Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable machine learning of formation energies and band gaps in hybrid perovskites through SHAP analysis and SISSO modelling

Christopher Xia
Saint Francis High School
H Zhang
Fudan University
Q. Y. Jin
qyjin@phy.ecnu.edu.cn
East China Normal University

Article

Keywords:

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11480170/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Interpretable machine learning of formation energies and band gaps in hybrid perovskites through SHAP analysis and SISSO modelling

Christopher Xia,[1] H. Zhang,[2] and Q. Y. Jin[3,*]

[1] *Saint Francis High School, Mountain View, CA*

[2] *College of Future Information Technology, Fudan University, Shanghai 200433, China*

[3] *School of Physics, East China Normal University, Shanghai 200062, China*

Abstract

Accurate and interpretable property prediction is essential for exploring hybrid organic–inorganic perovskites. Here, we combine machine-learning screening with SHAP analysis and SISSO descriptor discovery. Using 4,456 compounds, optimized XGBoost models achieved test-set R^2 values of 0.99 for formation energy and 0.93 for band gap and were applied to over 180,000 candidates. Comparison across approximately 17,000 common compounds revealed broadly overlapping distributions from XGBoost and CGCNN but substantially lower predictions from UPET, suggesting potential limitations in its transferability to hybrid organic–inorganic perovskites. SHAP analysis and correlation-based pruning retained eight primary features for each property, from which SISSO derived expressions comprising four nonlinear descriptors. These expressions couple molecular and inorganic properties for formation energy and emphasize relative atomic sizes and B-site characteristics for band gap. This framework supports efficient candidate prioritization and interpretable analysis of learned structures–property relationships in hybrid perovskites.

I. INTRODUCTION

As climate change continues to worsen, the demand for renewable energy is on the rise. In particular, solar energy is very abundant, leading to photovoltaics becoming one of the renewable energy sources with the most potential. Currently, the photovoltaic market is dominated by silicon based solar cells. Although they currently present a very respectable power-conversion efficiency (PCE), they also struggle with massive upfront production energy requirements and methods of quality recycling[1]. Among these alternatives, hybrid organic–inorganic perovskites have attracted considerable attention owing to their rapidly improving PCE[2], potentially low-cost fabrication and tunable optoelectronic properties.

The photovoltaic potential of these materials was established through early demonstrations of perovskite sensitizers, solid-state devices and sequential deposition methods[2–5]. Their molecular and inorganic constituents provide a broad compositional space for tailoring electronic properties and structural stability. Subsequent compositional engineering demonstrated that mixed-cation formulations can improve phase stability and device performance[6,7]. However, material degradation and the environmental risks associated with lead-containing compositions, including the potential release of lead upon degradation, remain major barriers to widespread deployment[8]. Addressing these challenges requires identifying alternative compositions that combine suitable band gaps with favourable energetics and reduced toxicity. The breadth of the available chemical space therefore motivates efficient screening strategies that can prioritize promising candidates and reveal the compositional relationships governing their properties.

Double perovskites expand this search space by accommodating distinct B-site cations. Experimental and computational studies of silver–bismuth and silver–indium halide double perovskites demonstrated lead-free substitution routes and substantial variation in their electronic properties[9–12]. These inorganic systems illustrate the opportunities offered by B-site chemistry, while hybrid double perovskites introduce additional molecular degrees of freedom[13]. The resulting diversity creates a large combinatorial space that is difficult to explore exhaustively. Identifying alternative compositions therefore requires balancing electronic properties and energetic favourability.

Formation energy and band gap provide complementary screening criteria. Formation energy measures energetic favourability relative to specified reference states, but does not

alone establish resistance to decomposition into competing phases[14]. Geometric compatibility, including the size relationships underlying tolerance-factor descriptions, also influences perovskite structure formation[15]. Molecular steric effects can modify octahedral tilting and metal–halide–metal bond angles, changing orbital overlap and band gaps[16]. For representative Pb- and Sn-based hybrid halide perovskites, electronic-structure calculations further demonstrate the importance of spin–orbit coupling and relativistic effects[17,18]. These coupled structural and electronic contributions motivate the use of molecular, elemental and geometric descriptors in predictive modelling.

ab-initio calculations are essential for evaluating such properties, but become computationally demanding across large candidate sets. Machine learning (ML) provides efficient surrogate models for prioritizing subsequent calculations. Composition-based frameworks and element-wize descriptors have enabled predictions of multiple materials properties[19,20]. Pilania et al. demonstrated descriptor-based band-gap prediction for double perovskite oxides[21], while Cai et al. combined machine learning and high-throughput calculations to explore hybrid double perovskites[13,22]. However, compositional descriptors may only indirectly capture local coordination and structural distortions. Graph neural networks (GNN) provide complementary representations of atomic environments. CGCNN learns properties from crystal graphs[23], whereas MEGNet and ALIGNN demonstrate further approaches to incorporating structural information, including explicit bond-angle information in ALIGNN[24–26]. More recently, PET-MAD-DOS, accessible through the universal PET (UPET) framework, has enabled electronic density-of-states prediction and band-gap estimation across diverse organic and inorganic systems[27]. Broad training coverage nevertheless does not guarantee uniform accuracy for every material family.

