Journal of Applied Physics ARTICLE pubs.aip.org/aip/jap

Predicting efficient and stable inorganic photovoltaic materials using interpretable machine learning combined with DFT calculations based on band edge orbital engineering

Cite as: J. Appl. Phys. 138, 023101 (2025); doi: 10.1063/5.0267853
Submitted: 26 February 2025 · Accepted: 18 June 2025 ·
Published Online: 8 July 2025

Ruo-Tong Chen, Zhihua Hu, and Hong-Jian Feng[a)]

AFFILIATIONS

School of Physics, Northwest University, Xi'an 710127, China

Note: This paper is part of the Special Topic on Integrating Data Science and Computational Materials Science.
[a)]**Author to whom correspondence should be addressed:** hjfeng@nwu.edu.cn

ABSTRACT

The discovery of new materials with high power conversion efficiency (PCE) is critical for the advancement of solar energy technologies. In this study, we combine traditional machine learning (ML) methods and deep learning methods to predict and categorize the inorganic photovoltaic materials, which are more stable than the organic counterpart. By employing gradient boosting trees, extremely randomized trees, random forests, backpropagation neural network, and convolutional neural network, we classify materials by bandgap (zero or non-zero) and predict their values with regression model. The stability, bandgaps, optical absorption, and PCE of these materials are validated through density functional theory (DFT) calculations, leading to the identification of promising candidates: $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, and $Mo_2Ba_5N_7$. The Shapley Additive Explanation method is applied to analyze feature interactions, intuitively establishing the relationship between band-edge orbitals and material properties, while uncovering hidden connections between structural and electronic properties. The results reveal that the delocalization of valence electrons, along with variations in atomic coordination environments, modifies the charge density distribution at the band edges, affecting the transition probabilities and atomic orbital connectivity. Among these, $Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$ stand out as the most promising materials for solar cell applications. Our findings provide a novel framework for accelerating the discovery of efficient photovoltaic materials using ML and DFT methods.

I. INTRODUCTION

Photovoltaic energy conversion is vital for sustainable development, so developing high power conversion efficiency (PCE) solar cells is urgent.[1] The performance of solar cells is determined by multiple factors, including material quality, device structure, and interface characteristics. However, the quality of the photovoltaic materials ultimately sets the upper limit for performance of solar cells. Over the past decades, the growing demand for various photovoltaic applications has led to the rapid emergence of organic photovoltaic (OPV) materials as a promising new technology, owing to their lightweight, flexibility, and low-cost characteristics.[2] However, the PCE and stability of OPV materials still lag behind those of inorganic materials, limiting their application in large-scale energy production.[2] In inorganic photovoltaic materials, GaAs has achieved very high PCE in photovoltaic applications due to its more suitable bandgap compared to crystalline silicon.[3] Meanwhile, emerging perovskite solar cells have also demonstrated significant potential for practical applications.[4] However, the high production cost of GaAs materials and the instability of perovskite materials have driven researchers to continuously explore new inorganic photovoltaic materials to overcome the limitations of existing technologies. It is worth noting that the promising photovoltaic materials reported so far should have high electronic dimensionality and that the high-dimensional crystal structures serve as the foundation for realizing high electronic dimensionality.[5] Therefore, the search for

high-performance photovoltaic materials with both structural and electronic three-dimensional characteristics, as well as an in-depth investigation of their microscopic electronic structures, is crucial for developing efficient solar cells.[6] For determining whether the material has excellent photovoltaic properties, the core is to have a full understanding and mastery of its bandgap (E_g). Therefore, density functional theory (DFT) is commonly used to predict the physical properties of materials and guide synthesis efforts toward the development of optimal materials.[7] Unfortunately, calculations based on DFT simulations tend to be both time- and cost-intensive, which significantly impedes the advancement of materials research.[8]

In recent years, the growing of materials databases has driven the integration of machine learning (ML) into materials science.[9] People can use ML to analyze the relationships between features and attributes in data samples, thereby revealing hidden patterns in multidimensional datasets.[10] For the past few years, the ML has significantly reduced the computational cost of DFT and accelerated the discovery of new materials.[9] ML can be broadly categorized into two main branches: deep learning (DL) and traditional ML, and they have obvious differences and unique advantages in methods and applications. Traditional ML has relatively low hardware requirements and performs well on many small and simple problems. DL utilizes multi-layer neural networks to automatically extract features, allowing for the representation of more detailed information and higher-level features, thereby improving the performance of the model.[11] Despite the many differences between DL and traditional ML, they are not mutually exclusive. In many cases, better results can be achieved by combining the two techniques. Currently, ML has been widely applied in the development of new energy materials, achieving significant breakthroughs in fields such as lithium-ion batteries,[12] thermoelectric materials,[13] and catalytic production.[14] Furthermore, our research has utilized ML to predict the 2D perovskite photovoltaic materials and guide the compositional engineering of $CsPbX_3$ (X = Cl, Br, I) perovskites.[15,16] In the field of inorganic materials research and its various applications, Park *et al.* successfully developed a thermodynamic regression model and a convex hull energy classifier to screen for thermodynamically stable MXenes.[17] In our previous research,[18] we utilized a New Light Harvesting Materials database from the Computational Materials Repository (CMR),[19] containing 2398 materials, to employ a Gradient Boosting Regression[20] ML model for predicting inorganic photovoltaic materials. By combining this approach with DFT calculations, we identified four promising photovoltaic candidates from the 3587 predicted materials: $Sr_8P_8Sn_4$, $Ba_8P_8Ge_4$, $Y_4Te_4Se_2$, and $Ba_4Te_{12}Ge_4$.[18] Furthermore, we discovered that strong *p*–*p* transitions near the band edges significantly contribute to enhancing the PCE, providing a guiding framework for the discovery of new materials.[18]

However, due to the "black-box" nature of ML models, interpretative knowledge is often scarce. Therefore, it is particularly important to improve the interpretability of ML models. Ritesh Kumar *et al.* combined the SHapley Additive exPlanation (SHAP) values[21] with the high-throughput method to better understand and screen stable photocatalytic materials conforming to the HSAB (Hard Soft Acid Base) principle.[22] SHAP values can be used for both global and local interpretation of features, revealing potential relationships between the model's predicted values and certain features.[21] Therefore, the combination of SHAP and ML can not only reveal hidden patterns in the existing data to accelerate the identification of new materials but also reveal the physics of the underlying processes.[23]

In this work, we utilized a new database—Screening for Photovoltaic and Photoelectrochemical (PV and PEC) materials[24] from the Open Quantum Materials Database (OQMD)—to construct the training dataset.[25] We also employed a Gradient Boosted Tree (GBT) algorithm,[20] along with traditional ML methods such as the Extremely Randomized Trees (EXRT)[26] algorithm and Random Forest (RF) algorithm,[27] as well as DL approaches including the Convolutional Neural Network (CNN)[28] and the Backpropagation Neural Network (BPNN).[29] These methods were collectively used to classify the prediction dataset into zero-bandgap materials ($E_g = 0$ eV) and non-zero-bandgap materials ($E_g \neq 0$ eV). Five regression methods are utilized to quantitatively determine the bandgap of non-zero bandgap materials, and the SHAP model is introduced for global and local interpretation of features to find hidden relationships between bandgaps and features. This method enhances prediction accuracy and facilitates model interpretability. The training dataset consists of 7235 materials, from which 101 distinct crystal structures were selected as initial templates. By substituting elements within the same group of the periodic table and performing DFT structural relaxation, a total of 7972 predicted structures were generated. The trained model was then applied to screen these structures and identify potential candidate materials. Finally, DFT calculations were performed on the candidate materials to further validate the stability, connectivity of atomic orbitals at the band edges, optical absorption, and theoretical PCE. $Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$ were successfully screened as highly promising materials for solar cells.

II. RESULTS AND DISCUSSION

A. The procedural framework

As illustrated in Fig. 1, the filtering framework comprises three key stages: dataset construction, ML algorithm application, and validation via DFT calculations. The target bandgap range for the ML model is 0.9–1.6 eV.[18] DFT is primarily used as a tool for validating our machine learning models and for the subsequent screening of materials. The input dataset sourced from a screening of PV and PEC materials in the OQMD. These data have been incorporated into the CMR database. Figure S1(a) in the supplementary material presents representative crystal structures and space groups from the training dataset, while Fig. S1(b) in the supplementary material illustrates the crystal structures of some initial materials used to construct the prediction dataset. In order to cover a broader range of material structures, we selected 101 unique crystal structures from the training dataset as initial templates. By substituting elements from the same group in the periodic table, we generated 7972 inorganic materials, which were then subjected to structural optimization. We utilized the Uniform Manifold Approximation and Projection (UMAP) technique to visualize and analyze the structural relationships between the training dataset and the prediction dataset. UMAP is a stochastic, non-linear dimensionality reduction algorithm that preserves both the local and global structural features of the data.[30] Through cluster analysis, we found that the sample distribution in the prediction dataset falls within the range of the training dataset, ensuring the model's generalization capability (see Fig. S2 in the supplementary material).

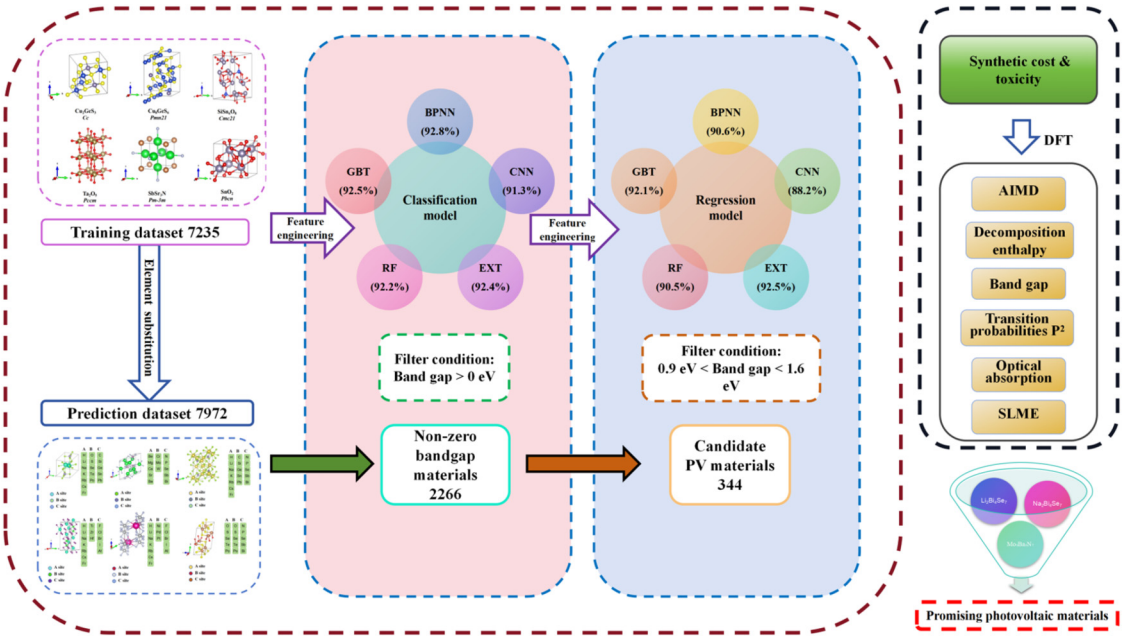

FIG. 1. The designed framework for predicting photovoltaic materials. The left panel (brown box) illustrates the ML workflow, including the construction of the training dataset (purple box) and prediction dataset (steel-blue box), as well as the selection criteria and outcomes for classification models (pink box) and regression models (light-sky-blue box). The right panel (black box) details the screening criteria and subsequent DFT calculations, encompassing transition-dipole moments, thermal stability, and theoretical photoelectric-conversion efficiency.

Once the crystal structures and bandgaps for the training dataset are obtained, features are extracted to explicitly characterize the relationships between input data and targets. A total of 271 initial primary features were constructed using the Voronoi tessellation method through the Material Agnostic Platform for Informatics and Exploration.[31,32] To prevent issues related to overfitting, the recursive feature elimination cross-validation (RFECV) method was applied to remove redundant features.[18] The process of feature engineering is illustrated in the purple box of Fig. S3 in the supplementary material.

After identifying the key features, we performed five rounds of training and validation using randomly generated subsamples. The hyperparameters for five classification models and five regression models were optimized through a cross-validation process to enhance prediction accuracy and reduce prediction instability. Finally, these models were trained and evaluated. Potential photovoltaic materials were selected from Prediction dataset using the trained models, and their results were validated through DFT calculations.

