Accurate Prediction of HSE06 Band Structures for a Diverse Set of Materials Using $\triangle$-learning

Santosh Adhikari,[†] Jacob Clary,[‡] Ravishankar Sundararaman,[¶] Charles Musgrave,[§] Derek Vigil-Fowler,[‡] and Christopher Sutton[*,†]

†*Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC, USA*

‡*National Renewable Energy Laboratory, Golden, CO, USA*

¶*Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA*

§*Department of Chemical and Biological Engineering, University of Colorado, Boulder, Boulder, CO, USA*

E-mail: cs113@mailbox.sc.edu

Abstract

We used machine learning (ML) to accurately predict eigenvalues of the hybrid HSE06 functional using eigenvalues computed by the less computationally expensive PBE and associated electronic features based on the k-point resolved atomic band character. The ML model was trained using eigenvalues from only one k-point for each of the 168 compounds in the training set. The HSE06 eigenvalues across all k-points were then predicted for a separate set of 169 compounds with a mean absolute error (MAE) of 0.13 eV, representing a significant improvement over the error of PBE-computed

eigenvalues relative to HSE06 (0.96 eV). These accurately predicted eigenvalues result in remarkably accurate predictions for the band structures, projected density of states and band gaps, even though the model was not explicitly trained on these other properties. Finally, we demonstrate that our ML model has a similar accuracy for both ternary and quaternary compounds well outside the initial training set and on systems with 112 and 160 atoms, demonstrating its potential to rapidly predict HSE06-quality electronic structures of complex materials that are practically unfeasible for HSE06.

Keywords

eigenvalues, band gaps, projected density of states, PBE, HSE06

Introduction

The electronic band structure describes several critically important materials properties, such as the allowed electron energy levels, nature of the band gap (direct/indirect, metallic/non-metallic), and carrier mobility via the curvature and alignment of the band edges. These properties are particularly relevant in, for example, the engineering of materials for opto-electronic[1–3] and catalytic[4] applications.

The band structure is most commonly calculated using density functional theory (DFT) with generalized gradient approximation (GGA) functionals.[5–7] However, GGA functionals suffer from known inaccuracies in describing materials due to their lack of derivative discontinuities,[8,9] and self-interaction[10] and delocalization errors.[11,12] These inaccuracies manifest through GGA functionals' tendency to underestimate band gap energies[13,14] and incorrectly align band edges.[15] Indeed, numerous previous reports have shown that GGAs predict inaccurate band gaps for diverse classes of materials, including small gap semiconductors, (e.g., PbS), semiconductors (e.g., ZnO, AlP), and wide gap semiconductors (e.g., NaF, MgO, AlN), as one would expect for a ground-state theory such as DFT.[16–21] Other important energies

and energy gaps, such as the energies of adsorbate frontier orbitals relative to catalyst surface states, can be poorly predicted by GGA functionals[16,22] and lead to poor predictions of surface energies, adsorption energies, and reaction barriers. For example, the PBE functional[23] incorrectly predicts the CO adsorption site on Pt(111).[24]

Although some GGA functionals such as DFT-1/2,[25,26] mBJ,[27] GLLB-SC,[28] TASK,[29] and, mTASK[30] have been specifically parameterized to improve band gap prediction accuracy, it remains difficult for a single approach to perform consistently well for a wide variety of materials. Alternatively, accuracy in predicted band gaps and adsorption energies can be improved by applying an on-site Coulomb term, the Hubbard U correction, to specific orbitals within the GGA+U framework.[31,32] However, these calculations may have limited transferability because there is no universally accepted approach to determine the magnitude of the +U correction. Moreover, the use of a different value chosen to more accurately predict one property may greatly affect other predicted properties. Additionally, because +U corrections are conventionally only applied to states with d or f character, such as those in strongly correlated materials, states with s and p character could still be described poorly.[33,34]

Ab-initio methods can be used to correct the issues observed for GGA functionals, e.g., through the use of hybrid DFT functionals (e.g., HSE06[35]) or through many-body perturbation theory via the GW approximation,[36] which remains the favored approach for predicting eigenvalues and band gaps with greater accuracy across various material classes.[17,18,21,37] Both approaches require a significantly higher computational cost compared to GGAs. This has traditionally limited their use in modeling systems that contain a large number of atoms or in high-throughput screening studies.

Machine learning (ML) can accelerate traditional computational materials discovery by rapidly estimating properties of some arbitrarily large chemical space using an ML model trained on a set of reference data (i.e., the training set) at orders of magnitude lower computational cost than QM methods. ML has recently been applied to learn various electronic

properties of materials, such as the nature of the band gap (direct/indirect, metallic/non-metallic),[38–40] band gap energies,[21,39–48] and the positions of the band edges.[48] Δ-learning[49] has been used to predict high-fidelity band gaps (e.g., HSE06, GW, or experiment) using lower-fidelity, less computationally demanding methods (typically semilocal DFT calculations).[21,44–48]

However, the majority of these studies used ML to predict a single property per sample, while high-fidelity quasi-particle GW eigenvalue shifts applied as corrections to DFT eigenvalues are not constant across all k- and band-indices.[22] This motivates Δ-learning *the complete band structure* by understanding how the k- and band-resolved eigenvalues of high-fidelity calculations shift from lower-fidelity calculations. A recent study in this direction explored applying the Δ-learning approach to predict G_0W_0@PBE band-structures for 286 non-magnetic 2D semiconductors using 3300 features that represent the local energy and radial distance between eigenstates.[50]

For periodic materials however, the need for detailed spatial information may not be as crucial, particularly for materials with few symmetrically unique atom sites. In this work, we show that it is possible to achieve accurate Δ-learning for the prediction of HSE06 band- and k-point resolved eigenvalues using only 14 features of the atomic band character (determined from nl-resolved projectors as defined in Methods section), 3 materials-specific features, and the PBE eigenvalues generated at one k-point per material, which do not explicitly include structure-based information. We demonstrate this approach on a diverse set of 337 bulk semiconducting/insulating binary compounds with a wide range of chemical compositions and show that it enables accurate electronic structure prediction and evidence that it can generalize to more complex materials, thus enabling faster materials discovery.

