# Optimized Artificial Neural Network with Improved Sparrow Search Algorithm for Predicting Bandgap and Stability of $A_2B^{1+}B^{3+}X_6$ Perovskites

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

**Keywords:** artificial neural network, double perovskite, bandgap, stability, sparrow search algorithm

**Received:** December 24, 2024

*With the rapid development of solar power generation technology, chalcogenide battery materials are becoming more and more important. However, there are many combinations of chemical formulae for chalcogenide materials, and the traditional calculation methods are too costly in terms of labor and time. Therefore, an artificial neural network-based prediction model for the band gap and stability of halogenated double chalcogenide ($A_2B^{1+}B^{3+}X_6$) is proposed in the study. The structural features of the material are extracted using the Voronoi diagram method. The sparrow search algorithm is improved by using the chaotic number generator, stochastic difference variant, dynamic allocation strategy, and nonlinear inertia factor. The number of algorithmic populations is set to 50. The maximum and minimum values of the ratio of discoverers and joiners are 3:7 and 1:9 respectively. The learning factor is 1 and the warning value is 0.5. The improved algorithm is used to optimize the artificial neural network. Experiments show that the optimal fitness value of the improved algorithm is 145, 128, and 53 lower than that of GBR, SVR, and XGBoost, and the running time is 0.035 s, 0.127 s, 0.022 s, and 1.212 s lower than that of GBR, SVR, and XGBoost respectively, indicating that the improved algorithm performs better in optimization problems. The root mean square error is 0.053, which is lower than SSA, GBR, SVR, and XGBoost algorithms by 0.036, 0.019, 0.101 and 0.038 respectively. The mean absolute error and root mean square error of the model are 0.0217 and 0.0354 lower than that of the XGBoost model, respectively, and the coefficient of determination is 5.46% higher. The mean absolute error and root mean square error distributions of the model are lower than that of the XGBoost model by 0.0217 and 0.0354, and the coefficient of determination is higher by 5.46%. The six $A_2B^{1+}B^{3+}X_6$ mines selected for the study meet the requirement of band gap between [1.3, 1.4], the band gap prediction error between [-0.047, 0.009], and the stability indexes of five of the materials meet the requirement of less than 0.05. It can be concluded that the study can effectively improve the screening speed of chalcogenide materials, reduce the screening cost, and provide more promising new materials for solar power generation.*

*Povzetek: Raziskava predstavlja izboljšan nevronski model ISSA-BP za napoved pasovnega prepovednega območja in stabilnosti $A_2B_1{}^+B_3{}^+X_6$ perovskitov, ki z Voronoi značilkami dosega visoko kvaliteto in hitro konvergenco.*

## 1 Introduction

As society continues to evolve, there is a growing demand for energy among people, and various traditional fossil fuels are widely used in various industries due to their huge reserves and ease of exploitation [1]. However, fossil fuels produce large amounts of greenhouse gases and pollutants such as sulfur dioxide during combustion, leading to climate problems such as global warming and acid rain [2]. Meanwhile, due to the non-renewable nature of traditional fossil fuels, relying solely on fossil fuels will only lead to energy depletion and ultimately result in an energy crisis [3]. Clean renewable energy such as solar energy is an effective solution to the energy problem, and chalcogenide material is an efficient alternative to traditional crystalline silicon solar cells, which can effectively improve the efficiency of solar power generation due to its high light absorption coefficient, adjustable bandgap, and low preparation cost [4]. The laboratory photoelectric conversion efficiency of chalcogenide solar cells can reach 25.7%, the theoretical efficiency of a single layer can reach 33%, and the double-layer material can reach more than 45% [5]. Moreover, the preparation process of chalcogenide cells is simpler, requiring only 5 or 6 procedures. The processing temperature is lower, which can effectively reduce its production cost. However, chalcogenide batteries suffer from poor stability in various environments, making them prone to decomposition. Controlling the preparation process over large areas is challenging, and the presence of lead in the material poses a risk of environmental pollution. To address these issues, halide bilayer chalcogenide emerges as an effective alternative to ordinary chalcogenide materials [6]. However, the

traditional screening methods for chalcogenide materials, such as high-throughput screening, first-principle calculations, and experimental trial and error, often consume excessive time and material costs. Alternatively, machine learning (ML) methods are used to assist the screening, but the ordinary ML methods require high data quality and quantity, and there is a risk of overfitting. Hu et al. proposed a new ML framework to further assess the ion adsorption of $A_2B^{1+}B^{3+}X_6$. This framework used first principles calculations to protect a dataset of 640 ions, sorted the Pearson correlation of their output values, evaluated them comprehensively using multiple feature selection methods, and then screened the virtual space predicted by ML. Experiments showed that the gradient boosting decision tree algorithm had the highest prediction accuracy in this framework [7]. Zhao et al. raised a new ML-based screening approach to improve the screening speed of perovskite and reduce screening costs. This method extracted important features grounded on the constraints of charge neutrality and Goldschmidt tolerance factor, extracted 16 out of 21 features to describe known perovskite compounds, and trained perovskite formability and stability models. Experiments showed that the accuracy and recall of the two models were 0.983 and 1.00 and 0.971 and 0.943, respectively, which could distinguish between perovskite and non-perovskite [8]. Wu et al. proposed a new framework combining high-throughput experiments, subgroup discovery, and support vector machines for predicting the properties and structures of material synthesis. This framework integrated multiple ML techniques to reveal hidden structural property relationships in high-throughput experiments and quickly screen materials with high synthesis feasibility. Experiments showed that this framework had a 40% higher success rate in the synthesis of inorganic hybrid perovskites compared to traditional methods [9].

Selvaratnam et al. proposed a new method for determining independent screening and sparse operators in order to develop interpretable ML models. This method used domain overlap as a criterion to find the descriptor with the highest information content in classification problems, and proposed the hypothesis of using decision trees as scoring functions to find the best descriptor, which could improve prediction performance. The experiment showed that this method could improve the accuracy of the test set by 0.86 [10]. Zhu et al. proposed an ML-based thermodynamic stability prediction model for perovskite materials in order to find highly stable materials. The model utilized four classification and four regression ML algorithms, and its performance was assessed through a five-fold cross-validation approach. Experiment outcomes indicated that the model effectively predicted the stability of perovskite and identified 23 compounds with appropriate band gaps [11]. Sharma et al. proposed a new method combining first principles calculations and ML algorithms to obtain low bandgap perovskite materials. This method identified and predicted the optimal dopant for BaZrS3 perovskite for photovoltaic devices, reducing the material bandgap. Experiments showed that this method could reduce the bandgap of materials from 1.75 eV to 1.26 eV, and doping Ca at the Ba point was superior to doping Ti at the Zr point [12]. Mattur et al. raised a novel ML approach based on the random forest algorithm to reduce the computational difficulty of material bandgap values and properties. The study used 5329 types of perovskite oxides and employed the random forest algorithm to predict the properties and bandgap values of the materials. The experiment indicated that the prediction accuracy of this approach reached 91%, and the prediction speed was significantly improved, which could quickly discover new materials in the perovskite family [13]. Yuan et al. raised a novel self-built ML assisted prediction method for predicting the structural dimension of perovskite materials. This method divided the dimensions of materials into zero dimensional, one-dimensional, and two-dimensional dimensions, and used the optimal K-nearest neighbor model to predict the dimensionality of low dimensional perovskite materials. Experiments showed that the method could achieve 92.3% prediction accuracy in the test set, and the key factor determining the structural dimension was found, namely, ATSC1pe and SlogP_VSA2 related to the surface polarity and electrostatic potential of the organic interlayer Vander Waals [14].

