Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 5, No. 1, 2025

Research on Optimization of Strip Steel Process Parameters Based on Random Forest Model

Haotong Wang[*, #], Xijun Gao[#], Mingze Gao[#]

School of Track Intelligent Engineering, Dalian Jiaotong University, Dalian, China

[*]**Corresponding author**: wang15842752121@163.com

[#]**These authors are co-first authors.**

Abstract: Aiming at the problem that it is difficult to accurately model the complex nonlinear relationship between process parameters and hardness in continuous annealing process of cold-rolled strip steel, a data-driven optimization method based on random forest regression model is proposed in this paper. Firstly, Pearson correlation coefficient analysis was used to quantify the correlation between process parameters and hardness, and the key influencing factors such as fast cooling furnace temperature, carbon content and soaking furnace temperature were identified. Secondly, a random forest regression model is established, and the prediction performance of the model is improved by grid search optimization. Finally, the data are preprocessed (including Z-score standardization and outlier processing), and the validity of the model is verified based on simulation experiments. The results show that the mean square error (MSE) of the optimized model on the test set is 285.97, and the coefficient of determination (R^2) is 0.384, which can better predict the hardness of the strip. This method provides theoretical support and decision-making basis for iron and steel enterprises to optimize production process and improve product quality.

Keywords: Cold Rolled Strip Steel, Process Parameter Optimization, Random Forest Regression, Data-Driven, Hardness Prediction.

1. Introduction

Cold rolled strip steel[1] is an important product of iron and steel enterprises, and its quality stability directly affects the efficiency of enterprises. Continuous annealing is the key process of cold rolled strip production, and its process parameters (such as fast cooling furnace temperature, carbon content, soaking furnace temperature, etc.) play a decisive role in product quality. However, due to the complex coupling and nonlinear relationship between process parameters, the traditional mechanism model is difficult to accurately describe its influence law, which brings great challenges to online quality control and process optimization. Therefore, it is of great significance to study how to establish a high-precision prediction model by data-driven method to realize real-time prediction of strip hardness and optimization of process parameters, so as to improve product quality and production efficiency.

Chen H et al.[2] proposed a steel surface defect detection attention network (SSFDANet) based on deep learning to solve the problem of real-time defect detection in the production of cold rolled galvanized automobile steel sheets. The SSFDANet can realize the real-time determination of product quality and the automatic classification of defect types, which effectively solves the problem of insufficient detection accuracy caused by unclear defect characteristics and high similarity between categories in actual production. Chen Y et al.[3] proposed a long-term and short-term memory network model (Attention-LSTM) with attention mechanism for the key quality index of cold rolled strip flatness, which solved the problem that the traditional mechanism model and machine learning method ignored the spatial dimension dependence in prediction. However, these studies mostly focus on the detection or prediction of a single quality index, and the complex relationship between process parameters and multi-dimensional quality indicators is insufficiently explored, and the feasibility of parameter optimization in actual production is lacking in-depth discussion.

Based on the above analysis, this paper proposes a data-driven optimization scheme based on random forest regression model for the complex relationship between process parameters and hardness in continuous annealing process of cold-rolled strip steel. Through data preprocessing, correlation analysis, model training and optimization, the real-time prediction of strip hardness is realized, and the key process parameters are identified.

