communications
materials

ARTICLE

Check for updates

https://doi.org/10.1038/s43246-023-00373-4 OPEN

Band gap predictions of double perovskite oxides using machine learning

Anjana Talapatra[1]✉, Blas Pedro Uberuaga[1], Christopher Richard Stanek[1] & Ghanshyam Pilania[1,2]

The compositional and structural variety inherent to oxide perovskites spawn wide-ranging applications. In perovskites, the band gap E_g, a key material parameter for these applications, can be optimally controlled by varying the composition. Here, we implement a hierarchical screening process in which two cross-validated and predictive machine learning models for band gap classification and regression, trained using exhaustive datasets that span 68 elements of the periodic table, are applied sequentially. The classification model separates wide band gap materials, with $E_g \geq 0.5$ eV, from materials which have zero or relatively small band gaps, namely $E_g < 0.5$ eV, and the second regression model quantitatively predicts the gap value of the wide band gap compounds. The study down-selects 13,589 cubic oxide perovskite compositions that are predicted to be experimentally formable, thermodynamically stable, and have a wide band gap. Of these, a subset of 310 compounds, which are predicted to be stable and formable with a confidence greater than 90%, are identified for further investigation. Our models are methodically analyzed via performance metrics and inter-dependence of model features to gain physical insight into the band gap prediction problem. Design maps to identify the variation of band gap with substitution of different elements are also presented.

[1] Materials Science and Technology Division, Los Alamos National Laboratory, Los Alamos, NM 87544, USA. [2] Present address: GE Research, Schenectady, NY 12309, USA. ✉email: atalapatra@lanl.gov

The band gap (E_g), which is non-existent for a metal but is positive for semiconductors and other insulators, is a fundamental property of a periodic solid and directly influences its electrical conductivity. Based on their band gap, functional electronic materials are used in a diverse array of applications such as field effect transistors[1], LEDs[2], photovoltaics[3], and scintillators[4]. By controlling the composition or structure of the material, systematic tuning of the band gap may be achieved, allowing for materials that are tailored to the desired application. Consequently, the band gap is widely used as a screening criterion in data sets generated via high-throughput calculations for application-based discovery[5–7].

The compositional and structural complexity afforded by perovskites of the form ABX_3 in conjunction with their fascinating electrical and magnetic properties such as piezoelectricity[8], optical properties[9], high-temperature superconductivity[10], ferroelectricity[11], and magnetostrictive effects[12] make them especially attractive candidates for band gap tuning. The prototypical ABX_3 cubic perovskite structure as indicated in Fig. 1a is composed of a three-dimensional BX_3 network of corner-sharing BX_6 octahedra. The A-site cations occupy the 12 coordinate sites formed by the octahedral units and each A cation is surrounded by 12 equidistant anions. The perovskite structure can accommodate 90% of the metallic ions in the periodic table[13] with a wide variety of different anions, increasing their utility in a wide range of applications. As we will show in this work, the band gap in these materials is closely correlated with their physical and chemical properties. Thus, varying the composition by replacing any of the atoms in these structures can be used to precisely tune the physical and/or chemical properties of interest and thereby the band gap. This tunability may be amplified by increasing the types of cations occupying the A- and/or B-sites, giving rise to double perovskites with formulae $A_xA'_{2-x}B_2X_6$, $A_2B_yB'_{2-y}X_6$ and, in the most general case, $A_xA'_{2-x}B_yB'_{2-y}X_6$ for $0 \le x, y \le 2$, as shown in Fig. 1a.

In recent years, to accelerate the discovery of novel electronic materials, tremendous strides have been made by using high-throughput Density Functional Theory (DFT) techniques to determine the electronic structure of materials. However, the accuracy of the computed band gap of a material depends on the particular exchange-correlation functional employed in the DFT technique that is used. If we consider the band gap to be an excitation energy, it is naive to expect an accurate description using ground state DFT. In fact, when using DFT computations with local and semi-local exchange-correlation functionals, the errors in semiconductor and insulator band gaps can be as large as 50%[14], which is especially significant in research in fields involving semiconductors, optical and photovoltaic materials, and thermoelectrics. This underestimation of the band gap is attributed to the connate lack of derivative discontinuity[15], self-interaction error (SIE)[16,17], and delocalization error[18] within conventional DFT functionals such as the local density approximation (LDA) or the generalized gradient approximation (GGA)[19]. Consequently, much attention has focused on solving the problem of the underestimation of the DFT band gap both within the Kohn-Sham DFT formalism and outside it[20,21]. However, most of these techniques, such as the GW approximation[22] and delta self-consistent-field (ΔSCF) method[20], as well the use of hybrid functionals[23], improve the accuracy of band gap estimation at increased computational cost which make them untenable for high-throughput calculation efforts. One feasible way to circumvent this problem is to screen large datasets using low fidelity GGA-type calculations and down-select to a tractable subset of potential candidates. High-fidelity band-gap estimation techniques or experiments can then be used to accurately determine the electronic structure of this smaller subset.

Machine learning (ML) has enjoyed much popularity in the field of materials science and condensed matter physics as an efficient tool to predict a physical property or quantity, particularly when the target property cannot be directly determined without the use of extensive resources, either experimental or computational[24–28]. In the last two decades, statistical learning frameworks in conjunction with regression techniques have been widely used to predict band gaps of large datasets in order to overcome the band gap estimation problem. Earliest among these was the work of Gu et al.[29] who used experimental band gaps of 25 binary compounds and 31 ternary compounds to construct support vector regression (SVR) and artificial neural network (ANN) models, using elemental predictors. Since then, SVR has featured prominently as the regression method of choice in band gap prediction[30–33] in addition to ordinary least squares regression (OLSR)[30,34–36] and least absolute shrinkage and selection operator (LASSO)[30,35,36] methods. Pilania et al.[37] presented a general formalism to discover decision rules that can be used to make ultra-fast, yet accurate, property predictions. They used ML to establish a mapping between chemo-structural fingerprints and the electronic charge density distribution of polymeric insulators

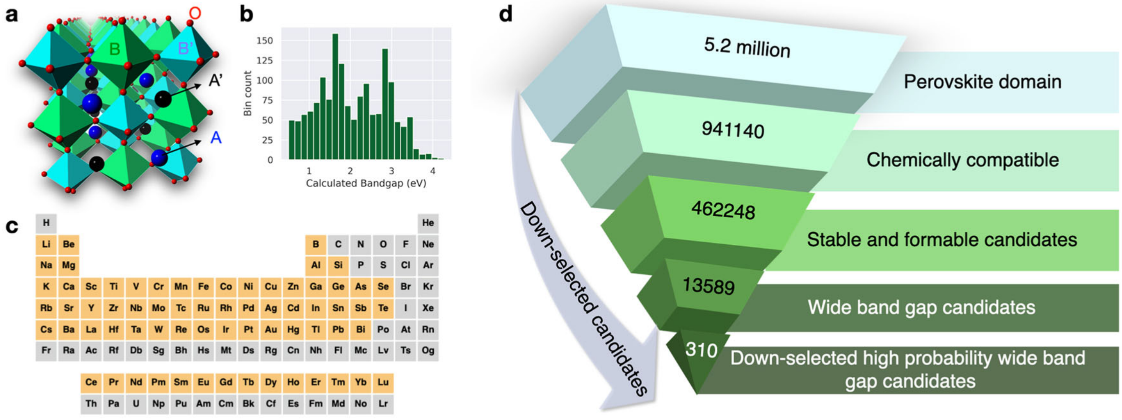
a
O
B
B'
A'
A
b
150
125
100
75
50
25
0
Bin count
1
2
3
4
Calculated Bandgap (eV)
c
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba La Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra Ac Rf Db Sg Bh Hs Mt Ds Rg Cn Nh Fl Mc Lv Ts Og
Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr
d
5.2 million
Perovskite domain
941140
Chemically compatible
462248
Stable and formable candidates
13589
Wide band gap candidates
310
Down-selected high probability wide band gap candidates
Down-selected candidates

Fig. 1 Hierarchical discovery framework for wide band gap double perovskites. **a** Double perovskite crystal structure with rocksalt ordering of both A- and B-site cations. **b** Histogram of DFT-calculated band gaps in the training dataset of oxide perovskite used in the development of the ML model presented in this work. **c** Chemical space of the perovskite oxides explored in the present study. Cations appearing at the A-site and/or the B-site are highlighted. **d** Hierarchical down-selection framework implemented in this work. Starting with more than 5 million potential chemistries, through a series of ML models we identify about 300 double perovskites that are likely to exhibit a wide band gap.

and their properties, including the band gap. In 2016, they demonstrated a robust learning framework for efficient and accurate predictions of electronic band gaps of double perovskites using a systematic feature-engineering approach[38]. They also implemented a multi-fidelity Gaussian process (GP)-based co-Kriging regression model to predict E_g for the class of elpasolites[39]. Recently, Na et al.[40] and Li et al.[41] reported multi-fidelity band gap prediction models based on graph neural networks (GNNs) using that can use band gaps data computed and/or measured at varying levels of fidelity to provide best estimates at the hhighest level of fidelity. Omprakash et al.[42] also used GNNs trained on experimental data gathered from literature to predict band gaps for a variety of different 2D, 3D, organometal and inorganic inorganic single perovskites. In a departure from traditional targets of high accuracy, Gladkikh et al.[43] used alternating conditional expectations (ACE), a ML technique suitable for small data sets which performs worse than more commonly used ML methods but presents its results in a graphic form, helping in interpretation, the lack of which is a core critique of ML applications today[44]. Kauwe et al.[45] used the example property of the band gap to demonstrate an ensemble learning approach which allows the efficient modeling of experimental data by combining models trained on otherwise disparate computational and experimental data. Most of these approaches favor the use of elemental predictors, lending credence to the idea of physics-informed models. Recently, Stanley et al.[33] used formation energy as a predictor to develop highly accurate band gap prediction models for ABO_3 perovskites. However, if intended to be applied to large prediction datasets, the use of properties which are relatively inexpensive to compute as model predictors can still be prohibitive.

Thus we see that prior research in the area has focused on exploring the potential of using different ML techniques to predict the band gap in organic as well as inorganic materials. While considerable work has been done on predicting the band gap in perovskites, and oxide perovskites in particular, these have either focused on single perovskites[32,43,46–48] or only small datasets of double perovskites[38,49,50], which limits the generalization of these models beyond certain chemistries and warrants more work to extend the models' applicabilities. Further, these works often demonstrate the accuracy of their models but do not take the next step of predicting new compounds that satisfy design criteria.