In summary, experts and scholars explored the bandgap and stability prediction of perovskite from multiple perspectives, and achieved certain research results. However, existing methods paid less attention to $A_2B^{1+}B^{3+}X_6$, and it was difficult to find new $A_2B^{1+}B^{3+}X_6$ materials through experiments, which required a lot of cost and time. To improve the screening efficiency of chalcogenide materials, the band gap prediction error was reduced to within $\pm 0.05$. Therefore, the study proposed an $A_2B^{1+}B^{3+}X_6$ bandgap and stability prediction model with an improved artificial neural network (ANN). The Voronoi diagram approach was innovatively used to extract material structural features, and the sparrow search algorithm (SSA) was improved using a chaotic number generator, random difference mutation, dynamic allocation strategy, and nonlinear inertia factor. The improved algorithm was then used to optimize the ANN. The research aimed to reduce the excavation cost of perovskite new materials, improve calculation speed, and search for more stable new materials.

Based on the above related studies, Table 1 summarizes the research methodology, root mean square error (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] | First principle computing and ML | 0.103 | 1073.5 | Calculation time too long |
| Literature [8] | ML and Goldschmidt tolerance factors | 0.072 | 237.4 | Not applicable to covalent compounds |
| Literature [9] | Subgroup discovery and support vector machines combined | 0.098 | 186.5 | Higher requirements for parameterization |
| Literature [10] | Decision trees and sparse operators | 0.105 | 98.8 | Difficult to handle high-dimensional sparse data |
| Literature [11] | Four regression ML algorithms | 0.087 | 109.2 | Easy to ignore important features |
| Literature [12] | First principle computing and ML combined | 0.095 | 983.7 | High computational cost and difficult to get high quality data |
| Literature [13] | Random forest algorithms | 0.142 | 153.6 | Need to traverse all decision trees to make predictions |
| Literature [14] | Self-constructed ML-assisted prediction methods | 0.085 | 105.3 | Noise, missing values or bias in the data can degrade the performance of the model. |
| This text | Voronoi diagrams and ANNs | 0.068 | 58.7 | / |

In Table 1, although the existing studies are able to achieve certain results in the face of chalcogenide material screening, their performance in terms of RMSE and computational efficiency is still unsatisfactory, and the requirements on data and model parameters are also more demanding. Therefore, the study innovatively proposes an ANN-based prediction model for the band gap and stability of halogenated bicalcite, which effectively improves the screening speed and accuracy of chalcogenide materials. The proposed method performs well in terms of the RMSE of the prediction results and the calculation time.

# 2 Methods and materials

## 2.1 Material structure feature extraction and model selection based on voronoi diagram method

$A_2B^{1+}B^{3+}X_6$ is a variant of perovskite structure, composed of two halide ions arranged in reverse. Its structural formula is $A_2B^{1+}B^{3+}X_6$, where A represents an alkali metal cation, two Bs are filled with monovalent and trivalent metal cations, and X is a halide anion [15]. The elements and their arrangement at each point of $A_2B^{1+}B^{3+}X_6$ are shown in Figure 1.

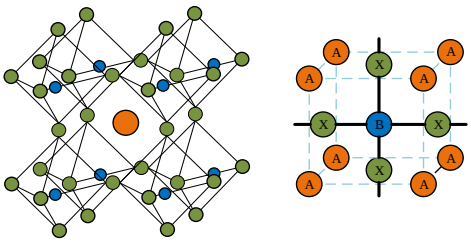


Figure 1 Specific structure of $A_2B^{1+}B^{3+}X_6$

In Figure 1, orange nodes represent alkali metal cations, green nodes represent halide anions, and blue nodes represent monovalent and trivalent metal cations. Different ions can form thousands of elemental structures, and the stability evaluation factors and tolerance factors of the perovskite structure are calculated as shown in Equation (1) [15].

$$t = \frac{r_A + r_X}{\left(\frac{r_{B^{1+}} + r_{B^{3+}}}{2} + r_X\right) \cdot \sqrt{2}} \tag{1}$$

In Equation (1), $t$ represents the tolerance factor, $r_A$ represents the ionic radius of the element at the A-site, $r_X$ represents the ionic radius of the element at the X position, $r_{B^{1+}}$ is the ionic radius of the element at the $B^{1+}$ site, and $r_{B^{3+}}$ represents the ionic radius of the element at the $B^{3+}$ site. The calculation of octahedral factor is shown in Equation (2) [16].

$$\mu = \frac{r_{B^{1+}} + r_{B^{3+}}}{2r_X} \tag{2}$$

In Equation (2), $\mu$ represents the octahedral factor. Equations (1) and (2) were chosen for the study to calculate the octahedral factor and tolerance factor of the material to characterize the structure of the chalcogenide material, which is suitable for evaluating the predicted properties of the material. In response to the problem that

traditional ML methods can only learn element features, this study adopts the Voronoi diagram method to extract the structural features of materials, further improving the accuracy of the model's bandgap prediction. The Voronoi diagram method can individually correspond all atoms in a material to polyhedra, and can perceive subtle changes in the material structure with low computational cost [17]. Compared to the traditional method, this approach enhances the material's feature extraction capabilities by simulating its microstructure, extracting geometric features, and analyzing mechanical properties. This not only provides robust support for the research and application of chalcogenide materials but also reduces computational complexity. The effective coordination number calculation of the Voronoi diagram method is shown in Equation (3).

$$N_V = \frac{\left(\sum_i S_i\right)^2}{\sum_i S_i^2} \tag{3}$$

In Equation (3), $N_V$ represents the effective coordination number, $S_i$ means the size of the $i$ th side of the polyhedron, and the average bond length is calculated as shown in Equation (4).

$$\overline{x} = \frac{\sum_i S_i \left\| r_i - r_j \right\|_2}{\sum_i S_i} \tag{4}$$

In Equation (4), $\overline{x}$ represents the average bond length, $r_i$ means the position of the atom on the $i$ th side, and $r_j$ means the position of the central atom. In the Voronoi diagram structure, adjacent atoms have the greatest impact on the effective coordination number and average bond length. The calculation of the nearest neighbor order parameter for Voronoi diagram is shown in Equation (5).

$$\varphi = 1 - \frac{\sum_i S_i \cdot \gamma}{C_t \sum_i S_i} \tag{5}$$

In Equation (5), $\varphi$ represents the nearest neighbor order parameter, $\gamma$ represents the Dirac function, and $C_t$ represents the concentration of $t$ atoms in the structure. The study uses Equations (3)-(6) for Voronoi diagram feature extraction, which can enhance the feature extraction ability of the model for materials and is suitable for material data processing. The nearest neighbor order parameter represents the degree of order of the system during the phase transition process. When the atomic arrangement in the structure is disordered, the order parameter is infinitely close to zero. When the atomic arrangement is ordered, the order parameter is one. The calculation of the local environmental differences of atoms in the structure is shown in Equation (6).

$$\Delta_e = \frac{\sum_i S_i \left| e_i - e \right|}{\sum_i S_i} \tag{6}$$

In Equation (6), $\Delta_e$ represents local environmental differences, $e_i$ represents the physical quantity of the atom on the $i$ th side, and $e$ means the physical quantity of the central atom. The local environmental differences are solved by weighted summation of polyhedral areas, and the closer the distance between adjacent atoms and the central atom, the greater the impact on their environmental differences. Due to the unequal and large number of atoms in $A_2B^{1+}B^{3+}X_6$ composed of all ions, the structural features generated by the four Voronoi diagram methods mentioned above cannot be directly used to train ANN models. After obtaining the data of the four structural features of all the atoms through the Voronoi diagram method, the study calculates the range of values, the average value, and the average absolute deviation value of all the structural features, which are used to train the corresponding ML models. ANNs, as a type of ML, have strong data processing capabilities and are widely used in various fields. Backpropagation neural networks (BPNNs), as one of the types of ANN networks, have excellent nonlinear data processing capabilities and can continuously optimize the model through error feedback during training to improve its predictive ability [18]. The BPNN has the superiority of low computational complexity and strong nonlinear mapping ability. Figure 2 illustrates the detailed training procedure.

