metals

MDPI

Article

Machine Learning Potential Model Based on Ensemble Bispectrum Feature Selection and Its Applicability Analysis

Jiawei Jiang [1], Li-Chun Xu [2,*], Fenglian Li [1,*] and Jianli Shao [3]

[1] College of Information and Computer, Taiyuan University of Technology, Jinzhong 030600, China
[2] College of Physics, Taiyuan University of Technology, Jinzhong 030600, China
[3] School of Mechatronic Engineer, Beijing Institute of Technology, Beijing 100084, China
* Correspondence: xulichun@tyut.edu.cn (L.-C.X.); lifenglian@tyut.edu.cn (F.L.)

Abstract: With the continuous improvement of machine learning methods, building the interatomic machine learning potential (MLP) based on the datasets from quantum mechanics calculations has become an effective technical approach to improving the accuracy of classical molecular dynamics simulation. The Spectral Neighbor Analysis Potential (SNAP) is one of the most commonly used machine learning potentials. It uses the bispectrum to encode the local environment of each atom in the lattice. The hyperparameter $j_{\max}$ controls the mapping complexity and precision between the local environment and the bispectrum descriptor. As the hyperparameter j_{max} increases, the description will become more accurate, but the number of parameters in the bispectrum descriptor will increase dramatically, increasing the computational complexity. In order to reduce the computational complexity without losing the computational accuracy, this paper proposes a two-level ensemble feature selection method (EFS) for a bispectrum descriptor, combining the perturbation method and the feature selector ensemble strategy. Based on the proposed method, the feature subset is selected from the original dataset of the bispectrum descriptor for building the dimension-reduced MLP. As a method application and validation, the data of Fe, Ni, Cu, Li, Mo, Si, and Ge metal elements are used to train the linear regression model based on SNAP for predicting these metals' atomic energies and forces them to evaluate the performance of the feature subsets. The experimental results show that, compared to the features of SNAP and qSNAP, the training complexity improvement of our EFS method on the qSNAP feature is more effective than SNAP. Compared with the existing methods, when the feature subset size is 0.7 times that of the original features, the proposed EFS method based on the SSWRP ensemble strategy can achieve the best performance in terms of stability, achieving an average stability of 0.94 across all datasets. The training complexity of the linear regression model is reduced by about half, and the prediction complexity is reduced by about 30%.

Keywords: machine learning potential; ensemble learning; feature select; descriptors

check for updates

Citation: Jiang, J.; Xu, L.-C.; Li, F.; Shao, J. Machine Learning Potential Model Based on Ensemble Bispectrum Feature Selection and Its Applicability Analysis. *Metals* **2023**, *13*, 169. https://doi.org/10.3390/met13010169

Academic Editor: Alain Pasturel

Received: 23 November 2022
Revised: 3 January 2023
Accepted: 11 January 2023
Published: 13 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, machine learning potential (MLP) [1–4] has been widely utilized in the fields of physics, chemistry, and materials science, and has gradually taken the place of the computational model paradigm. [5,6]. Instead of the complex mathematical model in the empirical potential function, MLP uses machine learning methods to establish the mapping relationship between material structure and properties. MLP has an accuracy close to density functional theory (DFT) and enables large-scale molecular dynamic simulations. It has made many breakthroughs in predicting the structure and evolution process of different elements [7,8]. The conventional calculation process of MLP can be divided into three steps [9,10]: Firstly, the local environment of atoms is encoded into atomic feature vectors using descriptors; secondly, the atomic feature vector is mapped to atomic energy using a regression model. Finally, the total energy of the structure is obtained by accumulating the atomic energy. Descriptors are key to establishing the effective machine learning potential.

At present, a variety of descriptors have been proposed, such as the Atom-Centered Symmetry Functions (ACSFs) [11], Smooth Overlap of Atomic Positions (SOAP) [12] kernel, and the rotational harmonic tensor [13]. Among them, ACSFs are composed of many functions with different parameters, which can be used for Behler–Parrinello neural networks [14], but they easily lead to redundancy because of their linear correlation. The Gaussian approximation potential [15] builds a regression model based on the Gaussian process and uses the SOAP kernel as the descriptor, which usually contains thousands of elements and incurs enormous computing costs. The moment tensor potential based on linear regression was proposed by Shapeev et al., which uses the rotation covariant tensor to describe the local environment of atoms [16]. The number of rotationally covariant tensors will sharply increase with the increase in the system size, and the calculation cost is expensive. The Spectral Neighbor Analysis Potential (SNAP) can be based on linear regression [17]. The local environment of each atom is characterized by a set of bispectrum descriptors, but there is a correlation between the descriptors, so it contains redundant information. The Quadratic Spectral Neighbor Analysis Potential (qSNAP) is a quadratic model based on SNAP, which contains more bispectrum descriptors and results in a greater computing complexity than SNAP. In summary, based on these typical descriptors, the size of the descriptors set is potentially enlarged for better defining the atomic structure, but at the expense of highly increased computational costs. The data dimension reduction strategy is an efficient way to reduce the expensive computational complexity caused by the sharp increase in the number of elements in the descriptor features set.

The high feature dimension of descriptors directly affects the computational complexity of machine learning models. The high-dimensional features of descriptors generally have the characteristics of sparseness, linear correlation, and redundancy. Reducing redundant information while ensuring its performance is an effective means to reduce computing costs. Feature selection, as a feature engineering, can be used to reduce redundant features. The purpose of this paper is to reduce descriptor redundancy using the feature selection method. That is, the problem of reducing the redundancy of descriptors is transformed into feature engineering, and the feature selection method is studied. The feature subset of descriptors that has a high correlation with the atom's performance is selected for the subsequent prediction model, which efficiently reduces the computational complexity during the training of the model without losing the prediction performance of the MLP model.

Feature extraction [18] and feature selection [19–21] are the two major techniques for feature reduction in machine learning. Feature extraction transforms the original feature space into another one. While feature selection acquires a subset from the original feature set. A subset of descriptor features set with a high correlation to the target can be chosen using feature selection, which could lower computing complexity while maintaining the prediction performance. In general, feature selection methods include the filter, wrapper, and embedded methods. The filter method is more suitable for high-dimensional data that selects the optimal subset by calculating the value of the features and sorting them. The computational performance of filtering methods is much better than the wrapper and embedded methods. Therefore, this paper studies a descriptor feature selection method based on the filter method.

Material attribute reduction has used a feature selection strategy. For instance, Xia Jun Fan et al. [22] conducted an ACSFs feature selection based on the Behler–Parrinello neural network using the Pearson correlation method. However, the performance of the feature subset obtained by this method was unstable and is easy to fall into local optimum. Imablzano et al. used feature selection methods such as the CUR decomposition, the Pearson correlation coefficient method, and the farthest point sampling to sparse the training set [23,24], but this method focuses on selecting training samples to reduce training costs. Recently, Li Wei et al. [21] applied the two-step feature recursive elimination method to the feature selection of inorganic magnetic materials, which uses cross-validation to evaluate the feature performances one by one, with a relatively high computational complexity. Ensemble learning [25–27] is a machine learning method that uses a series of base learn-

ers and aggregates the results of each base learner through an appropriate aggregation strategy to obtain a better performance. The improvement of existing ensemble methods is also the focus of current research. For example, Ivan Izonin et al. [25,26] improved an ensemble learner of two general regression neural networks, which used an extended-input successive geometric transformations model with a neural-like structure. The ensemble feature selection method combines ensemble learning [28] and feature selection [29] and is an efficient method that helps to obtain a stable feature subset with a good performance.

The highlights of this paper include:

1. A two-level ensemble feature selection (EFS) method is proposed to reduce the redundancy of the atomic bispectrum descriptor feature to obtain the feature subset of descriptors for predicting their energies and force.

2. The Stability Square Weight Rank Product (SSWRP) ensemble strategy is proposed to enhance the stability of selected feature subset.

3. The data of Fe, Ni, Cu, Li, Mo, Si, and Ge metal elements are used to obtain their feature sets of the qSNAP bispectrum descriptor, and the linear regression model is used to predict the atomic energies and force of the feature subset.

The rest of this paper is organized as follows: in Section 2, we introduce bispectrum and SNAP formalism. In Section 3, a two-level ensemble EFS method based on data perturbation and function perturbation is introduced, which combines the advantages of GBDT, RF, and Pearson, enhancing the stability and confidence of the results. In the EFS method, an ensemble strategy SSWRP is presented to enhance the stability. In Section 4, the popular SNAP and qSNAP are applied to the EFS method to reduce the feature dimension of bispectrum descriptors. Six different scenarios are designed to evaluate the prediction performance and computational complexity of the proposed EFS method. In Section 5 we discuss the result obtained. In Section 6, our conclusion is given.

2. Bispectrum and SNAP Formalism

The power spectrum [30], the Fourier transform of the autocorrelation function, quantifies how much energy the signal has in each frequency band, which should be invariant to translations of the signal intuitively. Unfortunately, in computing the spectrum, we lose all phase information. The bispectrum is a generalization of the power spectrum; the idea behind the bispectrum is a triple correlation to obtain a border set of invariants through the coupling of different angular momentum channels [17,31].

In earlier studies, the bispectrum and SNAP formalisms have been extensively discussed [32,33]. We will only provide a summary of the key concept here: The basic idea of the bispectrum is to map a 3D local atomic neighbor density into a set of coefficients that satisfy the invariant properties. The atomic neighbor density around atom i at location r is expressed as

$$\rho_i(r) = \delta(r) + \sum_{r_{ii'} < R_c} f_c(r_{ii'}) w_{i'} \delta(r - r_{ii'}), \tag{1}$$

where $w_{i'}$ is the dimensionless weight to distinguish the atom types, and i' denotes a neighbor atom. The weight is set to be one in this work as only one element is present. The cutoff function $f_c(r)$ ensures that the neighbor atomic density goes smoothly to zero when the distance $r_{ii'}$ is greater than the cutoff radius R_c.

The angular information in the 3D local density function can be projected onto spherical harmonic functions $Y_m^l(\theta, \phi)$. In the bispectrum approach, the radial component is converted into a third polar angle, defined by $\theta_0 = \theta_0^{\max} \frac{r}{R_c}$. Thus, the density function can be represented in the 3-sphere (θ, φ, θ_0) coordinates instead of (θ, φ, r). The density function defined on the 3-sphere can then be expanded using 4D hyperspherical harmonics as follows:

$$\rho(r) = \sum_{j=0,\frac{1}{2},\ldots}^{\infty} \sum_{m=-j}^{j} \sum_{m'=-j}^{j} u_{m,m'}^j U_{m,m'}^j(\theta, \phi, \theta_0), \tag{2}$$

where the coefficients $u^j_{m,m'}$ are obtained as the inner products between the density function and the basis, given by the following:

$$u^j_{m,m'} = U^j_{m,m'}(0,0,0) + \sum_{r_{ii'}<R_c} f_c(r_{ii'}) w_{i'} U^j_{m,m'}(\theta,\phi,\theta_0). \tag{3}$$

The bispectrum coefficients $B_{j_1,j_2,j}$ can then be obtained via the following equation:

$$B_{j_1,j_2,j} = \sum_{m_1,m_1'=-j_1}^{j_1} \sum_{m_2,m_2'=-j_2}^{j_2} \sum_{m,m'=-j}^{j} \left(u^j_{m,m'}\right)^* \times H^{jmm'}_{j_2 m_2 m_2'}{}^{j_1 m_1 m_1'} \; u^{j_1}_{m_1,m_1'} u^{j_2}_{m_2,m_2'}, \tag{4}$$

where the constants $H^{jmm'}_{j_2 m_2 m_2'}{}^{j_1 m_1 m_1'}$ are coupling coefficients and $\|j_1 - j_2\| \le j \le \|j_1 + j_2\|$, j_{max} limits these indices j_1, j_2, and j.

