Insights into oxygen vacancies from high-throughput first-principles calculations

Yu Kumagai,[1,2,3,*] Naoki Tsunoda,[1] Akira Takahashi,[1] and Fumiyasu Oba[1,2]

[1]*Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama 226-8503, Japan*
[2]*Materials Research Center for Element Strategy, Tokyo Institute of Technology, Yokohama 226-8503, Japan*
[3]*PRESTO, Japan Science and Technology Agency, Tokyo 113-8656, Japan*

(Received 20 September 2021; revised 15 November 2021; accepted 30 November 2021; published 27 December 2021)

Oxygen vacancies play significant roles in various properties of oxide materials. Therefore, insights into the oxygen vacancies can facilitate the discovery of better oxide materials. To achieve this, we developed codes for high-throughput point-defect calculations and applied them to characterize oxygen vacancies in 937 oxides. From the resulting large dataset, we analyzed the vacancy structures and formation energies and constructed machine-learning regression models to predict vacancy formation energies. We have found that the vacancy formation energies are predicted using the random forest regression models with accuracies of 0.27–0.44 eV depending on the charge states. Analyses of the importance of the descriptors show that the formation energies of the neutral vacancies are mainly determined by the orbital characteristics of the conduction-band minima, the oxide stability, and the band gaps, whereas those of the doubly charged defects are determined by factors related to electrostatic energy. These codes and datasets are publicly available, and a graphical user interface is available to analyze the calculation results.

DOI: 10.1103/PhysRevMaterials.5.123803

I. INTRODUCTION: IMPORTANCE OF HIGH-THROUGHPUT POINT DEFECT CALCULATIONS

Oxygen vacancies (V_O) play significant roles in various physical and chemical properties of metal oxides, including electronic conductivity, ion diffusion, superconductivity, catalytic and photocatalytic activity, and optical properties. These properties in turn determine the suitability of metal oxides for specific applications. For example, yttria-stabilized ZrO_2, HfO_2, TiO_2, and Sn-doped In_2O_3 are used as oxide-ion conductors [1], nonvolatile memory materials [2], (photo)catalysts [3], and transparent electronic conductors [4], respectively. Therefore, insights into oxygen vacancies, including their local structures and formation energies, can facilitate the discovery of better oxide materials. However, experimental investigations of V_O are challenging because the electronic and atomic structures of defects are not easy to determine, even today [5]. In contrast, accurate predictions of point-defect properties using first-principles calculations have become plausible with recent improvements in computational power and techniques.

Generally, only one to a few target oxides are involved in a theoretical study on V_O because point-defect calculations require complicated processes [6]. A few exceptions include the study reported by Deml *et al.* in which 45 neutral V_O were calculated in the prototypical binary and ternary oxides [7]. They provided a linear regression equation for the formation energies of neutral V_O ($E_f[V_\mathrm{O}]$) from simple bulk quantities. Another exception is the study reported by Linderälv *et al.*, who performed calculations of neutral and doubly ionized V_O in 26 prototypical binary and ternary oxides [8]. They found that the charge transition levels (CTLs) between the neutral and 2+ charge states are confined to a rather narrow energy range even while band gap and the electronic structure of the conduction band vary substantially.

General conclusions drawn from a limited variety of oxides are potentially inaccurate, and data from calculations of a much wider variety of oxides can be used to validate these previously reported conclusions. Furthermore, unexpected datacentric insights and findings can be gained by inspecting machine-learning (ML) regression models and outliers. Systematic calculations of various material properties, including topological electronic structures [9], elastic properties [10], and phonon properties [11], have been reported. These large datasets are publicly available and used by researchers to screen for superior materials [12] and to apply ML to these properties [13]. However, high-throughput calculations of point defects have not been reported although their importance has been emphasized by researchers [6,12].

In this study, we developed codes for high-throughput point-defect calculations and applied them to V_O in 937 oxides. By considering all the inequivalent O sites and the three vacancy charge states, the total number of calculated oxygen vacancies was up to ~6000, more than 100 times the dataset size from previous calculations [7,8]. Our calculations were mainly performed with the Perdew-Burke-Ernzerhof functional tuned for solids (PBEsol) [14], and underestimation of the band gap was resolved using the

[*] yuuukuma@gmail.com

non-self-consistent approach of a dielectric-dependent (nsc-dd) hybrid functional [15].

After evaluating the calculation accuracy using the results of representative oxides, we statistically analyze the V_O structures, $E_f[V_O]$, and CTL. To analyze the vacancy structures, we developed an atom pairing technique. Finally, we elucidate the origin of $E_f[V_O]$, a long-standing fundamental question in materials science, using ML. These findings will illuminate the way forward in searching for the best oxides for specific applications. All the data are accessible via our GitHub page [16], and the interactive graphic user interface can be retrieved locally.

II. METHODS

A. Target oxides

The targeted oxides were retrieved from the Materials Project database (MPD) [17] using its application programming interface implemented in PYMATGEN [18], at which the formation energies in the MPD were corrected using the empirical scheme to reproduce the experimental enthalpies at 298 K [19]. The selected oxides satisfied the following criteria: (i) they are stable against competing phases, (ii) their band gaps are larger than 0.3 eV in the MPD, (iii) they are nonmagnetic, and (iv) the number of atoms in their primitive cells is 30 or less in the MPD. To avoid complex systems comprising mixed anions and/or cations with partly occupied d and f orbitals, oxides containing H, He, C–Ne, P–Ar, Mn–Ni, Se–Kr, Tc–Rh, Te–Xe, Pr–Lu, Os, Ir, and Po–Lr were excluded in this study. Details of the screened oxides are provided in Supplemental Fig. S1 [20]. The number of target oxides was 1244 at this initial stage.

B. Symmetry identification and symmetrization of structures

We identified the space groups of unit cells and the point groups of oxygen vacancies using SPGLIB [21]. The distance and angle tolerances in Cartesian coordinates were set to 0.1 Å and 5°, respectively. The same parameters were used throughout the study for consistency. SPGLIB was also used to symmetrize the unit cells and vacancy structures.

C. Automation of the point-defect calculations

As will be shown later, the workflow for the point-defect calculations is complicated. Therefore, we developed two open-source codes for their complete automation. The first is VISE [22] in which the input files are created for first-principles calculations using VASP [23] and the output results are analyzed and visualized. The second is PYDEFECT [24] in which point-defect calculations are expedited for nonmetallic solids. PYDEFECT allows us to calculate chemical potential diagrams for the determination of chemical potential ranges in targeted materials, to determine supercell sizes used to model defects, to construct a set of input files required for point-defect calculations, to evaluate correction energies for spurious electrostatic interactions using the extended Freysoldt-Neugebauer–Van de Walle correction scheme [25,26], to analyze electronic and atomic structures for supercells containing defects, and to evaluate defect formation energies depending on the chemical potentials, with the aid of open-source packages, including PYMATGEN [18] and SPGLIB [21]. The point-defect programs such as PYCDT [6], PYLADA-DEFECT [27], and PYDEF [28] have been previously developed. The primary advantages of PYDEFECT compared to these are utilities of the analyses for atomic and electronic structures as shown later. Automation of the workflow is also supported by FIREWORKS [29] and CUSTODIAN [30].

D. Modeling the oxygen vacancies

We constructed the supercells by expanding the conventional unit cells. We first created a set of supercells by incrementing the scales along the lattice vectors. Exceptionally, rotating the supercells by 45° along the c axis was also considered for tetragonal systems. The number of atoms in the supercells was set to 60–500. We defined the supercell anisotropy as $\sum_{i=1,2,3} \frac{|s_i-\bar{s}|}{\bar{s}}$, where s_i and $\bar{s}$ are the lattice constants and their average, respectively, and adopted the supercell with the least anisotropy among the candidates.

To reduce the initial site symmetry of V_O, atoms within 1.3 times the shortest bond length from the initial vacancy site were displaced in the random direction by up to 0.2 Å.

E. Conditions of first-principles calculations

All the calculations were performed using the projector augmented-wave (PAW) method [31,32] implemented in VASP [23] version 6.2.0. The PAW dataset is detailed in Supplemental Table S1 [20]. The cutoff energy was set to 520 eV to optimize the lattice constants and to 400 eV for other calculations. All k-point samplings were centered at the Γ point. The k-point densities for the unit cells and supercells were respectively set to 2.5 and 1.8 Å^{-1}, and fractions of the numbers of k-points were rounded up. Body-centered orthorhombic and tetragonal systems were exceptions; although the reciprocal lattice lengths are different, the numbers of k-points must be the same to maintain lattice symmetry. Therefore, we calculated the geometric mean of the reciprocal lattice constants and adopted the average number of k-points along all the directions. To calculate the dielectric constant, the number of k-points was doubled along all the directions. Band paths in the band-structure calculations were generated using SEEKPATH [33] with a mesh distance of 0.05 Å^{-1}.