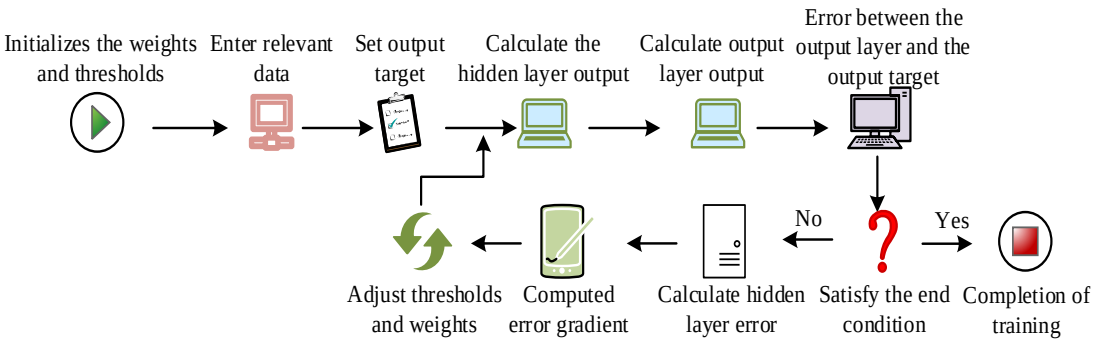

Figure 2: Specific training process of BPNN

The initialization of the neural network's weights and thresholds is depicted in Figure 2. Using the input data and the desired output, the outputs of each neuron in the hidden layer (HL) and output layer (OL) are computed separately. Depending on the discrepancy between the OL's result and the target value, a decision is made on whether to terminate the iteration. If the termination criteria are met, the training concludes. If not, the error values and gradients for the HL are computed, and the neural network's weights and thresholds are adjusted accordingly. The number of HLs and neurons in a BPNN greatly influences its computational power and prediction accuracy. It can effectively optimize the performance of the network and improve the prediction accuracy of the model, and the adjustment of the weights and thresholds are adjusted using the gradient descent method. A limited number of HL neurons may hinder the network's ability to fully capture data information, potentially leading to suboptimal solutions. Conversely, an excessive number of neurons can elevate model complexity, prolonging the training period. According to studies, a neural network with a single HL can approximate any function, and the optimal number of HL neurons is determined by Equation (7).

$$h = \sqrt{i + j} + a \tag{7}$$

In Equation (7), $h$ means the number of HL nodes, $i$ means the number of input layer nodes, $j$ means the number of OL nodes, and $a$ means a random integer between [1, 10]. The research focuses on addressing the issue of ordinary BPNNs being prone to falling into local optima. To improve this, the SSA is employed. Additionally, to further enhance the operational efficiency and accuracy of the SSA, a multi-strategy approach is utilized for its refinement.

## 2.2 Optimization of ANN model based on improved SSA

When using BPNN for bandgap prediction and stability analysis of $A_2B^{1+}B^{3+}X_6$, issues such as non-convex optimization and initial parameter selection may lead to gradient disappearance or guarantee, resulting in local optima. Therefore, the study adopts an improved SSA for parameter optimization of BPNN, solves the problem of local optima, and improves prediction accuracy. In the SSA, sparrow populations can be divided into discoverers, joiners, and vigilantes, and individuals are classified according to their fitness values. Individuals with higher fitness values are called discoverers, who are responsible for searching for food in different areas. After discovering food, discoverers guide joiners to search for food. When danger is detected, discoverers also switch to alert other individuals in the population and move to safe areas to continue foraging [19]. During the ongoing iterative optimization of the sparrow population, if the fitness of a joiner surpasses that of a discoverer, the joiner transitions into a discoverer role. The updated position calculation for the discoverer is provided in Equation (8).

$$S_{i,j}^{x+1} = \begin{cases} S_{i,j} \cdot \exp\left(-\dfrac{i}{\kappa \cdot it_{\max}}\right), R < ST \\ S_{i,j} + Q \cdot L, R \geq ST \end{cases} \tag{8}$$

In Equation (8), $S_{i,j}^{x+1}$ represents the updated position of the discoverer, $S_{i,j}$ means the position information of sparrow individual $i$ in the $j$ th dimension, $\kappa$ represents a random number between 0-1, $it_{\max}$ represents the max number of iterations, $R$ represents the alert value between 0-1, $ST$ represents the safety value between 0.5-1, $Q$ represents a random number, and $L$ represents a matrix with a length of 1 and a width of dimensionality, all of which are 1. The detailed steps of the SSA are illustrated in Figure 3.

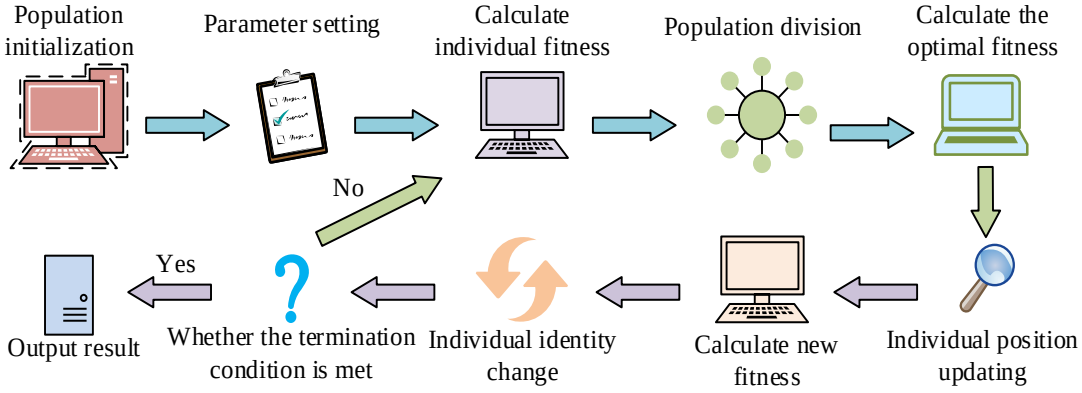


Figure 3: Specific flow of SSA

In Figure 3, the population is initialized first to determine the relevant algorithm parameters, including the number of populations, the maximum number of iterations, the ratio of discoverers and joiners, the learning factor and the warning value, etc. Reasonable parameter settings can effectively improve the algorithm's search efficiency, the speed of convergence, and the ability of the global search to avoid the problem of local optimization. Then the size of the fitness value of the individual is calculated and sorted according to the size. The population is categorized into discoverers and joiners in a ratio of 2:8, with the top 20% being designated as discoverers. The current optimal and worst positions and their fitness values are calculated, the positions of discoverers, joiners, and alert individuals are updated based on relevant formulas, new fitness values are calculated, and reordering is done. The joiners with high fitness values are transformed into discoverers. The determination of whether the iteration termination condition has been met is made, and the current optimal position is output. Although the SSA has high global search ability and fast search speed, its local search ability is weak. The code of individual identity in SSA is shown in Figure 4.

```
sorted_indices = np.argsort(fitness)
discoverers = population[sorted_indices[:num_discoverers]]
followers = population[sorted_indices[num_discoverers:]]
```

