Machine-Learning Prediction of Charged-Defect Formation Energies from Crystal Structures

Shin Kiyohara[1], Chisa Shibui[1], Soungmin Bae,[1] and Yu Kumagai[1,2,*]

[1]*Institute for Materials Research, Tohoku University, Sendai, Japan*
[2]*Organization for Advanced Studies, Tohoku University, Sendai, Japan*

Recent advances in materials informatics have expanded the number of synthesizable materials. However, screening promising candidates, such as semiconductors, based on defect properties remains challenging. This is primarily due to the lack of a general framework for predicting defect formation energies in multiple charge states from structural data. In this Letter, we present a protocol, namely data normalization, Fermi level alignment, and treatment of perturbed host states, and validate it by accurately predicting oxygen vacancy formation energies in three charge states using a single model. We also introduce a joint machine-learning model that integrates defect formation energies and band-edge predictions for virtual screening. Using this framework, we identify 89 hole-dopable oxides, including BaGaSbO, a potential ambipolar photovoltaic material. Our protocol is expected to become a standard approach for machine-learning studies on point defect formation energies.

Historically, superior materials that have contributed to human society's development have been discovered through experimental efforts. However, the cost and time required for such experiments pose significant constraints on material innovation. To mitigate these challenges, first-principles calculations have been adopted in materials science over the past few decades, especially thanks to the exponential growth in computational power [1]. Nowadays, high-throughput computational techniques have facilitated the development of extensive materials databases. They are currently employed in building machine learning (ML) surrogate models, which are then used to screen promising materials from millions of candidates [2].

Despite these advances, the number of theoretically identified materials that achieve commercial viability remains limited. One fundamental reason is that theoretical calculations usually assume defect-free materials, which can diverge from practical conditions where defects often play a crucial role, particularly in semiconductors. For instance, desirable photoabsorbers for solar cells should be free from point defects with deep levels [3]. In addition, the dopability may be limited by the compensation of native defects [4]. Thus, integrating defect-property-based screening into computational materials discovery is expected to

improve the chances of identifying commercially viable semiconductors.

To this end, Deml *et al.* constructed a linear regression model for the formation energies of neutral oxygen vacancies (V_O) in 45 binary and ternary oxides [5]. We also systematically calculated V_O not only in neutral charge state but in $+1$, and $+2$ charge states in ~ 1000 oxides that are composed of diverse crystal structures and cationic elements [6] and applied random forest (RF) regression to predict the V_O formation energies, hereafter referred to as $E_f[V_O]$. Subsequently, Witman *et al.* also calculated the neutral V_O in approximately 200 oxides and performed ML using graph-based neural network (GNN) models to predict their formation energies [7]. Frey *et al.* also calculated neutral defects in 158 two-dimensional materials and constructed a GNN model that predicts formation energies of defects [8]. While other research groups have also developed ML models for predicting defect formation energies, the atomic frameworks were fixed to specific types, such as perovskite [9] and zinc-blende structures [10,11].

Virtual screening requires ML models that predict defect formation energies in *various charge states* using *structural information alone*. In addition, when screening superior materials from a myriad of candidates, the models must accommodate *diverse crystal structures*. Furthermore, a single ML model should be able to predict multiple defect-charge combinations, analogous to a general-purpose ML potential. However, prior regression models fall short in these criteria. For example, our RF models require descriptors obtained from first-principles calculations for unit cells, such as dielectric constants and O-$2p$

---

center positions [6]. Also, the models developed by several groups focus only on the neutral defects [5,7,8].