To optimize the unit cells under the coarse (tight) condition, the convergence parameter for the decision of the self-consistent field (SCF), namely, the total-energy and eigenvalue change between two electronic steps, was set to 10^{-5} (10^{-8}) eV, and the atomic force criterion for structure optimization was set to 0.2 (0.001) eV/Å. Under the tight condition, support grids for the augmentation charge were added to reduce noise in the forces. The SCF parameter and the atomic force criterion were set to 10^{-5} eV and 0.03 eV/Å for the oxygen vacancy calculations. The SCF parameter was set to 10^{-5} and 10^{-6} eV for the band-structure and dielectric-constant calculations, respectively.

The dielectric constants and phonon band structures were calculated according to density functional perturbation theory (DFPT) [34,35]. PHONOPY [36] was used to plot the phonon band structures. Spin polarization was considered for the V_O calculations; when the absolute magnetization fell below $0.1\mu_B$ during structure optimization, it was switched off from the subsequent structure optimization.

F. Exchange-correlation interactions

Over the past decade, hybrid functionals have been routinely adopted for point-defect calculations because they tend to reproduce experimental band gaps more accurately than local or semilocal functionals for nonmetallic solids [37,38]. The computational cost is, however, too expensive, especially for large-number calculations. In addition, previously reported studies indicate that the estimated formation energies of defects are accurate if the band-edge positions are properly determined using hybrid functionals [39–41]. Therefore, we adopted the PBEsol functional [14] and Hubbard U corrections [42] for Cu and Zn d orbitals and Ce f orbitals with $U_{\mathrm{eff}} = 5$ eV. The band-edge positions were determined using the nsc-dd hybrid-functional calculations; our earlier work demonstrated that these calculations satisfactorily predict band edges in typical semiconductors and insulators, producing results comparable to those of the GW_0 approximation [15].

G. Formation energy of oxygen vacancies

The oxygen vacancy formation energies are evaluated according to

$$E_f[D^q] = \{E[D^q]+E_{\mathrm{corr}}[D^q]\} - E_P + \mu_O + q(\epsilon_{\mathrm{VBM}} + \Delta\epsilon_F),$$

where $E[V_O^q]$ is the total energy of the supercell with an O vacancy in charge state q, E_p is the total energy of the pristine supercell, μ_O is the O chemical potential, and $\Delta\epsilon_F$ is the Fermi level to the valence band maximum (VBM), ϵ_{VBM} [25]. To evaluate the correction term $E_{\mathrm{corr}}[V_O^q]$, the extended Freysoldt-Neugebauer–Van de Walle scheme [25,26] implemented in PYDEFECT was used. The dielectric screening effect was then evaluated using the sum of the electronic ($\varepsilon_{\mathrm{ele}}$) and ionic ($\varepsilon_{\mathrm{ion}}$) contributions of the dielectric constants.

The calculated binding energy of the O_2 molecule using PBEsol (3.3 eV/atom) was greater than that determined from experiments (2.6 eV/atom) [43]. Therefore, the standard state of the O chemical potential was determined from the sum of the calculated O-atom energy and the experimental O_2-molecule binding energy. The relative chemical potentials of the oxides with respect to those of the competing phases were determined from the stabilities available in the MPD [19]. Application of the corrected chemical potentials to the local or semilocal functional calculation results is known to show similar defect formation energies with those calculated by the hybrid functionals with the same corrections applied [39,41].

H. Identification of the perturbed host state

Because a hydrogenlike donor or acceptor state is a perturbed band-edge state distributed over a large area with a shallow donor or acceptor level [44], it is called a perturbed host state (PHS) [45]. For example, in case where V_O is a single shallow donor, V_O^0 is described as V_O^+ plus a hydrogenlike single donor electron positioned slightly below the conduction band minimum (CBM). The calculated formation energy of V_O^0 is almost the same as that of V_O^+ when the Fermi level is located at the CBM; the difference corresponds to the shallow donor level that is generally less than 0.1 eV. If one wants to calculate the shallow donor or acceptor level, one needs to adopt gigantic supercells, e.g., 64 000-atom supercells [44]. Conversely, when adopting supercells composed of several hundred atoms, the defect formation energies are erroneously calculated (see our previous papers [46,47] for details). Furthermore, the PHS perturbs the defect structure only slightly. Therefore, we excluded these defects from the data used for the statistical analyses of $E_f[V_O]$ and defect structures.

Although the PHS can be identified by manually scrutinizing electronic eigenvalues and orbital components in studies on small materials, we needed an algorithm to automate this process because of the large number of calculations. For this purpose, we determined the VBM- and CBM-like edge states in the supercells with V_O by comparing the eigenvalues and element-resolved orbital characteristics near the band edges in the defective supercells with those in the perfect supercell.

We initially describe the procedure used to determine the CBM-like edge state. The single-particle level (ε_e) and wavefunction of the CBM in the perfect supercell should be similar to those at the CBM-like edge state located at the same k-point. Here, we defined the *orbital dissimilarity* as

$$\Delta_m = \sum_{e,i} \left|\phi_{\mathrm{vacancy}}^{m,e,i} - \phi_{\mathrm{perfect}}^{e,i}\right|,$$

where $\phi_{\mathrm{vacancy}}^{(e,i)}$ is the sum of the projections on element e and orbital i ($i = s$, p, d, or f) at the mth band in the supercell with V_O and $\phi_{\mathrm{perfect}}^{(e,i)}$ is that at the CBM in the perfect supercell. We calculated Δ_m for the orbitals at the k-point where the CBM is located and defined the CBM-like edge state as the orbital with the lowest eigenvalue among those satisfying $\varepsilon_e > \varepsilon_{\mathrm{CBM}} - 0.5$ eV and $\Delta_m < 0.4$. Similarly, the VBM-like edge state was defined as the orbital with the highest eigenvalue among those satisfying $\varepsilon_e < \varepsilon_{\mathrm{VBM}} + 0.5$ eV and $\Delta_m < 0.4$, where $\phi_{\mathrm{perfect}}^{(e,i)}$ is the sum of the projections at the VBM in the perfect supercell.

Subsequently, vacancies with CBM-like (VBM-like) edge states occupied by more (less) than 0.2 (0.8) electrons were determined to have PHS. See Supplemental Fig. S5 for some examples [20].

I. Details on machine learning

To perform the ML regression, we need to choose the descriptors. In some cases, the information on crystal structures is directly converted to the descriptors using, e.g., the graph networks or symmetry functions [48,49]. However, these descriptors are generally difficult to use to interpret the physical and/or chemical origins of the target properties. We therefore adopted interpretable physical and chemical information on the oxides and the on-site and neighboring information at the oxygen sites.

Our descriptors are categorized into three types: The first type is the bulk oxide properties, namely, PBEsol(+U) band gap, spherically averaged $\varepsilon_{\mathrm{ele}}$ and $\varepsilon_{\mathrm{ion}}$, formation energies ($E_f$) that are retrieved from the MPD, and ratios of i-orbital ($i = s$, p, d, or f) components at the VBM and CBM. The