In the SNAP formula [33], the energy E_{SNAP} and force F^j_{SNAP} are related to the bispectrum coefficients B by the following equations:

$$E_{SNAP} = \beta_0 + \beta \cdot \sum_{i=1}^{N} B^i, \tag{5}$$

$$F^j_{SNAP} = -\beta \cdot \sum_{i=1}^{N} \frac{\partial B^i}{\partial r_j}, \tag{6}$$

where β_0 and the vector β are the coefficients in the linear model, which are fitted by the DFT data, and E_{SNAP} and F_{SNAP} are associated with the structural bispectral coefficients B and their derivatives $\frac{\partial B}{\partial r}$.

3. Two-Level Ensemble Feature Selection (EFS) Method of Bispectrum Descriptors

In Figure 1, we present an overview mapping of EFS-qSNAP from its structure and attributes (energy and force). First, the local environment of atoms in the structure is mapped to a bispectrum descriptor set, and then the dimension of the bispectrum descriptor is reduced and the feature subset of the bispectrum descriptor is obtained using ensemble learning and feature selection. Finally, the training and prediction of the linear regression model are conducted based on the feature subset obtained.

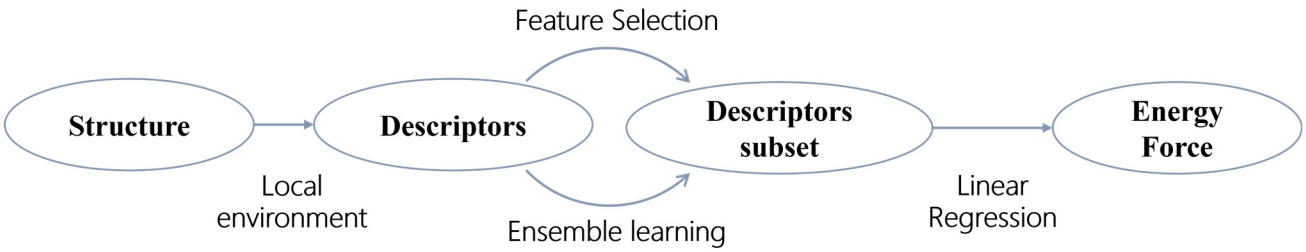

Figure 1. The mapping of EFS-qSNAP from crystal structure to its energy and force.

Theoretically, to predict energy and force directly based on a bispectrum, the descriptor feature set can achieve optimal prediction performance, but the computational complexity will increase sharply with the increase in descriptor feature dimension. Therefore, this paper proposes a feature selection method to obtain a subset of descriptors with better performance, simplify the training complexity of the model and reduce the computational cost. Assume that the number of atoms in a structure is d, and the atoms location information is $\{r_1, r_2, ..., r_d\}$. Map them to the bispectrum descriptor feature set $Fea_i = \{f_{ik}\}$, $i = 1, ..., d$, $k = 1, ..., K$, where K is the number of bispectrum descriptor features corresponding to the

analyzed material, and j_{max} limits the size of K. In SNAP, $K_{SNAP} = (j_{max} + 1)(j_{max} + 2)(j_{max} + 3/2)/3$; In qSNAP, $K_{qSNAP} = K_{qNAP}(K_{qNAP} + 1)/2$. The EFS method is used to select features from Fea_i, $i = 1, ..., d$, to obtain the subset $Fea_i' = \{f_{ik}'\}$, $k' = 1, ..., K'$, and $K' < K$. Then, the linear regression model is constructed to obtain the atomic energy $\{\varepsilon(r_1), \ldots, \varepsilon(r_d)\}$. Add these energies together to achieve the total energy of structure.

3.1. The Framework of the Two-Level Ensemble Feature Selection Method

Figure 2 shows the framework of the two-level EFS method. The proposed two-level EFS method fuses the feature weights of different datasets at level 1 and the feature weights of different feature selectors at level 2. The bootstrap method is used to realize data perturbation for obtaining different datasets based on the original data. Three kinds of different feature selectors are constructed as base feature learners. They are GBDT (Gradient Boosting Decision Tree), RF (Random Forest), and the Pearson correlation coefficient method. Both GBDT and RF are ensemble models based on a decision tree, but their data sampling strategies are different. GBDT is based on a boosting framework, which helps to improve the model prediction performance. Based on the improved bagging framework, RF randomly samples the input data and features to obtain feature subsets as training data, which helps to enhance the generalization ability of the model. The Pearson correlation coefficient method is a measure based on linear correlation. There is no dependency between these three models. This is a benefit for further improving the generalization performance of these models and it helps to enhance the stability of the selected features.

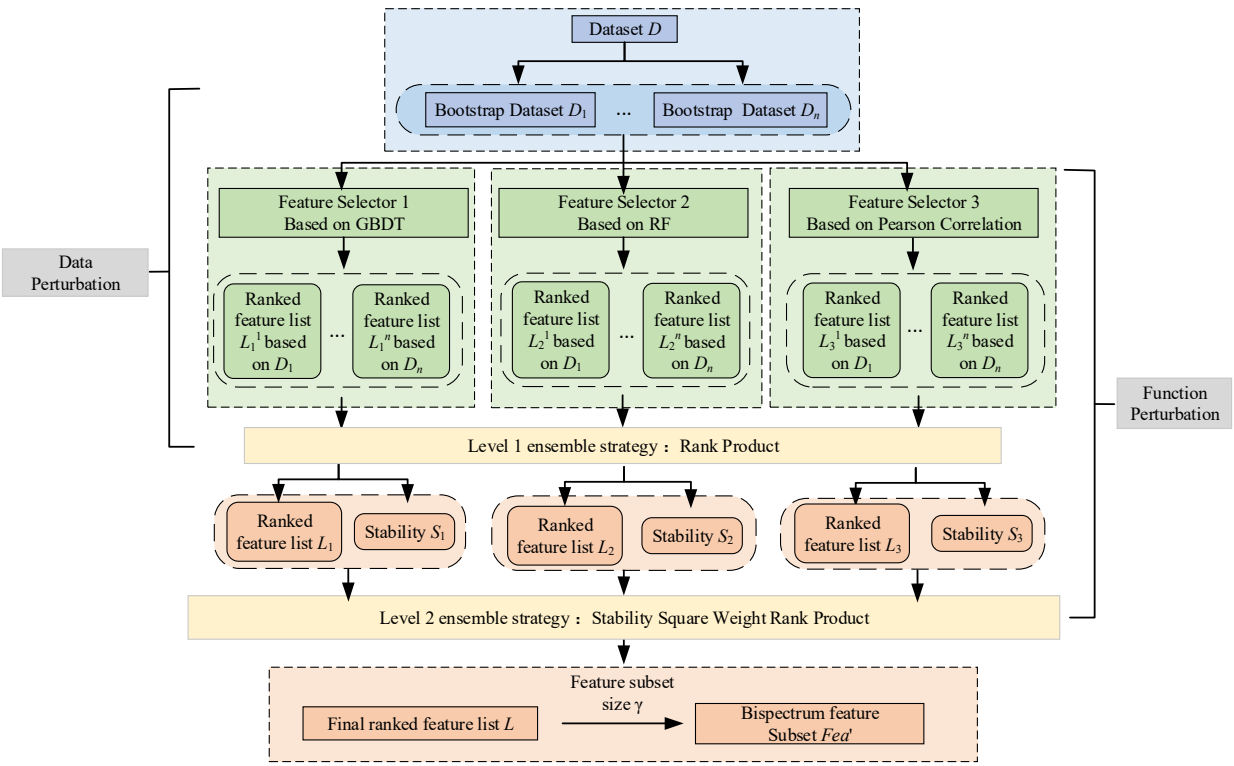

Figure 2. The framework of two-level EFS method.

Firstly, n bootstrap datasets D_1~D_n are generated as training data using data perturbation, and then n feature importance sequences L_m^l are obtained using three base feature selection models, where $m = 1, 2$, and 3 is the number of the base feature selector, $l = 1, ...,$ and n. Then, using the idea of function perturbation, a rank product (RP) strategy [28] is used to fuse the n feature sequences corresponding to all the feature selectors to obtain the important sequence list L_m containing each feature's importance. Moreover, in order

to improve the stability of a selected feature [29,30], the similarity of the feature subsets corresponding to each feature selection method under different datasets is further used as a weight. Here, the Stability Square Weight Rank Product (SSWRP) ensemble strategy is used to give a higher weight value to the feature selector with higher stability. Finally, according to the final feature importance value, sort the bispectrum descriptor features, and set the feature subset size parameter γ to obtain the feature subset Fea_i' as the input of the following linear regression model.

3.2. Data Perturbation and Function Perturbation

The bootstrap [34] method is introduced to realize data perturbation to obtain a stable and effective feature subset of the bispectrum descriptor, (q)SNAP. The bispectrum descriptor dataset D is taken as the original training set, and the bootstrap method is used to randomly select 70% of the structures in dataset D to generate n training subsets D_1~Dn, which are, respectively, used as the training data for the three feature selectors. n represents the index of the training subsets. In this paper, we set n to be 20. Based on the GBDT model, RF model, and the Pearson correlation coefficient method, the EFS method, which ensembles linear description and nonlinear description, is proposed, and three feature selectors 1, 2, and 3 are obtained, respectively. The feature importance analysis is carried out based on different mechanisms to realize the linear and nonlinear description of features and obtain more stable feature subsets. During the training of the model, this paper takes the bispectrum descriptor vector Fea_{i} of all the atoms in the analyzed structure as the input and sets the energy and force of the atomic structure obtained based on DFT, as the true values y and $y_predict$ represent the predicted value of each model.

Both GBDT [35–37] and RF [38,39] are ensemble models based on classification and the regression tree algorithm. This paper uses these two models as basic learners for feature selection. The basic idea of the feature importance calculation based on the GBDT and RF model was as follows: firstly, calculate the Gini reduction in the bispectrum descriptor feature f_k corresponding to the analyzed structure in a single decision tree [40], and then calculate the mean value of the Gini reduction in the bispectrum descriptor feature f_k in all decision trees to obtain the global importance V_{f_k} of the bispectrum descriptor feature f_k in the model, as shown in Equation (7) below.

$$V_{f_k} = \sum_{t \in T} \sum_{C \in c} V_{f_k,i,c}, \tag{7}$$

where $V_{f_k,i,c}$ represents the change of the Root Mean Square Error (RMSE) of the bispectrum descriptor feature f_k in the process of node c splitting in the decision tree i, c represents the set of all nodes in the decision tree t, and T represents the set of the decision tree in the ensemble model. After normalization, we can obtain the importance V_{f_k}' of feature f_k in selector one and selector two:

$$V_{f_k}' = V_{f_k} / \sum_{i=1}^{K} V_{f_k}. \tag{8}$$

Base feature selector one: the GBDT model. Set the learning rate of the GBDT model to be 0.1, the number of base learners of the model to be 100, the maximum depth of the tree to be five, the minimum number of samples contained by each non-leaf node to be two, and the minimum number of samples contained by each leaf node to be one.

Base feature selector two: the RF model. Set the number of base learners of the model to be 1000, the maximum depth of the tree to be five, the minimum number of samples contained in each non-leaf node to be two, and the minimum number of samples contained in each leaf node to be one. After training the model, the importance of all the bispectrum descriptor features of the analyzed material can be calculated according to Equations (7) and (8).