Table 1: Features used as inputs into our ML models

Type	Label	Definition	Unit
Eigenvalues	$\epsilon_{ik,PBE}$	PBE energy eigenvalue associated with ψ_{ik}	eV
Orbital-projections based features	$1s_{PBE}$, $2s_{PBE}$, $2p_{PBE}$, $3s_{PBE}$, $3p_{PBE}$, $3d_{PBE}$, $4s_{PBE}$, $4p_{PBE}$, $4d_{PBE}$, $5s_{PBE}$, $5p_{PBE}$, $5d_{PBE}$, $6s_{PBE}$, $6p_{PBE}$	nl-projectors of the PBE wavefunctions onto the spherical harmonics	-
Bader-based features	δ_{PBE}	average charge transfer, from Bader analysis on PBE charge density	e^-/at.
	P_{PBE}	dipole moment per unit volume, from Bader analysis on PBE charge density	$e^-/Å^2$
Atom-based features	Z_{pet}	a compound index based on modified Pettifor values[51] defined for constituent elements	-

Results and discussion

Model design and overall performance

For the "Δ" learning scheme, 18 features were used (Table 1), and three types of regression models were trained: linear (Linear) and kernel ridge regression using a Laplacian kernel (KRR), and a stacked Linear+KRR model (see Table 2)). We attempted the Linear+KRR model based on the observation that the Linear model alone reduces the MAE considerably relative to PBE. To establish a baseline for comparison, we trained two distinct linear models for unoccupied and occupied bands using PBE eigenvalues as the only feature. This technique, commonly known as a Scissors operator, is prevalent in GW calculations. The ML models were trained separately for the valence and conduction bands across all materials in the training set (50% of the 337 total compounds (Table S1); see Methods for a detailed description and learning curves). The mean absolute error (MAE) between the predicted and HSE06 eigenvalues was used in all cases as the loss function.

Eigenvalue accuracy across materials classes

The eigenvalue errors obtained from the PBE calculations exhibit a multi-modal distribution with a wide range of errors (MAE = 0.96 eV), particularly for the valence bands. This is attributed to the significant self-interaction error present in PBE calculations. The most significant errors in PBE calculations are predominantly associated with transition metals, chalcogenides, and halides (Figure S1 (a)). These materials constitute the majority of the

test set (out of a total of 169 compounds in the test set, 48, 67 and 51 compounds containing transition metals, halides, and chalcogenides). Despite its simple form, the Scissors model effectively corrects the approximately bimodal distribution observed in the PBE eigenvalue errors (Figure 1), thereby significantly improving the MAE (0.30 eV), and the 5^{th} and 95^{th} percentiles (Table 2).

However, the error distribution of the Scissors model remains non-uniform and exhibits long tails, which is primarily attributed to the same problematic material classes observed in the PBE error histogram (as depicted in Figure 1 and Figure S2). Indeed, the Scissors model results in an MAE $>$ 0.5 eV for 14 compounds (Table S2), 12 of which are either chalcogenides or halides. Figure 2 provides further confirmation that the eigenstates with a dominant d-orbital character exhibit significant errors in both the PBE and Scissors models. These results suggest that although the Scissors operator can mostly improve the errors in the PBE eigenvalues compared with HSE06, this approach cannot account for the bias that may occur due to different materials classes and types of bonding.

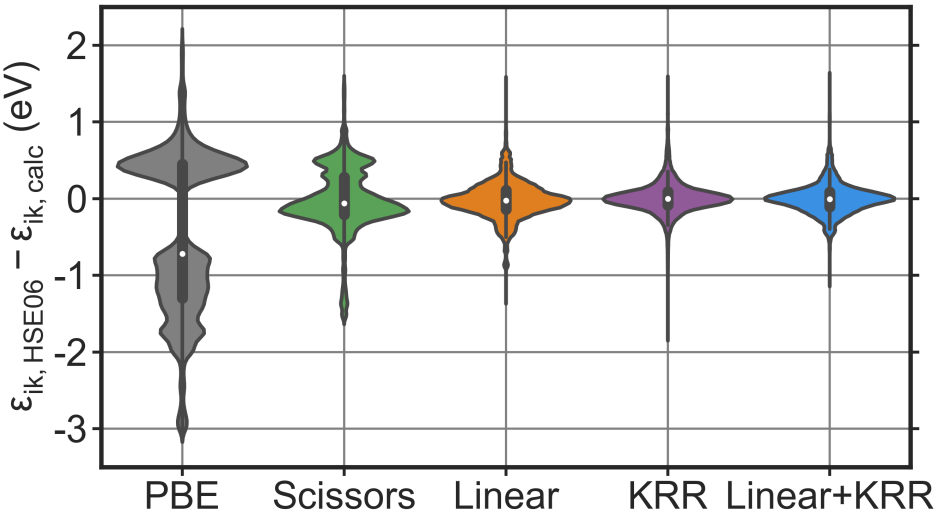


Figure 1: Violin plots showing the error distribution in the predicted PBE, Scissors, KRR, Linear, and Linear+KRR eigenvalues relative to the HSE06 eigenvalues. The shape of each violin displays the distribution of errors, while the thick lines and white circles inside each violin represent the range of the 25^{th} and 75^{th} percentiles and the 50^{th} percentile of the distribution, respectively.