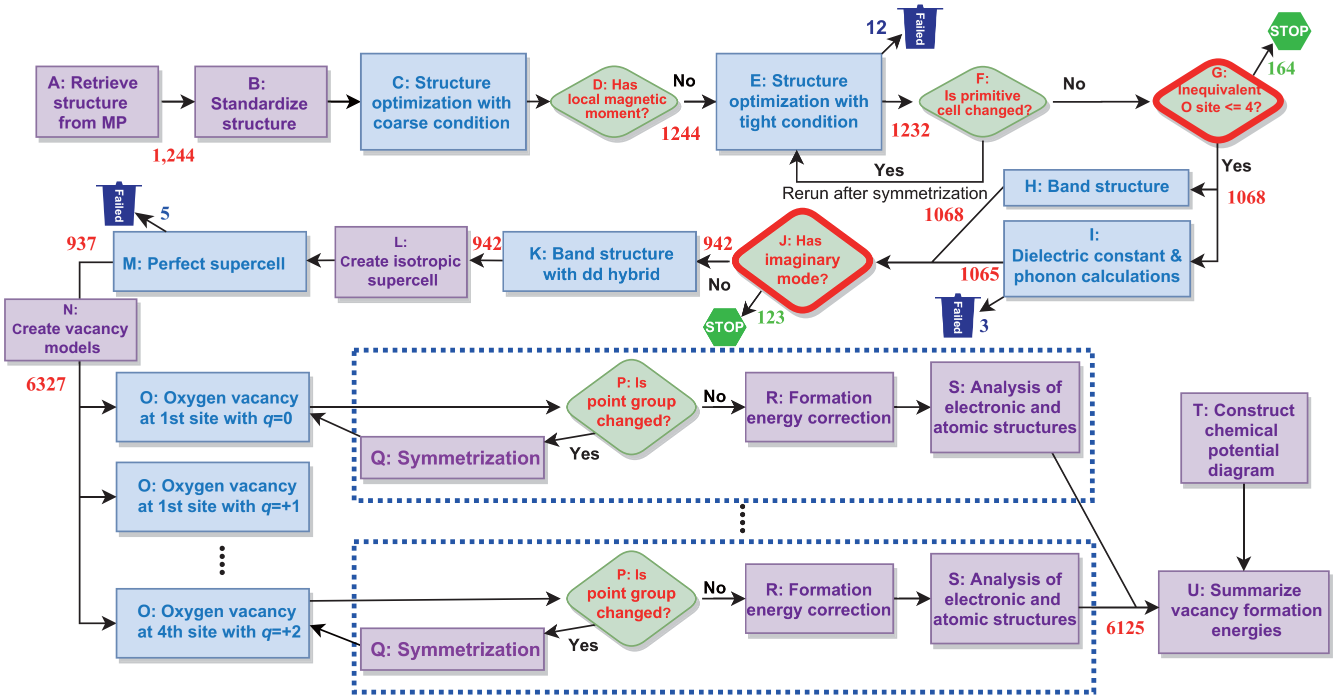
A: Retrieve structure from MP
1,244
B: Standardize structure
C: Structure optimization with coarse condition
D: Has local magnetic moment?
No
1244
E: Structure optimization with tight condition
12
Failed
1232
F: Is primitive cell changed?
No
Yes
Rerun after symmetrization
G: Inequivalent O site <= 4?
STOP
164
Yes
1068
H: Band structure
1068
I: Dielectric constant & phonon calculations
1065
Failed
3
J: Has imaginary mode?
No
STOP
123
942
K: Band structure with dd hybrid
942
L: Create isotropic supercell
M: Perfect supercell
937
Failed
5
N: Create vacancy models
6327
O: Oxygen vacancy at 1st site with q=0
O: Oxygen vacancy at 1st site with q=+1
O: Oxygen vacancy at 4th site with q=+2
P: Is point group changed?
No
Yes
Q: Symmetrization
R: Formation energy correction
S: Analysis of electronic and atomic structures
T: Construct chemical potential diagram
U: Summarize vacancy formation energies
6125

FIG. 1. Workflow for high-throughput point-defect calculations. Steps in the blue boxes and in the purple boxes require first-principles calculations and other complex processes, respectively. The green diamonds indicate decision steps, and those surrounded by bold red lines include a band-gap criterion ($E_g > 0.3$ eV). The number of compounds or vacancies successfully calculated (red), failed (blue), and eliminated (green) by decision steps are marked in different colors. The number of calculations at step O for each oxide is three times the number of inequivalent O sites because three charge states ($q = 0$, +, and 2+) are considered for each oxygen vacancy.

second type is the oxygen on-site information to be removed, namely, the Bader charge and volume [50], spherically averaged Born effective charge (BEC), and the difference between the center of the O-2p density of state and VBM at the O site that is proposed by Deml *et al.* [7]. The Bader charges were estimated using the code developed by Henkelman *et al.* [51]. Then, the atomic boundaries were determined from the all-electron charge density. The third type is the information on neighboring atoms. We defined the weights of surfaces pointing to neighboring X atoms (w_X) as fractions of the solid angles in the Voronoi cells, which were calculated using PYMATGEN [30]. In addition, the average of the Bader charges, BEC, and electronegativities of the neighboring atoms were calculated using w_X; the maximum and minimum values were determined from neighboring atoms with weights larger than 1/12. In total, 70 descriptors were considered in this study.

We used a random forest (RF) regression method [52] implemented in SCIKIT-LEARN [53]. The RF method is one of the ensemble learning techniques, which fits the target properties with a bundle of decision trees. It is known to generally show superior performance due to the nonlinear fitting and robustness against overfitting [52]. The hyperparameters in the RF technique, which controls the accuracy and overfitting, are the number of trees in the forests and the number of features in each forest. The former was fixed to 400, while the latter was determined via grid search in conjunction with fourfold cross validation (CV) and the mean-squared-error regression loss function. Other parameters were set to defaults in the random forest regressor in SCIKIT-LEARN version 0.24.1. The importance was evaluated with the permutation importance [52,54].

III. RESULTS AND DISCUSSION

A. Flow of high-throughput oxygen vacancy calculations

Figure 1 shows the workflow adopted in this study. We first performed structure optimization using the coarse conditions (step C). Because the exchange-correlation functional and Hubbard U parameters are different from those of the MPD, we verified the nonmagnetic states by running the calculations from the ferromagnetic spin configurations. Subsequently, structure optimization was performed with tighter conditions and the symmetries of the relaxed structures were identified (step E); when the symmetry increased, structure optimization was iterated from the symmetrized structure (step F). Consequently, 14 structures were found to show higher symmetries than those in the MPD (Supplemental Table S2 [20]), in which the structures were obtained using the Perdew-Burke-Ernzerhof (PBE) functional [55]. See Ref. [19] for details. We excluded oxides with more than four inequivalent O sites in step G. We also found that our calculated lattice constants were closer to the experimental results than those in the MPD, mainly originating from the use of PBEsol (Supplemental Fig. S3 [20]).

In steps H and I, we concurrently calculated the band structures and dielectric constants ($\varepsilon_{\mathrm{ele}}$ and $\varepsilon_{\mathrm{ion}}$). In the former step, the PBEsol(+U) band gap was determined and wavefunctions were created for the nsc-dd hybrid-functional calculations (step K). The average of $\varepsilon_{\mathrm{ele}}$ was used to determine the exchange interaction mixing parameter in the dd hybrid functional [15], whereas the sum of $\varepsilon_{\mathrm{ele}}$ and $\varepsilon_{\mathrm{ion}}$ was used to estimate the cell-size correction energies of $E_f[V_{\mathrm{O}}]$ [25]. The phonon frequencies at the Γ points, by-products of the $\varepsilon_{\mathrm{ion}}$ calculations, were used to eliminate unstable oxides (step J); because phonon frequencies in the full reciprocal space were not calculated, oxides with unstable modes at other q points were not eliminated at this step.

We then constructed and calculated the supercells (steps L and M); five of these calculations failed because the electronic energies were diverged during the SCF iteration. Finally, calculations for 20 (1.6%) of the 1244 oxides failed before reaching step N, and 287 oxides were excluded because of the criteria. Supercells for V_{O} at all the inequivalent O sites with V_{O} typical charge states, namely, $q = 0$, +, or 2+, were then created at step N. After structure optimization, iterations similar to those in steps E and F were also implemented at steps O, P, and Q. The correction energies and the electronic and atomic structures were automatically analyzed using PYDEFECT at steps R and S.

A total of 6327 calculations were performed to characterize the V_{O} in 937 oxides, of which 122 calculations (1.9%) failed because structure optimization did not converge. Moreover, the formation energies of 21 V_{O} in $K_2Cd_2O_3$, $Ba_2Ti(GeO_4)_2$, and Cs_2O were too low because the structures of the host materials were unstable (Supplemental Fig. S4 [20]). These results were removed from the following analyses but remained in the publicly accessible data for verification purposes.

As written in Sec. II H, we need to exclude the oxygen vacancies with PHS. In this study, the band edges in 59 V_{O} (0.96%) were not determined within our fixed criteria on account of large modulation of the band edges caused by V_{O}. Such issues may be resolved when using larger supercells. Because these V_{O} constituted a small percentage, their results were removed from the data used for the statistical analyses. Finally, 924 V_{O} were identified as defects with the PHS, and 5201 V_{O} remained.

We aggregated the calculated $E_f[V_{\mathrm{O}}]$ at step U to draw the defect formation energy diagrams.