In this work, we make use of a ML-based screening framework, employing the low-fidelity semi-local exchange-correlation Perdew–Burke–Ernzerhof (PBE) flavor[51] of the GGA functional, to down-select a tractable number of promising compounds from large candidate datasets. Specifically, we investigate a previously identified[52] exhaustive chemical space of formable and stable cubic single and double oxide perovskites. These predictions of formability and stability have been independently validated by other researchers who have synthesized some of the compositions that were originally predicted in our past work[53,54]. Starting from these, we first identify novel cubic compositions that are likely to be have relatively wide band gaps ($E_g \geq 0.5$ eV) and then predict their DFT band gaps with high accuracy using ML models based on Random Forests (RF)[55]. We use this two-step strategy of separating the materials which have or are predicted to have no band gap or a very small band gap (within the precision limits of PBE calculated band gaps) from the materials which have a significant band gap. We use a threshold value of 0.5 eV for the sake of illustration. In principle, the framework may be modified to account for any application-specific cutoff or a targeted range. We employ a very large dataset of DFT-based band gaps that are generated in-house for training and implement a hierarchical screening process, wherein we build two ML models independently, the first being a classification model to separate wide band gap materials from narrow band gap materials, and, the second, a regression model that quantitatively predicts the band gap of designated wide band gap materials. The trained models are then applied sequentially to our large candidate dataset. We intentionally do this to screen out a large number of materials which have a vanishing band gap to avoid biasing the subsequent regression model. To implement this scheme, we use a threshold band gap of 0.5 eV to demarcate narrow ($E_g < 0.5$ eV) and wide ($E_g \geq 0.5$ eV) band gap materials. Thus, the regression model is applied to only those candidates that are likely to have a band gap greater than 0.5 eV as determined by the classification model. It is to be noted that the terminology of narrow and wide band gap materials used in this manuscript to differentiate between the two material classes is distinct from the similar terminology used with reference to semiconductor materials. A histogram of DFT-calculated band gaps used in the development of the ML model presented in this work and the associated chemical space are shown in Fig. 1b and c, respectively. In Fig. 1d, the number of down-selected candidates are shown at each step, details of which are discussed further in the manuscript. Since we focus on a very large chemical space, the developed models are deemed generalizable to the entire space of perovskites and double perovskite oxides, potentially containing millions of compounds.

Results

Hierarchical Down-selection workflow. We note that when we refer to oxide perovskite structures used for model training as well as the novel perovskite compositions that are predicted to have a wide band gap, we only consider the cubic variants. In our previous work[52], we consider the formability and stability of oxide perovskites, and in building our models, use a stability criterion of energy above hull (Ec) ≤50 meV/atom. We showed that cubic thermodynamic stability is a very conservative criterion for perovskite synthesizability, and in many cases, it is also possible to further reduce the total energy (leading to an increased thermodynamic stability) by relaxing to several lower energy phases (e.g., tetragonal, orthorhombic, rhombohedral etc.) that are commonly found in perovskites. Therefore, if a compound is predicted to be stable in a cubic symmetry, it is only going to be more stable in these other possible reduced symmetry phases. Furthermore, these stabilizing local distortions are also known to slightly widen the bandgap in perovskites. Accounting for all the possible reduced symmetry configurations for a given double perovskite composition can lead to several tens of lower symmetry phases. Therefore, this assumption of restricting ourselves to only cubic symmetry allows us to keep the number of computations in this study limited to a practically-feasible level. However, a downside of this assumption is possible omission of some promising compounds whose PBE bandgaps are greater than 0.5 eV in the lower symmetry phases but less than 0.5 eV in the cubic phase.

The novelty of this complete hierarchical framework is multi-fold in that no prior work exists, to the best of our knowledge, in which multiple ML models have been used to connect the stability, formability, insulating nature, and band gap for such a large perovskite oxide chemical space. We consider all possible double oxide perovskite combinations for 68 elements from the periodic table, resulting in a very large chemically diverse set of candidate materials, which we contend is the largest and most diverse that has been evaluated till date. Owing to this heterogeneous and large dataset, our machine learning models are able to attain an very high prediction accuracy over such a vast chemical space (we note that in past this level of accuracy is only demonstrated on chemical spaces which exhibit a rather limited chemical diversity, eventually limiting the exploration potential of the developed surrogate models). The

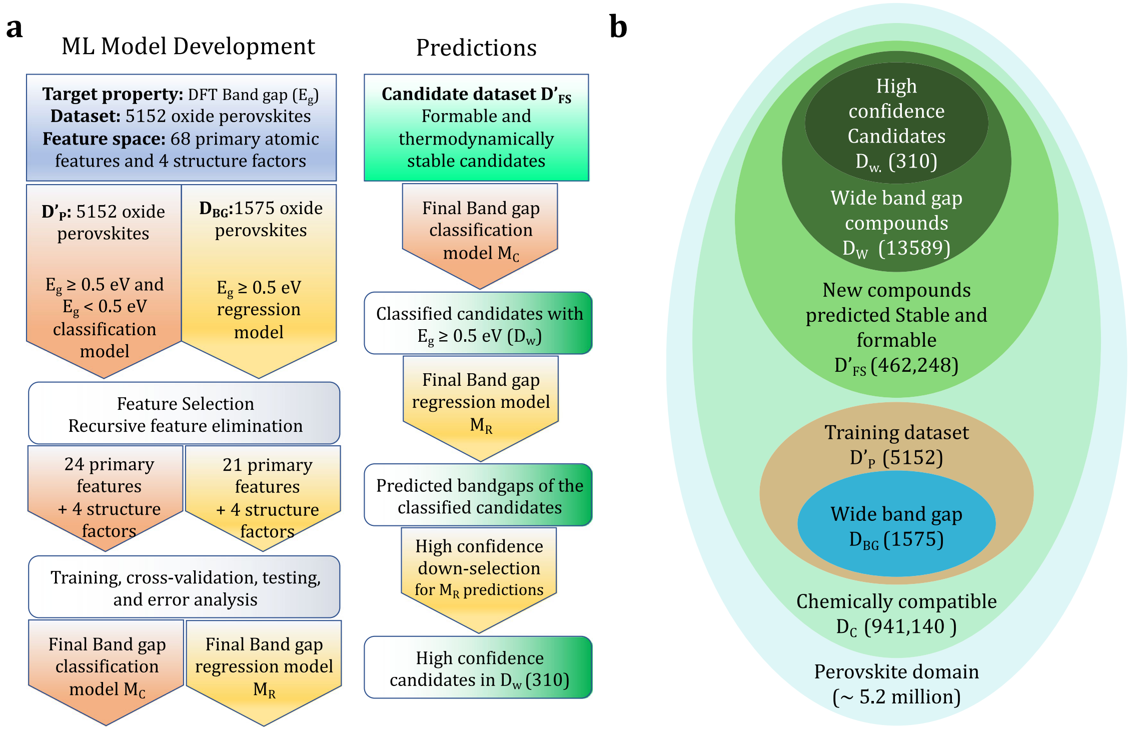
a
ML Model Development
Predictions
Target property: DFT Band gap (E_g)
Dataset: 5152 oxide perovskites
Feature space: 68 primary atomic features and 4 structure factors
D'_P: 5152 oxide perovskites
E_g ≥ 0.5 eV and E_g < 0.5 eV classification model
D_BG:1575 oxide perovskites
E_g ≥ 0.5 eV regression model
Feature Selection
Recursive feature elimination
24 primary features + 4 structure factors
21 primary features + 4 structure factors
Training, cross-validation, testing, and error analysis
Final Band gap classification model M_C
Final Band gap regression model M_R
Candidate dataset D'_FS
Formable and thermodynamically stable candidates
Final Band gap classification model M_C
Classified candidates with E_g ≥ 0.5 eV (D_w)
Final Band gap regression model M_R
Predicted bandgaps of the classified candidates
High confidence down-selection for M_R predictions
High confidence candidates in D_w (310)
b
High confidence Candidates D_w. (310)
Wide band gap compounds D_W (13589)
New compounds predicted Stable and formable D'_FS (462,248)
Training dataset D'_P (5152)
Wide band gap D_BG (1575)
Chemically compatible D_C (941,140)
Perovskite domain (~ 5.2 million)

Fig. 2 Schematic representation of workflow and associated datasets. a Computational workflow for model development and predictions. DFT calculated band gaps wserve as the foundation of two ML models, one for classification of insulators and a second regression model that predicts the band gaps of the insulating compounds. The two ML models are then applied sequentially to candidate datasets $D_F S$ and D_W respectively. **b** Venn diagram representation of training dataset and candidate datasets used in this work.

models are trained adaptively in order to achieve highly accurate and efficient predictive performance during the model building stages and are analyzed rigorously via performance metrics and inter-dependence of model features in an effort to gain physical insight into the band gap prediction problem. Our ML models allow for instant band gap predictions in the vast perovskite chemical space and screening for a variety of applications. The exhaustive design space that we explore here lends insight into design rules and dopant selection for band gap and band edge engineering[56].

Figure 2a shows an overview of the model building and model application workflows adopted in this work. First, to build the ML models, DFT is used to compute the band gaps of more that 5000 materials to compile a training dataset. These materials are then classified as having either a narrow or wide band gap; in this work, a threshold value of 0.5 eV is used to separate them. Note that this value is chosen for the sake of illustration only and, in principle, any cutoff value depending on a target application can be implemented within the workflow. Simultaneously, the training descriptors are generated for this dataset and the descriptors along with the band gaps are used to build two ML models: i) a wide/narrow band gap classification model (M_C) trained on both wide and narrow band gap data and ii) a band gap regression model (M_R) trained on only wide band gap data. Then, the models are applied to the large chemical space of potentially formable and thermodynamically stable perovskite materials sequentially to first identify wide band gap candidates via M_C, and then predict the band gap of those wide band gap candidates using M_R.

Training data, prediction data and descriptors. To build training and candidate datasets, the 68 elements highlighted in the periodic table in Fig. 1c were considered and all possible ABO_3, $A_2BB'O_6$, $AA'B_2O_6$ and $AA'BB'O_6$ compounds that could be formed by substituting them at the A- and B-sites were enumerated. Considering all possible combinations and accounting for charge neutrality, this resulted in a set of 946,292 unique single and double perovskite compositions (some of which have multiple valence combinations). From these, 5152 oxide perovskite compounds were adaptively selected to form the training dataset D_P and the structures were optimized (while constrained to remain cubic) and their band gaps were calculated with DFT. Further technical details of our DFT calculations are provided in the methods section. Initially, all experimentally known oxide perovskites were calculated to form the training dataset. The initial classification model predictions of wide and narrow band gap materials and their corresponding band gaps were then used to adaptively augment this training dataset.

In total, 5152 compounds were evaluated to create a robust training dataset D_P for the wide/narrow band gap classification model. Structures with calculated band gaps equal to or greater than 0.5 eV were labeled as wide band gap materials (insulators) while those with band gaps less than 0.5 eV were labeled as narrow band gap materials. Applying this 0.5 eV threshold criterion, of the 5152 calculated perovskites, 1575 (i.e. about 30%) were found to have a wide band gap, while the remaining 3577 were found to have a narrow band gap. These 1575 structures encompass the training dataset D_{BG} for the band gap prediction regression model and the distribution of their band gaps is shown in Fig. 1b. As is self-evident, D_{BG} is a subset of D_P. The various training datasets used and referenced in this work are listed in Table 1 and are also represented in a Venn diagram in Fig. 2b. The complete training dataset is included in Supplementary Data 1.

Candidate dataset for prediction. The exhaustive dataset of 946,292 unique single and double perovskite compositions less the 5152 compounds that comprise the training dataset, results in the foundational chemically compatible candidate dataset D_C comprising 941,140 perovskite oxide compounds. ML models described in[52]

Table 1 Definition of training and candidate oxide perovskite datasets used in this work.

