https://doi.org/10.31449/inf.v49i19.7913 Informatica **49** (2025) 37–52 **37**

Improved GBRT Algorithm and Transfer Learning for Band Gap Regulation and Screening of Perovskite Materials

Xueshuang Deng*, Jiaojiao Chen
Electronic Information and Electrical College of Engineering, Shangluo University, Shangluo 726000, China
E-mail: slxy20251223dx@163.com
*Corresponding author

Keywords: transfer learning, GBRT, halogen perovskite, band gap modulation, screening for new materials

Received: December 27, 2024

As society continues to evolve, there is an increasing demand for energy. Solar energy is a clean and renewable alternative, but the photovoltaic conversion efficiency of calcite solar cells is low. Therefore, the study proposes a new material screening method for chalcocite using an improved gradient boosting regression tree (GRBT) model and migration learning. It adopts weighted averaging instead of the initial simple averaging to make complete use of the information between all the data, and at the same time introduces an adaptive step reduction to optimize the algorithm, which is fused with the support vector machine (SVM) algorithm is fused to construct a hybrid model, using hierarchical migration learning to divide the source domain data and train the model separately. Experiments showed that the astringent loss function values of the improved method were 0.115 and 0.160 lower than those of the GBRT algorithm and SVM, respectively. Moreover, and the root mean square errors and coefficients of determination for predicting the caliche band gap of the hybrid model were 0.017 and -2.2 and 0.023 and -3.7% lower than those of GBRT and SVM, respectively. The average pairwise decision error, root mean square error, and coefficient of determination of the improved transfer learning method were 0.0097, 0.0205, and -5.06% lower, respectively, than those of the ordinary method, and the running speed was 1.92 s faster than that of the ordinary method. The study screened out six halogenated bis-calcitonite new materials with band gap values in the range of [1.14-1.62] eV, and the formation energies were all below 0.05. It can be concluded that the improved method can effectively enhance the screening accuracy and speed of perovskite materials, and promote the high speed development of solar cells.

Povzetek: Članek se osredotoča na izboljšan algoritem Gradient Boosting Regression Tree (GBRT) in prenosno učenje za obvladovanje in iskanje novih perovskitnih materialov s prilagodljivimi širjenji pasovnih lukenj. Eksperimenti so pokazali, da izboljšani model znižuje napako napovedi in povečuje učinkovitost iskanja novih materialov, s poudarkom na halogeniranih perovskitih.

1 Introduction

With the continuous development of human society, people's demand for various energy sources is also increasing dramatically, and traditional fossil energy sources such as coal, oil, and natural gas are non-renewable resources, and their reserves are being depleted at an alarming rate [1]. According to the international energy agency (IEA), the world's fossil energy resources will last less than a century at the current rate of extraction [2]. At the same time, according to the United Nations report, there are 685 million people around the world who face the dilemma of not having electricity, and there is a short-term problem of energy. Furthermore, solar energy as an almost endless energy source, complete can become a key means to solve the human energy shortage [3]. However, ordinary solar power generation materials are usually monocrystalline or polycrystalline silicon materials, whose photoelectric conversion efficiency is about 25%, while its preparation environment requires high requirements, which cannot be promoted on a large scale [4]. As an important material for new solar cells, perovskite has attracted much attention. Its photoelectric conversion efficiency can reach 25.7%. Its preparation cost is lower than that of traditional silicon solar materials, and it is not restricted by the preparation environment [5]. However, its stability is poor, so there is a need to screen for more superior combinations of components among the perovskite fraction. The traditional experimental trial-and-error and first-principle calculations are too time-consuming and costly. Instead, computerized techniques can be used to quickly screen perovskite materials with suitable band gap and stability [6]. In an attempt to improve the model's prediction performance, Yang et al. developed a gradient boosting regression tree (GRBT) model based on the enhanced sparrow search algorithm to solve the GRBT algorithm improvement challenge. The sparrow search algorithm was improved by the model using chaotic sinusoidal mapping and Student t distribution mutation. The gradient-enhanced regression number model was then optimized using the revised algorithm to increase the model prediction accuracy (PA). In comparison to the unimproved model, experiments

showed that the improved model's correlation coefficient increased by 0.015 and its root mean square error (RMSE) decreased by 0.09 [7]. When data is few, Zhang et al. suggested a novel way to increase PA by adding missing values for gradient-enhanced regression numbers. The method generated different copies of the masked dataset according to the degree of missing data in the preprocessing stage and pre-interpolated them, and regressed the time residuals in the input stage to input the missing values into the masked dataset. According to experiments, the technique could successfully increase PA in the absence of data [8].

To address the screening problem of new perovskite materials, Diao et al. proposed a new high-throughput data calculation method in order to find out the efficient, stable and non-toxic combinations among many perovskite materials. The method introduced the data mining algorithm into the high-throughput calculation to investigate the photovoltaic conversion performance and stability of 42 materials. The experiments showed that 39 of the perovskite materials had high stability with tolerance coefficients between 0.8 and 1.1, and three ideal cell materials were chosen wrongly [9]. To increase the conversion efficiency of solar cells, Zhi et al. suggested a machine learning (ML)-based approach for perovskite molecular characterisation and conversion efficiency. The method constructed a ML model using relevant data of 19 materials. It was shown that hydrogen bond donors, hydrogen atoms, and octane-water partition coefficients were important features for the selection of perovskite materials [10]. Liu et al. proposed a new ML-based screening method in order to enhance the screening efficiency of perovskite passivation materials. The method mapped the relationship between conversion efficiency and interfacial passivation materials at the atomic level and utilized density flood theory for high-throughput prediction. The results of the experiments showed that the high-performance materials could successfully contribute to the significant passivation effect, and the method offered screening guidelines for interfacial materials at the atomic level interface [11]. Lai proposed a new high-throughput screening strategy in an attempt to enhance the flux and reduce the time cost of the new halide perovskite. To spatially encode particle size and composition, respectively, the approach combined defect-engineered anion-exchange techniques with evaporative crystallization polymer pen lithography (EC-PPL). To quickly create 3-particle libraries, it selectively changed the defect concentration of particular particles. Experiments demonstrated that this strategy was effective in faster screening of new materials [12]. Liu et al. proposed a new ML based model in an attempt to reduce the trial and error cost of double perovskite oxides. The model utilized the band gap data of 236 perovskite materials to form a dataset, and used features such as ionic radius to screen stable perovskite materials with suitable band gaps from a variety of candidate combinations. Experiments demonstrated that the model was able to effectively reduce the cost of trial and error for new materials [13].

In summary, existing studies have explored the improvement of the GRBT algorithm and the screening of new perovskite materials from various aspects, and have achieved certain results. However, the existing methods for screening halogenated chalcogenides continue to face challenges, including inadequate PA. The root-mean-square error in band gap prediction remains substantial, necessitating the reduction of this error to below 0.05 to enhance the accuracy of chalcogenide predictions. Therefore, the study proposes a perovskite screening model based on GBRT algorithm and transfer learning, which innovatively adopts weighted averaging instead of the initial simple averaging to utilize all the information between the data completely. Meanwhile, adaptive reduction step size is introduced to optimize the algorithm, which is fused with support vector machine (SVM) algorithm to construct the hybrid model. Hierarchical transfer learning is used to divide the source domain data (SDA) and train the model separately. The goal of study is to improve the screening accuracy and speed of perovskite new materials, promote the rapid development of solar power generation, and reduce carbon emissions.

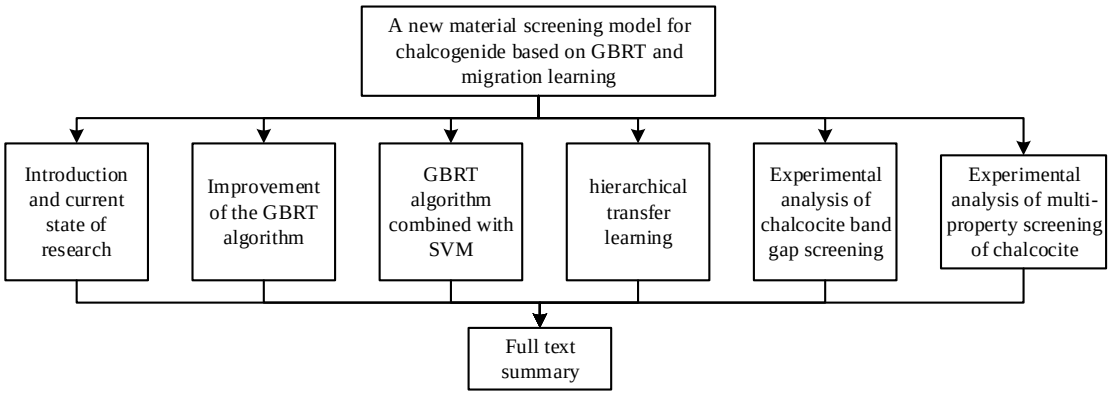
Based on the above related studies, Table 1 summarizes the research methodology, RMSE, calculation time, and shortcomings of the related studies.