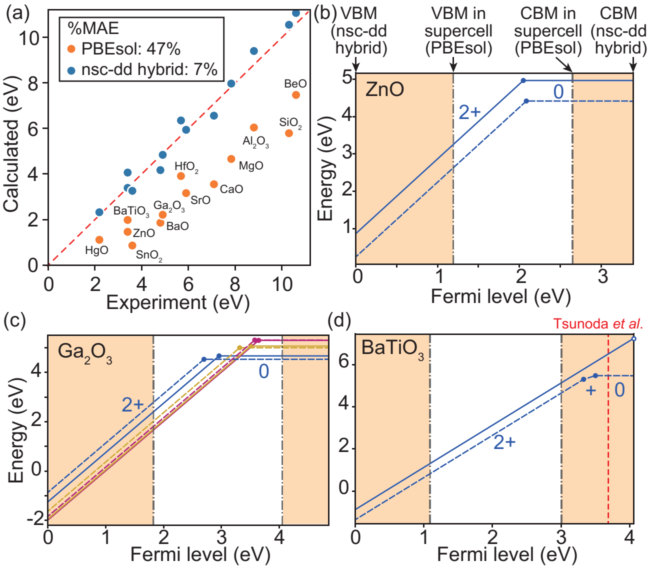

FIG. 2. Accuracy of band gaps and oxygen vacancy formation energies. (a) Band gaps from the PBEsol($+U$) and nsc-dd hybrid-functional calculations along with the experimental values [56–61]. The mean absolute percent errors (%MAE) are also shown. (b)–(d) Calculated formation energies of oxygen vacancies in ZnO, Ga_2O_3, and $BaTiO_3$ under the O-rich condition as a function of the Fermi level compared with the results of previous hybrid-functional calculations [62–64] (dashed lines). The VBM is aligned and set to zero in each oxide. Although our calculated band gaps for ZnO and Ga_2O_3 agree with those in Refs. [62,63] within 0.03 eV, that for $BaTiO_3$ (4.06 eV) is 0.37 eV larger than the gap reported in Ref. [64] (3.69 eV). Therefore, the CBM reported by Tsunoda *et al.* [64] is shown by the red dashed vertical line. The VBM and CBM in the perfect supercells by PBEsol($+U$) are marked with vertical dash-dotted lines, and the energy ranges between those and the band edges determined by the nsc-dd hybrid-functional calculations are colored in orange. The charge states are also described, and the transition levels are designated with solid circles. In (c), three inequivalent O sites are distinguished with different colors. An open circle in (d) indicates the CTL associated with a perturbed host state.

B. Validation of the calculation results for prototypical oxides

The formation energies of charged defects are linearly dependent on the Fermi level, and its controllability is linked to the band gap. Therefore, we compared the band gaps of prototypical oxides obtained from the nsc-dd hybrid-functional calculations with experimental values [56–61]. As shown in Fig. 2(a), most of the band gaps are drastically improved when the nsc-dd hybrid-functional calculations are used. The increases in the band gaps are more related to VBM shifts than to CBM shifts (Supplemental Fig. S6 [20]). Herein, VBM and CBM refer to the nsc-dd hybrid values unless otherwise mentioned.

In Figs. 2(b)–2(d), we show our calculated $E_f[V_{\mathrm{O}}]$ for three prototypical oxides (ZnO, Ga_2O_3, and $BaTiO_3$) under the O-rich condition with previously reported results using hybrid functionals [62–64]. The CTL in ZnO is in good agreement with the previous result despite the large band-edge shift (~2 eV) from the PBEsol+U to the nsc-dd hybrid-functional calculations. $E_f[V_{\mathrm{O}}]$ was, however, constantly overestimated by 0.5 eV, which was partly ascribed to the differences in the O chemical potential. $E_f[V_{\mathrm{O}}]$ for Ga_2O_3, including the positional relationship among the three different O sites, also match the previous results.

However, the CTL located between the band edges of the supercells and those evaluated by the nsc-dd hybrid-functional calculations [orange area in Figs. 2(b)–2(d)] is not described properly. As an example, the calculation results for $BaTiO_3$ are shown in Fig. 2(d). Although $E_f[V_{\mathrm{O}}]$ for $q = 2$ is calculated with adequate accuracy, the CTL found in the previous hybrid-functional calculations [64] is not reproduced and $E_f[V_{\mathrm{O}}]$ for $q = 2$ is linearly extended to the CBM; the highly localized polaronic electrons at $q = 0$ or + reported in Ref. [64] were not stabilized when using the PBEsol functional. Instead, the donor-type PHS appeared in our PBEsol calculations. Indeed, it has been reported that such small polaronic behavior is not reproduced in some titanium oxides such as rutile TiO_2 [65] and $SrTiO_3$ [66] when using the

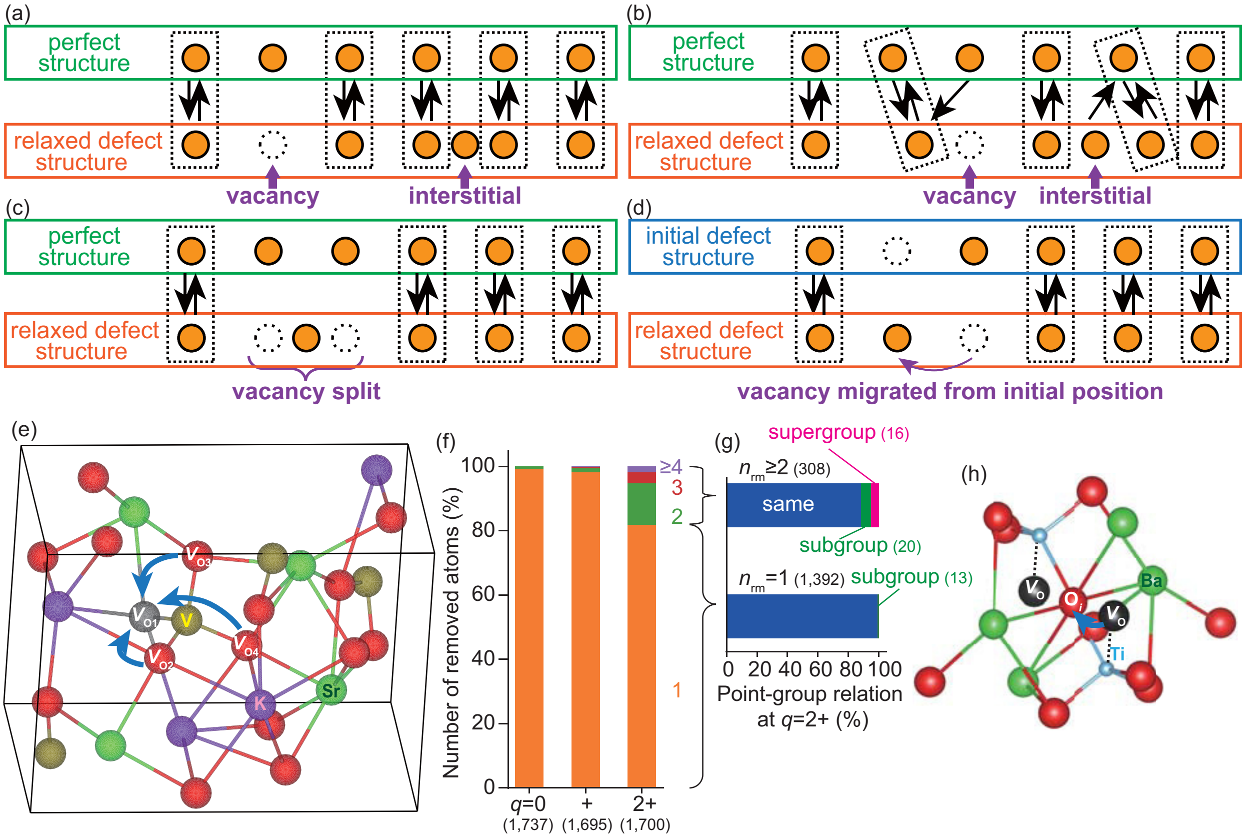

FIG. 3. Analysis of oxygen vacancy structures. (a)–(d) Schematic representation of the atom pairing technique. The orange balls represent the same element in the supercell. One-to-one correspondence of atoms is identified (a)–(c) between the perfect and relaxed defect structures and (d) between the initial and relaxed defect structures. Paired atoms with bidirectional arrows are enclosed in dashed boxes. (e) Crystal structure of $KSrVO_4$. The four inequivalent O sites surrounding a V ion are shown, and vacancy migration from three sites is indicated by the blue arrows. (f) Percentages of the number of removed atoms (n_{rm}) in the oxygen vacancy structure for each q defined by the atom pairing technique with $r_{\mathrm{cutoff}} = 1$ Å (see text for details). The total number of oxygen vacancies at each q is shown in parentheses. (g) Percentages of the point-group relationships in the oxygen vacancy structures compared with the original O site symmetries for $q = 2+$. The bottom bar is for $n_{\mathrm{rm}} = 1$, while the top bar is for $n_{\mathrm{rm}} \geqslant 2$. (h) Local structure of the split-type oxygen vacancy in Ba_2TiO_4. When the left-hand vacancy is introduced, the neighboring right-hand O ion migrates to the high-symmetry site.

semilocal functionals. Addition of the Hubbard U correction to the Ti $3d$ orbitals tends to reproduce the polaronic states, but it may erroneously calculate the ground-state structures; for example, the structure of $BaTiO_3$ is converged to the cubic phase, inconsistent with experiments [64], when relaxed using the PBEsol functional with $U_{\mathrm{eff}} = 3$ eV at the Ti $3d$ orbitals. From this example, we should keep in mind that the CTL may not be properly calculated especially when the CBM is largely underestimated.