Base feature selector three: the Pearson correlation coefficient. The Pearson correlation coefficient [41,42] is usually used to measure the degree of correlation between two vectors, whose value range is [−1, 1]. If its value is greater than zero, it means that the two vectors are positively correlated; if its value is less than zero, the two vectors are negatively

correlated; if its value is equal to zero, there is no correlation between the two vectors. The Pearson correlation coefficient between the bispectrum descriptor feature f_k and the output variable y is defined as follows:

$$\rho_{f_k y} = \frac{Cov(f_k, y)}{\sqrt{Var(f_k)}\sqrt{Var(y)}}. \tag{9}$$

In Equation (9), *Var* represents the variance, and *Cov* (f_k, y) represents the covariance of the bispectrum descriptor features f_k and y. Normalize the absolute value of the Pearson correlation coefficient, then we obtain the Pearson correlation between the bispectrum descriptor feature f_k and y, named P_{f_k}. As Equation (10):

$$P_{f_k} = \left|\rho_{fy}\right| / \sum\nolimits_{f_k \in Fea} \left|\rho_{fy}\right|. \tag{10}$$

3.3. Two-Level Ensemble Strategy

First-level fusion strategy is used to aggregate the results of the data perturbation and obtain the feature importance sequence of each base feature selector. In this paper, the RP strategy is used to aggregate feature importance during data perturbation. For the base feature selector m, the importance of its feature f_k is calculated as follows (11):

$$R_{f_k,m} = \prod_{l=1}^{n} R^l_{f_k,m}. \tag{11}$$

where $R_{f_k,m}$ represents the importance of the bispectrum descriptor feature f_k in the base feature selector m, and $R^l_{f_k,m}$ represents the importance of the bispectrum descriptor feature f_k in the base feature selector m in the dataset D_l.

Second-level fusion strategy is used to aggregate the results of the function perturbation to obtain the ensemble feature importance of all bispectrum descriptor features. In this paper, the SSWRP strategy is proposed to aggregate the feature importance sequences L_1 to L_3. This strategy can make the feature sequences with higher stability S_n have greater weights. The ensemble importance of features based on the SSWRP strategy is defined in Equation (12):

$$R_{f_k} = \prod_{m=1}^{3} \left(R_{f_k,m}\right)^{(1-Sm)^2}, \tag{12}$$

$$R_{f_k} = \prod_{m=1}^{3} \left(R_{f_k,m}\right)^{1-Sm}. \tag{13}$$

where R_{fk} represents the ensemble importance of the bispectrum descriptor feature f_k, and S_m represents the feature stability of the base feature selector m.

To illustrate the performance of the proposed SSWRP feature importance strategy, further compare its performance with the existing Stability-Weighted Rank Product (SWRP). The definition of the SWRP strategy is shown in Equation (13). Comparing Equations (12) and (13), it can be seen that, when the feature importance and stability of the base feature selector are equal, the ensemble feature importance obtained by the proposed SSWRP ensemble strategy in this paper is greater than that of the SWRP [43]. Figure 3 shows the increasing trend comparison of ensemble importance R_{fk} with the increasing stability S_m and $R_{f_k,m}$, based on the SSWRP and SWRP strategies. Apparently, the ensemble importance R_{fk} obtained by the SSWRP is bigger than the SWRP strategy. Furthermore, Figure 4 shows the comparison of the ensemble importance multiplier obtained using the SSWRP and SWRP strategies when the feature important is 0.33 in Figure 4a and the stability is 0.75 in Figure 4b. It can be seen that, when the importance or stability of each basic feature selector is equal, the ensemble importance obtained by the SSWRP is bigger than that of the SWRP. When the stability of each base feature selector is close to one, the SSWRP

makes each multiplier close to one too, which effectively avoids the problem of overfitting feature selection. The experimental results show that, when the feature importance of the bispectrum descriptor redundant feature is close to zero, the SSWRP effectively improves the removal ability of these redundant features and can identify redundant features more effectively.

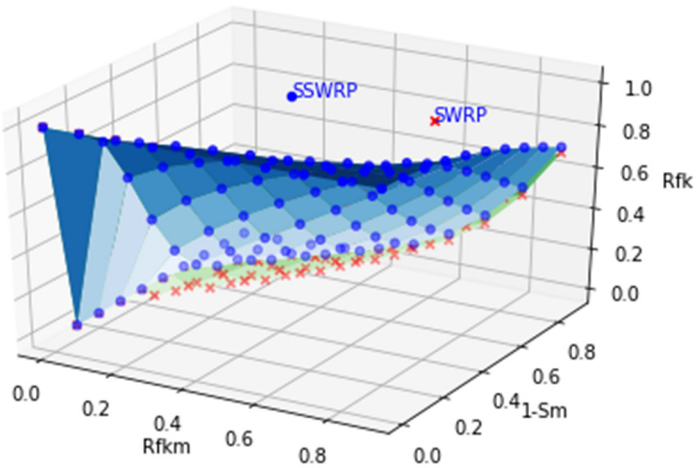

Figure 3. Variation of R_{fk} for SSWRP and SWRP.

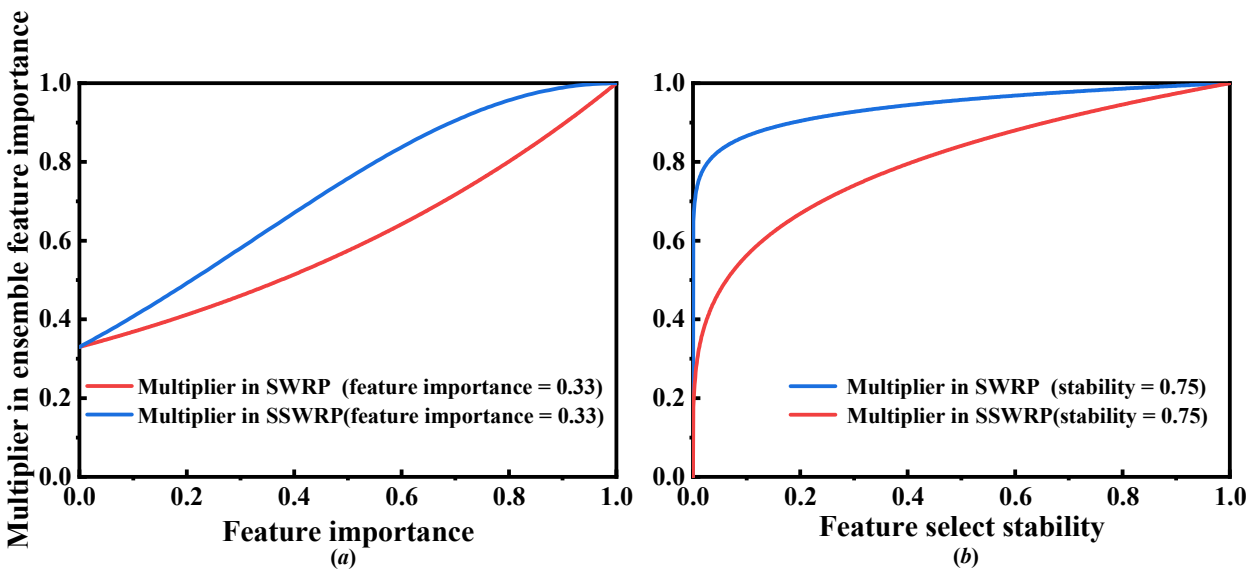

Figure 4. Multiplier comparison of ensemble feature importance between SSWRP and SWRP. (**a**) Feature important = 0.33. (**b**) stability = 0.75.

After obtaining the ensemble feature importance R_{fk} of the bispectrum descriptor, it is necessary to further determine the size of the feature subset. This paper introduces γ to measure the size of the feature subset. The relationship between the original feature scale K and the feature subset scale K' is as follows (14):

$$K' = \gamma K. \tag{14}$$

where γ is between [0, 1]. We can see that, the value of γ determines the size of the feature subset. Apparently, the bigger the value of γ, the bigger the size of the feature subset. In this paper, γ is set as *0.7*. If γ is 0.7, it means that 70% of the most important features were selected to form a feature subset.

3.4. EFS Method for LR Model Calculation Complexity Analysis

3.4.1. Model Training Complexity Analysis

In the linear regression model $Y = X\theta^T$ [44,45], K represents the feature dimension of the input matrix X; u represents the samples dimension; and θ represents the coefficient matrix. The output matrix Y can be expressed in Equation (15). The linear regression model's solution can be expressed in Equation (16).

$$X = \begin{bmatrix} x^{(0)} \\ x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(u)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_0^{(1)} & x_1^{(1)} & \dots & x_K^{(1)} \\ x_0^{(2)} & x_1^{(2)} & \dots & x_K^{(2)} \\ \dots & \dots & \dots & \dots \\ x_0^{(u)} & x_1^{(u)} & \dots & x_K^{(u)} \end{bmatrix}, \theta^T = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \dots \\ \theta_K \end{bmatrix}, Y = \begin{bmatrix} y^{(0)} \\ y^{(1)} \\ y^{(2)} \\ \dots \\ y^{(u)} \end{bmatrix}, \tag{15}$$

$$\theta^T = \left(X^T X\right)^{-1} X^T Y. \tag{16}$$

The calculation complexity of the linear regression model is $O(K^2(u + K))$ based on the feature dimension K.

In this paper, the EFS method reduces the original K-dimension bispectrum feature to the K'-dimension. When the number of samples u is far greater than K, the calculation complexity of the linear regression model on the feature subset is shown in Equation (17):

$$O(K'^2(u + K')) = O(\gamma^2 K^2(u + \gamma K) \approx \gamma^2 O(K^2(u + K)). \tag{17}$$

It means that the calculation complexity to train the linear regression model on the feature subset is $\gamma^2 O(K^2(u + K))$.

3.4.2. Model Prediction Complexity Analysis

The formula to predict energy and force based on the linear regression model is shown in Equation (18):

$$Y = X\theta^T. \tag{18}$$

In this paper, the EFS method reduces the original K-dimension bispectrum feature to the K'-dimension, so the complexity based on the linear regression model is shown in Equation (19):

$$O(uK') = O(\gamma uK) = \gamma O(uK). \tag{19}$$

The complexity of the model to predict energy and force is $\gamma O(uk)$ based on the feature dimension K'.

When γ is 0.7, the calculation complexity to train the linear regression model on the feature subset is $0.49O(K^2(u + K))$. It indicates that when the size of feature subset γ is set to be 0.7, the calculation complexity to train the linear regression model will be reduced by 51%, and the complexity of the model to predict energy and force will be reduced by 30%.

3.5. Stability Index and Evaluation Criteria

This paper uses stability [46] to measure the performance of the feature subset obtained by the feature selection method. Stability is defined in Equation (20):

$$S_m = \frac{\sum_{f \in F}(\omega(f)/n)}{|F|}. \tag{20}$$

where S_m represents the stability of the feature selector m, and n represents the number of feature importance sequences generated; F represents the set of all features and $|F|$ represents the number of features which appear in F; $\omega(f)$ represents the frequency of feature f appearing in n importance sequences. For all base feature selectors, the feature subset size γ is set to be 0.7 to calculate stability S_m.