B. MODEL EVALUATION

The model was trained and evaluated by splitting the dataset into two parts: 80% for training and 20% for testing. The training set consists of 5788 materials (2781 zero bandgap and 3007 non-zero bandgap), and the validation set consists of 1447 materials (696 zero bandgap and 751 non-zero bandgap). To assess feature importance and its relationship to the target, we applied RFECV to our classification model and tuned the hyperparameters for each estimator. Neural networks typically do not provide interpretable importance scores. Therefore, we employed three tree-based algorithms (EXRT, GBT, and RF) as estimators within an RFECV framework. The RFECV process identified optimal feature subsets of 65, 61, and 83 features for EXRT, GBT, and RF, respectively. Subsequently, the performance of five classification models was evaluated on each subset, and the results indicated that the 61-feature subset selected by GBT yielded the highest average accuracy across all models. Therefore, these 61 features were ultimately selected as the descriptor dataset (as shown in Fig. S4 in the supplementary material). The feature correlation heatmap in Fig. 2(a) shows that the correlation between most features is very low, with only a handful showing stronger relationships, indicating that the selected key features are largely independent, which is beneficial for improving model performance. Figure S4 in the supplementary material indicates that the average melting temperature, maximum packing efficiency, and the melting temperature of elements in the first coordination shell together account for approximately 60% of the total feature importance. Nevertheless, the remaining features with lower weights played a crucial role in enabling the model to capture more comprehensive information and complex relationships within the data. As illustrated in Fig. 2(b), the confusion matrix's diagonal pattern supports the model's ability to effectively distinguish between zero-bandgap and non-zero-bandgap materials. A true positive is a material that is correctly predicted as a non-zero bandgap. A True Negative is a

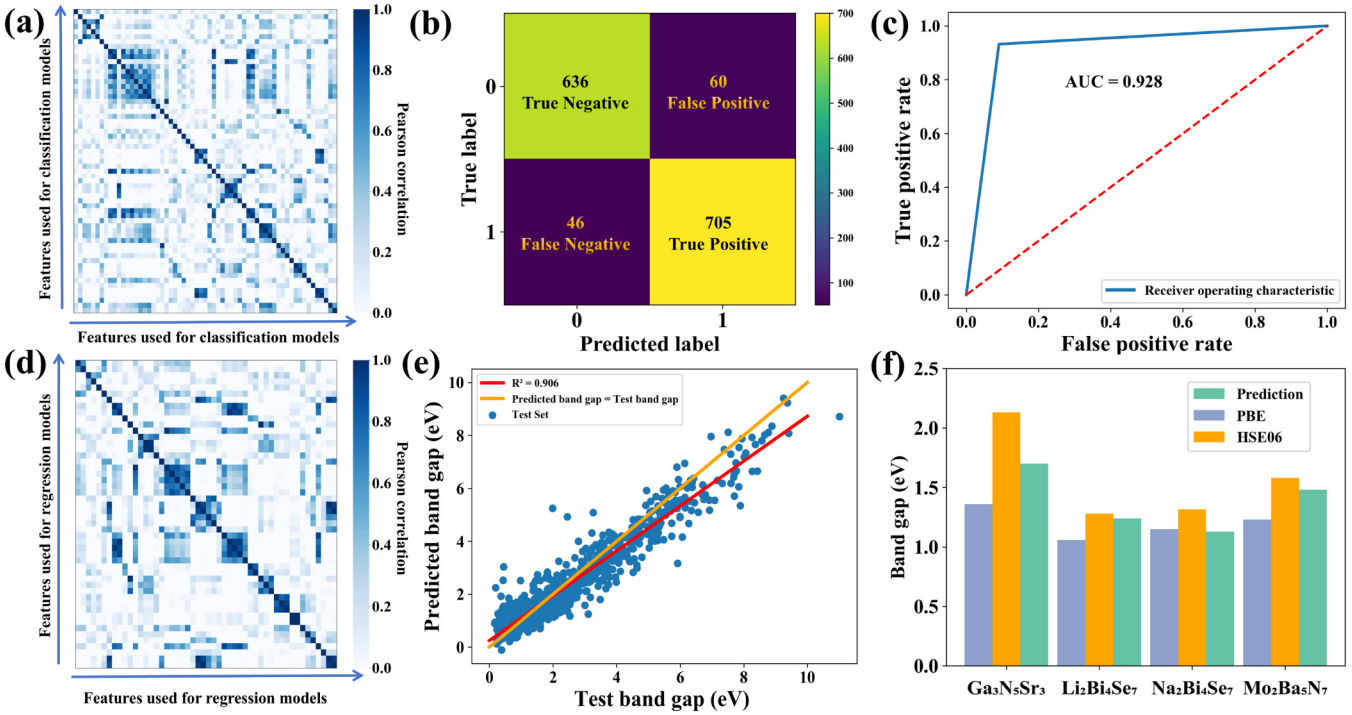

FIG. 2. (a) Correlation heatmap among features for the classification model, (b) confusion matrix for the BPNN classification model, and (c) ROC curve for the BPNN classification model, (d) correlation heatmap among features for the regression model, (e) the performance curve for the BPNN regression model; and (f) comparison of ML predictions with DFT calculations.

material that is correctly predicted as a zero bandgap. The receiver operating characteristic (ROC) curves further demonstrate the model's capability in this task as shown in Fig. 2(c). Perfect separation is represented by an area under the ROC curve (AUC) equal to 1.0, while an AUC of 0.5 (denoted by the dashed line) corresponds to random guessing. For the BPNN classification model, the AUC for distinguishing zero-bandgap from non-zero-bandgap materials was 0.928, indicating excellent classification performance. In comparison, the AUCs for the GBT, RF, EXRT, and CNN classification models were 0.925, 0.922, 0.924, and 0.914, respectively (as detailed in Fig. S6 in the supplementary material). By applying all five trained classification models to the prediction dataset, a total of 2266 compounds were identified as non-zero bandgap materials by all five models in consensus. The bandgaps of these non-zero-bandgap materials were then predicted using regression models.

Since the regression model focuses on fitting continuous bandgap values, we trained it using 3758 samples with non-zero bandgaps. Similarly, because the two tasks have different objectives and evaluation metrics, the ultimately selected feature subsets may differ. For the regression model, we trained using 3758 samples with non-zero bandgaps and employed RFECV to re-evaluate feature importance, while also fine-tuning the hyperparameters for each model. RFECV identified optimal feature subsets of 50, 106, and 36 features for EXRT, GBT, and RF, respectively. Across all five regression models, a subset of 50 features yielded the highest mean predictive accuracy. Therefore, 50 key features were selected and ranked by their importance (as detailed in Fig. S5 in the supplementary material). Figure 2(d) presents a heatmap showing low correlation among most features, which contributes to enhancing the regression model's performance. The evaluation of the BPNN regression model is presented in Fig. 2(e), where the red line represents the model's fitted prediction line and the yellow line indicates the ideal prediction line. The blue dots scattered around the yellow line suggest that the BPNN regression model achieves a strong agreement between the test bandgap and the predicted bandgap. The R^2 for this model is 0.906, which demonstrates the robustness and accuracy of our ML model. In comparison, the R^2 values for the GBT, RF, EXRT, and CNN regression models were 0.921, 0.905, 0.925, and 0.882, respectively (as detailed in Fig. S7 in the supplementary material). Figure 2(f) shows the alignment between the ML-predicted bandgap and the DFT-calculated bandgap for four candidate materials, highlighting the atomic-level precision of our predictions. Through ML, the range of critical features can be derived by analyzing the relationships between critical features and bandgaps (as shown in Fig. S8 in the supplementary material).

C. Geometry, stability, and electronic structures

According to the screening criteria outlined above, 344 candidate materials with suitable bandgaps were obtained. Since materials containing toxic or radioactive elements pose significant threats to

human health and are unsuitable for commercial applications, these candidates were excluded from further consideration. In addition, although an appropriate bandgap is a key determinant of light absorption, it alone does not guarantee optimal absorption performance. We first employed the Perdew–Burke–Ernzerhof (PBE) functional[33] to calculate the band structure and transition dipole moment. Materials exhibiting parity-forbidden transitions at the band edges were excluded, as such phenomena significantly hinder photon absorption (as detailed in Fig. S9 in the supplementary material). Ultimately, eight materials remained: Sb_2S_2, $Na_2P_4Se_7$, $Li_4Sn_2Se_5$, $Ga_3Ca_3N_5$, $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, $Ga_3N_5Sr_3$, and $Mo_2Ba_5N_7$ (as detailed in Fig. S10 in the supplementary material). The corresponding lattice constants of the above materials are listed in Table S1 in the supplementary material. To investigate the stability of these materials, we conducted further evaluations of their thermal stability using DFT calculations.

For the candidate materials, we then employed the Nosé–Hoover method to perform *ab initio* molecular dynamics (AIMD) simulations within an NPT ensemble.[34] The result, shown in Figs. 3(a)–3(c) and Fig. S11 in the supplementary material,

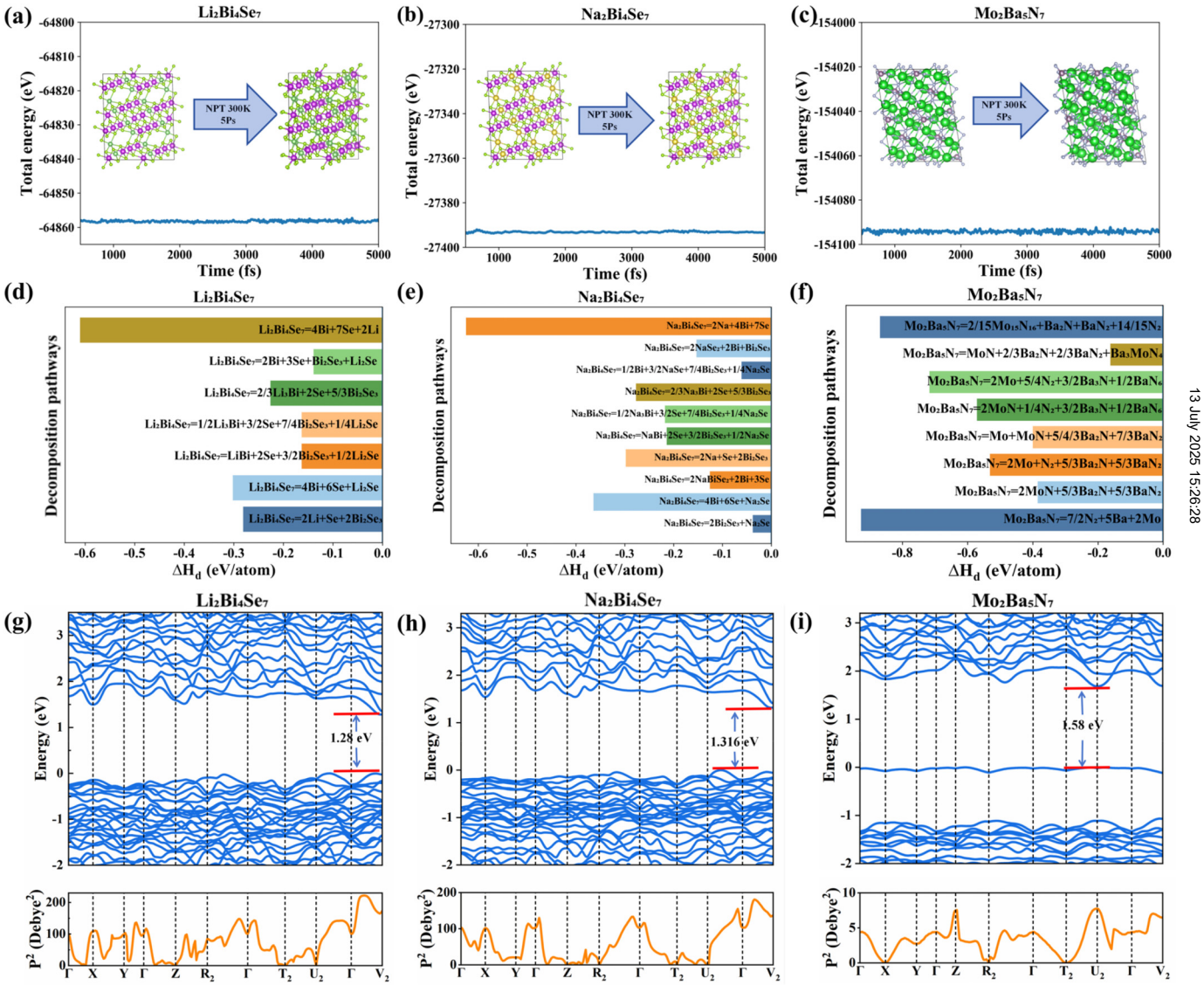

FIG. 3. Total energy changes from *ab initio* molecular dynamics simulations at 300 K for (a) $Li_2Bi_4Se_7$, (b) $Na_2Bi_4Se_7$, and (c) $Mo_2Ba_5N_7$. The decomposition energy path of the three candidate materials, i.e., (d) $Li_2Bi_4Se_7$, (e) $Na_2Bi_4Se_7$, and (f) $Mo_2Ba_5N_7$. Band structure and P^2 for (g) $Li_2Bi_4Se_7$, (h) $Na_2Bi_4Se_7$, and (i) $Mo_2Ba_5N_7$ were calculated using the HSE06 functional.

indicates that the total energy of five materials—$Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, $Mo_2Ba_5N_7$, $Ga_3Ca_3N_5$, and $Ga_3N_5Sr_3$—remained relatively stable at 300 K, suggesting that their structures are stable at room temperature. In contrast, the other three materials exhibited significant structural changes at room temperature, indicating instability, and were therefore excluded from further consideration. To further assess thermal stability, we calculated the decomposition enthalpy (ΔH_d) for the five stable materials along their respective decomposition pathways, as illustrated in Figs. 3(d)–3(f) and Fig. S12 in the supplementary material. A negative ΔH_d value signifies that the candidate materials are thermodynamically stable. For $Ga_3Ca_3N_5$, the positive ΔH_d value suggests the presence of potential decomposition pathways; hence, it was excluded from further consideration.

Additionally, PBE calculations typically underestimate the bandgap values. To improve the accuracy of predictions, the Heyd–Scuseria–Ernzerhof (HSE06)[35] functional was used to evaluate the band structure of the candidate materials with greater precision. The transmission peaks of P^2, which appear near the high symmetry points of the Valence Band Maximum (VBM) or Conduction Band Minimum (CBM), indicate the parity-allowed optical transition probability. From Figs. 3(g)–3(i), we observe that there are no parity-forbidden transitions near the band edges for the three candidate materials, and relatively large P^2 is present. When the photon energy matches the electronic energy levels of the material, the larger P^2 leads to stronger photon–electron interactions, thereby enhancing light absorption efficiency. However, the bandgap of $Ga_3N_5Sr_3$ exceeds 1.6 eV, which may significantly narrow its absorption range of the solar spectrum, thereby reducing its PCE. Therefore, $Ga_3N_5Sr_3$ is excluded from the subsequent verification (as shown in Fig. S13 in the supplementary material).

