ACS NANO

www.acsnano.org

ARTICLE

Interpretable Surrogate Learning for Electronic Material Generation

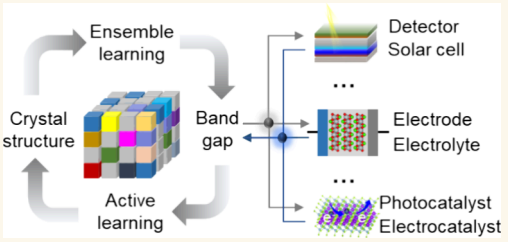
Zhilong Wang,[†] Sixian Liu,[†] Kehao Tao, An Chen, Hongxiao Duan, Yanqiang Han, Fengqi You,* Gang Liu,* and Jinjin Li*

Cite This: https://doi.org/10.1021/acsnano.4c12166

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: **Despite many accessible AI models that have been developed, it is an open challenge to fully exploit interpretable insights to enable effective materials design and develop materials with desired properties for target applications. Here, we introduce an interpretable surrogate learning framework that can actively design and generate electronic materials (EMGen), akin to producing updated materials with requirements by screening all possible elements and fractions. Taking the materials system with required band gaps as a case study, EMGen exhibits a benchmarking predictive error and a running time of 1.7 min for designing and producing a structure with a desired band gap. Using EMGen, we establish a large hybrid functional band gap database, and more uplifting is that the proposed EMGen effectively designs Ga_xO_y with a wide band gap (>5.0 eV) for deep ultraviolet (DUV) optoelectronic devices, enabling a breakthrough extension of the applicability of Ga_xO_y films in photodetectors to DUV light below 240 nm. The augmented band gap also helps improve the breakdown voltage and the heat resilience performance of the amorphous Ga_xO_y film, thereby achieving considerable potential within the realm of power electronics applications. The proposed EMGen, as a specialized, interpretable AI model for the generation of electronic materials, is demonstrated to be an essential tool for on-demand semiconductor materials design.**



KEYWORDS: *machine learning, electronic materials, band gap, active learning, ensemble learning, first-principles calculations*

INTRODUCTION

With the continuous advancement of high-throughput experiments and calculations, materials data have been rapidly growing, and the discovery of materials knowledge from big data has become a key means of research. At this stage, materials informatics has been developed rapidly.[1−3] With the recent surge of artificial intelligence (AI), the fusion of AI with materials science has significantly promoted and advanced the development and innovation of materials informatics.[2,4−6] Despite the accuracy of AI methods, there is often a trade-off between model accuracy and model interpretability. The most accurate and flexible AI models such as deep neural networks (DNNs) are often referred to as "black boxes". This lack of interpretability limits the usability of AI models for general scientific tasks, such as understanding hidden causal relationships and generating scientific hypotheses.

To address the challenge of model interpretability, many interdisciplinary efforts have been devoted to the development of interpretable AI methods, to obtain the prediction of materials properties and mine the physical factors closely related to materials properties.[7] For example, tree-based machine learning (ML) models (extreme boosting decision tree (XGBoost), gradient boosting decision tree (GBDT), and random forest (RF)) can rank the importance of input features and thus have interpretability, which has been applied in the band gap (E_g) prediction of spinel,[8] garnet,[9] double perovskite,[10] and other systems.[11,12] For the crystal graph convolutional neural network (CGCNN) model,[13] Gao et al. have developed explainable methods for node feature vectors and discovered chemical factors closely related to the effective mass of semiconductors.[14] These studies show that the key factors discovered by AI can be further used to assist computational and experimental scientists in materials design. Nevertheless, effective methods can typically be designed to accommodate not more than two key factors due to the complexity of manually assessing the nonlinear relationships between multiple factors and materials properties. This has led to a lot of existing works that only identify the key chemical factors based on interpretable or explainable methods, and further materials

Received: August 31, 2024
Revised: October 19, 2024
Accepted: October 24, 2024

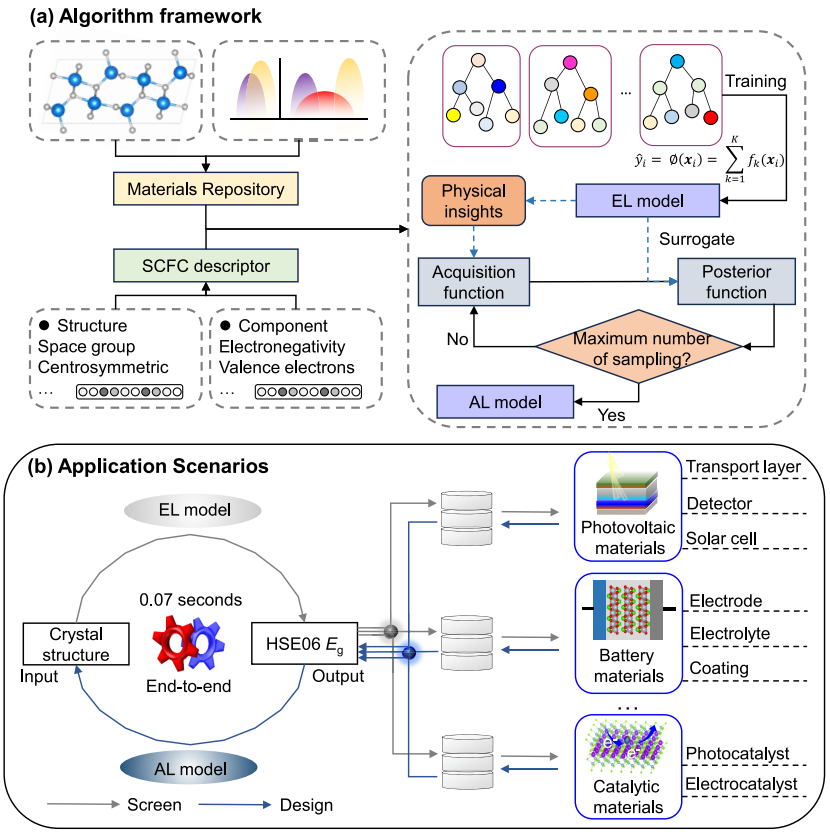


Figure 1. Schematic workflow of EMGen. (a) Algorithm framework, including the materials repository, structural and component fusion coding (SCFC) descriptor, ensemble learning (EL) model, and active learning (AL) model. Crystal structures and the corresponding HSE06 E_g were compiled as the data set. The SCFC descriptor integrates structural and component information. The EL model, which is established by using ensemble learning with physical interpretability, serves as a surrogate model for the AL model. (b) EMGen process for directed design of materials. The proposed EL and AL form a bidirectional intelligence platform for the discovery and design of electronic materials with suitable E_g.

design processes guided by these key chemical factors are missing.

Determining how to fully leverage model interpretability to efficiently guide materials design for achieving specific properties remains an open challenge. It has to be said that the most widely studied materials property, E_g, the energy difference between the conduction band minimum (CBM) and valence band maximum (VBM), reflects the electronic conductivity of the materials to some extent. As one of the most critical properties of electronic materials, E_g largely determines the properties of materials and the performance of downstream devices. The suitable E_g plays a key role in electrode materials (<1.0 eV),[15,16] photovoltaic conversion layers (0.9−1.6 eV),[17,18] solid-state electrolytes (>3.0 eV),[9,19] and other systems. In recent years, materials engineering has begun to adjust E_g at the energy level through recombination, doping, alloying, and other technologies to achieve meaningful applications.[20−23] For example, the E_g of g-C_3N_4 is up to 2.7 eV, which can be partially reduced (∼1.9 eV) by proper defect doping.[24] The E_g of transition metal chalcogenides represented by MoS_2 and WS_2 can be changed by elemental doping, but the range of regulation is limited by their structures (<2.1 eV).[25] De Bastiani et al. experimentally fabricated five perovskites with different E_g (1.59, 1.62, 1.65, 1.68, and 1.7 eV) by varying the ratios of different halogen compounds to gradually find the candidates with E_g close to 1.7 eV.[26] Tao et al. used an F-type pseudohalogen to regulate perovskite films, which effectively inhibited carrier recombination, reduced charge transfer loss, and inhibited phase separation. They obtained an inverted 1.67 eV perovskite and achieved a power conversion efficiency (PCE) of >20%.[27] E_g regulation represents an important future research direction in nanomaterials, which determines the practical applications in areas such as batteries, catalysts, supercapacitors, etc.

ML frameworks have been demonstrated to significantly enhance traditional trial-and-error experiments and high-throughput density functional theory (DFT) calculations, effectively reducing the discovery cycle for materials with appropriate E_g.[6,8,28−31] Zhuo et al. applied a support vector machine (SVM) to predict the E_g of inorganic solids by using the experimental data, enabling researchers to efficiently screen the inorganic phase space.[32] For nonmagnetic 2D semiconductors, Knøsgaard and Thygesen used 286 DFT data points to predict the GW E_g, where G represents Green's function and W denotes the screened Coulomb approximation, using a gradient boosting algorithm for the acquisition of high-precision GW E_g estimates through relatively cost-effective DFT calculations.[33] We previously developed a CGCNN-TL framework that employs transfer learning (TL) to predict the E_g from Perdew−Burke−Ernzerhof (PBE, low-precision) to hybrid functional (HSE06, high-precision) models. This framework demonstrates that the accuracy of E_g predictions can be enhanced from a low-precision to high-precision level at a reduced cost.[34] ML models have been trained to screen

