbor distance variation (MNDV) [Fig. 8(b)], which was proposed and used for constructing ML models by Ward *et al.* [68]. The MNDV is determined from Voronoi tessellation by

$$\text{MNDV} = \frac{\sum_{\kappa} \frac{\sum_{\kappa'} w_{\kappa\kappa'} \|r_{\kappa} - \bar{r}_{\kappa}\|}{\bar{r}_{\kappa} \sum_{\kappa'} w_{\kappa\kappa'}}}{N_{\text{atom}}}, \quad (8)$$

where N_{atom} , $w_{\kappa\kappa'}$, $r_{\kappa\kappa'}$, and \bar{r}_{κ} are the number of atoms in the unit cell, area of the facet of the Voronoi polygon, standardized distance between atom κ and neighbor atom κ' , and average of $r_{\kappa\kappa'}$ weighted by $w_{\kappa\kappa'}$, respectively. The MNDV indicates to what extent the atoms are displaced from high-symmetry sites and to what extent the lattice is distorted against high-symmetry structures.

For example, the MNDV of the rocksalt structure is exactly zero [Fig. 9(a)]. Conversely, the MNDV of the cubic perovskite structure is slightly positive (MNDV = 0.073) because the positions of the oxide ions have noncubic site symmetry [Fig. 9(b)]. The MNDV also increases when the lattice is distorted from the cubic structure, e.g., the MNDV

of tetragonal perovskite PbTiO_3 is 0.088. The material in our database that possesses the largest MNDV is tetragonal B_2O_3 [Fig. 9(c)], in which the coordination of atoms is chainlike and Voronoi tessellation assigns very distant atoms in the direction perpendicular to the bonding orientation as neighbor atoms. The MNDV has a negative correlation with ϵ_{ion} ($r = -0.5$) [Fig. S8(b) in the Supplemental Material [38]]. As in the case of the SDR, we estimated the correlation of the MNDV with the lowest phonon frequency and average of the BEC to investigate the origin of the negative correlation. The correlation coefficients of the MNDV with the lowest optical phonon frequency and average of the BEC are -0.2 and -0.4 , respectively [Fig. S8(d) and S8(f) in the Supplemental Material [38]]. As mentioned above, low optical phonon frequencies and large BECs increase ϵ_{ion} . Therefore, the negative correlation between ϵ_{ion} and MNDV originates from the implicit correlation of the MNDV with the BEC rather than the optical phonon frequency. Using the band-by-band decomposition of the BEC, Ghosez *et al.* [80] demonstrated that the displacement and distortion from cubic symmetry drastically decreases the BECs of the Ti and O ions in BaTiO_3 . They also showed that the BECs in BaTiO_3 are sensitive to the degree of anisotropy of the Ti environment, but not to the isotropic volume change [61]. For KNbO_3 , the BECs and ϵ_{ion} are also lower for both the tetragonal and rhombohedral structures than for the cubic structure, as previously reported by Wang *et al.* [81]. As another example, for several polymorphs of perovskite oxides, Kersch and Fischer [82] reported that the dielectric constants from DFPT calculations tend to be large in high-symmetry structures. These previous findings for perovskite oxides appear to be consistent with the correlation between MNDV and ϵ_{ion} found in this study.

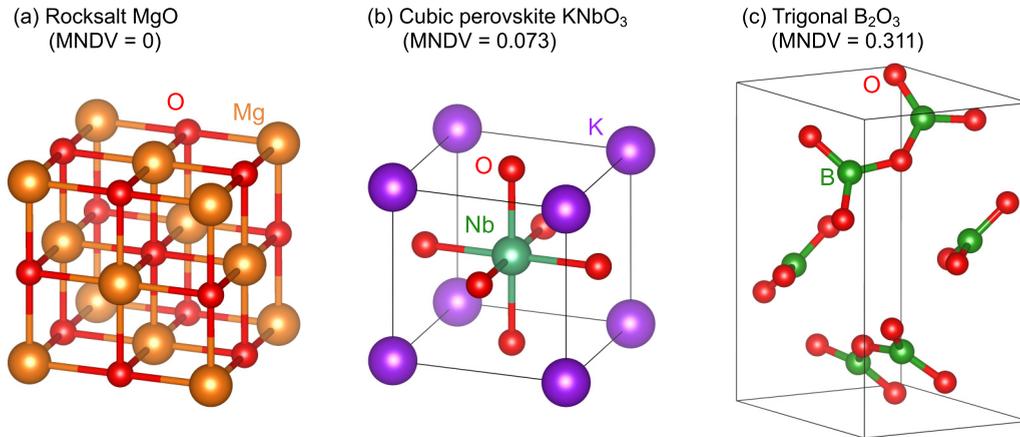


FIG. 9. Crystal structures and MNDVs of (a) MgO, (b) KNbO₃, and (c) B₂O₃.