Figure 4 Code of individual identity in SSA

The study addresses the problem that the random generation of the initial population affects the search accuracy by using a chaotic number generator, which generates pseudo-random numbers by means of chaotic properties. These numbers have good randomness and traversal, which can optimize the random operation in the algorithm and thus improve the global search ability and convergence speed of the algorithm. The one-dimensional mapping random allocation method of the chaotic model is shown in Equation (9).

$$\begin{cases} s_{n+1} = \sin\left(2/s_n\right), n = 0,1,...,N \\ -1 \le s_n \le 1, s_n \ne 0 \end{cases} \tag{9}$$

In Equation (9), $s_{n+1}$ means the position of the $n+1$ th individual, $s_n$ represents the position of the $n$ th individual, and $s_n$ is not equal to 0 to ensure that no constant points are generated within the interval. To expedite the algorithm's search speed and enhance its global search capability in the initial stages and local search capability in the later stages, the population's proportion of discoverers and joiners is dynamically adjusted. In the early stages, the algorithm boosts the number of discoverers to broaden the search scope, while in the later stages, it increases the number of joiners to intensify local search and elevate calculation precision. The allocation ratio of the sparrow population progressively decreases as the iteration count rises, with the decreasing trend calculated as shown in Equation (10).

$$R = R_{\max} - \left(R_{\max} - R_{\min}\right)\left(\frac{t}{T}\right)^2 \tag{10}$$

In Equation (10), $R$ represents the current allocation ratio between discoverers and joiners, $R_{\max}$ represents the set max allocation ratio, $R_{\min}$ represents the set min allocation ratio, $t$ means the current iteration count, and $T$ means the max iteration count. The study addresses the issue of individual sparrow discoverers becoming trapped in local optima. To increase their likelihood of escaping these local optima, it introduces a random difference variant. Specifically, for the first half of the individuals, Gaussian distribution perturbation is applied to their fitness values. This high-speed distribution perturbation aids the individuals in jumping out of local extreme points, thereby enhancing their diversity. Simultaneously, it boosts the algorithm's local search capability, as outlined in Equation (11).

$$X_t^* = X_t + N\left(0,1\right)\cdot X_t \tag{11}$$

In Equation (11), $X_t^*$ means the position of the individual after mutation, $X_t$ means the position of the individual before mutation, and $N\left(0,1\right)$ means a random number that follows a normal distribution between 0-1. Individuals with fitness values in the bottom half are perturbed with a Cauchy distribution, which can mutate individuals to produce random numbers far from the origin and enhance their global search ability, as calculated in Equation (12).

$$X_t^* = X_{tbest} + cauchy\left(0,1\right)\cdot X_t \tag{12}$$

In Equation (12), $X_{tbest}$ represents the optimal position found by the discoverer, and $cauchy\left(0,1\right)$ represents the standard Cauchy distribution. In the SSA, the discoverer sends a signal to summon the joiners after finding the area where food exists. The joiners then conduct a more accurate search within the reduced area to obtain the global optimal solution. However, when the joiner moves towards the area where the discoverer is located, the directional coefficients of traditional algorithms are only 1 and -1, which results in insufficient search of local areas and reduces search accuracy. Therefore, the study introduced a nonlinear inertia weight factor to optimize the position update direction of the joiner, as calculated in Equation (13).

$$w = w_{\max} - \left(w_{\max} - w_{\min}\right)\left(\frac{\pi t}{T}\right)^2 \quad (13)$$

In Equation (13), $w$ represents the nonlinear inertia weight factor, $w_{\max}$ means the max value of the nonlinear inertia weight factor, and $w_{\min}$ means the min value of the nonlinear inertia weight factor. The updated position of the fitness value of the individual after improvement in the first half of the population is shown in Equation (14).

$$S_{i,j}^{x+1} = Q \cdot \exp\left(\frac{S_w - S_{i,j}}{i^2}\right) \quad (14)$$

In Equation (14), $S_w$ represents the individual farthest from the center in the population, and $i^2$ means the square of the index of the $i$ th sparrow in the current iteration. The position update of the individual fitness value of the joiner in the latter half of the population is shown in Equation (15).

$$S_{i,j}^{x+1} = S_b + \left|S_{i,j} - S_b\right| \cdot L \cdot L^* \quad (15)$$

In Equation (15), $S_b$ represents the individual closest to the center in the population, $L^*$ represents a matrix with a length of 1 and a width of dimensionality, and the number in the matrix is a random number of -1 or 1. The study uses Equations (9)-(15) for SSA improvement to enhance its search speed and local search accuracy, which is suitable for model optimization processing. The improved SSA runs as shown in Figure 5.

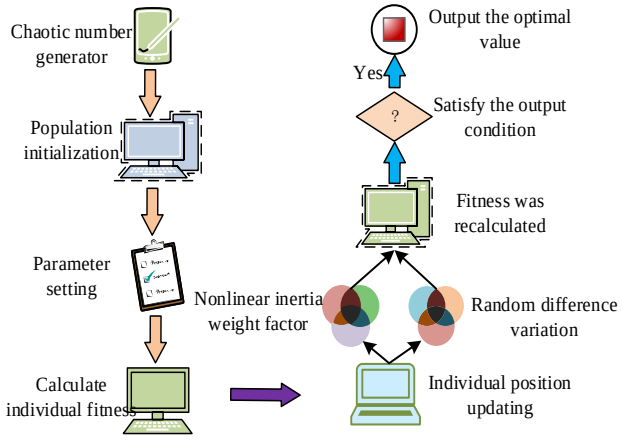


Figure 5: Operation flow of the improved SSA

In Figure 5, the study first uses a chaotic number generator to initialize the population after replacing the original population processing, sets the maximum value of the allocation ratio of discoverers and joiners to 3: 7 and the minimum value to 1: 9, and calculates the fitness of all individuals and ranks them. The position of the discoverer is updated, and likewise, the position of the joiner is adjusted using a nonlinear inertia weight factor to broaden the algorithm's local search capabilities. Additionally, stochastic differential variation is applied to the discoverer to prevent it from getting stuck in a local optimum. Once the position update is complete, all individuals' positions and fitness values are recalculated. Subsequently, it is determined whether the termination conditions have been met, and if so, the relevant results are outputted. The improved SSA has faster search speed in the early iteration, higher search accuracy in the late iteration, and is not easy to fall into local optimal solutions. The ISSA-BP model prediction process based on improved SSA optimized BPNN is shown in Figure 6.

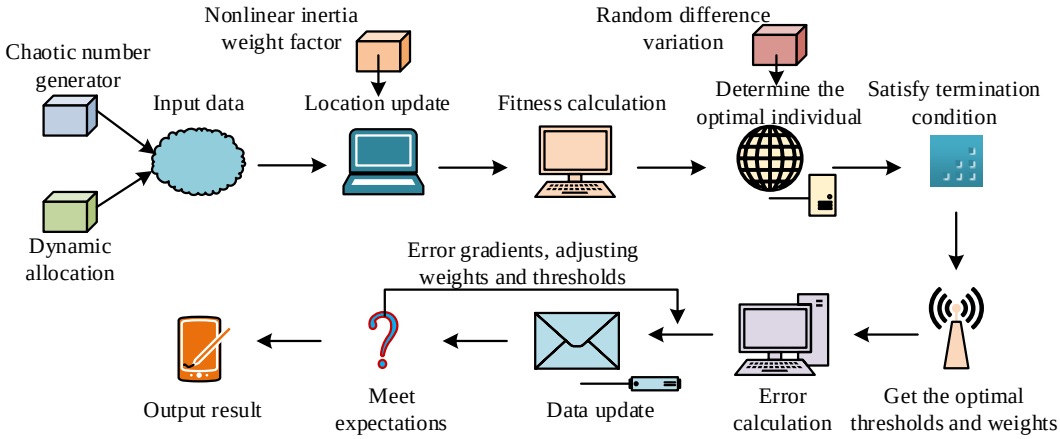


Figure 6: Prediction flow of optimized BPNN based on improved SSA

In Figure 6, the ISSA-BP model initially optimizes the SSA by setting its corresponding parameters. Specifically, the algorithm population is set to 50, with the ratio of discoverers to joiners ranging from a minimum of 1: 9 to a maximum of 3: 7. The learning factor is set to 1, and the early warning value is set to 0.5, among other configurations. Moreover, the optimized algorithm is used to obtain the optimal weights and optimal thresholds of the BPNN and set the corresponding desired output targets. The study calculates the error between the output value of the output layer and the desired output value, updates the weights and thresholds, determines whether the desired

output value is satisfied, and calculates the error gradient if it is not satisfied, and adjusts the weights and thresholds according to the gradient. The modulation of weights and thresholds can effectively optimize the performance of the network and improve the prediction accuracy of the model, both of which are adjusted using the gradient descent method. Until the desired output is satisfied or the maximum number of iterations is reached, the final output value is the bandgap prediction value of the chalcogenide material.