Abbreviation	Size	Description
Training datasets		
D_P	5152	Adaptively built training dataset for band gap classification model used in this work
D_{BG}	1575	Training dataset of wide band gap materials used for band gap regression model
Prediction datasets		
D_C	941,140	Chemically compatible oxide perovskite candidates less training dataset D_P
D_{FS}	462,248	Chemically compatible, formable and thermodynamically stable candidates (subset of D_C)
D_W	13,589	Subset of D_{FS} with candidates predicted to have wide band gaps ($E_g > 0.5$ eV)
D_w	310	Subset of D_W, selected high probability (>90%) wide band gap candidates

Table 2 List of descriptors used in the ML models.

Abbreviation	Feature
Elemental	
HOMO	Highest Occupied Molecular Orbital (eV)
LUMO	Lowest Unoccupied Molecular Orbital (eV)
IE	Ionization energy (kJ/mol)
X	Pauling Electronegativity
Z radius	Zunger's Pseudopotential radius (a.u.)
EA	Electron affinity (kJ/mol)
Geometric	
t	Tolerance factor
μ	Octahedral factor
$\bar{\mu}$	Mismatch factor

that predict formability and theromodynamic stability of oxide perovskites were then applied to D_C, to identify the subset D_{FS} of 462,248 oxide perovskites (using a cutoff probability of 0.5) that are predicted to be formable and thermodynamically stable. This dataset D_{FS} is then our candidate dataset to which we apply the classification model M_C and band gap regression model M_R built in this work sequentially to identify the oxide perovskites that are likely to have a wide band gap and then predict the actual band gaps. It is worth repeating that in order to maintain zero overlap between the training and candidate datasets, all compounds that comprise the training dataset were removed from the D_C dataset and hence do not feature in the D_{FS} database. Additionally, as mentioned earlier, all experimentally known oxide perovskites are part of the training dataset. Thus the compounds in the candidate dataset D_{FS} are all perovskite oxide compositions that have never been experimentally synthesized to the best of our knowledge. The nested venn diagram in in Fig. 2b delineates the relationship between the training dataset, the candidate dataset and the sets of down-selected candidates at each step in our hierarchical strategy.

Descriptors for machine-learning. In this work, we use a combination of geometric and atomic descriptors to train the ML models. As far as possible, the intention was to use the simplest possible inputs and minimize computational overhead necessary to compile inputs while achieving high prediction accuracies by effectively incorporating notions of chemical similarity across different chemistries. To this end, elemental properties corresponding to the A- and B-site cations were used as descriptors for the single perovskites and their symmetric and anti-symmetric compound variants were used for the double perovskites. For a double perovskite $AA'BB'O_6$, for a given property P, the symmetric compound feature is calculated as $P^{A+} = (P_A + P'_A)/2$ and the anti-symmetric compound feature as $P^{A-} = |P_A - P'_A|/2$ for the A-site, where P_A and P'_A are the elemental properties of A and A'; similar descriptors are defined for the B-site. Such compound descriptors have been previously adopted by the authors[38,52,57] and others[58] and are an effective and well-known technique to account for the multiple cations in a single site scenario. We intentionally do not include compound properties such as total energy, formation energy or electronic charge densities as descriptors even though they may be calculated relatively easily since the goal is to create robust ML models which may be universally applied to large prediction datasets.

All descriptors were normalized to ensure zero mean and standard deviation of unity. In the preliminary stages of model building, we started out with a very comprehensive list of 68 structural and chemical properties that may be indicators of the insulator nature of the perovskites and the band gap. Using Pearson correlation[59] values as a first step screening process and then using the recursive feature elimination (RFE)[60] technique during model development using the open-source Scikit-learn[61] python package, the least important descriptors were pruned and the final set of relevant descriptors were identified. RFE works to select features by recursively removing those features which exhibit the smallest weight that are assigned by an extra trees classifier. The estimator is first trained on the initial set of features and the importance of each feature is determined. The most important features are then retained from the current set of features. This procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.

The final set of atom-specific descriptors include the Highest Occupied Molecular Orbital (HOMO) and Lowest Unoccupied Molecular Orbital (LUMO) energies[62], ionization energy (IE), electro-negativity[63], Zunger's pseudopotential radius (Z radius), and electron affinity (EA). For the Zunger's pseudopotential radius the sum of the radii for the s and p orbitals was used. As mentioned earlier, symmetric and anti-symmetric compound descriptors for the A- and B-sites were used for these 6 atom-specific descriptors resulting in 24 compound descriptors. We further included perovskite-specific geometric descriptors – the Goldschmidt tolerance factor (t)[64], octahedral factor (μ)[65], and mismatch factors[52,66] for the A-site ($\bar{\mu_A}$) and B-sites ($\bar{\mu_B}$) which are defined as:

$$\mu = \frac{r_B}{r_O}, \tag{1}$$

$$t = \frac{r_A + r_O}{\sqrt{2}(r_B + r_O)}, \tag{2}$$

$$\bar{\mu_A} = |r_A - r_{A'}|/2r_O, \tag{3}$$

$$\bar{\mu_B} = |r_B - r_{B'}|/2r_O, \tag{4}$$

where $r_A, r'_A\, r_B, r'_B$ and r_O are the coordination dependent Shannon's ionic radii[67] of the A-site cations, the octahedrally coordinated B cations and the oxygen anion respectively. Including these four geometric descriptors, we were left with a total of 28 descriptors. Using RFE, for the insulator classification model, it was found that all 28 descriptors are important, while for the band gap regression model, 21 atom-specific descriptors

Table 3 Performance metrics for M_C averaged over 100 different runs for varying percentages of training training/test data splits.

Training % Data	Training			Test		
	Accuracy	Precision	Recall	Precision	Recall	Accuracy
50	0.88 ∓ 0.02	0.90 ∓ 0.02	0.70 ∓ 0.05	0.89 ∓ 0.01	0.91 ∓ 0.01	0.74 ∓ 0.01
60	0.89 ∓ 0.02	0.91 ∓ 0.02	0.72 ∓ 0.05	0.90 ∓ 0.02	0.91 ∓ 0.03	0.77 ∓ 0.04
70	0.90 ∓ 0.01	0.92 ∓ 0.01	0.75 ∓ 0.03	0.91 ∓ 0.02	0.89 ∓ 0.04	0.78 ∓ 0.07
80	0.91 ∓ 0.01	0.93 ∓ 0.01	0.77 ∓ 0.03	0.91 ∓ 0.04	0.88 ∓ 0.05	0.79 ∓ 0.09
90	0.91 ∓ 0.01	0.94 ∓ 0.01	0.79 ∓ 0.04	0.95 ∓ 0.02	0.93 ∓ 0.03	0.87 ∓ 0.03

Also listed are standard deviations which are computed as the square root of the estimated variance of the metrics obtained over the 100 runs.

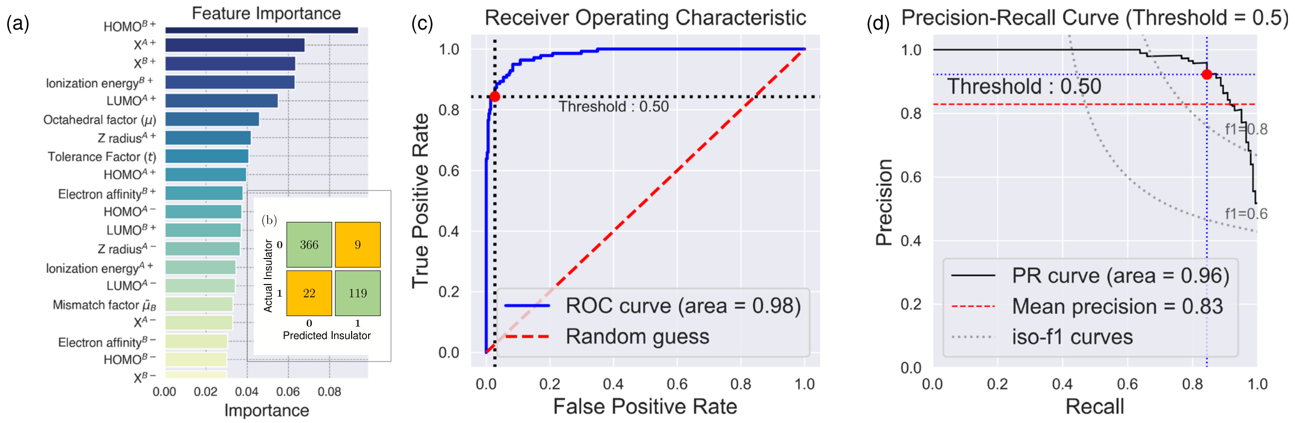

Fig. 3 Results for perovskite insulator character using M_C. **a** Feature importance plot for all the descriptors with significant values, **b** Confusion matrix, **c** Receiver operating characteristic (ROC) curves, and **d** Precision-recall curves of the cross-validated RF classification on test data using M_C.

and all the geometric descriptors were relevant. The specific descriptors identified for both models are indicated in Table 2. Values of these 28 descriptors for the D'_P database are provided in Supplementary Data 1. The dataset also includes the DFT-calculated band gap of the compounds. The HOMO and LUMO Kohn-Sham levels of isolated atoms were computed using DFT without spin polarization, in a large orthorhombic supercell (15 Å × 14.5 Å × 14 Å, to break the cubic symmetry) with respect to the vacuum level[38,68].

The two ML models M_C and M_R were built using the descriptors listed in Table 2 and the respective training databases D'_P and D_{BG}. To gauge the robustness of the models and to identify the minimum significant value for the feature importance, both models were also tested using random descriptors. The models were then tested on the test datasets created using the 90/10 shuffle split as described previously. Subsequent to the training and testing phases, the models were applied to the D_{FS} dataset of predicted stable and formable perovskites. In the following sub-sections, we will discuss the models and their performance in detail. Technical details regarding the ML models used may be found in the methods section.

Wide/narrow band gap classification. The accuracy, precision, and recall metrics of M_C were evaluated using five-fold cross validation for different combinations of training/test splits; the values are shown in Table 3. For the final model, we used a 90/10 training/test split and achieved a model accuracy of 0.94 and 0.95 and model precision of 0.95 and 0.93 on the training and test sets, respectively.

In Fig. 3a, we present the 24 most important descriptors for the perovskite insulator classification problem. It is seen that the symmetric HOMO energy, electronegativity, and ionization energy for the B-site are the most important descriptors, followed by the symmetric electronegativity and LUMO energy for the A-site, in differentiating between compounds that are likely to have a wide or narrow band gap. Figure 3b shows the average test confusion matrix for 100 runs of the model. The off-diagonals (in red) indicate the False Positives (FP) and False Negatives (FN) while the diagonal elements (in green) indicate the True Positives (TP) and True Negatives (TN).

In Fig. 3c, we plot the receiver operating characteristic (ROC)[69] curves of the cross-validated classification model M_C on the test data. The closer the curve is to the ideal point (top left corner on the plot), the greater the area under the curve (AUC), and the better the performance of the classifier. The plot emphasizes the value of the curve for a threshold value of 0.5, i.e., compounds having a probability of ≥0.5 of being an insulator are classified as wide band gap materials while those with a probability <0.5 are classified as narrow band gap materials. For this classifier, we achieve a very good AUC value of 0.98. Figure 3(d) showcases the precision-recall (PR) curves for the classifier, which is a plot of the precision rate on the ordinate and the recall rate on the abcissa for varying threshold values. The point (1,1) on this plot implies a perfect predictive model, and the closer the curve is to this point, the more robust is the model. As is evident, the PR curve for a threshold value of 0.5, is very close to the ideal (1,1) point yielding an AUC of 96, resulting in a very skillful model. The $f1$-score is the harmonic mean of the precision and recall rates for a particular probability threshold value. On the PR curve, we indicate the iso-curves for various $f1$-score values.