It is worth noting that our previous work revealed that strong p–p transitions near the band edge play a significant role in enhancing PCE.[18] As shown in Fig. S5 in the supplementary material, in our ML model, the contribution of the p-orbital valence fraction to feature importance is significant, accounting for 26.2%, while the d-orbital valence fraction contributes 6.2%. Therefore, p-orbitals are crucial for achieving high PCE materials, while d-orbitals also play a role in photon absorption and overall photovoltaic performance. Additionally, considering the interaction and dependency effects between features, other characteristics such as the ionic nature of interatomic bonds play a crucial role in enhancing the model's accuracy for bandgap prediction.

Based on this conclusion, we used SHAP values to provide both global and local interpretability of features, further elucidating the "black box" nature of the ML model and its ability to uncover hidden relationships between features and target properties.

D. The Shapley additive explanation

Figure 4(a) displays the scatterplot of feature density for all samples, illustrating the impact of different features on the model output. Features are sorted by their mean absolute SHAP values, with the color bar on the right representing the magnitude of each feature's value. The results confirm that the p-orbital and d-orbital valences are the most influential characteristics, as valence electron orbitals directly affect the material's energy-level transitions, which, in turn, have a significant influence on the bandgap. We use the SHAP dependency plots to explore how the p-orbital valence fraction and d-orbital valence fraction jointly influence the bandgap, as shown in Fig. S14(a) in the supplementary material. The results indicate a positive correlation between the p-valence electron fraction and the bandgap, and a negative correlation between the d-valence electron fraction and the bandgap. As the occupancy of p-orbital valence fraction increases, the predicted bandgap typically increases. This may be attributed to the enhanced interaction at the band edges by the p orbital electrons, leading to a larger band gap.

For the prediction dataset with suitable bandgaps, a SHAP value model was further established. As shown in Fig. 4(b), it was found that, in addition to the p-orbital valence fraction, the minimum effective coordination number has the highest importance. We used a SHAP dependence plot to explore how the minimum effective coordination number and p-orbital valence fraction jointly affect the bandgap, illustrating their synergistic interaction, as shown in Fig. 4(c). The results show that when the minimum effective coordination number is less than 6, a lower effective coordination number tends to increase the bandgap, while a higher effective coordination number tends to decrease it. Perhaps the most significant reason is that as the coordination number decreases, the number of surrounding atoms or ions is reduced, and the proportion of valence electrons in p-orbitals increases, possibly leading to enhanced electron coupling. These factors collectively enhance the interactions at the band edges, thereby causing an increase in the bandgap. Once the minimum effective coordination number exceeds 6, the predicted bandgap value becomes nearly independent of it. This could be attributed to the fact that, when the coordination number increases to a certain level, the bonding interactions involving the atoms or ions in the local environment become nearly fully saturated. At this stage, further increases in the coordination number have little impact on the orbital interactions or electron distribution. Under these conditions, the variation in the bandgap with coordination number can be considered negligible.

Figures 4(d)–4(f) provide locally interpretable SHAP plots for $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, and $Mo_2Ba_5N_7$. In all cases, the fraction of p-orbital valence shifts the predicted values lower than the baseline (the average bandgap values from the training dataset). Figures 4(g)–4(i) further explore the contribution of local features to material properties. The bars start from zero (baseline) at the bottom and illustrate how the predicted target value increases (red) or decreases (blue) with different features to reach the predicted bandgap values. For $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, and $Mo_2Ba_5N_7$, the fraction of p-orbital valence has a significant influence, which is consistent with the SHAP global feature importance map. However, for $Mo_2Ba_5N_7$, "CanFormIonic" also has a highly significant impact on the predicted bandgap values. In general, a larger electronegativity difference between elements increases the energy gap between their atomic orbitals, resulting in a higher bandgap for the compound [as shown in Fig. S14(b) in the supplementary material].[6] Additionally, the minimum effective coordination number pushes the predicted bandgap values higher for $Mo_2Ba_5N_7$, while for other candidate materials, it drives the predicted bandgap values lower. These results indicate that the coordination environment of atoms in the material can significantly affect the interaction strength among internal atoms, thereby influencing electronic behavior and

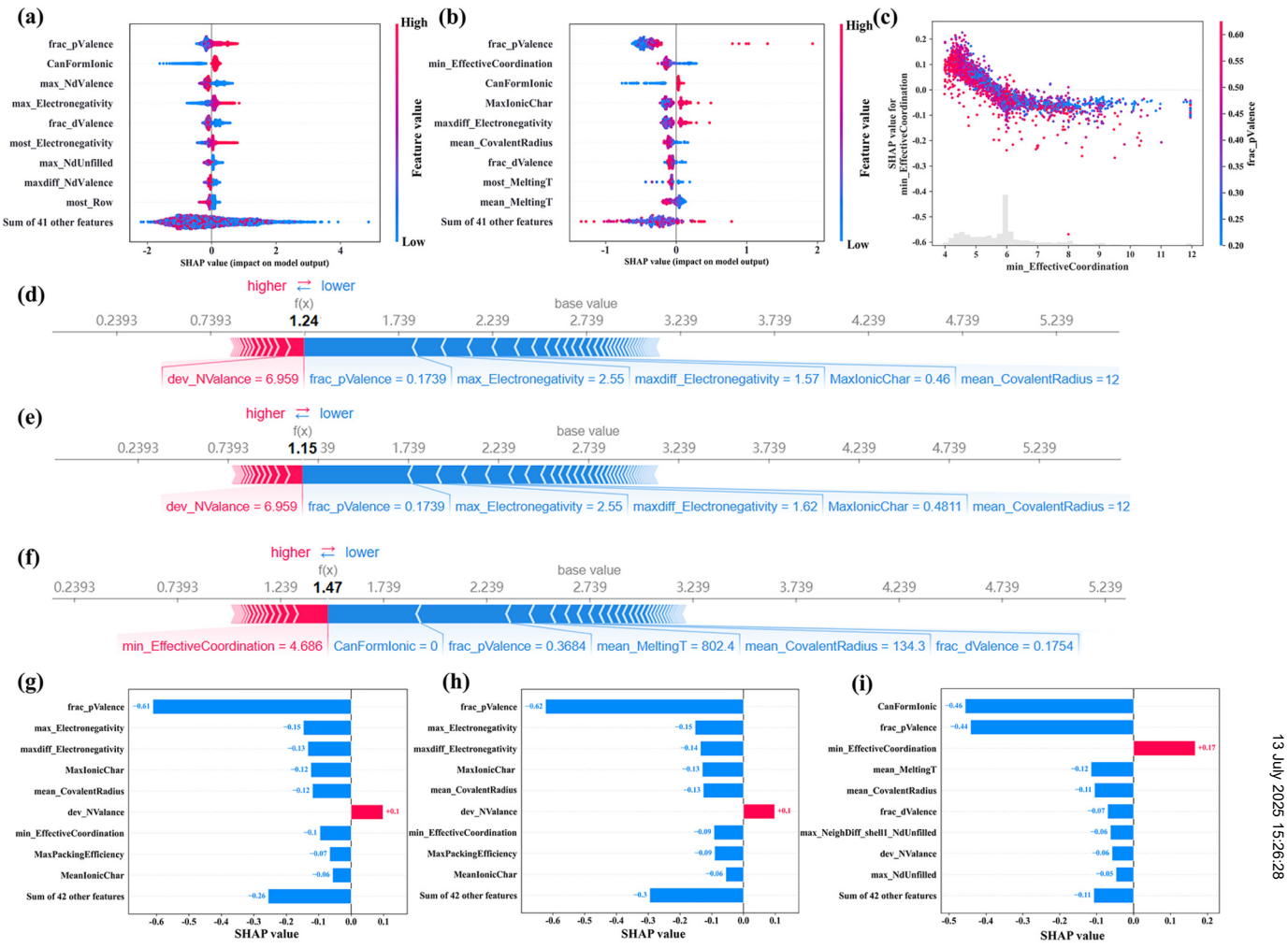

FIG. 4. (a) Global interpretability of E_g regression in the training dataset using SHAP. (b) Global interpretability of E_g regression in the prediction dataset using SHAP. (c) SHAP dependence plots for min_Effective Coordination (Minimum effective coordination number) features for E_g. Local interpretability into E_g regression using SHAP. Individual SHAP plots using the E_g for (d) $Li_2Bi_4Se_7$, (e) $Na_2Bi_4Se_7$, and (f) $Mo_2Ba_5N_7$. Local feature importance plot using the E_g for (g) $Li_2Bi_4Se_7$, (h) $Na_2Bi_4Se_7$, and (i) $Mo_2Ba_5N_7$.

modulating the bandgap. SHAP analysis provides insights into how ML models recognize target attributes (e.g., bandgaps) based on features. This not only aids in a deeper understanding of the decision-making mechanisms within the model but also broadens its applicability and interpretability in the study of inorganic compounds.

E. DOS, absorption spectra, and Spectroscopic Limited Maximum Efficiency

To explore the electronic structure of these materials, Figs. 5(a)–5(c) present the density of states (DOS) for the three candidates. For $Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$, the VBM is primarily contributed by the *p*-orbitals of Se atoms, while the CBM is dominated by the *p*-orbitals of Bi atoms [Figs. 5(a) and 5(b)]. A similar phenomenon is also observed in the stibnite family of semiconductor metal chalcogenides (A_2B_3, where A = Sb, Bi and B = S, Se).[36] For $Mo_2Ba_5N_7$ [Fig. 5(c)], both the VBM and CBM are primarily contributed by the *d*-orbitals of Mo and Ba atoms. A higher DOS near the band edges suggests a greater number of available electronic states, allowing more electrons to be excited by a given photon energy, thereby enhancing light absorption. Therefore, all three materials have a high potential for light absorption. The electrons in the *p*- and *d*-orbitals play a key role in contributing to the VBM and CBM of these materials, aligning with the main features identified by the SHAP model.

Furthermore, DFT calculations were performed on the three candidate materials to verify the connectivity of atomic orbitals at the band edges, optical absorption properties, and theoretical PCE.

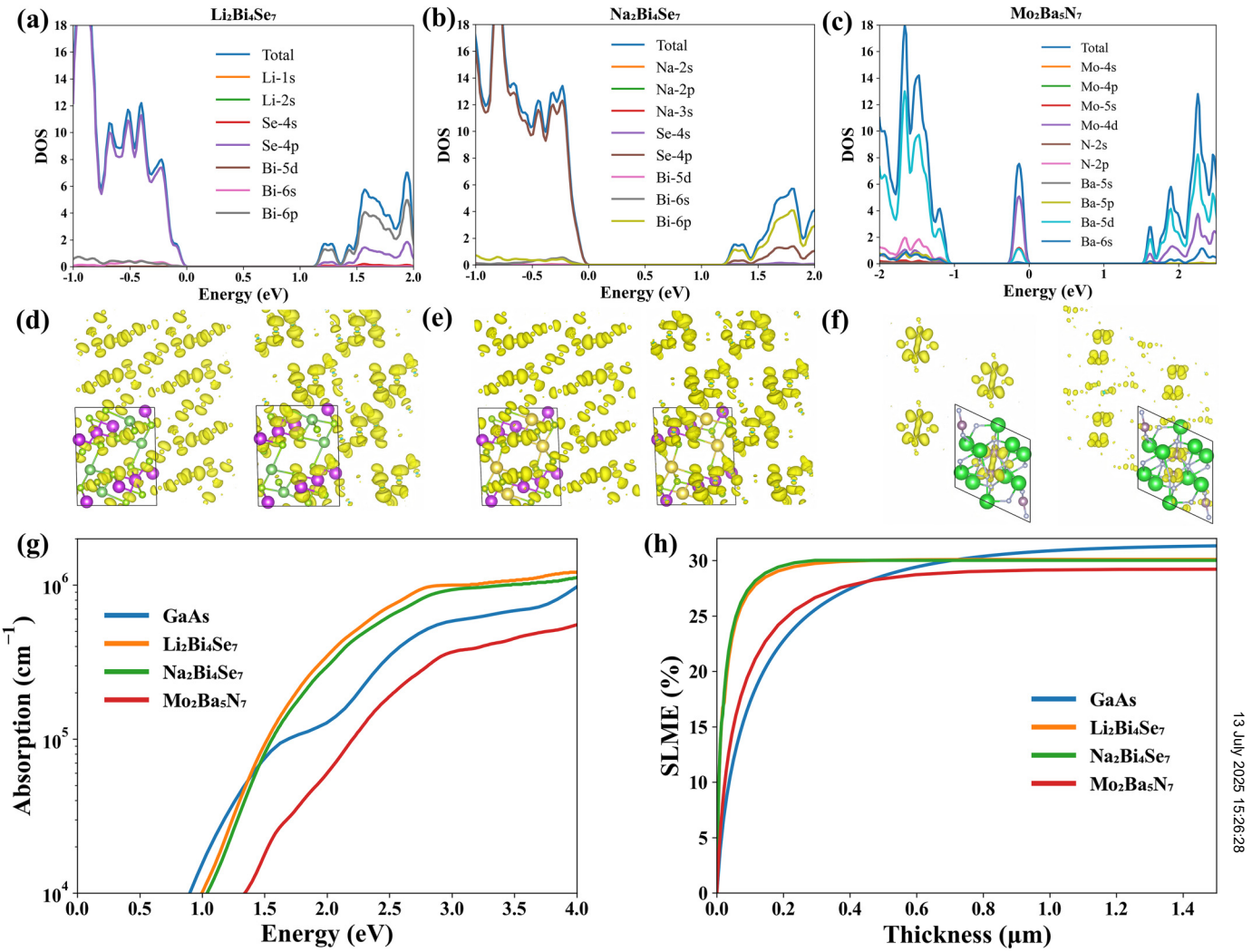

FIG. 5. Projected density of states for (a) $Li_2Bi_4Se_7$, (b) $Na_2Bi_4Se_7$, and (c) $Mo_2Ba_5N_7$ were calculated using the HSE06 functional, The corresponding wave function distributions of the VBM and CBM for (d) $Li_2Bi_4Se_7$, (e) $Na_2Bi_4Se_7$, and (f) $Mo_2Ba_5N_7$. (g) Photon absorption spectra of GaAs, $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, and $Mo_2Ba_5N_7$. (h) Spectroscopic Limited Maximum Efficiency (SLME) of GaAs, $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, and $Mo_2Ba_5N_7$.