The Linear, KRR and Linear+KRR models have significantly lower MAEs compared

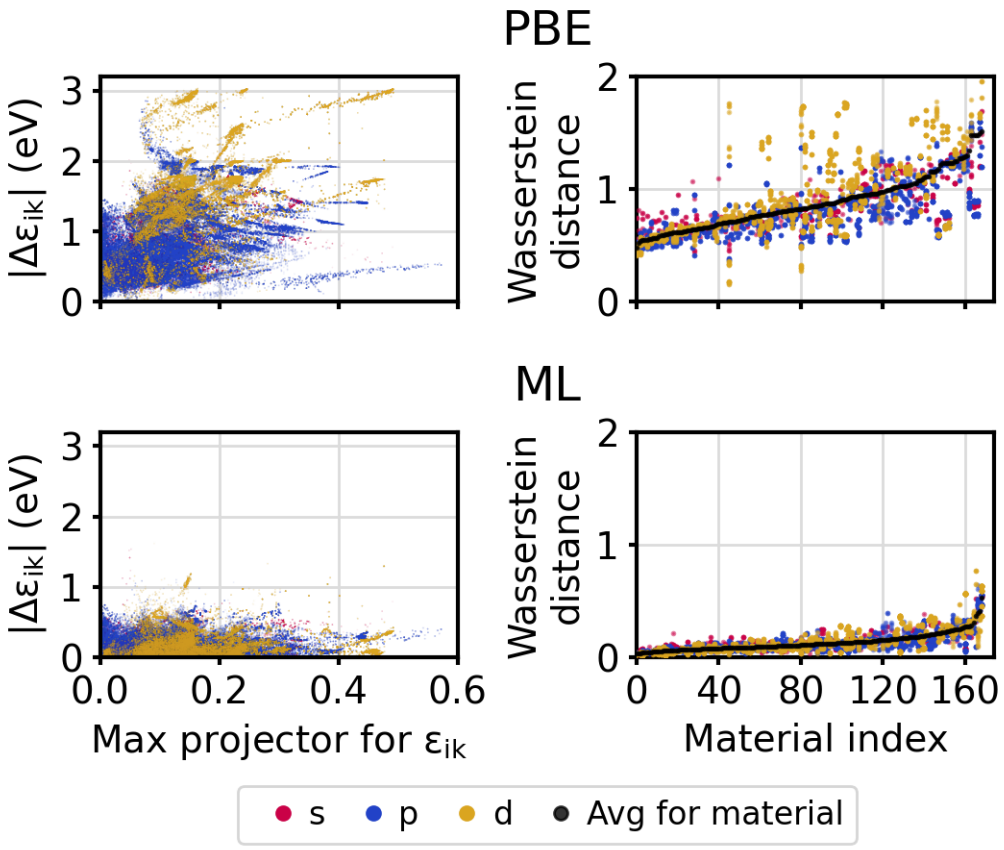


Figure 2: The left column shows the PBE (top left) or ML (bottom left) eigenvalue errors relative to HSE06 eigenvalues ($|\Delta\epsilon_{ik}|$) as a function of the maximum value of the PBE s (red), p (blue), and d (gold) orbital projectors for a given eigenvalue. The PBE eigenvalues are inaccurate for all orbitals while the ML model eigenvalues exhibits similar accuracy for all orbitals. Plots of eigenvalue errors for all models considered in this work are shown in Figure S4. The right column shows the Wasserstein distances of the PDOSs generated using PBE and ML eigenvalues with PBE projectors as compared to HSE06 eigenvalues with PBE projectors. The compounds are sorted by average (black) error across their s (red), p (blue), and d (gold) orbitals. The PBE PDOSs are inaccurate for all orbitals while the ML model eigenvalues produce PDOS plots with similar accuracy for all orbitals. Plots of Wasserstein distances for all models considered in this work are shown in Figure S5.

Table 2: Comparison of errors in the eigenvalues predicted from the Δ-learning ML model types (KRR, Linear, Linear+KRR) and the Scissors operator. Each model type was re-trained 21 times using different random seeds. The standard deviations for the 21 models are shown in parentheses. The summary statistics (i.e., fifth (Q5) and ninety-fifth percentile (Q95) MAEs, all in eV, and R2-scores) are shown for the model with the median error out of 21 models. The MAEs of the model-predicted band gaps (E_g) (in eV) are also shown, which were computed as the difference between the predicted eigenvalues at the same k-point/band indices as the conduction band minimum (CBM) and valance band maximum (VBM) in the PBE calculation.

Model	Eigenvalue predictions (eV)				E_g prediction (eV)
	MAE	R2-score	Q5	Q95	MAE
PBE	0.96	0.970	-1.99	0.77	1.20
Scissors	0.30 (0.01)	0.996	-0.54	0.59	0.62
Linear	0.17 (0.01)	0.999	-0.41	0.37	0.25
KRR	0.14 (0.01)	0.999	-0.31	0.33	0.25
Linear+KRR	0.13 (0.01)	0.999	-0.27	0.33	0.25

with PBE eigenvalues relative to HSE06 (0.17 eV, 0.14 eV, and 0.13 eV respectively, for the test set; see Table 2). Note that the subsequent discussion primarily centers on the results from the Linear+KRR model alone. The marked decrease in the MAEs is clearly evident in the violin plot (Figure 1) representing the distribution of eigenvalues across all material classes. Notably, the range of eigenvalue errors is significantly narrower than that observed for PBE, with the width of the 5^{th} and 95^{th} percentile errors being reduced by a factor of approximately 5 (as illustrated in Table 2). In contrast to the multimodal distribution of the PBE eigenvalue errors with a long tail in the valence states (Figure 1), all these models show a unimodal distribution centered around 0 eV (see Figure S1(b)). Moreover, the errors in eigenstates characterized by projectors with high d character are comparable to those of eigenvalues with high s or p character. This result indicates that the ML model improves the bias in the PBE and Scissors eigenvalue errors in the eigenstates with particular orbital characters (see Figure 2, left column).

The consistent performance of the ML model yields comparable accuracy for a diverse range of chemistries and bonding environments. This includes compounds containing transi-

tion metals (MAE = 0.14 eV), alkali metals (MAE = 0.11 eV), alkaline earth metals (MAE = 0.13 eV), or other elements primarily belonging to groups 13 and 14 of the periodic table (MAE = 0.13 eV), where the MAEs for these specific groups of compounds are similar to the overall test set MAE (0.13 eV), as shown in Figure S3. Similarly, consistent errors are calculated when grouping compounds by their anion, such as for halides (MAE = 0.11 eV) and chalcogenides (MAE = 0.11 eV), which represent the most common materials classes in the test set, constituting 67 and 51 out of 169 compounds, respectively (Figure S3). The largest eigenvalue MAEs for the ML model are for nitrides (MAE = 0.20 eV) and oxides (MAE = 0.20 eV), respectively. The larger errors for nitrides are attributed to the fact that the dataset itself contains fewer nitrides overall (9/168 training compounds and 6/69 test compounds). The relatively higher errors observed for oxides (e.g., ZnO MAE = 0.35 eV, OsO_4 MAE = 0.43 eV, WO_3 MAE = 0.45 eV) is not surprising, considering that providing accurate depictions of transition metal oxide electronic structures remains a persistent obstacle for computational methods, and Δ-learned corrections will inherit those challenges.

Because the feature set used in this work does not explicitly include structure-based information, a key question is the generalizability of the ML models to other compositions or polymorphs (i.e., different structures not included in the original training set). To test the generalizability of the ML models, we generated three distinct test sets, each comprising: (a) unique stoichiometries that are in the test set only, (b) polymorphs with at least one compound in the training set, or (c) polymorphs within the test set only. The ML model exhibits similar performance (within $\pm$0.02 eV of the overall test set eigenvalue MAE) for each of the three categories (Figure S6). The exceptional performance of the ML model in predicting eigenvalues for categories featuring previously unseen data is noteworthy given that the model did not employ any structure-based features (such as density, bond distances, nearest neighbors, etc.), which are typically more effective in discriminating polymorphs during model training. This observation indicates that the feature set utilized in this study facilitates the development of a robust ML model capable of accurately describing various

chemistries and bonding environments in different polymorphs.