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

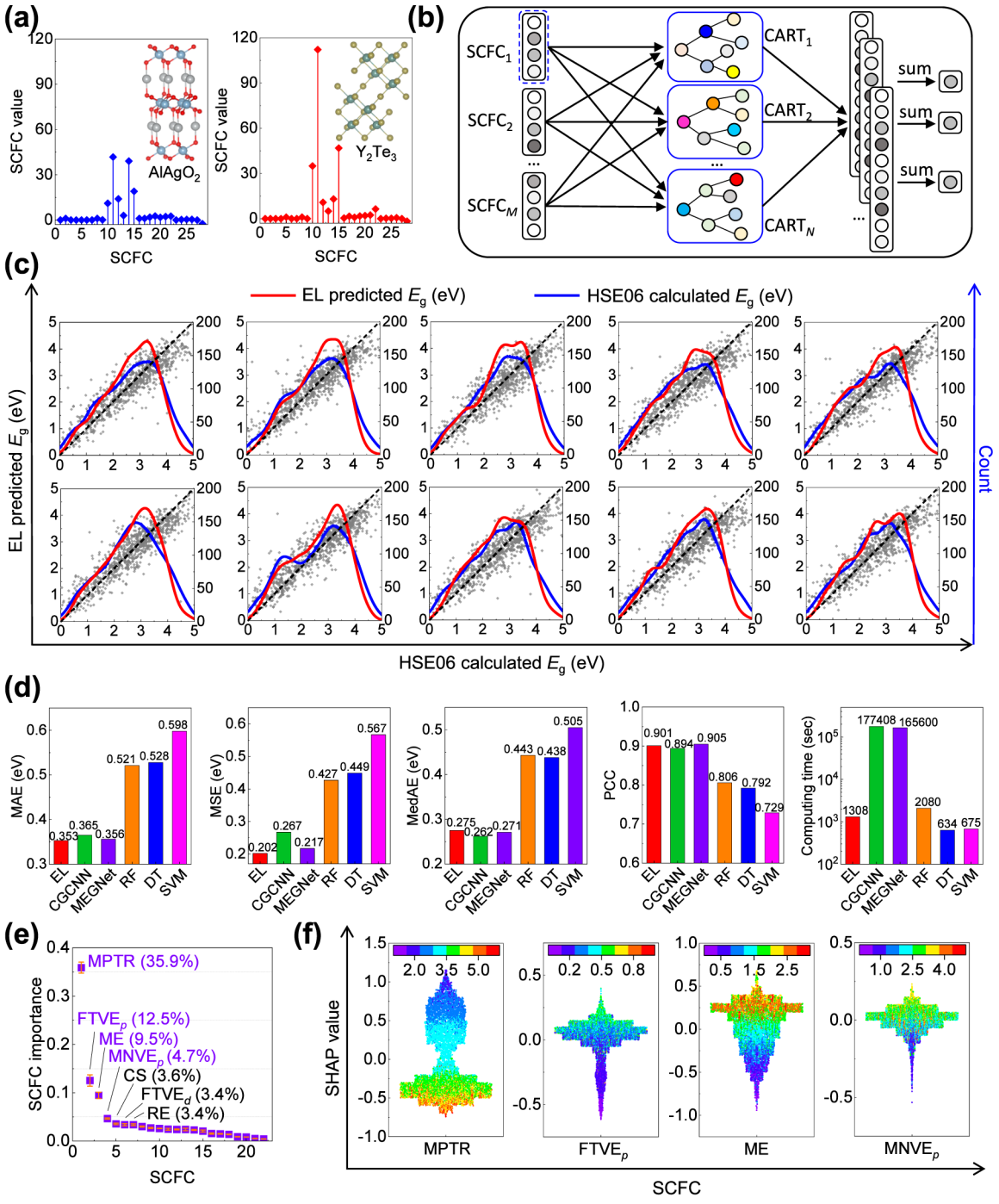


Figure 2. Insights from the EL model. (a) SCFC descriptors of $AlAgO_2$ and Y_2Te_3. (b) Architecture of the EL model. (c) Comparisons of the EL predicted E_g and the calculated HSE06 E_g based on 10-fold cross-validations. (d) Performance comparisons (model accuracy and computing cost) of EL, CGCNN, MEGNet, RF, DT, and SVM. (e) Ranking of feature importance revealed by the EL model. (f) SHAP analyses of the most four important features; the color bars denote the different feature values.

materials with suitable E_g. If the E_g of material is deemed unsuitable for a specific application, then it is either engineered through methods such as doping or modification or discarded. Additionally, the exploration of target materials on a large scale is a decentralized effort in a vast chemical space.

Here, we proposed an interpretable surrogate learning framework, EMGen, which actively designs and generates electronic materials by reconfiguring their components to ensure that their functions meet specifically designed requirements. We took E_g as an example to carry out directed design of electronic materials so that the designed E_g can meet different applications in photovoltaic, electrode, and insulating materials. EMGen was built by a developed structural and component fusion coding (SCFC) descriptor and an interpretable surrogate learning framework. Here, an ensemble learning (EL) model for predicting the E_g was established with a mean squared error of 0.202 eV and a predictive time of 0.7 s. Then, we developed a high-precision HSE06 E_g database containing 116,742 entries (http://aimslab.cn/tools#/query_hse06_band_gap), making it possible to discover electronic

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

materials for specific applications at a large scale. Furthermore, an active learning (AL) model based on the surrogate optimization method was integrated into the EMGen and validated through first-principles (FP) calculations. The proposed EMGen framework can design materials with suitable E_g in a limited cycle of approximately 149 prediction iterations (~1.7 min), facilitating on-demand materials design. Through experimental feedback and verification, we successfully designed wide-E_g (>5.0 eV) Ga_xO_y for deep ultraviolet (DUV) optoelectronic devices, which is poised to extend the use of Ga_xO_y films within photodetectors to DUV light below 240 nm. The proposed EMGen framework significantly shortens the cycle of trial-and-error experiments and computations, bridging the research gap in the targeted design of high-performance electronic materials and providing the insights for on-demand electronic materials design. In addition to electrical parameters, EMGen's capabilities can be easily extended to include the design of mechanical and magnetic parameters, with potential applications in various materials design fields, such as enhancing catalytic activity and improving thermal conductivity.

RESULTS AND DISCUSSION

Proposed EMGen Framework. The schematic workflow is depicted in Figure 1. It encompasses several stages: data compilation, structural and component fusion coding (SCFC) descriptor design, the establishment of an EL model, physical interpretation, AL model establishment, and subsequent applications. We first compiled a data set that contains inorganic crystal structures with the corresponding HSE06 E_g, as shown in Figure 1a (see Note S1 and Figure S1). Integrating the structural (space group and centrosymmetric state) and component information (atomic weight and electronegativity), a descriptor (SCFC) was proposed for representing these structures, which has the advantages of easy availability, low dimension, and physical interpretability. Then, taking SCFC descriptors as inputs and E_g as outputs, an EL model was trained by using a tree-based ML model, surpassing the mainstream methods, such as the GBDT, RF, SVM, CGCNN, MEGNet, etc. (see Figure 2). Using the well-trained EL model, we constructed a large database that contains the inorganic structures and their HSE06 E_g, which can be easily accessed. Importantly, the quantitative and qualitative physical insights can be evaluated from the EL model, which guides further directed materials design. Importantly, an AL model is integrated into EMGen, which is used to directed design E_g for a given material and be successfully applied in photovoltaic, battery, and catalytic materials. AL optimizes the materials with suitable chemical components according to the target HSE06 E_g by using an efficient surrogate optimization model.

In Figure 1b, EMGen constructed a high-precision HSE06 E_g database containing 116,742 entries and selected 60 structures with poor E_g performance for component modification to be used in photovoltaic materials (E_g > 1.6 eV or E_g < 0.9 eV), electrode materials (E_g < 1.0 eV), and insulating materials (E_g > 3.0 eV). The results demonstrate that EMGen not only quickly discovers suitable materials but also designs materials with desired functionalities by component variation. This capacity has significant potential for the future development of photo/electronic materials.

Ensemble Learning Model. Designing materials descriptors to establish a reliable descriptor−property mapping relationship is a central focus and challenge in materials informatics.[7,35,36] Descriptors are crucial for the efficiency and precision of ML models. In general, the dimensionality of descriptors should be balanced, not too high, to avoid overfitting and training difficulties, nor too low, to prevent an incomplete description of the material. To enhance model interpretability, descriptors should incorporate component and structural information, reflecting the physical essence as much as possible. Missing important components or structural information in the descriptor decreases prediction accuracy. Here, we proposed a descriptor for representing inorganic solids, termed SCFC. The SCFC descriptor concatenates the structural feature vector ($\boldsymbol{v}_i$) and component feature vector ($\boldsymbol{l}_i$), including the mean number of atoms per volume (MNAV), mean volume per atom (MVA), mean atomic weight (MAW), mean periodic table column (MPTC), mean periodic table row (MPTR), range and mean of the atomic number (RAN and MAN), range and mean of the atomic radius (RAN and MAN), range and mean of electronegativity (RE and ME), mean number and fraction of valence electrons in the orbital ($MNVE_i$ and $FTVE_i$, i = s, p, d, and f), and band center (BC). More details are presented in Table S1, Note S2, and Figure S2.

Using the proposed SCFC as descriptors, we constructed a high-level ML model using >8,000 crystal structures along with their corresponding HSE06 E_g (0−5 eV) derived from the high-throughput FP calculations.[37] Although experiment-based band gaps have the highest fidelity, large-scale experimental data containing structural information are scarce. For the two compounds $AlAgO_2$ and Y_2Te_3, we visualized the SCFC descriptors in Figure 2a, where significant numerical variations are observed, with a range of approximately 45 for $AlAgO_2$ and 110 for Y_2Te_3, respectively. Given that the SCFC features are all physically significant, it is crucial to select a suitable model that is not affected by the scale. Using this fusion coding, we applied an EL model (Figure 2b) that integrates many classification and regression trees (CARTs) and adopts an additive strategy to form the final prediction model.[38] The EL model allocates each training SCFC descriptor to many different CARTs, utilizing features as tree nodes to facilitate sample division based on the entropy value. The final predicted values were obtained by aggregating the weights of all the leaf nodes in the CARTs. The EL model is trained by minimizing the errors between the predicted and actual values as the optimization objective. Meanwhile, the feature importance can be evaluated according to the frequency of different features used as the basis for classification in nonleaf nodes. Due to the large data scale of more than 8,000 data points, the number of training times and calculation cost of the leave-one-out method will also increase. Therefore, we adopted 10-fold cross-validations to evaluate the performance of the EL model, as shown in Figure 2c. The performance of the EL model is stable in the 10 training sessions, and the gray scatter points are evenly concentrated around the black diagonal. The coincidence degree of the predicted E_g and the calculated E_g is very high, which statistically indicates that the predicted values are in good agreement with the calculated values. Detailedly, four criteria were selected to compute the deviations between EL predicted and calculated E_g. The mean absolute error (MAE), mean squared error (MSE), median absolute error (MedAE), and Pearson correlation coefficient (PCC) of 10-fold cross-validations are 0.353 eV, 0.202 eV^2, 0.275 eV, and 0.901, respectively (see Table S2). The proposed model with such a low prediction error leads the