Indeed, a general ML framework for defect formation energies at various q values has not yet been established. Our purpose is to present one here using $E_f[V_O]$ as a case study. The key aspects include pruning inadequate data, normalization of the defect formation energies with different q values, and determination of the Fermi level (ϵ_F), which linearly shifts formation energies of charged defects. In addition, we discuss how to manage the defects with perturbed host states (PHS). We confirmed that all these factors improve the prediction of $E_f[V_O]$ by crystal graph convolution neural network models (CGCNN) [12]. Using a single ML model and only structural information, we achieved improved accuracies of 0.29, 0.22, and 0.37 eV for the neutral, $+1$, and $+2$ charge states, respectively, outperforming our previous models. We further introduce an ML model that predicts band edge position and apply a joint model combining defect and valence-band maximum (VBM) ML models to virtually screen particularly rare hole-dopable oxides among recently proposed stable structures [13–15]. Consequently, we identified 89 hole-dopable oxides, including BaGaSbO as a potential ambipolar photovoltaic material.

Charged defect formation energies—Since defect formation energies depend on ϵ_F when $q \neq 0$, we must align ϵ_F across compounds. The simplest choice is to set $\epsilon_F = \epsilon_{VBM}$ of each compound. However, this is inadequate because ϵ_{VBM} varies significantly across materials. For example, oxides with valence bands composed of lone-pair orbitals or transition-metal d orbitals generally have higher VBMs than those dominated by O-$2p$ orbitals [6]. Therefore, defect formation energies at $\epsilon_F = \epsilon_{VBM}$ reflect both information on the defect formation energy and ϵ_{VBM}, complicating prediction using a *single* GNN. Instead, we propose using core potentials as references. When a homogeneous charge is placed in the core potential region, the electrostatic energy remains constant under this alignment. Thus, aligning ϵ_F using core potentials ensures consistent electrostatic contributions to the defect formation energies (Fig. S1 in the Supplemental Material [16]).

However, arbitrariness remains in determining ϵ_F. During training, standardizing target values is known to improve weight updates [17]. A difficulty arises when training on formation energies of defects with mixed charge states. Figure 1(a) shows the distributions of $E_f[V_O]$ at $q = 0$, $+1$, and $+2$, where ϵ_F of all the oxides are set to the VBM of ZnO via oxygen core potential alignment; ZnO was selected as the reference example, while noting that the choice of the Fermi level is, in principle, arbitrary. The energy distributions for different q depend on ϵ_F; for example, as ϵ_F decreases from the VBM of ZnO, each energy distribution becomes more separated. To train the GNN effectively, the distributions should overlap as much
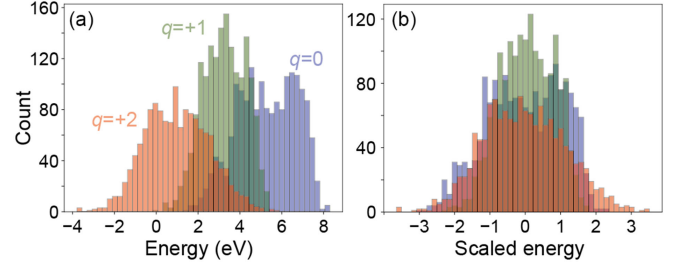

FIG. 1. (a) Distributions of oxygen vacancy formation energies at $q = 0$, $+1$, and $+2$ with ϵ_F aligned to the VBM of ZnO via core potentials. Data for the vacancies with perturbed host states are already removed. (b) Normalized distributions with minimized mean differences between charge states (see text for details).

as possible. We define the difference in means (Δ) across charge states as

$$\Delta(\epsilon_F) = \sum_q \sum_{q'>q} \left| \mu(q, \epsilon_F) - \mu(q', \epsilon_F) \right|, \quad (1)$$

where μ is the mean of the dataset at given q and ϵ_F. For V_O, q and q' can be 0, $+1$, or $+2$. We choose an arbitrary ϵ_F to minimize Δ, then constantly shift all $E_f[V_O]$ and standardize each distribution. As shown in Fig. 1(b), the resulting distributions overlap well, as desired.