Interpretability provides a complementary basis for evaluating these predictions. SHAP attributes feature contributions to individual predictions, with tree-based implementations also supporting the analysis of global model behaviour[13,28,29]. Descriptor selection is itself central to reliable materials modelling[30]. The Sure Independence Screening and Sparsifying Operator (SISSO) and its C++ implementation, SISSO++, identify sparse analytical relationships by searching nonlinear combinations of primary features[31,32]. The SISSO-derived tolerance factor of Bartel et al. illustrates its ability to construct compact descriptors for perovskite formability[33,34]. Combining SHAP-guided selection with correlation-based pruning can reduce input redundancy and restrict the SISSO search space, linking feature attribution

to explicit empirical expressions.

In this work, we combine predictive modelling, model comparison and descriptor discovery for hybrid perovskites. Regression models are assessed using 4,456 compounds, and optimized XGBoost models are applied to 180,057 candidates[35]. Band-gap predictions from XGBoost, CGCNN and UPET are compared across 17,001 common compounds. SHAP analysis and correlation-based pruning select eight primary features for each property, from which SISSO derives expressions comprising four nonlinear descriptors. These expressions relate molecular electronic properties, relative atomic sizes and B- and X-site characteristics to the learned formation-energy and band-gap trends, supporting efficient candidate prioritization and interpretable analysis.

II. DATA PREPARATION AND METHODOLOGY

The training data set for bandgap and formation energy consists of 4456 double hybrid organic-inorganic perovskite (HOIP) crystal compounds and has the same physical properties presented in the candidate dataset[13], but with the inclusion of the band gap and formation energy as well in each respective data set[13]. The training dataset was first imputed to replace all NaN's with 0's, then went through dimensionality reduction with PCA method[36], then underwent an 80-20 train test split before training the regression models. The prediction dataset contains over 180,000 HOIP crystal candidates[13]. Many physical properties of the individual ions and bonds in each candidate were noted, including ionic radii, the Mendeleev number, dipole polarizability, etc. These features are used primarily used in the training and inference of the regression models, but the chemical composition of the crystals is used through every other step of screening.

The framework of our work consists of three layers of screening, with each successive layer providing higher accuracy at the cost of increased computational expense. The screening process begins with trained machine-learning regression models for predicting formation energies and band gaps, respectively, followed by more accurate recalculations using graph neural network (GNN) models. Finally, the SISSO method is employed to derive empirical formula for the formation energies and band gaps of perovskite materials, providing physical insights into the underlying structure-property relationships, and further enabling more efficient and interpretable materials screening.

Regression models are computationally inexpensive. However, their accuracy is somewhat lacking compared to other methods. As a result, they are very effective for the first step of filtering out the majority of candidates that are either invalid or do not have a good bandgap. The intention was to train two regression models, one for predicting the bandgap and one for predicting the formation energy of the compound. Using the two regression models, we performed an initial screening to find materials that showed promising bandgap and formation energy.

Although the regressions prove to be a useful form of initial screening, they are simple regression models and do not consider the structures of the perovksites as a feature for its predictions. Therefore, we used two deep learning models, i.e. CGCNN[23] and UPET[37], as an intermediate step to recalculate the bandgap with greater accuracy than the regression model but still at a faster rate than first principle calculations. These models also take in the structure of the HOIP crystals to make a more accurate prediction of both formation energy and bandgap. We trained a CGCNN model on 3489 organic-inorganic perovskite materials with their crystal structures and bandgaps, then fed the model with the crystal files of the screened candidates into the neural network[13]. The crystal files for our candidates were made by taking a template perovskite and swapping the elements in the A, B, and X positions, which allows for a fast way to generate these crystal structure files. Simultaneously, we used PET-MAD-DOS[37], to predict bandgaps of generated HOIPs. Although it is not specifically trained on perovskite crystals, it is trained on the Massive Atomic Diversity Dataset, creating a robust model that would be a good reference to compare alongside CGCNN.

For the interpretable descriptor-discovery stage, the SISSO method was employed to establish compact analytical relationships between the input features and target material properties[31,32]. Starting from the primary descriptors, SISSO generates a large feature space by recursively applying a predefined set of mathematical operators, including $+$, $-$, $\times$, $\div$, x^{-1}, x^2, $\sqrt{x}$, and e^x. Sure independence screening is then used to reduce the dimensionality of the generated feature space, followed by the sparsifying operator to identify a small subset of optimal descriptors for property prediction. We set the minimum and maximum values of each feature to be 10^{-3} to 10^5 before it is discarded. Furthermore, we allowed a nonzero intercept, used the L0 sparse regression method, and used RMSE to grade each dimension. Once the model has been returned, we predicted the bandgaps of 180,000 candidates and compared them to the predictions of XGBoost through correlation analysis.