2. Principle of correlation method

2.1. Correlation analysis based on Pearson correlation coefficient

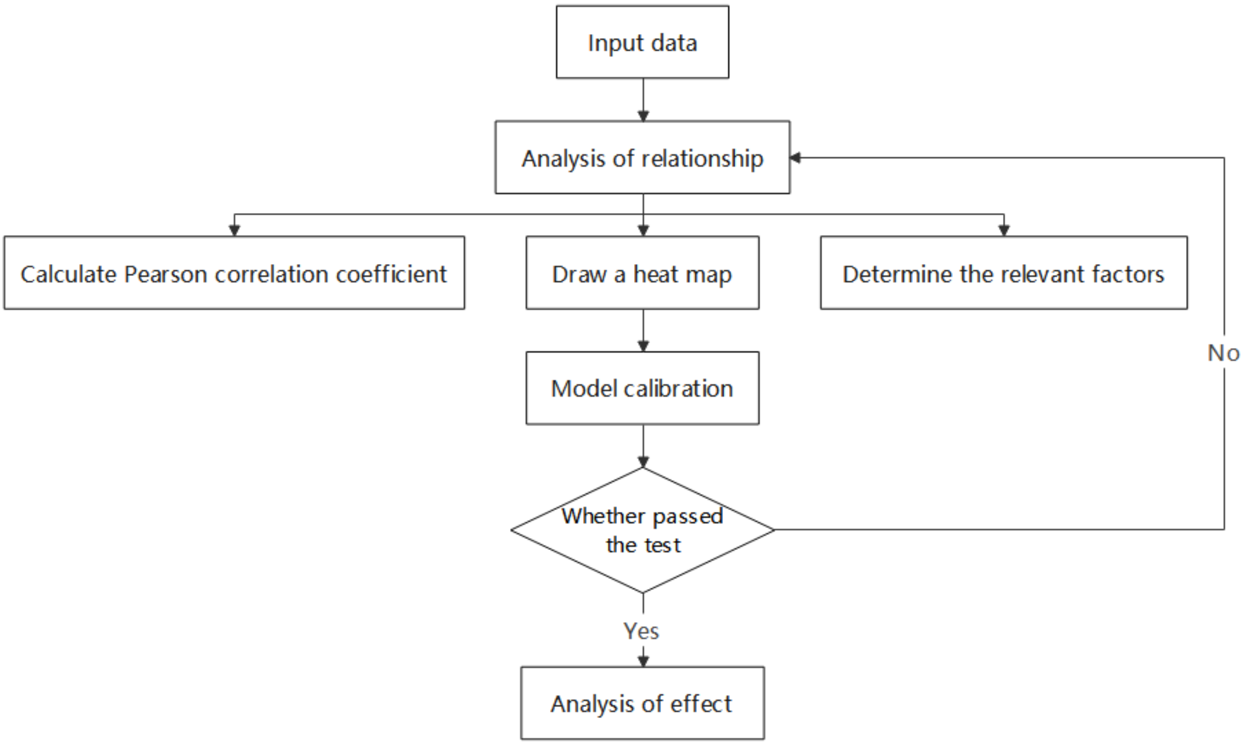

Figure 1. Correlation analysis process of Pearson correlation coefficient

Correlation analysis based on Pearson correlation coefficient[4] is a statistical method used to evaluate the strength and direction of the linear relationship between two continuous variables. Its core is to quantify the correlation between variables by calculating the correlation coefficient (r), and to meet the conditions of normal distribution and linear relationship. In order to determine which parameters have an important influence on the mechanical properties of strip steel, this paper analyzes the Pearson correlation coefficient between the specification data, process parameters and performance index data of strip steel products and the hardness of strip steel products.

The mathematical definition of the Pearson correlation coefficient is the quotient of the covariance of two variables multiplied by their respective standard deviations, i.e.:

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

(1)

where X_i and Y_i are the observed values of the two variables in the sample, and are the mean values of the two variables, respectively, and n is the sample size.

Table 1. Pearson correlation coefficient description

$r = 1$	It indicates that there is a completely positive linear correlation between X and Y. That is, when one variable increases, the other variable also increases in a fixed proportion.
$r = -1$	Indicates that there is a completely negative linear correlation between X and Y. That is, when one variable increases, the other variable decreases in a fixed proportion.
$r = 0$	Indicates that there is no linear correlation between X and Y. But it does not mean that there is no relationship between them, just that there is no linear relationship between them. There may be other types of relationships (such as nonlinear relationships) between them.
$0 < \|r\| < 1$	It indicates that there is a certain degree of linear correlation between X and Y, but it is not completely correlated. The closer to 1, the higher the degree of correlation; the closer to 0, the lower the correlation.