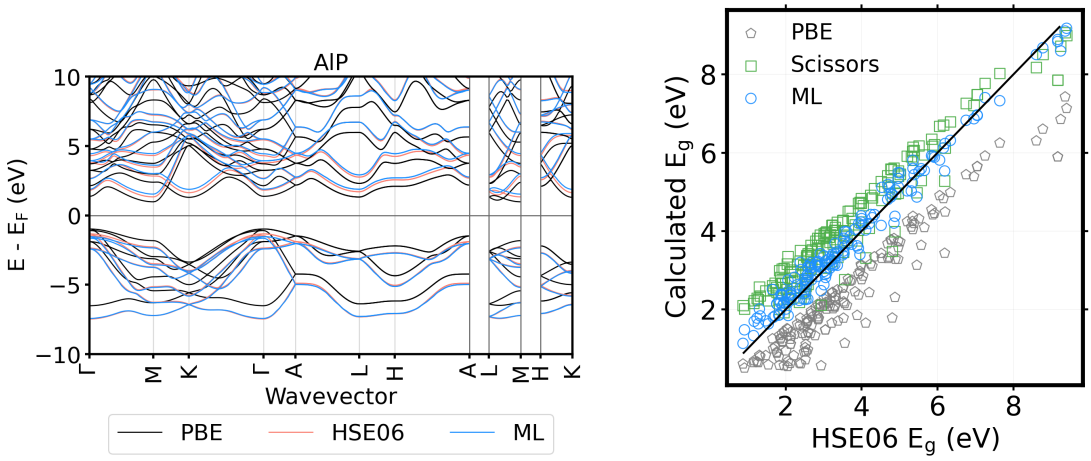
Prediction of electronic structures using ML eigenvalues

After confirming our ML model's capability to accurately predict HSE06 eigenvalues, we use the model to generate band structures at essentially HSE06 for every compound in the dataset.

In order to guarantee that all samples were included in the test set, the initial training and test sets, each representing ∼50% of the dataset, were swapped and a new model was re-trained (see the SI for details, the zipped folder is provided via a link below). We selected 6 systems and compared them with the results from HSE06 calculations (Figure 3 (a) and S7). The choice of these selected compounds was based on the following criteria: (a) representation of the diversity of cations and anions in our dataset, (b) coverage of a wide range of band gaps (including low to typical ranges for compounds such as ZnO, AlP, and GaSe, as well as wide band gaps for MgO and AlN), and (c) prior literature reports indicating significant underestimation of band gaps by PBE. The highest accuracies in ML-predicted band structures occur in materials with minimal d orbital contributions to the bands (eg. AlP, AlN, NaH, GaSe, PbI_2, Figures 3(a) and S7). In a small amount of cases (eg. ZnO, Figure S7), that involve dominant d orbital contributions to the bands, occasional kinks are observed in ML-predicted band structures where multiple bands from PBE are nearly degenerate. Overall these results demonstrate that ML can effectively produce band structures with a level of accuracy comparable to HSE06 irrespective of materials class, band gap ranges, and method of k-space sampling.

Unsurprisingly, the high accuracies achieved by ML-predicted eigenvalues and band structures lead to more accurate electronic properties, such as the band gaps, regardless of the material class. Compared with HSE06 band gaps, the MAE for the ML model is 0.25 eV, which is a factor of ∼5 lower error than PBE band gaps (MAE = 1.20 eV) (see Table 2, Fig. 3 (b)). Despite being trained on a different target property these results are comparable to

the performance of an ML model trained directly to predict the HSE06 band gaps (MAE = 0.23 eV) for the same train-test split and using the same features (Fig. 3 (b), band gaps predicted from all models are reported in Table S2 (see Table S1 for PBE and HSE06 band gaps across the whole dataset).



(a) (b)

Figure 3: Figures showing (a) the PBE (black), HSE06 (red), and ML (blue) band structures of AlP (mp-8880), and (b) the HSE06 band gaps (black diagonal line), and band gaps predicted by PBE (gray), Scissors (green), and ML model (Linear+KRR, blue). The band gap MAEs for the KRR (0.25 eV) and Linear models (0.25 eV) are similar to that of Linear+KRR (0.25 eV) model. For reference, a KRR model trained explicitly for band gaps using the same compounds in training and test set predicts the HSE06 band gaps with an MAE of 0.23 eV. The band gaps predicted by all models are provided in Table S2.

In addition to the band gap and band structure, we also evaluated the accuracy of the model-predicted eigenvalues for reproducing HSE06 projected density of states (PDOS) using HSE06 eigenvalues and PBE nl -resolved projectors, which are used to calculate the atomic band character (as defined in Methods section). We quantified the similarity between the ML-predicted PDOS curves and their HSE06 counterparts for all compounds in the test set using the first Wasserstein distances (WD). This metric estimates the minimum "work" needed to transform one curve into another, with smaller distances indicating greater similarity between the curves. In order to ensure fair comparison between different models, we utilized PBE projectors for all plots because none of the models were designed to predict new projectors

in addition to new eigenvalues.

The plot of WD as a function of orbital type suggests that the ML models exhibit approximately three times higher accuracy in predicting HSE06 PDOS curves compared to PBE (Figure 2 right column, Figure S5). The low WD observed across all orbital types and compounds indicates that the ML models have a consistent accuracy across different material classes and can reliably predict the PDOS of diverse materials, regardless of their composition. The significant improvement of the ML model's WD for PDOS curves of d states as compared to PBE and the Scissors model shows that these models excel even for compounds known to be poorly described by PBE.

Model performance for systems with a large number of atoms/compositions

Although the dataset used for training/testing of the ML model encompasses a variety of chemical families, we restricted our dataset to only binary compositions that include ≤ 6 atoms per unit cell for convenience. To deploy our ML model to more complicated systems, we selected five ternary and quaternary systems from the Materials Project database of widely studied perovskites ($CaTiO_3$ (mp-5827), $CsPbI_3$ (mp-1069538, mp-540839)), materials relevant for battery applications ($VCoO_4$ (mp-771137), $LiVCoO_4$ (mp-753151)) and two binary systems that contain a large number of atoms per unit cell for which HSE06 is impractical (Nb_2O_5 (mp-556048, 112 atoms), and Tl_2Cl_3 (mp-680294, 160 atoms). These systems were downloaded from the Materials Project and the HSE06 eigenvalues were calculated.