The charge densities corresponding to the CBM (right) and VBM (left) of $Li_2Bi_4Se_7$ [Fig. 5(d)] and $Na_2Bi_4Se_7$ [Fig. 5(e)] are observed to be more uniformly distributed across the band edges throughout the entire crystal. This uniform distribution indicates higher electronic dimensionality, combined with large band edge dispersion, which, in turn, leads to smaller effective masses for charge carriers. These properties facilitate carrier transport, improve carrier separation efficiency, and provide greater opportunity of achieving high photovoltaic performance. Compared to the benchmark material GaAs for solar cells, $Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$ exhibit comparable optical absorption coefficients in the visible light range [Fig. 5(g)]. Their theoretical PCEs are relatively high, with $Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$ reaching 30.1% and 30.0%, respectively [Fig. 5(h)]. This is likely due to their bandgaps being nearer to the ideal value (1.34 eV) as predicted by the Shockley–Queisser limit, enabling more effective photon absorption in the visible spectrum.[18] In contrast, the Mo-4d and Ba-5d atomic orbitals in $Mo_2Ba_5N_7$ are not connected at the band edges, indicating a 0D electronic dimensionality [Fig. 5(f)]. The band structure is non-dispersive at the VBM, resulting in a very low hole mobility for $Mo_2Ba_5N_7$. Interestingly, for $Mo_2Ba_5N_7$, the effective coordination number feature shifts the predicted values to the right of the baseline [Fig. 4(f)]. Its coordination number, 4.6, is lower than those of the other two materials, $Li_2Bi_4Se_7$ (6.32) and $Na_2Bi_4Se_7$ (6.07). A low coordination number leads to more localized electron clouds, which weakens the connectivity of atomic orbitals at the CBM and VBM, blocking the

transport channels. Consequently, the carrier mobility in $Mo_2Ba_5N_7$ is relatively low. Therefore, these two materials ($Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$), with high PCEs and ideal bandgaps, are promising candidates for high-performance solar cell applications.

III. CONCLUSION

In summary, we explore a ML- and DL-assisted framework to accelerate the discovery of high-efficiency inorganic photovoltaic materials, focusing on predicting bandgaps for potential solar cell materials. We employed several ML models (GBT, RF, and EXRT) and DL models (BPNN and CNN) to classify inorganic materials based on whether their bandgap is zero or non-zero. For materials with non-zero bandgaps, regression models are subsequently applied to quantify the bandgap values. The model's performance was enhanced through dimensionality reduction and hyperparameter optimization, while SHAP value analysis provided interpretability in the model prediction process. By screening 7972 inorganic materials, the study identified two promising candidates—$Li_2Bi_4Se_7$ and $Na_2Bi_4Se_7$—with PCEs of 30.1% and 30.0%, respectively. The inorganic solar cell material is more stable than the organic one. The results validate the potential of the ML-assisted framework in discovering novel inorganic photovoltaic materials with ideal solar cell performance and provide a reference for experiments.

SUPPLEMENTARY MATERIAL

See the supplementary material for the details of the calculations, including the representative crystal structures of the training and prediction datasets, feature importance rankings for classification and regression models, prediction results and their feature descriptions, visualizations of training data for classification and regression models, AIMD simulations of the total energy changes in candidate materials, decomposition energy pathways, band structures, and the corresponding lattice constants.

ACKNOWLEDGMENTS

This work was supported by the NSFC under Grant No. 12475270 and the Innovation Capability Support Program of Shaanxi under Grant No. 2025RS-CXTD-029.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

R.-T.C. and Z.H. contributed equally to this paper.

Ruo-Tong Chen: Data curation (lead); Formal analysis (equal); Investigation (lead); Methodology (equal); Software (lead); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Zhihua Hu:** Data curation (equal); Formal analysis (equal); Investigation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Hong-Jian Feng:** Funding acquisition (lead); Project administration (lead); Supervision (lead); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]A. Polman, M. Knight, E. C. Garnett, B. Ehrler, and W. C. Sinke, "Photovoltaic materials: Present efficiencies and future challenges," Science. **352**(6283), aad4424 (2016).
[2]M. Riede, D. Spoltore, and K. Leo, "Organic solar cells—The path to commercial success," Adv. Energy Mater. **11**(1), 2002653 (2021).
[3]I. Massiot, A. Cattoni, and S. Collin, "Progress and prospects for ultrathin solar cells," Nat. Energy **5**(12), 959–972 (2020).
[4]J. Jeong, M. Kim, J. Seo, H. Lu, P. Ahlawat, A. Mishra, Y. Yang, M. A. Hope, F. T. Eickemeyer, M. Kim, Y. J. Yoon, I. W. Choi, B. P. Darwich, S. J. Choi, Y. Jo, J. H. Lee, B. Walker, S. M. Zakeeruddin, L. Emsley, U. Rothlisberger, A. Hagfeldt, D. S. Kim, M. Grätzel, and J. Y. Kim, "Pseudo-halide anion engineering for α-FAPbI3 perovskite solar cells," Nature **592**(7854), 381–385 (2021).
[5]Z. Xiao, W. Meng, J. Wang, D. B. Mitzi, and Y. Yan, "Searching for promising new perovskite-based photovoltaic absorbers: The importance of electronic dimensionality," Mater. Horiz. **4**(2), 206–216 (2017).
[6]G. Tang, P. Ghosez, and J. Hong, "Band-edge orbital engineering of perovskite semiconductors for optoelectronic applications," J. Phys. Chem. Lett. **12**(17), 4227–4239 (2021).
[7]Q. He, B. Yu, Z. Li, and Y. J. E. Zhao, "Density functional theory for battery materials," Energy Environ. Mater. **2**(4), 264–279 (2019).
[8]R. A. Friesner, "*Ab initio* quantum chemistry: Methodology and applications," Proc. Natl. Acad. Sci. **102**(19), 6648–6653 (2005).
[9]L. Zhu, J. Zhou, and Z. Sun, "Materials data toward machine learning: Advances and challenges," J. Phys. Chem. Lett. **13**(18), 3965–3977 (2022).
[10]M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science **349**(6245), 255–260 (2015).
[11]Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
[12]C. Lv, X. Zhou, L. Zhong, C. Yan, M. Srinivasan, Z. W. Seh, C. Liu, H. Pan, S. Li, Y. Wen, and Q. Yan, "Machine learning: An advanced platform for materials development and state prediction in lithium-ion batteries," Adv. Mater. **34**(25), 2101474 (2022).
[13]Y. Gan, G. Wang, J. Zhou, and Z. Sun, "Prediction of thermoelectric performance for layered IV–V–VI semiconductors by high-throughput *ab initio* calculations and machine learning," npj Comput. Mater. **7**(1), 176 (2021).
[14]A. Chen, X. Zhang, and Z. Zhou, "Machine learning: Accelerating materials development for energy storage and conversion," Infomat **2**(3), 553–576 (2020).
[15]H.-J. Feng and P. Ma, "Machine learning prediction of 2D perovskite photovoltaics and interaction with energetic ion implantation," Appl. Phys. Lett. **119**(23), 231902 (2021).
[16]Z.-H. Sun, L.-D. Zhang, and H.-J. Feng, "Composition engineering guided experimental fabrication of $Cs_{(1-n)}AnPb_{(1-m)}B_mX_3$ via machine learning for high-efficiency solar cells," Phys. Lett. A **529**, 130065 (2025).
[17]J. Park, M. Kim, H. Kim, J. Lee, I. Lee, H. Park, A. Lee, K. Min, and S. Lee, "Exploring the large chemical space in search of thermodynamically stable and mechanically robust MXenes via machine learning," Phys. Chem. Chem. Phys. **26**(14), 10769–10783 (2024).
[18]H.-J. Feng, K. Wu, and Z.-Y. Deng, "Predicting inorganic photovoltaic materials with efficiencies >26% via structure-relevant machine learning and density functional calculations," Cell Rep. Phys. Sci. **1**(9), 100179 (2020).
[19]See https://cmr.fysik.dtu.dk for "Computational Materials Repository," which includes the Screening for Photovoltaic and Photoelectrochemical (PV and PEC) materials database.

[20]J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat. **29**(5) 1189–1232 (2001).
[21]S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst. **30**, 4768–4777 (2017).
[22]R. Kumar and A. K. Singh, "Chemical hardness-driven interpretable machine learning approach for rapid search of photocatalysts," npj Comput. Mater. **7**(1), 197 (2021).
[23]S. Zhang, T. Lu, P. Xu, Q. Tao, M. Li, and W. Lu, "Predicting the formability of hybrid organic–inorganic perovskites via an interpretable machine learning strategy," J. Phys. Chem. Lett. **12**(31), 7423–7430 (2021).
[24]K. Kuhar, M. Pandey, K. S. Thygesen, and K. W. Jacobsen, "High-throughput computational assessment of previously synthesized semiconductors for photovoltaic and photoelectrochemical devices," ACS Energy Lett. **3**(2), 436–446 (2018).
[25]J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," JOM **65**(11), 1501–1509 (2013).
[26]P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Mach. Learn. **63**, 3–42 (2006).
[27]L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).
[28]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**(11), 2278–2324 (1998).
[29]D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature **323**(6088), 533–536 (1986).
[30]L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform manifold approximation and projection," J. Open Source Softw. **3**(29), 861 (2018).
[31]L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, "Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations," Phys. Rev. B **96**(2), 024104 (2017).
[32]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**(1), 1–7 (2016).
[33]J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**(18), 3865–3868 (1996).
[34]S. Nosé, "A unified formulation of the constant temperature molecular dynamics methods," J. Chem. Phys. **81**(1), 511–519 (1984).
[35]A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, "Influence of the exchange screening parameter on the performance of screened hybrid functionals," J. Chem. Phys. **125**(22), 224106 (2006).
[36]M. R. Filip, C. E. Patrick, and F. Giustino, "GW quasiparticle band structures of stibnite, antimonselite, bismuthinite, and guanajuatite," Phys. Rev. B **87**(20), 205125 (2013).

Supplementary Information

Predicting Efficient and Stable Inorganic Photovoltaic Materials Using Interpretable Machine Learning Combined with DFT Calculations Based on Band Edge Orbital Engineering

Ruo-Tong Chen,[#] Zhihua Hu,[#] and Hong-Jian Feng,[*]

School of Physics, Northwest University, Xi'an 710127, China

#Equal contribution

Computational details

In the scikit-learn package,[1] **Gradient Boosting Decision Trees, Random Forest, and Extremely Randomized Trees** machine learning algorithms are implemented. The Keras interface in TensorFlow is used to build **Back propagation neural network** and **Convolutional neural network** deep learning algorithms.[2]

Gradient Boosted Decision Trees

Gradient Boosting Decision Trees is an ensemble machine learning algorithm that uses gradient boosting to combine multiple weak tree learners into a strong learner.[3] The model is trained iteratively, with each iteration generating a weak learner. This process gradually reduces the loss function value on the training data until a predefined stopping criterion is met.

Random forest

Random Forest is an ensemble model composed of numerous decision trees. Its core concept involves inputting the training data into the model, where a sampling process is applied to the training dataset. Different subsets of data and feature attributes are used to build multiple smaller decision trees, which are then combined into a more powerful model.[4]

Extremely Randomized Trees

The Extra-Trees algorithm constructs a collection of unpruned decision or regression trees based on the classical top-down process.[5]

Back propagation neural network

The process of Back Propagation Neural Network is mainly divided into two stages, the first stage is the forward propagation of the signal, from the input layer through the hidden layer, and finally to the output layer; The second phase is the backpropagation of errors to reverse-update the weights, thereby enabling the classification and regression of samples.[6]

Convolutional neural network

Convolutional Neural Network is a deep learning model, the principle of which is mainly to extract the features of the input data through convolutional operations and pooling operations, and to learn and classify the features through the multi-layer neural network structure.[7]

Voronoi tessellation

Voronoi tessellation divides a crystal into small regions determined by the nearest neighboring atoms.[8] This partitioning is influenced solely by the crystal structure. Since each face of the

Voronoi polyhedron is related to its corresponding nearest neighbor atom, Voronoi tessellation can clearly describe the local crystal structure of materials and identify structural changes occurring in polycrystalline structures. Through Voronoi tessellation, various properties of crystals can be represented, including effective coordination number, structural inhomogeneity, chemical order, maximum packing efficiency, local environment, and its compositional attributes. These characteristics can distinctly differentiate the structural and compositional properties of different crystals.

Feature Construction

The 271 features in our manuscript are drawn from two principal sources—composition-dependent attributes based on elemental properties and Voronoi-tessellation-derived attributes based on crystal structure and can be summarized as follows. The composition-dependent features fall into four categories: (1) stoichiometric attributes, which depend only on the relative fractions of each element (not their identities); (2) elemental property attributes, computed as the mean, maximum, minimum, range, mode, and mean absolute deviation of 22 distinct elemental properties; (3) valence-orbital attributes, reflecting the fractional contributions of s, p, d, and f electrons of the constituent elements; and (4) ionicity attributes, including both the feasibility of forming a neutral ionic compound under a given composition and a "likeness" metric based on electronegativity differences. The Voronoi-based structural features comprise five categories: (1) crystal-system descriptors; (2) effective coordination numbers, defined as functions of the Voronoi face areas; (3) structural heterogeneity metrics, which quantify the variation in local bonding-environment shapes; (4) chemical ordering attributes, derived from Warren-Cowley parameters that measure deviations from a random atomic distribution; (5) maximum packing efficiency, given by the sum of the largest possible atomic spheres divided by the cell volume; and (6) local-environment attributes, calculated as the face-area-weighted average of the absolute differences in elemental properties between each atom and its neighbors.