The study adopts the Voronoi diagram method to extract the structural features of the material to further improve the band gap prediction accuracy of the model. The Voronoi diagram method uniquely maps each atom in the material to individual polyhedra, enabling the detection of subtle structural changes with minimal computational expense. This provides robust support for the research and application of chalcocite materials. The study adopts chaotic number generator, nonlinear inertia weight factor and random difference variation to improve the SSA, which can effectively improve the global search speed and local search accuracy of the SSA. Then the improved algorithm is used to optimize the ANN, and the optimized ANN is used to carry out the prediction and analysis of the new materials of chalcocite. Finally the best new materials of chalcocite were obtained.

# 3 Results

## 3.1 Performance analysis of improving sparrow algorithm

The algorithm ran in an Intel Core i7-12600 K environment with a frequency range of 3.7 GHz-4.9 GHz, a GeForce GTX 3060 GPU, and 16 GB of memory. The max number of iterations for the algorithm was set to 200, and the comparative algorithms used in the experiment included Gradient Boosting Regression (GBR), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost). The comparison of optimization curves for different algorithms is shown in Figure 7.

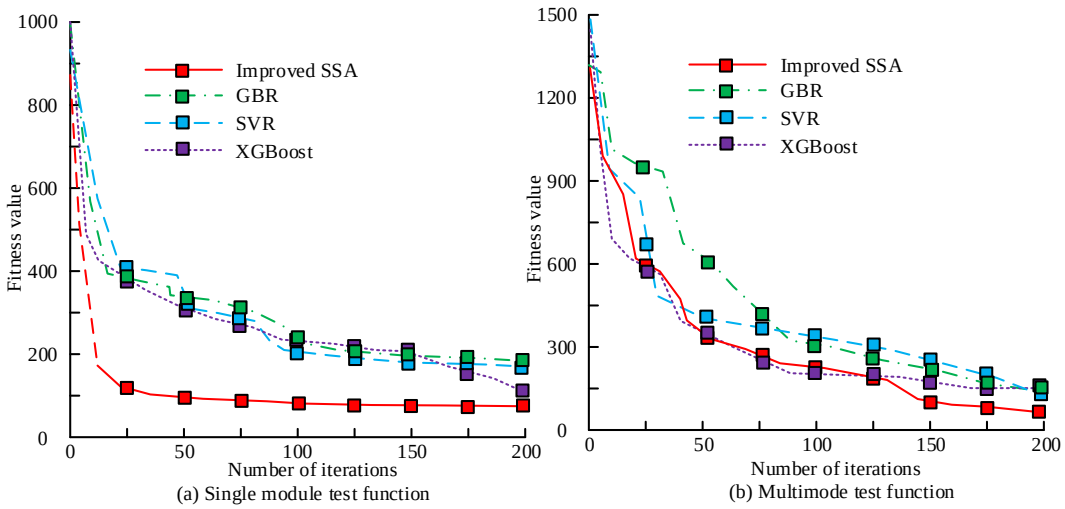


Figure 7: Comparison of optimal fitness value finding curves for different algorithms

In Figure 7 (a), the optimization curve of the improved SSA decreased the fastest, completing most of the optimization tasks in 0-30 iterations. The fitness value decreased from 870 to 84, a decrease of 786. The other three algorithms reached their optimal fitness values after 123, 104, and 197 iterations, respectively. The optimal fitness values of the improved SSA were 145, 128, and 53 lower than the other three algorithms, respectively. In Figure 7 (b), the optimization curve of the improved SSA slowed down and reached the optimal fitness value only after 154 iterations. The fitness value decreased from 1294 to 136, a decrease of 1158, while the optimization speed of the other three algorithms decreased even faster. The study selected common single-mode and multi-mode test functions for the performance check of the algorithms. The single-mode test function has only one globally optimal solution, and the multi-mode test function has multiple locally optimal solutions and one globally optimal solution. The complexity comparison of different algorithms is shown in Table 2.

Table 2: Compares the complexity of different algorithms

| Test function | Algorithm | Number of iterations | Run time (s) | RMSE |
|---|---|---|---|---|
| Single module test function | Improved SSA | 28 | 0.037 | 0.048 |
| | SSA | 84 | 0.072 | 0.074 |
| | GBR | 123 | 0.164 | 0.065 |
| | SVR | 104 | 0.059 | 0.127 |
| | XGBoost | 197 | 1.249 | 0.073 |

| | | | | |
|---|---|---|---|---|
| Multimode test function | Improved SSA | 154 | 0.204 | 0.053 |
| | SSA | 214 | 0.427 | 0.089 |
| | GBR | 174 | 0.875 | 0.072 |
| | SVR | 322 | 1.537 | 0.154 |
| | XGBoost | 259 | 5.624 | 0.091 |

In Table 2, among different test functions, improving SSA required fewer iterations, with 56 and 60 fewer iterations respectively compared to before optimization. In the single-mode test function, the improved SSA reduced the running time by 0.035 s, 0.127 s, 0.022 s, and 1.212 s compared to SSA, GBR, and XGBoost algorithms, respectively. In the multi-mode test function, the improved SSA's running time increased by 0.167 s, while the running time of other algorithms increased faster. The RMSE of the improved algorithm was 0.053, which was lower than the other four algorithms, 0.036, 0.019, 0.101, and 0.038, respectively. Ablation experiments were performed on the SSA and the comparison results are shown in Table 3.

Table 3: SSA ablation experiments

| Test function | Algorithm | Number of iterations | Run time (s) | RMSE |
|---|---|---|---|---|
| Single module test function | Improved SSA | 28 | 0.037 | 0.048 |
| | Remove Chaotic Number Generator | 105 | 0.104 | 0.184 |
| | Removal of Random Difference Anomalies | 117 | 0.157 | 0.143 |
| | Remove dynamic allocation | 95 | 0.095 | 0.357 |
| | BPNN | 76 | 0.125 | 0.236 |

In Table 3, upon removing the optimization module from the improved SSA, the algorithm's performance improved across all cases. Additionally, removing the chaotic number generator, random difference anomalies, and dynamic allocation increased the RMSE of the algorithm by 0.136, 0.097, and 0.309, respectively. Notably, the improved SSA's RMSE was 0.188 lower than that of the standard BPNN model.

## 3.2 $A_2B^{1+}B^{3+}X_6$ belt conveyor and fixed value prediction analysis

The study obtained the corresponding molecular expressions of $A_2B^{1+}B^{3+}X_6$ from existing literature, and composed an experimental dataset of important features, including the bandgap characteristic atomic radius, potential occupancy rate, and highest and lowest molecular orbitals of each ion. This data acquisition method can effectively enhance the richness and diversity of the dataset, while the literature data have been rigorously experimentally validated and peer-reviewed, with high quality and reliability. After the element combination is completed, the existing perovskite compound data was deleted in the study. The comparison of bandgap prediction results after five fold cross validation using different algorithms is shown in Figure 8.