Table 1: Summary of relevant information of relevant studies

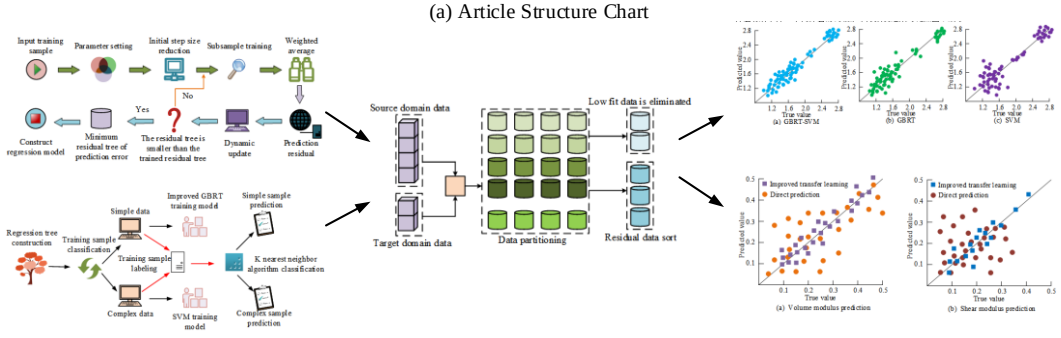
Author	Research methods	RMSE	Calculation time (ms)	Insufficient
Literature [7]	Improvement of GBRT based on sparrow search algorithm	0.072	97.5	Reduced computation speed when data volume is too large
Literature [8]	Pre-interpolation GBRT	0.066	86.2	Higher requirements for pre-interpolation
Literature [9]	High-throughput data computation	0.094	247.5	Higher hardware requirements
Literature [10]	Machine learning models	0.107	89.4	Requires more precise parameterization
Literature [11]	Machine learning and density functional theory	0.065	155.2	Requires high quality data for training
Literature [12]	EC-PPL and defect-engineered anion exchange techniques	0.058	586.4	Higher equipment requirements and more difficult to operate

Literature [13]	Machine learning models	0.079	58.3	Insufficient generalization ability when facing new materials
This text	GBRT and transfer learning	0.032	62.5	/

As illustrated in Table 1, while extant studies have attained specific outcomes in the context of chalcogenide material screening, their RMSE and computational efficiency performances remain unsatisfactory. Moreover, the experimental requirements are more stringent. Therefore, this study innovatively proposes a chalcogenide material screening method based on GBRT and migration learning, which effectively improves the screening speed and accuracy of new chalcogenide materials. To realize the screening and band gap regulation of new chalcogenide materials, the study proposes a chalcogenide screening model based on GBRT and migration learning. The general structure of which is shown in Figure 1.



(a) Article Structure Chart



(b) Article Flowchart

Figure 1: Overall architecture and flowchart of the article

In Figure 1(a), the study first explains the relevant research background, the current status of domestic and international research, and the innovation of the manuscript. The first chapter focuses on how to improve the GBRT algorithm and use it for chalcocite band gap screening. The second chapter focuses on the hierarchical migration learning of the model and the screening of chalcocite materials for multi-properties. The third chapter carries out a comprehensive test on the above methods. The fourth chapter carries out a full text summary as well as the future prospects. In Figure 1(b), the GBRT algorithm is selected as the fundamental prediction model, with the KNN algorithm employed to enhance it. The weighted average of the KNN classification algorithm is capable of emphasizing the correlation between variables, thereby enhancing the model's PA. Concurrently, the adaptive reduced step size optimization algorithm is implemented, which can employ a larger learning rate to expedite the identification of the optimal solution in the initial training phase and progressively diminish the learning rate in the subsequent phase to circumvent oscillatory behavior near the optimal solution or the occurrence of overfitting. The improved GBRT model is combined with the SVM model to improve the complex data processing capability and robustness of the hybrid model, and to achieve stable prediction results. The target domain data is input into the hierarchical migration model using a sliding window for training to obtain the initial training model, and the hierarchical migration can effectively reduce the computation time. The optimized GBRT-SVM hierarchical migration learning model first uses the low-level features of the data obtained by migration learning as input to the GBRT, then initializes the SVM with the final weights obtained by migration learning, and finally uses the features generated by the GBRT as input to the SVM for classification. Finally, the two models are used to screen the chalcogenide materials separately to obtain the best screening results.

2 Methods and materials

2.1 Improved GBRT algorithm for screening hybrid perovskite band gap

In solar cells, the band gap of perovskite materials is one of the direct influences on the photovoltaic conversion efficiency. The width of the band gap determines the wavelength range of light absorbed by perovskite materials, which in turn affects the photovoltaic conversion [14]. A hybrid perovskite is a heterogeneous material that combines organic and inorganic substances, which reacts when the two substances are interspersed, resulting in its unique properties. The general formula is ABX_3. The A site (A-S) is usually a cation with a large radius, such as Cs^+, Rb^+, and MA^+. The B site (B-S) is usually a relatively small radius cation, such as Pb^{2+}, Sn^{2+}, and Ge^{2+}. The X site (X-S) is for anions, commonly halogenated elements such as I., Cl., and Br.. Solar cells made of hybrid perovskite material have the advantages of simple process and low cost. Their photoelectric conversion efficiency is generally more than 20%, but there is still room for further improvement. The material components can be continuously adjusted to change its band gap width to enhance the photoelectric conversion efficiency [15]. Therefore, corresponding ML techniques are needed to screen perovskite materials with suitable band gaps, and the GBRT algorithm is selected for the study. Because GBRT can achieve high PA by integrating multiple weak learners, and when combined with SVM models, it can handle nonlinear relationships and complex data structures, making it suitable for scenarios with more features and complex relationships. Lower PA results from the platform GBRT algorithm's overly simplistic prediction function on the leaf node, which is also susceptible to changes in data quality. Therefore, by replacing the original simple average with the weighted average of the K-nearest neighbor (KNN) classification algorithm, the study enhances the technique and fully utilizes the information between all the data. The KNN algorithm first needs to calculate the distance from the training sample to the predicted sample in each node. A sample is selected which is closer to the predicted sample and the output variable of the sample is shown in Equation (1) [16].

$$\left(f_1, f_2, ..., f_{x-1}, f_x \right) \tag{1}$$

In Equation (1), f_x denotes the output variable of the x th training sample. The weights calculated for each sample weighting are shown in Equation (2).

$$w_i = \left(e^{\frac{d_i^2}{2}} \right) \Big/ e^{-\frac{d_1^2}{2}} + e^{-\frac{d_2^2}{2}} + ... + e^{-\frac{d_i^2}{2}} \tag{2}$$

In Equation (2), w_i is the weight of the training sample. d_i is the distance from the training sample to the predicted sample. The weighted average is calculated as shown in Equation (3).

$$W = \sum_{i=1}^{x} w_i f_i \tag{3}$$

In Equation (3), W denotes the weighted average of the samples. The study choose Equations (1)-(3) to compute the weighted average of the KNN algorithm instead of the simple average, which improves the utilization of data and is suitable for initial data processing. The K value of the KNN algorithm by plotting the classification error rate curve corresponding to different K values, the error rate in a certain range of K values will first decrease and then increase, when the lowest point corresponding to the K value is the best choice. The study uses the weighted average of the KNN classification algorithm to be able to focus on the correlation between variables and effectively enhance the PA of the model. The reduction step size of GBRT algorithm, i.e., the learning rate, is usually kept constant, which increases the risk of overfitting of the algorithm, as well as reduces the training efficiency and so on. The study introduces an adaptive shrinkage step to optimize the algorithm. Firstly, Equation (4) illustrates how the algorithm's loss function (LF) is defined [17].

$$L = \sum_{i=1}^{n} \frac{1}{N} \left[f_i - \left(H_j(x_i) + \lambda h_{j+1}(x_i) \right) \right]^2 \tag{4}$$

In Equation (4), L denotes the LF of the algorithm. N denotes the number of random subsamples. $H_j(x_i)$ denotes the strong learner consisting of the first j residual trees. λ denotes the reduced step size. $h_{j+1}(x_i)$ denotes the weak learner of the $j+1$ th residual tree. When the strong and weak learners are deterministic values, the LF is used to derive the reduced step size, as shown in Equation (5).

$$\frac{\partial L}{\partial \lambda} = -\frac{2}{N} \sum_{i=1}^{n} \left[f_i - \left(H_j(x_i) + \lambda h_{j+1}(x_i) \right) \right] \cdot h_{j+1}(x_i) \tag{5}$$

In Equation (5), the value of the adaptive reduction step can be computed using Equation (6) when the derivative equals zero.

$$\lambda = \frac{\sum_{i=1}^{n} \left[f_i - H_j(x_i) \right] \cdot h_{j+1}(x_i)}{\sum_{i=1}^{n} h_{j+1}^2(x_i)} \tag{6}$$

Adaptive scaling of the step size allows a larger learning rate to be used in the early stages of training to quickly solve the optimal solution, and a progressively smaller learning rate in the later stages to avoid oscillation around the optimal solution or overfitting. The study

selects Equations (4)-(6) to improve the GBRT algorithm with adaptive reduced step size, which can improve the training efficiency and reduce the risk of overfitting of the algorithm, and is applicable to the training process of the GBRT algorithm. The specific operation flow of GBRT algorithm improved with KNN is shown in Figure 2.