The ML model effectively corrects the errors in the eigenvalues obtained from PBE calculations (MAE = 1.02 eV) to more accurately compare with HSE06 eigenvalues (MAE = 0.27 eV), see Fig. 4 (a). The results for these systems (Table S3) are consistent with our other results described above. Thus, despite the complexity in the band structures of these additional systems, the ML model eigenvalues show a consistent four-fold improvement over PBE, which is remarkable considering that these systems were completely unknown during the model training/testing. The accuracies achieved in the ML model's prediction of the

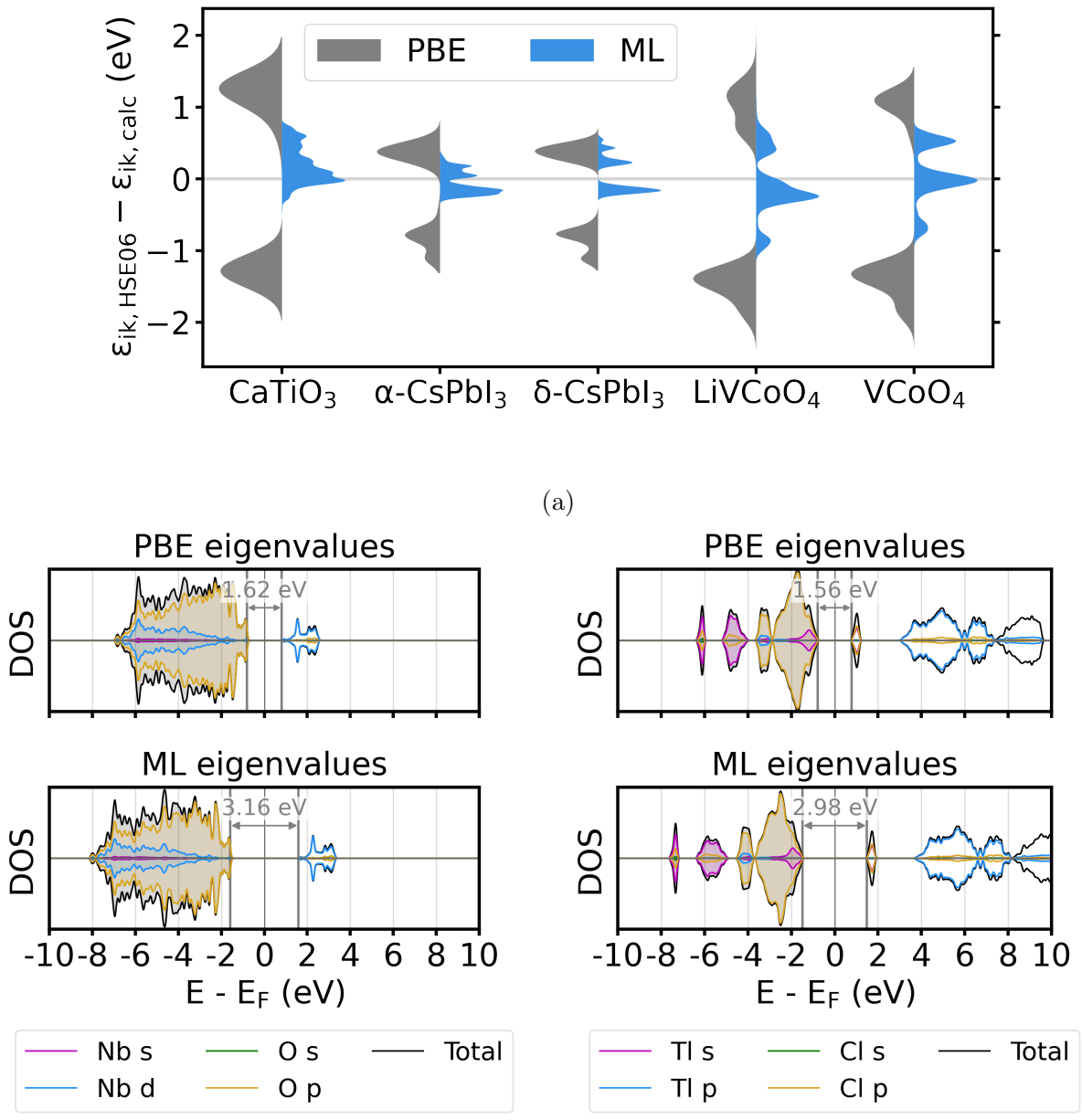


Figure 4: Figures showing (a) the violin plots of the PBE (gray) and ML (blue) eigenvalue errors relative to the HSE06 eigenvalues for $CaTiO_3$ (mp-5827), α-$CsPbI_3$ (mp-1069538), δ-$CsPbI_3$ (mp-540839), $LiVCoO_4$ (mp-753151), and $VCoO_4$ (mp-771137), and PBE and ML (Linear+KRR) projected density of states (PDOS) of (b) Nb_2O_5 (mp-556048), and (c) Tl_2Cl_3 (mp-680294).

HSE06 eigenvalues are also apparent in the band structures of $CaTiO_3$ (Fig. S8 (a)) and $CsPbI_3$ (Fig. S8 (b)).

However, unlike the results for the ML model applied to binary compounds, we observe that the eigenvalue error distribution of the ML model remains multimodal in $CaTiO_3$, $LiVCoO_4$, and $VCoO_4$ (Fig. 4 (a)). We attribute this to the averaging of the orbital projectors across various atom types, which is done to maintain a compact 14-value representation. Hence, in ternary and quaternary compounds, the model only perceives an "effective" cation, unlike in binary compounds where a single unique cation is present.

Finally, the ML model was applied to two systems, Nb_2O_5 and Tl_2Cl_3, which contain 112 and 160 atoms per unit cell, respectively. Rather than showing the complicated band structures for these compounds, the HSE06-quality PDOS are generated using the eigenvalues predicted from our ML model (Fig. 4 (b and c)). This result is especially noteworthy as the ML model prediction has essentially no computational cost while performing HSE06 calculations on such large systems necessitates running calculations on the largest existing supercomputers.

Conclusions

In this work, we used ML to accurately predict the HSE06 eigenvalues for a dataset of 337 semiconducting/insulating bulk solids. Δ-learning for the prediction of HSE06 band- and k-resolved eigenvalues was performed using only 18 features per k-resolved eigenvalue and training on only data at one k-point for each material. The MAE of the ML model is 0.13 eV across all k-points (i.e., the entire band structure) for a test set of 169 compounds previously unseen to the model. This error is a factor of 7 lower than PBE eigenvalues relative to HSE06 (0.96 eV). The features used in the ML model are primarily related to the atomic character of given electronic states. Our results indicate that the ML model developed here consistently predicts accurate eigenvalues across materials classes, indicating that the feature

set is robust and describes many different chemistries and bonding environments. Indeed, the need for data only at one k-point for each compound shows that electronic structure information is largely redundant across k-space.