IV. CONCLUSION

High-throughput first-principles calculations and ML techniques have been used to develop prediction models of the static dielectric constants of metal oxides with the ground-state crystal structures. First, we constructed a computational database by performing DFT and DFPT calculations of 1266 oxides, which covered various chemical compositions and structure types. Because the quality of the database generally affects the reliability of the ML results, we cautiously designed a workflow to evaluate the dielectric constants and excluded unconvincing computational results. We then used the RF regression method and constructed two prediction models for each of the electronic and ionic contributions to the static dielectric constants: one that only requires compositional descriptors and the other that uses both compositional and structural descriptors. The ML models perform well even when structural information is not available, as in the search for novel materials, particularly in prediction of the electronic contributions. Further, we extracted the importance of the descriptors to investigate the determining factors of the dielectric constants. The MAM and mass density were found to be the most important descriptors for the regression models of the electronic contributions without and with structural information, respectively. Their importance can be understood in terms of the correlations with the band gap and ionic polarizability. Regarding the ionic contributions, the structural descriptors more clearly improve the prediction accuracy than the case of the electronic contributions, indicating a stronger relevance to the crystal structure. The most important descriptors were found to be the SDR and MNDV in prediction of the

ionic contributions without and with structural information, respectively. Their correlations with theoretically known key factors for determining the ionic contributions, namely, the optical phonon frequencies and BECs, were discussed. We thus demonstrated that accurate prediction models and the controlling factors of the dielectric constants of oxides can be obtained using ML techniques. We believe that our models and findings will be useful to design and explore future dielectric oxides, as well as to better understand currently used materials.

The computational database and machine learning prediction models used in this study are available at Ref. [83].

ACKNOWLEDGMENTS

This work was supported by KAKENHI Grants No. JP19K15275, No. JP19H02416, No. JP20H00302, and No. JP18J21747 from the Japan Society for the Promotion of Science, PRESTO (Grant No. JPMJPR16N4) and CREST (Grant No. JPMJCR17J2) from the Japan Science and Technology Agency, and the Creation of Life Innovation Materials for Interdisciplinary and International Researcher Development Project of the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computing resources of the Academic Center for Computing and Media Studies at Kyoto University and the Research Institute for Information Technology at Kyushu University were used. The crystal structures in Fig. 9 were visualized using the VESTA code [84].

- [1] J. Robertson, *J. Vac. Sci. Technol.* **31**, 050821 (2013).
 [2] Q. Zhang, M. M. Tavakoli, L. Gu, D. Zhang, L. Tang, Y. Gao, J. Guo, Y. Lin, S.-F. Leung, S. Poddar, Y. Fu, and Z. Fan, *Nat. Commun.* **10**, 727 (2019).
 [3] S. Addanki, I. Amiri, and P. Yupapin, *Results Phys.* **10**, 743 (2018).
 [4] H. K. Raut, V. A. Ganesh, A. S. Nair, and S. Ramakrishna, *Energy Environ. Sci.* **4**, 3779 (2011).
 [5] V. A. Gritsenko, T. V. Perevalov, and D. R. Islamov, *Phys. Rep.* **613**, 1 (2016).
 [6] C. C. Evans, C. Liu, and J. Suntivich, *Opt. Express* **23**, 11160 (2015).
 [7] M. M. Abdullah, F. M. Rajab, and S. M. Al-Abbas, *AIP Adv.* **4**, 027121 (2014).
 [8] Z. Liu, X. Zhang, T. Murakami, and A. Fujishima, *Sol. Energy Mater. Sol. Cells* **92**, 1434 (2008).