Band gap regression. As mentioned in prior sections, the band gap prediction model was continually improved by adaptively augmenting the size of the training dataset until the MAE for the model on the test data was below 0.2 eV. The regression coefficient (R^2) and MAE values for different combinations of training/

test data splits evaluated using five-fold cross validation are shown in Table 4. The band gap prediction model results are shown in Fig. 4. To visualize the accuracy of our band gap predictions, the machine-learning predictions of the band gap as a function of the calculated band gap for the training and test data are shown in Fig. 4a and b, respectively. Regression coefficient (R^2) values of 0.97 and 0.86 and MAE values of 0.07 eV and 0.18 eV were achieved for the training and test data, respectively, averaged over 100 different runs, each with randomly selected 90% training and 10% test sets. These values compare very favorably with and are at least as good as previously reported band gap prediction models[38–40] which are trained on much smaller datasets.

Figure 4c, d shows the learning curves for the regression model for the R^2 and MAE performance metrics. Learning curves are a widely used diagnostic tool in ML algorithms to indicate the incremental improvement in learning with respect to an increase in the size of the training dataset. The model is evaluated for a chosen evaluation metric on the training dataset and on a hold-out validation dataset after each update during training; the measured performance when plotted comprise the learning curves. Here, we plot the learning curves for our chosen metrics of R^2 (Fig. 4c and the MAE (Fig. 4d). In both these plots, we see that the test curve follows the trajectory of the training curve with a moderate gap between them - indicating that our model exhibits low bias and moderate variance, as is desired.

Table 4 Performance metrics (regression coefficient R^2 and MAE in eV) for M_{R}, averaged over 100 different runs for different combinations of training and test set sizes.

Training % Data	Training		Test	
	R^2	MAE	R^2	MAE
50	0.96 ∓ 0.04	0.09 ∓ 0.04	0.71 ∓ 0.04	0.29 ∓ 0.04
60	0.97 ∓ 0.06	0.09 ∓ 0.04	0.74 ∓ 0.06	0.27 ∓ 0.04
70	0.97 ∓ 0.04	0.08 ∓ 0.04	0.77 ∓ 0.04	0.24 ∓ 0.04
80	0.97 ∓ 0.02	0.08 ∓ 0.02	0.80 ∓ 0.02	0.23 ∓ 0.04
90	0.97 ∓ 0.02	0.07 ∓ 0.01	0.86 ∓ 0.02	0.18 ∓ 0.02

Also listed are standard deviations which are computed as the square root of the estimated variance of the metrics obtained over the 100 runs.

Figure 4e indicates the regression feature importances. Here, we see that the geometric factors $(t, \bar{\mu_B})$ feature very low on the importance list while descriptors such as the electronegativities, HOMO and LUMO energies, Zunger's pseudopotential radius, and ooctahedral factor (μ) prominently affect the band gap. Overall, We see that the features corresponding to the B-site appear more predominantly. This is expected since, in general, we observe (as indicated in the SI) that for a given oxide perovskite altering the A-site cation results in a smaller band gap variation as compared to an alteration in the B-site cation. This observation may be further attributed to the ionic nature of the perovskites, wherein the valence band maximum (VBM) states are predominantly oxygen p states while the conduction band minimum (CBM) states are predominantly B element d states. Thus, if keeping the anion chemistry constant, one would expect the nature of the B cations to have the greatest effect on the band gap.

To gain additional insight into the interplay of the model, the model's features, and their effect on the band gap, we also calculate the two dimensional (binary) partial dependence for the most important features identified. Partial dependence plots (PDP)[70] which are calculated after the model has been fit to the data, show the marginal effect a descriptor has on the predicted outcome of the ML model. PDPs are useful in determining whether the relationship between the target property and the descriptors is linear, monotonic, or more complex. This is done by varying the value of the descriptor of interest and using the model to predict the target value for multiple instances; the average target value is then plotted as a function of the descriptor. Thus, the partial dependence method considers all instances and gives a statement about the global relationship of a feature with the predicted outcome. While the single feature PDPs capture the average trend of a given feature with the target property, the binary PDPs visualize the partial dependence of two descriptors simultaneously. Note that

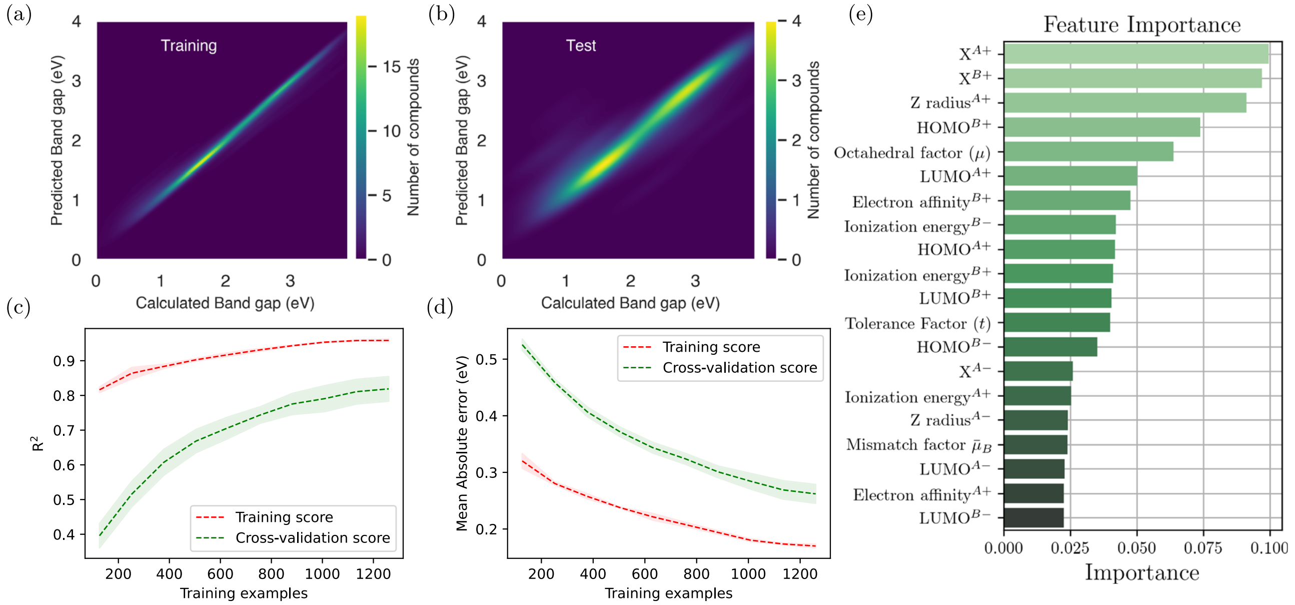

Fig. 4 Results for prediction of band gaps of oxide perovskites using regression model M_R. a Machine-learning predictions for training data and (**b**) machine-learning predictions for test data. Learning curves representing the variation in (**c**) regression coefficient (R^2) and (**d**) MAE in eV with increase in size of training dataset. Each step represents the average R^2 and MAE of 100 iterations of shuffle split cross-validation with 10% of the data for validation. The shaded regions indicate the averaged standard deviation. **e** Feature importance plot for all the descriptors with significant values.

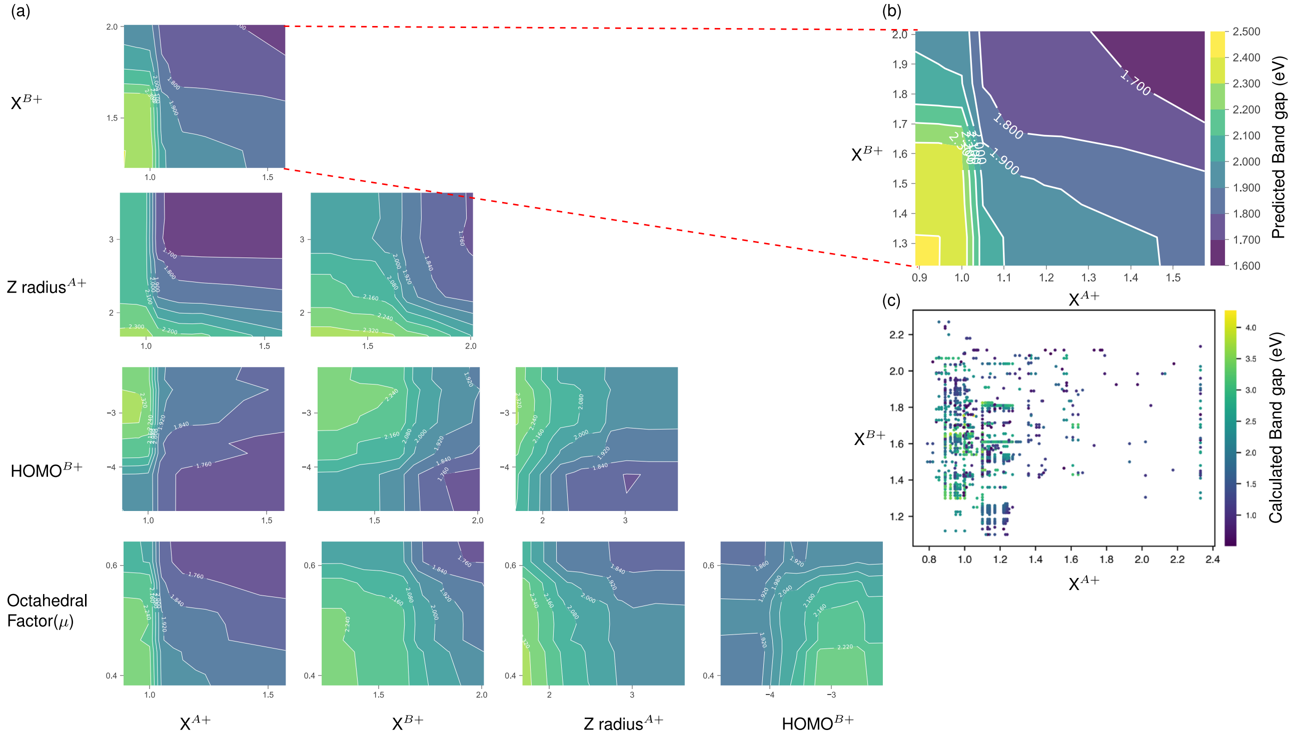

Fig. 5 Partial dependence plots (PDPs) for model features. a PDPs for the five top ranked features of the band gap prediction regression model. The units of the features used here are consistent with those reported in the Table 2. **b** Expanded PDP for the top two features i.e., X^{B+} and X^{A+}, and **c** the marginalized DFT calculated band gaps for varying values of X^{B+} on the ordinate and X^{A+} on the abcissa.

partial dependence works by marginalizing the machine learning model output over the distribution of the remaining features, so that the function shows the relationship between the features we are interested in and the predicted outcome. Marginalization is a method that requires summing over the possible values of one variable to determine the marginal contribution of another. By marginalizing over the other features, we get a function that depends only on the features of interest.