Additionally, we show that these accurately predicted eigenvalues translate to more accurate band gaps and projected densities of states, for which the ML model was not explicitly trained. For example, the ML model band gap MAE is 0.25 eV relative to HSE06 band gaps, which is a factor of $\sim$5 lower error than the PBE band gaps (MAE = 1.20 eV)

Finally, the ML model is shown to perform accurately for more complex ternary and quaternary systems such as $VCoO_4$, $LiVCoO_4$, $CaTiO_3$, and $CsPbI_3$ compared to HSE06. Because of the computational efficiency of the ML model, it was also applied to predict the band structures and PDOS of Nb_2O_5 and Tl_2Cl_3, which contain 112 and 160 atoms per unit cell, respectively. At these large system sizes, HSE06 calculations would be impractical.

Future models using these features should be generalized in order to predict eigenvalue shifts for metallic/near-metallic compounds for which higher fidelity methods could change the occupancy of a KS orbital. These models could also incorporate more explicit structural features that can generalize into more complex compounds and potentially surface structures as well.

Methods

DFT calculations

DFT calculations were performed using the Vienna Ab initio Simulation Package (VASP)[52,53] version 5.4.4 with periodic boundary conditions. Standard projector augmented wave (PAW) pseudopotentials[54,55]were used for all elements. We performed geometry optimizations using the PBE[23] GGA functional and planewave cutoffs equal to 30% higher than the largest recommended cutoff of any pseudopotential within a material. *One-shot* HSE06[35] (which we refer to as HSE06 throughout this letter) was performed using the PBE-optimized geometries

and electron densities. The Γ-centered Monkhorst-Pack k-point grids used contained at least 1000 k-points per reciprocal atom. Band structure k-paths were generated using the SeeK-path Python package.[56,57]

Model training and selection

scikit-learn[58] version 1.1.1 was used to train the linear and kernel ridge regression, KRR, (using the Laplacian kernel) models. The hyperparameters for KRR were tuned via grid search using 5-fold cross-validation and the features were normalized using the min-max technique. Train and test set partition was based on the number of compounds. For example, if 168 random compounds were selected for training, the remaining 169 compounds were reserved for testing purposes. All models were initiated with different random seeds and trained 21 times using randomly selected k-points and compounds for each run. Out of these 21 runs, the models with the median MAE out of the 21 runs are discussed. The standard deviation of the calculated errors on the test set of these independent runs is reported in Table 2. Variations of model accuracy with the number of training compounds (Figure S9) and k-points (Table S4) were studied to determine their optimal values. Because most practical applications of accurate electronic structure predictions are concerned with states near E_F, our analysis focused on eigenvalues within $E_F \pm 10$ eV, whereas the models are trained on eigenvalues within $E_F \pm 15$ eV. The focus on a narrower range of energies closer to the Fermi level leads to an MAE ca. 0.01 eV larger compared with the larger range of $E_F \pm 15$ eV in the test set (see SI for details on model training, training set selection, and learning curves).

Dataset

The dataset for this study consists of 337 binary solids from the Materials Project (MP) database[5] (queried in April 2022) spanning 52 elements across the periodic table. These compounds were selected based on these criteria: (a) decomposition enthalpies of less than

50 meV/atom; (b) PBE band gap greater than 0.5 eV; (c) no more than 6 atoms in the unit cell; and (d) no f-block elements. The dataset spans 52 elements across the periodic table and incorporates diverse chemical families such as oxides, nitrides, phosphides, halides, chalcogenides, etc. (Figure 5). The most common cation classes are transition metals (108/337) followed by alkaline earth metals (72/337) and alkali metals (71/337). Si, Ge, and Sn are considered to be cations except when bonded with calcium.

The dataset comprises 227 distinct compositions, with 75 of those represented by 2 or more (average of 2.46) polymorphs. Most of the polymorphs are chalcogenides, featuring rock salt (RS), zinc blende (ZB), wurtzite (WZ), and Ni-As (NA) structures. A list of all compounds used in this study along with the corresponding **mp-id**, number of atoms per unit cell, PBE and HSE06 bandgaps are available in Table S1.

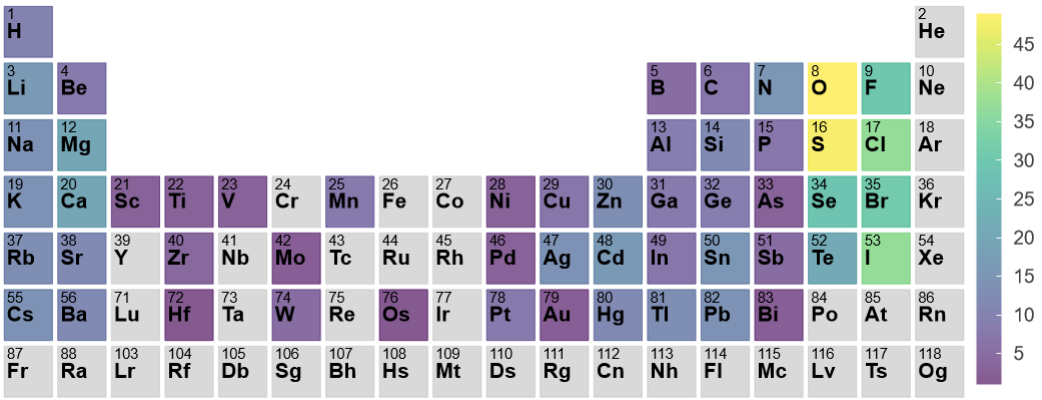


Figure 5: Distribution and frequency of the various elements in our dataset. Out of 337 compounds, halides (135), chalcogenides (98), and oxides (49) are the three most abundant chemical families with transition metals (108), alkaline earth metals (72), and alkali metals (71) the most frequent cations.

Orbital eigenvalues and projector features

For each Kohn-Sham eigenfunction ($\psi_{i\mathbf{k}}$) associated with a band index i and k-point label $\mathbf{k}$, we define the projectors, $f_{nl}^{sa}[\psi_{i\mathbf{k}}]$, onto the spherical harmonics, χ_{nlm}^{sa}, for all atoms, a, of

a given element type, s, in the system as:

$$f_{nl}^{sa}[\psi_{i\mathbf{k}}] \equiv \sum_{m=-l}^{l} |\langle \chi_{nlm}^{sa} | \psi_{i\mathbf{k}} | \rangle|^2, \tag{1}$$

where n, l, and m are the valence principal, angular and magnetic/azimuthal quantum numbers of the atom, respectively. To account for differing numbers of atoms in the unit cell between materials, the projectors from Eq. 1 were averaged over the total number of atoms of type s, N_a^s, in the unit cell to produce an averaged projector for each atom type, $f_{nl}^{s}[\psi_{i\mathbf{k}}]$:

$$f_{nl}^{s}[\psi_{i\mathbf{k}}] \equiv \sum_{a=1}^{N_a^s} \frac{f_{nl}^{sa}[\psi_{i\mathbf{k}}]}{N_a^s}. \tag{2}$$

In this work, Eq. 2 was evaluated for $1 \leq n \leq 6$ and $0 \leq l \leq 2$ (excluding $6d$, $n = 6$ and $l = 3$) as input features. This results in 14 orbital features for each material and atom type for that material, namely projectors for the *1s, 2s, 2p, 3s, 3p, 3d, 4s, 4p, 4d, 5s, 5p, 5d, 6s*, and *6p* orbitals. These projectors and their associated eigenvalues, ϵ_{ik} were calculated using the PBE functional.