DFT Calculations

First-principles calculations were performed using the PWmat package.[9] Our DFT calculations were performed using the Projector Augmented Wave (PAW)[10] method. The exchange-correlation functionals employed were the Perdew-Burke-Ernzerhof (PBE)[11] and

Heyd-Scuseria-Ernzerhof (HSE06)[12] functionals for different electronic structure calculations. The charge density distribution at the material's band edges was qualitatively analyzed using the Quantum ESPRESSO code.[13] The plane wave basis has a cutoff energy of 450 eV and is convergent for all DFT calculations. An energy convergence threshold of 10^{-5} eV is used for all structural optimizations. k-point sampling is performed using a 2 × 2 × 2 M-P grid. The atomic force criterion during structural relaxation is 0.005 eV/Å. To evaluate the thermodynamic stability of the predicted materials, we performed AIMD simulations with 5000 fs using PWmat. The time step is 1 fs. 2 × 2 × 2 supercells are used for AIMD simulations. Constant pressure NPT system is used and the temperature is controlled at 300 K.

Absorption Coefficient

In order to calculate the optical properties of the material, the absorption coefficient is defined as follows:[14]

$$\alpha(E)=\frac{2E}{\hbar c}\sqrt{\frac{\sqrt{\left(\varepsilon^{(1)}(E)\right)^2}+\sqrt{\left(\varepsilon^{(2)}(E)\right)^2}-\left(\varepsilon^{(1)}(E)\right)}{2}} \quad \text{(S1)}$$

where c is the speed of light, $\varepsilon_{\alpha\beta}^{(1)}$ is the real part of the dielectric tensor and $\varepsilon_{\alpha\beta}^{(2)}$ is the imaginary part of the dielectric tensor.

Spectroscopic limited maximum efficiency (SLME)

The Python3 implementation of the Spectral Limit Maximum Efficiency (SLME) analysis based on solar energy absorbers.[15, 16] SLME η is the ratio of the maximum output power density P_{max} to the total incident solar energy density P_{in}. The defining equation is as follows

$$\eta=\frac{P_{\max}}{P_{in}}=\frac{\max\left\{\left(J_{sc}-J_0\left(e^{eV/kT}-1\right)\right)V\right\}_V}{\int_0^{\infty}EI_{sun}(E)dE} \quad \text{(S2)}$$

Where I_{sun} represents the photo flux, T is the temperature, J denotes the current density, and V is the voltage.

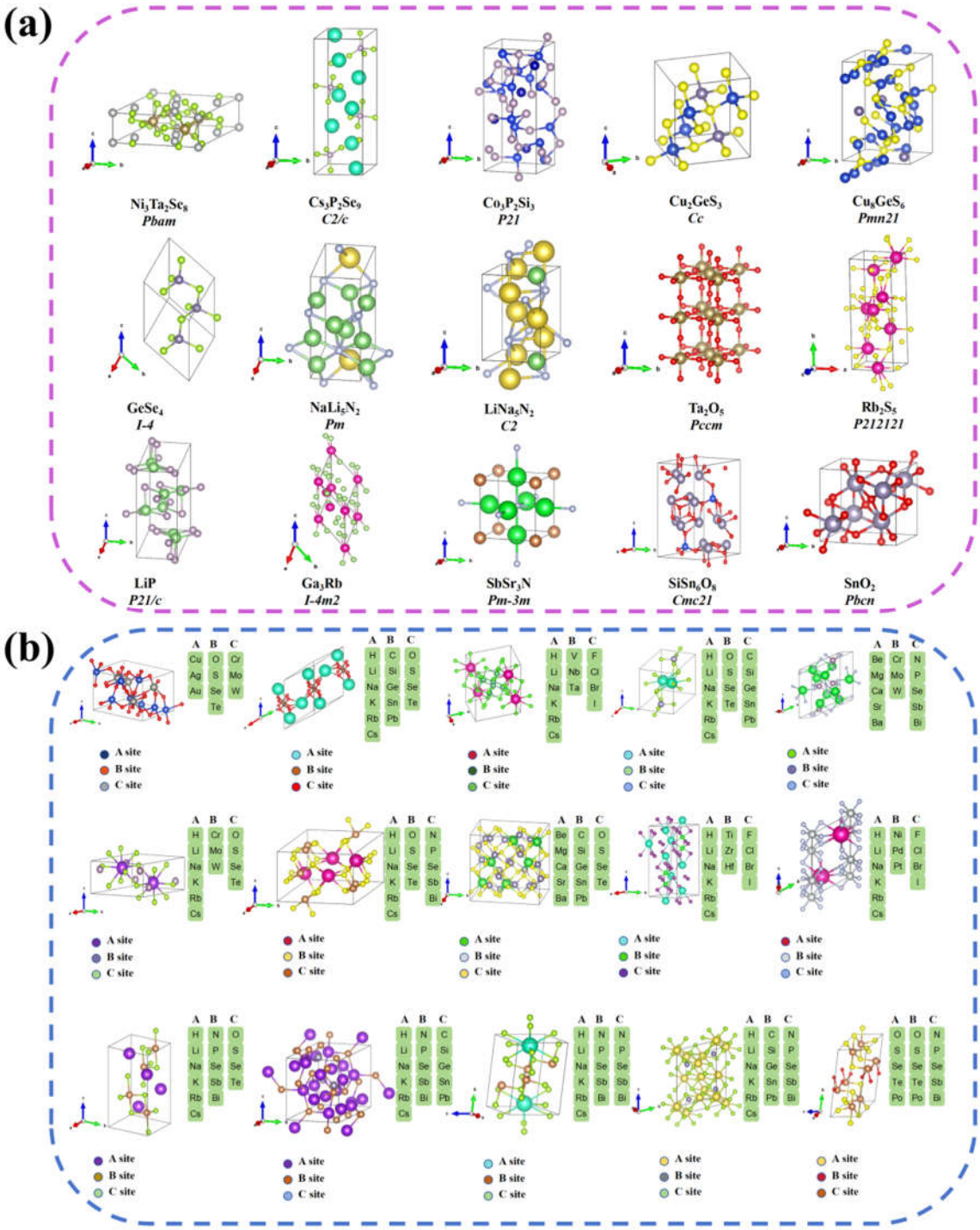

Figure S1. Representative crystal structures are shown for (a) the training datasets (purple box) and (b) the prediction datasets (blue box).

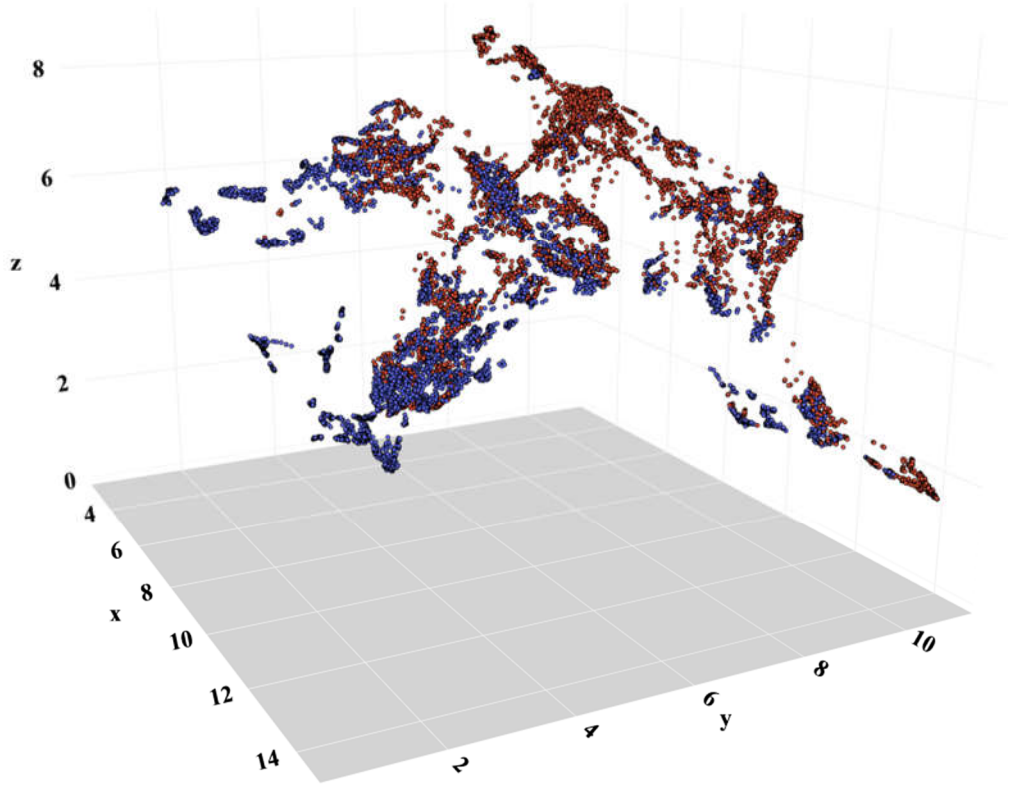

Figure S2. The training dataset (red) and test dataset (blue) are visualized using the UMAP method.

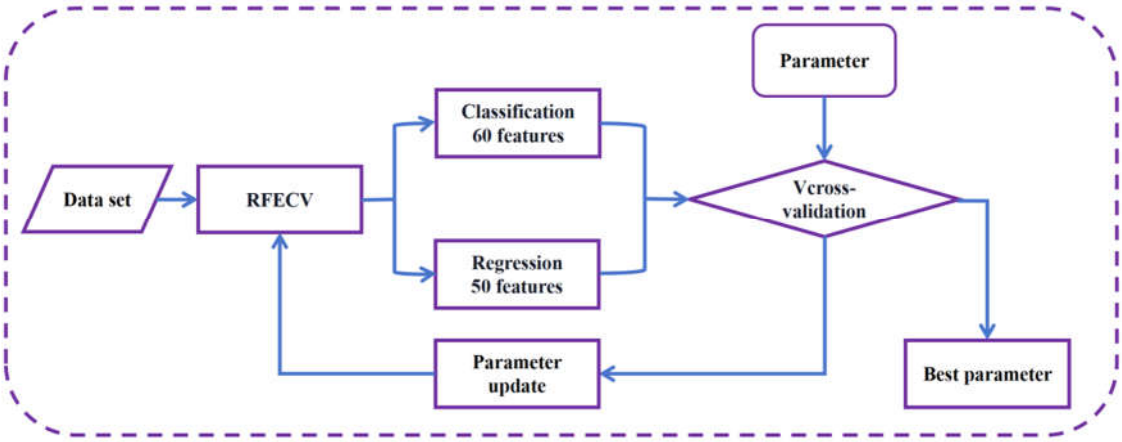

Figure S3. Feature engineering is performed based on the target attributes, and after obtaining the key attributes, the hyper-parameters of the algorithm are determined by a cross-validation procedure.

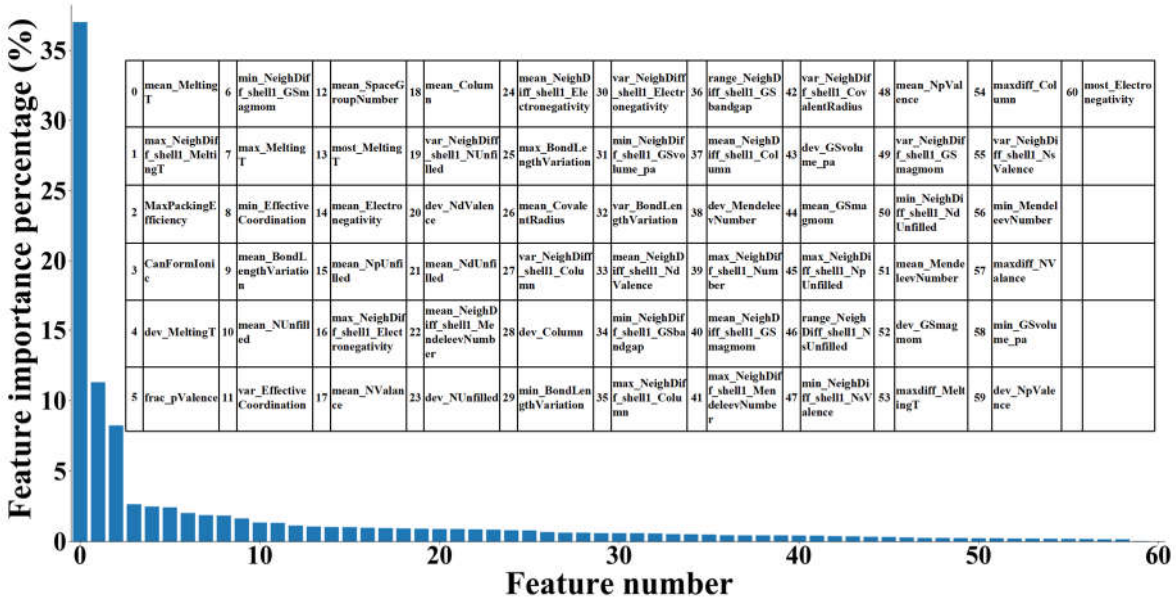

Figure S4. The ranking of importance of 61 features for classification models.