Commonly used indicators to measure the regression performance include the RMSE (Root Mean Square Error), MSE (Mean Square Error), and MAE (Mean Absolute Error), etc. However, when we use the MSE as a prediction indicator, it is necessary to change the unit of energy and force. In addition, the MAE is difficult to use to distinguish the prediction performance in the context of minimizing prediction errors and avoiding outliers in this paper. After comprehensive consideration, the RMSE was selected as the prediction indicator in this paper, which is shown in Equation (21):

$$RMSE = \sqrt{\sum_{1}^{Z}(y - y_predict)^2/Z}. \tag{21}$$

where $y_predict$ represents the predicting value of atomic energy or its force of the test structure, y represents the value of atomic energy or force of the test structure based on DFT, and Z represents the number of test samples.

4. Experiments and Results Analysis

In this section, the proposed EFS method is used to analyze the performance of potential function qSNAP, mainly including prediction accuracy analysis, training efficiency analysis, material parameter prediction, and feature selection stability analysis. We first introduce the dataset used in this paper.

The experimental environment of this paper is: a 2-way Xeon E5-2620V2 (2.1 GHz/6-core) processor and a 128 GB (8 * 16 GB) DDR3-1333 MHz-1.35 V low-voltage memory operating environment. Based on the Lammps molecular dynamics simulation software architecture, it is implemented using the Python programming language.

4.1. Dataset Preparation

To investigate the applicability of the EFS method in this paper, this paper uses two types of metal datasets, including the public dataset in the published paper [32] and the self-built dataset. Table 1 shows the dataset information used in this paper in detail, which contains three self-built datasets of the Fe element (bcc, fcc, and hcp), and six public datasets of the Ni, Cu, Li, Mo, Si, and Ge elements. Each of the Fe elements contains 3000 structures in the training dataset and 1000 structures in the test dataset. The number of structures in the Ni, Cu, Li, Mo, Si, and Ge training datasets is 263, 243, 241, 194, 194, and 241, respectively, and the number of structures in the test datasets is 31, 32, 29, 23, 23, and 29, respectively. The feature dimension is 56 for all elements on SNAP, except for Mo, which is 30. Similarly, the feature dimension is 1509 for all elements on qSNAP, except for Mo, which is 465. The number of training samples varies from 30,455 to 289,545, depending on the structure of different metal elements. Accordingly, the number of test samples varies from 3589 to 96,515. The variation range of both the sample number and feature dimension is large, which can meet the analysis needs of this paper. The public data used in this work are published as an open source on Github (https://github.com/materialsvirtuallab/mlearn) accessed on 12 January 2022.

The self-built dataset relates to different crystal structures of the iron element, which mainly reflects the potential description of the transformation behavior of different structures based on the EFS feature selection method. This dataset is constructed using the ab initio molecular dynamics (AIMD) simulation for structure sampling. Based on three typical crystal structures of iron (bcc, fcc, and hcp), the initial configurations are obtained by cell expansion and perturbation of cell parameters and then the AIMD are implemented at 300K. The Schrödinger equation with Kohn–Sham approximate methods is used for every change of atomic coordinates and configurations in the simulation, to obtain accurate atomic energies and forces, and thus obtain the self-built dataset of the iron element.

Table 1. Dataset Information.

Datasets	Training Structure	Testing Structure	j_{max}	Feature Dimension (SNAP)	Feature Dimension (qSNAP)	Training Sample	Test Sample
bcc Fe	3000	1000	4	56	1509	146,265	48,755
fcc Fe	3000	1000	4	56	1509	289,545	96,515
hcp Fe	3000	1000	4	56	1509	145,530	48,510
Ni	263	31	4	56	1509	82,523	9505
Cu	243	32	4	56	1509	82,510	9565
Li	241	29	4	56	1509	34,969	3988
Mo	194	23	3	30	465	30,455	3589
Si	194	23	4	56	1509	39,913	4600
Ge	241	29	4	56	1509	42,444	4729

Figure 5 shows the energy vs. volume distribution for the Fe bcc, hcp, and fcc structures. It can be seen that, the energy shows a decreasing trend as the volume of the structure increases. Figure 6 shows the energy and volume distribution histograms for the Fe bcc, hcp, and fcc structures, where Figure 6a is the energy distribution histogram, and Figure 6b is the volume distribution histogram. Apparently, it is approximately normal distribution on both energy and volume for the Fe bcc, hcp, and fcc structures. At the same time, considering the influence of sample size and sample distribution on the performance of the model, the datasets of different sizes and different distributions are further adopted.

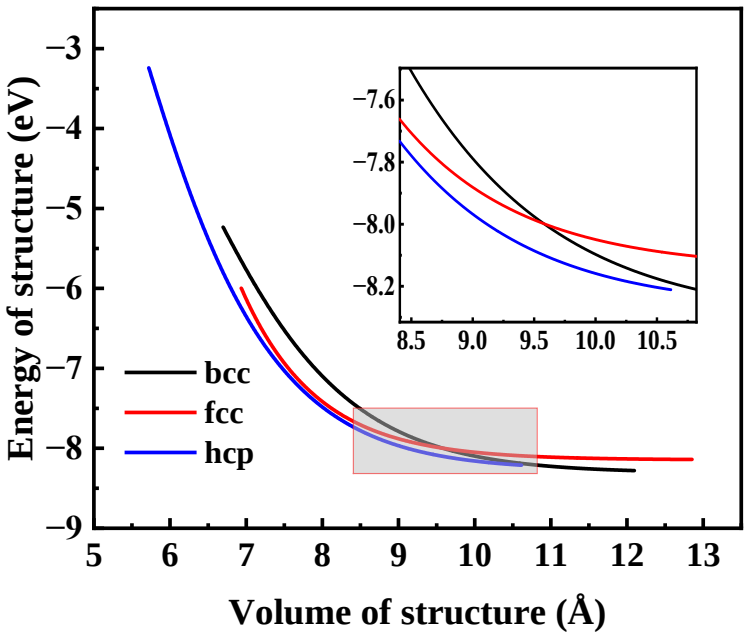

Figure 5. Energy vs. volume relations for bcc, hcp, and fcc Fe.

From the volume–energy relationship, it can be seen that the self-built dataset only involves the structures near the ground state of the potential energy surface. There are many structural samples in the low energy region, among which there are 2750, 3349, and 2418 bcc, fcc, and hcp structures in the [−8.5, −7.5] energy region, accounting for 68.75%, 83.72%, and 60.45% of each dataset, respectively. The amount of data in the structural samples in the high energy region is small; there are only 103 bcc structures in the energy range of [−4.5, −3.5], which conforms to the sampling characteristics of molecular dynamics at limited temperature.

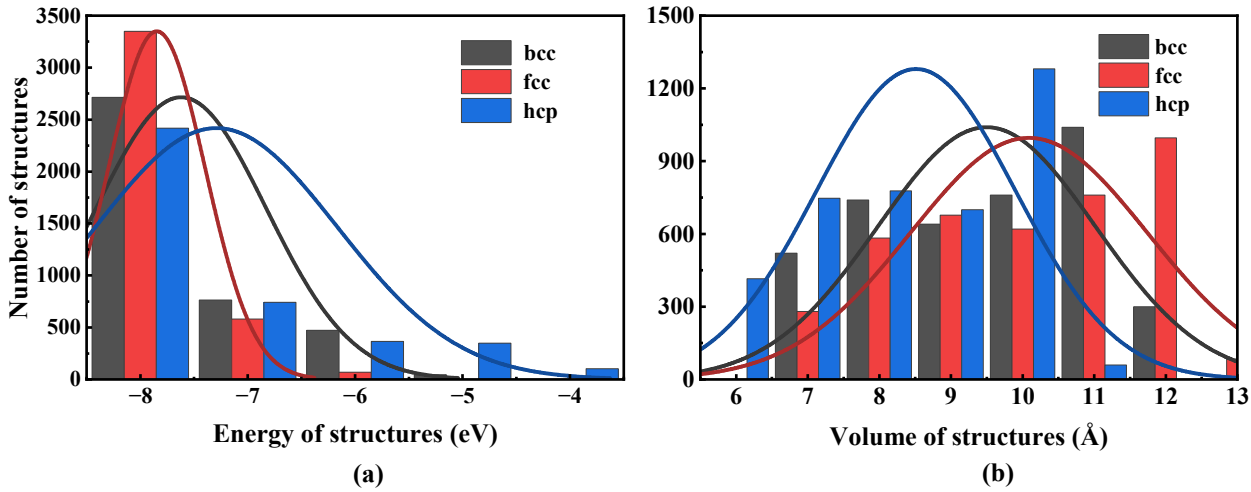

Figure 6. Distribution for bcc, hcp, and fcc Fe. (**a**) Energy distribution histogram; (**b**) volume distribution histogram.

For the public dataset, this paper uses the datasets provided in the literature [2], which cover six metal elements, namely Ni, Cu, Li, Mo, Si, and Ge. The sample size is approximately 1/10 of the self-built datasets. On datasets of different sizes, we can analyze the effect of the sample size and sample distribution on the model's performance. Despite the small size of the dataset (fewer than 300), the samples have rich and strong differential localized atomic features, including crystal ground state structures, strain structures, surface structures, structures at different temperatures, and defect structures with vacancies, which further increases the difficulty of feature selection.

4.2. Model Training and Parameter Optimization

The key to SNAP is to train a good linear regression model, in which, the sample weight is an important hyperparameter that affects the performance of the multivariate linear model, and it provides different weights to different samples [47]. In this paper, we set the sample weight of force to be one and set sample weight of energy to be ω_e, and its cost function is as follows:

$$J = \frac{1}{\omega_e \cdot S_{energy} + S_{force}} \left(\omega_e \sum_{1}^{S_{energy}} \left(y_{energy} - y\right)^2 + \sum_{1}^{S_{force}} \left(y_{force} - y\right)^2 \right). \tag{22}$$

where ω_e represents the sample weight of energy; S_{energy} and S_{force} represent the number of energy samples and force samples, respectively. y_{energy} and y_{force} represent the predicted value, respectively, y represents the true value.

The Bayesian optimization algorithm [48] is used to search for an optimal sample weight of energy and train the final potential model. The search space for the energy sample weight is set to be [50, 500], and the number of iterations is set to be 40.

4.3. Prediction Performance Analysis of the EFS Method

4.3.1. Scenario 1: Prediction Performance Comparison between qSNAP and EFS-qSNAP on Different Metal Datasets

This part mainly presents the RMSE comparison of prediction performance for energy and force between the EFS-qSNAP and qSNAP features. Here, EFS-qSNAP means the feature subset qSNAP, which is obtained based on the EFS method, and qSNAP means the original SNAP or qSNAP feature. Tables 2 and 3 give the RMSE comparison results for the energy and force.

Table 2. RMSE comparison of energy between qSNAP and EFS-qSNAP.

Dataset	RMSE of Energies/eV			
	SNAP	**EFS-SNAP**	**qSNAP**	**EFS-qSNAP**
bcc Fe	0.06	0.06 (0.0%)	0.03	0.03 (0.0%)
fcc Fe	0.22	0.23 (−4.5%)	0.08	0.07 (12.5%)
hcp Fe	0.11	0.11 (0.0%)	0.04	0.04 (0.0%)
Fe (bcc, fcc, and hcp)	0.23	0.25 (−8.7%)	0.10	0.10 (0.0%)
Ni	1.17	1.24 (−6.0%)	1.04	0.97 (6.7%)
Cu	0.87	0.93 (−6.7%)	1.16	1.14 (1.7%)
Li	1.31	1.32 (−0.8%)	0.85	0.86 (+1.2%)
Mo	9.06	9.17 (−1.2%)	3.96	3.94 (0.5%)
Si	8.06	8.14 (−1.0%)	6.28	5.75 (8.4%)
Ge	10.96	9.75 (+11.0%)	10.55	9.24 (12.4%)

Table 3. RMSE comparison of force between qSNAP and EFS-qSNAP.