Perturbed host states—We here focus on PHS using donor-type defects as examples, which has rarely been discussed despite its significance; the same applies to acceptor-type defects. Figure 2 schematically illustrates eigenvalues in supercell models with defects. When defects
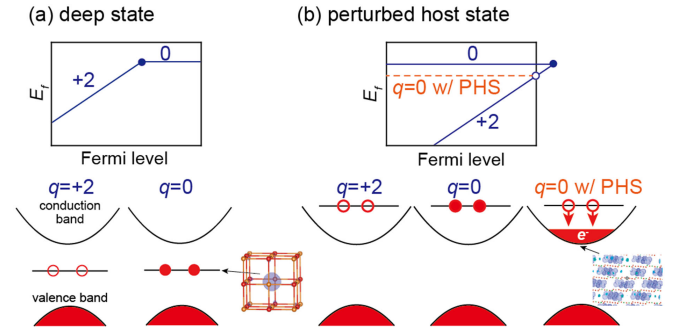

FIG. 2. Schematic illustration of defects with localized occupied states (a) inside the band gap and (b) above the CBM. In (b), the electrons are consequently dropped to the CBM, resulting in the delocalized PHS. The formation energy of a defect with PHS is indicated by a dashed line. Donor transition levels are schematically indicated in the formation energy diagrams as a function of the Fermi level. Transition levels involving deep localized states are accurately calculated from first principles (solid circles). In contrast, shallow levels related to the PHS are shown qualitatively (open circle), as they cannot be computed with realistically sized supercells (see text for details). Squared wave functions of these states in neutral oxygen vacancies in MgO and BaTiO$_3$ are also shown as examples.

exhibit localized in-gap states, they remain spatially confined and can be described using supercells composed of ~100 atoms [Fig. 2(a)]. However, when defects possess the localized occupied states above the conduction band minimum (CBM) [Fig. 2(b)], the electrons drop into the CBM and become loosely trapped by defect centers electrostatically, which are called PHS or shallow states [18–20]. Their weak binding causes them to extend over thousands of atoms or more [21].

The formation energy of a defect D in charge state q (D^q) with a single PHS, calculated using a supercell, is given by

$$E_f[D^q]|_{\epsilon_F=\epsilon_{CBM}} = E_f[D^{q+1}]|_{\epsilon_F=\epsilon_{CBM}} + E_{bind} + E_{overlap}$$
$$+ \int_{\epsilon_{CBM}}^{\infty} (\epsilon - \epsilon_{CBM})D(\epsilon)f(\epsilon)d\epsilon + \epsilon_{CBM},$$
$$(2)$$

where ϵ_{CBM} is the eigenvalue of the CBM, and $E_f[D^q]|_{\epsilon_F=\epsilon_{CBM}}$ is the formation energy of D^q at $\epsilon_F = \epsilon_{CBM}$. D and f are the density of states and orbital occupation fraction in the *supercell*, respectively. Note that the first term on the right-hand side is a defect formation energy without PHS at $\epsilon_F = \epsilon_{CBM}$. The second term corresponds to the binding energy of the carrier electron to the positively charged defect center, typically ranging from a few tens to a few hundred meV [21]. The third term represents the energy change from the overlap between PHSs in a supercell with periodic boundary conditions, generally negative, analogous to the stabilization observed in alkali metals composed of positively charged ion cores and free electrons. The fourth term arises from Burstein-Moss-like effects when multiple k-point sampling is used, being positive due to electron occupancy above the CBM. Note that the third and fourth terms arise from using small supercells, making accurate calculation of these small binding energies infeasible. In some studies, the formation energies of defects with PHS have been evaluated by removing the fourth term [22–24]; however, this is insufficient due to the contribution of the third term, which can correspond to a few hundred meV, resulting in the deep transition levels of the PHS [18–20]. Therefore, we usually present PHS-related transition levels qualitatively [14,25,26], as shown by an open circle in Fig. 2(b). Note that when the formation energies of deep-level defects are accurately predicted by ML models, the PHS-related transition levels can still be qualitatively described, as the transition levels are predicted to lie above the CBM, as shown in Fig. 2(b).