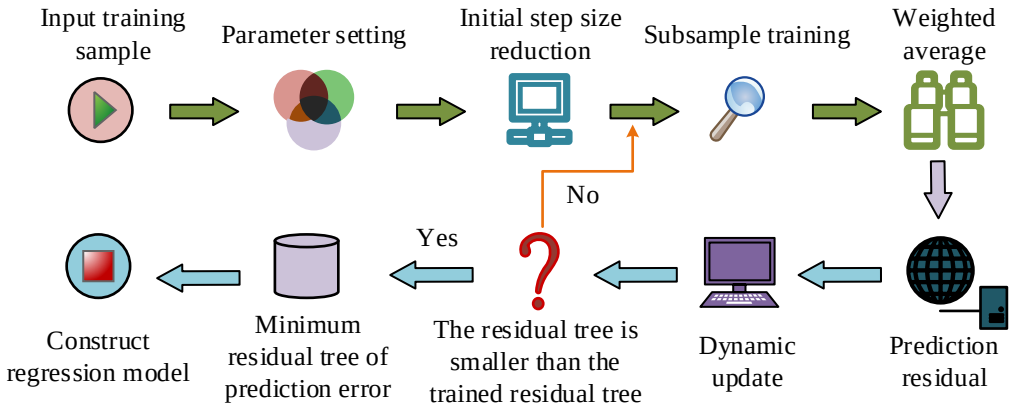


Figure 2: Specific operation flow of improved GBRT algorithm

In Figure 2, the training samples are entered first, the parameters such as the number of training times and the random sampling rate of the residual tree are set, and the initial reduction step is set to 0.01. The smaller reduction step modulates the model by only a small amount each time, which can effectively prevent the model from overfitting during training. Relevant studies have shown that although a larger reduction step can converge quickly, it is only applicable to simple data sets. Moreover, in the screening of chalcocite materials, the accuracy of the model with a reduction step of 0.01 is usually greater than that with a reduction step of 0.1, and the samples are initialized after the reduction step setting is completed. A sub-sample is randomly selected and trained on the residual tree, and a weighted average is used to find the prediction function. The output variable values are updated using the predicted residuals, and the size of the reduction step is dynamically updated. When the number of remaining residual trees is smaller than the number of trained residual trees, the residual tree with the smallest prediction error is extracted. This residual tree and all previous residual trees are used to form a regression model. To address the modeling problem when the relevant sample data is insufficient and to enhance the model PA, the research adopts the fusion of SVM algorithm and GBRT algorithm to construct the hybrid model. This approach effectively improves the complex data processing ability and robustness of the hybrid model and provides stable prediction results through the stronger generalization ability and high-dimensional data processing ability of SVM. The SVM algorithm, through the nonlinear mapping function, maps the low-dimensional. The SVM algorithm maps the low-dimensional training sample data into the high-dimensional space by means of a nonlinear mapping function, and performs linear regression of the data in the high-dimensional space, and the regression constructor of the SVM is shown in equation (7).

$$f(x)=\omega\varphi(x)+b \tag{7}$$

In Equation (7), $f(x)$ denotes the regression constructor. ω is the weight vector. $\varphi(x)$ denotes the NMF. S denotes the bias term. The weight vector and bias term are solved as shown in Equation (8).

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+C\sum_{i+1}^{n}(\zeta+\zeta^*) \tag{8}$$

In Equation (8), C denotes the regularization constant. Both ζ and ζ^* denote slack variables. Equation (9) is used to determine the solution high spatial data.

$$f(X)=\sum(q_i-q_i^*)\kappa(X_i,X_j)+b \tag{9}$$

In Equation (9), both q_i and q_i^* denote Lagrange multipliers. $\kappa(X_i,X_j)$ is the kernel function (KF). The selection of the kernel function is contingent upon the particular circumstances of each case. This study posits that the radial basis kernel function requires a smaller number of parameters and is adept at effectively addressing nonlinear, differentiable problems by mapping the data points of the input feature. Space into an infinite-dimensional feature space. At the same time, the kernel function has a wider scope of applicability, and it can make the SVM model have a wide range of accuracy, so it is selected as the kernel function of SVM. Which is calculated as shown in Equation (10).

$$\kappa(x_i,x_j)=\exp\left(-\sigma\|x_i-x_j\|^2\right) \tag{10}$$

In Equation (10), $\kappa(x_i,x_j)$ denotes the radial basis KF. σ denotes the width of the radial basis kernel.

$\|x_i - x_j\|$ denotes the Euclidean distance between two points. The study selects Equations (7)-(10) for SVM and GBRT fusion, which can improve the model's high-dimensional data processing ability and is suitable for the analysis of complex data. Figure 3 depicts the hybrid model's operational flow.

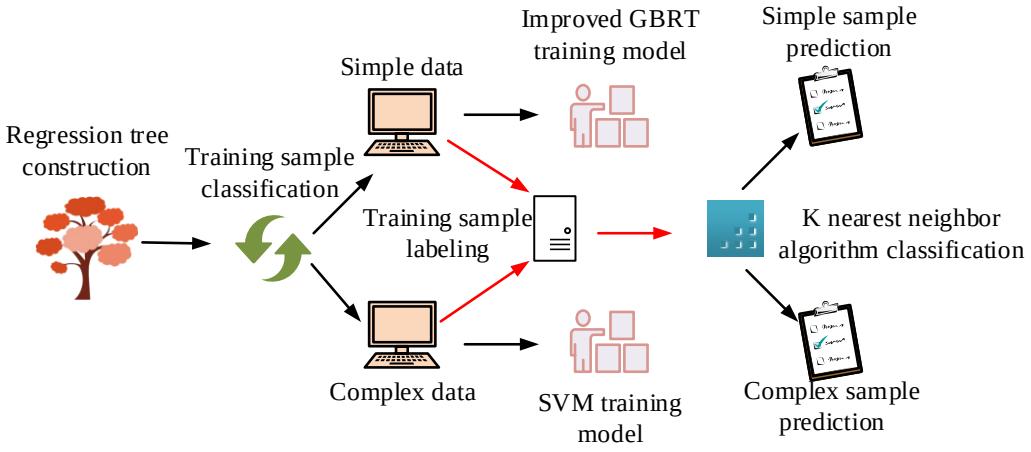


Figure 3: Operation flow of the hybrid model

In Figure 3, after constructing the regression tree using the GBRT algorithm, the training samples are divided into simple data and complex data based on the prediction deviation of the leaf nodes. The first 1/2 leaf nodes with large prediction deviation are categorized as complex data, and the remaining 1/2 leaf nodes are categorized as simple data. In the study, the simple data is input to the GBRT model for training, and the complex data is input to the SVM model for training. After labeling the two training data with categories, the test samples are classified using the classification algorithm based on K-nearest neighbor algorithm, which classifies the test samples into simple and complex data, and then the two models are used to predict the test samples after the training is completed.

2.2 Multi-property screening of halogen double perovskite based on hierarchical transfer learning

Despite the advantages of low manufacturing cost and high photovoltaic conversion efficiency, the presence of heavy metals such as lead in their composition renders heterogeneous chalcogenide materials environmentally and biotoxic. Leakage of these materials can result in soil and water contamination, thereby affecting the ecosystem and human health. Adverse effects may include neurological damage, reproductive toxicity, and hematologic disorders. Therefore, there is a need to screen non-toxic and environmentally friendly chalcogenide materials [18]. Halogen double perovskite is an effective alternative to lead-based perovskite materials due to its lead-free nature and high stability and dimmability. The general formula for the structure of halogen double perovskite is $A_2BB'X_6$. Among them, the A-S is an inorganic cation. The B-S is a monovalent metal ion (MI), the B' site is a trivalent MI, and the X-S is a halogen ion (HI). The B' site is a trivalent MI and the X-S is a HI. The X-S is a HI. The octahedral factor of $A_2BB'X_6$ is calculated as shown in Equation (11).

$$\mu = \frac{r_B}{r_X} \tag{11}$$

In Equation (11), μ denotes octahedral factor. r_B denotes radius of B-S ion. r_X is the radius of the X-S HI. The calculation of tolerance factor is shown in Equation (12).

$$TF = \frac{\left(r_A + r_X\right)}{\sqrt{2}\left(r_B + r_X\right)} \tag{12}$$

In Equation (12), TF denotes the tolerance factor. r_A denotes the radius of the A-S ion. Equations (11) and (12) are chosen for the study to evaluate the stability in chalcogenide structures. However, the available data samples of $A_2BB'X_6$ materials are small, and most of the ML methods are unable to obtain the prediction results with high accuracy in the limited data. Therefore, the study uses transfer learning algorithm for $A_2BB'X_6$ material screening. Transfer learning can improve the model's generalization capacity, lessen reliance on a lot of labeled data, and adapt to the target domain by using the knowledge of the source domain. Figure 4 illustrates the fundamental architecture of transfer learning in artificial neural networks.