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

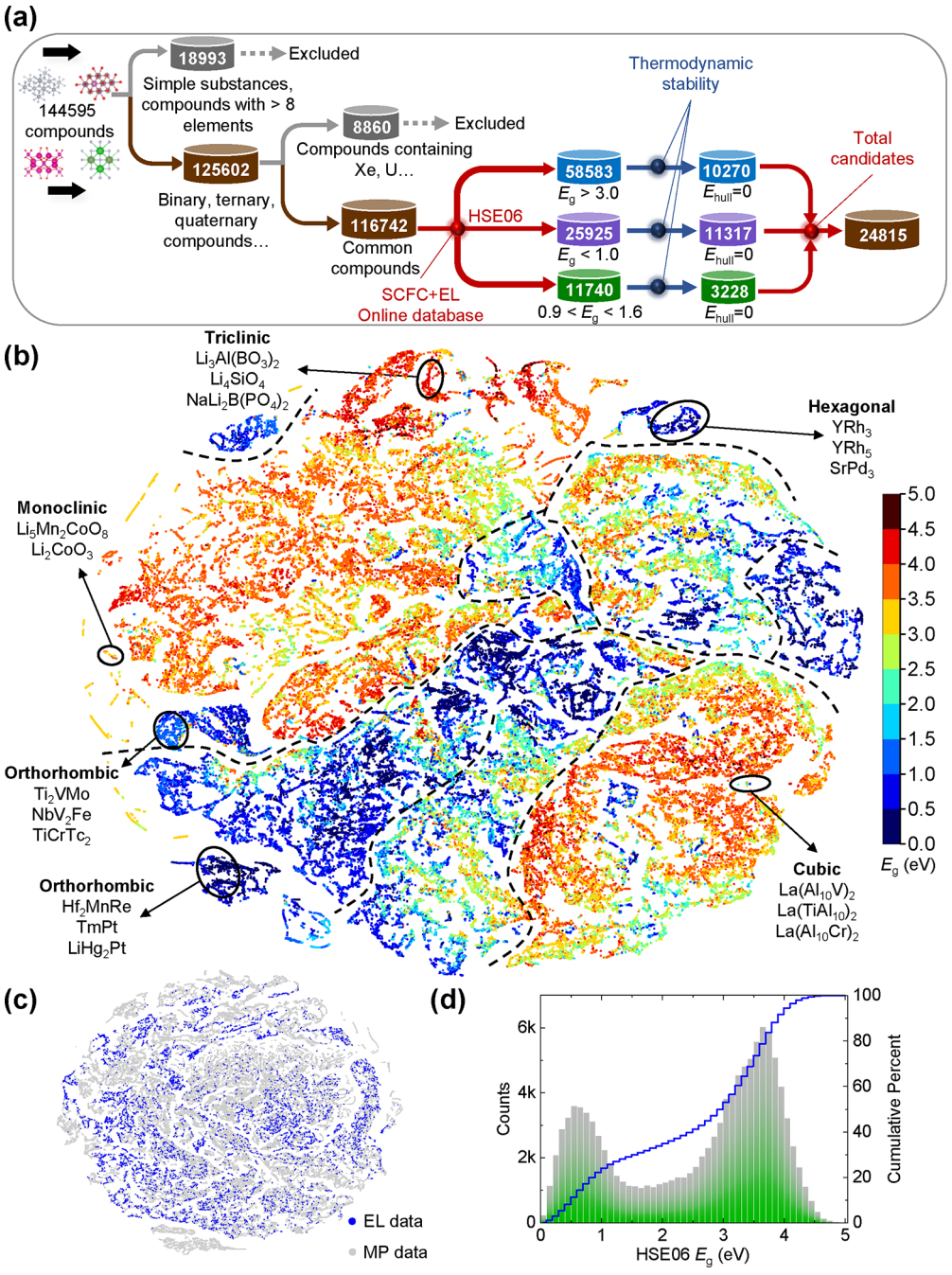


Figure 3. Large-scale materials screening and discovery based on the EL model. (a) Materials screening from the MP database by SCFC descriptors and the EL model. (b) 2D t-SNE plot of the SCFC descriptors; the color bar denotes the predicted HSE06 E_g values. (c) 2D t-SNE plots of the EL data (training/known data, blue dots) and MP data (prediction/unknown data, gray dots). (d) Histogram of predicted HSE06 E_g of 116,742 compounds.

way among models that can predict any compound (not a particular system). In the goal of predicting the E_g of any inorganic materials, the performance of the proposed EL model is better than the previous GBDT model based on a materials fragment descriptor, which has an MSE of 0.26 eV^2,[39] and an SVM model based on the only component descriptor, which has an MSE of 0.203 eV^2.[32] This suggests that missing important structural or component information in the descriptor results in low accuracy of the prediction model, and the error in the E_g prediction of all inorganic materials by the EL model can be used as a benchmark. We applied RF, decision tree (DT), and SVM to perform the E_g prediction, as shown in Figure 2d and Table S3. Although these models maintain some predictive ability based on SCFC descriptors, their predictive errors are much larger than those of the EL model. Moreover, we compared the EL model with the popular and universal crystal graph-based neural networks (CGCNN[13] and MEGNet,[40] see Figures S3 and S4 and Note S4). Overall, there is no significant difference in model accuracy between EL, CGCNN, and MEGNet. However, the training cost of EL is only 1308 s, far less than that of CGCNN (177,408 s) and MEGNet (165,600 s). It is noted that the training time is highly dependent on the computer used, so all the comparisons were made on the same computational resource. The above

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

results show that the EL model and SCFC descriptor proposed are superior to the existing models, including GBDT, RF, DT, SVM, CGCNN, and MEGNet, when considering prediction accuracy and time. Despite the increasing computational power of supercomputers, this combination of high-performance and cost-effective modeling still retains significant advantages and continues to be crucial in model optimization.

Physical Insights Based on the EL Model. The importance of features in SCFC was evaluated according to the frequency of different features used as the basis for classification in nonleaf nodes, as shown in Figure 2e. As calculated in Figure S2e and Table S1, the strong linear relationship between MPTR and the E_g (PCC = −0.59) was also captured by the EL model, with an importance of 35.9%. FTVE$_p$, ME, and MNVE$_p$ are also very important, each accounting for more than 4% importance, indicating their strong correlation with the E_g of materials. These show that the periodic table rows, the number of valence electrons in the p orbitals, and the electronegativity are closely related to the E_g of the materials. These statistically derived insights are reliable and can be used as a reference for designing electronic materials.

Figure 2f presents more detailed quantitative analyses of the four most important features by carrying out Shapley additive explanations (SHAP).[41] SHAP values illustrate the degree of influence of each feature on the E_g. A concentration of data points near SHAP = 0 indicates a minimal influence, whereas positive SHAP values suggest a positive correlation with the E_g, and vice versa. For MPTR, most data points are predominantly clustered around SHAP = 0.5 and −0.5, particularly at −0.5, indicating a more prominent influence on E_g when the MPTR values exceed 3.5. This aligns with the observed trend that higher MPTR values (represented by red points) correspond to lower E_g, consistent with the PCC of −0.59 in Table S1. Typically, elements in higher periods of the same group have lower electron affinity and ionization energy, resulting in a smaller gap between their valence and conduction bands. For example, oxides of chalcogen elements have larger E_g compared to sulfides and selenides because oxygen is positioned higher in the periodic table and has a stronger ability to bind valence electrons. For FTVE$_p$ and MNVE$_p$, the number of valence electrons in unfilled p orbitals has a greater impact on the E_g, and the smaller the number of valence electrons (blue points), the smaller the E_g. This finding can be interpreted as follows: when the elements constituting the material have fewer valence electrons in their unfilled p orbitals, the covalent bonds formed between atoms in the material may be weaker, thereby facilitating electron transitions and resulting in a reduced E_g. A larger E_g typically necessitates higher energy for electron transitions, but in cases where there are fewer valence electrons, the required transition energy is lower, thus leading to a smaller E_g. For ME, compounds that combine more electronegative elements overall have a greater effect on the wide E_g (red points). In addition, when there is a significant electronegativity difference between the elements that make up the material (that is RE), the material tends to have a larger E_g. For instance, oxides (such as Al_2O_3) form highly polar chemical bonds due to the high electronegativity of oxygen and the low electronegativity of metal elements, leading to a wider E_g.

These qualitative and quantitative analyses of composition-based features should be fully considered in the subsequent regulation of the E_g of electronic materials. For example, for experimental and computational communities, the fractions of chemical formulas can be controlled in batches and the SCFC descriptors can be calculated to achieve the purpose of designing materials, or these materials can be optimized based on active learning, which is discussed in the next section.

Materials Screening Based on the EL Model. By using the well-trained EL model, we further predicted the HSE06 E_g of crystal structures in the Materials Project (MP) database.[42] As shown in Figure 3a, from the 144,595 crystal structures collected from MP, we selected 125,602 compounds with a suitable number of element types (2−7) (binary, ternary, quaternary compounds, etc.). Since the training data of the EL model did not contain rare and radioactive elements (Xe, U, etc.), which pose a challenge to the experimental synthesis, we further refined our selection to 116,742 common compounds (see Supplementary Data 1). Then, we calculated the SCFC descriptors for these 116,742 compounds and applied the EL model to predict their HSE06 E_g, where 58,583 compounds were predicted to have E_g > 3.0 eV, suitable for use as semiconductor materials or electronic insulation materials. In addition, 25,925 compounds were predicted to have E_g < 1.0 eV, and they serve as electronic conducting materials, while 11,740 compounds with E_g ranging from 0.9 to 1.6 eV are promising for photovoltaic applications. In addition, considering the materials' processability, we used thermodynamic stability (energy above hull, E_{hull}) as a screening criterion, identifying 24,815 compounds (10,270 compounds with E_g > 3.0 eV, 11,317 compounds with E_g < 1.0 eV, and 3,228 photovoltaics with E_g in the range of 0.9−1.6 eV) with specific E_g values (see Supplementary Data 2−4). For applications such as photovoltaic materials, electrodes, and catalytic materials, electronic conductivity and thermodynamic stability are essential yet not exhaustive indicators. In this work, we focused on the accurate prediction of E_g. When EMGen is applied in the screening and design of other materials with different functions, additional performance indicators, such as light absorption intensity, ionic conductivity, and kinetic activity, should also be considered.

Figure 3b plots the 2D t-distributed stochastic neighbor embedding (t-SNE) of 116,742 compounds alongside their predicted HSE06 E_g. The distribution between SCFC descriptors and E_g displays a regular pattern, as illustrated by the dashed line divisions, which show a gradient from narrow E_g (blue dots) to wide E_g (red dots). In addition, some of the clustered compounds have the same crystal structures or chemical formulas, suggesting that the proposed SCFC descriptors can effectively capture the structural and component information on the compounds. Figure 3c also gives the distribution of EL data (training data) and MP data. The uniform data coverage in Figure 3c ensures that the EL model will not have serious extrapolation during the prediction stage. We calculated the histogram of the predicted HSE06 E_g in Figure 3d and found that the distribution E_g values of compounds approximate a Gaussian distribution primarily falling within the ranges of 0−2.0 and 2.0−5.0 eV, showing the value distribution of electronic conductivity at a large scale. We set up an online query tool (http://aimslab.cn/tools#/query_hse06_band_gap) that enables users to search for compounds with suitable E_g by giving the materials ID in the MP.