Figure 5a shows the binary PDPs for the top 5 descriptors, in which the yellow regions indicate the largest values of predicted band gaps. Figure 5b shows the expanded PDP for the top two features i.e., X^{B+} and X^{A+}. This plot indicates that the maximum values of band gaps may be found for approximate value ranges of 0 > X^{B+} > 1.6 and 0 > X^{A+} > 1.0. Such feature value ranges may be obtained for various feature pairs, or even single features if one refers to unary PDPs. Figure 5c shows the corresponding scatter plot for the training data, i.e. the marginalized DFT calculated band gaps for varying values of X^{B+} on the ordinate and X^{A+} on the abcissa. On comparison of Fig. 5b, c, we see that the training data does not include a significant number of data points in the corresponding high band gap region predicted by the PDP. Thus, while PDPs may be used to delineate ranges of feature values to obtain specific ranges of band gaps, it is imperative to keep in mind the limitation that the predictions are averaged values over the remaining features and the training data rarely cover the entire targeted feature ranges in a uniform manner.

The statistical nature of predictions obtained using ML is concomitant with uncertainties in the predictions themselves. Hence, while it is important to ensure high prediction accuracy while building ML models, it is equally if not more important to explain a prediction, and the confidence that we have in the prediction made by a model. Also, predictions are interpolative for data points that lie within the range of previously seen data, and extrapolative for data points that fall beyond the range of previously seen data. The uncertainty arising due to this distance from the domain of the training data needs to be quantified and hence confidence intervals for the test data using our band gap regression model were calculated and are and shown in Fig. 6a. Here, we see that for a large number of data points, the error bars cross the prediction-equals-calculation diagonal, indicating the lack of residual noise in the predicted band gaps and that our features that constitute the regression model adequately describe the oxide perovskite band gaps. The variation in the lengths of the error bars indicates that our model is more confident about some predictions than others. Reassuringly, we observe that, in general, the model is more confident about those predictions for which the predicted and calculated band gaps are similar.

Prediction of new oxide perovskites with wide band gaps. Subsequent to the training and testing phases, the wide/narrow band gap classification and band gap regression models were applied to the dataset D_{FS} of predicted formable and stable oxide perovskites to predict new wide band gap materials and their band gaps respectively. This prediction dataset D_{FS} consists of 462,248 distinct compounds. Using a 50% probability cutoff, this was reduced to 13,589 (<3%) compounds that were predicted to possess a wide band gap; this set of compounds is designated D_W. The distribution of the wide band gap prediction probabilities for these 13,589 compounds is shown in Fig. 6b. As discussed earlier, the candidate compounds do not contain any experimentally known perovskites, and hence these 13,589 oxide perovskite compositions are novel compounds that have never been synthesized to the best of our knowledge. To down-select a tractable number of wide band gap candidates for further exploration, from these 13,589 compounds, we retained those compounds (D_w) for which the prediction probabilities of experimental formability, thermodynamic stability, and wide band gap nature were greater than 90% which amounted to 310 compounds. The inset in Fig. 6b reflects the wide band gap prediction probabilities

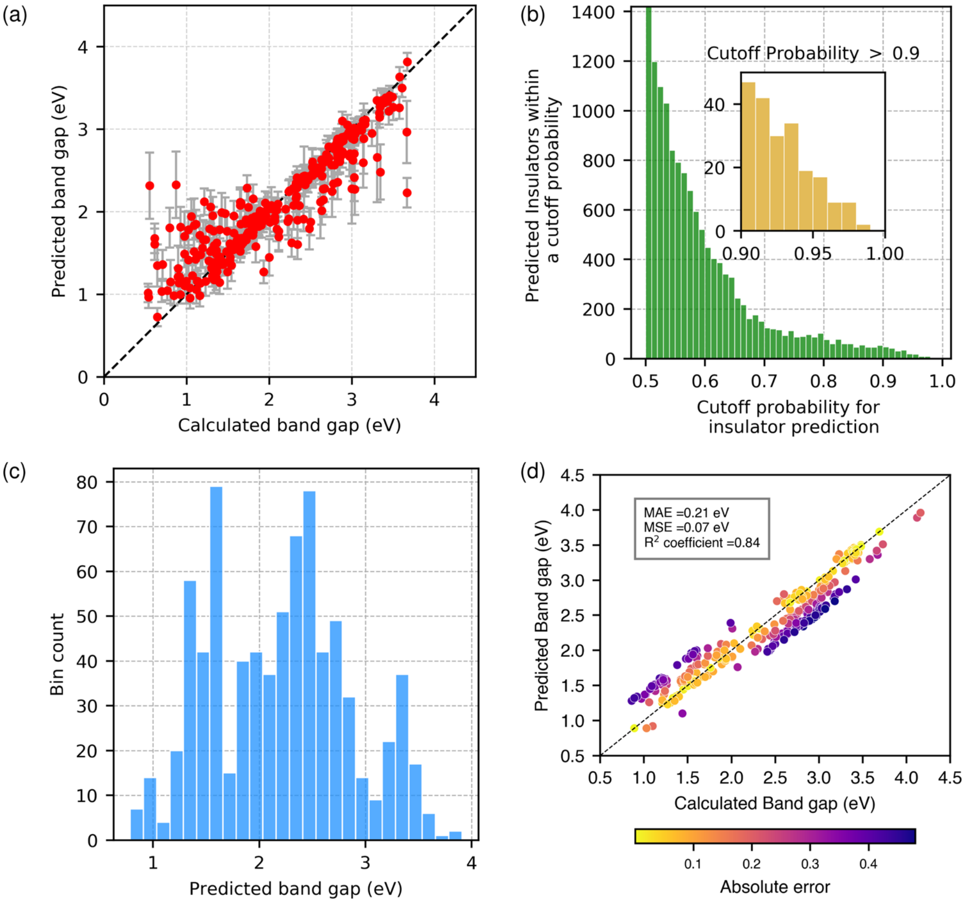

Fig. 6 Band gap prediction results. a Confidence intervals for test data for the band gap prediction regression. **b** Histogram of the predicted probabilities of candidate compounds that are likely to be wide band gap materials, binned by 0.01. The inset indicates the zoomed-in histogram for for the 310 oxide perovskites with a predicted wide band gap probability between 0.9 and 1. **c** Histogram of predicted band gaps for the 310 oxide perovskites, binned in 0.125 eV intervals. **d** Parity plot of calculated vs. predicted band gaps for 310 down-selected candidates.

for these 310 compounds. For these 310 compounds, we then predicted the band gaps using M_R, the distribution of which is shown in Fig. 6c. The 310 compounds are listed in Table 5 and complete descriptions of these compounds along with their predicted band gaps are included in Supplementary Data 2.

The DFT band gaps of these 310 compounds were calculated in an effort to computationally confirm our predictions, as shown in Fig. 6d. Calculations showed that all of these 310 compounds are indeed wide band gap materials as defined in this work, i.e., all their band gaps were greater than 0.5 eV, thus proving that our insulator classification model is indeed very accurate. Further, the calculated band gaps agree very well with the predicted band gaps, with a maximum error of 0.48 eV, a MAE of 0.21 eV, a mean square error (MSE) of 0.07 and a R^2 value of 0.84. Absolute difference in predicted and calculated band gaps for these compounds are also denoted in eV in parentheses in Table 5. In these calculations shown in Fig. 6d, we examine 31 distinct *AA'* pairs. For a given *AA'* pair, the band gap shows wide variation with variation in the *BB'* pair. For example, for $A = A' =$ Ba, we examine 32 distinct *BB'* pairs and find that the calculated band gap varies from 1.57–4.1 eV. In Supplementary Note 1 and Supplementary Note 2 we include detailed heat maps for different combinations of *AA'* and *BB'*.

Discussion

As mentioned in prior sections, with a 50% probability cutoff, 13,589 compounds were predicted to be wide band gap materials, i.e., materials with a DFT band gap greater than 0.5 eV. These 13,589 compounds are double perovskites with $A_2BB'O_6$, $AA'B_2O_6$ and $AA'BB'O_6$ compositions. Thus our predictions enable us to compile design maps which can give an insight into the band gap variation and associated trends of a basic ABO_3 structure when another element (A'/B') replaces half of the *A*-site, *B*-site or *A*- and *B*-site cations. Figure 7 indicates one such design map for the $Ba_2BB'O_6$ for selected formable and stable combinations of *B* and *B'*. The white squares indicate combinations which are predicted to be narrow band gap materials ($E_g < 0.5$ eV) and hence do not have an accompanying band gap prediction. Elements are arranged on the x- and y-axes in order of increasing atomic number (left to right and bottom to top respectively). Here, we see that Ta, Sb, In, La on the *B*-site result in larger band gaps across the board, with the specific combination of Ta and In resulting in the largest band gap of 3.9 eV. On the other hand, the presence of Bi on the *B*-site always lowers the band gap. Supplementary Note 3 includes Design maps of predicted band gaps with respect to average values of the five top ranked features individually and pairwise are included in Supplementary Note 4 and Supplementary Note 5 respectively. Average values of variation in band gap for base single perovskite oxides with addition of element in A- and B-sublattices are shown in Supplementary Note 5.

Thus, a hierarchical screening strategy (summarized in Fig. 1d is presented to identify 310 novel formable, stable double oxide perovskites exhibiting wide band gaps via a ML-guided exploration of a very large fraction of the double perovskite chemical space. We employ a large training dataset spanning a vast chemical space to

Table 5 List of 310 predicted wide band gap oxide perovskite compounds satisfying cutoff probabilities >0.9 for stability, formability and wide/narrow band gap prediction models.