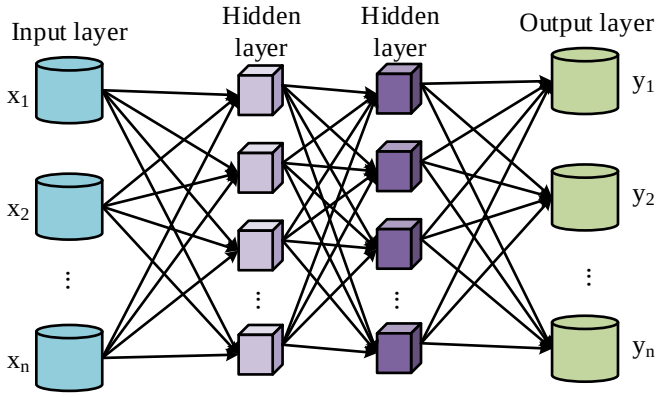


Figure 4: Basic structure of transfer learning neural network

In Figure 4, the neural network is divided into five layers in total, which are the input layer, two hidden layers, one output layer, and one error calculation layer. Equation (11) is used to connect the data in the input layer amongst all of the neurons in each layer, and the two hidden layers are used to extract the input data's features layer by layer. Then, the nonlinear inertia in the data is learned and expressed by the activation function. Moreover, the study selects the ReLU activation function that possesses computational simplicity, can avoid gradient vanishing, as well as performs well in deep networks. Finally, it converts the data into predictions through the output layer. However, the ordinary transfer learning algorithm needs to be carried out step by step many times, which takes a long time. Therefore, the algorithm is improved hierarchically. The hierarchical transfer learning firstly needs to divide the SDA, calculate the fit of the SDA and sort it according to the size. The target domain data (TDA) is input into the GBRT-SVM model using a sliding window for training, and the initial training model is obtained. The SDA partitioning process for hierarchical transfer learning is shown in Figure 5.

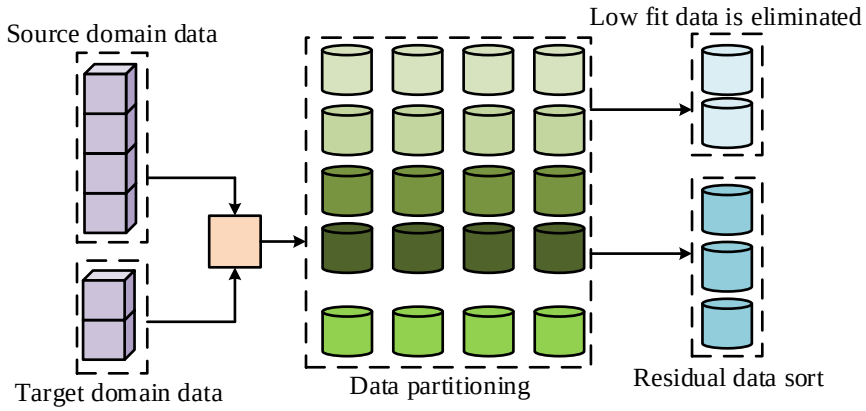


Figure 5: Source domain data partitioning process of hierarchical transfer learning

In Figure 5, the study also used a sliding window to input the SDA into the model. Among them, a single prediction is available in each window, and R^2, the goodness of fit, is used to indicate the degree of correctness of the prediction results. The study sets the fit threshold at 0.7, removes data with a fit less than the threshold, and uses the size of the R^2 to rank the data, with the higher the fit the higher the ranking. The study is based on the number of layers of the hierarchical migration model. The SDA that meets the fit requirements is divided into a total of K groups of data, the K1 group of data with a lower fit is used to train the K1 layer of the model. The weights of the training results are recorded, the rest of the layer is frozen, and the K2 group of data is used to train the K2 layer of the model. Moreover, the K2 layer of the training is used to load the previous layer of the weights and to determine the loss of prediction results in the target dataset in whether or not to decline. If it decreases, unfreeze the current layer and all previous layers, and vice versa, continue training until all layers of the model are all unfrozen. The prediction effect of each layer is judged and the weight of the layer with the largest increase in prediction effect is recorded. The operation flow of hierarchical transfer learning is shown in Figure 6.

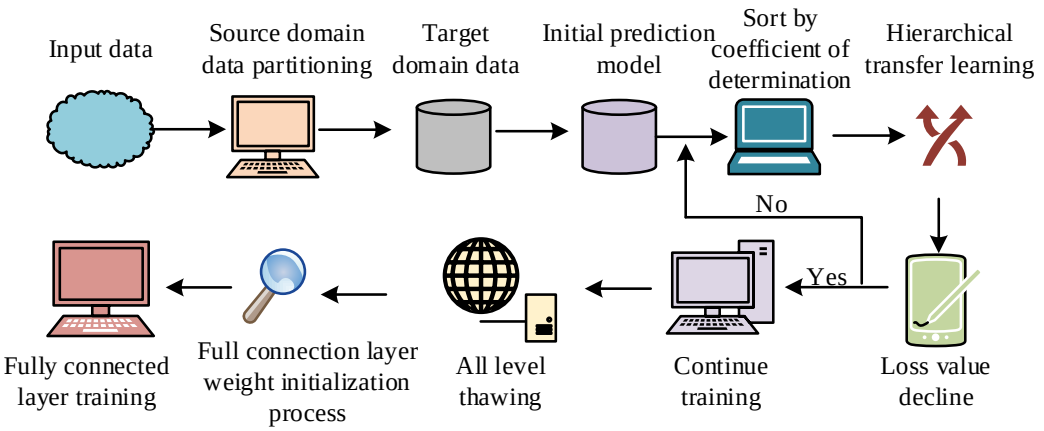


Figure 6: Operation flow of hierarchical transfer learning

In Figure 6, the input data are SDA, TDA, and GBRT-SVM hybrid model. The study first divides the SDA, trains the TDA in the model to get the initial prediction model, and sorts the SDA according to the coefficient of determination and divides them into K groups. The study uses the data of the division number for hierarchical transfer learning, using the loss value as a judgment criterion for the prediction effect. The training is continued if the loss value decreases until all layers of the model are fully unfrozen. After the model transfer learning is completed, the study uses the TDA to adjust the model parameters, initialize the final fully connected layer weights, and set a smaller learning rate to train only the fully connected layer. The optimized GBRT-SVM model with hierarchical migration learning first uses the low-level features of the data obtained from migration learning as inputs to the GBRT, then initializes the SVM with the final weights obtained from migration learning, and finally uses the feature inputs generated by the GBRT in the SVM for classification. The study, in an effort to deduce the complexity of the model and the occurrence of overfitting situations, uses a feature subset-based approach for feature dimensionality reduction after comprehensive consideration. The method utilizes the Pearson coefficient between different feature subsets to reduce the linear feature correlation. The Pearson coefficient is calculated as shown in Equation (13) [19].