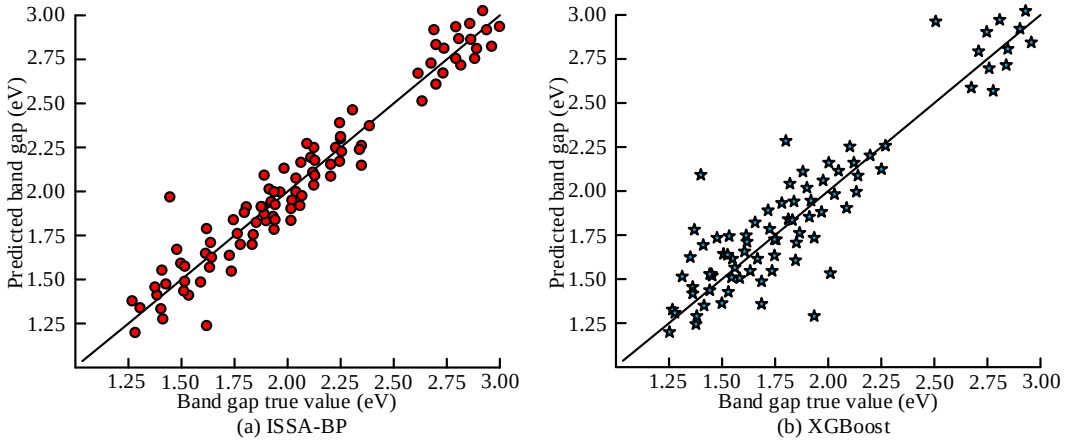


Figure 8: Comparison of bandgap prediction 5-fold cross-validation results for different algorithms

In Figure 8, the 5-fold cross-validation was to divide the dataset into five parts, four of which were used as the training set and one as the test set. The test was repeated five times, each subset was used as the test set, and the average of the results of the five tests were took. The data points indicated the predicted values of the model, and the

predicted values of the ISSA-BP model all converged near the diagonal of the horizontal and vertical coordinates, and there were no prediction points with large deviations. The closer the predicted values of the model were to the diagonal line, the higher the consistency between the predictions and the actual values, indicating a high prediction accuracy of the model. The mean absolute error (MAE), RMSE and coefficient of determination of the ISSA-BP model were 0.0524, 0.0683, and 98.87%, respectively, of which the distribution of the MAE and the RMSE were lower than that of the XGBoost model by 0.0217 and 0.0354, and the coefficient of determination was lower than the XGBoost model by 0.0217, 0.0354, and 0.0354, respectively. The coefficient of determination was 5.46% higher than that of the XGBoost model. The ISSA-BP model had a better prediction effect without underfitting or overfitting linearity, and the P-value of both models was less than 0.05, and the results were statistically significant. The relationship between the octahedral factor, tolerance factor, and bandgap of perovskite is shown in Figure 9.

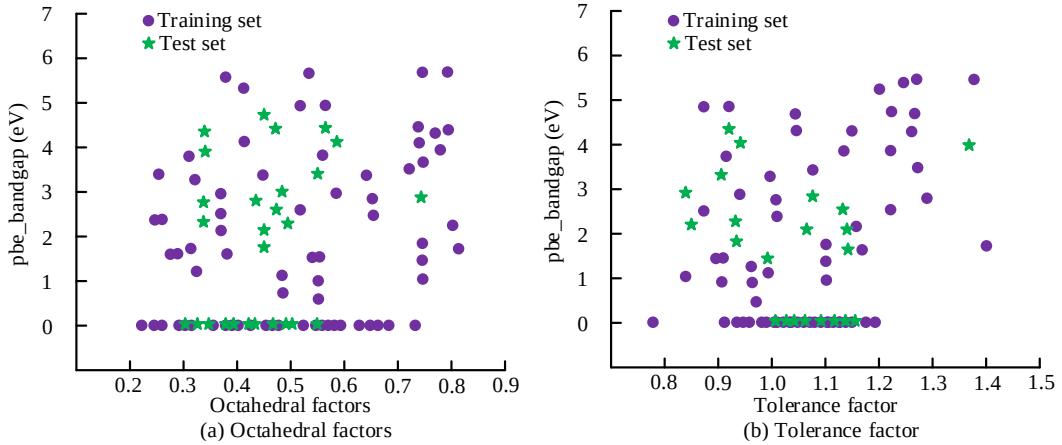


Figure 9: Effect of octahedral factor and tolerance factor on band gap in chalcogenide materials

In Figure 9 (a), the octahedral factor affected the bandgap mainly by changing the octahedral tilt and lattice distortion, while the tolerance factor affected the bandgap by regulating the stability of the structure and the phase transition. Both the octahedral factor and the tolerance factor showed a certain linear relationship with the bandgap value of the material, and the distribution of the octahedral factor was more dispersed in the training set, but most of them were distributed between 0.3-0.6 $O_f$ in the test set. In Figure 9 (b), the majority of samples in the test set were distributed between 0.9-1.2 $T_f$. The comparison of stability prediction results after five fold cross validation is shown in Figure 10.

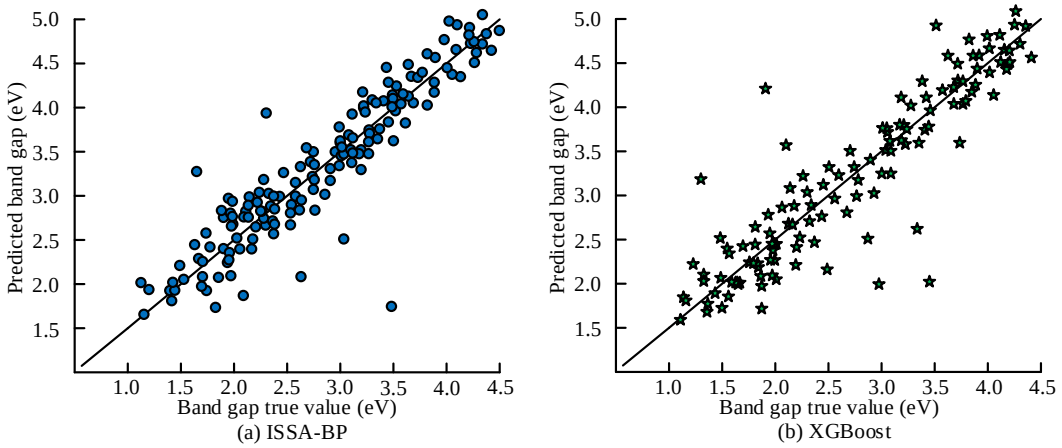


Figure 10: Comparison of 5-fold cross-validation results of different algorithms for stability prediction of chalcogenide materials

In Figure 10, the data points were the predicted values of the model for the stability of the material, and the diagonal line was the true value. Stability indicated the ability of the chalcogenide material to keep its intact structure and properties unchanged under different conditions. The ISSA-BP model fitted better, and most of the predicted values converged around the diagonal line, and only individual predictions were far from the diagonal line. The MAE and RMSE of the stability prediction of the model were lower than that of XGBoost by 0.0182 and 0.0154, respectively, and the coefficient of determination of the model was 1.42% higher than that of XGBoost, and

the ISSA-BP model was better in predicting the stability of chalcocite samples. A comparison of the band gap value prediction results of the ISSA-BP model for chalcogenide materials in different data sets is shown in Table 4.