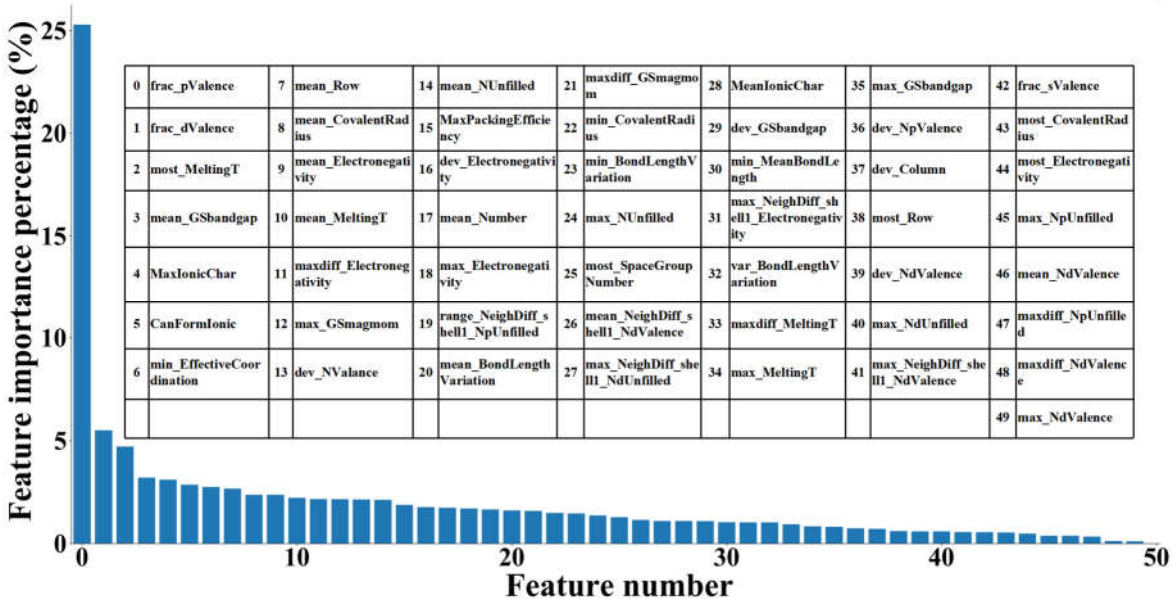

Figure S5. The ranking of importance of 50 features for regression models.

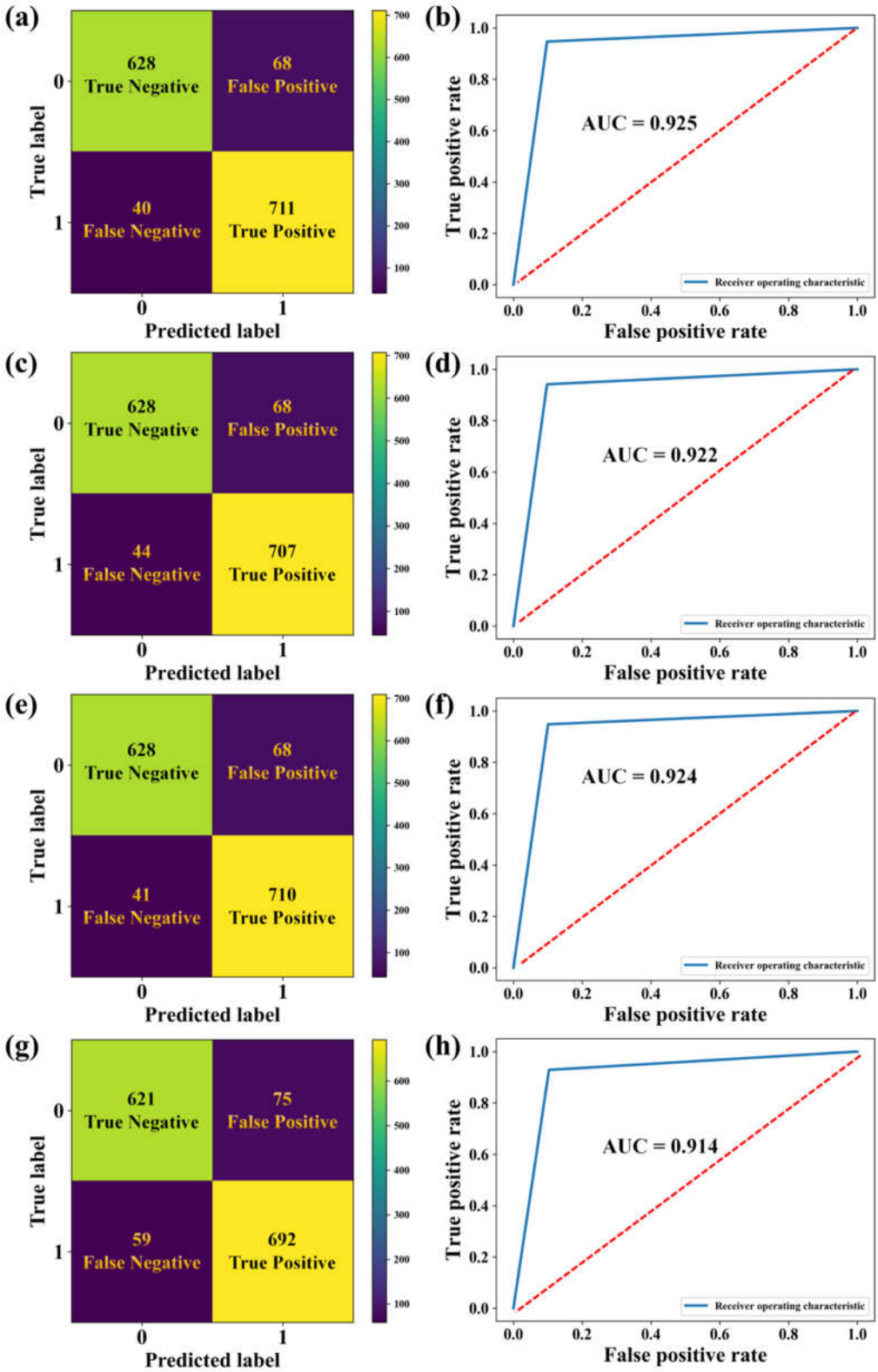

Figure S6. (a) Confusion matrix for the EXTC model. **(b)** Subject job characteristics (ROC) curves for the EXTC model. **(c)** Confusion matrix for the GBTC model. **(d)** Subject job characteristics (ROC) curves for the GBC model. **(e)** Confusion matrix for the RFC model. **(f)** Subject job characteristics (ROC) curves for the RFC model. **(g)** Confusion matrix for the CNNC model. **(h)** Subject job characteristics (ROC) curves for the CNNC model.

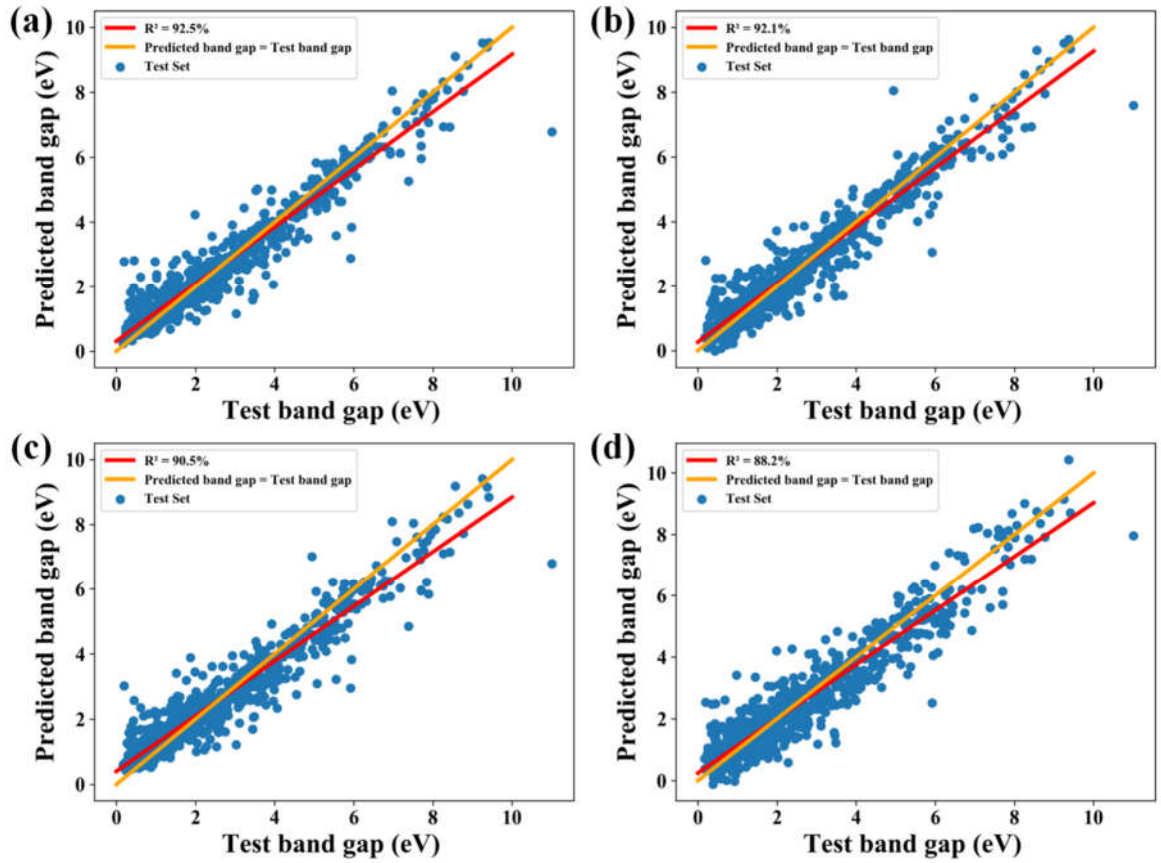

Figure S7. (a) Evaluation of EXTR models. **(b)** Evaluation of GBTR models. **(c)** Evaluation of RFR models. **(d)** Evaluation of CNNR model.

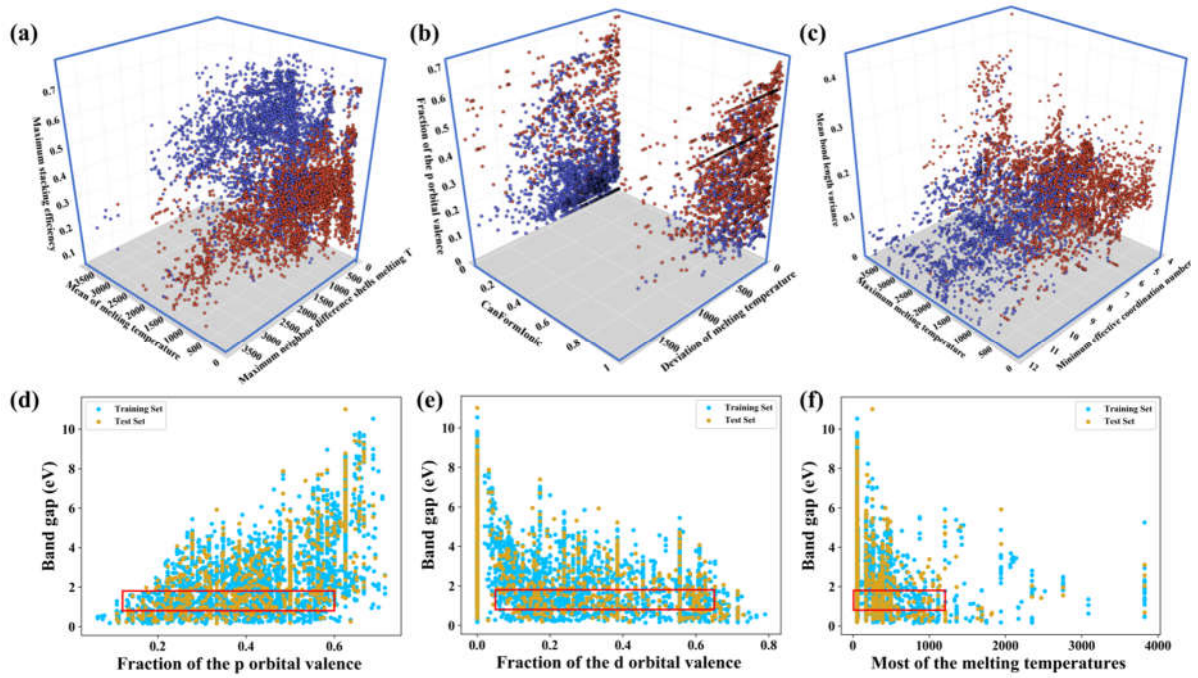

Figure S8. Visualization of classification model training data (red: bandgap≠0; purple: bandgap=0). Visualization of training data for the regression model. Red boxes indicate the range of values for (d) fraction of p-orbital valence, (e) fraction of d-orbital valence, and (f) most of the melting temperatures

within the selected bandgap, respectively.