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}} \quad (13)$$

In Equation (13), r denotes the Pearson coefficient. X_i is the i th observation of variable (OV) X. $\bar{X}$ is the mean of variable (MV) X. Y_i is the i th OV Y. $\bar{Y}$ is the MV Y. n is the total quantity of observations. The Pearson's coefficient takes values between [-1, 1], when it takes a value of 1, it indicates a fully positive linear correlation. When it takes a value of -1, it indicates a fully negative linear correlation, and when it takes a value of 0, it indicates no linear relationship. The study calculates the Pearson coefficient of each characteristic with respect to the target variable and retains the characteristics with high linear correlation. The study retains features with Pearson coefficients greater than 0.5 as a subset of features, and feature dimensionality can be reduced. The data are normalized to ensure comparability between data magnitudes. The study uses the normalization method to convert the initial data to standard data, as calculated in Equation (14) [20].

$$X_{std} = \frac{X - \mu}{\gamma} \quad (14)$$

In Equation (14), X_{std} denotes the normalized feature data (FD). X denotes the initial data. μ the mean value of the FD. γ is the standard deviation of the FD. After the screening of the materials is completed, the study employs the formation energy (FE) to validate the materials with first-principle calculations. The relevant parameters of the materials are the same as the model in the training set. The FE of a perovskite compound is the energy released or absorbed by an atom to form a compound from a free state. The FE of a compound is calculated as shown in Equation (15).

$$E_b = \frac{E\left(A_m B_n C_k\right) - m \cdot E(A) - n \cdot E(B) - k \cdot E(C)}{m + n + k} \quad (15)$$

In Equation (15), E_b is the FE of the compound. $E\left(A_m B_n C_k\right)$ denotes the total energy of the compound. $E(A)$, $E(B)$, and $E(C)$ denotes the energy of the free atom A, B, and C, respectively. m, n, and k denotes the number of atoms A, B, and C, respectively. The study selects Equations (13)-(15) for data feature reduction and normalization, which can reduce the complexity of the model, ensure the comparability between data magnitudes, and apply to data processing of different dimensions. The pseudo-code for the improved GBRT-SVM hybrid model is shown in Figure 7.

```
# Import necessary libraries
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, accuracy_score

# Assume the dataset is already prepared
# X: Feature matrix
# y: Target variable
# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 1: Feature extraction using GBRT
gbrt = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
gbrt.fit(X_train, y_train)

# Extract features from the training and testing sets using the GBRT model
X_train_features = gbrt.apply(X_train).reshape(X_train.shape[0], -1)
X_test_features = gbrt.apply(X_test).reshape(X_test.shape[0], -1)

# Step 2: Classification or regression using SVM
svm = SVC(kernel='rbf', gamma='scale', C=1.0)  # For classification tasks
# For regression tasks, use SVR instead
# svm = SVR(kernel='rbf', gamma='scale', C=1.0)

# Train the SVM model
svm.fit(X_train_features, y_train)

# Step 3: Evaluate model performance
# For classification tasks
y_pred = svm.predict(X_test_features)
accuracy = accuracy_score(y_test, y_pred)
print(f"Test set accuracy: {accuracy:.4f}")

# For regression tasks
# y_pred = svm.predict(X_test_features)
# rmse = mean_squared_error(y_test, y_pred, squared=False)
# print(f"Test set RMSE: {rmse:.4f}")
```

Figure 7: Pseudo-code for improved GBRT-SVM hybrid modeling

3 Results

3.1 Experimental analysis of screening of hybrid perovskite band gap

In order to verify the prediction effect of the GBRT-SVM hybrid model, the study adopts the Perovskite Database, which contains more than 16,000 chalcocite-related thesis data, covering more than 42,400 chalcocite materials with detailed information, which is of high quality and good reliability, and the database can effectively simplify the process of literature search and data analysis to improve the experimental efficiency. The database can effectively simplify the process of literature search and data analysis, and improve the experimental efficiency. Table 2 Advantages of the two datasets used in the experiment over other datasets.

Table 2: Advantages of the two datasets used in the experiment over other datasets

Characterization	Perovskite Database	Materials Project
Scale	Covering more than 42,400 chalcogenide materials	Over 1 million inorganic materials
Type	To include the structure, properties and other data of chalcogenide materials	Provide crystal structure, energy properties, electronic structure, thermodynamic properties, etc.
Update Frequency	Continuously updated	The data is continuously updated to keep it up-to-date
Openness	Open Data	Open Data
Collaboration	Data sharing	Encourage user collaboration and cooperative research
User Interface	Interactive graphical interface, easy to use	User-friendly interface, easy to search and browse

In Table 2, the two datasets selected for the experiment possess greater advantages in terms of data size, data type, and openness, and the data are well-targeted and easy to find. The filtered, dimensionality

reduced, and normalized data are divided into training set, validation set, and test set in 8:1:1 manner. The penalty parameter of the SVM model is set to 0.1 and the kernel function parameter is set to 0.001. The comparison algorithms used in the study include SVM and GBRT algorithms. Figure 8 compares the variations in loss values of several algorithms.

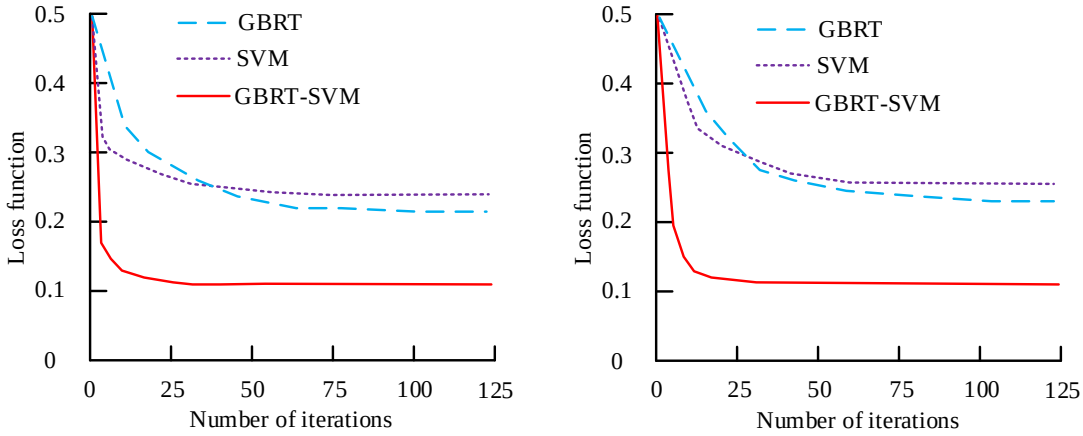


Figure 8: Comparison of loss value changes of different algorithms

In Figure 8(a), the LF of the GBRT-SVM hybrid model decreases faster in the training set and is able to obtain the astringent LF value in 30 iterations. The minimum value is 0.108, which is 0.115 and 0.160 lower than the GBRT and SVM algorithms, respectively. The optimization speed is faster than the other two algorithms by 35 and 20 iterations, respectively. In Figure 8(b), the astringent loss function value of all three algorithms increases when faced with the test set data. However, the GBRT-SVM hybrid model has the smallest increase in loss value, with an astringent loss function value of 0.125, which is lower than the other two algorithms by 0.139 and 0.173, and its increase is lower than the other two algorithms by 0.024 and 0.013, respectively. The optimization speed of the hybrid model remains basically unchanged, while the other algorithms have decreased. Comparison of perovskite band gap prediction results of different algorithms is shown in Figure 9.

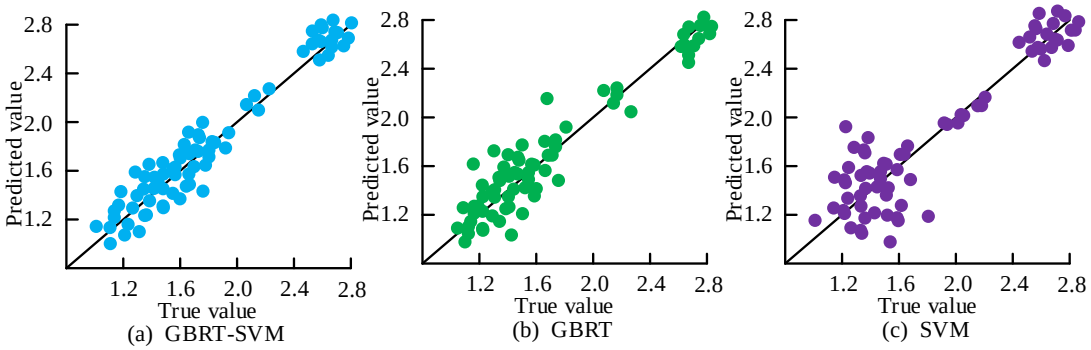


Figure 9: Comparison of prediction results of perovskite band gap by different algorithms

In Figure 9(a), the RMSE and coefficient of determination of the prediction of chalcocite band gaps of the GBRT-SVM hybrid model are 0.032 and 99.4%, respectively. Furthermore, the predicted values all converge to the vicinity of the diagonal of the horizontal and vertical coordinates, and there are no prediction points with large deviations. The more the model's predicted values dissolve the diagonal line, the closer the predicted values are to the true values and the higher the model's PA. In Figure 9(b), there are some points with large prediction deviations between the true band gap values [1.2, 1.8], and the RMSEand coefficient of determination of the GBRT model are 0.049 and 97.2%, respectively. In Figure 9(c), there are points that deviate from the convergence diagonal for band gap values below 2.0, and the RMSE and coefficient of determination of the SVM model are 0.055 and 95.7%, respectively. UV-Vis diffuse reflectance spectroscopy is used to determine the band gap of chalcogenide materials using a UV-Vis spectrophotometer and an integrating sphere attachment. A chalcogenide sample is first prepared as a thin film and placed in an integrating sphere where the reflectance spectra in the UV-Vis region are recorded and then converted to absorption spectra by the Kubelka-Munk function to measure the band gap value of the sample. Table 3 displays the hybrid model's band gap prediction results for various hybrid perovskite samples.