III. RESULTS AND DISCUSSION

A. Machine/Deep-learning models for formation energies and bandgaps

There were multiple types of regression algorithms tested for formation energy: Gradient Boosting Regressor (GBR), Kernel Ridge Regressor (KRR), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), K Neighbors Regressor (KNR), Support Vector Regressor (SVR), Multi-Layer Perceptron Regressor (MLP), Extra Trees Regressor (ETR), and XGBoost (XGB) methods[36].

Initially, as shown in Figure 1(a), the top performing regression models for formation energies of perovskites include XGBoost ($R^2 = 0.99$), ETR ($R^2 = 0.98$), RFR ($R^2 = 0.98$), and GBR ($R^2 = 0.97$) . These four models were all also able to undergo hyperparameter optimization. However, the changes in accuracy were marginal, and both ETR and RFR even saw a slight decrease in fit. The biggest increase in R^2 was seen in GBR, which increased its R^2 by 0.02 to 0.99. After the optimization, the best performing model overall is XGBoost with an R^2 of 0.99.

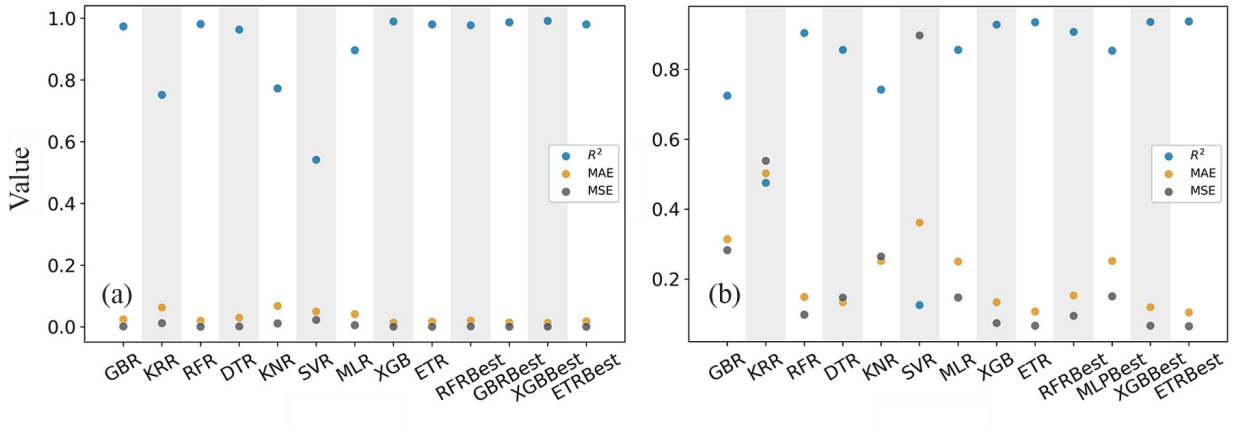


FIG. 1: Performance of various models on predicting (a) formation energies and (b) bandgaps of hybrid perovskites. For RFR, GBR, ETR, and XGBoost, hyperparameter optimization was also conducted.

As for bandgap calculations, as shown in Figure 1(b), we compared the same regression models as for calculating formation energy. The models were not as accurate. Initially, the best performing regression models were RFR ($R^2 = 0.89$), ETR ($R^2 = 0.94$), and XGBoost ($R^2 = 0.91$). After performing hyperparameter optimization on XGB, RFR, MLP, and ETR,

there was some improvement, but not by much. The best regression model was still ETR ($R^2 = 0.94$) with XGBoost as a close second ($R^2 = 0.93$).

As a result, XGBoost was chosen as it was the best formation energy regression model and was very close to being the best bandgap regression model. Then, with one XGBoost regression model trained for bandgaps (E_g) and another for formation energies (E_f), we predicted bandgaps and formation energies for each of the 180,057 crystals[13]. An initial formation-energy filter ($E_f < -0.1$ eV/atom) was applied to identify energetically favourable candidates, followed by screening using three machine-learning and deep-learning models.