C. Oxygen vacancy structures

Historically, nontrivial defects, represented by split interstitials in Si [67] and DX centers in III–V semiconductors [68], have been investigated on the atomic scale using first-principles calculations. High-throughput calculations should play a complementary role. However, identifying nontrivial defects from large datasets requires a method that quantifies the defect structures.

To this end, we developed an atom pairing technique that is composed of the following steps. First, we calculate the distances from each atom in the perfect supercell structure (PSS) to atoms of the same element in the relaxed defect structure (RDS), where the farthest atom from the defect position in the initial defect structure (IDS) is taken as an anchorage point. The atom with the shortest distance in the RDS within a cutoff radius (r_{cutoff}) is then marked as the pointed atom [atom to which a down arrow is pointing in Fig. 3(a)]. Next, we perform the same calculations from each atom in the RDS to atoms in the PSS and find the pointed atom in the reverse direction (atom to which an up arrow is pointing). The atoms bi-pointed to each other are then defined as the host atoms in the RDS, whereas unpaired atoms in the PSS and RDS are defined as removed atoms (vacancies) and inserted atoms (interstitials), respectively [Fig. 3(a)]. Note that such bidirectional pairing is essential to identify pairs because an atom could be pointed to from multiple atoms, as shown in Fig. 3(b).

The nontrivial defects associated with large atomic reconstruction were detected using this technique. For instance, a split-type vacancy shows the removal of two atoms and the insertion of one atom at least, as shown in Fig. 3(c). We also identified the defects that migrated from their initially assigned positions to other sites during structure optimization by performing the same analysis between the IDS and the RDS [Fig. 3(d)].

Specifically, the number of removed and inserted atoms ($n_{\rm rm}$ and $n_{\rm in}$) in the $V_{\rm O}$ supercells satisfies $n_{\rm rm} = n_{\rm in} + 1$, and the number of removed and inserted atoms against the IDS ($n_{\rm rm}^{\rm IDS}$ and $n_{\rm in}^{\rm IDS}$) satisfies $n_{\rm rm}^{\rm IDS} = n_{\rm in}^{\rm IDS}$. Furthermore, when a vacancy migrates to another O site, $n_{\rm rm}= n_{\rm rm}^{\rm IDS}$; otherwise, $n_{\rm rm} = n_{\rm rm}^{\rm IDS} + 1$. Therefore, $n_{\rm in}$, $n_{\rm rm}^{\rm IDS}$, and $n_{\rm in}^{\rm IDS}$ are uniquely derived from $n_{\rm rm}$ if vacancy migrations are properly detected.

In this study, we applied this technique with $r_{\rm cutoff}$ = 1 Å, which is approximately 70% of the O ionic radius (1.35–1.42 Å) [69], and found that 58 $V_{\rm O}$ migrated to other O sites during structure optimization. For example, in $\rm KSrVO_4$ [Fig. 3(e)], three out of the four inequivalent $V_{\rm O}^{2+}$ (the superscript is the charge state q) surrounding a V ion transferred to the O1 site. In contrast to the K and Sr ions, the V ion participates in covalent bonding with the O ions. Therefore, the O ions can easily rotate around the V ion and migrate to the most energetically favorable position. This was confirmed even without local structure perturbation, meaning the energy barriers are intrinsically absent. However, these migrations were absent for $q = 0$ and rare for $q = +$ (5 $V_{\rm O}^{+}$ vs 53 $V_{\rm O}^{2+}$) mainly because the electrons localized at the vacant sites prevent migration; indeed, the defect-induced electrons in the five migrated $V_{\rm O}^{+}$ are located at the d^0 transition-metal ions rather than the vacant sites. Because the migrated $V_{\rm O}$ are duplicates of the other $V_{\rm O}$, we excluded them from the following analyses (see Supplemental Table S3 [20]).

The statistics for $n_{\rm rm}$ as a function of q are shown in Fig. 3(f). Most $V_{\rm O}$ with $q = 0$ and $+$ show trivial vacancy structures ($n_{\rm rm} = 1$), whereas 20% of $V_{\rm O}^{2+}$ show large atom reconfiguration ($n_{\rm rm} \geqslant 2$) because vacancy-induced localized electrons are absent. Figure 3(g) shows the symmetry relation between the $V_{\rm O}^{2+}$ structures and the initial O-site symmetries. For $n_{\rm rm} = 1$, more than 99% of $V_{\rm O}^{2+}$ have the same symmetries with their O sites. In contrast, for $n_{\rm rm} \geqslant 2$, 16 (5.2%) and 20 (6.5%) $V_{\rm O}^{2+}$ show supergroup and subgroup relations, respectively. The site symmetry affects the defect concentration because site degeneracy is proportional to the inverse of the number of symmetry operations. In an extreme case, when the $m\bar{3}m$ site symmetry is reduced to 1, the number of equivalent sites, which linearly correlates to the defect concentration, is increased by 48.

There are a variety of reasons for symmetry lowering; convergence failure during structure optimization is potentially a technical reason if the potential energy surfaces are nearly flat. Therefore, these results need to be handled with care (Supplemental Table S4 [20]). Conversely, because an increase in site symmetry represents a merging of the Wyckoff positions, all the defects with supergroup relations are associated with split-type $V_{\rm O}$, which were found only for $q = 2$. As an example, the calculation results for $\rm Ba_2TiO_4$ [70] are shown in Fig. 3(h). When an O ion was removed from the Wyckoff position 4e, the Ti-O coordination number at the neighboring Ti ions is reduced from 4 to 3, but the neighboring O ion is migrated to the merged Wyckoff position 2a with inversion symmetry during the structure optimization and the original coordination number is recovered.

D. Oxygen vacancy formation energies

Analyzing the large dataset from the calculations could help us better understand the origin of $E_f[V_{\rm O}]$, which is a fundamental question in defect chemistry. Figures 4(a)–4(c) show the distributions of $E_f[V_{\rm O}]$ for each q. For comparison, the O chemical potential is set to the O standard state (see Sec. II G). The energy distribution ranges of $E_f[V_{\rm O}]$, with standard deviation in brackets, were 6.9 (1.41), 5.3 (1.01), and 9.3 (1.53) eV for $q = 0$, $+$, and 2+, respectively. The lowest $E_f[V_{\rm O}]$ were calculated for $\rm Au_2O_3$ ($q = 0$), $\rm MgCu_2O_4$ ($q = +$), and $\rm V_2O_5$ ($q = 2+$), whereas the largest $E_f[V_{\rm O}]$ were determined for $\rm La_2Zr_2O_7$ ($q = 0$), $\rm NaAlO_2$ ($q = +$), and $\rm KBO_2$ ($q = 2+$). Early studies reported band gaps (E_g) and E_f are highly correlated to $E_f[V_{\rm O}]$ [7,71–73]. Although $E_f[V_{\rm O}]$ at $q = 0$ agree with these reported results, it is not true for $E_f[V_{\rm O}]$ at $q = 2+$ (Supplemental Fig. S9 [20]).

To properly understand the origin of $E_f[V_{\rm O}]$ without preconception, we performed ML regressions for $E_f[V_{\rm O}]$ using the random forest model [52] (see Sec. II I for details). For ML, only $E_f[V_{\rm O}]$ for $n_{\rm rm} \leqslant 2$ were included because use of the descriptors on the local environment is inappropriate when large atomic reconfiguration exists. The dataset was split into training and test datasets (see Sec. II I for more details).

In some oxides, the local environments were almost identical between different O sites. When their $E_f[V_{\rm O}]$ are split into the training and test datasets, the prediction accuracy is erroneously enhanced, known as data leakage. Therefore, we divided the dataset by the oxides rather than by the O sites. The total numbers of oxides were 824, 764, and 750 for $q = 0$, $+$, and 2+, respectively; 22, 70, 220, and 700 oxides were selected for the training sets, and 48 oxides for the test sets. Conversely, the statistical errors were evaluated from the errors for the O sites.

Figure 5(a) shows the ML prediction performances for each q; the accuracy is rather high considering the distribution of $E_f[V_{\rm O}]$. The error for $q = +$ is small compared with the others, primarily ascribed to the narrower energy distribution. Following Ref. [74], the data were fitted using $aN^{-0.5} + b$, where N is the size of the training dataset; the mean absolute error (MAE) is lowered to 0.19 ($q = 0$), 0.18 (+), and 0.33 (2+) at $N \to \infty$ when the values at $N = 220$ and 700 are extrapolated. The large error at $q = 2+$ mainly stems from the spurious defect charge interaction errors that remained even after the corrections [25].