To simplify the discussion, we ignore the second and third terms, which are on the order of a few tenths of an eV and assume that the PHS occupies the CBM (i.e., no band filling effect). Then, we obtain $E_f[D^q]|_{\epsilon_F=\epsilon_{CBM}} = E_f[D^{q+1}]|_{\epsilon_F=\epsilon_{CBM}} + \epsilon_{CBM}$. This shows that $E_f[D^q]$ consists of $E_f[D^{q+1}]$ and the single-particle level of the CBM,

meaning that the defect formation energies with PHS qualitatively differ from those without PHS. Indeed, excluding defects with PHS from the dataset improves prediction accuracy, as shown later.

Accuracy evaluation—To validate our data handling framework, we built a GNN model using an $E_f[V_O]$ database with mixed q values [6]. Crystal structures are modeled using the CGCNN framework, as shown in Fig. S2 in the Supplemental Material [16] (see also Method for details). Other GNN models can also be used; the choice usually depends on the dataset size. We firstly encoded the atomic species and bond lengths, then passed them to convolutional layers. Subsequently, we extracted features at O sites at a pooling layer, and concatenated the charge state q. Finally, the data were fed into a fully connected neural network to predict an $E_f[V_O]$. (see Methods and Fig. S2 in the Supplemental Material [16]).

Figure 3(a) shows the parity plot of $E_f[V_O]$ for the test sets with the average mean absolute errors (MAEs). The prediction errors are 0.29, 0.22, and 0.37 eV for $q = 0, +1$, and $+2$, respectively, which are lower than in our previous study [6], despite using only structural information and a single GNN model here. Note that learning defect
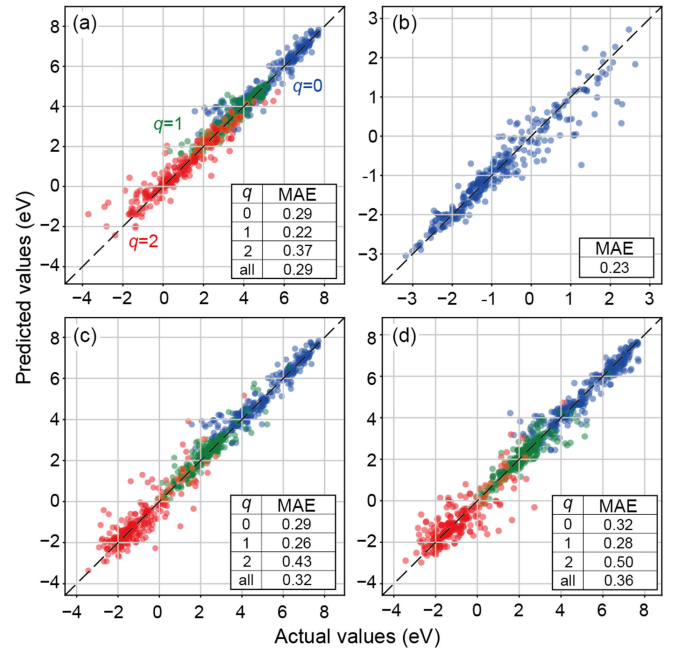

FIG. 3. Parity plots of (a), (c), (d) oxygen vacancy formation energies ($E_f[V_O]$) and (b) VBM positions. (a) $E_f[V_O]$ with the Fermi level (ϵ_F) aligned to the VBM of ZnO. (c),(d) $E_f[V_O]$ with ϵ_F aligned to the VBM of each compound, predicted using (c) a joint model and (d) a single model (see text for details). The x axis shows values from first-principles calculations, while the y axis shows our crystal graph convolution neural network predictions. Results are for test sets composed of 140 oxides. Mean absolute errors are shown in eV in the insets. Owing to site-dependent core potentials, the VBM in (b) also vary by site.

formation energies for each charge state q may improve accuracy, but reduces generality, as ML models must be built for every element-charge state combination. We confirmed that including data with PHS reduces accuracy by 0.02–0.03 eV (Fig. S3 in the Supplemental Material [16]), even though the dataset size increases by 10%.