Predicted novel oxide perovskite candidates			
Ba_2BiErO_6 (0.09)	Ba_2BiHoO_6 (0.08)	Ba_2BiLaO_6 (0.08)	Ba_2BiNdO_6 (0.04)
Ba_2BiScO_6 (0.07)	Ba_2BiTbO_6 (0.1)	Ba_2BiYO_6 (0.07)	Ba_2NbAlO_6 (0.01)
Ba_2NbBiO_6 (0.3)	Ba_2NbPmO_6 (0.13)	Ba_2NbPrO_6 (0.15)	Ba_2NbTbO_6 (0.02)
Ba_2SbDyO_6 (0.06)	Ba_2SbErO_6 (0.05)	Ba_2SbGdO_6 (0.01)	Ba_2SbHoO_6 (0.06)
Ba_2SbLaO_6 (0.01)	Ba_2SbTmO_6 (0.04)	Ba_2SbYO_6 (0.02)	Ba_2TaDyO_6 (0.01)
Ba_2TaErO_6 (0.07)	Ba_2TaGdO_6 (0.03)	Ba_2TaHoO_6 (0.03)	Ba_2TaInO_6 (0.24)
Ba_2TaLaO_6 (0.06)	Ba_2TaNdO_6 (0.01)	Ba_2TaPmO_6 (0.04)	Ba_2TaPrO_6 (0.17)
Ba_2TaSmO_6 (0.02)	Ba_2TaTbO_6 (0.02)	Ba_2TaTmO_6 (0.05)	Ba_2TaYO_6 (0.08)
$BaCaAuDyO_6$ (0.16)	$BaCaAuErO_6$ (0.1)	$BaCaAuHoO_6$ (0.13)	$BaCaBiDyO_6$ (0.01)
$BaCaNbYO_6$ (0.09)	$BaCaSbDyO_6$ (0.15)	$BaCaSbTbO_6$ (0.05)	$BaCaSbYO_6$ (0.04)
$BaCaTaDyO_6$ (0.06)	$BaCaTaErO_6$ (0.04)	$BaCaTaGdO_6$ (0)	$BaCaTaHoO_6$ (0.1)
$BaCaTaTbO_6$ (0.07)	$BaCaTaTmO_6$ (0.01)	$BaCaTaYO_6$ (0.03)	$BaSrAuDyO_6$ (0.03)
$BaSrAuErO_6$ (0.00)	$BaSrAuGdO_6$ (0.04)	$BaSrAuHoO_6$ (0.01)	$BaSrAuNdO_6$ (0.08)
$BaSrAuPmO_6$ (0.08)	$BaSrBiDyO_6$ (0.19)	$BaSrBiTbO_6$ (0.11)	$BaSrHfSnO_6$ (0.22)
$BaSrNbDyO_6$ (0.14)	$BaSrNbSmO_6$ (0.26)	$BaSrSbDyO_6$ (0.02)	$BaSrSbErO_6$ (0.01)
$BaSrSbHoO_6$ (0.02)	$BaSrSbTbO_6$ (0.03)	$BaSrSbYO_6$ (0.01)	$BaSrTaDyO_6$ (0.27)
$BaSrTaErO_6$ (0.32)	$BaSrTaHoO_6$ (0.05)	$BaSrTaNdO_6$ (0.28)	$BaSrTaPmO_6$ (0.16)
$BaSrTaTbO_6$ (0.05)	Ca_2MoMgO_6 (0.04)	Ca_2MoZnO_6 (0.09)	Ca_2WMgO_6 (0.02)
Ca_2WZnO_6 (0.13)	$DyCsHfGeO_6$ (0.08)	$DyCsHfSnO_6$ (0.22)	$DyCsTaInO_6$ (0.38)
$DyKTaInO_6$ (0.08)	$DyRbTaInO_6$ (0.18)	$ErCsHfGeO_6$ (0.2)	$ErCsHfTiO_6$ (0.22)
$ErCsTaInO_6$ (0.23)	$ErCsTaScO_6$ (0.21)	$ErKHfSnO_6$ (0.20)	$ErKTaInO_6$ (0.21)
$ErRbHfSnO_6$ (0.26)	$ErRbTaInO_6$ (0.22)	$HoCsTaInO_6$ (0.22)	$HoKHfSnO_6$ (0.26)
$HoKTaInO_6$ (0.27)	$HoRbHfSnO_6$ (0.2)	$HoRbTaInO_6$ (0.31)	$LaCsAuDyO_6$ (0.34)
$LaCsAuErO_6$ (0.04)	$LaCsAuHoO_6$ (0.03)	$LaCsAuNdO_6$ (0.35)	$LaCsAuPmO_6$ (0.36)
$LaCsBiDyO_6$ (0.2)	$LaCsBiErO_6$ (0.08)	$LaCsBiHoO_6$ (0.17)	$LaCsBiNdO_6$ (0.19)
$LaCsBiPmO_6$ (0.1)	$LaCsBiSmO_6$ (0.11)	$LaCsBiTbO_6$ (0.01)	$LaCsMoYbO_6$ (0.06)
$LaCsNbDyO_6$ (0.17)	$LaCsNbErO_6$ (0.03)	$LaCsNbHoO_6$ (0.26)	$LaCsSbDyO_6$ (0.04)
$LaCsSbErO_6$ (0.02)	$LaCsSbGdO_6$ (0.18)	$LaCsSbHoO_6$ (0.35)	$LaCsSbPmO_6$ (0.12)
$LaCsSbTbO_6$ (0.23)	$LaCsTaDyO_6$ (0.37)	$LaCsTaErO_6$ (0.34)	$LaCsTaHoO_6$ (0.39)
$LaCsTaNdO_6$ (0.05)	$LaCsTaPmO_6$ (0.40)	$LaCsTaSmO_6$ (0.28)	$LaCsTaTbO_6$ (0.38)
$LaKAuDyO_6$ (0.34)	$LaKAuErO_6$ (0.25)	$LaKAuHoO_6$ (0.31)	$LaKAuNdO_6$ (0.06)
$LaKAuPmO_6$ (0.42)	$LaKBiDyO_6$ (0.31)	$LaKBiGdO_6$ (0.41)	$LaKBiHoO_6$ (0.28)
$LaKHfZrO_6$ (0.26)	$LaKNbYO_6$ (0.22)	$LaKSbDyO_6$ (0.28)	$LaKSbTbO_6$ (0.05)
$LaKSbYO_6$ (0.34)	$LaKTaDyO_6$ (0.38)	$LaKTaErO_6$ (0.37)	$LaKTaHoO_6$ (0.42)
$LaKTaNdO_6$ (0.16)	$LaKTaPmO_6$ (0.4)	$LaKTaSmO_6$ (0.41)	$LaKTaTbO_6$ (0.32)
$LaKTaYO_6$ (0.42)	$LaRbAuDyO_6$ (0.39)	$LaRbAuErO_6$ (0.29)	$LaRbAuHoO_6$ (0.36)
$LaRbAuNdO_6$ (0)	$LaRbAuPmO_6$ (0.01)	$LaRbBiDyO_6$ (0.41)	$LaRbCeZrO_6$ (0.3)
$LaRbNbYO_6$ (0.41)	$LaRbSbDyO_6$ (0.12)	$LaRbSbTbO_6$ (0.11)	$LaRbSbYO_6$ (0.18)
$LaRbTaDyO_6$ (0.37)	$LaRbTaErO_6$ (0.35)	$LaRbTaHoO_6$ (0.4)	$LaRbTaNdO_6$ (0.44)
$LaRbTaPmO_6$ (0.43)	$LaRbTaSmO_6$ (0.41)	$LaRbTaTbO_6$ (0.45)	$LaRbZrSnO_6$ (0.41)
$LuCsHfSnO_6$ (0.26)	$NdCsAuDyO_6$ (0.03)	$NdCsAuErO_6$ (0.03)	$NdCsAuGdO_6$ (0.21)
$NdCsAuTmO_6$ (0.05)	$NdCsAuYO_6$ (0.14)	$NdCsBiGdO_6$ (0.23)	$NdCsBiYO_6$ (0.1)
$NdCsTaDyO_6$ (0.48)	$NdCsTaErO_6$ (0.48)	$NdCsTaGdO_6$ (0.41)	$NdCsTaHoO_6$ (0.44)
$NdCsTaInO_6$ (0.15)	$NdCsTaTmO_6$ (0.41)	$NdCsTaYO_6$ (0.21)	$NdKAuDyO_6$ (0.4)
$NdKAuErO_6$ (0.34)	$NdKAuGdO_6$ (0.08)	$NdKAuHoO_6$ (0.38)	$NdKAuLuO_6$ (0.22)
$NdKAuTmO_6$ (0.3)	$NdKAuYO_6$ (0.41)	$NdKBiDyO_6$ (0.42)	$NdKBiGdO_6$ (0.18)
$NdKBiHoO_6$ (0.39)	$NdKBiTmO_6$ (0.37)	$NdKBiYO_6$ (0.21)	$NdKHfSnO_6$ (0.21)
$NdKSbDyO_6$ (0.43)	$NdKSbErO_6$ (0.38)	$NdKSbGdO_6$ (0.42)	$NdKSbHoO_6$ (0.41)
$NdKSbTbO_6$ (0.23)	$NdKSbTmO_6$ (0.22)	$NdKTaDyO_6$ (0.09)	$NdKTaErO_6$ (0.23)
$NdKTaGdO_6$ (0.06)	$NdKTaHoO_6$ (0.39)	$NdKTaInO_6$ (0.4)	$NdKTaTmO_6$ (0.48)
$NdKTaYO_6$ (0.26)	$NdRbAuDyO_6$ (0.12)	$NdRbAuErO_6$ (0.05)	$NdRbAuGdO_6$ (0.07)
$NdRbAuHoO_6$ (0.09)	$NdRbAuLuO_6$ (0.16)	$NdRbAuTmO_6$ (0.38)	$NdRbAuYO_6$ (0.12)
$NdRbBiDyO_6$ (0.07)	$NdRbBiTmO_6$ (0.02)	$NdRbBiYO_6$ (0.14)	$NdRbCeZrO_6$ (0.4)
$NdRbHfSnO_6$ (0.2)	$NdRbSbDyO_6$ (0.3)	$NdRbSbErO_6$ (0.25)	$NdRbSbGdO_6$ (0.36)
$NdRbSbHoO_6$ (0.27)	$NdRbSbTbO_6$ (0.39)	$NdRbSbTmO_6$ (0.11)	$NdRbTaBiO_6$ (0.34)
$NdRbTaDyO_6$ (0.43)	$NdRbTaErO_6$ (0.45)	$NdRbTaGdO_6$ (0.4)	$NdRbTaHoO_6$ (0.4)
$NdRbTaInO_6$ (0.45)	$NdRbTaTmO_6$ (0.48)	$NdRbTaYO_6$ (0.43)	$PmCsAuDyO_6$ (0.02)
$PmCsAuErO_6$ (0.08)	$PmCsAuGdO_6$ (0.06)	$PmCsAuHoO_6$ (0.01)	$PmCsAuTmO_6$ (0.04)
$PmCsAuYO_6$ (0.06)	$PmCsTaDyO_6$ (0.42)	$PmCsTaErO_6$ (0.42)	$PmCsTaGdO_6$ (0.3)
$PmCsTaHoO_6$ (0.39)	$PmCsTaInO_6$ (0.08)	$PmCsTaTmO_6$ (0.35)	$PmCsTaYO_6$ (0.46)
$PmKAuGdO_6$ (0)	$PmKAuYO_6$ (0.42)	$PmKHfSnO_6$ (0.17)	$PmKTaGdO_6$ (0.41)
$PmKTaTmO_6$ (0.47)	$PmRbAuGdO_6$ (0.36)	$PmRbAuYO_6$ (0.04)	$PmRbTaGdO_6$ (0.43)
$PmRbTaInO_6$ (0.05)	$PmRbTaTmO_6$ (0.43)	$PmRbVInO_6$ (0.08)	$PrCsAuDyO_6$ (0.42)
$PrCsAuErO_6$ (0.42)	$PrCsAuHoO_6$ (0.43)	$PrCsAuLuO_6$ (0.07)	$PrCsBiDyO_6$ (0.12)

Table 5 (continued)