- [9] J. Li, Y. Lu, P. Lan, X. Zhang, W. Xu, R. Tan, W. Song, and K.-L. Choy, *Sol. Energy* **89**, 134 (2013).
- [10] X. Gonze, *Phys. Rev. A* **52**, 1086 (1995).
- [11] 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, *APL Mater.* **1**, 011002 (2013).
- [12] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, *Comput. Mater. Sci.* **58**, 227 (2012).
- [13] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, *npj Comput. Mater.* **1**, 15010 (2015).
- [14] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira, and M. Scheffler, *npj Comput. Mater.* **3**, 46 (2017).
- [15] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, *Nat. Mater.* **12**, 191 (2013).
- [16] F. Oba and Y. Kumagai, *Appl. Phys. Express* **11**, 060101 (2018).
- [17] K. Yim, Y. Yong, J. Lee, K. Lee, H.-H. Nahm, J. Yoo, C. Lee, C. Seong Hwang, and S. Han, *NPG Asia Mater.* **7**, e190 (2015).
- [18] M. Lee, Y. Youn, K. Yim, and S. Han, *Sci. Rep.* **8**, 14794 (2018).
- [19] I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T. D. Schladt, K. A. Persson, and F. B. Prinz, *Sci. Data* **4**, 160134 (2017).
- [20] Y. Umeda, H. Hayashi, H. Moriwake, and I. Tanaka, *Jpn. J. Appl. Phys.* **57**, 11UB01 (2018).
- [21] G. Petretto, S. Dwarakanath, H. P. C. Miranda, D. Winston, M. Giantomassi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier, and G.-M. Rignanese, *Sci. Data* **5**, 180065 (2018).
- [22] F. Naccarato, F. Ricci, J. Suntivich, G. Hautier, L. Wirtz, and G.-M. Rignanese, *Phys. Rev. Mater.* **3**, 044602 (2019).
- [23] P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon, and K. Rajan, *Comput. Mater. Sci.* **83**, 185 (2014).
- [24] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, *Sci. Rep.* **3**, 2810 (2013).
- [25] Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, *J. Phys. Chem. Lett.* **9**, 1668 (2018).
- [26] A. Seko, T. Maekawa, K. Tsuda, and I. Tanaka, *Phys. Rev. B* **89**, 054303 (2014).
- [27] V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, *npj Comput. Mater.* **4**, 29 (2018).
- [28] K. Morita, D. W. Davies, K. T. Butler, and A. Walsh, Modelling the dielectric constants of crystals using machine learning, *J. Chem. Phys.* **153**, 024503 (2020).
- [29] Y. Noda, M. Otake, and M. Nakayama, *Sci. Technol. Adv. Mater.* **21**, 92 (2020).
- [30] Y. Umeda, H. Hayashi, H. Moriwake, and I. Tanaka, *Jpn. J. Appl. Phys.* **58**, SLLC01 (2019).
- [31] K. Choudhary, K. F. Garrity, V. Sharma, A. J. Biacchi, A. R. Hight Walker, and F. Tavazza, *npj Comput. Mater.* **6**, 64 (2020).
- [32] P. E. Blöchl, *Phys. Rev. B* **50**, 17953 (1994).
- [33] G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- [34] G. Kresse and D. Joubert, *Phys. Rev. B* **59**, 1758 (1999).
- [35] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, *Phys. Rev. Lett.* **100**, 136406 (2008).
- [36] V. I. Anisimov, J. Zaanen, and O. K. Andersen, *Phys. Rev. B* **44**, 943 (1991).
- [37] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton, *Phys. Rev. B* **57**, 1505 (1998).
- [38] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevMaterials.4.103801> for more information on the additional figures and tables.
- [39] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [40] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, *Comput. Mater. Sci.* **152**, 60 (2018).
- [41] A. Togo and I. Tanaka, SpgLib: A software library for crystal symmetry search, [arXiv:1808.01590](https://arxiv.org/abs/1808.01590).
- [42] Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba, and I. Tanaka, *Comput. Mater. Sci.* **128**, 140 (2017).
- [43] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013).
- [44] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, *Concurr. Comput. Pract. Exp.* **27**, 5037 (2015).
- [45] K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. heng Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S. P. Ong, K. Persson, and A. Jain, *Comput. Mater. Sci.* **139**, 140 (2017).
- [46] M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, and F. Bechstedt, *Phys. Rev. B* **73**, 045112 (2006).
- [47] S. Baroni and R. Resta, *Phys. Rev. B* **33**, 7017 (1986).
- [48] L. Breiman, *Mach. Learn.* **45**, 5 (2001).
- [49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *J. Mach. Learn. Res.* **12**, 2825 (2011).
- [50] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Mach. Learn.* **46**, 389 (2002).
- [51] M. Werner and W. Hans, *Springer Handbook of Condensed Matter and Materials Data* (Springer Science & Business Media, New York, 2006).
- [52] M. Otfried, *Semiconductors: Data Handbook* (Springer, Berlin, 2004).
- [53] Z. Burshtein and Y. Shimony, *Opt. Mater.* **20**, 87 (2002).
- [54] C. Randall T. and L. Antonio C., *Am. Mineral.* **71**, 758 (1986).
- [55] Y. Hinuma, Y. Kumagai, I. Tanaka, and F. Oba, *Phys. Rev. Mater.* **2**, 124603 (2018).
- [56] Y. Hinuma, T. Gake, and F. Oba, *Phys. Rev. Mater.* **3**, 084605 (2019).
- [57] K. A. Müller and H. Burkard, *Phys. Rev. B* **19**, 3593 (1979).
- [58] H. Weaver, *J. Phys. Chem. Solids* **11**, 274 (1959).
- [59] I. Petousis, W. Chen, G. Hautier, T. Graf, T. D. Schladt, K. A. Persson, and F. B. Prinz, *Phys. Rev. B* **93**, 115151 (2016).
- [60] A. Walsh, J. L. F. Da Silva, S.-H. Wei, C. Körber, A. Klein, L. F. J. Piper, A. DeMasi, K. E. Smith, G. Panaccione, P. Torelli,