Dataset	RMSE of Atomic Forces/eV/Å			
	SNAP	**EFS-SNAP**	**qSNAP**	**EFS-qSNAP**
bcc Fe	0.01	0.01 (0.0%)	0.01	0.01 (0.0%)
fcc Fe	0.17	0.18 (−5.9%)	0.14	0.13 (+7.1%)
hcp Fe	0.03	0.03 (0.0%)	0.02	0.02 (0.0%)
Fe (bcc, fcc, and hcp)	0.13	0.15 (−15.3%)	0.06	0.06 (0.0%)
Ni	0.08	0.08 (0.0%)	0.07	0.06 (+14.3%)
Cu	0.08	0.08 (0.0%)	0.05	0.05 (0.0%)
Li	0.04	0.04 (0.0%)	0.04	0.04 (0.0%)
Mo	0.37	0.40 (−8.1%)	0.33	0.33 (0.0%)
Si	0.34	0.39 (−16.6%)	0.29	0.26 (+10.3%)
Ge	0.29	0.30 (−3.4%)	0.20	0.18 (+10.0%)

As can be seen from Table 2, compared with SNAP, the change of energy RMSE of EFS-SNAP on 10 datasets is between −8.7% to 11%. The metal element Fe (bcc, fcc, and hcp) performs the worst. Its RMSE is about 8.7% larger than that of SNAP. The metal element Ge performs the best. Its RMSE is about 11.0% lower than that of SNAP. On the bcc Fe and hcp Fe datasets, both the energy and RMSE remain unchanged. In addition, the energy RMSE of EFS-qSNAP also shows a better performance than qSNAP for most datasets. The fcc Fe and Ge are 12.5% and 12.4%, respectively, which are the two optimal datasets. The prediction performance of EFS-qSNAP on bcc Fe, hcp Fe, and Fe (bcc, fcc, and hcp) datasets remains the same as qSNAP.

Table 3 shows that, the force RMSE of EFS-SNAP varies between −16.6% and 0% on the 10 datasets, compared to SNAP. For datasets of bcc Fe, Fe hcp, Ni and Cu, the force RMSE remains the same as SNAP. Fe (bcc, fcc, and hcp) and Si datasets are −15.3% and −16.6%, respectively, which are the two worst datasets. Similar to energy, EFS-qSNAP also outperforms qSNAP in terms of force RMSE on most datasets. Especially on the Ni dataset, the maximum performance improvement is obtained, which is 14.3%. EFS-qSNAP maintains the same force RMSE as qSNAP on bcc Fe, hcp Fe, Fe (bcc, fcc, and hcp), Cu, Li, and Mo datasets.

In summary, for energy and force, the EFS-qSNAP shows a better prediction performance than the qSNAP on most datasets. Compared with SNAP, the prediction performance of EFS-SNAP shows a descending trend on most datasets. The reason for this may be that the qSNAP feature dimension is 1509 on all datasets, except for Mo, which is 465. It's much more than SNAP, which is 56 on all datasets except for Mo, which is 30. This means that, there is more redundant information in the qSNAP features of these datasets. Compared to SNAP, feature selection is more suitable for qSNAP features. This achieves a good conclusion for the next application of material analysis.

4.3.2. Scenario 2: Prediction Performance Comparison of qSNAP and EFS-qSNAP on Fe Different Scale Training Structures

Figure 7 shows the RMSE trend comparison of the energy between EFS-qSNAP and qSNAP as the number of Fe training structure increases on bcc Fe, fcc Fe, hcp Fe, and Fe (bcc, fcc, and hcp). The training data size increases from 100 to 3000 Fe structures.

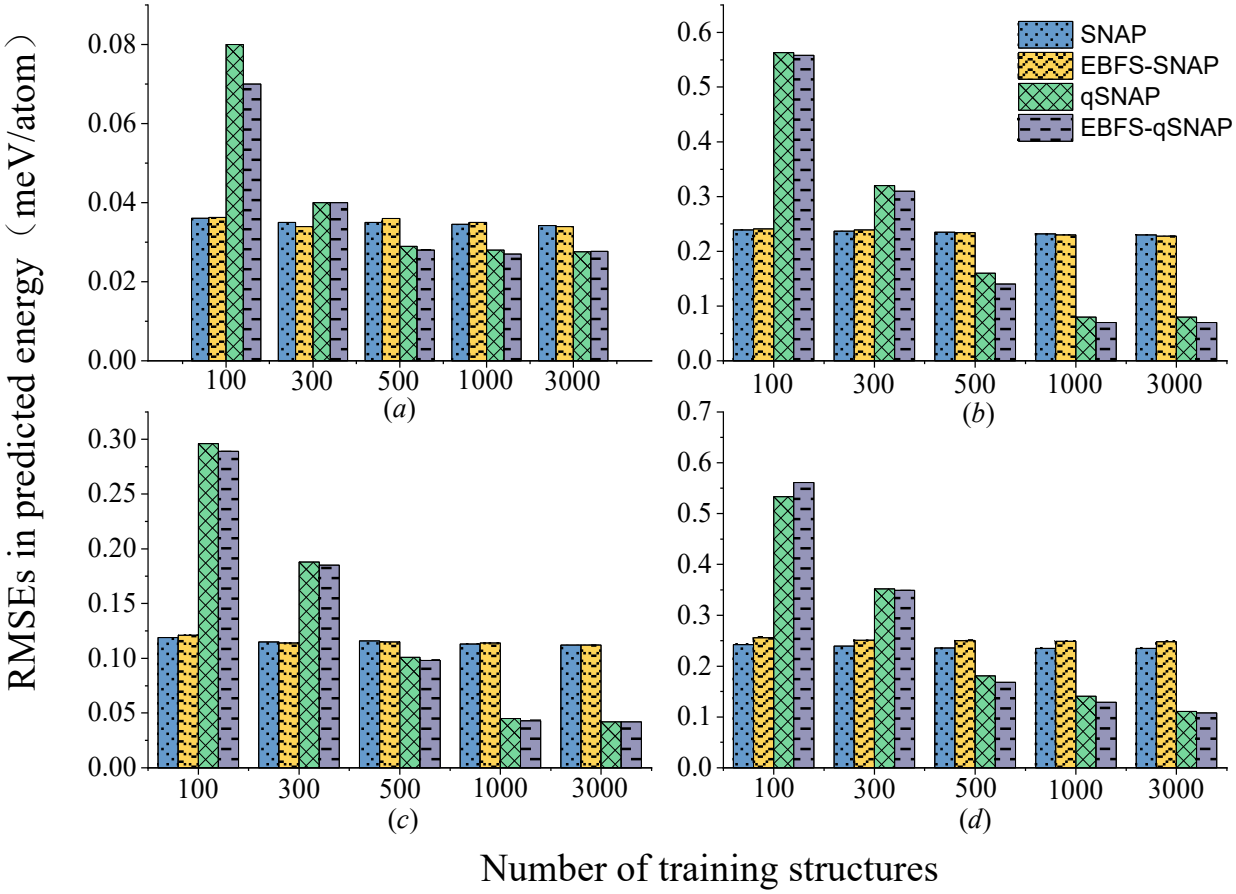

Figure 7. RMSE comparison of energy between EFS-qSNAP and qSNAP, with the number increase in Fe training structures (**a**) bcc Fe; (**b**) fcc Fe; (**c**) hcp Fe; and (**d**) bcc, fcc, and hcp Fe.

It can be seen that, as the number of Fe training structures increases, the prediction performance of SNAP for all Fe datasets hardly changes. Our EFS-SNAP method slightly improves the prediction performance for fcc Fe and hcp Fe, as shown in Figure 7b,c, while it degrades for Fe (bcc, fcc, and hcp), as shown in Figure 7d, but the performance changes for both cases are small. The reason for this may be that the feature dimension of SNAP for these datasets is 56, it always keeps less than the training samples for these datasets, even if the training structure is 100. After feature dimension reduction, this condition still meets. It further verifies that for the SNAP feature, the feature dimension reduction is not helpful for improving its prediction performance.

The prediction performance of qSNAP gradually improves as the number of sample training structures increases for these Fe datasets. The reason for this is that, for the qSNAP features, the number of training samples is even smaller than the feature dimension when the training structure number is 100, leading to a sharp drop in prediction performance. When the number of training dataset structures is 300 and the number of samples is close to the feature dimension of qSNAP, the prediction performance is better than 100. However, the data still have a certain degree of sparsity. The proposed EFS-qSNAP method has the best performance on all datasets with different training structures, except when the number of Fe (bcc, fcc, and hcp) structures is 100. The reason for this is that, with the number

increase in training dataset structures, the feature dimension of qSNAP is much less than the training sample number, so the model can be trained efficiently, which is helpful for maintaining good prediction performance. Our EFS method makes the feature dimension reduction in qSNAP further alleviate the redundant information, causing an improvement of the prediction performance.

Therefore, when we execute a high-dimension simulation, the qSNAP feature is more suitable for large-scale Fe datasets. In addition, our EFS method is more suitable for high-dimension qSNAP features and dimensionality reduction.

4.3.3. Scenario 3: Prediction Performance Comparison on Different Feature Select Methods

In order to verify the prediction performance of the feature subset selected by the EFS method, *t*-test, Lasso, Pearson, GBDT, and RF feature selection methods are used as benchmarks. The *t*-test focuses on describing the significance of features, and the Lasso method focuses on selecting features with multicollinearity. Tables 4–7 show the RMSE comparisons of energy and force prediction results for our EFS method and these benchmarks. The feature subset obtained by all methods uses the same size (γ = 0.7) and the linear regression is used as the prediction model.

Table 4. Energy RMSE comparison between EFS and benchmark methods on SNAP.

Dataset	RMSE of Atomic Energies/eV					
	***t*-Test**	**Lasso**	**Pearson**	**RF**	**GBDT**	**EFS**
bcc Fe	0.07	0.08	0.07	0.07	0.06	0.06
fcc Fe	0.26	0.25	0.26	0.22	0.22	0.23
hcp Fe	0.12	0.12	0.12	0.11	0.11	0.11
Fe (bcc, fcc, and hcp)	0.28	0.29	0.28	0.26	0.24	0.25
Ni	1.26	1.25	1.25	1.25	1.24	1.24
Cu	0.97	0.95	0.96	0.94	0.93	0.93
Li	1.35	1.35	1.34	1.31	1.32	1.32
Mo	9.17	9.17	9.20	9.21	9.23	9.17
Si	8.16	8.19	8.17	8.16	8.14	8.14
Ge	9.76	10.11	9.78	9.62	9.71	9.45

Table 5. Force RMSE comparison between EFS and benchmark methods on SNAP.

Dataset	RMSE of Atomic Forces/eV/Å					
	***t*-Test**	**Lasso**	**Pearson**	**RF**	**GBDT**	**EFS**
bcc Fe	0.01	0.01	0.01	0.01	0.01	0.01
fcc Fe	0.19	0.20	0.20	0.18	0.19	0.18
hcp Fe	0.03	0.04	0.03	0.03	0.03	0.03
Fe (bcc, fcc, and hcp)	0.20	0.19	0.18	0.17	0.18	0.16
Ni	0.09	0.09	0.09	0.09	0.08	0.08
Cu	0.10	0.09	0.08	0.09	0.08	0.08
Li	0.04	0.05	0.04	0.04	0.04	0.04
Mo	0.41	0.45	0.43	0.39	0.41	0.40
Si	0.41	0.43	0.42	0.39	0.41	0.39
Ge	0.34	0.33	0.30	0.31	0.33	0.30

Table 4 shows the RMSE comparison of energy for the SNAP potential based on the EFS and these benchmark methods. It can be seen that, for most datasets, the RMSE performance based on the feature subset of the EFS method maintains the best result, except for the sets fcc Fe, Fe (bcc, fcc, and hcp) that exhibit a small decline.