The formation energies of charged defects depend linearly on the Fermi level, which varies within the band gap. Therefore, the band edge positions obtained from core potentials in unit cell calculations must also be determined. Since unit cells are not computed in structure-based virtual screening, separate ML models to predict the band edges are needed. In this study, we constructed a CGCNN model on a VBM dataset for the same oxides used in constructing oxygen vacancy database. Figure 3(b) shows a parity plot for the VBM, showing high accuracy of our CGCNN model.

Using the CGCNN models for both $E_f[V_O]$ and VBM, referred to as a joint model, we predicted $E_f[V_O]$ at the VBMs, as shown in Fig. 3(c). For comparison, Fig. 3(d) shows $E_f[V_O]$ predicted directly by a single CGCNN model based on the datasets of $E_f[V_O]$ at the VBM. The joint model achieves higher accuracy, with slight improvements of 0.02–0.07 eV. It is, however, noted that since the VBM can be obtained from unit cell calculations, unlike point defects, its prediction accuracy can be significantly improved using large-scale datasets like the Materials Project [2], which is a major advantage of this joint model.

Virtual screening—Using the joint model, we screened as-yet-unsynthesized hole-dopable oxides, since hole doping is typically hindered by compensation from positively charged V_O [14]. Our targets are the stable oxides predicted by Merchant *et al.* [27]. We firstly screen the oxides with the same criteria as the training set, identifying 1,809 candidate oxides [Fig. 4(a)]. Using our joint model, we screened oxides where V_O do not act as hole killers, that is, all V_O are stable in neutral or positive charge states with positive formation energies at the VBM in the O-rich conditions. The number of remaining oxides is 89 (4.9%), indicating how difficult it is to make *p*-type oxides.

To identify promising *p*-type oxides for applications, we excluded oxides containing highly toxic or expensive elements, leaving 32 candidates (see Methods). For these, we calculated the electron and hole effective masses, electronic and ionic dielectric constants, and optical absorption spectra using dielectric-dependent hybrid (DDH) functionals [28,29], which accurately predict the band gaps. Consequently, we have identified several prospective oxides as solar cell absorbates and transparent conducting oxides as shown in Fig. 4(b).

BaGaSbO is a promising photovoltaic material among them. As shown in the inset of Fig. 4(b), it exhibits very low electron and hole effective masses in the *ab* plane (0.34 and $0.19m_0$, where m_0 is the free-electron rest mass, respectively), a steep optical absorption onset [Fig. 4(c)] and slight indirect gap [Fig. 4(d)], which mitigates carrier
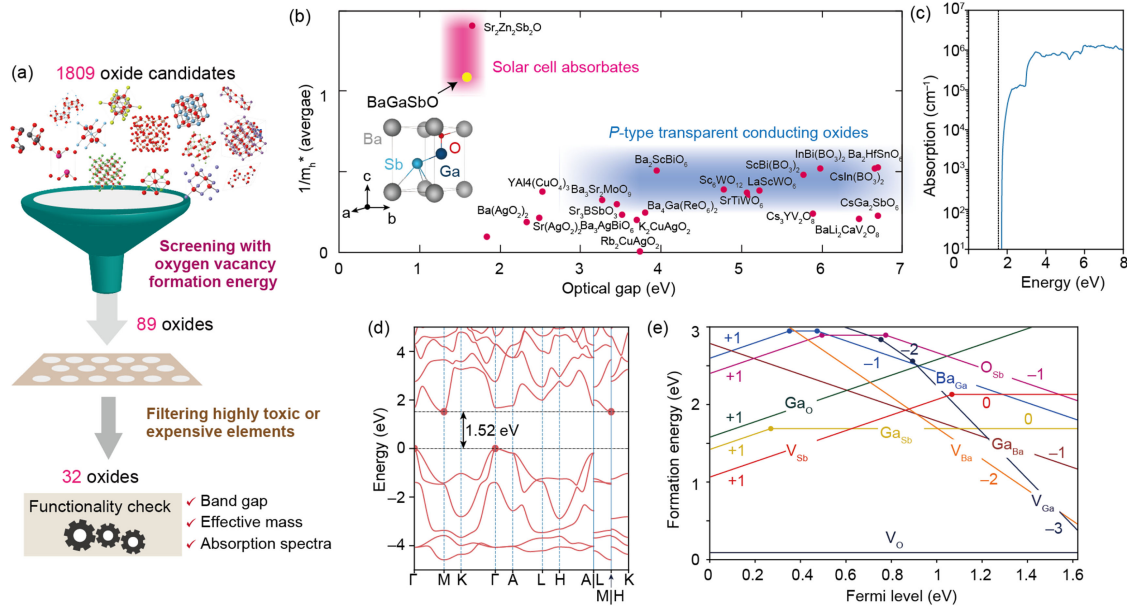