Active Learning Model Based on Physical Insights. The well-trained EL model provides an effective and high-precision protocol for large-scale electronic material screening. Moreover, efficiently directed designing of solid compounds

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

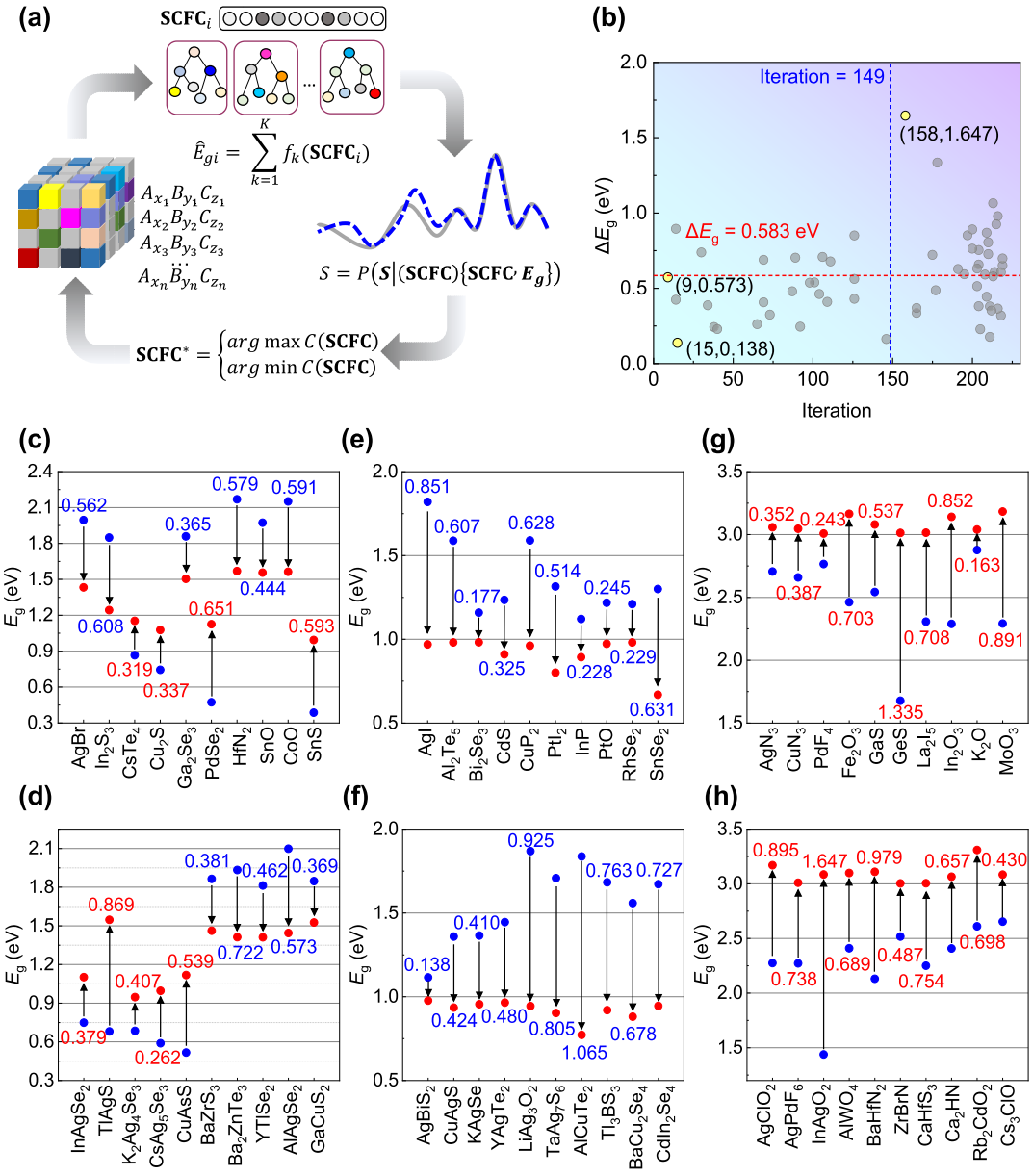


Figure 4. AL model. (a) Process of the AL model. (b) Number of iterations and range of E_g regulation by the AL model. (c,d) Directed design of photovoltaic materials (binary and ternary compounds). (e,f) Directed design of electrode materials (binary and ternary compounds). (g,h) Directed design of insulating materials (binary and ternary compounds). The blue dots denote the original E_g values of compounds, and the red dots are the designed E_g values.

with appropriate E_g presents a challenge for both experimental and computational communities. Here, motivated by the SCFC importance and the SHAP analysis, which show that the top four important features are all based on component information accounting for 62.6% (see Figure 2e), we proposed a global AL model to fully exploit the component information and determine appropriate E_g (GE_g). As shown in Figure 4a, we initiated the process using various chemical components as the initial chemical space, and the E_g values of different components were predicted by the EL model based on the calculated SCFC descriptors. Subsequently, the potential spatial distribution between SCFC descriptors and E_g was simulated according to the surrogate function (S). The posterior probability provided by the surrogate model automatically mines the next potential chemical component, effectively identifying chemical components with appropriate E_g through active explorations and exploitations. More details of the AL model can be found in the Methods section. The design of a material with a suitable GE_g is crucial in determining its application in various systems. For example, Han et al. illustrated that increasing the E_g (approximately more than 1.9 eV), rather than further increasing the ionic conductivity of solid-state electrolytes (SSEs), is critical for inhibiting the dendrite formation in all-solid-state Li batteries.[19] The cathode coatings used to stabilize SSEs must possess a wide E_g to withstand the high voltage of the cathode at the SSE/coating interface. On the contrary, electrode materials require excellent electronic conductivity, typically characterized by a very narrow E_g (generally less than 1.0 eV), to ensure efficient electron transport. Materials with an E_g within the visible wavelength range (0.9−1.6 eV) are considered to have high optical conversion efficiencies and

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

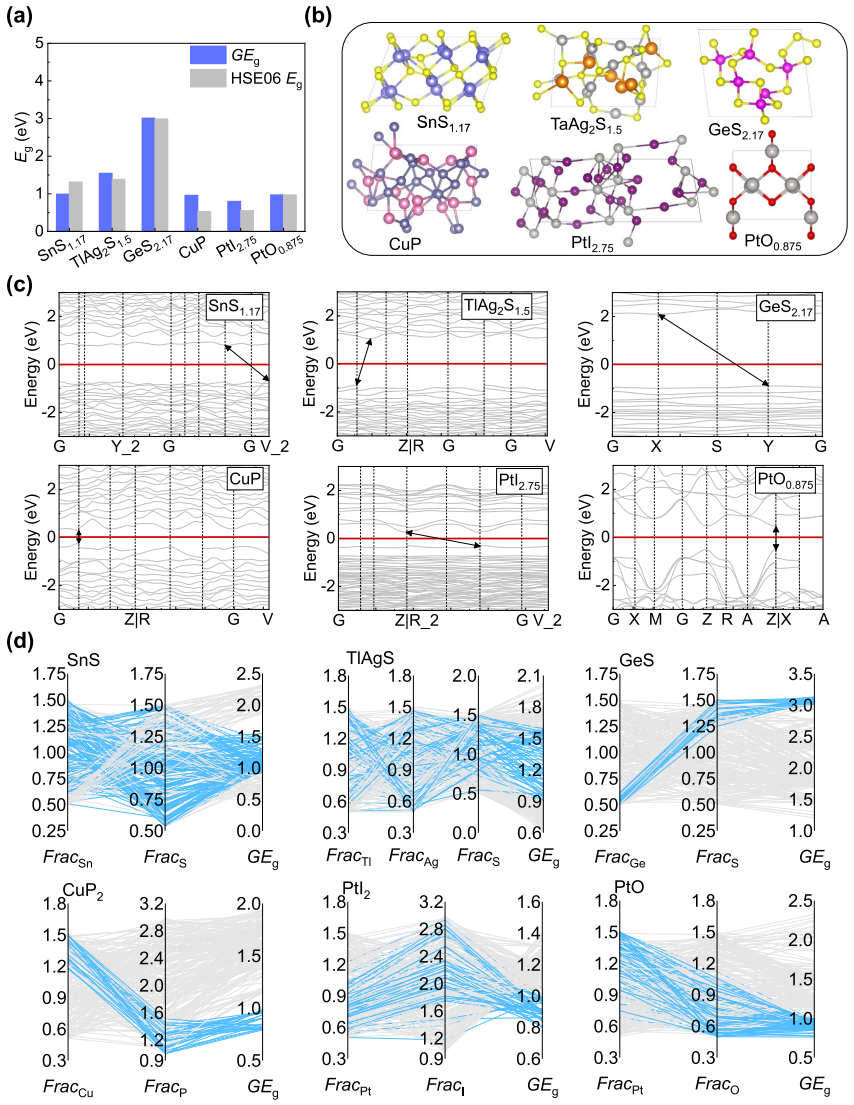


Figure 5. Comparison and validation of the AL model. (a) Comparison between GE_g and FP calculated E_g. (b) Optimized structures of six compounds by the PBE functional. (c) Band structures of six compounds by the HSE06 functional. (d) Visualization of component regulation by the AL model.

Table 1. Comparison between GE_g and HSE06 Calculated E_g[a]

no.	formula	supercell formula	optimized formula	GE_g (eV)	HSE06 E_g (eV)	deviation (eV)	FP time (h)
1	$Sn_{1.215}S_{1.432}$	$Sn_{12}S_{12}$	$Sn_{12}S_{14}$	0.993	1.314	0.321	151.42
2	$Tl_{0.574}Ag_{1.17}S_{0.944}$	$Tl_{12}Ag_{12}S_{12}$	$Tl_6Ag_{12}S_9$	1.547	1.384	0.163	120.40
3	$Ge_{0.5692}S_{1.307}$	$Ge_{12}S_{12}$	Ge_6S_{13}	3.013	2.99	0.023	9.07
4	$Cu_{1.361}P_{1.416}$	Cu_8P_{16}	$Cu_{14}P_{14}$	0.962	0.530	0.432	56.14
5	$Pt_{0.8433}I_{2.238}$	Pt_8I_{16}	Pt_8I_{22}	0.801	0.556	0.245	91.33
6	$Pt_{0.8168}O_{0.6303}$	Pt_8O_8	Pt_8O_7	0.973	0.973	0.000	132.38

[a]The FP time denotes the approximate computing hours by using the HSE06 functional on a supercomputer node (24-core Intel Xeon E5-2685 2.6 GHz processer). The average deviation between the GE_g and HSE06 calculated E_g values is only 0.197 eV.

are promising for photovoltaic applications, such as (double) perovskites and quaternary chalcogenides.[10,43,44] π-Conjugated polymers with narrow E_g (< 1.6 eV) absorbing in the near-infrared range (760−1100 nm) are an interesting family of semiconductor materials for a variety of functional optoelectronic applications, such as organic field-effect transistors, photodetectors, and electrochromic devices.[45] Using typical examples like photovoltaic materials (thin films, HTLs, and ETLs, 0.9 < E_g < 1.6 eV), electrode materials (battery electrodes and electrocatalysts, E_g < 1.0 eV), and insulating materials (solid-state electrolytes, E_g > 3.0 eV) as examples, we randomly selected 20 compounds (binary and ternary compounds, totaling 60 compounds) from the original data for targeted design. For photovoltaic materials, we selected the compounds with E_g > 1.6 eV or E_g < 0.9 eV and applied the bidirectional design to regulate the E_g. Similarly, we selected the compounds with E_g > 1.0 eV and E_g < 3.0 eV as the cases