Predicted novel oxide perovskite candidates			
$PrCsSbDyO_6$ (0.06)	$PrCsTaDyO_6$ (0.41)	$PrCsTaErO_6$ (0.43)	$PrCsTaHoO_6$ (0.47)
$PrCsTaTbO_6$ (0.46)	$PrKAuDyO_6$ (0.39)	$PrKAuErO_6$ (0.32)	$PrKAuHoO_6$ (0.36)
$PrKAuLuO_6$ (0.19)	$PrKAuScO_6$ (0.2)	$PrKBiDyO_6$ (0.42)	$PrKBiErO_6$ (0.38)
$PrKSbDyO_6$ (0.06)	$PrKSbErO_6$ (0.4)	$PrKSbHoO_6$ (0.04)	$PrKSbTbO_6$ (0.18)
$PrKTaDyO_6$ (0.42)	$PrKTaErO_6$ (0.43)	$PrKTaHoO_6$ (0.3)	$PrKTaYO_6$ (0.42)
$PrNdDyCoO_6$ (0.01)	$PrNdErCoO_6$ (0.18)	$PrNdHoCoO_6$ (0.14)	$PrRbAuDyO_6$ (0.03)
$PrRbAuErO_6$ (0.39)	$PrRbAuHoO_6$ (0.06)	$PrRbBiDyO_6$ (0.14)	$PrRbBiErO_6$ (0.09)
$PrRbBiHoO_6$ (0.11)	$PrRbSbDyO_6$ (0.09)	$PrRbSbErO_6$ (0.03)	$PrRbSbHoO_6$ (0.06)
$PrRbSbTbO_6$ (0.17)	$PrRbSbYO_6$ (0.06)	$PrRbTaDyO_6$ (0.48)	$PrRbTaErO_6$ (0.45)
$PrRbTaHoO_6$ (0.44)	$PrRbTaTbO_6$ (0.48)	Sr_2AuErO_6 (0.07)	Sr_2AuNdO_6 (0.16)
Sr_2BiErO_6 (0.06)	Sr_2BiHoO_6 (0.18)	Sr_2BiYO_6 (0.31)	Sr_2MoZnO_6 (0.09)
Sr_2SbErO_6 (0.01)	Sr_2SbScO_6 (0.24)	Sr_2SbTmO_6 (0.03)	Sr_2TaInO_6 (0.2)
Sr_2WMgO_6 (0.05)	Sr_2WZnO_6 (0.03)	$TbCsTaInO_6$ (0.08)	$TbCsVInO_6$ (0.11)
$TbRbTaInO_6$ (0.01)	$YCsHfSnO_6$ (0.22)		

These compounds have never been synthesized. Absolute difference in predicted and calculated band gaps in eV is also denoted in parentheses in blue for the 310 oxide perovskites for which the predicted band gaps have been validated.

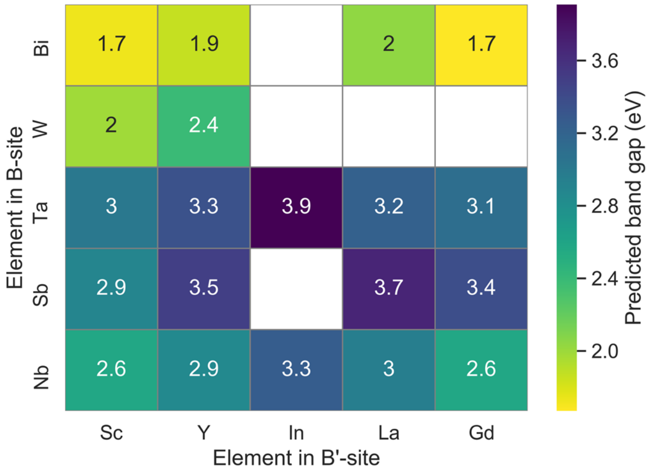

Fig. 7 Band gap prediction map for selected combinations of $Ba_2BB'O_6$. White squares indicate compositions that are predicted to be narrow band gap materials and hence do not have an accompanying band gap prediction.

train the ML models and achieve high predictive accuracies. Confidence intervals were derived for the quantitative band gap prediction model and our predictions are validated with DFT computations. We find excellent agreement between our predictions and these validating calculations.

The calculated band gap data that comprises the training data as well as the predicted band gaps may be used to generate design maps for desired combinations of elements on the *A*- and *B*- sites that can offer insight into the variation of the band gap due to doping and be used to identify preliminary candidates for specific applications. The 310 identified candidates are novel chemistries that have not been explored experimentally to date. These candidates are predicted to have band gaps ranging from 0.5 to 4 eV and, consequently, can potentially find application in a wide range of areas ranging from infrared radiation detection, solar cells, and other light emitting devices (LEDs), to scintillator materials.

The efficiency of our multi-step hierarchical screening approach, which may be generalized to investigate other classes of materials in addition to the oxide perovskites examined here, provides further impetus to the application of physics-based ML models to the discovery of novel functional materials. This approach - of creating models and then using them to identify specific novel candidates that will be of value to the community at large and that are suggested for follow-on experimental studies - is not always followed in the prior literature. We note that this hierarchical strategy, where we first classify materials with perovskite structure as either insulators or metals and then train regression models on the insulators only with such high accuracies, is unique in the literature and provides a route for improved models for these types of properties.

Lastly, we note that although the absolute DFT gaps are underestimated due to our choice of the low-fidelity but less expensive PBE functional, the relative changes in band gap as a function of chemistry are well estimated by the PBE-GGA and they correlate with the experimental band gaps (since the underestimation is systemic and generally proportional to the band gap itself) and therefore chemical trends, which in themselves are very useful, are expected to be well captured using our approach. A high-throughput study such as the present one necessitates a low-cost technique and recent studies[56,71] have indicated that ground state properties calculated using the PBE-GGA functional are sufficiently accurate, particularly for changes in electronic structure with chemistry. However, to ensure that this is true for the current application, 100 materials were randomly selected from our training dataset of wide band gap materials and their HSE band gaps were calculated; the results are included in the Supplementary Note 6. It is seen that PBE-GGA indeed systematically underestimates the band gap for most materials.

Methods

DFT calculation details. The Vienna ab initio simulation package (VASP)[72,73] implementation of the DFT framework was used in this work to calculate the band gaps of the training data. The parameterization proposed by Perdew, Burke, and Ernzerhof [51] of the GGA[74] approach was used. A Monkhorst-Pack mesh was used to perform the Brillouin zone integrations with at least 5000k points per reciprocal atom. The structures were fully relaxed using the Methfessel–Paxton smearing method[75] of order one and a final self-consistent static calculation was carried out. The calculations were spin polarized and we used a cutoff energy of 533 eV for all of the structures. All relaxations were carried out until changes in the total energy between relaxation steps were within 1×10^{-6} eV and atomic forces on each of the atoms were smaller than 0.01 eV/Å.

ML models using random forests. The choice to use RFs to build our ML models was made to leverage the inherent robustness, low bias, and moderate variance afforded by the technique. The RF (also known as Random Decision Trees) is a *bagging*-based ensemble learning method. It has been shown that using ensembles of trees, where each tree in the ensemble is grown in accordance with the realization of a random vector, results in consequential gains in classification or prediction accuracy. Final predictions are obtained by aggregate voting over the ensemble using equal weights in most cases. RFs seek to induce randomness by using subsets of descriptors drawn at random to determine the optimal split of a given node of a tree, thereby reducing the correlation between the quantities being averaged and consequently enhancing the variance gains. We used the Scikit-learn package to implement our RF models.

For the RF classification model M_C, the training and test dataset selections were stratified over the wide/narrow band gap chemistries using a 90/10 training/test split. Thus, we used 90% of the dataset D_P' to train the classification model M_C and then tested on the remaining 10%. The maximum tree depth was set at 25 and the number of estimators or trees was 200 for the classification model. For the RF regression model M_R, we also used 90% of the dataset D_{BG} to train the regression model M_R and then tested on the remaining 10%. The maximum tree depth was set to 50 and the number of estimators was chosen to be 200. To maximize accuracy while minimizing the standard deviation on unseen data, we used 5-fold cross-validation using a 90% training subset to determine the hyper-parameters for both classification and regression.

Uncertainty quantification. RF models are difficult to interpret owing to their black box nature, and it is difficult to quantify the associated modeling and input uncertainties. Two methods are most widely used for the quantification of confidence intervals in RF-based regression models: i) U-Statistic-based RFs[76] and ii) bootstrap[77], jackknife-after-bootstrap[78], and infinitesimal jackknife[79] based methods, which we use in this work.

U-statistics[80] is a class of statistics in which a minimum-variance unbiased prediction is derived by drawing a predetermined number of times through all combinatorial selections of the training data set and then averaging over the possible results of these sub-samples. Mentch and Hooker[76] showed that under a strict sub-sampling scheme, predictions for individual feature vectors are asymptotically normal, allowing for application of U-statistics to RF predictions. Relevant statistical measures can then be used to quantify the uncertainty related to the reducible error of the RF prediction and construct confidence intervals.

Bootstrap sampling and jackknifing rely on estimating the variance of a prediction by using the variability between re-samples rather than using statistical distributions. In bootstrapping, numerous prediction models are developed by randomly excluding varied small subsets of the training data and the mean and variance of the predictions is estimated from these bootstrapped models. This methods attempts to quantify the sensitivity of the model with respect to slight perturbations in the training data. Jackknifing is another re-sampling technique predating bootstrapping, in which each training data point is systematically left out, the model is trained on the remaining data and an estimate is calculated; the jackknife estimate is found by evaluating the average of these calculations. These two ideas may be combined in a jackknife-after-bootstrap method which is used to find the an error estimate (for example variance) to a bootstrap estimate. As opposed to jackknifing and jackknifing-after-bootstrapping, where the behavior of a prediction is studied after excluding one or more observations at a time, the infinitesimal jackknife (IJ) looks at the effect on a prediction after down-weighting each observation by an infinitesimal amount. In 2014, Wager et al.[81] demonstrated that both the jackknife-after-bootstrap and the infinitesimal jackknife methods suffer from considerable Monte Carlo bias, and they proposed a bias corrected version of the method, the implementation of which is used to calculate the confidence intervals for the test data using our band gap regression model.

Data availability

The training dataset used in this study is described in Supplementary Data 1 and the predicted candidates are listed in Supplementary Data 2. Additional data related to this work is available on reasonable request.

Code availability

The source code used in this study may be downloaded from GitHub (https://github.com/anjanatalapatra/perovskite_oxide_discovery).