FIG. 4.   (a) Schematic illustration of our screening process for *p*-type oxides. (b) Reciprocal of the averaged hole effective mass (m_h^*) in units of the free-electron mass plotted against the optical gap for 32 screened oxides. Candidates for solar cell absorbers and transparent conducting oxides are highlighted. The crystal structure of BaGaSbO is also shown in the inset. (c) Optical absorption coefficient and (d) band structure of BaGaSbO. The band edge positions are marked with filled circles. (e) Formation energies of native defects in BaGaSbO under the Ga-rich condition. The charge states are also described. V_i denotes a vacancy of element i. These calculations were performed using the dielectric-dependence hybrid functionals (see the main text for details).

recombination. Although $Sr_2Zn_2Sb_2O$ exhibits more favorable effective masses, its large indirect-direct band gap difference (0.7 eV) would unfortunately cause significant efficiency loss (see Fig. S4 in the Supplemental Material [16]).

Figure 4(e) presents the calculated defect formation energies in BaGaSbO with the DDH functional. As predicted by our joint model, the oxygen vacancies are stable in the neutral charge state even at $\epsilon_F = \epsilon_{VBM}$, indicating no hole compensation. Furthermore, there are no defects with both low formation energies and deep levels, and it also shows p-type (n-type) behavior using Mg or Zn (La or Y) as dopants, respectively (Fig. S5 in the Supplemental Material [16]), indicating its remarkable potential as a photovoltaic material.

Conclusion—This study introduces an ML framework for point defect formation energies across multiple charge states. Focusing on V_O, we demonstrate how to prune inadequate data and normalize formation energies using the ϵ_F alignment. Our single CGCNN model achieves high accuracy merely with structural information. We further introduce the joint model that combines the models predicting the formation energies of defects and VBMs. With this model, we virtually screened stable oxides and identified 89 hole-dopable candidates, highlighting BaGaSbO as a promising ambipolar semiconductor for solar energy applications.

Finally, we discuss the inherent limitations and the scope of applicability of the proposed procedure. In the present protocol, the Fermi level is aligned using the core potentials. This procedure is also applicable to antisite defects and extrinsic substitutional dopants, since the core potentials of the original atoms in the unit cell are well defined. In contrast, for interstitial defects, no atoms occupy the corresponding positions in the unit cell, and thus the protocol cannot be directly applied in such cases. In addition, the error introduced by the ML model adds to the intrinsic errors of first-principles calculations. Therefore, higher-level reference data, such as those obtained from hybrid-functional calculations, are desirable to reduce the overall inaccuracy. The ML model is not intended to replace first-principles calculations but rather to complement them, serving as a prescreening tool to identify promising materials among numerous candidates in which point defects play crucial roles, as demonstrated in our screening of hole-dopable oxides.