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

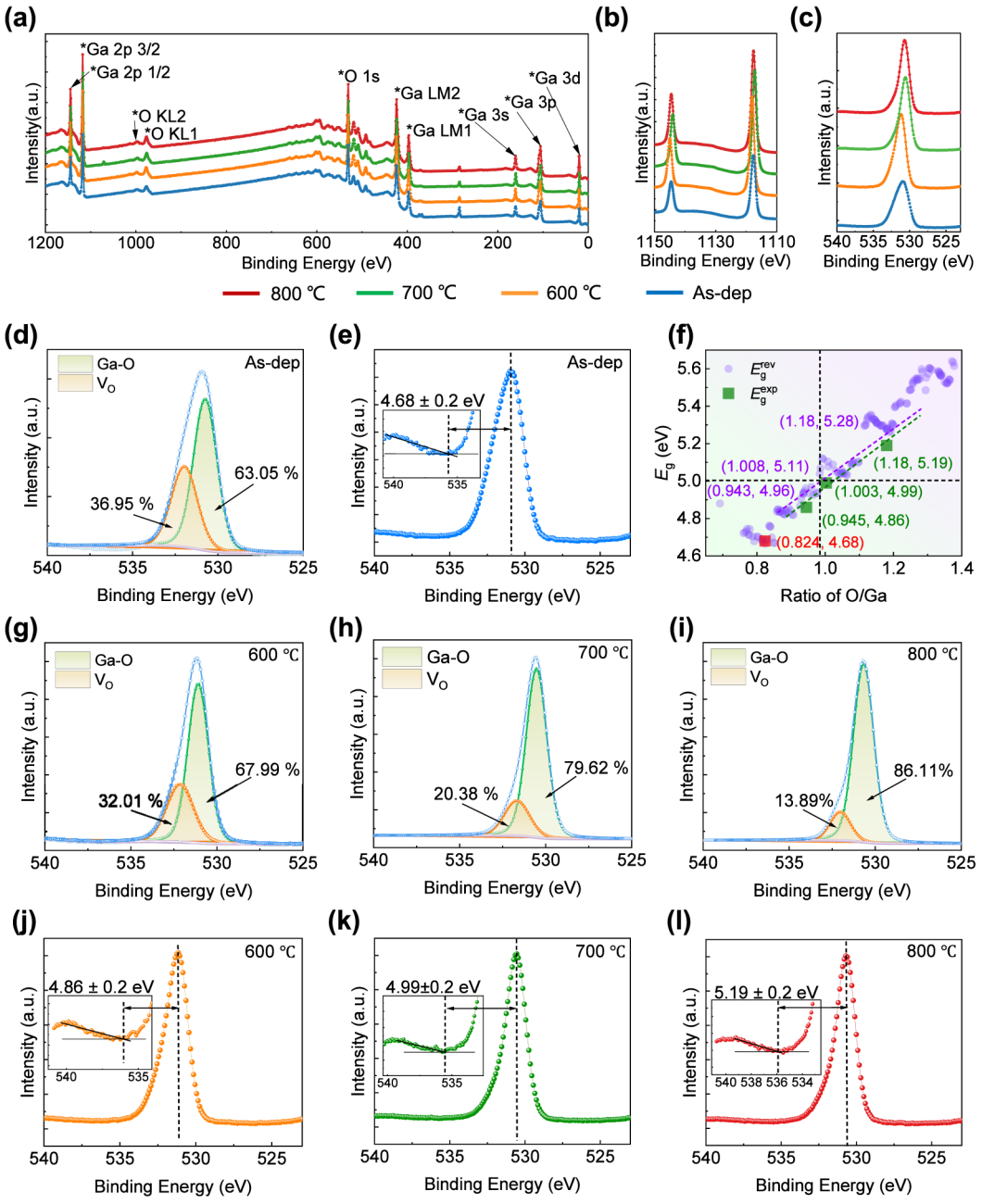


Figure 6. Directed design by using EMGen with experimental validations. (a) Full XPS spectra of Ga_xO_y samples annealed at different temperatures. (b) Ga 2p spectra. (c) O 1s spectra. (d) Loss structure in the As-dep state. (e) Analysis of O 1s spectra of Ga_xO_y films in the As-dep state. (f) Comparison of the predicted and experimental E_g, where the red point refers to the experimental reference point for the E_g revision. (g–i) Loss structure at 600, 700, and 800 °C. (j–l) Analysis of O 1s spectra of Ga_xO_y films at 600, 700, and 800 °C.

for designing electrode materials and insulating materials, respectively.

As shown in Figure 4b, in these 60 cases, the average number of iterations is 149 (the minimum is 9, each iteration takes approximately 0.7 s), with E_g adjustments ranging from 0.138 to 1.647 eV (mean = 0.583 eV). In essence, the AL model can regulate the E_g of a material in approximately 1.7 min, with changes up to 0.583 eV. Figure 4c–h shows the original E_g (blue dots) and GE_g (red dots) of specific compounds, together with the changes of E_g. Figure 4c,d provides the directional design of potential photovoltaic materials. For a compound with E_g > 1.6 eV, the optimization objective is minimizing the E_g (indicated by the downward arrow), whereas for a compound with E_g < 0.9 eV, the optimization objective is maximizing the E_g (indicated by the upward arrow). Figure 4e,f addresses potential electrode materials, and Figure 4g,h discusses potential insulating materials. We further applied FP calculations (the PBE functional for structure relaxations and the HSE06 functional for electronic structure calculations) to evaluate the E_g of six designed compounds. A comparison between GE_g and FP calculated E_g is presented in Figure 5a, with relevant statistics summarized in Table 1. Excellent agreement is found between the GE_g and FP calculated E_g values (with an average deviation of 0.197 eV), verifying the great superiority of the current EL and AL technologies. Figure 5b provides the optimized

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

structures of six compounds, whose initial structures were obtained by establishing supercells and modifying the chemical component (see Note S6). Figure 5c gives the band structures by the HSE06 functional, and Figure S5 provides the density of states (DOSs) and local DOS of each element for these materials. The visualization of directed design is plotted in Figure 5d and Figures S6−S9, where the fraction of components plays an important role in changing the E_g. For example, appropriately reducing the ratio of Ge to S (resulting in Ge defects) can increase the E_g. Importantly, it takes only approximately 1.7 min for the design of a given compound, whereas if the FP calculation is adopted, it will take approximate 93.46 h for each structure (150 optimizations would take ∼14,000 h). Therefore, we concluded that our current EL-AL scheme provides a possibility of achieving high-precision FP accuracy and has priority in complex material systems.

Design DUV Materials. The proposed EMGen facilitates efficient, directed design focused on the chemical component of electronic materials, serving as an excellent toolbox for designing optoelectronic materials with specific E_g. Here, gallium oxide (Ga_2O_3) was selected to assess the practical capabilities of EMGen. Ga_2O_3 is recognized as a crucial semiconductor with a wide E_g, holding substantial promise in the field of DUV optoelectronic devices (see the atomic structure in Figure S10). As the energy band structure is a key parameter that controls the performance of Ga_2O_3,[46,47] an in-depth examination of the alterations in the chemical state and band structure of Ga_xO_y via EMGen and experiments was performed. Although Ga_2O_3 has been extensively studied, here, we would like to highlight the potential of EMGen in regulating electronic materials by combining a mature experimental process. It should be noted that even high-precision HSE06 E_g predictions might differ from experimental measurements; therefore, we initially utilized an experimental reference point to investigate the deviation between prediction at the HSE06 level and the experimental result. More discussions are provided in Note S5. Figure 6a shows full X-ray photoelectron spectroscopy (XPS) of Ga_xO_y in the As-deposited state (blue line), Figure 6b,c shows that the Ga $2p_{3/2}$ and Ga $2p_{1/2}$ peaks are around 1144 and 1117 eV, respectively, and the O 1s peak is concentrated around 530−531 eV (blue line), giving that the composition ratio of O and Ga (O/Ga) in the deposited state is 0.824 that is $GaO_{0.824}$. Although $GaO_{0.824}$ is completely different from Ga_2O_3, it has a common crystal to monoclinic Ga_2O_3 (close to the parent compound) (see the X-ray diffraction (XRD) pattern in Figure S11). Further evidence comes from the analysis of the $O_{II}/(O_I + O_{II})$ ratio, as shown in Figure 6d. Through Gaussian fitting analysis, each O 1s nuclear energy level spectrum can be deconvolved into two components: O_I represents the contribution of the Ga−O bond of Ga_2O_3, and O_{II} can be attributed to O^{2-} ions in the oxygen-deficient region.[48] Therefore, the $O_{II}/(O_I + O_{II})$ intensity ratio serves as an indicator of the oxygen vacancy density in the Ga_xO_y film.

According to the calculations of the O 1s nuclear energy level and loss structure (Figure 6e), the experimental E_g (E_g^{exp}) of the unannealed film is 4.68 eV. For $GaO_{0.824}$, the predicted E_g (E_g^{pre}) of 2.42 eV was predicted by the EL model (which was based on the HSE06 computational data), resulting in a deviation of 2.26 eV (see Note S5 for more discussion). Therefore, according to this experimental reference point, a weighted formula (the revised E_g^{pre}, E_g^{rev}) that related to the O/Ga is constructed to bridge this bias: $E_g^{rev} = [(E_g^{pre} + 2.26/(O/Ga/0.824) + (E_g^{pre} + 2.26)]/2$. Subsequently, the AL model was used for the directed designing of Ga_xO_y compounds, obtaining the corresponding ratio of O/Ga and the E_g^{rev}, as shown in Figure 6f (purple points). It can be seen that as the oxygen vacancy decreases, that is, as the ratio of O/Ga increases, the E_g^{rev} roughly shows a monotonically rising trend. When the ratio of O/Ga approaches 1.0, the E_g^{rev} can reach ∼5.0 eV, providing clear guidance for achieving Ga_xO_y films with enhanced insulating properties. Following the EMGen results, Ga_xO_y films were prepared at varying annealing temperatures (600, 700, and 800 °C). Figure 6a−c shows the full XPS spectrum, where the binding energy of Ga $2p_{3/2}$ and Ga $2p_{1/2}$ peaks remained largely unaffected by the annealing process,[49] and Ga remained predominantly in its highest oxidation state (Ga^{3+}). As the annealing temperature increases, the composition ratios of O and Ga of the annealed sample are 0.945, 1.033, and 1.180, respectively. This clearly shows that these Ga_xO_y films lack oxygen atoms, and additional O_2 can improve their stoichiometric composition. From Figure 6g−i, we observed that as the annealing temperature increases, the $O_{II}/(O_I + O_{II})$ ratio shows monotonic decreasing trends of 32.01, 20.38, and 13.89%, respectively. This trend signifies that higher annealing temperatures energized oxygen atoms, fostering increased oxygen adsorption and diffusion to mitigate oxygen vacancies within the Ga_xO_y film, thereby enhancing the chemical state.[50]

According to the calculations of the O 1s nuclear energy level and loss structure (Figure 6j−l), changes in the energy band structure as the annealing temperature increases can be observed. From Figures S12 and S13, it can be observed that the energy difference of the unannealed film is 2.25 eV, and the energy difference of the films annealed is larger, indicating enhanced insulation performance. The E_g^{exp} of the film annealed at 600, 700, and 800 °C are 4.86, 4.99, and 5.19 eV, respectively. This trend of increasing E_g^{exp} can be attributed to the reduction of oxygen vacancy defects, resulting in better insulating properties of Ga_xO_y films. The experimental results confirm the trends predicted by using the proposed EMGen (Figure 6f). Combining changes in the optical E_g and improvements in the chemical state, it was deduced that the film annealed at 800 °C yielded the greatest E_g enhancement. This development is poised to further expand the theoretical detection range of Ga_xO_y films within photodetectors, extending their applicability to DUV light below 240 nm. At the same time, the augmented E_g also helps the improvement in breakdown voltage and the heat resilience performance of the amorphous Ga_xO_y film,[51] thereby unlocking considerable potential within the realm of power electronics applications.