Table 4: Comparison of band gap value prediction results of ISSA-BP model for chalcogenide materials in different datasets

| Dataset | Model | MAE | RMSE | Coefficient of determination (%) |
|---|---|---|---|---|
| Perovskite Database | ISSA-BP | 0.0531 | 0.0627 | 97.62 |
| | XGBoost | 0.0947 | 0.1025 | 90.58 |
| Materials Project | ISSA-BP | 0.0582 | 0.0695 | 97.05 |
| | XGBoost | 0.1072 | 0.1254 | 91.37 |

In Table 4, the MAE, RMSE, and coefficient of determination of the ISSA-BP model in the Perovskite Database dataset were 0.0531, 0.0627, and 97.62%, respectively, which were lower than those of XGBoost by 0.0416, 0.0398, and -7.04%, respectively. The ISSA-BP model still had a high prediction performance in the more complicated Materials Project dataset, with an RMSE of only 0.0695. The comparison of the importance of each feature on the model prediction outcomes is shown in Figure 11.

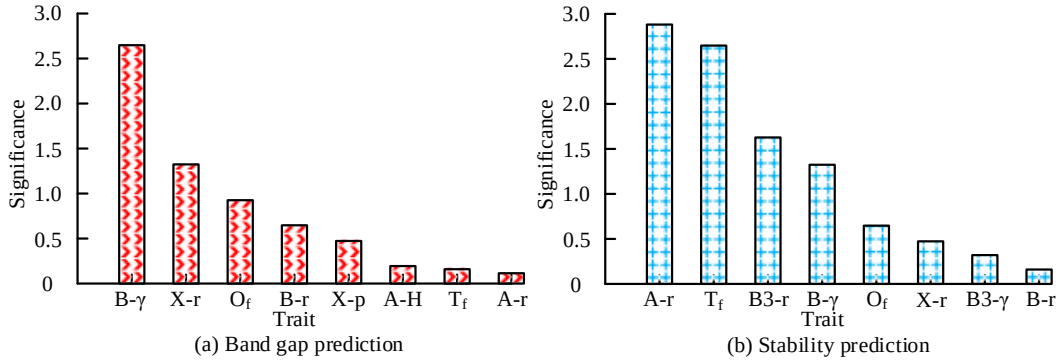


Figure 11: Comparison of the degree of influence of different features on band gap prediction of chalcogenide materials

In Figure 11 (a), B-γ denotes the electronegativity of the B ion, X-r denotes the radius of the X-site ion, $O_f$ denotes the octahedral factor, B-r denotes the radius of the B ion, X-p denotes the number of outermost orbital electrons, A-H denotes the ordinal number of the A-site ion, $T_f$ denotes the tolerance factor, A-r denotes the radius of the A-site ion, B3-r denotes the radius of the B3-site ion, and B3-γ denotes the electronegativity of the B3-site ion. The electronegativity of the B atom had the greatest impact on the predicted bandgap of the model, with an importance of 2.62. The X-site ion radius, octahedral factor, B-site ion radius, and outermost orbital electrons had importance above 0.5. In Figure 11 (b), those with an importance of 2.5 or above included the A-site ion radius and tolerance factor, while those with an importance of 1 or above included the B3 ion radius and B atom electronegativity. The predicted bandgap and stability of different perovskite samples by the model are shown in Table 5.

Table 5: Bandgap prediction results of different perovskite samples

| Sample chemical formula | pbe_bandgap (eV) | ml_bandgap (eV) | Error value (eV) | pbe_Ehull (eV/atom) | ml_Ehull (eV/atom) | Error value (eV/atom) |
|---|---|---|---|---|---|---|
| $K_2NaTiI_6$ | 1.378 | 1.374 | 0.004 | 0.086 | 0.094 | -0.008 |
| $Rb_2LiIrI_6$ | 1.384 | 1.391 | -0.007 | 0.028 | 0.034 | -0.006 |
| $Rb_2NaAmI_6$ | 1.351 | 1.372 | -0.021 | 0.021 | 0.027 | -0.006 |
| $Cs_2LiSbBr_6$ | 1.358 | 1.365 | -0.007 | 0.023 | 0.025 | -0.002 |
| $Cs_2NaCoI_6$ | 1.304 | 1.351 | -0.047 | 0.042 | 0.057 | -0.015 |
| $Cs_2TlAsBr_6$ | 1.317 | 1.308 | 0.009 | 0 | 0.012 | -0.012 |

In Table 5, pbe_bandgap (eV) denotes the value of material bandgap calculated by density flood theory, i.e., the actual bandgap value, ml_bandgap (eV) denotes the value of material bandgap predicted by the model,

pbe_Ehull (eV/atom) denotes the value of energy of the material above the convex packet, i.e., it denotes the stability of the material calculated by density flood theory, ml _Ehull (eV/atom) denotes the energy value of the material energy above the convex packet obtained by model prediction. The ISSA-BP model had a predicted bandgap error of [-0.047, 0.009] for different $A_2B^{1+}B^{3+}X_6$ mineral samples, which was within the allowable range. According to relevant studies, the optimal bandgap for solar cell materials was 1.34 eV, so the best bandgap was between [1.3, 1.4]. The six selected $A_2B^{1+}B^{3+}X_6$ in the study all met the requirements. The evaluation index for material stability was usually expressed as Energy above the convex hull (Ehull), and relevant studies showed that Ehull values below 0.05 met the requirements. In the experiment, except for $K_2NaTiI_6$, all other samples met the requirements. A histogram of the model's predictions of bandgap and stability for different chalcocite samples is shown in Figure 12.

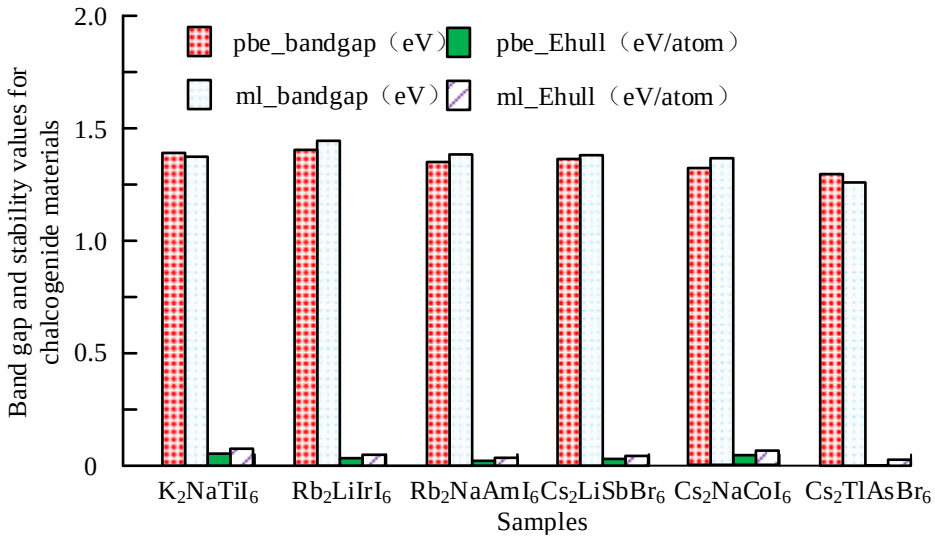


Figure 12: Histogram of model predictions of band gap and stability for different chalcocite samples

## 4 Discussion

This paper presented a prediction model of band gap and stability of perovskite materials based on improved SSA and improved ANN. The model was applied to the actual analysis of perovskite materials, and the validity of the prediction model was verified by relevant experimental analysis. Isa-bp model had faster convergence rate on both single-mode and multi-mode test functions, and the final convergence value was smaller than other methods, because the chaotic number generator and dynamic allocation strategy could effectively improve the search speed and search accuracy of the model in the early stage. Compared with the improved methods of Hu [7] and Zhao [8], this method could significantly improve the calculation speed while ensuring the prediction accuracy. The RMSE and coefficient of determination of band gap prediction of ISA-BP model were better than those of other methods, and the predicted value was more consistent with the real value, and the prediction accuracy was higher than that of ML model of Wu [9]. In cross-validation, the MAE and RMSE of stability prediction of ISA-BP model were lower than that of XGBoost by 0.0182 and 0.0154 respectively, which could meet the requirements related to prediction error. Compared with Selvaratnam [10], the proposed method could effectively improve the computational efficiency. The cost of screening new materials was reduced. ISSA-BP model could effectively improve the discovery speed of new perovskite materials through its powerful predictive analysis ability, and quickly identify material combinations with potential high performance. This will effectively promote the rapid development of new energy power generation technology.