Table 5 shows the RMSE comparison of force for the SNAP potential based on the EFS and these benchmarks. Apparently, the RMSE performance of force based on the EFS method still remains the best for all datasets, except for the Mo set, which achieves a RMSE of 0.4, which slightly decreases compared to the RF method with an RMSE of 0.39.

Table 6. Energy RMSE comparison between EFS and benchmark methods on qSNAP.

Dataset	RMSE of Atomic Energies/eV					
	***t*-Test**	**Lasso**	**Pearson**	**RF**	**GBDT**	**EFS**
bcc Fe	0.03	0.04	0.03	0.03	0.03	0.03
fcc Fe	0.09	0.10	0.09	0.08	0.08	0.07
hcp Fe	0.04	0.05	0.04	0.04	0.04	0.04
Fe (bcc, fcc, and hcp)	0.12	0.11	0.12	0.09	0.11	0.10
Ni	1.07	1.11	1.09	1.00	1.01	0.97
Cu	1.18	1.21	0.19	1.14	1.16	1.14
Li	1.03	1.02	1.03	0.89	0.92	0.86
Mo	4.03	4.04	4.04	3.94	4.00	3.94
Si	5.69	5.70	5.66	5.47	5.37	5.45
Ge	9.17	9.19	9.12	8.98	8.91	8.86

Table 7. Force RMSE comparison between EFS and benchmark methods on qSNAP.

	RMSE of Atomic Forces/eV/Å					
	***t*-Test**	**Lasso**	**Pearson**	**RF**	**GBDT**	**EFS**
bcc Fe	0.01	0.01	0.01	0.01	0.01	0.01
fcc Fe	0.15	0.15	0.15	0.13	0.14	0.13
hcp Fe	0.02	0.02	0.02	0.02	0.02	0.02
Fe (bcc, fcc, and hcp)	0.07	0.07	0.07	0.06	0.06	0.06
Ni	0.07	0.08	0.08	0.08	0.06	0.06
Cu	0.07	0.07	0.07	0.06	0.06	0.05
Li	0.05	0.06	0.05	0.04	0.04	0.04
Mo	0.35	0.34	0.35	0.33	0.32	0.33
Si	0.20	0.20	0.17	0.19	0.18	0.18
Ge	0.11	0.12	0.11	0.10	0.10	0.10

Table 6 shows the RMSE comparison of energy for the qSNAP potential based on the EFS and these benchmarks. Expect for the datasets Fe (bcc, fcc, and hcp) and Si, the EFS method still provides the best RMSE of energy for all datasets. For the dataset Fe (bcc, fcc, and hcp), the RMSE of the EFS method is 0.10, which is slightly higher than that of the RF method, which is 0.09. For the dataset Si, the GBDT method provides the best energy RMSE of 5.37, but our EFS method achieves a higher energy RMSE of 5.45.

Table 7 shows the force RMSE comparison for qSNAP based on the EFS and these benchmarks. The RMSE of force obtained by our EFS method is generally at a lower level than most benchmarks for all datasets, except for datasets Mo and Si. This means that the force accuracy predicted by our EFS method is the best for most datasets. Therefore, the feature subset obtained by the EFS method has a better prediction performance than the single feature selection method in general.

In this scenario, the *t*-test, Lasso, Pearson, GBDT, and RF are used as benchmark feature selection methods to illustrate the difference in their prediction performance on their respective feature subsets. From the above results, it can be seen that the linear regression model based on the feature subset of the EFS method has the best prediction performance on most datasets. However, it only slightly improves or remains the same compared to these baseline methods.

4.3.4. Scenario 4: Comparison of Prediction Performance for Different Feature Subset Sizes

For illustrating the relationship between the feature subset size of the EFS method and the prediction performance on energy and force, we provide performance change trends with the increase in the feature subset size. Figure 8 shows the performance trend curves on these datasets. Here, we only provide the prediction performance on SNAP. The horizontal axis represents the proportion of the feature subset size to the original bispectrum descriptor feature set. The vertical axis is the relative RMSE of energy and

force. Relative RMSE means that the RMSE of each subset size is normalized with respect to the value at size 1.0. When γ = 1.0, it means that the feature subset size is equal to the feature dimension of the original bispectrum descriptor, that is, there is no dimension reduction. When γ = 0.3, it means that the feature subset size after feature selection is only 30% of the original bispectrum descriptor feature set. When γ is in the range of 1.0~0.7, the energy RMSE on most datasets generally shows a trend of steady change or slow rise. It indicates that when the feature subset scale is reduced to about 70%, the change in energy and force prediction performance is not significant. If γ is in the range of 0.7~0.3, the RMSE of energy and force tends to increase significantly, especially on the Cu, Mo, Si, and hcp Fe datasets. On the fcc Fe dataset, however, there are different laws. After its energy RMSE slowly increases with the size of the feature subset in the range of 1.0~0.5, the energy RMSE becomes smaller and smaller as the size of the feature subset continues to decrease. In addition, the RMSE corresponding to its force keeps a very gentle increasing trend in this process. It indicates that fcc Fe may have more serious feature redundancy. It needs to be verified in future work.

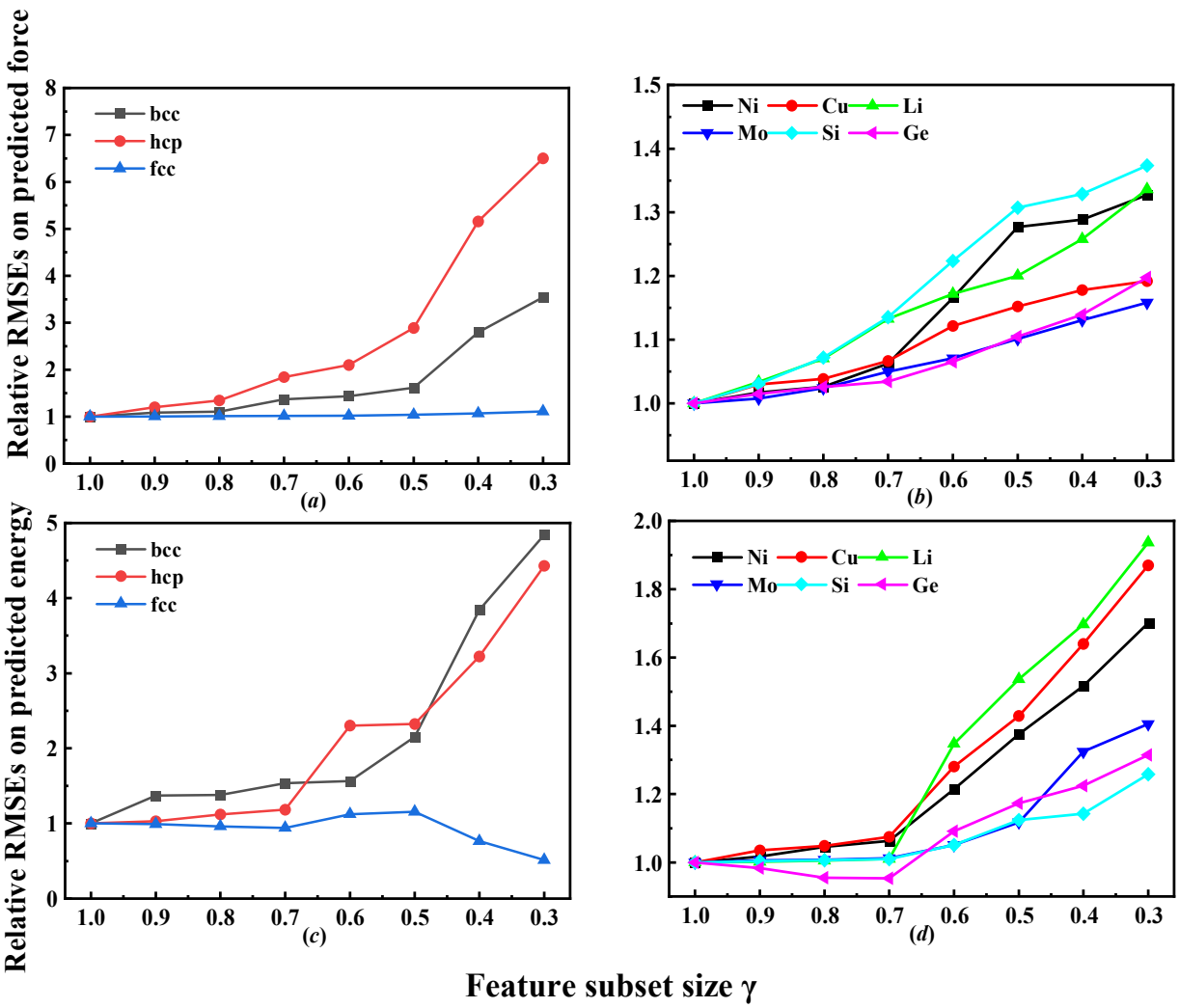

Figure 8. Trend analysis of the influence of feature subset size on prediction performance. (**a**) Force RMSE on Fe (bcc, hcp, and fcc); (**b**) force RMSE on Ni, Cu, Li, Mo, Si, and and Ge. (**c**) Energy RMSE on Fe (bcc, hcp, and fcc); (**d**) energy RMSE on Ni, Cu, Li, Mo, Si, and Ge.

In summary, when γ is set to be 0.7, the energy and force prediction performance can be guaranteed, which helps to reduce the computational complexity.

4.4. Stability and Complexity Analysis of the EFS Method

4.4.1. Stability Analysis

Scenario 5: Stability Comparison between the EFS and Benchmarks

Here, we compare the stability of the EFS method proposed in this paper with five benchmark feature selection methods, namely the *t*-test, Lasso, RF, GBDT, and the Pearson correlation method, respectively. For Fe element, the data of four structures obey normal distribution. The stability results of the proposed EFS method are close to those of benchmark methods. Therefore, only the stability comparisons of the Ni, Cu, Li, Mo, Si, and Ge datasets are presented here. The feature subset size γ is set to be 0.7, and their feature stabilities are calculated according to Equation (8).

Table 8 shows the stability S comparison between our EFS method and all benchmarks. We can see that the stability of our EFS method is more than 0.9 on all six datasets, maintaining the best result. RP+SWRP is an ensemble learning method, which is based on the same basic feature selectors as the EFS method but uses RP and SWRP ensemble strategies. The RP+SWAP method provides the second-best stability on six datasets. Then, the third is the stability of the Pearson method, which is better than that of the *t*-test. The Pearson method only considers the linear correlation between the bispectrum descriptor features of the input data and the variables to be predicted. If the bispectrum descriptor features do not obey normal distribution, or the relationship between the bispectrum descriptor features and variables to be predicted is nonlinear, this method may lead to the failure of the feature selection. The stability of the *t*-test is generally between 0.85 and 0.90 on most datasets, which is less than those of the EFS, RP+SWAP, and Pearson methods. The reason for this may be that the *t*-test only considers the significance of features and lacks the description of data details. The stability of the Lasso method is generally less than 0.87, which may be related to the regularization parameters of the Lasso method. The stability of the GBDT and RF methods has a large fluctuation range, and the stability of the GBDT method is the worst, with an average stability of 0.81, which is between 0.73 and 0.89. The reason for this may be that there are differences in the training datasets of each base learner, resulting in unstable feature subsets.