Then the mean value is calculated, and the calculation formula is as follows:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (2)$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \quad (3)$$

2.2. Hardness prediction of cold-rolled strip steel based on random forest regression model

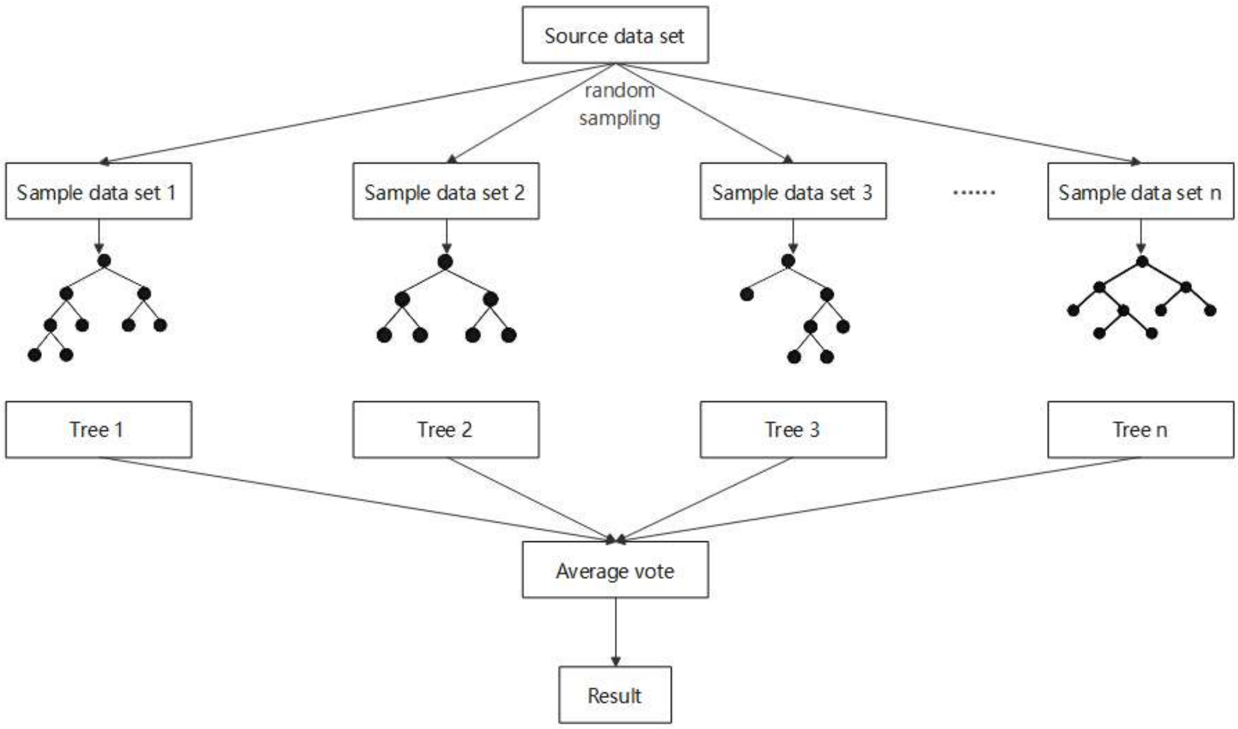

Figure 2. Random forest model diagram

Random forest[5] is an ensemble learning algorithm based on decision tree construction. The random forest regressor improves the overall prediction accuracy and stability by combining the prediction results of multiple decision tree regressors. The following are the main principles of random forest regressors:

Decision tree: Decision tree[6] is a basic machine learning model, which predicts the value of the target variable through a series of judgment rules. Each leaf node outputs not a category, but a continuous value, which is usually the average of the target variables of all training samples on the node. Assuming that the sample set contained in the leaf node L is S_L, the predicted value $\hat{y}_L$ of the node is:

$$\hat{y}_L = \frac{1}{|S_L|}\sum_{i\in S_L} y_i \tag{4}$$

here, $|S_L|$ represents the number of samples in the sample set S_L, and y_i represents the target variable value of the i th sample.

Ensemble learning: Ensemble learning[7] improves overall performance by combining predictions of multiple simple models (such as decision trees). Random forests use this strategy to create multiple decision trees and integrate their prediction results by voting or averaging. The final prediction result $\hat{y}_{RF}$ is the average of all decision tree prediction results:

$$\hat{y}_{RF}(x) = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_m(x) \tag{5}$$

among them, M is the number of decision trees in the random forest, and $\hat{y}_m(x)$ is the prediction result of the m-th decision tree for sample x.