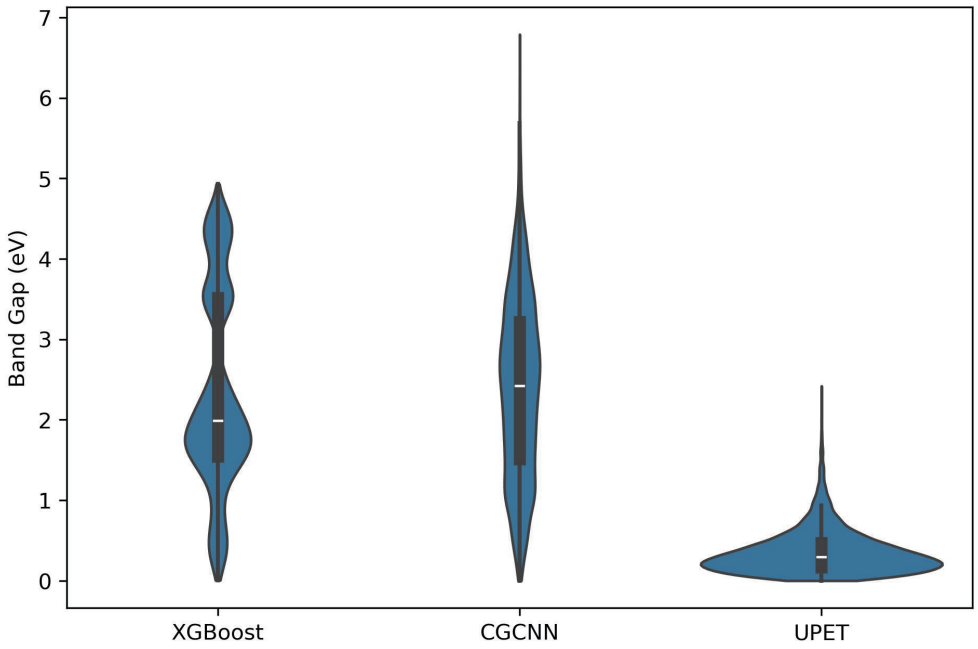


FIG. 2: Comparison of band gaps predicted by XGBoost, CGCNN and UPET for a shared dataset containing over 17,000 perovskites.

Figure 2 compares the distributions predicted by XGBoost, CGCNN and UPET for a shared set of 17,001 perovskite compounds. This set was obtained from 18,407 chemical formulas common to all three prediction datasets by excluding compounds with $E_g \leq 0$ eV predicted by any model. The violin widths represent the estimated probability density, the internal boxes indicate the interquartile ranges, and the white markers denote the medians. XGBoost produces a multimodal distribution with a median near 2.0 eV. The CGCNN distribution substantially overlaps with that of XGBoost but has a higher median of approximately 2.4 eV and a longer upper tail. In contrast, UPET predictions are concentrated at much lower energies, with a median near 0.3 eV.

The broadly similar distributions obtained from XGBoost and CGCNN, both trained on perovskite datasets[13], differ markedly from that predicted by UPET, which was trained on general inorganic crystals[37]. This discrepancy suggests possible limitations in transferring UPET to hybrid organic–inorganic perovskites, where bandgaps depend on the coupling between molecular composition, inorganic coordination and electronic structure. UPET's lower-energy distribution may reflect insufficient representation of these hybrid environments in its training data, although differences in the electronic-structure methods used to generate training labels could also contribute to the discrepancy. These results highlight the need to validate and, where necessary, calibrate UPET against consistent reference calculations before using it for bandgap screening in this chemical space.

B. SHAP analysis and SISSO-based empirical formula for formation energies and bandgaps

To elucidate the feature dependencies underlying formation energy and bandgap, we employed the SISSO method to derive compact empirical expressions for both properties. As described above, the most influential features were first identified through SHAP analysis of the XGBoost models, and highly correlated features were removed to reduce redundancy. SISSO then recursively combined the retained features using mathematical operators to construct a large space of candidate descriptors that explicitly capture nonlinear combinations of the input features. Sure independence screening reduced this space to a tractable subset, from which sparse regression identified low-dimensional descriptor combinations for each target property. The resulting expressions relate formation energy and bandgap to a small number of composite descriptors, providing an interpretable representation of the learned structure-property relationships.

Because the SISSO descriptor space grows combinatorially with the number of primary features[31], we restricted the input set to the most influential features identified by SHAP analysis of the XGBoost models. For formation energy as shown in Figure 3(b), the ten highest-ranked features in descending order of importance were X_electron_affinity, X-averB, B1_Firstionization, Lumo, Polarizability, B2_mendeleev_number, X_Ionicradius, B1_mendeleev_number, B1_proton_affinity and B2_Firstionization. Pairwise correlations among these features were subsequently examined to assess redundancy, as shown in Fig-