The case of Ga_xO_y in this work provides a good foundation and proof for the application of EMGen in other materials systems. The surrogate model that we developed is benchmarked against theoretical calculations, which inevitably introduces discrepancies compared to experimental results. To address this, we employed an experimental reference point to calibrate the model's predictions. This step is essential for applying AI models to practical materials screening. Moreover, we consider this to be a representative framework for achieving a closed-loop system that integrates computation, AI, and experimentation. When designing other materials with specific E_g, a similar sequential and single experimental reference point can be used to calibrate the model's predictions, thus reducing the cost of materials design based on experiments. For other

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

material properties, such as ionic conductivity or mechanical properties, the same architecture as EMGen can be applied, achieving the materials directed design with desired properties.

CONCLUSIONS

In this paper, a protocol was developed for discovering and designing the electronic materials with suitable E_g. We first compiled a high-fidelity data set and designed a general descriptor SCFC that encapsulates structure and component information. Subsequently, we constructed a high-precision and efficient model EL for E_g prediction of any inorganic crystal structure. We demonstrated that the EL model has lower errors than the advanced graph neural network models. Leveraging the low cost of the EL model, we built an E_g database of more than 116,700 entries (https://aimslab.cn/#/query_hse06_band_gap) facilitating large-scale prediction and screening. Further, interpreting the EL model at the physical level enabled us to use key component information to regulate the E_g of the compounds. We adopted a global optimization model AL for the directed design of E_g, which was validated by the subsequent theoretical calculations.

Our work shows how material data, descriptors, models, and further materials design can be synergistically combined, transcending the typical constraints of discovering materials, a focus of many current studies. In addition to the design of E_g of electronic materials, EMGen can be adapted to explore a variety of material functions, such as magnetism, mechanical properties, kinetic activity, and optical properties. Many interpretable methods have been developed to extract the physical factors closely related to material properties, to understand hidden causal relationships, and to generate scientific hypotheses. However, the high-dimensional nonlinear relationships between multiple factors and materials properties directly hinder the development of materials generation processes guided by these key chemical factors. How to fully utilize model interpretability and achieve efficient, directed materials generation to obtain materials with specific properties remain an open challenge. Therefore, in this work, we proposed an active learning model based on physical insights provided by interpretable ensemble learning methods to optimize materials for desired properties. This approach is versatile and not limited to any specific type of material or materials property. By integrating these physical insights derived from interpretability into the next optimization model, we can achieve targeted materials design. Moreover, at present, the *ab initio* crystal structure prediction based on chemical components can be integrated with EMGen to facilitate *ab initio* crystal materials design. The prediction accuracy and speed of EMGen foster optimism that more high-precision material function prediction models and databases will be established in the future. The intelligent design of materials using EMGen will be an important development trend in materials science.

METHODS

SCFC Descriptor. Inspired by our physical intuition and previous studies[28,44,52] that structure and component information were key factors in materials property modeling, a descriptor was proposed in this work for inorganic compounds, that is, the structure and component fusion coding (SCFC). The SCFC descriptor of a material can be simply obtained by concatenating the structural feature vector ($\boldsymbol{v}_i$) and component feature vector ($\boldsymbol{l}_i$):

$$\mathrm{SCFC}_i = (\boldsymbol{v}_i \oplus \boldsymbol{l}_i) \tag{1}$$

The structural feature vector $\boldsymbol{v}_i$ describes the crystal skeleton information on the material that consists of four attributes: the crystal system (CS), symmetric state (SS), mean atomic volume (MAV), and mean volume atoms (MNAV). Seven CSs (cubic, hexagonal, trigonal, tetragonal, orthorhombic, monoclinic, and triclinic) were encoded by seven-dimensional one-hot codes, for example, [0, 0, 0, 1, 0, 0, 0] for the tetragonal structure. SS was used for distinguishing the centrosymmetric structures (CSS) and noncentrosymmetric structures (non-CSS) and is coded by one character (0 or 1). MAV and MNVA describe the arrangement of atoms within the structure, and their values reflected the atomic coordination number and chemical bond information. The component feature vector $\boldsymbol{l}_i$ was extracted from chemical formulas by using the isolated elemental properties. Here, eight general elemental properties were selected according to the previous material modeling:[9,13] the atomic weight (AW), row number and column number in the periodic table, atomic number, atomic radius, electronegativity, valence electrons in the orbital, and band center. Considering the different numbers of atoms and atomic types in the compound, the original properties were mathematically coded. For the ith material having j elements, we calculated the mean atomic weight (MAW) as

$$\mathrm{MAW}_i = \sum_{k=1}^{j} \mathrm{AW}_k{}^{*}\mathrm{frac}_k \tag{2}$$

where frac denotes the fraction of the kth element. Similarly, we calculated the MPTR, MPTC, MAN, MAR, ME, $\mathrm{MNVE_s}$, $\mathrm{MNVE_p}$, $\mathrm{MNVE_d}$, and $\mathrm{MNVE_f}$, and the computational details are provided in Note S2. Moreover, to extract the difference information between elements, we counted the RAN, RAR, and RE.

$$\mathrm{MR}_i = \sum_{k=1}^{j} \sum_{t=1}^{j} \max(R_k - R_t) \tag{3}$$

where MR represents the mean RAN, RAR, or RE, and R is the AN, AR, or RE. Furthermore, the band center (BC) was closely related to the density of states according to the d-band theory, which may be useful for the E_g prediction. We estimated the BC using the geometric mean of electronegativity:

$$\mathrm{BC}_i = \prod_{k=1}^{j} E_k{}^{*}\mathrm{frac}_k \tag{4}$$

The vectors of $\boldsymbol{v}_i$ and $\boldsymbol{l}_i$ were concatenated to form a 28-dimensional fusion code, which was low-dimensional, physically interpretable, computationally simple, and easily extensible. From the input of the crystal structure to the formation of an SCFC descriptor, the computational time was approximately 0.07 s on a single-core CPU. In addition, model training costs on ML algorithms and even deep neural networks were very low due to relatively low dimensionality. SCFC was calculated with the Matminer tool.[53]

EL Model. Based on 8409 high-fidelity data, we established the E_g prediction model by using the fusion coding and ensemble learning model (EL). The EL model was implemented by the extreme gradient boosting (XGBoost) algorithm, which integrated multiple classification and regression trees (CART) and calculated the final prediction result by an additive strategy:[38]

$$\hat{y}_i = \varnothing(\mathrm{SCFC}_i) = \sum_{k=1}^{K} f_k(\mathrm{SCFC}_i), f_k \in \mathcal{F} \tag{5}$$

where $\mathcal{F} = \{f(\mathrm{SCFC}) = w_q(\mathrm{SCFC})\}$ ($q: \mathbb{R}^m \to T$, $w \in \mathbb{R}^T$) is the space of CART, q denotes the structure of each CART that maps an entry to the corresponding leaf node, and T is the number of leaves in the CART. f_k was decided by q and leaf weight w. This shows that EL has a higher generalization performance than a single CART, and the comprehensive investigation of multiple CARTs can greatly reduce the influence of the overfitting phenomenon. To learn the set of f_k used in the EL model, we minimized the following regularized objective:

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

$$\mathcal{H}(\varnothing) = \sum_i \Omega(\hat{y}_i, y_i) + \sum_k \mathcal{L}(f_k) \quad (6)$$

$$\mathcal{L}(f) = \Upsilon T + \frac{1}{2}\lambda \|w\|^2 \quad (7)$$

where Ω measured the difference between the prediction results and true values and $\mathcal{L}$ is a regularization item with parameters Υ and λ that penalizes the complexity of CARTs. In this work, we adopted 10-fold cross-validations to evaluate the model performance, that is, completing 10 training sessions, each using 90% of that data for training and 10% for validations. Hyperparameters for the tree booster were optimized by using a random searching method: step size shrinkage used in the update to prevent overfitting (0.04), subsample ratio of the training instances (0.98), subsample ratio of columns when building each CART (0.8), and maximum depth of CART (8).

AL Model. The establishment of the EL model provides an efficient predictor for finding materials with suitable E_g on a large scale. At a deeper level, embedded feature engineering reveals the key factors affecting material E_g qualitatively and quantitatively, including linearity and nonlinearity. This has implications for the reverse or directional design of materials with desired E_g. Taking E_g of materials as the optimization objective and SCFCs as the independent variables, an optimization problem is naturally formed. Although EL provides critical physical factors, setting up the structures or combination is a scattershot endeavor. EMGen proposed an AL model based on the Bayesian optimization method combining fusion coding and active learning.[54] For a given material, the space of its chemical composition is unknown, and its mapping to E_g is ambiguous, which was defined as the objective function C. The AL model starts by initializing a surrogate function S in place of the C and constructs an acquisition function Acq to sample the next different components of the material. Then, the data set $\{\boldsymbol{X}, \boldsymbol{Y}\}$ is constructed by the collected sample points, the E_g values are predicted by the EL model ($\varnothing(\boldsymbol{x})$), and the S is updated to obtain the posterior distribution (as the prior distribution for the next sampling) and is getting closer to the C:

$$S = P(\boldsymbol{S}(X)|\{\boldsymbol{X}, \boldsymbol{Y}\}) \quad (8)$$

Then, such an active learning model can help design materials with suitable E_g. For example, when we want to design a material with E_g > 3.0 eV or E_g < 1.0 eV, we can set a global maximizer or minimizer:

$$x^* = \begin{cases} arg\ \max C(x),\ x \in \chi \\ arg\ \min C(x),\ x \in \chi \end{cases} \quad (9)$$

where χ is the space of SCFC for a given material and x denotes the designed material. During the AL model, a total of 220 samplings were performed to find the material components with suitable E_g, including 200 explorations for the predictions on the uncertainly in the S and 20 exploitations for the predictions on the optimal objectives.