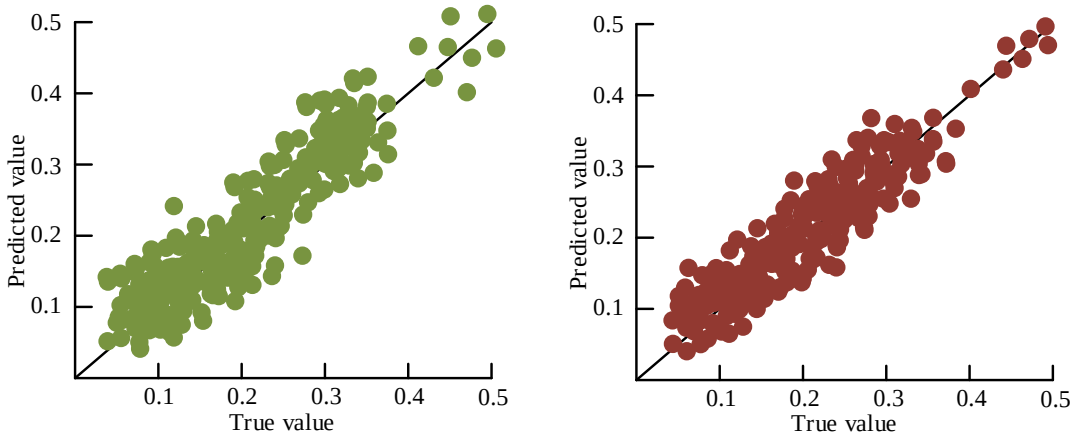
Table 3: Band gap prediction results of mixed models for different hybrid perovskite samples

Serial number	Pbe_band gap	Ml_band gap (eV)	Prediction error
1	1.534	1.520	0.014
2	0.015	0.004	0.011
3	0.000	0.013	-0.013
4	2.745	2.713	0.032
5	1.036	1.107	-0.071
6	0.000	0.005	-0.005
7	1.527	1.535	-0.008

In Table 3, the band gap prediction values of GBRT-SVM hybrid model for hybrid perovskite are all between [0.004, 2.713]. The highest and lowest value of the prediction error is 0.071 and 0.005, which both satisfy the requirement that the error range is less than 0.1 eV. Among them, there are only two that satisfy the band gap requirement of solar cell materials, which are serial numbers 1 and 7, with BGVs in the range of 1.1eV-1.7eV. The prediction errors of the hybrid model for sample #1 are 0.031 and 0.28 lower than those of the GBRT and SVM models, respectively, and 0.004 and -0.0042 lower for sample #2.

3.2 Experimental analysis of multi-property screening of halogen double perovskite

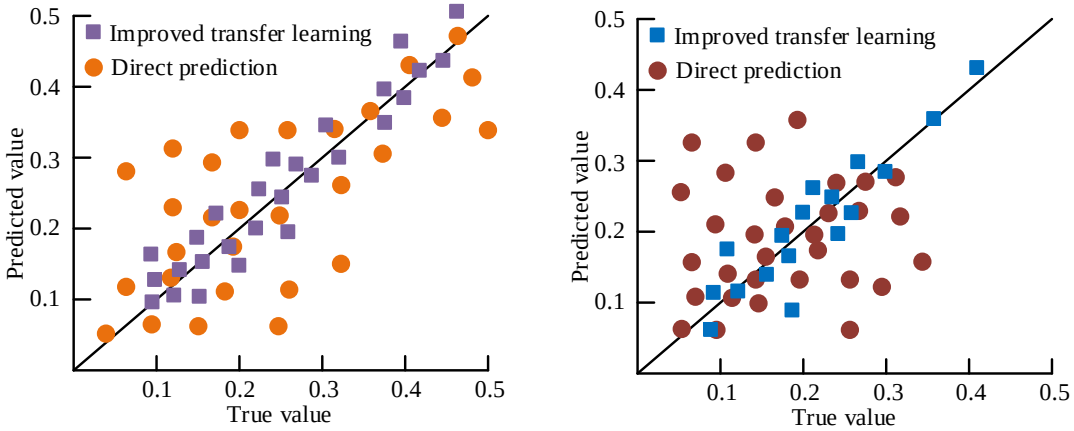
The inorganic halogen double perovskite dataset used for the experiments is Materials Project. Other parameters and treatments are the same as in Section 2.1, and the transfer learning model is used to compare the results of FE, bulk modulus, and shear modulus predictions of perovskite materials, respectively. Figure 10 compares the FE prediction outcomes of perovskite materials using various techniques.



(a) Transfer learning (b) Improved transfer learning

Figure 10: Comparison of prediction results of perovskite material formation by different methods

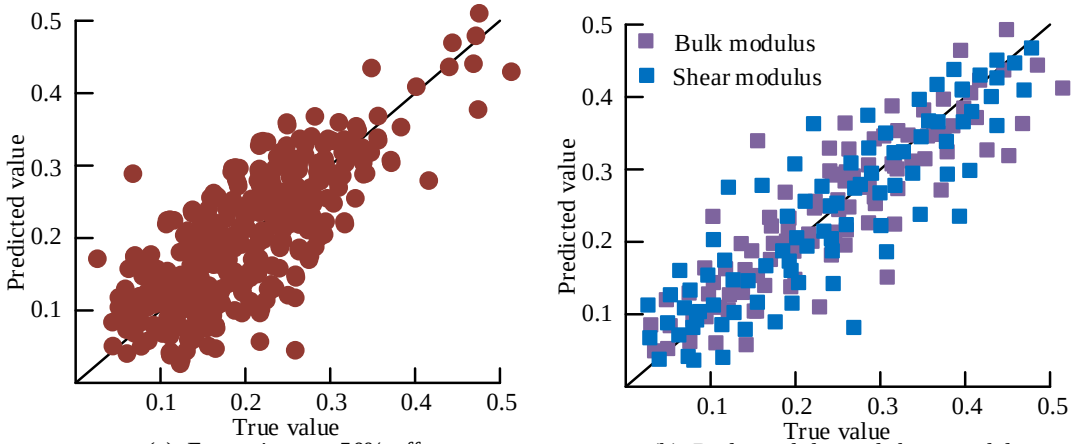
In Figure 10(a), the ordinary transfer learning method can achieve a certain PA for the FE prediction of perovskite. Its average decision error, RMSE, and coefficient of determination are 0.0472, 0.0609, and 94.35%, respectively. However, transfer learning takes a long time, running once for 2.89 s. In Figure 10(b), the average decision error, RMSE, and coefficient of determination of the improved transfer learning method are lower than that of the ordinary method by 0.0097, 0.0105, and -5.06%, respectively, and the running speed is faster than that of the ordinary method by 1.92 s. Figure 11 compares the modulus and shear modulus prediction findings of various techniques for perovskite materials.



(a) Volume modulus prediction (b) Shear modulus prediction

Figure 11: Comparison of prediction results of bulk modulus and shear modulus of perovskite materials by different methods

In Figure 11(a), the halogen double perovskite material has less relevant data and the feature dimension reaches 132 dimensions, which leads to lower PA using direct prediction methods. The RMSE and coefficient of determination are 0.092 and 87.25%, respectively. The RMSE of the bulk modulus measured using the improved transfer learning method is 0.037 lower and the coefficient of determination is 11.18% higher than the direct prediction. The prediction results of ordinary transfer learning methods are more scattered and further away from the diagonal, indicating that the predicted values are less consistent with the true values and their PA is lower. In Figure 11(b), the RMSE and coefficient of determination of shear modulus prediction by the improved transfer learning method are 0.057 and 98.65%, which are 0.042 and -12.36% lower than the direct prediction, respectively. To quantify the uncertainty of the model, the study uses 5-fold cross-validation for model evaluation. Among them, the experimental data set is randomly divided into five mutually exclusive subsets of equal size, each of which is an independent test set. The results of the 5-fold cross-validation of the formation energy and modulus of the material are shown in Figure 12.



(a) Formation can 50% off cross validation results (b) Body modulus and shear modulus 5 fold cross validation results

Figure 12: Formation energy and modulus of materials 5-fold cross validation results

In Figure 12(a), the RMSE of the improved transfer learning method after 5-fold cross-validation is 0.0417. It is 0.0013 higher than the validation result of the independent test set, but still less than 0.05, which meets the relevant requirements. In Figure 12(b), the RMSE of body modulus after 5-fold cross-validation is 0.057, and the RMSE of shear modulus is 0.059, which is higher than that of the independent test set. However, both are within the acceptable range, indicating that the model is more reliable. A comparison of the importance of component features on perovskite materials modulus and shear modulus is shown in Figure 13.