We compared our ML model for $q = 0$ with the linear regression model proposed by Deml *et al.* [7]. In their model, the descriptors are E_f, E_g, an average of electronegativity at the nearest-neighboring cations (χ'_n), and the difference between the center of the O-2p density of state and VBM ($\Delta_{\text{O-2}p}$). Note that χ'_n are different from the averaged electronegativity used in our ML model, which includes both cations and anions including oxygen atoms, and the weight of that average depends on the solid angles of the Voronoi cells (see Sec. II I for the definition). For comparison, we determined the linear fitting parameters using the same descriptors (see Supplemental Table S6 [20]). Consequently, the MAE of our ML model (0.32 eV) is less than half the MAE of Deml *et al.*'s model (0.71 eV) (Supplemental Fig. S10 [20]).

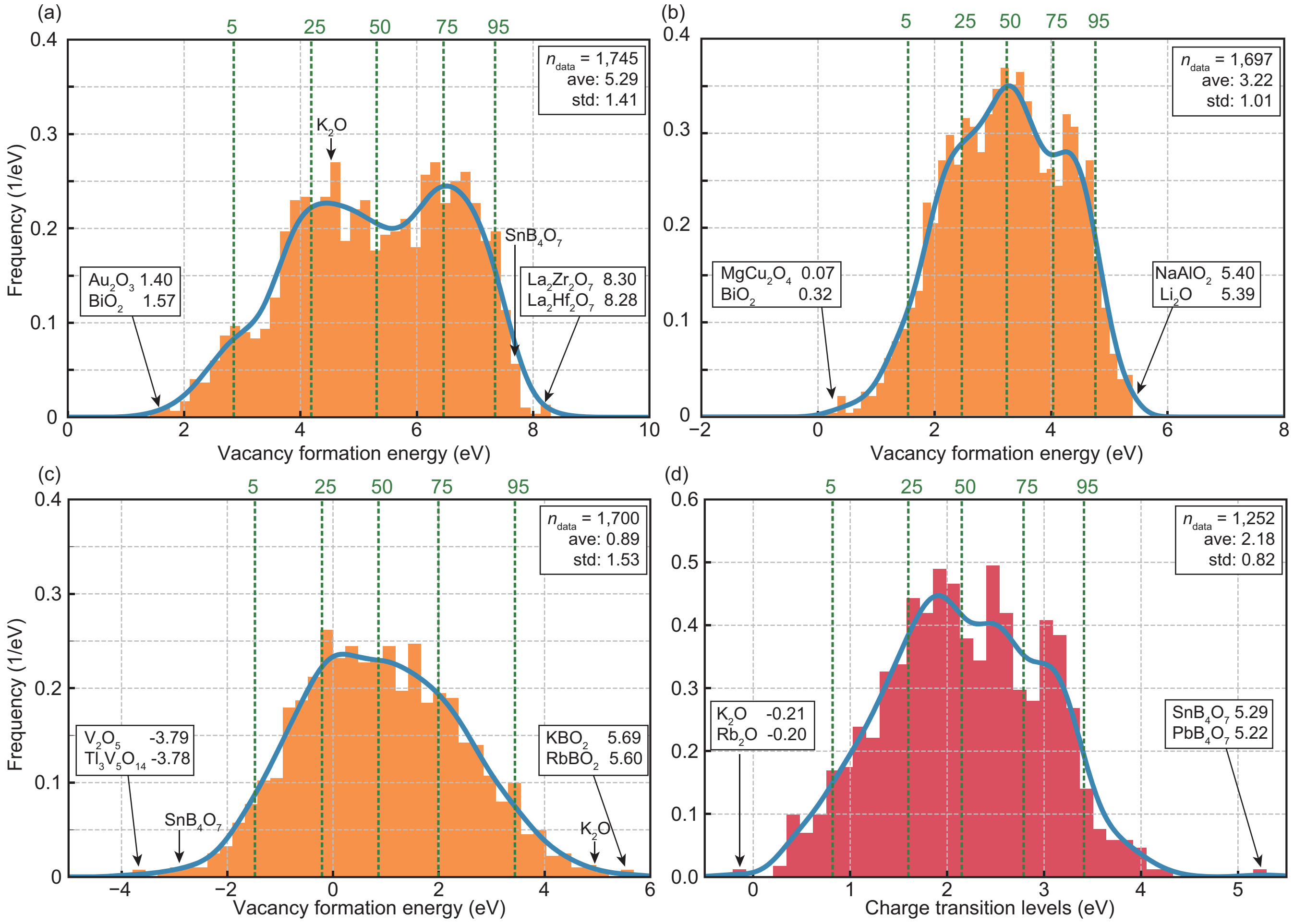

FIG. 4. Statistical analysis of vacancy formation energies. (a)–(c) Distribution of the vacancy formation energies at $q = 0$, +, and 2+. (d) Charge transition levels between $q = 0$ and 2+. The O chemical potential is set to the O standard state. Because the values of $E_f[V_O]$ at $q = +$ and 2+ depend on the Fermi level, we use the O-site local potential as reference for their alignment, and the origin of the Fermi level is set to the VBM in ZnO. The number of data (n_{data}), and the average (ave) and standard deviation (std) of the data are shown in each upper right inset (the latter two are in eV). The oxides with the lowest and highest values are also shown with their values. Five different percentiles are marked with green dashed vertical lines.

The importance of the significant descriptors for $q = 0$ and 2+ is shown in Figs. 5(b) and 5(c). Although E_f, E_g, and the maximum of $\boldsymbol{\chi}_n$ are listed as important for $q = 0$, the p-type orbital component ratio at the CBM, absent in Deml *et al.*'s model, is the most important descriptor. Yin *et al.* reported both the localized defect states and the CTL are qualitatively determined from the orbital characteristics at the CBM [75], which agrees with our results. However, their model has a limitation because the localized states do not always strongly hybridize with the CBM but float at the vacant sites as electrides, as exemplified in the inset of Fig. 5(b).

Notably, the important descriptors differ substantially between $q = 0$ and 2+. As for $q = 2+$, $\varepsilon_{\mathrm{ele}}$ and the Born effective charges are listed as important, in addition to the average of $\boldsymbol{\chi}_n$, mostly related to finite q; $\varepsilon_{\mathrm{ele}}$ determines the screening strength of the electrostatic potential caused by the vacancy charge, whereas the Born effective charges determine the spatial charge distribution [76]. The weight of neighboring surfaces pointing to O atoms (w_O) is also listed probably because positively charged vacancies are preferentially located near the negatively charged O ions (see Supplemental Figs. S12 and S13 for their scatter plots against $E_f[V_O]$ [20]).

The formation energies for $q = 2+$ are found to be mainly dominated by the electrostatics-related quantities. However, it should be emphasized that this result is not self-evident without high-throughput first-principles calculations; the factors of the charged defect formation energies are determined by both the bond strength and electrostatics, and which is dominant depends on the defect species and charge states. Indeed, $E_f[V_O^+]$ are dominated by a mixture of both as shown in Supplemental Fig. S10 [20].

The distribution of the CTLs between $q = 0$ and 2+ is shown in Fig. 4(d). Linderälv *et al.* [8] have reported that the standard deviation of the CTLs averaged over 22 oxides is 0.66 eV, which is slightly smaller than our standard deviation (0.82 eV). However, unlike the conclusion that the CTLs are confined to a rather narrow energy range, we emphasize that there are a certain number of exceptions, which could be attributed to the difference of the important descriptors between the neutral and doubly charged defects. Oxides showing the lowest and highest CTL are K_2O and SnB_4O_7, respectively.

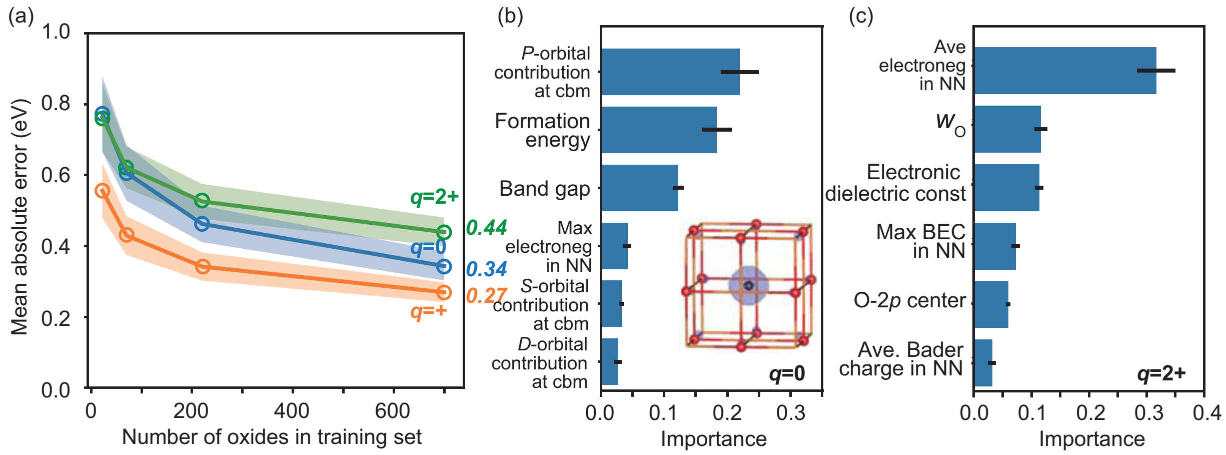

FIG. 5. Machine learning of oxygen vacancy formation energies. (a) RF model convergence against the number of oxides in the training set. The model creation process was repeated 100 times with a different training set. Mean absolute errors and standard deviations are estimated from the test sets, which are represented by the circles and shaded areas, respectively. (b), (c) Average of the importance of the significant descriptors in the RF models for (b) $q = 0$ and (c) 2+. The error bars represent the standard deviations. w_O is the weight of neighboring surfaces pointing to O atoms in Voronoi polyhedra. The inset in (b) shows the spatial distribution of the localized vacancy state of V_O^0 in MgO.