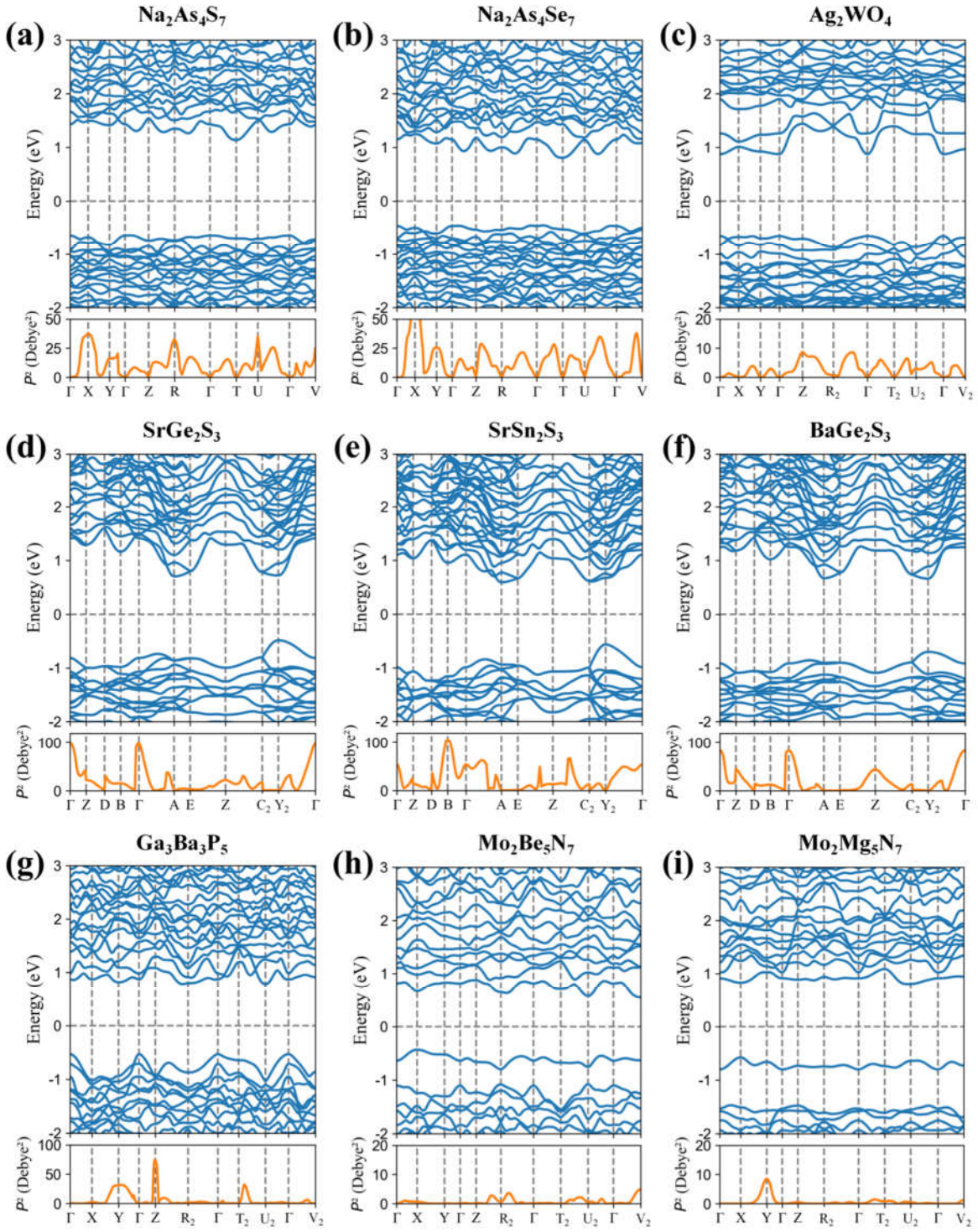

Figure S9. Band structure and P^2 of materials exhibiting parity-forbidden transitions at the band edges were calculated using the PBE functional.

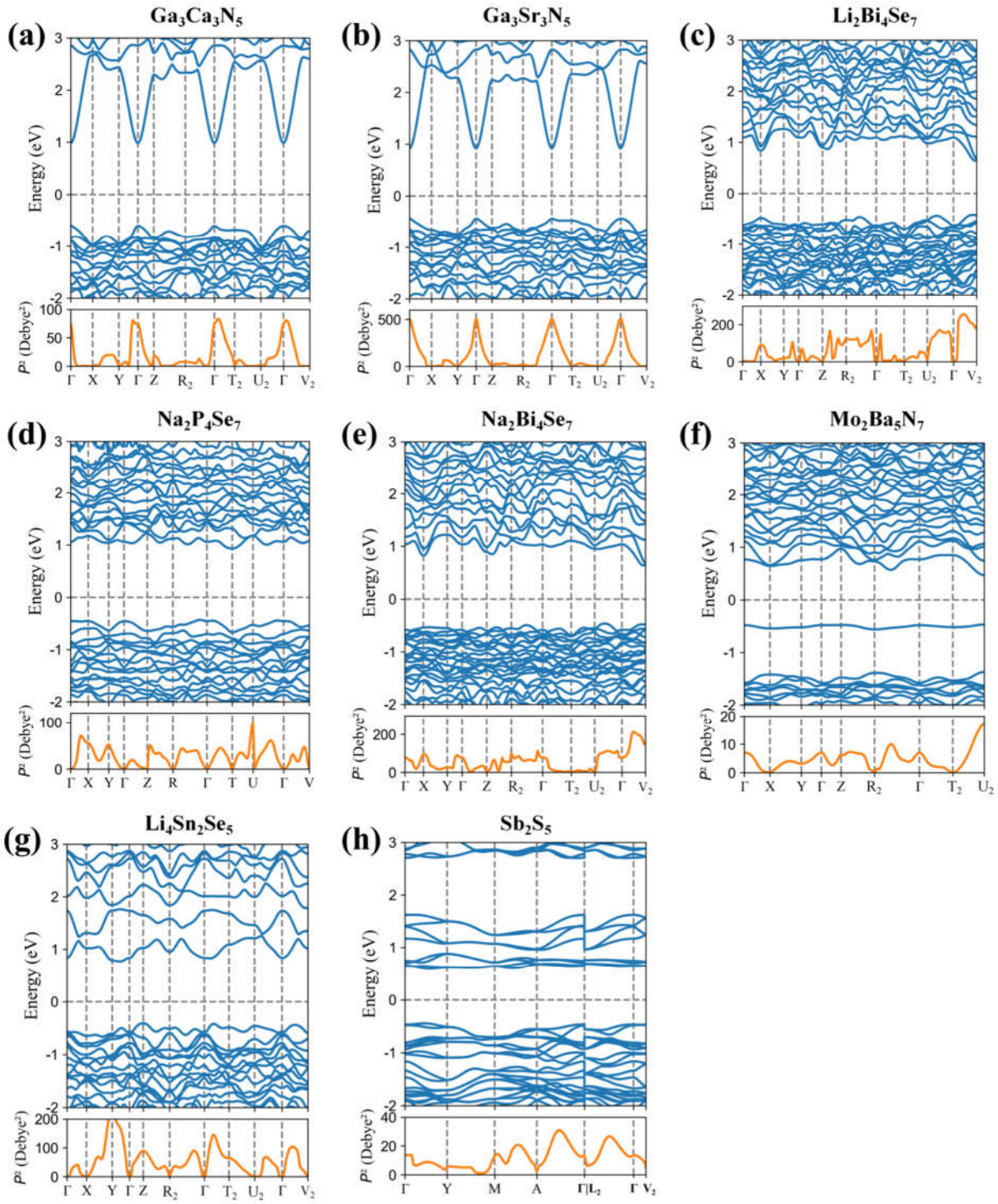

Figure S10. Band structure and P^2 of candidate materials without parity-forbidden transitions at the band edges were calculated using the PBE functional.

.

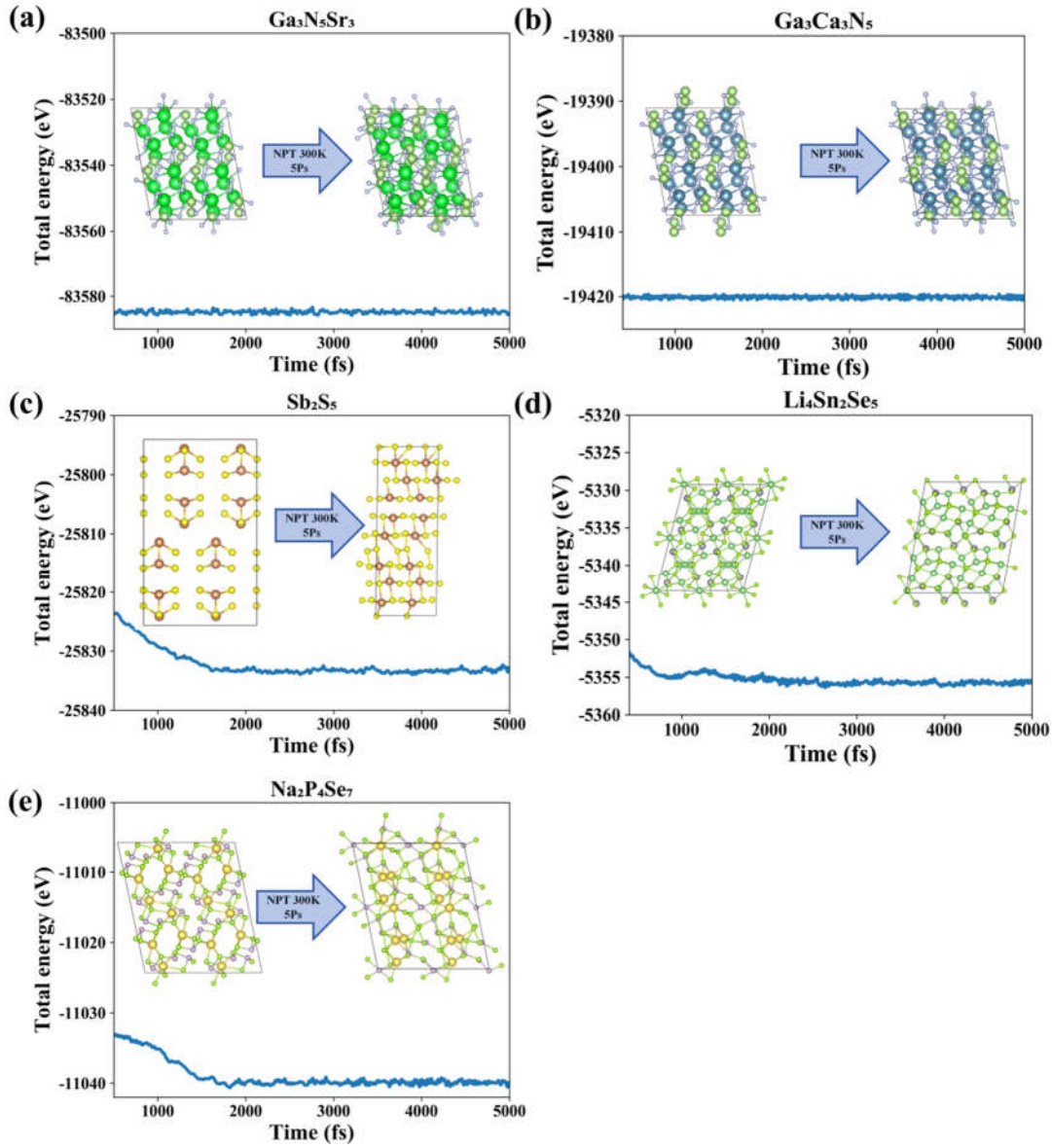

Figure S11. Total energy changes from ab initio molecular dynamics simulations at 300 K for **(a)** $Ga_3N_5Sr_3$, **(b)** $Ga_3Ca_3N_5$, **(c)** Sb_2S_2, **(d)** $Li_4Sn_2Se_5$, and **(e)** $Na_2P_4Se_7$. The inset shows the corresponding tectonic evolution.

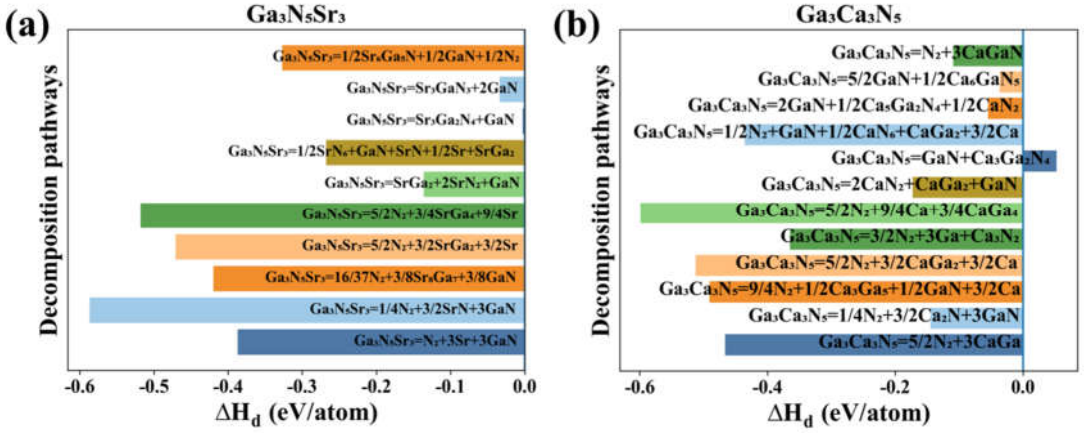

Figure S12. The decomposition energy path of the (a) $Ga_3N_5Sr_3$ and (b) $Ga_3Ca_3N_5$.

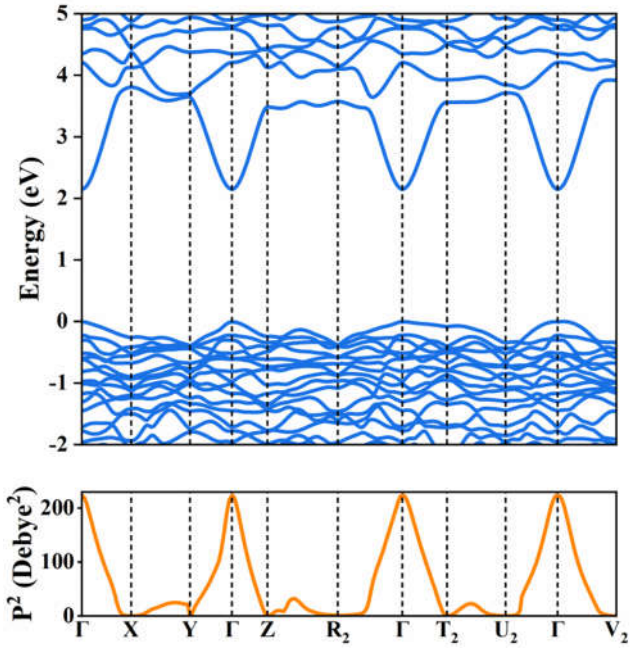

Figure S13. Band structure and P^2 for the $Ga_3N_5Sr_3$ calculated using the HSE06 hybrid functional.