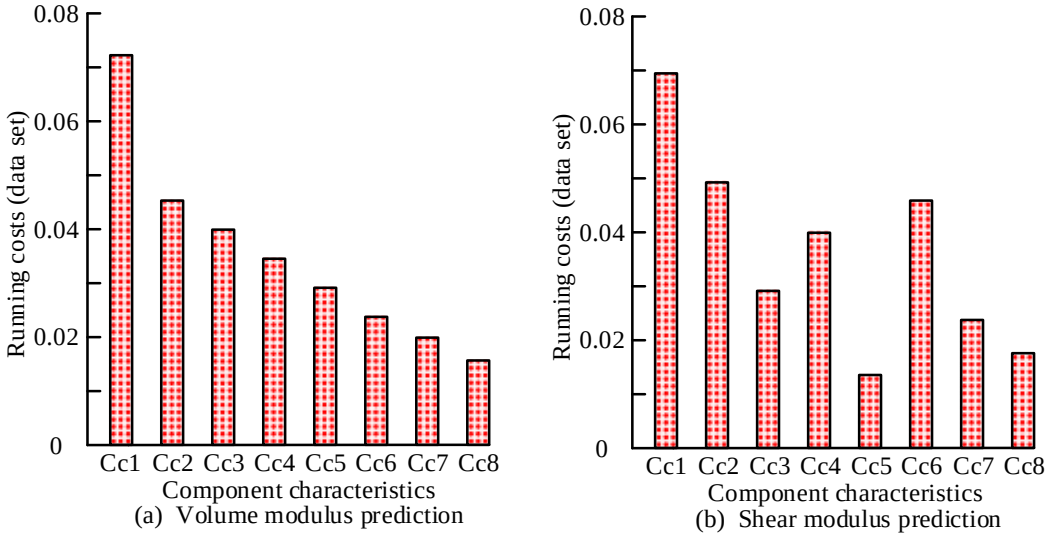


Figure 13: Comparison of the importance of component characteristics to the bulk modulus and shear modulus of perovskite materials

In Figure 13(a), the factors affecting the material's importance from Cc1 to Cc8 are Mendeleev number, electronegativity, ionic radius, octahedral factor, p-orbital valence electrons, total valence electrons, tolerance factor, and s-orbital valence electrons, respectively. Among them. The greatest degree of influence on the modulus of perovskite materials is the Mendeleev number. The Mendeleev number is most important because of its ability to quickly identify combinations of materials with characteristic properties, and electronegativity because of its role in the formation and stability of chemical bonds in materials. In Figure 13(b), the degree of influence on the shear modulus of the materials above 0.04 includes Mendeleev number, electronegativity, octahedral factor, total valence electrons. The study screens new chalcogenide materials with compliant elemental compositions based on the thermal stability, band gap value, and B/G ratio of the materials. The screened new halogenated bis-chalcogenide materials that meet the relevant conditions are shown in Table 4.

Table 4: Comparison of properties of halogen double perovskite new materials

Element combination	Band gap value	Formation energy	Bulk modulus	Shear modulus	B/G
Cs_2AuNiF_6	1.52	0.042	52.64	6.87	7.66
K_2CuInF_6	1.14	0.018	66.51	9.95	6.68
Cs_2AuOsF_6	1.55	0.049	60.75	11.09	5.48
$K_2InMoCl_6$	1.25	0.037	27.95	4.87	5.74
Cs_2CuRhF_6	1.62	0.054	79.27	15.93	4.98
K_2InCrF_6	1.44	0.012	70.25	14.98	4.69
Rb_2InRuF_6	1.21	0.009	61.54	16.89	3.64

In Table 4, the BGVs of the seven halogen double perovskite new materials are between [1.14-1.62] eV, while the optimum photoelectric conversion efficiency of perovskite solar cells is around 1.4 eV, and all of them can be used for solar cells at 1.1-1.7 eV. Formation energies below 0.05 all meet the stability requirements of the material. Except for Cs_2CuRhF_6, the other six materials meet the stability requirements. The ratios of shear modulus to bulk modulus of the new materials are all greater than 1.75, indicating that the new materials have high toughness and can be obtained with good ductility at room temperature. The comparison results of ablation experiments with different adaptive algorithms and improved migration learning methods are shown in Table 5.

Table 5: Comparison results of ablation experiments with different adaptive algorithms and improved migration learning methods

Algorithms	Ablation Module	RMSE	Coefficient of determination (%)	Predictive accuracy (%)
Adaptive GBRT-SVM	/	0.032	99.4	85.2
Adaptive SVM	/	0.057	94.8	73.5
Adaptive RF	/	0.074	90.3	69.8
Improved migration learning	/	0.040	99.4	84.3
	Normalization	0.073	92.6	79.2
	Feature dimensionality reduction	0.056	95.7	80.6

In Table 5, the adaptive GBRT-SVM achieves the optimal values for all metrics, and its RMSE is lower than that of the adaptive SVM and adaptive random forest (RF) by 0.025 and 0.042, respectively. Moreover, the coefficients of determination are higher than those of the two methods by 4.4% and 9.1%, respectively. For the ablation experiment of the improved transfer learning algorithm, after removing the normalization processing and feature dimensionality reduction module, the performance of the algorithm decreases in all aspects. Its RMSE increases by 0.033 and 0.016, and the coefficient of determination decreases by 6.8% and 3.7%, respectively. It indicates that the normalization processing has a greater impact on the performance of the model, and the feature dimensionality reduction mainly affects the computation speed of the model. The relevant terms and variables used in the manuscript are shown in Table 6.

Table 6: Manuscript-related variables and their interpretation

Serial number	Term	Detailed Information
1	GRBT	Gradient Boosting Regression Tree
2	EC-PPL	Evaporation Crystallization-Polymer Pen Lithography
3	SVM	Supported Vector Machine
4	KNN	K-NearestNeighbor
5	B/G	Ratio of shear modulus to bulk modulus
6	f_x	Output variables of the training samples
7	w_i	Weights of the training samples
8	d_i	Distance from training samples to predicted samples
9	W	Weighted average of samples
10	$H_j(x_i)$	j Strong learners consisting of residual trees
11	$h_{j+1}(x_i)$	Weak Learner for $j+1$ Residual Trees
12	λ	step-down
13	$f(x)$	regression constructor
14	$\varphi(x)$	nonlinear mapping function
15	b	bias entry (computing)
16	C	Regularization constants
17	ζ and ζ^*	slack variable (math)
18	q_i and q_i^*	Lagrange multiplier (math)
19	$\kappa(X_i, X_j)$	kernel function (math)
20	$\kappa(x_i, x_j)$	radial basis kernel function (math)
21	σ	Width of radial base core
22	$\|x_i - x_j\|$	Euclidean distance between two points
23	μ	octahedral factor
24	r_B	The radius of the B-site ion
25	r_X	The radius of the X halogen ion
26	TF	Tolerance factor
27	r_A	The radius of the A-position ion
28	r	Pearson coefficient
29	X_{std}	The feature data after normalization
30	E_b	The formation energy of the compound
31	$E(A_mB_nC_k)$	Total energy of the compound
32	$E(A)$	The energy of the free atom A
33	$E(B)$	The energy of the free atom B
34	$E(C)$	The energy of the free atom C

4 Discussion

A band gap prediction model for chalcocite materials based on the improved GBRT algorithm and migration learning was proposed and applied to the actual analysis of chalcocite samples, and the validity of the prediction model was verified by relevant experimental analysis. The GBRT-SVM hybrid model converged faster in both the training and test sets, and the final convergence value was smaller than that of the other methods, as the adaptive reduction of the step size could effectively improve the pre-optimization speed and post-optimization accuracy of the model. Compared with the improved methods of Yang [7] and Zhang [8], the proposed method could significantly improve the computational speed while guaranteeing the PA. The RMSE and coefficient of determination of the band gap prediction of the hybrid model were better than those of the basic method, and the consistency between the predicted value and the true value was higher, with higher PA, compared with the RMSE model of Zhi [10]. In cross-validation, the RMSE of the improved migration learning method was 0.0417, which was 0.0013 higher than that of the independent test set validation results. However, it still less than 0.05, which could meet the relevant requirements, and the proposed method could effectively improve the computational efficiency and reduce the cost of screening new materials compared with Liu [13]. The GBRT-SVM hybrid model, through its powerful classification and regression analysis capabilities, could effectively improve the discovery efficiency of new chalcogenide materials and quickly identify material combinations with potential high performance. This could promote the rapid development of solar cells.