Table 8. Stability comparison between EFS and benchmarks.

Datasets	*t*-Test	Lasso	Base Feature Selector			RP+SWRP	EFS
			RF	GBDT	Pearson		
Ni	0.87	0.87	0.92	0.78	0.84	0.93	0.93
Cu	0.86	0.82	0.86	0.82	0.91	0.93	0.95
Li	0.83	0.86	0.84	0.89	0.90	0.93	0.95
Mo	0.88	0.73	0.79	0.73	0.90	0.89	0.91
Si	0.90	0.78	0.87	0.78	0.90	0.95	0.96
Ge	0.86	0.84	0.86	0.83	0.87	0.96	0.96
Average	0.88	0.82	0.86	0.81	0.89	0.93	0.94

The EFS method proposed in this paper combines the advantages of the three feature selectors and describes the features more comprehensively from linear and nonlinear aspects, respectively, making the stability much more than 0.9 on the six datasets, which reflects that the EFS method proposed in this paper has good stability in feature selection.

Scenario 6: Stability Comparison of the Different Feature Subset Size

In order to illustrate the changing trend of stability with the size of the feature subset, the Ni, Li, and Ge datasets are selected to show the relationship between them. Figure 9 provides the relationship between stability and feature subset size. The results show that when the size of the feature subset γ is in the range of 0.1 to 0.3, the stability of the GBDT method is higher than that of the other benchmark methods, but as γ increases, the stability of the GBDT method decreases on the three datasets. When the size of the feature subset

γ is in the range of 0.7–0.9, the RF method has higher stability among the benchmark methods. The Pearson method has better performances in all stages compared with the other benchmark methods for the Li and Ge datasets. Compared with the RP+ SWRP method, the EFS method in this paper has a larger stability when γ is in the range of 0.1 to 0.5, and so there is an obvious stability improvement. When γ is in the range of 0.6 to 1.0, the stability of the RP+SWRP method and the EFS method is very close on the Ni and Ge datasets, while the stability of the EFS method is better than that of the RP+SWRP method on the Li dataset.

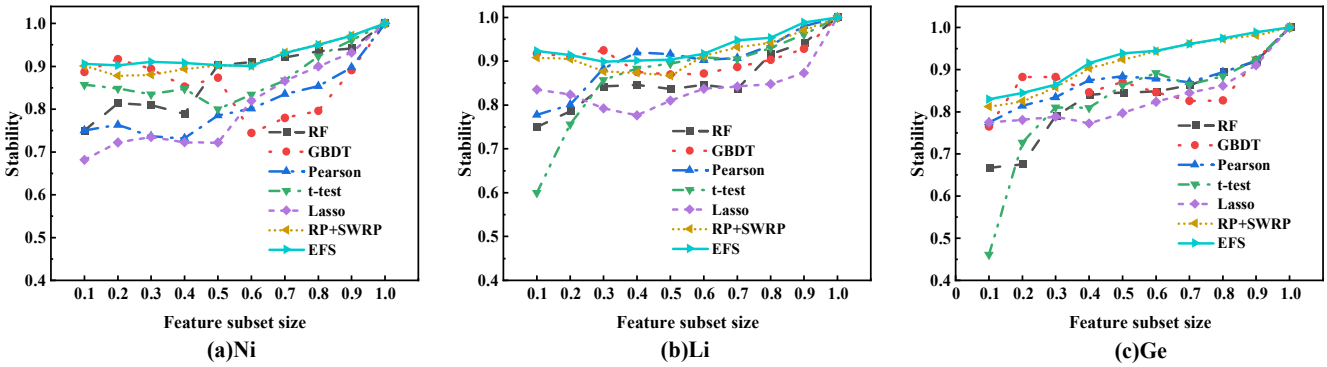

Figure 9. The relationship between stability and feature subset size.

In summary, compared with all the benchmark methods, the stability of the EFS method always remains at the maximum when the size of the feature subset γ is larger than 0.5 for the three datasets. The reason for this may be that the EFS is an ensemble method, which combines the advantages of three base feature selection methods. This enables our EFS method to have a greater feature subset selection ability.

4.4.2. Performance Analysis on the Computational Complexity of the EFS Method

The influence of the EFS feature selection method on the qSNAP model training time is given in Table 9. Where the qSNAP column represents the training time of the model when building a linear regression potential model based on the original bispectrum descriptor feature. The EFS-qSNAP column represents the training time for constructing the linear regression prediction model based on the bispectrum descriptor feature subset obtained by the proposed EFS method.

Table 9. Computational complexity comparison between qSNAP and EFS-qSNAP.

Dataset	Model Training Time/s			
	SNAP	EFS-SNAP	qSNAP	EFS-qSNAP
bcc Fe	1.94	1.39 (28.35%)	32.92	18.73 (43.10%)
fcc Fe	2.03	1.55 (23.65%)	41.52	26.95 (35.09%)
hcp Fe	1.89	1.36 (28.04%)	31.18	19.21 (38.39%)
Fe (bcc, fcc, and hcp)	7.62	5.12 (32.81%)	167.42	91.67 (45.25%)
Ni	0.35	0.22 (37.14%)	2.99	1.60 (46.49%)
Cu	0.30	0.15 (50.00%)	3.87	2.35 (39.28%)
Li	0.17	0.11 (35.29%)	1.72	0.97 (43.60%)
Mo	0.08	0.05 (37.50%)	1.98	1.16 (41.41%)
Si	0.10	0.06 (40.00%)	7.13	3.54 (50.35%)
Ge	0.15	0.12 (20.00%)	2.08	1.09 (47.60%)
Average improvement	-	33.28%	-	43.06%

The model training time based on the proposed EFS method is significantly reduced for all datasets. Compared with SNAP, the model training time of the EFS-SNAP feature subset

is reduced by 33.28% on average on all datasets. Especially for the Cu dataset, the training time of the model based on the EFS-SNAP feature subset is reduced by 50% compared with the SNAP feature subset. For the Si dataset, its model training time on the EFS-SNAP feature subset is 40% shorter than for SNAP. The smallest improvement is for the Ge dataset, whose model training time on the EFS-SNAP feature subset is reduced by 20% compared to SNAP.

Compared with qSNAP, the EFS-qSNAP feature subset reduces the model training time by 43.06% on average on all datasets. By comparing the reduction in training time of all datasets, fcc Fe is the least, which is reduced by 35.09%. The training time of the Si dataset was reduced the most, and its model training time on the EFS-qSNAP feature subset was reduced by 50.35%.

In summary, compared to the features of SNAP and qSNAP, the training complexity improvement of our EFS method on the qSNAP feature is more effective than SNAP. The reason may be that the dimension of the qSNAP feature is much higher than SNAP. Our EFS method is more suitable for feature reduction in high dimension bispectrum descriptor features.

5. Discussion

Feature selection is an effective method to obtain a feature subset by reducing redundant information, which is beneficial for further constructing prediction models with lower computational complexity. It is suitable for optimizing high-dimensional MLP in molecular dynamics simulations. For this purpose, a two-level ensemble feature selection method is proposed, and the above experiments are executed. The *t*-test, Lasso, RF, GBDT, and Pearson correlation feature selection methods and RP+ SWRP ensemble strategy are used in different scenarios as benchmarks. Based on the experimental results, we propose the following discussion, which can be used as a reference in subsequent research.

(1) In this paper, stability is used as a factor for obtaining the ensemble feature importance. For obtaining the feature subset size with good stability, we give the changing trend of stability with the size of the feature subset, on the Ni, Li, and Ge datasets, and compare the stability difference of the EFS method with the benchmark methods. The experimental results show that, compared with all the benchmark methods, the EFS method has the largest stability when the feature subset γ is greater than 0.5, which is conducive to the better feature subset selection ability of the EFS method. The stability of our EFS method is more than 0.9 on all six datasets when the feature subset γ is set to be 0.7, maintaining the best results compared with all the benchmark methods.

(2) In order to illustrate the relationship between the feature subset size of the EFS method and the prediction performance of energy and force, the trend of performance based on EFS-SNAP with the feature subset size increase is discussed. It can be seen that, when the feature subset size γ is in the range of 0.7 to 1.0, the prediction performance of energy and force on most datasets shows a trend of little change. It indicates that we can reduce the feature subset size to around 70%, which can reduce the computational complexity while guaranteeing the energy and force prediction performance. Further experimental results show that, on average, the model training time of the EFS-SNAP feature subset is reduced by 33.28%, and the model training time of the EFS-qSNAP feature subset is reduced by 43.06% on all datasets. Generally speaking, the EFS method can effectively reduce the training complexity of SNAP and qSNAP, and the training complexity improvement of the EFS method on the qSNAP features is more effective than the SNAP method.

(3) For illustrating the prediction performance of the EFS method, the SNAP and qSNAP bispectrum descriptors are used in the paper. The experimental results show that, for energy and force, EFS-qSNAP shows a better prediction performance than qSNAP on most datasets. Compared with SNAP, the prediction performance of EFS-SNAP shows a descending trend on most datasets. It means that there is more redundant information in the qSNAP features of these datasets. Compared to SNAP, feature selection is more

suitable for the qSNAP features. This achieves a good conclusion for the next application of material analysis.

(4) The prediction performance of our EFS method on a different number of Fe training structures is evaluated on the SNAP and qSNAP bispectrum descriptors. The results show that, with the increase in Fe training structures, the prediction performance of SNAP is almost unchanged for all Fe datasets. The prediction performance of our EFS-SNAP method is similar to that of SNAP, which means that there is almost no redundant information. Again, it is verified that for the SNAP features, feature dimensionality reduction does not help improve their prediction performance. In summary, compared to the feature of SNAP and qSNAP, the training complexity improvement of our EFS method on the qSNAP feature is more effective than SNAP. The reason for this may be that the dimension of the qSNAP feature is much higher than for SNAP. Our EFS method is more suitable for the feature reduction in the high dimension bispectrum descriptor feature.

(5) In order to illustrate the predictive performance of the EFS method compared to all benchmark methods from a feature selection perspective, we further present the predictive performance of our EFS method and other benchmark methods to SNAP. The experimental results show that, for energy and force, the EFS method shows a better prediction performance than all the benchmarks method on most datasets. This shows that the EFS method is more efficient in evaluating the bispectrum features and the feature subset obtained by this method can predict the energy and force more accurately.

6. Conclusions

This paper proposes a two-level ensemble feature selection method for effectively reducing the redundancy of the bispectrum descriptor feature based on the perspective of feature engineering. The idea of data perturbation and function perturbation is firstly introduced, and the GBDT model, RF model, and Pearson correlation coefficient method are used as the base feature selectors. A stability-weighted square sequence product aggregate strategy (SSWRP) is proposed for the ensemble feature selector, which is considered an important part of the feature selection method. The proposed ensemble bispectrum descriptor feature select method has excellent diversity and high stability, which can achieve the purpose of effective feature selection for the bispectrum descriptor, reduce information redundancy, and provide input data with lower dimensions for the prediction model. The experimental results show that the proposed EFS method can perform feature selection effectively and train the potential function model with comparable performance. Moreover, the proposed method can achieve effective feature reduction using the SNAP and qSNAP analysis without losing prediction accuracy. This indicates that the proposed EFS method can guarantee the material parameter prediction performance and is beneficial for reducing the computational complexity of the subsequent material simulation. When the feature subset size is 0.7 times that of the original features, the proposed EFS method based on the SSWRP ensemble strategy can achieve the best performance in terms of stability, achieving an average stability of 0.94 across all datasets. The training complexity of the linear regression model is reduced by about half, and the prediction complexity is reduced by about 30%. It should be noted that this paper only provides a preliminary explanation of the importance and correlation of the bispectrum descriptor to predict variables, and there is still a lack of an in-depth exploration of their relationship. The EFS lacks a theoretical derivation based on mathematics, and its practicability needs to be verified by a simulation of its physical processes.