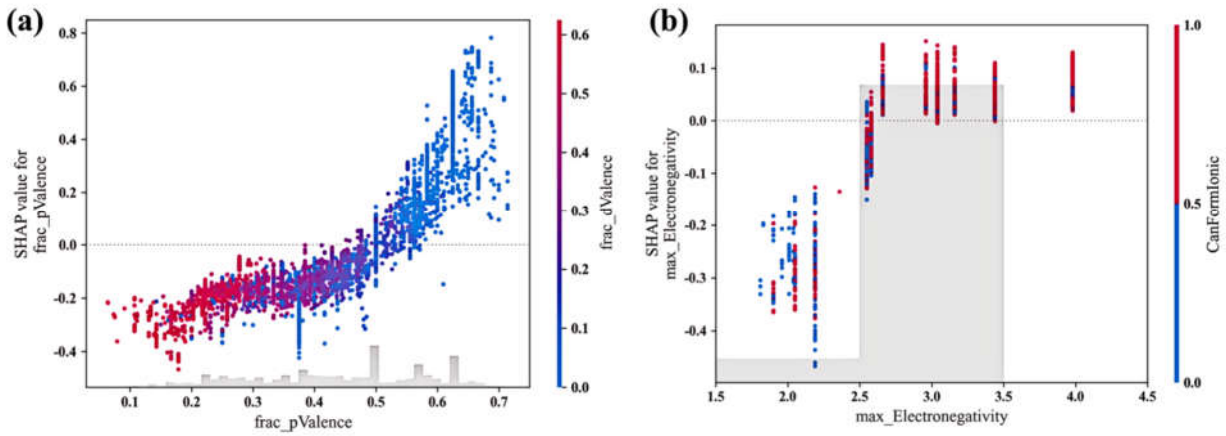

Figure S14. SHAP dependence plots for (a) frac_pValence and (b) max_Electronegativity features for E_g.

Table S1 The lattice parameters for the Materials, i.e., $Li_2Bi_4Se_7$, $Na_2Bi_4Se_7$, $Ga_3N_5Sr_3$ and $Mo_2Ba_5N_7$, Sb_2S_2, Bi_2SeO_5, Bi_2TeO_5, $Na_2P_4Se_7$.

Materials	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)
$Li_2Bi_4Se_7$	9.413	11.261	7.195	90.23	104.28	91.18
$Na_2Bi_4Se_7$	9.581	11.494	7.303	89.88	103.94	92.42
$Ga_3N_5Sr_3$	6.029	7.260	8.783	108.23	103.96	103.96
$Mo_2Ba_5N_7$	6.781	9.664	9.771	65.01	65.01	90.24
Sb_2S_5	7.142	21.379	6.031	90.00	114.61	90.00

$Li_4Sn_2Se_5$	7.233	7.929	10.700	74.91	69.31	86.34
$Ga_3Ca_3N_5$	5.957	6.957	8.523	107.24	106.02	95.16
$Na_2P_4Se_7$	9.485	10.957	7.084	77.27	99.88	103.61

Table S2. The descriptions of 61 features for classification models.

Features	Description
var_EffectiveCoordination	Variance of effective coordination number.
min_EffectiveCoordination	Minimum effective coordination number.
mean_BondLengthVariation	Average bond length variation.
var_BondLengthVariation	Variance of bond length variation.
min_BondLengthVariation	Minimum bond length variation.
max_BondLengthVariation	Maximum bond length variation.
MaxPackingEfficiency	Maximum packing efficiency.
max_NeighDiff_shell1_Number	Maximum number of neighbors in the first coordination shell.
mean_NeighDiff_shell1_MendeleevNumber	Average Mendeleev number of nearest neighbors in the first shell.
max_NeighDiff_shell1_MendeleevNumber	Maximum Mendeleev number of nearest neighbors in the first shell.
max_NeighDiff_shell1_MeltingT	Maximum melting temperature of nearest neighbors in the first shell.
mean_NeighDiff_shell1_Column	Average group number of nearest neighbors in the first shell.
var_NeighDiff_shell1_Column	Variance of group numbers of nearest neighbors in the first shell.
max_NeighDiff_shell1_Column	Maximum group number of nearest neighbors in the first shell.
var_NeighDiff_shell1_CovalentRadius	Variance of covalent radii of nearest neighbors in the first shell.
mean_NeighDiff_shell1_Electronegativity	Average electronegativity of nearest neighbors in the first shell.
var_NeighDiff_shell1_Electronegativity	Variance of electronegativity of nearest neighbors in the first shell.
max_NeighDiff_shell1_Electronegativity	Maximum electronegativity of nearest neighbors in the first shell.
var_NeighDiff_shell1_NsValence	Variance in the number of s-orbital valence electrons of nearest neighbors in the first shell.
min_NeighDiff_shell1_NsValence	Minimum number of s-orbital valence electrons of nearest neighbors in the first shell.
mean_NeighDiff_shell1_NdValence	Average number of d-orbital valence electrons of nearest neighbors in the first shell.

range_NeighDiff_shell1_NsUnfilled	Range of unfilled s-orbital electrons of nearest neighbors in the first shell.
max_NeighDiff_shell1_NpUnfilled	Maximum number of unfilled p-orbital electrons of nearest neighbors in the first shell.
min_NeighDiff_shell1_NdUnfilled	Minimum number of unfilled d-orbital electrons of nearest neighbors in the first shell.
var_NeighDiff_shell1_NUnfilled	Variance in the number of unfilled electrons in the first shell.
min_NeighDiff_shell1_GSvolume_pa	Minimum ground-state volume (in Pascals) of nearest neighbors in the first shell.
min_NeighDiff_shell1_GSbandgap	Minimum ground-state bandgap of nearest neighbors in the first shell.
range_NeighDiff_shell1_GSbandgap	Range of ground-state bandgap of nearest neighbors in the first shell.
mean_NeighDiff_shell1_GSmagmom	Average magnetic moment of nearest neighbors in the first shell.
var_NeighDiff_shell1_GSmagmom	Variance of magnetic moment of nearest neighbors in the first shell.
min_NeighDiff_shell1_GSmagmom	Minimum magnetic moment of nearest neighbors in the first shell.
mean_MendeleevNumber	Average Mendeleev number of elements in the material.
dev_MendeleevNumber	Standard deviation of the Mendeleev number (atomic number).
min_MendeleevNumber	Minimum Mendeleev number (atomic number) of elements in the material.
mean_MeltingT	Average melting temperature of the material.
maxdiff_MeltingT	Maximum difference in melting temperature.
dev_MeltingT	Standard deviation of the melting temperature.
max_MeltingT	Maximum melting temperature.
most_MeltingT	Most common melting temperature.
mean_Column	Average periodic table group number.
maxdiff_Column	Maximum difference in group numbers of the periodic table.
dev_Column	Standard deviation of the group numbers of the periodic table.
mean_CovalentRadius	Average covalent radius.
mean_Electronegativity	Average electronegativity.
most_Electronegativity	Most common electronegativity.
mean_NpValence	Average p-orbital valence electrons.
dev_NpValence	Standard deviation of p-orbital valence electrons.
dev_NdValence	Standard deviation of d-orbital valence electrons.
mean_NValance	Average valence electrons.

maxdiff_NValance	Maximum difference in valence electrons.
mean_NpUnfilled	Average number of unfilled p-orbital electrons.
mean_NdUnfilled	Average number of unfilled d-orbital electrons.
mean_NUnfilled	Average number of unfilled electrons.
dev_NUnfilled	Standard deviation of unfilled electrons.
dev_GSvolume_pa	Standard deviation of the ground-state volume(in Pascals).
min_GSvolume_pa	Minimum ground-state volume (in Pascals).
mean_GSmagmom	Average ground-state magnetic moment.
dev_GSmagmom	Standard deviation of the ground-state magnetic moment.
mean_SpaceGroupNumber	Average space group number.
frac_pValence	Fraction of p-orbital valence electrons.
CanFormIonic	Ability to form ionic bonds.

Table S3. The descriptions of 50 features for regression models.

Features	Description
min_EffectiveCoordination	Minimum effective coordination number.
min_MeanBondLength	Minimum mean bond length.
mean_BondLengthVariation	Mean bond length variation.
var_BondLengthVariation	Variance in bond length variation.
min_BondLengthVariation	Minimum bond length variation.
MaxPackingEfficiency	Maximum packing efficiency.
max_NeighDiff_shell1_Electronegativity	Maximum electronegativity of nearest neighbors in the first shell.
mean_NeighDiff_shell1_NdValence	Average d-orbital valence electron number of nearest neighbors in the first shell.
max_NeighDiff_shell1_NdValence	Maximum d-orbital valence electron number of nearest neighbors in the first shell.
range_NeighDiff_shell1_NpUnfilled	Range of unfilled p-orbital electrons of nearest neighbors in the first shell.
max_NeighDiff_shell1_NdUnfilled	Maximum number of unfilled d-orbital electrons of nearest neighbors in the first shell.
mean_Number	Mean atomic number of elements in the material.
mean_MeltingT	Average melting temperature.
maxdiff_MeltingT	Maximum difference in melting temperature.
max_MeltingT	Maximum melting temperature.
most_MeltingT	Most common melting temperature.
dev_Column	Standard deviation of the periodic table column number.

mean_Row	Mean row number in the periodic table.
most_Row	Most common row number in the periodic table.
mean_CovalentRadius	Mean covalent radius of elements in the material.
min_CovalentRadius	Minimum covalent radius of elements in the material.
most_CovalentRadius	Most common covalent radius in the material.
mean_Electronegativity	Mean electronegativity of elements in the material.
maxdiff_Electronegativity	Maximum difference in electronegativity of elements.
dev_Electronegativity	Standard deviation of electronegativity of elements in the material.
max_Electronegativity	Maximum electronegativity of elements in the material.
most_Electronegativity	Most common electronegativity of elements in the material.
dev_NpValence	Standard deviation of the number of p-orbital valence electrons.
mean_NdValence	Mean d-orbital valence electron number.
maxdiff_NdValence	Maximum difference in the number of d-orbital valence electrons.
dev_NdValence	Standard deviation of the number of d-orbital valence electrons.
max_NdValence	Maximum number of d-orbital valence electrons.
dev_NValance	Standard deviation of the valence electrons of all orbitals.
maxdiff_NpUnfilled	Maximum difference in unfilled p-orbital electrons.
max_NpUnfilled	Maximum number of unfilled p-orbital electrons.
max_NdUnfilled	Maximum number of unfilled d-orbital electrons.
mean_NUnfilled	Mean number of unfilled electrons in all orbitals.
max_NUnfilled	Maximum number of unfilled electrons in all orbitals.
mean_GSbandgap	Mean ground-state bandgap.
dev_GSbandgap	Standard deviation of the ground-state bandgap.
max_GSbandgap	Maximum ground-state bandgap.
maxdiff_GSmagmom	Maximum difference in ground-state magnetic moment.
max_GSmagmom	Maximum ground-state magnetic moment.
most_SpaceGroupNumber	Most common space group number.
frac_sValence	Fraction of s-orbital valence electrons in the material.
frac_pValence	Fraction of p-orbital valence electrons in the material.
frac_dValence	Fraction of d-orbital valence electrons in the material.
CanFormIonic	Ability to form ionic bonds.
MaxIonicChar	Maximum ionic character.

MeanIonicChar	Average ionic character.

References

(1) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Scikit-learn: Machine learning in Python. *J MACH LEARN RES* **2011**, *12*, 2825-2830.

(2) F. a. o. Chollet, Keras. *GitHub repository* **2015**. https://github.com/fchollet/keras

(3) J. H. Friedman, Greedy function approximation: a gradient boosting machine. *ANN STAT* **2001**, 1189-1232.

(4) Leo, B. Random forests. *Mach learn* **2001**, *45*, 5-23.

(5) P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Mach learn* **2006**, *63*, 3-42.

(6) D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **1986**, *323* (6088), 533-536.

(7) Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, 86 (11), 2278-2324.

(8) L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **2017**, *96* (2), 024104.

(9) W. L. Jia, J. Y. Fu, Z. Y. Cao, L. Wang, X. B. Chi, W. G. Gao, L. W. Wange, Fast plane wave density functional theory molecular dynamics calculations on multi-GPU machines. *J Comput Phys* **2013**, *251*, 102-115.

(10) P. E. Blöchl, Projector augmented-wave method. *Phys. Rev. B.* **1994**, *50* (24), 17953.

(11) J. P. Perdew, K. Burke, M. J. Ernzerhof, Generalized gradient approximation made simple. *Phys. Rev. B.* **1996**, *77* (18), 3865-3868.

(12) A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, G. E. Scuseria, Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **2006**, 125 (22), 224106.

(13) P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, et al., QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter* **2009**, *21* (39), 395502.

(14) K. Choudhary, M. Bercx, J. Jiang, R. Pachter, D. Lamoen, F. Tavazza, Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. *Chem. Mater*. **2019**, *31* (15), 5900-5908.

(15) SL3ME. *GitHub repository*. https://github.com/ldwillia/SL3ME.

(16) L. Yu, A. Zunger, Identification of Potential Photovoltaic Absorbers Based on First-Principles Spectroscopic Screening of Materials. *Phys. Rev. Lett.* **2012**, *108* (6), 068701.