Received: 28 April 2022; Accepted: 22 May 2023;
Published online: 10 June 2023

References

1. Ueno, K. et al. Field-effect transistor based on ktao 3 perovskite. *Appl. Phys. Lett.* **84**, 3726–3728 (2004).
2. Schubert, E. F. & Kim, J. K. Solid-state light sources getting smart. *Science* **308**, 1274–1278 (2005).
3. Goetzberger, A. & Hebling, C. Photovoltaic materials, past, present, future. *Solar Energy Mater. Solar Cells* **62**, 1–19 (2000).
4. Van Loef, E., Dorenbos, P., Van Eijk, C., Krämer, K. & Güdel, H.-U. High-energy-resolution scintillator: Ce 3+ activated labr 3. *Appl. Phys. Lett.* **79**, 1573–1575 (2001).
5. Yu, L. & Zunger, A. Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. *Phys. Rev. Lett.* **108**, 068701 (2012).
6. Castelli, I. E. et al. New light-harvesting materials using accurate and efficient bandgap calculations. *Adv. Energy Mater.* **5**, 1400915 (2015).
7. Huo, Z., Wei, S.-H. & Yin, W.-J. High-throughput screening of chalcogenide single perovskites by first-principles calculations for photovoltaics. *J. Phys. D Appl. Phys.* **51**, 474003 (2018).
8. Uchino, K. Glory of piezoelectric perovskites. *Sci. Technol. Adv. Mater.* **16**, 046001 (2015).
9. DiDomenico Jr, M. & Wemple, S. Optical properties of perovskite oxides in their paraelectric and ferroelectric phases. *Phys. Rev.* **166**, 565 (1968).
10. Galasso, F. Perovskite type compounds and high t c superconductors. *JOM* **39**, 8–10 (1987).
11. Towler, M., Dovesi, R. & Saunders, V. R. Magnetic interactions and the cooperative Jahn-Teller effect in $KCuF_3$. *Phys. Rev. B* **52**, 10150 (1995).
12. Visser, D., Ramirez, A. & Subramanian, M. Thermal conductivity of manganite perovskites: colossal magnetoresistance as a lattice-dynamics transition. *Phys. Rev. Lett.* **78**, 3947 (1997).
13. Dulian, P. Solid-state mechanochemical syntheses of perovskites. In *Perovskite Materials: Synthesis, Characterisation, Properties, and Applications*, 1 (eds Pan, L. & Guang, Z.) (BoD–Books on Demand, 2016).
14. Sham, L. J. & Schlüter, M. Density-functional theory of the energy gap. *Phys. Rev. Lett.* **51**, 1888 (1983).
15. Perdew, J. P. & Levy, M. Physical content of the exact kohn-sham orbital energies: band gaps and derivative discontinuities. *Phys. Rev. Lett.* **51**, 1884 (1983).
16. Anisimov, V. I., Aryasetiawan, F. & Lichtenstein, A. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the lda+ u method. *J. Phys. Condensed Matter* **9**, 767 (1997).
17. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA+U framework. *Phys. Rev. B* **73**, 195107 (2006).
18. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Fractional charge perspective on the band gap in density-functional theory. *Phys. Rev. B* **77**, 115123 (2008).
19. Martin, R. M. *Electronic structure: basic theory and practical methods* (Cambridge University Press, 2020).
20. Chan, M. & Ceder, G. Efficient band gap prediction for solids. *Phys. Rev. Lett.* **105**, 196403 (2010).
21. Crowley, J. M., Tahir-Kheli, J. & Goddard III, W. A. Resolution of the band gap prediction problem for materials design. *J. Phys. Chem. Lett.* **7**, 1198–1203 (2016).
22. Aryasetiawan, F. & Gunnarsson, O. The GW method. *Rep. Progr. Phys.* **61**, 237 (1998).
23. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
24. Pilania, G. Machine learning in materials science: from explainable predictions to autonomous design. *Comput. Mater. Sci.* **193**, 110360 (2021).
25. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 1–13 (2017).
26. Morgan, D. & Jacobs, R. Opportunities and challenges for machine learning in materials science. *Ann. Rev. Mater. Res.* **50**, 71–103 (2020).
27. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
28. Choudhary, K. et al. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **8**, 1–26 (2022).
29. Gu, T., Lu, W., Bao, X. & Chen, N. Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors. *Solid State Sci.* **8**, 129–136 (2006).
30. 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. *Phys. Rev. B* **93**, 115104 (2016).
31. Huang, Y. et al. Band gap and band alignment prediction of nitride-based semiconductors using machine learning. *J. Mater. Chem. C* **7**, 3238–3245 (2019).
32. Li, J., Pradhan, B., Gaur, S. & Thomas, J. Predictions and strategies learned from machine learning to develop high-performing perovskite solar cells. *Adv. Energy Mater.* **9**, 1901891 (2019).
33. Stanley, J. C., Mayr, F. & Gagliardi, A. Machine learning stability and bandgaps of lead-free perovskites for photovoltaics. *Adv. Theory Simul.* **3**, 1900178 (2020).
34. Setyawan, W., Gaume, R. M., Lam, S., Feigelson, R. S. & Curtarolo, S. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS Combinatorial Sci.* **13**, 382–390 (2011).
35. Dey, P. et al. Informatics-aided bandgap engineering for solar materials. *Comput. Mater. Sci.* **83**, 185–195 (2014).
36. Khmaissia, F. et al. Accelerating band gap prediction for solar materials using feature selection and regression techniques. *Comput. Mater. Sci.* **147**, 304–315 (2018).

37. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 1–6 (2013).
38. Pilania, G. et al. Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 1–10 (2016).
39. Pilania, G., Gubernatis, J. E. & Lookman, T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).
40. Na, G. S., Jang, S., Lee, Y.-L. & Chang, H. Tuplewise material representation based machine learning for accurate band gap prediction. *J. Phys. Chem. A* **124**, 10616–10623 (2020).
41. Li, X.-G. et al. Graph network based deep learning of bandgaps. *J. Chem. Phys.* **155**, 154702 (2021).
42. Omprakash, P. et al. Graph representational learning for bandgap prediction in varied perovskite crystals. *Comput. Mater. Sci.* **196**, 110530 (2021).
43. Gladkikh, V. et al. Machine learning for predicting the band gaps of ABX_3 perovskites from elemental properties. *J. Phys. Chem. C* **124**, 8905–8918 (2020).
44. Baker, N. et al. *Workshop report on basic research needs for scientific machine learning: core technologies for artificial intelligence*. Technical Report, (USDOE Office of Science (SC), 2019).
45. Kauwe, S. K., Welker, T. & Sparks, T. D. Extracting knowledge from dft: experimental band gap predictions through ensemble learning. *Integr. Mater. Manuf. Innov.* **9**, 213–220 (2020).
46. Li, W. et al. Predicting band gaps and band-edge positions of oxide perovskites using density functional theory and machine learning. *Phys. Rev. B* **106**, 155156 (2022).
47. Zhang, S. et al. Predicting the formability of hybrid organic–inorganic perovskites via an interpretable machine learning strategy. *The J. Phys. Chem. Lett.* **12**, 7423–7430 (2021).
48. Liu, H. et al. Screening stable and metastable abo3 perovskites using machine learning and the materials project. *Comput. Mater. Sci.* **177**, 109614 (2020).
49. Yang, Z. et al. Machine learning accelerates the discovery of light-absorbing materials for double perovskite solar cells. *J. Phys. Chem. C* **125**, 22483–22492 (2021).
50. Wu, Y., Lu, S., Ju, M.-G., Zhou, Q. & Wang, J. Accelerated design of promising mixed lead-free double halide organic–inorganic perovskites for photovoltaics using machine learning. *Nanoscale* **13**, 12250–12259 (2021).
51. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
52. Talapatra, A., Uberuaga, B. P., Stanek, C. R. & Pilania, G. A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides. *Chem. Mater.* **33**, 845–858 (2021).
53. Bondzior, B., Vu, T., Stefańska, D., Winiarski, M. & Dereń, P. Tunable broadband emission by bandgap engineering in (ba, sr) 2 (mg, zn) wo6 inorganic double-perovskites. *J. Alloys Compounds* **888**, 161567 (2021).
54. Jia L, Lloyd M, Lees M, Huang L, Walton R. Limits of solid solution and evolution of crystal morphology in (La1-x RE x) FeO3 perovskites by low temperature hydrothermal crystallization. *Inorg. Chem.* **62**, 4503–4513 (2023).
55. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
56. Yadav, S. K., Uberuaga, B. P., Nikl, M., Jiang, C. & Stanek, C. R. Band-gap and band-edge engineering of multicomponent garnet scintillators from first principles. *Phys. Rev. Appl.* **4**, 054012 (2015).
57. Pilania, G., Balachandran, P. V., Gubernatis, J. E. & Lookman, T. Data-based methods for materials design and discovery: basic ideas and general methods. *Synth. Lect. Mater. Optics* **1**, 1–188 (2020).
58. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
59. Pearson, K. & Lee, A. Mathematical contributions to the theory of evolution. viii. on the inheritance of characters not capable of exact quantitative measurement. part i. introductory. part ii. on the inheritance of coat-colour in horses. part iii. on the inheritance of eye-colour in man. *Philos. Trans. R. Soc. Lond. Ser. A* **195**, 79–150 (1900).
60. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
61. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
62. Beurich, H., Madach, T., Richter, F. & Vahrenkamp, H. Experiments on the HOMO-LUMO nature of metal-metal bonds. *Angew. Chemie Int. Ed. Engl.* **18**, 690–691 (1979).
63. Zunger, A. A pseudopotential viewpoint of the electronic and structural properties of crystals. *Struct. Bond. Cryst.* **1**, 73–135 (1981).
64. Goldschmidt, V. M. Die gesetze der krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
65. Li, C. et al. Formability of ABX$_3$ (X= F, Cl, Br, I) Halide Perovskites. *Acta Crystallogr. Sect. B Struct. Sci.* **64**, 702–707 (2008).
66. Filip, M. R. & Giustino, F. The geometric blueprint of perovskites. *Proc. Natl Acad. Sci.* **115**, 5397–5402 (2018).
67. Shannon, R. T. & Prewitt, C. Revised values of effective ionic radii. *Acta Crystallogr. Sect. B Struct. Crystallogr. Crystal Chem.* **26**, 1046–1048 (1970).
68. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
69. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143**, 29–36 (1982).
70. Friedman, J., Hastie, T. & Tibshirani, R. *The elements of statistical learning*, vol. 1, Springer series in statistics (Springer, 2001).
71. Fasoli, M. et al. Band-gap engineering for removing shallow traps in rare-earth Lu$_3$Al$_5$O$_{12}$ garnet scintillators using ga 3+ doping. *Phys. Rev. B* **84**, 081102 (2011).
72. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
73. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50 (1996).
74. Perdew, J. P. & Wang, Y. Accurate and simple analytic representation of the electron-gas correlation energy. *Phys. Rev. B* **45**, 13244 (1992).
75. Methfessel, M. & Paxton, A. High-precision sampling for brillouin-zone integration in metals. *Phys. Rev. B* **40**, 3616 (1989).
76. Mentch, L. & Hooker, G. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.* **17**, 841–881 (2016).
77. Efron, B., Halloran, E. & Holmes, S. Bootstrap confidence levels for phylogenetic trees. *Proc. Natl Acad. Sci.* **93**, 13429–13429 (1996).
78. Efron, B. Jackknife-after-bootstrap standard errors and influence functions. *J. R. Stat. Soc. Ser. B* **54**, 83–111 (1992).
79. Efron, B. Estimation and accuracy after model selection. *J. Am. Stat. Assoc.* **109**, 991–1007 (2014).
80. Hoeffding, W. A class of statistics with asymptotically normal distribution. In *Breakthroughs in statistics*, 308–334 (Springer, 1992).
81. Wager, S., Hastie, T. & Efron, B. Confidence intervals for random forests: the jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **15**, 1625–1651 (2014).

Acknowledgements

Research presented in this paper was supported by the Laboratory Directed Research and Development program of Los Alamos National Laboratory under project numbers 20190656PRD4 and 20190043DR. Computational support for this work was provided by LANLs high-performance computing clusters. This work was supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001).

Author contributions

B.P.U and G.P. proposed and supervised the entire project and C.R.S. was involved in the initial conceptualization. A.T. worked on the development and testing of the model and performed DFT simulations. B.P.U., G.P. and A.T. analyzed and discussed theory results. A.T. prepared the final draft of the manuscript which was then reviewed and edited by all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43246-023-00373-4.

Correspondence and requests for materials should be addressed to Anjana Talapatra.

Peer review information Communications Materials thanks Felix Mayr, Seunghun Jang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: Milica Todorović and Aldo Isidori. A peer review file is available.