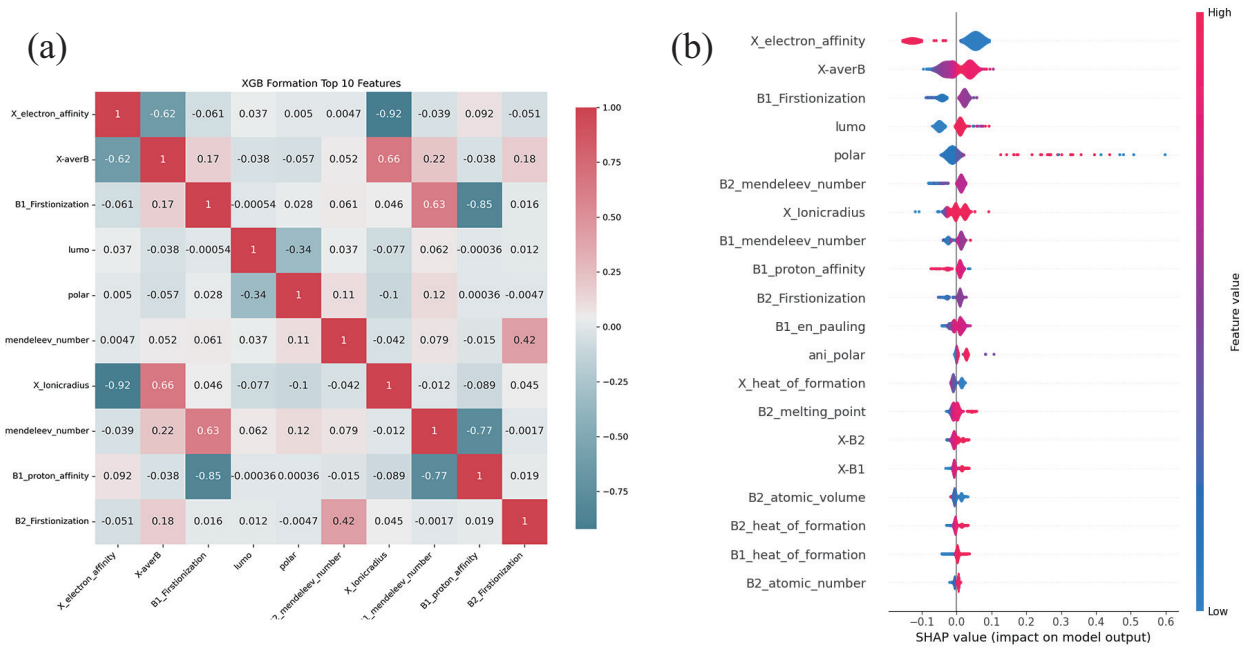


FIG. 3: (a) Correlation analysis and (b) SHAP analysis of the most important features influencing the formation energies of perovskites.

ure 3(a). The electron affinity and ionic radius of the X-site element were strongly correlated; we therefore retained electron affinity, which ranked higher in the SHAP analysis, and excluded ionic radius. Similarly, the strong correlation between B1_Firstionization and B1_proton_affinity led us to retain the former and exclude the latter. The remaining eight most influential features were retained as inputs for the subsequent SISSO analysis.

Similarly, for bandgaps of perovskites as shown in Figure 4(b), the ten highest-ranked features identified by SHAP analysis were Polarizability, X-averB, X/averB, B2_melting_point, B1_Firstionization, Of, B1_atomic_volume, B1_melting_point, B1_mendeleev_number and B2_Firstionization, in descending order of importance. Pairwise correlations among these features were subsequently examined to assess redundancy, as shown in Figure 4(a). The octahedral factor (Of) was strongly correlated with X-averB and X/averB, with correlation coefficients of -0.96 and -0.95, respectively, and was therefore excluded. X-averB and X/averB were also strongly correlated, with a coefficient of 0.95. We retained X-averB, which ranked higher in the SHAP analysis, and excluded X/averB. The remaining eight most influential features were retained as inputs for the subsequent SISSO analysis.

For formation energy, we performed SISSO on a dataset of 4,456 perovskite samples[13], with a maximum descriptor dimension of four. The selected model expresses formation en-

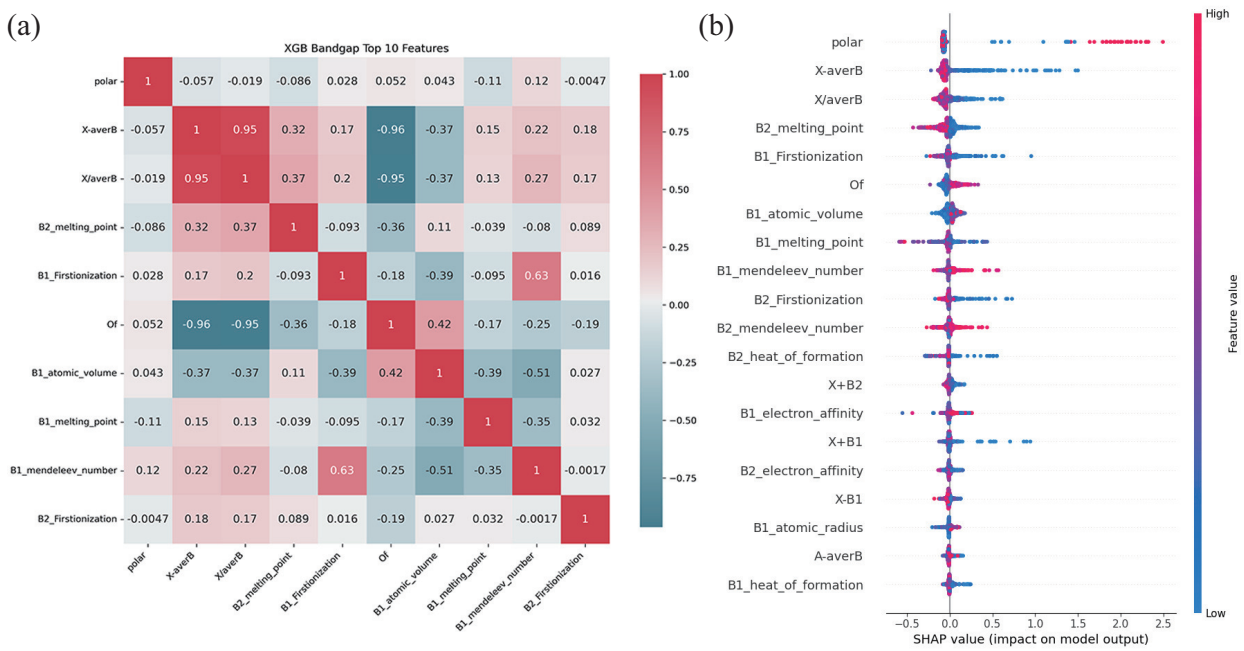


FIG. 4: (a) Correlation analysis and (b) SHAP analysis of the most important features influencing the bandgaps of perovskites.