Author Contributions: Conceptualization, L.-C.X. and F.L.; methodology, F.L., J.J. and J.S.; software, J.J. and L.-C.X.; validation, J.J., F.L. and L.-C.X.; writing—original draft preparation, J.J., F.L., L.-C.X. and J.S.; writing—review and editing, J.J., F.L. and L.-C.X. and J.S.; project administration, F.L. and J.S.; funding acquisition, L.-C.X. and F.L. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China-NSAF Grant Nos. U2030117 and partly by the National Natural Science Foundation of China Grant Nos. 62171307.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We acknowledge the administrative and technical support for this journal and all public datasets used in the paper for the experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Keith, J.A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816–9872. [CrossRef] [PubMed]
2. Mortazavi, B.; Novikov, I.S.; Podryabinkin, E.V.; Roche, S.; Rabczuk, T.; Shapeev, A.V.; Zhuang, X. Exploring phononic properties of two-dimensional materials using machine learning interatomic potentials. *Appl. Mater. Today* **2020**, *20*, 100685. [CrossRef]
3. Mueller, T.; Hernandez, A.; Wang, C. Machine learning for interatomic potential models. *J. Chem. Phys.* **2020**, *152*, 050902. [CrossRef] [PubMed]
4. Zuo, Y. Machine Learning Towards Large-Scale Atomistic Simulation and Materials Discovery. Ph.D. Thesis, University of California, San Diego, CA, USA, 2021.
5. Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **2021**, *6*, 655–678. [CrossRef]
6. Manna, S.; Loeffler, T.D.; Batra, R.; Banik, S.; Chan, H.; Varughese, B.; Sasikumar, K.; Sternberg, M.; Peterka, T.; Cherukara, M.J.; et al. Learning in continuous action space for developing high dimensional potential energy models. *Nat. Commun.* **2022**, *13*, 368. [CrossRef]
7. Botu, V.; Batra, R.; Chapman, J.; Ramprasad, R. Machine Learning Force Fields: Construction, Validation, and Outlook. *J. Phys. Chem. C* **2017**, *121*, 511–522. [CrossRef]
8. Deng, Z.; Chen, C.; Li, X.-G.; Ong, S.P. An electrostatic spectral neighbor analysis potential for lithium nitride. *npj Comput. Mater.* **2019**, *5*, 75. [CrossRef]
9. Bereau, T.; DiStasio, R.A.; Tkatchenko, A.; von Lilienfeld, O.A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706. [CrossRef]
10. Smith, J.S.; Isayev, O.; Roitberg, A.E. ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192–3203. [CrossRef]
11. Eckhoff, M.; Behler, J. High-dimensional neural network potentials for magnetic systems using spin-dependent atom-centered symmetry functions. *Npj Comput. Mater.* **2021**, *7*, 170. [CrossRef]
12. Caro, M.A. Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials. *Phy. Rev. B* **2019**, *100*, 024112. [CrossRef]
13. Novikov, I.; Grabowski, B.; Körmann, F.; Shapeev, A. Magnetic Moment Tensor Potentials for collinear spin-polarized materials reproduce different magnetic states of bcc Fe. *NPJ Comput. Mater.* **2022**, *8*, 13. [CrossRef]
14. Schran, C.; Uhl, F.; Behler, J.; Marx, D. High-dimensional neural network potentials for solvation: The case of protonated water clusters in helium. *J. Chem. Phys.* **2017**, *148*, 102310. [CrossRef] [PubMed]
15. Dragoni, D.; Daff, T.D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.* **2018**, *2*, 013808. [CrossRef]
16. Gubaev, K.; Podryabinkin, E.V.; Hart, G.L.W.; Shapeev, A.V. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Comput. Mater. Sci* **2019**, *156*, 148–156. [CrossRef]
17. Wood, M.A.; Thompson, A.P. Extending the accuracy of the SNAP interatomic potential form. *J. Chem. Phys.* **2018**, *148*, 241721. [CrossRef]
18. Taguchi, Y.H.; Turki, T. Principal component analysis- and tensor decomposition-based unsupervised feature extraction to select more reasonable differentially methylated cytosines: Optimization of standard deviation versus state-of-the-art methods. *bioRxiv* **2022**. bioRxiv:486807. [CrossRef]
19. Ali, A.; Gravino, C. Improving software effort estimation using bio-inspired algorithms to select relevant features: An empirical study. *Sci. Comput. Program* **2021**, *205*, 102621. [CrossRef]
20. Cersonsky, R.K.; Helfrecht, B.A.; Engel, E.A.; Kliavinek, S.; Ceriotti, M. Improving sample and feature selection with principal covariates regression. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 035038. [CrossRef]
21. Li, W.; Long, L.C.; Liu, J.Y.; Yang, Y. Classification of magnetic ground states and prediction ofmagnetic moments of inorganic magneticmaterials based on machine learning. *Acta Phys. Sin.* **2022**, *71*, 278–285.
22. Xia, J.F.; Zhang, Y.L.; Jiang, B. Efficient selection of linearly independent atomic features for accurate machine learning potentials. *Chin. J. Chem. Phys.* **2021**, *34*, 695–703. [CrossRef]

23. Imbalzano, G.; Anelli, A.; Giofré, D.; Klees, S.; Behler, J.; Ceriotti, M. Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. *J. Chem. Phys.* **2018**, *148*, 241730. [CrossRef] [PubMed]
24. Liu, J.; Guo, J.; Xu, D. APSNet: Toward Adaptive Point Sampling for Efficient 3D Action Recognition. *IEEE Trans. Image Process.* **2022**, *31*, 5287–5302. [CrossRef] [PubMed]
25. Izonin, I.; Tkachenko, R.; Verhun, V.; Zub, K. An approach towards missing data management using improved GRNN-SGTM ensemble method. *Eng. Sci. Technol. Int. J.* **2021**, *24*, 749–759. [CrossRef]
26. Izonin, I.; Tkachenko, R.; Vitynskyi, P.; Zub, K.; Tkachenko, P.; Dronyuk, I. Stacking-based GRNN-SGTM Ensemble Model for Prediction Tasks. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8–9 November 2020; pp. 326–330.
27. Zhang, W.; Li, H.; Han, L.; Chen, L.; Wang, L. Slope stability prediction using ensemble learning techniques: A case study in Yunyang County, Chongqing, China. *J. Rock Mech. Geotech.* **2022**, *14*, 1089–1099. [CrossRef]
28. Zhou, K.; Yang, Y.; Qiao, Y.; Xiang, T. Domain Adaptive Ensemble Learning. *IEEE Trans. Image Process.* **2021**, *30*, 8008–8018. [CrossRef]
29. Zhang, X.; Jonassen, I. An Ensemble Feature Selection Framework Integrating Stability. In Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA, 18–21 November 2019; pp. 2792–2798.
30. Gow, A.D.; Byrnes, C.T.; Cole, P.S.; Young, S. The power spectrum on small scales: Robust constraints and comparing PBH methodologies. *J. Cosmol. Astropart. Phys.* **2021**, *2021*, 002. [CrossRef]
31. Chen, C.; Deng, Z.; Tran, R.; Tang, H.; Chu, I.-H.; Ong, S.P. Accurate force field for molybdenum by machine learning large materials data. *Phy. Rev. Mater.* **2017**, *1*, 043603. [CrossRef]
32. Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A.V.; Thompson, A.P.; Wood, M.A.; et al. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *J. Phys. Chem. A* **2020**, *124*, 731–745. [CrossRef]
33. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phy. Rev. B* **2019**, *99*, 014104. [CrossRef]
34. Poland, D.; Rychkov, S.; Vichi, A. The conformal bootstrap: Theory, numerical techniques, and applications. *Rev. Mod. Phys.* **2019**, *91*, 015002. [CrossRef]
35. Zhang, W.; Yu, J.; Zhao, A.; Zhou, X. Predictive model of cooling load for ice storage air-conditioning system by using GBDT. *Energy Rep.* **2021**, *7*, 1588–1597. [CrossRef]
36. Yu, Z.; Wang, Z.; Zeng, F.; Song, P.; Baffour, B.A.; Wang, P.; Wang, W.; Li, L. Volcanic lithology identification based on parameter-optimized GBDT algorithm: A case study in the Jilin Oilfield, Songliao Basin, NE China. *J. Appl. Geophys.* **2021**, *194*, 104443. [CrossRef]
37. Lv, S.; Liu, G.; Bai, X. Multifeature pool importance fusion based GBDT (MPIF-GBDT) for short-term electricity load prediction. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *702*, 012012. [CrossRef]
38. Antoniadis, A.; Lambert-Lacroix, S.; Poggi, J.-M. Random forests for global sensitivity analysis: A selective review. *Reliab. Eng. Syst. Saf.* **2021**, *206*, 107312. [CrossRef]
39. Khan, M.A.; Memon, S.A.; Farooq, F.; Javed, M.F.; Aslam, F.; Alyousef, R. Compressive Strength of Fly-Ash-Based Geopolymer Concrete by Gene Expression Programming and Random Forest. *Adv. Civ. Eng.* **2021**, *2021*, 6618407. [CrossRef]
40. Aria, M.; Cuccurullo, C.; Gnasso, A. A comparison among interpretative proposals for Random Forests. *Mach. Learn. Appl.* **2021**, *6*, 100094. [CrossRef]
41. Edelmann, D.; Móri, T.F.; Székely, G.J. On relationships between the Pearson and the distance correlation coefficients. *Stat. Probabil. Lett.* **2021**, *169*, 108960. [CrossRef]
42. Jebli, I.; Belouadha, F.-Z.; Kabbaj, M.I.; Tilioua, A. Prediction of solar energy guided by pearson correlation using machine learning. *Energy* **2021**, *224*, 120109. [CrossRef]
43. Kou, G.; Yang, P.; Peng, Y.; Xiao, F.; Chen, Y.; Alsaadi, F.E. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. *Appl. Soft Comput.* **2020**, *86*, 105836. [CrossRef]
44. Bergh, D.v.d.; Clyde, M.A.; Gupta, A.R.K.N.; de Jong, T.; Gronau, Q.F.; Marsman, M.; Ly, A.; Wagenmakers, E.-J. A tutorial on Bayesian multi-model linear regression with BAS and JASP. *Behav. Res. Methods* **2021**, *53*, 2351–2371. [CrossRef]
45. Huang, X.; Wang, H.; Luo, W.; Xue, S.; Hayat, F.; Gao, Z. Prediction of loquat soluble solids and titratable acid content using fruit mineral elements by artificial neural network and multiple linear regression. *Sci. Hortic.* **2021**, *278*, 109873. [CrossRef]
46. Khaire, U.M.; Dhanalakshmi, R. Stability of feature selection algorithm: A review. *J. King Saud. Univ.-Com.* **2022**, *34*, 1060–1073. [CrossRef]
47. Liu, S.; Lu, M.; Li, H.; Zuo, Y. Prediction of Gene Expression Patterns With Generalized Linear Regression Model. *Front. Genet.* **2019**, *10*, 120. [CrossRef]
48. Torun, H.M.; Swaminathan, M.; Davis, A.K.; Bellaredj, M.L.F. A Global Bayesian Optimization Algorithm and Its Application to Integrated System Design. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2018**, *26*, 792–802. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.