- D. J. Payne, A. Bourlange, and R. G. Egdel, *Phys. Rev. Lett.* **100**, 167402 (2008).
- [61] P. Ghosez, J.-P. Michenaud, and X. Gonze, *Phys. Rev. B* **58**, 6224 (1998).
- [62] Y. Mochizuki, Y. Kumagai, H. Akamatsu, and F. Oba, *Phys. Rev. Mater.* **2**, 125004 (2018).
- [63] Y. Mochizuki, H.-J. Sung, A. Takahashi, Y. Kumagai, and F. Oba, *Phys. Rev. Mater.* **4**, 044601 (2020).
- [64] S. A. Mabud and A. M. Glazer, *J. Appl. Cryst.* **12**, 49 (1979).
- [65] S. Kotochigova, Z. H. Levine, E. L. Shirley, M. D. Stiles, and C. W. Clark, *Phys. Rev. A* **55**, 191 (1997).
- [66] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, *npj Comput. Mater.* **2**, 16028 (2016).
- [67] A. M. Deml, R. Ohayre, C. Wolverton, and V. Stevanovic, *ChemInform* **47** (2016).
- [68] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, *Phys. Rev. B* **96**, 024104 (2017).
- [69] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, and A. Gamst, *Sci. Rep.* **6**, 34256 (2016).
- [70] P. A. Cox, *The Electronic Structure and Chemistry of Solids* (Oxford University Press, Oxford, 1987).
- [71] D. R. Penn, *Phys. Rev.* **128**, 2093 (1962).
- [72] P. Hervé and L. Vandamme, *Infrared Phys. Technol.* **35**, 609 (1994).
- [73] V. Kumar and J. K. Singh, *Indian J. Pure Appl. Phys.* **48**, 571 (2010).
- [74] P. Atkins and T. Overton, *Shriver and Atkins' Inorganic Chemistry* (Oxford University Press, Oxford, 2010).
- [75] L. Pauling, *Proc. Roy. Soc. A (London)* **114**, 181 (1927).
- [76] S. T. Pantelides, *Phys. Rev. Lett.* **35**, 250 (1975).
- [77] M. M. Braun and L. Pilon, *Thin Solid Films* **496**, 505 (2006).
- [78] O. I. Malyi, M. Boström, V. V. Kulish, P. Thiyam, D. F. Parsons, and C. Persson, *Phys. Chem. Chem. Phys.* **18**, 7483 (2016).
- [79] N. E. R. Zimmermann and A. Jain (unpublished).
- [80] P. Ghosez, X. Gonze, P. Lambin, and J.-P. Michenaud, *Phys. Rev. B* **51**, 6765 (1995).
- [81] C.-Z. Wang, R. Yu, and H. Krakauer, *Phys. Rev. B* **54**, 11161 (1996).
- [82] A. Kersch and D. Fischer, *J. Appl. Phys.* **106**, 014105 (2009).
- [83] https://github.com/takahashi-akira-36m/oxi_diel_db.
- [84] K. Momma and F. Izumi, *J. Appl. Crystallogr.* **44**, 1272 (2011).