Methods—To test the performance of our protocol, we used an V_O database comprising nonmagnetic 932 oxides with 2090 inequivalent sites [6]. Note that the PBEsol functional [30] with Hubbard U corrections [31] for Cu and Zn d orbitals and Ce f orbitals with $U_{eff} = 5$ eV was used to estimate the formation energies of oxygen vacancies, while the band edge positions were determined using the non-self-consistent DDH functional [28]. Note that, although the spin-orbit coupling was not taken into account, our CGCNN model can be combined with the band-edge positions determined with spin-orbit coupling in an *ad hoc* manner.

Initially, we excluded V_O that migrate to different sites or accompany neighboring atoms to higher-symmetry interstitials. We also removed V_O in dynamically unstable host structures (see Supplemental Note 1 in the Supplemental Material [16]). The numbers of V_O with $q = 0$, $+1$, and $+2$ are 1734 (2045), 1676 (2014), and 1416 (1637) after (before) removing defects with PHS. To identify V_O with PHS, we used the algorithm based on eigenvalues and orbital components of the single-particle levels near the band edges [6]. To avoid the data leakage, we divided the dataset by the oxides rather than by the O sites as we did previously [6]. The dataset was split into training, validation, and test sets in a 0.7:0.15:0.15 ratio for evaluating the accuracies. When constructing the CGCNN models, elemental information is embedded using original CGCNN descriptors, and bonding is represented via Gaussian filters (see Supplemental Note 2 in the Supplemental Material [16]) [7]. The model is illustrated in Fig. S2, and the tuned hyperparameters with optuna [32] are also tabulated in Tables S1 and S2 in the Supplemental Material [16]. $E_f[V_O]$ was evaluated by setting the oxygen chemical potential to half the total energy of an O_2 molecule in the spin-triplet state.

For virtual screening, the model was retrained on the full dataset using the same hyperparameters. In the structural database provided by Merchant *et al.* [27], slightly metastable structures are also included. Here, however, we focus on oxides stable against competing phases to enhance the likelihood of synthesizing the predicted candidates. For the screening by $E_f[V_O]$, we assumed oxygen-rich conditions. To determine these, we retrieved the total energies of the competing phases from the Materials Project [2] and constructed the chemical potential diagrams accordingly. In the next step, compounds with highly toxic (Tl, Hg, As, Cd, Pb, Cr) or expensive (Au, Pt, Pd) elements are excluded. We used DDH to evaluate the band gaps of screened oxides and the defect formation energies in BaGaSbO. Details of the first-principles calculations are described in the Supplemental Note 3 in the Supplemental Material [29,33–38].

Data availability—The data that support the findings of this article are openly available [39].

[1] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, Nat. Mater. **12**, 191 (2013).

[2] A. Jain *et al.*, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[3] W. Shockley and W. T. Read, Statistics of the recombinations of holes and electrons, Phys. Rev. **87**, 835 (1952).

[4] S. B. Zhang, S.-H. Wei, and A. Zunger, Overcoming doping bottlenecks in semiconductors and wide-gap materials, Physica B **273**, 976 (1999).

[5] 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).

[6] Y. Kumagai, N. Tsunoda, A. Takahashi, and F. Oba, Insights into oxygen vacancies from high-throughput first-principles calculations, Phys. Rev. Mater. **5**, 123803 (2021).

[7] M. D. Witman, A. Goyal, T. Ogitsu, A. H. McDaniel, and S. Lany, Defect graph neural networks for materials discovery in high-temperature clean-energy applications, Nat. Comput. Sci. **3**, 675 (2023).

[8] N. C. Frey, D. Akinwande, D. Jariwala, and V. B. Shenoy, Machine learning-enabled design of point defects in 2D materials for quantum and neuromorphic information processing, ACS Nano **14**, 13406 (2020).

[9] X. Wu, H. Chen, J. Wang, and X. Niu, Machine learning accelerated study of defect energy levels in perovskites, J. Phys. Chem. C **127**, 11387 (2023).