## 5 Conclusion

A prediction model based on ANN $A_2B^{1+}B^{3+}X_6$ is proposed to address the high trial and error costs and time-consuming nature of perovskite new materials. The experiment showed that the optimal fitness values of improved SSA were 145, 128, and 53 lower than SSA, GBR, and XGBoost algorithms, respectively, and the optimization speed was faster. The running time was 0.035 s, 0.127 s, 0.022 s, and 1.212 s lower than the other three algorithms, respectively. In bandgap prediction, the MAE and RMSE distribution of the ISSA-BP model were 0.0217 and 0.0354 lower than those of the XGBoost model, and the coefficient of determination was 5.46% higher than that of the XGBoost model. The octahedral factor and tolerance factor showed a certain linear relationship with the bandgap, and were distributed between 0.3-0.6 $O_f$ and 0.9-1.2 $T_f$ in the test set, respectively. In stability prediction, the MAE and RMSE of the model were 0.0182 and 0.0154 lower than XGBoost, respectively, and the coefficient of determination was 1.42% higher than XGBoost. The B atom electronegativity and A-site ion radius had the greatest impact on the predicted bandgap and stability of the model, with importance values of 2.62 and 2.81, respectively. The six selected $A_2B^{1+}B^{3+}X_6$ minerals in the

study all met the requirements, with bandgap prediction errors between -0.047 and 0.009. Among them, the stability indicators of five materials met the requirements. There are still some problems in this research, for example, only the first principle was used in the validation to check the difference between the actual bandgap and stability of the material and the predicted value, and there was still some error between the bandgap value calculated by the first principle and the actual bandgap value, and the experimental synthesis can be added subsequently to enhance the persuasive power of the model. Meanwhile, the ISSA-BP model needs to be optimized and trained through many iterations, and the computational cost is high when dealing with large-scale datasets, and the fusion model can be simplified subsequently.

**List of abbreviations**

$t$ : Tolerance factor

$\mu$ : Octahedral factor

$N_V$ : Effective coordination number

$\overline{x}$ : Average key length

$\varphi$ : Nearest neighbor order parameter

$\gamma$ : Mirac function

$C_t$ : Concentration of t atoms in the structure

$\Delta_e$ : Local environmental differences

BPNN: Backpropagation Neural Network

$R$ : Discoverer and accession distribution ratio

$R_{\max}$ : Maximum distribution ratio

$R_{\min}$ : Minimum Distribution Ratio

$X_{tbest}$ : Discoverer searches for the optimal position

$cauchy(0,1)$ : Standard Cauchy distribution

$w$ : Nonlinear inertia weighting factors

SSA: Sparrow Search Algorithm

ISSA-BP: Improved Sparrow Search Algorithm Backpropagation

SVR: Support Vector Regression

XGBoost: Extreme Gradient Boosting

GBR: Gradient boosting regression

# Funding

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

# References

[1] De Angelis F. The impact of machine learning in energy materials research: the case of halide perovskites. ACS Energy Letters, 2023, 8(2): 1270-1272. DOI 10.1021/acsenergylett.3c00182

[2] Mannodi-Kanakkithodi A, Chan M K Y. Accelerated screening of functional atomic impurities in halide perovskites using high-throughput computations and machine learning. Journal of Materials Science, 2022, 57(23): 10736-10754. DOI 10.1007/s10853-022-06998-z

[3] Yang C, Chong X, Hu M, Yu W, He J, Zhang Y, Wang L W. Accelerating the discovery of hybrid perovskites with targeted band gaps via interpretable machine learning. ACS Applied Materials & Interfaces, 2023, 15(34): 40419-40427. DOI 10.1021/acsami.3c06392.s001

[4] Howard J M, Wang Q, Srivastava M, Gong T, Lee E, Abate A, Leite M S. Quantitative predictions of moisture-driven photoemission dynamics in metal halide perovskites via machine learning. The Journal of Physical Chemistry Letters, 2022, 13(9): 2254-2263. DOI 10.1021/acs.jpclett.2c00131.s001

[5] Biswas M, Desai R, Mannodi-Kanakkithodi A. Screening of novel halide perovskites for photocatalytic water splitting using multi-fidelity machine learning. Physical Chemistry Chemical Physics, 2024, 26(35): 23177-23188. DOI 10.1039/d4cp02330g

[6] Yang J, Mannodi-Kanakkithodi A. High-throughput computations and machine learning for halide perovskite discovery. MRS Bulletin, 2022, 47(9): 940-948. DOI 10.1557/s43577-022-00414-2

[7] Hu W, Zhang L, Pan Z. Designing two-dimensional halide perovskites based on high-throughput calculations and machine learning. ACS Applied Materials & Interfaces, 2022, 14(18): 21596-21604. DOI 10.1021/acsami.2c00564.s001

[8] Zhao J, Wang X. Screening perovskites from abo3 combinations generated by constraint satisfaction techniques using machine learning. ACS omega, 2022, 7(12): 10483-10491. DOI 10.1021/acsomega.2c00002

[9] Wu Y, Wang C F, Ju M G, Jia Q, Zhou Q, Lu S, Wang J. Universal machine learning aided synthesis approach of two-dimensional perovskites in a typical laboratory. Nature Communications, 2024, 15(1): 138-157. DOI 10.1038/s41467-023-44236-5

[10] Selvaratnam B, Oliynyk A O, Mar A. Interpretable machine learning in solid-state chemistry, with applications to perovskites, spinels, and rare-earth intermetallics: Finding descriptors using decision trees. Inorganic Chemistry, 2023, 62(28): 10865-10875. DOI 10.1021/acs.inorgchem.3c01153.s004

[11] Zhu Y, Zhang J, Qu Z, Jiang S, Liu Y, Wu Z, Dai Y. Accelerating stability of ABX3 perovskites analysis with machine learning. Ceramics International, 2024, 50(4): 6250-6258. DOI 10.1016/j.ceramint.2023.11.349

[12] Sharma S, Ward Z D, Bhimani K, Sharma M, Quinton J, Rhone T D, Koratkar N. Machine learning-aided band gap engineering of BaZrS3 chalcogenide perovskite. ACS Applied Materials & Interfaces, 2023, 15(15): 18962-18972. DOI 10.1021/acsami.3c00618.s001

[13] Mattur M N, Nagappan N, Rath S, Thomas T. Prediction of nature of band gap of perovskite oxides (ABO3) using a machine learning approach. Journal of Materiomics, 2022, 8(5): 937-948. DOI 10.1016/j.jmat.2022.04.006

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

[16] 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

[17] 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

[18] Zhang J, Li C, Yin Y, Zhang J, Grzegorzek M. Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer. Artificial Intelligence Review, 2023, 56(2): 1013-1070. DOI 10.1007/s10462-022-10192-7

[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