5 Conclusion

Aiming at the problem of insufficient PA of existing screening methods for perovskite materials, this study proposed the use of an improved GRBT algorithm for screening of hybrid perovskite band gap and hierarchical transfer learning for halogen double perovskite multi-property screening. Experiments indicated that the LF of the GBRT-SVM hybrid model decreased faster, with the astringent LF values being 0.115 and 0.160 lower than those of the GBRT and SVM algorithms, respectively. The optimization speed was 35 and 20 iterations faster than the other two algorithms, respectively. The pedestrian band gap prediction RMSE and coefficient of determination of GBRT-SVM hybrid model were 0.032 and 99.4%, respectively. The predicted values all converge neared the diagonal of the horizontal and vertical axes, with no significant deviation from the predicted points. They were 0.017 and -2.2, and 0.023 and -3.7% lower than GBRT and SVM, respectively. The maximum prediction error of the GBRT-SVM hybrid model for the band gap of hybrid perovskite was 0.071, and the minimum error was 0.005. The mean absolute error, RMSE, and coefficient of determination of the improved transfer learning method were 0.0097, 0.0205, and -5.06% lower than those of the ordinary method, respectively, and the running speed was 1.92s faster. The RMSE of bulk modulus measured by the improved transfer learning method was 0.037 lower than directly predicted, and the coefficient of determination was 11.18% higher. Factors that had an impact on the shear modulus of materials above 0.04 included Mendeleev number, electronegativity, octahedral factor, and total valence electrons. The BGVs of seven halogen double perovskite new materials were between [1.14-1.62] eV, with formation energies all below 0.05, and the ratio of shear modulus to bulk modulus greater than 1.75. The present study has identified several areas that necessitate further refinement. For instance, although the development of a screening model for new chalcogenide materials has been achieved using a small-scale dataset, enhancing the screening accuracy and efficiency to a certain extent, the screening accuracy of the model can be further enhanced through the integration of density flooding calculation in future iterations.

List of abbreviations

GRBT: Gradient Boosting Regression Tree

EC-PPL: Evaporation Crystallization-Polymer Pen Lithography

SVM: Supported Vector Machine

KNN: K-NearestNeighbor

B/G: Ratio of shear modulus to bulk modulus

f_x : Output variables of the training samples

w_i : Weights of the training samples

d_i : Distance from training samples to predicted samples

W : Weighted average of samples

$H_j(x_i)$: j Strong learners consisting of residual trees

$h_{j+1}(x_i)$: Weak Learner for $j+1$ Residual Trees

λ : Step-down

$f(x)$: Regression constructor

$\varphi(x)$: Nonlinear mapping function

b : Bias entry (computing)

C : Regularization constants

ζ and ζ^* : Slack variable (math)

q_i and q_i^* : Lagrange multiplier (math)

$\kappa(X_i, X_j)$: Kernel function (math)

$\kappa(x_i, x_j)$: Radial basis kernel function (math)

σ : Width of radial base core

$\|x_i - x_j\|$: Euclidean distance between two points

μ : Octahedral factor

r_B : The radius of the B-site ion

r_X : The radius of the X halogen ion

TF : Tolerance factor

r_A : The radius of the A-position ion

r : Pearson coefficient

X_{std} : The feature data after normalization

E_b : The formation energy of the compound

$E\left(A_mB_nC_k\right)$: Total energy of the compound

$E\left(A\right)$: The energy of the free atom A

$E\left(B\right)$: The energy of the free atom B

$E\left(C\right)$: The energy of the free atom C

Funding

This work is partially supported by the Natural Science Foundation of Shangluo University (21SKY114, 22KYZX14).

References

[1] Chen H, Cheng Q, Liu H, Cheng S, Wang S, Chen W, & Li Y. Organic-semiconductor-assisted dielectric screening effect for stable and efficient perovskite solar cells. Science Bulletin, 2022, 67(12): 1243-1252. DOI: 10.1021/acsaem. 2c04166.s001

[2] Shyur H J, & Shih H S. Resolving rank reversal in TOPSIS: A comprehensive analysis of distance metrics and normalization methods. Informatica, 2024, 35(4): 837-858. DOI: 10.15388/24-INFOR576

[3] Kim H W, Han J H, Ko H, Samanta T, Lee D G, Jeon D W, & Cho S B. High-Throughput screening on halide Perovskite derivatives and rational design of Cs_3LuCl_6. ACS Energy Letters, 2023, 8(8): 3621-3630. DOI: 10.1021/acsenergylett. 3c01207.s002

[4] Brociek R, Goik M, Miarka J, Pleszczyński M, & Napoli C. Solution of inverse problem for diffusion equation with fractional derivatives using metaheuristic optimization algorithm. Informatica, 2024, 35(3): 453-481. DOI: 10.15388/24-INFOR563

[5] Nguyen X B, Nguyen T H, Nguyen K V T, Nguyen T T T, & Nguyen D D. Efficient prediction of axial load-bearing capacity of concrete columns reinforced with FRP bars using GBRT model. Journal of Materials and Engineering Structures «JMES», 2023, 10(4): 551-568. DOI: 10.55228/jtst.12(2).38-47

[6] Abdelbasset W K, Elsayed S H, Alshehri S, Huwaimel B, Alobaida A, Alsubaiyel A M, & Abourehab M A. Development of GBRT model as a novel and robust mathematical model to predict and optimize the solubility of decitabine as an anti-cancer drug. Molecules, 2022, 27(17): 5676. DOI: 10.3390/molecules27175676

[7] Yang H, Liu X, & Song K. A novel gradient boosting regression tree technique optimized by improved sparrow search algorithm for predicting TBM penetration rate. Arabian Journal of Geosciences, 2022, 15(6): 461-482. DOI: 10.1007/s12517-022-09665-4

[8] Zhang W, Li R, Zhao J, Meng X, & Li Q. Miss-gradient boosting regression tree: A novel approach to imputing water treatment data. Applied Intelligence, 2023, 53(19): 22917-22937. DOI: 10.1007/s10489-023-04828-6

[9] Diao X, Diao Y, Tang Y, Zhao G, Gu Q, Xie Y, & Zhang L. High-throughput screening of stable and efficient double inorganic halide perovskite materials by DFT. Scientific Reports, 2022, 12(1): 12633. DOI: 10.21203/rs.3.rs-1565189/v1

[10] Zhi C, Wang S, Sun S, Li C, Li Z, Wan Z, & Liu Z. Machine-learning-assisted screening of interface passivation materials for perovskite solar cells. ACS Energy Letters, 2023, 8(3): 1424-1433. DOI: 10.1021/acsenergylett. 2c02818.s001

[11] Liu W, Lu Y, Wei D, Huo X, Huang X, Li Y, & Song D. Screening interface passivation materials intelligently through machine learning for highly efficient perovskite solar cells. Journal of Materials Chemistry A, 2022, 10(34): 17782-17789. DOI: 10.1021/acsenergylett. 2c02818.s001

[12] Lai M, Shin D, Jibril L, & Mirkin C A. Combinatorial synthesis and screening of mixed halide perovskite megalibraries. Journal of the American Chemical Society, 2022, 144(30): 13823-13830. DOI: 10.1021/jacs.2c05082

[13] Liu H, Feng J, & Dong L. Quick screening stable double perovskite oxides for photovoltaic applications by machine learning. Ceramics International, 2022, 48(13): 18074-18082. DOI: 10.1016/j.ceramint.2022.02.258

[14] Yuan S, Liu Y, Lan J, Yang W, Long H, Li W, & Fan J. Accurate dimension prediction for low-dimensional organic–inorganic halide perovskites via a self-established machine learning strategy. The Journal of Physical Chemistry Letters, 2023, 14(32): 7323-7330. DOI: 10.1021/acs.jpclett. 3c01915.s001

[15] Kurani A, Doshi P, Vakharia A, & Shah M. A comprehensive comparative study of artificial neural network (ANN) and support vector machines (SVM) on stock forecasting. Annals of Data Science, 2023, 10(1): 183-208. DOI: 10.1007/s40745-021-00344-x

[16] Santos C F G D, & Papa J P. Avoiding overfitting: A survey on regularization methods for convolutional neural networks. ACM Computing Surveys (CSUR), 2022, 54(10s): 1-25. DOI: 10.1145/3510413

[17] Chen Z. Research on internet security situation awareness prediction technology based on improved RBF neural network algorithm. Journal of Computational and Cognitive Engineering, 2022, 1(3): 103-108. DOI: 10.47852/bonviewjcce149145205514

[18] Hebbi C, & Mamatha H. Comprehensive dataset building and recognition of isolated handwritten Kannada characters using machine learning models. Artificial Intelligence and Applications, 2023, 1(3):179-190. DOI: 10.47852/bonviewaia3202624

[19] Wang G, Jia Q S, Zhou M C, Bi J, Qiao J, & Abusorrah A. Artificial neural networks for water quality soft-sensing in wastewater treatment: a review. Artificial Intelligence Review, 2022, 55(1): 565-587. DOI: 10.1007/s10462-021-10038-8

[20] Mena-Yedra R, López-Redondo J, Pérez-Sánchez H, & Martinez-Ortigosa P. Boosting pairwise molecular contrasts with scalable methods. Informatica, 2024, 35(3): 617-648. DOI: 10.15388/24-INFOR558