[10] J. B. Varley, A. Samanta, and V. Lordi, Descriptor-based approach for the prediction of cation vacancy formation energies and transition levels, J. Phys. Chem. Lett. **8**, 5059 (2017).

[11] A. Mannodi-Kanakkithodi, X. Xiang, L. Jacoby, R. Biegaj, S. T. Dunham, D. R. Gamelin, and M. K. Y. Chan, Universal machine learning framework for defect predictions in zinc blende semiconductors, Patterns **3**, 100450 (2022).

[12] T. Xie and J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. **120**, 145301 (2018).

[13] K. Yim, Y. Youn, M. Lee, D. Yoo, J. Lee, S. H. Cho, and S. Han, Computational discovery of p-type transparent oxide semiconductors using hydrogen descriptor, npj Comput. Mater. **4**, 17 (2018).

[14] Y. Kumagai, Computational screening of p-type transparent conducting oxides using the optical absorption spectra and oxygen-vacancy formation energies, Phys. Rev. Appl. **19**, 034063 (2023).

[15] T. N. H. Vu and Y. Kumagai, Investigation of hole dopability in oxygen-2p-dominated bands, Chem. Mater. (2025), 10.1021/acs.chemmater.5c02032.

[16] See Supplemental Material at http://link.aps.org/supplemental/10.1103/h66h-y5k6 for providing details of a CGCNN-based model that predicts charged oxygen-vacancy formation energies directly from crystal structures. It also explains how defect energies are aligned across materials (via O core potentials or VBMs) and how a consistent Fermi-level reference is chosen, and presents the model architecture, graph features, and optimized hyper-parameters, and documents the DFT workflow used to generate the training data, along with the pruning rules for removing anomalous vacancy relaxations.

[17] K. Cabello-Solorzano, I. O. de Araujo, M. Peña, L. Correia, and A. J. Tallón-Ballesteros, 18th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2023), Lect. Notes Netw. Syst. 344 (2023).

[18] 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 (2006).

[19] 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 (2018).

[20] 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).

[21] 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).

[22] H. R. Gopidi, L. Vashist, and O. I. Malyi, Physics of band-filling correction in defect calculations of solid-state materials, RSC Adv. **14**, 17675 (2024).

[23] S. Lany and A. Zunger, Anion vacancies as a source of persistent photoconductivity in II-VI and chalcopyrite semiconductors, Phys. Rev. B **72**, 035215 (2005).

[24] C. Persson, Y.-J. Zhao, S. Lany, and A. Zunger, n-type doping of $CuInSe_2$ and $CuGaSe_2$, Phys. Rev. B **72**, 035211 (2005).

[25] Y. Kumagai, S. R. Kavanagh, I. Suzuki, T. Omata, A. Walsh, D. O. Scanlon, and H. Morito, Alkali mono-pnictides: A new class of photovoltaic materials by element mutation, PRX Energy **2**, 043002 (2023).

[26] Y. Kumagai, K. Harada, H. Akamatsu, K. Matsuzaki, and F. Oba, Carrier-induced band-gap variation and point defects in Zn3N2 from first principles, Phys. Rev. Appl. **8**, 014015 (2017).

[27] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, Nature (London) **624**, 80 (2023).

[28] 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).

[29] J. H. Skone, M. Govoni, and G. Galli, Self-consistent hybrid functional for condensed systems, Phys. Rev. B **89**, 195112 (2014).

[30] 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).

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

[32] A. Teredesai *et al.*, Optuna, Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 2623 (2019).

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

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

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

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

[37] J. Heyd, G. E. Scuseria, and M. Ernzerhof, Hybrid functionals based on a screened Coulomb potential, J. Chem. Phys. **118**, 8207 (2003).

[38] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, Influence of the exchange screening parameter on the performance of screened hybrid functionals, J. Chem. Phys. **125**, 224106 (2006).

[39] https://github.com/kumagai-group/ML_charged_defects.