ergy as a linear combination of four nonlinear descriptors constructed from the eight primary features retained after SHAP ranking and correlation-based pruning. At each descriptor dimension, the sure independence screening (SIS) step retained the 100 highest-ranked candidate features (nf_sis = 100) for subsequent sparse regression. The resulting SISSO-derived empirical expression for formation energy is

$$E_f^{SISSO} \approx c_0 + \sum_{i=1}^{4} c_i d_i, \tag{1}$$

where $c_0 = 1.00$, $c_1 = -7.14$, $c_2 = 15.88$, $c_3 = 1.79$ and $c_4 = -0.01$. The nonlinear descriptors are defined as

$$d_1 = \frac{f_1}{f_2 f_5}, \qquad d_2 = \exp\left(\frac{f_4}{f_3}\right), \qquad d_3 = \frac{f_1 f_3}{f_4}, \qquad d_4 = \frac{f_5}{\sqrt{f_6}}, \tag{2}$$

where f_1, f_2, f_3, f_4, f_5 and f_6 denote the normalized values for the electron affinity of the X-site element, the first ionization energy of the B_1-site element, the LUMO energy of the A-site molecule, the polarizability of the A-site molecule, the Mendeleev number of the B_2-site element and the Mendeleev number of the B_1-site element, respectively.

The SISSO expression in Eq. (1) describes formation energy through coupled combinations of molecular and elemental properties. The first descriptor, d_1, combines the electron affinity of the X-site element with the first ionization energy of the B_1-site element and the Mendeleev number of the B_2-site element. Its negative coefficient, c_1, indicates that increasing this ratio lowers the predicted formation energy when the other descriptors are held fixed, corresponding to greater energetic favourability relative to the reference states. Similarly, the negative coefficient c_4 indicates that increasing d_4, the ratio of the B_2-site Mendeleev number to the square root of the B_1-site Mendeleev number, favours a lower predicted formation energy under the same condition. The second and third descriptors (d_2 and d_3) couple the LUMO energy and polarizability of the A-site molecule, with the third additionally incorporating the X-site electron affinity. These terms link the molecular electronic properties to those of the inorganic constituents, highlighting their combined contributions to the energetic trends captured by the model.

For bandgaps of perovskites, we performed SISSO on a dataset of 923 perovskite samples, with a maximum descriptor dimension of four. O It expresses bandgap as a linear combination of three nonlinear descriptors, together with an intercept, constructed from the eight primary features retained after SHAP ranking and correlation-based pruning. At each descriptor dimension, the SIS step retained the 100 highest-ranked candidate features (nf_sis = 100) for subsequent sparse regression. The resulting SISSO-derived empirical expression is

$$E_{\mathrm{g}}^{SISSO} \approx a_0 + \sum_{i=1}^{4} a_i d_i, \tag{3}$$

where $a_0 = 3.70$, $a_1 = -1.85 \times 10^{-4}$, $a_2 = -0.01$, $a_3 = 3.89 \times 10^{-3}$ and $a_4 = -1.15$. The nonlinear descriptors are defined as

$$d_1 = g_3 g_6, \qquad d_2 = g_2 g_5 g_7, \qquad d_3 = \sqrt{g_3} g_1, \qquad d_4 = \frac{g_1}{g_7}. \tag{4}$$

where g_1 denotes the normalized values of the difference between the atomic radius of the X-site element and the average atomic radius of the B_1- and B_2-site elements, corresponding to X-averB. g_2 denotes the normalized values of the quotient of the X-site element and the average atomic radius of the B_1 and B_2-site elements. g_3 denotes the normalized values of the first ionization energy of the B_1-site element. g_4 represents the normalized values of the

radii of the B_1 site elements, and g_5 denotes the normalized values of the first ionization energy of the B_2 site elements; and g_6 denotes the melting temperature of the B_2 site element in its crystalline form.

Similarly, the SISSO expression in Eq. (3) describes bandgap through coupled combinations of geometric and elemental properties as well. The first descriptor, d_1, combines the first ionization energy of the B_1-site element with its Mendeleev number. Its negative coefficient, a_1, indicates that increasing this product lowers the predicted bandgap when the other descriptors are held fixed, capturing an empirical dependence on B-site elemental identity and electronic character. The second descriptor, d_2, couples the melting temperatures of the B_1- and B_2-site elements in their elemental crystalline forms with g_7. Its negative coefficient, a_2, likewise associates an increase in this product with a lower predicted bandgap under the same condition. The third and fourth descriptors incorporate the difference between the X-site atomic radius and the average B-site atomic radius, weighted by the square root of the B_1-site first ionization energy and the inverse of g_7, respectively. Although a_3 is positive and a_4 is negative, the net dependence on the radius difference is governed by $a_3\sqrt{g_3} + a_4/g_7$, rather than either coefficient alone. Together, these descriptors provide an interpretable representation of the compositional dependencies captured by the model, without directly quantifying orbital interactions or structural distortions.