Randomness:

Sample randomness: When constructing each tree, the random forest performs random sampling (called self-sampling) from the original data set. This means that each tree is trained on different data subsets.

Feature randomness: on each node of the tree, the random forest algorithm does not consider all features to find the best segmentation point, but randomly selects a feature subset and then finds the best segmentation point from the subset.

The mathematical function formula of random forest can be expressed as:

$$F(x) = \frac{1}{M}\sum_{m=1}^{M} f_m(x) \tag{6}$$

among them, $F(x)$ is the prediction result of sample x by random forest, M is the number of decision trees in random forest, and $f_m(x)$ is the prediction result of sample x by the m-th decision tree.

3. Simulation analysis

3.1. Data pre-processing

3.1.1. Z-score standardization

In order to preprocess the specification data, process parameters and performance index data of strip steel products, including missing value processing, outlier detection and normalization. In this paper, the Z-score standardization[8] method is used to calculate the z quantile. The z score is the standard difference between the data point and the average value. The calculation formula is:

$$z = \frac{x-\mu}{\sigma} \tag{7}$$

here, x is the original data point, μ is the mean value (mean value) of the data, and σ is the standard deviation of the data. The z score can be used to identify outliers in the data, that is, values that differ from the mean value by more than a certain standard deviation.

3.1.2. Judgment and processing of outliers

According to the set z-score threshold (3 times the standard deviation), determine whether there are outliers in the data set. In general, values greater than 3 standard deviations can be considered outliers.

Continuous outliers are detected, which may indicate that there are missing values in the data. In this case, due to the large number of samples, the impact of the lack of some data can be ignored, so this paper will delete these missing values.

The detected outliers are not continuous, which may indicate that there are abnormal observations in the data. In this case, this paper chooses to replace the outliers with the median of the column, which is a common outlier processing method.

3.2. Correlation analysis results

According to the correlation analysis of Pearson correlation coefficient, the correlation coefficient and the absolute value of the correlation coefficient are as follows:

Table 2. The absolute value of correlation coefficient of each parameter

parameter	The absolute value of correlation coefficient
Fast cooling furnace temperature	0.398455
Carbon content	0.355307
Homogenizing furnace temperature	0.312227
Quenching furnace temperature	0.281479
Strip speed	0.160937
Furnace temperature	0.127605
Overaging furnace temperature	0.112443
Slow cooling furnace temperature	0.111541
Strip thickness	0.049760
Tension of temper mill	0.020577
Strip width	0.005457
Silica content	0.003550

The degree of correlation with the hardness of strip steel products obtained by is as follows: fast cooling furnace temperature>carbon content>soaking furnace temperature>quenching furnace temperature>strip speed>heating furnace temperature>overaging furnace temperature>slow cooling furnace temperature>strip thickness>temper mill tension>strip width>silicon content.

Finally, in order to better observe its correlation, the heat map is drawn. The results are as follows:

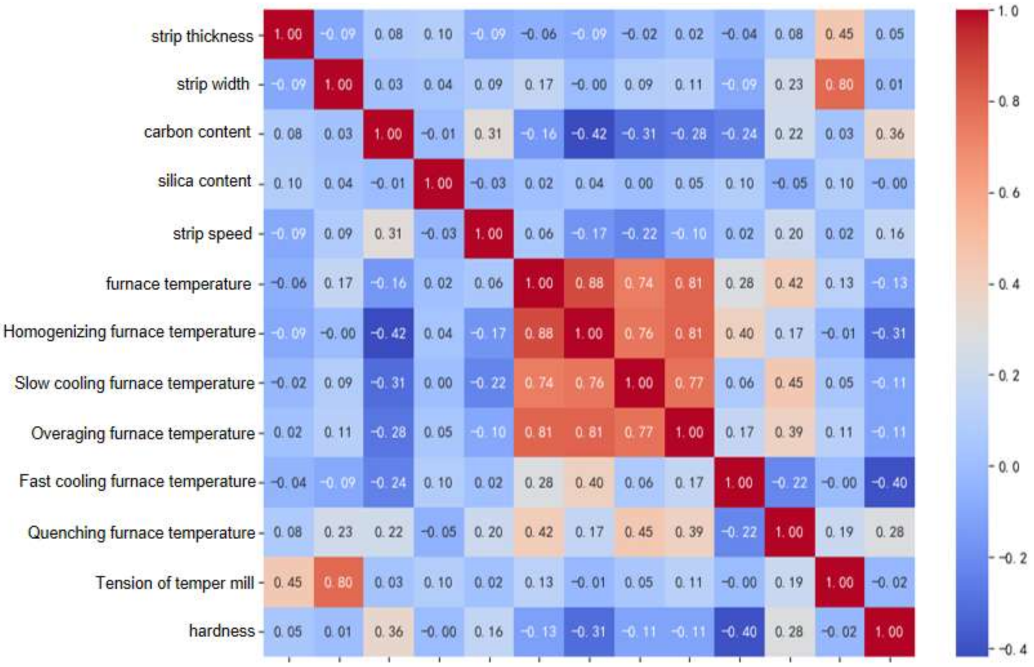

Figure 3. Pearson correlation coefficient heat map

3.3. Random forest model prediction

3.3.1. Division of data sets

In this paper, the data set is divided into training set and test set in order to evaluate the performance of the model. Follow the following steps:

The data set is divided into training set and test set. Use 80 % of the data as the training set and the remaining 20 % as the test set. The training set contains 800 samples and 12 features, and the test set contains 200 samples and 12 features.

Use the training set to train a regression model.

We use the test set to evaluate the performance of the model.

3.3.2. Model training and parameter tuning

Next, this paper will use a random forest regressor for training and evaluation. After using the random forest regressor, the performance of the model on the test set: mean square error (MSE)[9] is 334.59. The determination coefficient (R^2) is 0.374, which indicates that the model can explain about 37.4 % of the hardness change.

In order to train a random forest regression model with good performance, this paper uses a grid search (GridSearchCV)[10] method to find the optimal model parameters. Firstly, a parameter grid is defined, including the number of decision trees (n _ estimators) and the maximum depth of the tree (max _ depth). Then, a random forest regressor instance is initialized, and cross-validation (cv = 5) and scoring criteria (R^2 score) for grid search are set. Through grid search, a set of optimal parameters are found, and the optimal parameters are: 'max _ depth': None, 'n _ estimators': 200, and these parameters are used to retrain the random forest model.

3.3.3. Model evaluation

After training, the performance of the model was evaluated on the test set, and the mean square error (MSE) and R^2 score were calculated to quantify the prediction accuracy and explanatory power of the model. The mean square error (MSE) is 285.972975, and the coefficient of determination (R^2) is 0.3843556112893157.

3.3.4. Model prediction and visual analysis

Through the above process, this paper realizes the training and parameter optimization of random forest regression model. The mean square error of the model on the test set is 285.972975, and the R^2 score is 0.3843556112893157, indicating that the model has good prediction performance and explanatory power. Finally, this paper selects ten sets of data, uses the trained model for prediction, conducts visual analysis, and compares it with the test set. It is found that the model has better performance. The results are as follows:

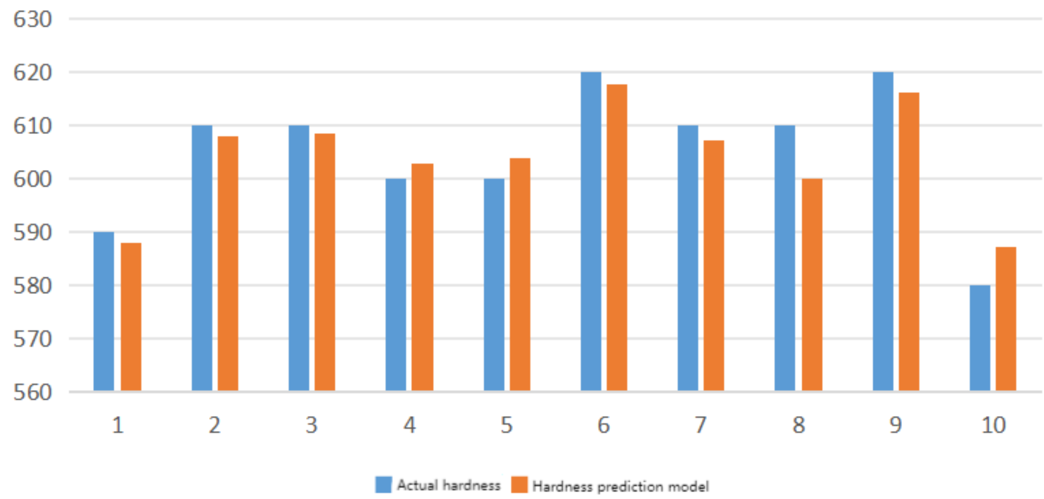

Figure 4. Model prediction results

4. Conclusions

Based on the random forest regression model, this study analyzed and optimized the complex relationship between key parameters and hardness in the continuous annealing process of cold rolled strip steel. Through data preprocessing, correlation analysis and model training, it is found that the temperature of fast cooling furnace, carbon content and soaking furnace temperature have the most significant influence on the hardness of strip steel. The optimized model achieves better prediction performance on the test set, with MSE of 285.97 and R^2 of 0.384, which verifies the effectiveness of the method.

The main advantage of this paper is that the data-driven method overcomes the limitations of the traditional mechanism model and can more accurately capture the nonlinear relationship between process parameters and hardness. At the same time, the key parameters are defined by feature importance analysis, which provides theoretical support for actual production optimization. In addition, the model further improves the prediction accuracy and stability by grid search tuning.

Future research can be carried out from the following aspects: First, introduce more advanced deep learning models, such as attention mechanism[11] or graph neural network[12], to further improve prediction performance ; the second is to explore the real-time application of the model in the actual production control system to realize the dynamic optimization of process parameters. The third is to consider the multi-objective optimization problem and comprehensively improve the multi-dimensional quality indicators such as hardness and flatness of the strip; fourth, expand data sources and enhance the generalization ability of the model. These improvements will help promote the intelligent production of the steel industry and provide more reliable technical support for enterprises to improve product quality and production efficiency.

References

[1] Wang P, Yan Z, Li X, et al. Edge drop control of cold rolled silicon steel strip based on model predictive control[J]. Journal of Manufacturing Processes, 2022, 82: 88-95.

[2] Chen H, Nie Z, Xu Q, et al. Intelligent detection and classification of surface defects on cold-rolled galvanized steel strips using a data-driven faulty model with attention mechanism[J]. Journal of Computing and Information Science in Engineering, 2023, 23(4): 041001.

[3] Chen Y, Peng L, Wang Y, et al. Prediction of tandem cold-rolled strip flatness based on Attention-LSTM model[J]. Journal of Manufacturing Processes, 2023, 91: 110-121.

[4] Šverko Z, Vrankić M, Vlahinić S, et al. Complex Pearson correlation coefficient for EEG connectivity analysis[J]. Sensors, 2022, 22(4): 1477.

[5] Hu J, Szymczak S. A review on longitudinal data analysis with random forest[J]. Briefings in bioinformatics, 2023, 24(2): bbad002.

[6] Costa V G, Pedreira C E. Recent advances in decision trees: An updated survey[J]. Artificial Intelligence Review, 2023, 56(5): 4765-4800.

[7] Zhang Y, Liu J, Shen W. A review of ensemble learning algorithms used in remote sensing applications[J]. Applied Sciences, 2022, 12(17): 8654.

[8] Schober P, Mascha E J, Vetter T R. Statistics from A (agreement) to Z (z score): a guide to interpreting common measures of association, agreement, diagnostic accuracy, effect size, heterogeneity, and reliability in medical research[J]. Anesthesia & Analgesia, 2021, 133(6): 1633-1641.

[9] Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. Peerj computer science, 2021, 7: e623.

[10] Ahmad G N, Fatima H, Ullah S, et al. Efficient medical diagnosis of human heart diseases using machine learning techniques with and without GridSearchCV[J]. Ieee Access, 2022, 10: 80151-80173.

[11] Soydaner D. Attention mechanism in neural networks: where it comes and where it goes[J]. Neural Computing and Applications, 2022, 34(16): 13371-13385.

[12] Veličković P. Everything is connected: Graph neural networks[J]. Current Opinion in Structural Biology, 2023, 79: 102538.