Electronic Structure Calculation. All first-principles calculations were accomplished using the Vienna *ab initio* simulation package (VASP).[55] The PBE generalized gradient approximation (GGA) functionals and project-augmented wave (PAW) atom potentials were employed to perform geometric structure optimizations.[56,57] The PBE +U method was used to yield a reasonable electronic and magnetic ground state for bulk materials with transition metal elements, and the U value of 3.0 eV was selected for Sn and Pt, and 4.0 eV for Cu and Ta. The cutoff energy for the plane-wave basis was set as 500 eV. The structure optimization process ended when the energy convergence was lower than 10^{-5} eV and the atomic force was less than 0.01 eV/Å. Further, the hybrid functional (HSE06) was applied for calculating the accurate electronic structures.[58] The high symmetry points of electrons were generated by using the VASPKIT tool.[59]

In the optimization process, for each material, we selected the compound whose charge is closest to zero (i.e., neutral). Since a change in chemical composition inevitably results in a charge imbalance, the adjustable ratio in the optimization process takes into account extreme imbalances, making the optimized material similar to a material with a point defect. In addition, when using DFT for structural optimization, we randomly constructed three structures for each material and selected the lowest energy structure as a stable structure for subsequent electronic structure calculation.

Deposition of Ga_2O_3 Films. First, the p-Si substrate was sequentially immersed in acetone, alcohol, and water and ultrasonically cleaned for 10 min. Subsequently, in an atmosphere of high-purity argon gas (Ar, >99.99% purity), Ga_2O_3 thin films were deposited by radio frequency (RF) magnetron sputtering using a 2 in.-diameter, 99.99% pure Ga_2O_3 target. Oxygen gas (O_2, >99.99% purity) was introduced to compensate for any oxygen loss during the sputtering process. All films were deposited at a base pressure of approximately 1×10^{-4} Pa with a sputtering working pressure of 1.0 Pa, an RF power of 75 W, and a deposition time of 60 min, maintaining a constant argon-to-oxygen ratio (Ar:O_2 = 6:1).

Testing and Characterization. Following film deposition, the samples were annealed at 600, 700, and 800 °C, respectively, in an air muffle furnace (KSL-1100X-S, HF-Kejing) for 2 h and cooled to room temperature. Subsequently, various characterization methods were employed for surface analysis. X-ray photoelectron spectroscopy (XPS, AXIS Ultra DLD) was performed to confirm the bonding states of the elements in the gallium oxide thin films. The XPS measurements were initially conducted on the as-deposited films exposed to ambient conditions. The binding energy was calibrated using the C 1s peak at 284.8 eV, and data analysis was carried out using the CasaXPS software package. Gaussian and Lorentzian mixed line shapes, along with Shirley background subtraction, provided excellent peak fitting. The phase identification and crystal structure of the deposited films were determined by X-ray diffraction (XRD, D8 Advance, Bruker AXS) using Cu Kα radiation in the 5−80° range. All tests were conducted at room temperature.

ASSOCIATED CONTENT

Data Availability Statement

The database of HSE06 band gaps is available at https://aimslab.cn/#/query_hse06_band_gap.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsnano.4c12166.

HSE06 E_g of 116,742 compounds (XLSX)
Identified 10,270 compounds with E_g > 3.0 eV (XLSX)
Identified 11,317 compounds with E_g < 1.0 eV (XLSX)
Identified 3,228 photovoltaics with E_g in the range of 0.9−1.6 eV (XLSX)
Materials repository; computational details and analysis of the SCFC descriptor; parameters of CGCNN and MEGNet; discussion of deviation between the EL prediction and experiment measurement; statistics of continuous features in the SCFC descriptor; performance of the EL model evaluated by 10-fold cross-validations; comparison of EL, CGCNN, MEGNet, RF, DT, and SVM on predictive performance; design materials by the AL model; DOSs of six validated materials by using HSE06 calculations; XRD patterns and band shift diagram of Ga_xO_y films (PDF)

Accession Codes

The implementations of the codes used for calculating the SCFCs, training the EL model, and performing the AL models are publicly available at https://github.com/CodingWZL/EMGen. The codes of Bayesian optimization are available at https://github.com/durong/BayesianOptimization. The codes of SHAP analysis are available at https://github.com/slundberg/shap. The codes of CGCNN and MEGNet are available at https://github.com//txie-93//cgcnn and https://github.com/materialsvirtuallab/megnet.

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Authors

Fengqi You − *Systems Engineering and Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0001-9609-4299; Email: fengqi.you@cornell.edu

Gang Liu − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology, Shanghai Jiao Tong University, Shanghai 200240, China;* orcid.org/0000-0002-3142-9323; Email: gang.liu@sjtu.edu.cn

Jinjin Li − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology, Shanghai Jiao Tong University, Shanghai 200240, China;* orcid.org/0000-0003-4661-4051; Email: lijinjin@sjtu.edu.cn

Authors

Zhilong Wang − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology and Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China; Systems Engineering and Robert Frederick Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, New York 14853, United States;* orcid.org/0000-0002-1910-3654

Sixian Liu − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology and Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;* orcid.org/0000-0003-2788-8137

Kehao Tao − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology and Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

An Chen − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology and Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;* orcid.org/0000-0003-0470-688X

Hongxiao Duan − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology and Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

Yanqiang Han − *National Key Laboratory of Advanced Micro and Nano Manufacture Technology and Department of Micro/Nano Electronics, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;* orcid.org/0000-0001-5454-0617

Complete contact information is available at: https://pubs.acs.org/10.1021/acsnano.4c12166

Author Contributions

†Zhilong Wang and Sixian Liu contributed equally to this study. Zhilong Wang: conceptualization, methodologies, visualization, data curation, and writing of the original draft. Sixian Liu: methodologies, visualization, and writing of the original draft. Hongxiao Duan, An Chen, Kehao Tao, and Yanqiang Han: methodologies, visualization, and data curation. Fengqi You, Gang Liu, and Jinjin Li: conceptualization, methodologies, supervision, resources, and review and editing.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Jinjin Li and Gang Liu are grateful for financial support provided by the National Key R&D Program of China (nos. 2021YFC2100100 and 2022YFB4700102), the National Natural Science Foundation of China (nos. 32301040, 61974090, and 62111540271), and the Shanghai Science and Technology Project (nos. 21JC1403400 and 23JC1402300).

REFERENCES

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547−555.

(2) Batra, R.; Song, L.; Ramprasad, R. Emerging Materials Intelligence Ecosystems Propelled by Machine Learning. *Nat. Rev. Mater.* **2021**, *6* (8), 655−678.

(3) Wang, Z.; Chen, A.; Tao, K.; Han, Y.; Li, J. MatGPT: A Vane of Materials Informatics from Past, Present, to Future. *Adv. Mater.* **2024**, *36* (6), 2306733.

(4) Rao, Z.; Tung, P.-Y.; Xie, R.; Wei, Y.; Zhang, H.; Ferrari, A.; Klaver, T. P. C.; Körmann, F.; Sukumar, P. T.; Kwiatkowski da Silva, A.; Chen, Y.; Li, Z.; Ponge, D.; Neugebauer, J.; Gutfleisch, O.; Bauer, S.; Raabe, D. Machine Learning−Enabled High-Entropy Alloy Discovery. *Science* **2022**, *378* (6615), 78−85.

(5) López, C. Artificial Intelligence and Advanced Materials. *Adv. Mater.* **2023**, *35* (23), 2208683.

(6) Wang, Z.; Chen, A.; Tao, K.; Cai, J.; Han, Y.; Gao, J.; Ye, S.; Wang, S.; Ali, I.; Li, J. AlphaMat: A Material Informatics Hub Connecting Data, Features, Models and Applications. *Npj Comput. Mater.* **2023**, *9* (1), 130.

(7) Zhong, X.; Gallagher, B.; Liu, S.; Kailkhura, B.; Hiszpanski, A.; Han, T. Y.-J. Explainable Machine Learning in Materials Science. *Npj Comput. Mater.* **2022**, *8* (1), 204.

(8) Wang, Z.; Zhang, H.; Li, J. Accelerated Discovery of Stable Spinels in Energy Systems via Machine Learning. *Nano Energy* **2021**, *81*, No. 105665.

(9) Wang, Z.; Lin, X.; Han, Y.; Cai, J.; Wu, S.; Yu, X.; Li, J. Harnessing Artificial Intelligence to Holistic Design and Identification for Solid Electrolytes. *Nano Energy* **2021**, *89*, No. 106337.

(10) Wang, Z.; Han, Y.; Lin, X.; Cai, J.; Wu, S.; Li, J. An Ensemble Learning Platform for the Large-Scale Exploration of New Double Perovskites. *ACS Appl. Mater. Interfaces* **2022**, *14* (1), 717−725.

(11) Chen, A.; Wang, Z.; Gao, J.; Han, Y.; Cai, J.; Ye, S.; Li, J. A Data-Driven Platform for Two-Dimensional Hybrid Lead-Halide Perovskites. *ACS Nano* **2023**, *17* (14), 13348−13357.

(12) Chen, A.; Cai, J.; Wang, Z.; Han, Y.; Ye, S.; Li, J. An Ensemble Learning Classifier to Discover Arsenene Catalysts with Implanted Heteroatoms for Hydrogen Evolution Reaction. *J. Energy Chem.* **2023**, *78*, 268−276.

(13) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), No. 145301.

(14) Gao, J.; Wang, Z.; Han, Y.; Gao, M.; Li, J. CEEM: A Chemically Explainable Deep Learning Platform for Identifying Compounds with Low Effective Mass. *Small* **2024**, *20* (4), 2305918.

(15) Konuma, I.; Goonetilleke, D.; Sharma, N.; Miyuki, T.; Hiroi, S.; Ohara, K.; Yamakawa, Y.; Morino, Y.; Rajendra, H. B.; Ishigaki, T.; Yabuuchi, N. A near Dimensionally Invariable High-Capacity Positive Electrode Material. *Nat. Mater.* **2023**, *22*, 225−234.

(16) Ko, S.; Obukata, T.; Shimada, T.; Takenaka, N.; Nakayama, M.; Yamada, A.; Yamada, Y. Electrode Potential Influences the Reversibility of Lithium-Metal Anodes. *Nat. Energy* **2022**, *7* (12), 1217−1224.