Out of the 3 models trained with various parameters for formation energy, the RMSE reached a convergence, ranging from 0.1260-0.1360. The selected model achieving an RMSE of 0.1260. As for bandgaps, out of the 7 models trained on various parameters, the RMSE ranges from 0.7455-0.8399. The selected model had an RMSE of 0.8340, as it removed an extra highly correlated feature.

IV. CONCLUSION

In this work, we developed a multistage framework combining XGBoost, CGCNN, UPET and SISSO to screen 180,000 hybrid organic–inorganic perovskites. XGBoost achieved R^2 values of 0.99 for formation energy and 0.93 for bandgap. SHAP analysis identified the dominant molecular, elemental and geometric features, while correlation analysis removed redundant inputs. SISSO then derived compact analytical expressions from the retained features. The formation-energy SISSO-induced formulae captures coupled contributions from

the organic molecule and inorganic framework, whereas the SISSO-induced formulae emphasizes relative atomic sizes and B-site electronic properties, enabling efficient and interpretable materials discovery and screening.

ACKNOWLEDGEMENTS

The authors acknowledge the use of ChatGPT (OpenAI) for assistance in improving the language and readability of the manuscript.

FUNDING

Not applicable.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

DATA AVAILABILITY

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

AUTHOR CONTRIBUTIONS

Q. J. conceived the project and supervised the research. C. X. developed and trained the models and analyzed the results. C. X. wrote the original manuscript. C. X. and H. Z. contributed to the discussion of the results.

---

* qyjin@phy.ecnu.edu.cn

[1] Swapnil Dubey, Nilesh Y Jadhav, and Betka Zakirova. Socio-economic and environmental impacts of silicon based photovoltaic (pv) technologies. *Energy procedia*, 33:322–334, 2013.

[2] Akihiro Kojima, Kenjiro Teshima, Yasuo Shirai, and Tsutomu Miyasaka. Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *Journal of the American Chemical Society*, 131(17):6050–6051, 2009.

[3] Michael M. Lee, Joël Teuscher, Tsutomu Miyasaka, Takurou N. Murakami, and Henry J. Snaith. Efficient Hybrid Solar Cells Based on Meso-Superstructured Organometal Halide Perovskites. *Science*, 338(6107):643–647, 2012.

[4] Julian Burschka, Norman Pellet, Soo-Jin Moon, Robin Humphry-Baker, Peng Gao, Mohammad K. Nazeeruddin, and Michael Grätzel. Sequential deposition as a route to high-performance perovskite-sensitized solar cells. *Nature*, 499(7458):316–319, 2013.

[5] Md Helal Miah, Mayeen Uddin Khandaker, Md Bulu Rahman, Mohammad Nur-E-Alam, and Mohammad Aminul Islam. Band gap tuning of perovskite solar cells for enhancing the efficiency and stability: issues and prospects. *RSC advances*, 14(23):15876–15906, 2024.

[6] Nam Joong Jeon, Jun Hong Noh, Woon Seok Yang, Young Chan Kim, Seungchan Ryu, Jangwon Seo, and Sang Il Seok. Compositional engineering of perovskite materials for high-performance solar cells. *Nature*, 517(7535):476–480, 2015.

[7] Michael Saliba, Taisuke Matsui, Ji-Youn Seo, Konrad Domanski, Juan-Pablo Correa-Baena, Mohammad Khaja Nazeeruddin, Shaik M. Zakeeruddin, Wolfgang Tress, Antonio Abate, Anders Hagfeldt, and Michael Grätzel. Cesium-containing triple cation perovskite solar cells: improved stability, reproducibility and high efficiency. *Energy & Environmental Science*, 9(6):1989–1997, 2016.

[8] Junming Li, Hai-Lei Cao, Wen-Bin Jiao, Qiong Wang, Mingdeng Wei, Irene Cantone, Jian Lü, and Antonio Abate. Biological impact of lead from halide perovskites reveals the risk of introducing a safe threshold. *Nature communications*, 11(1):310, 2020.

[9] Adam H. Slavney, Te Hu, Aaron M. Lindenberg, and Hemamala I. Karunadasa. A Bismuth-Halide Double Perovskite with Long Carrier Recombination Lifetime for Photovoltaic Applications. *Journal of the American Chemical Society*, 138(7):2138–2141, 2016.

[10] Eric T. McClure, Molly R. Ball, Wolfgang Windl, and Patrick M. Woodward. Cs2AgBiX6 (X = Br, Cl): New Visible Light Absorbing, Lead-Free Halide Perovskite Semiconductors. *Chemistry of Materials*, 28(5):1348–1354, 2016.