$E_f[V_O^0]$ for K_2O is located in the middle of the distribution, whereas V_O^{2+} is quite high in energy [Figs. 4(a) and 4(c)], which is primarily ascribed to the small electronegativity of K (Supplemental Fig. S12(f) [20]). Conversely, SnB_4O_7 shows a small $E_f[V_O^{2+}]$ and large $E_f[V_O^0]$, which would be affected by the large electronegativity of B (Supplemental Fig. S12(f) [20]) and the large band gap (Supplemental Fig. S9(a) [20]), respectively. The effective correlation energy U for electrons trapped at the vacancy sites is also discussed in Supplemental Fig. S15 [20].

IV. SUMMARY

We conducted high-throughput calculations of the oxygen vacancies in 937 oxides with band gaps. For this purpose, we initially developed programs to automate the calculations and analyze the atomic and electronic structures of point defects based on the programmable algorithms.

After showing the details of the calculation flow, we checked the accuracies of the band gaps and the vacancy formation energies in comparison with experiments and previously reported hybrid-functional calculations, respectively. It was found that the band gaps are predicted with an accuracy of 7% MAE and our vacancy formation energies are similar to those calculated by the hybrid functionals, including the positional relationship among the different O sites. However, it should be noted that the transition levels associated with the small polarons may not be properly calculated using PBEsol(+U).

We next proposed the atom mapping technique to identify the oxygen vacancy structures. With this technique, we found the doubly charged vacancies tend to show large atomic reconstructions such as the split-type vacancies or those migrated to another site probably because they do not have the localized electrons at the vacant sites.

Finally, we showed the distribution of the oxygen vacancy formation energies and the results of the RF regressions. The vacancy formation energies are predicted with accuracies of 0.27–0.44 eV depending on the charge state. Based on the importance of the descriptors, we have found the formation energies of the neutral vacancies are mainly determined by the orbital characteristics of the conduction band minima, the oxide stability, and the band gaps, whereas those of the doubly charged defects are determined by the factors related to the electrostatic energy. Note that these are not self-evident without high-throughput first-principles calculations because the formation energies of the charged defects are determined by a balance of the bond strength and electrostatics.

Although we mainly discussed the tendency and outliers of V_O, our large dataset should be useful for finding oxide materials suitable for specific applications. For instance, oxides in which only neutral V_O are stable at the Fermi level in the band gaps are good semiconductors because neither Fermi level pinning nor carrier trapping is caused by the oxygen vacancies within these oxides.

Furthermore, the codes and techniques developed in this research are easily adapted to other types of defects and other nonmetallic materials. Indeed, the oxygen vacancies are not only relevant defects in oxides, but cation vacancies, interstitials, and impurities such as ubiquitous hydrogens may also play vital roles in oxides. We hope the results presented herein will facilitate high-throughput point-defect calculations in the community and promote the discovery of new superior materials where point defects play significant roles, such as semiconductors, catalysts, and ion batteries.

ACKNOWLEDGMENTS

Funding for this study was provided by the MEXT Element Strategy Initiative to Form Core Research Center (Grant No. JPMXP0112101001), PRESTO (JPMJPR16N4) from the Japan Science and Technology Agency, KAKENHI Grant No. 19H02416, “STAR” and Challenging Research Award from the Tokyo Institute of Technology, and Design & Engineering by Joint Inverse Innovation for Materials Architecture, MEXT.

APPENDIX: DATA VISUALIZATION

We have developed the graphical user interface to visualize the calculated results with the aid of Crystal Toolkit [77]. We show an example of the graphical user interface distributed in the GitHub page [78] (see Supplemental Figs. S16 and S17 [20]). Since the data would be updated continuously, please refer the web page for details.

[1] B. Steele, Oxygen ion conductors and their technological applications, Mater. Sci. Eng. B **13**, 79 (1992).

[2] M. H. Park, Y. H. Lee, T. Mikolajick, U. Schroeder, and C. S. Hwang, Review and perspective on ferroelectric HfO_2-based thin films for memory applications, MRS Commun. **8**, 795 (2018).

[3] X. Pan, M.-Q. Yang, X. Fu, N. Zhang, and Y.-J. Xu, Defective TiO_2 with oxygen vacancies: Synthesis, properties and photocatalytic applications, Nanoscale **5**, 3601 (2013).

[4] R. A. Afre, N. Sharma, M. Sharon, and M. Sharon, Transparent conducting oxide films for various applications: A review, Rev. Adv. Mater. Sci. **53**, 79 (2018).

[5] F. Gunkel, D. V. Christensen, Y. Z. Chen, and N. Pryds, Oxygen vacancies: The (in)visible friend of oxide electronics, Appl. Phys. Lett. **116**, 120505 (2020).

[6] D. Broberg, B. Medasani, N. E. Zimmermann, G. Yu, A. Canning, M. Haranczyk, M. Asta, and G. Hautier, Pycdt: A python toolkit for modeling point defects in semiconductors and insulators, Comput. Phys. Commun. **226**, 165 (2018).

[7] A. M. Deml, A. M. Holder, R. P. O'Hayre, C. B. Musgrave, and V. Stevanović, Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides, J. Phys. Chem. Lett. **6**, 1948 (2015).

[8] C. Linderalv, A. Lindman, and P. Erhart, A unifying perspective on oxygen vacancies in wide band gap oxides, J. Phys. Chem. Lett. **9**, 222 (2017).

[9] M. G. Vergniory, L. Elcoro, C. Felser, N. Regnault, B. A. Bernevig, and Z. Wang, A complete catalogue of high-quality topological materials, Nature (London) **566**, 480 (2019).

[10] M. d. Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. K. Ande, S. v. d. Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, and M. Asta, Charting the complete elastic properties of inorganic crystalline compounds, Sci. Data **2**, 150009 (2015).

[11] G. Petretto, S. Dwaraknath, H. P. Miranda, D. Winston, M. Giantomassi, M. J. v. Setten, X. Gonze, K. A. Persson, G. Hautier, and G.-M. Rignanese, High-throughput density-functional perturbation theory phonons for inorganic materials, Sci. Data **5**, 180065 (2018).

[12] G. Hautier, A. Miglio, G. Ceder, G.-M. Rignanese, and X. Gonze, Identification and design principles of low hole effective mass *p*-type transparent conducting oxides, Nat. Commun. **4**, 2292 (2013).

[13] C. J. Bartel, S. L. Millican, A. M. Deml, J. R. Rumptz, W. Tumas, A. W. Weimer, S. Lany, V. Stevanović, C. B. Musgrave, and A. M. Holder, Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry, Nat. Commun. **9**, 4168 (2018).

[14] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces, Phys. Rev. Lett. **100**, 136406 (2008).

[15] Y. Hinuma, Y. Kumagai, I. Tanaka, and F. Oba, Band alignment of semiconductors and insulators using dielectric-dependent hybrid functionals: Toward high-throughput evaluation, Phys. Rev. B **95**, 075302 (2017).

[16] Y. Kumagai, Oxygen vacancy database (2021), https://github.com/kumagai-group/oxygen_vacancies_db.

[17] 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).

[18] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, Comput. Mater. Sci. **97**, 209 (2015).

[19] A. Wang, R. Kingsbury, M. McDermott, M. Horton, A. Jain, S. P. Ong, S. Dwaraknath, and K. A. Persson, A framework for quantifying uncertainty in DFT energy corrections, Sci. Rep. **11**, 15496 (2021).

[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.5.123803 for the statistical information for our high-throughput calculations and technical examples to supplement our newly developed techniques used in this study.

[21] A. Togo and I. Tanaka, Spglib: A software library for crystal symmetry search, arXiv:1808.01590.

[22] Y. Kumagai, VISE (2021), https://github.com/kumagai-group/vise.

[23] G. Kresse and J. Furthmuller, Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set, Phys. Rev. B **54**, 11169 (1996).

[24] Y. Kumagai, PYDEFECT (2021), https://github.com/kumagai-group/pydefect.