Charge and stoichiometry-based features

We used the Bader[59] charge density partitioning method to compute two charge transfer-based features, namely, the average charge transfer (δ) and dipole moment per unit volume (P), as utilized by Ref.[60]

To numerically distinguish various cations in the compound, we utilized a modified Pettifor index[51] (Z_{pet}); a unique value assigned to each element in the periodic table that encodes the extent of its replaceability in the crystal structure.

Author Information

Author Contributions

Santosh Adhikari and Jacob Clary contributed equally to this work.

Acknowledgement

Financial support for this work was provided by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under Award No. DE-SC0022247 and departmental start-up funds at the University of South Carolina. The research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory.

Supporting Information Available

The following file is available free of charge.

- SI.pdf: contains figures and tables that support the letter
- PBE_vs_ML_band_structure_plots.zip : zip file containing PBE and ML band structure plots for all 169 compounds in the test set

References

(1) Schubert, E. F.; Kim, J. K. Solid-state light sources getting smart. *Science* **2005**, *308*, 1274–1278.

(2) Polman, A.; Knight, M.; Garnett, E. C.; Ehrler, B.; Sinke, W. C. Photovoltaic materials: Present efficiencies and future challenges. *Science* **2016**, *352*, aad4424.

(3) Soe, C. M. M.; Stoumpos, C. C.; Kepenekian, M.; Traoré, B.; Tsai, H.; Nie, W.; Wang, B.; Katan, C.; Seshadri, R.; Mohite, A. D., et al. New type of 2D perovskites with alternating cations in the interlayer space,(C (NH2) 3)(CH3NH3) n Pb n I3 n+ 1: Structure, properties, and photovoltaic performance. *Journal of the American Chemical Society* **2017**, *139*, 16297–16309.

(4) Pinaud, B. A.; Chen, Z.; Abram, D. N.; Jaramillo, T. F. Thin films of sodium birnessite-type MnO2: optical properties, electronic band structure, and solar photoelectrochemistry. *The Journal of Physical Chemistry C* **2011**, *115*, 11830–11838.

(5) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G., et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **2013**, *1*, 011002.

(6) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials* **2015**, *1*, 1–15.

(7) Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O., et al. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science* **2012**, *58*, 218–226.

(8) Perdew, J. P.; Parr, R. G.; Levy, M.; Balduz Jr, J. L. Density-functional theory for fractional particle number: derivative discontinuities of the energy. *Physical Review Letters* **1982**, *49*, 1691.

(9) Perdew, J. P.; Levy, M. Physical content of the exact Kohn-Sham orbital energies: band gaps and derivative discontinuities. *Physical Review Letters* **1983**, *51*, 1884.

(10) Perdew, J. P.; Zunger, A. Self-interaction correction to density-functional approximations for many-electron systems. *Physical Review B* **1981**, *23*, 5048.

(11) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Insights into current limitations of density functional theory. *Science* **2008**, *321*, 792–794.

(12) Mori-Sánchez, P.; Cohen, A. J.; Yang, W. Localization and delocalization errors in density functional theory and implications for band-gap prediction. *Physical Review Letters* **2008**, *100*, 146401.

(13) Bredow, T.; Gerson, A. R. Effect of exchange and correlation on bulk properties of MgO, NiO, and CoO. *Physical Review B* **2000**, *61*, 5194.

(14) Liu, P.; Franchini, C.; Marsman, M.; Kresse, G. Assessing model-dielectric-dependent hybrid functionals on the antiferromagnetic transition-metal monoxides MnO, FeO, CoO, and NiO. *Journal of Physics: Condensed Matter* **2019**, *32*, 015502.

(15) Feibelman, P. J.; Hammer, B.; Nørskov, J. K.; Wagner, F.; Scheffler, M.; Stumpf, R.; Watwe, R.; Dumesic, J. The CO/Pt (111) Puzzle. *The Journal of Physical Chemistry B* **2001**, *105*, 4018–4025.

(16) Shishkin, M.; Kresse, G. Self-consistent G W calculations for semiconductors and insulators. *Physical Review B* **2007**, *75*, 235102.

(17) Moses, P. G.; Miao, M.; Yan, Q.; Van de Walle, C. G. Hybrid functional investigations of band gaps and band alignments for AlN, GaN, InN, and InGaN. *The Journal of Chemical Physics* **2011**, *134*, 084703.

(18) Yan, Q.; Rinke, P.; Janotti, A.; Scheffler, M.; Van de Walle, C. G. Effects of strain on the band structure of group-III nitrides. *Physical Review B* **2014**, *90*, 125118.

(19) Kirchner-Hall, N. E.; Zhao, W.; Xiong, Y.; Timrov, I.; Dabo, I. Extensive benchmarking of DFT+ U calculations for predicting band gaps. *Applied Sciences* **2021**, *11*, 2395.

(20) Shih, B.-C.; Xue, Y.; Zhang, P.; Cohen, M. L.; Louie, S. G. Quasiparticle band gap of ZnO: High accuracy from the conventional G 0 W 0 approach. *Physical Review Letters* **2010**, *105*, 146401.

(21) Borlido, P.; Schmidt, J.; Huran, A. W.; Tran, F.; Marques, M. A.; Botti, S. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *npj Computational Materials* **2020**, *6*, 1–17.

(22) Strocov, V.; Claessen, R.; Aryasetiawan, F.; Blaha, P.; Nilsson, P. Band-and k-dependent self-energy effects in the unoccupied and occupied quasiparticle band structure of Cu. *Physical Review B* **2002**, *66*, 195104.

(23) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77*, 3865–3868.

(24) Schimka, L.; Harl, J.; Stroppa, A.; Grüneis, A.; Marsman, M.; Mittendorfer, F.; Kresse, G. Accurate surface and adsorption energies from many-body perturbation theory. *Nature Materials* **2010**, *9*, 741–744.

(25) Ferreira, L. G.; Marques, M.; Teles, L. K. Approximation to density functional theory for the calculation of band gaps of semiconductors. *Physical Review B* **2008**, *78*, 125116.