[11] George Volonakis, Marina R. Filip, Amir Abbas Haghighirad, Nobuya Sakai, Bernard Wenger, Henry J. Snaith, and Feliciano Giustino. Lead-Free Halide Double Perovskites via Heterovalent

Substitution of Noble Metals. *The Journal of Physical Chemistry Letters*, 7(7):1254–1259, 2016.

[12] George Volonakis, Amir Abbas Haghighirad, Rebecca L. Milot, Weng H. Sio, Marina R. Filip, Bernard Wenger, Michael B. Johnston, Laura M. Herz, Henry J. Snaith, and Feliciano Giustino. Cs2InAgCl6: A New Lead-Free Halide Double Perovskite with Direct Band Gap. *The Journal of Physical Chemistry Letters*, 8(4):772–778, 2017.

[13] Xia Cai, Yiming Zhang, Zejiao Shi, Ying Chen, Yujie Xia, Anran Yu, Yuanfeng Xu, Fengxian Xie, Hezhu Shao, Heyuan Zhu, Desheng Fu, Yiqiang Zhan, and Hao Zhang. Discovery of Lead-Free Perovskites for High-Performance Solar Cells via Machine Learning: Ultrabroadband Absorption, Low Radiative Combination, and Enhanced Thermal Conductivities. *Advanced Science*, 9(4):2103648, 2022.

[14] G. P. Nagabhushana, Radha Shivaramaiah, and Alexandra Navrotsky. Direct calorimetric verification of thermodynamic instability of lead halide hybrid perovskites. *Proceedings of the National Academy of Sciences*, 113(28):7717–7721, 2016.

[15] W. Travis, E. N. K. Glover, H. Bronstein, D. O. Scanlon, and R. G. Palgrave. On the application of the tolerance factor to inorganic and hybrid halide perovskites: a revised system. *Chemical Science*, 7(7):4548–4556, 2016.

[16] Marina R. Filip, Giles E. Eperon, Henry J. Snaith, and Feliciano Giustino. Steric engineering of metal-halide perovskites with tunable optical band gaps. *Nature Communications*, 5(1):5757, 2014.

[17] Jacky Even, Laurent Pedesseau, Jean-Marc Jancu, and Claudine Katan. Importance of Spin–Orbit Coupling in Hybrid Organic/Inorganic Perovskites for Photovoltaic Applications. *The Journal of Physical Chemistry Letters*, 4(17):2999–3005, 2013.

[18] Paolo Umari, Edoardo Mosconi, and Filippo De Angelis. Relativistic GW calculations on CH3NH3PbI3 and CH3NH3SnI3 Perovskites for Solar Cell Applications. *Scientific Reports*, 4(1):4467, 2014.

[19] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):16028, 2016.

[20] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications*, 8(1):15679, 2017.

[21] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, and T. Lookman. Machine learning bandgaps of double perovskites. *Scientific Reports*, 6(1):19375, 2016.

[22] Xia Cai, Yan Li, Jianfei Liu, Hao Zhang, Jianguo Pan, and Yiqiang Zhan. Discovery of all-inorganic lead-free perovskites with high photovoltaic performance via ensemble machine learning. *Materials Horizons*, 10(11):5288–5297, 2023.

[23] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14):145301, 2018.

[24] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9):3564–3572, 2019.

[25] Kamal Choudhary and Brian DeCost. Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

[26] Jaehyung Lee, Charles Rhys Campbell, Akshaya Ajith, Sergei V Kalinin, Christopher Wolverton, and Kamal Choudhary. Alignn 2.0: A unified line-graph neural network framework for materials screening, force fields, inverse design, spectroscopy, and microscopy. *arXiv preprint arXiv:2609.19487*, 2026.

[27] Wei Bin How, Pol Febrer, Sanggyu Chong, Arslan Mazitov, Filippo Bigi, Matthias Kellner, Sergey Pozdnyakov, and Michele Ceriotti. A universal machine learning model for the electronic density of states. *Digital Discovery*, 5(4):1635–1649, 2026.

[28] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[29] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):56–67, 2020.

[30] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big Data of Materials Science: Critical Role of the Descriptor. *Physical Review Letters*, 114(10):105503, 2015.

[31] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, 2018.

[32] Thomas Purcell, Matthias Scheffler, Christian Carbogno, and Luca M Ghiringhelli. Sisso++: A c++ implementation of the sure-independence screening and sparisifying operator approach. *The Journal of Open Source Software*, 7(71), 2022.

[33] Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave, Luca M. Ghiringhelli, and Matthias Scheffler. New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances*, 5(2):eaav0693, 2019.

[34] Xiaoyu Yang, Wenyue Zhao, Xin He, Weidong Fei, Yuanhui Sun, Yu Zhao, and Lijun Zhang. Octahedral-motif-guided design of optoelectronic semiconductors via interpretable machine learning. *Nature Communications*, 2026.

[35] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[37] Wei Bin How, Pol Febrer, Sanggyu Chong, Arslan Mazitov, Filippo Bigi, Matthias Kellner, Sergey Pozdnyakov, and Michele Ceriotti. A universal machine learning model for the electronic density of states, 2025.