[25] Y. Kumagai and F. Oba, Electrostatics-based finite-size corrections for first-principles point defect calculations, Phys. Rev. B **89**, 195205 (2014).

[26] C. Freysoldt, J. Neugebauer, and C. G. Van de Walle, Fully *Ab Initio* Finite-Size Corrections for Charged-Defect Supercell Calculations, Phys. Rev. Lett. **102**, 016402 (2009).

[27] A. Goyal, P. Gorai, H. Peng, S. Lany, and V. Stevanović, A computational framework for automation of point defect calculations, Comput. Mater. Sci. **130**, 1 (2017).

[28] A. Stoliaroff, S. Jobic, and C. Latouche, PyDEF 2.0: An easy to use post-treatment software for publishable charts featuring a graphical user interface, J. Comput. Chem. **39**, 2251 (2018).

[29] 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, FireWorks: A dynamic workflow system designed for high-throughput applications, Concurrency Comput.: Pract. Exper. **27**, 5037 (2015).

[30] 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, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. **68**, 314 (2013).

[31] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B **59**, 1758 (1999).

[32] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B **50**, 17953 (1994).

[33] Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba, and I. Tanaka, Band structure diagram paths based on crystallography, Comput. Mater. Sci. **128**, 140 (2017).

[34] S. Baroni and R. Resta, *Ab initio* calculation of the macroscopic dielectric constant in silicon, Phys. Rev. B **33**, 7017 (1986).

[35] X. Gonze and C. Lee, Dynamical matrices, Born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory, Phys. Rev. B **55**, 10355 (1997).

[36] A. Togo and I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. **108**, 1 (2015).

[37] F. Oba and Y. Kumagai, Design and exploration of semiconductors from first principles: A review of recent advances, Appl. Phys. Express **11**, 060101 (2018).

[38] C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, First-principles calculations for point defects in solids, Rev. Mod. Phys. **86**, 253 (2014).

[39] C. Freysoldt, B. Lange, J. Neugebauer, Q. Yan, J. L. Lyons, A. Janotti, and C. G. Van de Walle, Electron and chemical reservoir corrections for point-defect formation energies, Phys. Rev. B **93**, 165206 (2016).

[40] A. Alkauskas, P. Broqvist, and A. Pasquarello, Defect Energy Levels in Density Functional Calculations: Alignment and Band Gap Problem, Phys. Rev. Lett. **101**, 046405 (2008).

[41] H. Peng, D. O. Scanlon, V. Stevanović, J. Vidal, G. W. Watson, and S. Lany, Convergence of density and hybrid functional defect calculations for compound semiconductors, Phys. Rev. B **88**, 115201 (2013).

[42] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton, Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDAU study, Phys. Rev. B **57**, 1505 (1998).

[43] L. Wang, T. Maxisch, and G. Ceder, Oxidation energies of transition metal oxides within the GGA+U framework, Phys. Rev. B **73**, 195107 (2006).

[44] G. Zhang, A. Canning, N. Gronbech-Jensen, S. Derenzo, and L.-W. Wang, Shallow Impurity Level Calculations in Semiconductors Using Ab Initio Methods, Phys. Rev. Lett. **110**, 166404 (2013).

[45] Y. Zhang, A. Mascarenhas, and L.-W. Wang, Systematic approach to distinguishing a perturbed host state from an impurity state in a supercell calculation for a doped semiconductor: Using GaP:N as an example, Phys. Rev. B **74**, 041201(R) (2006).

[46] N. Tsunoda, Y. Kumagai, A. Takahashi, and F. Oba, Electrically Benign Defect Behavior in Zinc Tin Nitride Revealed from First Principles, Phys. Rev. Appl. **10**, 011001(R) (2018).

[47] Y. Kumagai, M. Choi, Y. Nose, and F. Oba, First-principles study of point defects in chalcopyrite $ZnSnP_2$, Phys. Rev. B **90**, 125202 (2014).

[48] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. **31**, 3564 (2019).

[49] J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, J. Chem. Phys. **134**, 074106 (2011).

[50] R. F. W. Bader, *Atoms in Molecules: A Quantum Theory* (Oxford University Press, New York, 1990).

[51] W. Tang, E. Sanville, and G. Henkelman, A grid-based Bader analysis algorithm without lattice bias, J. Phys.: Condens. Matter **21**, 084204 (2009).

[52] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Muller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. **12**, 2825 (2011).

[54] L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author), Stat. Sci. **16**, 199 (2001).

[55] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. **77**, 3865 (1996).

[56] Y. Hinuma, T. Gake, and F. Oba, Band alignment at surfaces and heterointerfaces of Al_2O_3, Ga_2O_3, In_2O_3, and related group-III oxide polymorphs: A first-principles study, Phys. Rev. Mater. **3**, 084605 (2019).

[57] Y. Hinuma, Y. Kumagai, I. Tanaka, and F. Oba, Effects of composition, crystal structure, and surface orientation on band alignment of divalent metal oxides: A first-principles study, Phys. Rev. Mater. **2**, 124603 (2018).

[58] M. Batzill and U. Diebold, The surface and materials science of tin oxide, Prog. Surf. Sci. **79**, 47 (2005).

[59] S. Maj, Energy gap and density in SiO_2 polymorphs, Phys. Chem. Miner. **15**, 271 (1988).

[60] M. Balog, M. Schieber, M. Michman, and S. Patai, The chemical vapour deposition and characterization of ZrO_2 films from organometallic compounds, Thin Solid Films **47**, 109 (1977).

[61] S. H. Wemple, Polarization fluctuations and the optical-absorption edge in $BaTiO_3$, Phys. Rev. B **2**, 2679 (1970).

[62] Y. K. Frodason, K. M. Johansen, T. S. Bjørheim, B. G. Svensson, and A. Alkauskas, Zn vacancy as a polaronic hole trap in ZnO, Phys. Rev. B **95**, 094105 (2017).

[63] J. B. Varley, J. R. Weber, A. Janotti, and C. G. Van de Walle, Oxygen vacancies and donor impurities in Î²β-Ga_2O_3, Appl. Phys. Lett. **97**, 142106 (2010).

[64] N. Tsunoda, Y. Kumagai, and F. Oba, Stabilization of small polarons in $BaTiO_3$ by local distortions, Phys. Rev. Mater. **3**, 114602 (2019).

[65] A. Janotti, C. Franchini, J. B. Varley, G. Kresse, and C. G. V. d. Walle, Dual behavior of excess electrons in rutile TiO_2, Phys. Status Solidi **7**, 199 (2013).

[66] M. Choi, F. Oba, Y. Kumagai, and I. Tanaka, Antiferrodistortive-like oxygen-octahedron rotation induced by the oxygen vacancy in cubic $SrTiO_3$, Adv. Mater. **25**, 86 (2013).

[67] R. J. Needs, First-principles calculations of self-interstitial defect structures and diffusion paths in silicon, J. Phys.: Condens. Matter **11**, 10437 (1999).

[68] D. J. Chadi and K. J. Chang, Theory of the Atomic and Electronic Structure of DX Centers in GaAs and $Al_xGa_{1-x}As$ Alloys, Phys. Rev. Lett. **61**, 873 (1988).

[69] R. D. Shannon, Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides, Acta Crystallogr. Sect. A **32**, 751 (1976).

[70] J. A. Bland, The crystal structure of barium orthotitanate, Ba_2TiO_4, Acta Crystallogr. **14**, 875 (1961).

[71] I. Tanaka, F. Oba, K. Tatsumi, M. Kunisu, M. Nakano, and H. Adachi, Theoretical formation energy of oxygen-vacancies in oxides, Mater. Trans. **43**, 1426 (2002).

[72] T. Yamamoto and T. Mizoguchi, First principles study on oxygen vacancy formation in rock salt-type oxides MO (M: Mg, Ca, Sr and Ba), Ceram. Int. **39**, S287 (2013).

[73] A. Murat and J. E. Medvedeva, Composition-dependent oxygen vacancy formation in multicomponent wide-band-gap oxides, Phys. Rev. B **86**, 085123 (2012).

[74] M. Rupp, A. Tkatchenko, K.-R. Muller, and O. A. von Lilienfeld, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett. **108**, 058301 (2012).

[75] W.-J. Yin, S.-H. Wei, M. M. Al-Jassim, and Y. Yan, Prediction of the chemical trends of oxygen vacancy levels in binary metal oxides, Appl. Phys. Lett. **99**, 142109 (2011).

[76] N. Tsunoda, Y. Kumagai, M. Araki, and F. Oba, One-dimensionally extended oxygen vacancy states in perovskite oxides, Phys. Rev. B **99**, 060103(R) (2019).

[77] M. Horton, Crystal Toolkit (2021), https://docs.crystaltoolkit.org, version 1.6.0.

[78] https://github.com/kumagai-group/oxygen_vacancies_db.