(26) Ferreira, L. G.; Marques, M.; Teles, L. K. Slater half-occupation technique revisited: the LDA-1/2 and GGA-1/2 approaches for atomic ionization energies and band gaps in semiconductors. *AIP Advances* **2011**, *1*, 032119.

(27) Tran, F.; Blaha, P. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Physical Review Letters* **2009**, *102*, 226401.

(28) Kuisma, M.; Ojanen, J.; Enkovaara, J.; Rantala, T. T. Kohn-Sham potential with discontinuity for band gap materials. *Physical Review B* **2010**, *82*, 115106.

(29) Aschebrock, T.; Kümmel, S. Ultranonlocality and accurate band gaps from a meta-generalized gradient approximation. *Physical Review Research* **2019**, *1*, 033082.

(30) Neupane, B.; Tang, H.; Nepal, N. K.; Adhikari, S.; Ruzsinszky, A. Opening band gaps of low-dimensional materials at the meta-GGA level of density functional approximations. *Physical Review Materials* **2021**, *5*, 063803.

(31) Liechtenstein, A.; Anisimov, V. I.; Zaanen, J. Density-functional theory and strong interactions: Orbital ordering in Mott-Hubbard insulators. *Physical Review B* **1995**, *52*, R5467.

(32) Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C.; Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+ U study. *Physical Review B* **1998**, *57*, 1505.

(33) Bajaj, A.; Janet, J. P.; Kulik, H. J. Communication: Recovering the flat-plane condition in electronic structure theory at semi-local DFT cost. *The Journal of Chemical Physics* **2017**, *147*, 191101.

(34) Moore, G. C.; Horton, M. K.; Ganose, A. M.; Siron, M.; Persson, K. A. High-throughput determination of Hubbard U and Hund J values for transition metal oxides via linear response formalism. *arXiv preprint arXiv:2201.04213* **2022**,

(35) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *The Journal of Chemical Physics* **2006**, *125*, 224106.

(36) Hedin, L. New method for calculating the one-particle Green's function with application to the electron-gas problem. *Physical Review* **1965**, *139*, A796.

(37) Borlido, P.; Aull, T.; Huran, A. W.; Tran, F.; Marques, M. A.; Botti, S. Large-scale

benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *Journal of Chemical Theory and Computation* **2019**, *15*, 5069–5079.

(38) Mattur, M. N.; Nagappan, N.; Rath, S.; Thomas, T., et al. Prediction of nature of band gap of perovskite oxides (ABO3) using a machine learning approach. *Journal of Materiomics* **2022**,

(39) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters* **2018**, *9*, 1668–1673.

(40) Gladkikh, V.; Kim, D. Y.; Hajibabaei, A.; Jana, A.; Myung, C. W.; Kim, K. S. Machine learning for predicting the band gaps of ABX3 perovskites from elemental properties. *The Journal of Physical Chemistry C* **2020**, *124*, 8905–8918.

(41) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-aided bandgap engineering for solar materials. *Computational Materials Science* **2014**, *83*, 185–195.

(42) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2016**, *2*, 1–7.

(43) Wang, T.; Zhang, K.; Thé, J.; Yu, H. Accurate prediction of band gap of materials using stacking machine learning model. *Computational Materials Science* **2022**, *201*, 110899.

(44) Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B* **2016**, *93*, 115104.

(45) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B.; Ramprasad, R.; Gubernatis, J.;

Lookman, T. Machine learning bandgaps of double perovskites. *Scientific Reports* **2016**, *6*, 1–10.

(46) Pilania, G.; Gubernatis, J. E.; Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science* **2017**, *129*, 156–163.

(47) Patra, A.; Batra, R.; Chandrasekaran, A.; Kim, C.; Huan, T. D.; Ramprasad, R. A multi-fidelity information-fusion approach to machine learn and predict polymer bandgap. *Computational Materials Science* **2020**, *172*, 109286.

(48) Li, W.; Wang, Z.; Xiao, X.; Zhang, Z.; Janotti, A.; Rajasekaran, S.; Medasani, B. Predicting band gaps and band-edge positions of oxide perovskites using density functional theory and machine learning. *Physical Review B* **2022**, *106*, 155156.

(49) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *Journal of Chemical Theory and Computation* **2015**, *11*, 2087–2096.

(50) Knøsgaard, N. R.; Thygesen, K. S. Representing individual electronic states for machine learning GW band structures of 2D materials. *Nature Communications* **2022**, *13*, 1–10.

(51) Glawe, H.; Sanna, A.; Gross, E.; Marques, M. A. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New Journal of Physics* **2016**, *18*, 093011.

(52) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science* **1996**, *6*, 15–50.

(53) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B* **1996**, *54*, 11169.

(54) Blöchl, P. E. Projector augmented-wave method. *Physical Review B* **1994**, *50*, 17953.

(55) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B* **1999**, *59*, 1758.

(56) Hinuma, Y.; Pizzi, G.; Kumagai, Y.; Oba, F.; Tanaka, I. Band structure diagram paths based on crystallography. *Computational Materials Science* **2017**, *128*, 140–184.

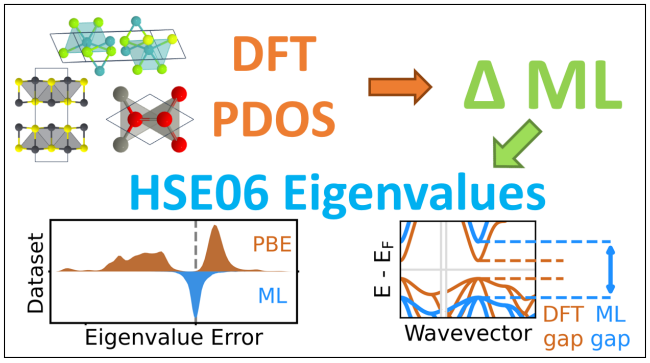
(57) Togo, A.; Tanaka, I. *Spglib*: a software library for crystal symmetry search. *arXiv preprint arXiv:1808.01590* **2018**,

(58) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(59) Henkelman, G.; Arnaldsson, A.; Jónsson, H. A fast and robust algorithm for Bader decomposition of charge density. *Computational Materials Science* **2006**, *36*, 354–360.

(60) Adhikari, S.; Bartel, C. J.; Sutton, C. Machine learning to improve the accuracy of semi-local density functionals for the prediction of formation energies. *unpublished* **2023**,

TOC Graphic


DFT
PDOS
Δ ML
HSE06 Eigenvalues
Dataset
PBE
ML
Eigenvalue Error
E - E_F
Wavevector
DFT gap
ML gap