(17) Bu, T.; Ono, L. K.; Li, J.; Su, J.; Tong, G.; Zhang, W.; Liu, Y.; Zhang, J.; Chang, J.; Kazaoui, S.; Huang, F.; Cheng, Y.-B.; Qi, Y. Modulating Crystal Growth of Formamidinium−Caesium Perovskites for over 200 Cm2 Photovoltaic Sub-Modules. *Nat. Energy* **2022**, *7* (6), 528−536.
(18) Ruiz-Preciado, M. A.; Gota, F.; Fassl, P.; Hossain, I. M.; Singh, R.; Laufer, F.; Schackmar, F.; Feeney, T.; Farag, A.; Allegro, I.; Hu, H.; Gharibzadeh, S.; Nejand, B. A.; Gevaerts, V. S.; Simor, M.; Bolt, P. J.; Paetzold, U. W. Monolithic Two-Terminal Perovskite/CIS Tandem Solar Cells with Efficiency Approaching 25%. *ACS Energy Lett.* **2022**, *7* (7), 2273−2281.
(19) Han, F.; Westover, A. S.; Yue, J.; Fan, X.; Wang, F.; Chi, M.; Leonard, D. N.; Dudney, N. J.; Wang, H.; Wang, C. High Electronic Conductivity as the Origin of Lithium Dendrite Formation within Solid Electrolytes. *Nat. Energy* **2019**, *4* (3), 187−196.
(20) Li, G.; Yoon, K.-Y.; Zhong, X.; Wang, J.; Zhang, R.; Guest, J. R.; Wen, J.; Zhu, X.-Y.; Dong, G. A Modular Synthetic Approach for Band-Gap Engineering of Armchair Graphene Nanoribbons. *Nat. Commun.* **2018**, *9* (1), 1687.
(21) Pickup, L.; Sigurdsson, H.; Ruostekoski, J.; Lagoudakis, P. G. Synthetic Band-Structure Engineering in Polariton Crystals with Non-Hermitian Topological Phases. *Nat. Commun.* **2020**, *11* (1), 4431.
(22) Chaves, A.; Azadani, J. G.; Alsalman, H.; da Costa, D. R.; Frisenda, R.; Chaves, A. J.; Song, S. H.; Kim, Y. D.; He, D.; Zhou, J.; Castellanos-Gomez, A.; Peeters, F. M.; Liu, Z.; Hinkle, C. L.; Oh, S.-H.; Ye, P. D.; Koester, S. J.; Lee, Y. H.; Avouris, Ph.; Wang, X.; Low, T. Bandgap Engineering of Two-Dimensional Semiconductor Materials. *Npj 2D Mater. Appl.* **2020**, *4* (1), 29.
(23) Ortstein, K.; Hutsch, S.; Hambsch, M.; Tvingstedt, K.; Wegner, B.; Benduhn, J.; Kublitski, J.; Schwarze, M.; Schellhammer, S.; Talnack, F.; Vogt, A.; Bäuerle, P.; Koch, N.; Mannsfeld, S. C. B.; Kleemann, H.; Ortmann, F.; Leo, K. Band Gap Engineering in Blended Organic Semiconductor Films Based on Dielectric Interactions. *Nat. Mater.* **2021**, *20* (10), 1407−1413.
(24) Faye, O.; Eduok, U.; Szpunar, J. A. Boron-Decorated Graphitic Carbon Nitride (g-C_3N_4): An Efficient Sensor for H_2S, SO_2, and NH_3 Capture. *J. Phys. Chem. C* **2019**, *123* (49), 29513−29523.
(25) Zhao, F.; Feng, Y.; Wang, Y.; Zhang, X.; Liang, X.; Li, Z.; Zhang, F.; Wang, T.; Gong, J.; Feng, W. Two-Dimensional Gersiloxenes with Tunable Bandgap for Photocatalytic H_2 Evolution and CO_2 Photoreduction to CO. *Nat. Commun.* **2020**, *11* (1), 1443.
(26) De Bastiani, M.; Mirabelli, A. J.; Hou, Y.; Gota, F.; Aydin, E.; Allen, T. G.; Troughton, J.; Subbiah, A. S.; Isikgor, F. H.; Liu, J.; Xu, L.; Chen, B.; Van Kerschaver, E.; Baran, D.; Fraboni, B.; Salvador, M. F.; Paetzold, U. W.; Sargent, E. H.; De Wolf, S. Efficient Bifacial Monolithic Perovskite/Silicon Tandem Solar Cells via Bandgap Engineering. *Nat. Energy* **2021**, *6* (2), 167−175.
(27) Tao, J.; Liu, X.; Shen, J.; Han, S.; Guan, L.; Fu, G.; Kuang, D.-B.; Yang, S. F-Type Pseudo-Halide Anions for High-Efficiency and Stable Wide-Band-Gap Inverted Perovskite Solar Cells with Fill Factor Exceeding 84%. *ACS Nano* **2022**, *16* (7), 10798−10810.
(28) Han, Y.; Ali, I.; Wang, Z.; Cai, J.; Wu, S.; Tang, J.; Zhang, L.; Ren, J.; Xiao, R.; Lu, Q.; Hang, L.; Luo, H.; Li, J. Machine Learning Accelerates Quantum Mechanics Predictions of Molecular Crystals. *Phys. Rep.* **2021**, *934*, 1−71.
(29) Chen, X.; Liu, X.; Shen, X.; Zhang, Q. Applying Machine Learning to Rechargeable Batteries: From the Microscale to the Macroscale. *Angew. Chem., Int. Ed.* **2021**, *60* (46), 24354−24366.
(30) Wang, Z.; Han, Y.; Cai, J.; Chen, A.; Li, J. Vision for Energy Material Design: A Roadmap for Integrated Data-Driven Modeling. *J. Energy Chem.* **2022**, *71*, 56−62.
(31) Yao, N.; Chen, X.; Fu, Z.-H.; Zhang, Q. Applying Classical, Ab Initio, and Machine-Learning Molecular Dynamics Simulations to the Liquid Electrolyte for Rechargeable Batteries. *Chem. Rev.* **2022**, *122* (12), 10970−11021.
(32) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9* (7), 1668−1673.
(33) Knøsgaard, N. R.; Thygesen, K. S. Representing Individual Electronic States for Machine Learning GW Band Structures of 2D Materials. *Nat. Commun.* **2022**, *13* (1), 468.
(34) Wang, Z.; Wang, Q.; Han, Y.; Ma, Y.; Zhao, H.; Nowak, A.; Li, J. Deep Learning for Ultra-Fast and High Precision Screening of Energy Materials. *Energy Stor Mater.* **2021**, *39*, 45−53.
(35) Wang, Z.; Sun, Z.; Yin, H.; Liu, X.; Wang, J.; Zhao, H.; Pang, C. H.; Wu, T.; Li, S.; Yin, Z.; Yu, X.-F. Data-Driven Materials Innovation and Applications. *Adv. Mater.* **2022**, *34* (36), 2104113.
(36) Li, Z.; Yoon, J.; Zhang, R.; Rajabipour, F.; Srubar, W. V., III; Dabo, I.; Radlińska, A. Machine Learning in Concrete Science: Applications, Challenges, and Best Practices. *Npj Comput. Mater.* **2022**, *8* (1), 127.
(37) Kim, S.; Lee, M.; Hong, C.; Yoon, Y.; An, H.; Lee, D.; Jeong, W.; Yoo, D.; Kang, Y.; Youn, Y.; Han, S. A Band-Gap Database for Semiconducting Inorganic Materials Calculated with Hybrid Functional. *Sci. Data* **2020**, *7* (1), 387.
(38) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016. .
(39) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8* (1), 15679.
(40) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31* (9), 3564−3572.
(41) Lundberg, S.; Lee, S. I. A Unified Approach to Interpreting Model Predictions. In *NIPS*; 2017.
(42) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. a. The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1* (1), No. 011002.
(43) Ahmadi, M.; Ziatdinov, M.; Zhou, Y.; Lass, E. A.; Kalinin, S. V. Machine Learning for High-Throughput Experimental Exploration of Metal Halide Perovskites. *Joule* **2021**, *5* (11), 2797−2822.
(44) Wang, Z.; Cai, J.; Wang, Q.; Wu, S.; Li, J. Unsupervised Discovery of Thin-Film Photovoltaic Materials from Unlabeled Data. *Npj Comput. Mater.* **2021**, *7* (1), 128.
(45) Kim, S.; Lee, D.; Lee, J.; Cho, Y.; Kang, S.-H.; Choi, W.; Oh, J. H.; Yang, C. Diazapentalene-Containing Ultralow-Band-Gap Copolymers for High-Performance near-Infrared Organic Phototransistors. *Chem. Mater.* **2021**, *33* (18), 7499−7508.
(46) Chen, X.; Ren, F.; Gu, S.; Ye, J. Review of Gallium-Oxide-Based Solar-Blind Ultraviolet Photodetectors. *Photonics Res.* **2019**, *7* (4), 381−415.
(47) Liu, Z.; Tang, W. A Review of Ga_2O_3 Deep-Ultraviolet Metal−Semiconductor Schottky Photodiodes. *J. Phys. Appl. Phys.* **2023**, *56* (9), No. 093002.
(48) Gao, C.; Wang, Y.; Fu, S.; Xia, D.; Han, Y.; Ma, J.; Xu, H.; Li, B.; Shen, A.; Liu, Y. High-Performance Solar-Blind Ultraviolet Photodetectors Based on β-Ga_2O_3 Thin Films Grown on p-Si(111) Substrates with Improved Material Quality via an AlN Buffer Layer Introduced by Metal−Organic Chemical Vapor Deposition. *ACS Appl. Mater. Interfaces* **2023**, *15* (32), 38612−38622.
(49) Liu, X.; Wang, S.; He, L.; Jia, Y.; Lu, Q.; Chen, H.; Ma, F.; Hao, Y. Growth Characteristics and Properties of Ga_2O_3 Films Fabricated by Atomic Layer Deposition Technique. *J. Mater. Chem. C* **2022**, *10* (43), 16247−16264.
(50) Jesenovec, J.; Weber, M. H.; Pansegrau, C.; McCluskey, M. D.; Lynn, K. G.; McCloy, J. S. Gallium Vacancy Formation in Oxygen Annealed β-Ga_2O_3. *J. Appl. Phys.* **2021**, *129* (24), 245701.
(51) Zhang, J.; Dong, P.; Dang, K.; Zhang, Y.; Yan, Q.; Xiang, H.; Su, J.; Liu, Z.; Si, M.; Gao, J.; Kong, M.; Zhou, H.; Hao, Y. Ultra-Wide Bandgap Semiconductor Ga_2O_3 Power Diodes. *Nat. Commun.* **2022**, *13* (1), 3900.
(52) Chen, A.; Wang, Z.; Zhang, X.; Chen, L.; Hu, X.; Han, Y.; Cai, J.; Zhou, Z.; Li, J. Accelerated Mining of 2D van Der Waals

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX

Heterojunctions by Integrating Supervised and Unsupervised Learning. *Chem. Mater.* **2022**, *34* (12), 5571−5583.
(53) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(54) Snoek, J., Larochelle, H.; Adams, R. P.; Pereira, F. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*; 2012 Vol. *25*, 2951−2959 (Curran Associates Inc., 2012). A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning.
(55) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169−11186.
(56) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50* (24), 17953−17979.
(57) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865−3868.
(58) Heyd, J.; Scuseria, G. E. Efficient Hybrid Density Functional Calculations in Solids: Assessment of the Heyd−Scuseria−Ernzerhof Screened Coulomb Hybrid Functional. *J. Chem. Phys.* **2004**, *121* (3), 1187−1192.
(59) Wang, V.; Xu, N.; Liu, J.-C.; Tang, G.; Geng, W.-T. VASPKIT: A User-Friendly Interface Facilitating High-Throughput Computing and Analysis Using VASP Code. *Comput. Phys. Commun.* **2021**, *267*, No. 108033.

https://doi.org/10.1021/acsnano.4c12166
ACS Nano XXXX, XXX, XXX−XXX