

Supplementary Materials

Industrial big data analysis strategy based on automatic data classification and interpretable knowledge graph

Bingtao Ren, Chenchong Wang, Yuqi Zhang, Xiaolu Wei*, Wei Xu

State Key Laboratory of Rolling and Automation, Northeastern University, Shenyang 110819, Liaoning, China.

***Correspondence to:** Dr. Xiaolu Wei, State Key Laboratory of Rolling and Automation, Northeastern University, NO. 3-11, Wenhua Road, Heping District, Shenyang 110819, Liaoning, China, E-mail: weixl@smm.neu.edu.cn

Supplementary Table 1 represents the composition characteristics of each subdataset.

Supplementary Table 1. The composition features of each subdataset

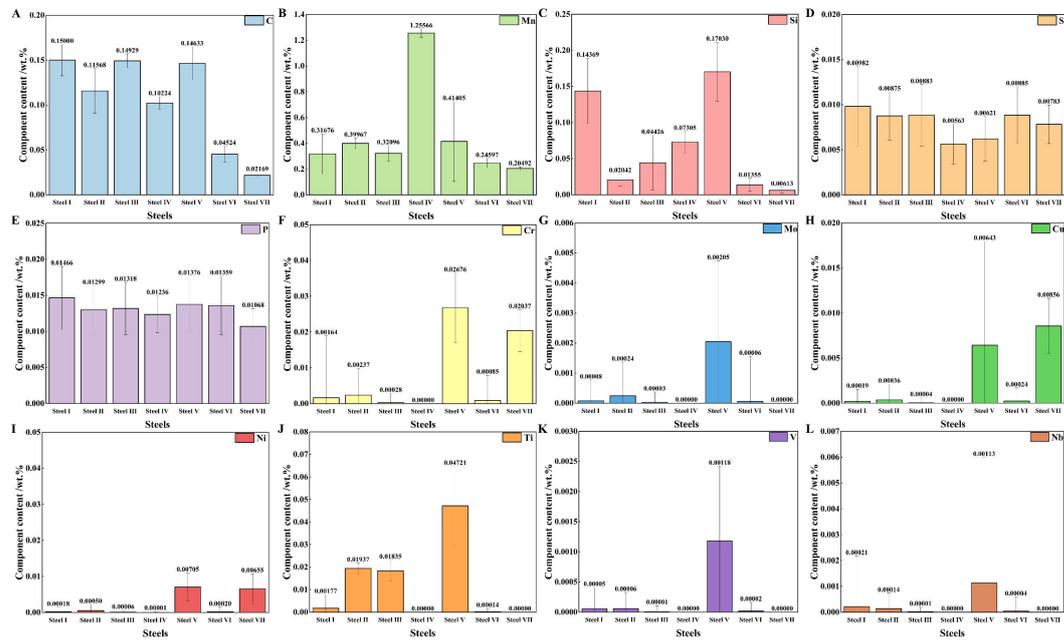
	Cluster1	Cluster2	Cluster3	Cluster4
	Carbon	Carbon	Carbon	Carbon
	Manganese	Manganese	Manganese	Manganese
	Silicon	Silicon	Silicon	Silicon
		Chromium	Chromium	Chromium
			Molybdenum	Molybdenum
Composition features	Copper	Copper	Copper	Copper
	Nickel	Nickel	Nickel	Nickel
		Titanium	Titanium	Titanium
	Vanadium		Vanadium	Vanadium
	Niobium	Niobium		Niobium

Further analysis of the predicted results was conducted using statistical analysis. Statistical analysis provided additional information in the study. The MAE and ER of the GCN and CNN models were examined by the confidence-interval method, with the results presented in Supplementary Table 2. The analysis revealed that the confidence intervals for MAE and ER differed between the GCN and CNN models. Specifically, the GCN model had a lower MAE confidence interval but a higher ER confidence interval. This suggests that the GCN model outperforms the CNN model significantly in terms of MAE and ER across all four subdatasets. It further underscores that the superiority of the GCN model over the CNN model is statistically significant and not merely a result of random variations.

Supplementary Table 2. Confidence interval analysis of Mean Absolute Error (MAE) and Effective Ratio (ER) in GCN and CNN models

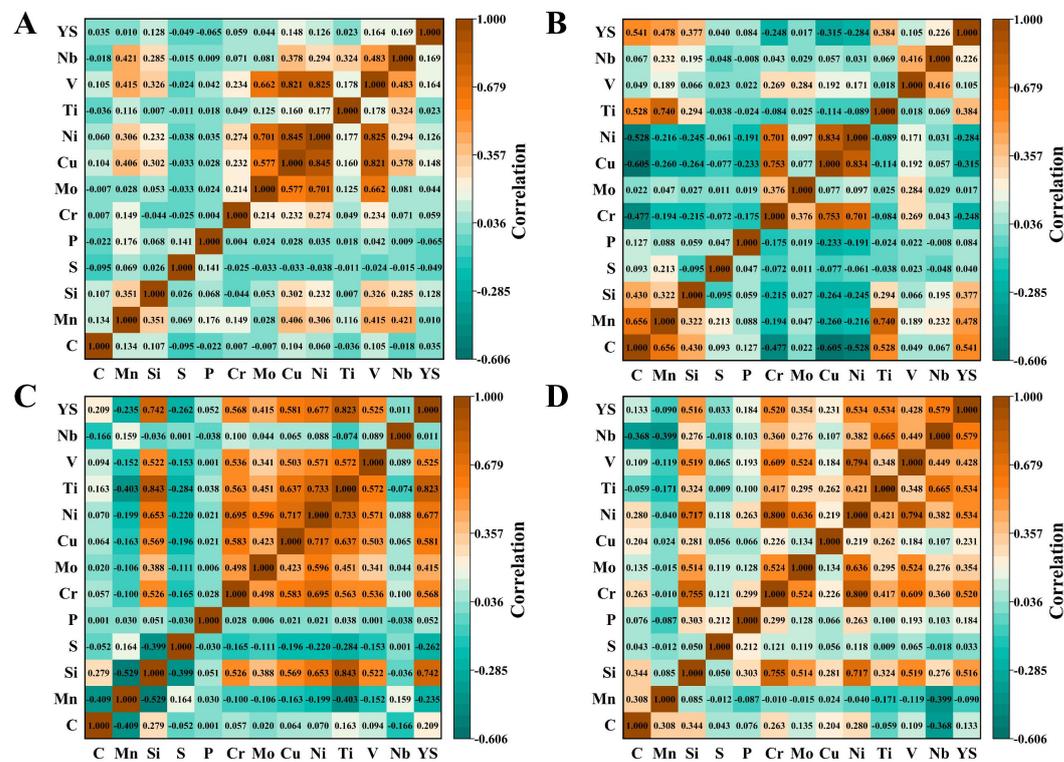
	GCN		CNN	
	MAE /MPa	ER /%	MAE /MPa	ER /%
Cluster1	(21.86, 21.89)	(76.3, 76.5)	(22.34, 22.37)	(74.8, 75.0)
Cluster2	(24.23, 24.27)	(56.2, 56.4)	(25.30, 25.39)	(55.9, 56.2)
Cluster3	(19.84, 19.91)	(80.1, 80.3)	(20.33, 20.38)	(78.6, 78.7)
Cluster4	(19.88, 20.05)	(84.3, 84.6)	(20.06, 20.20)	(82.2, 82.5)

The composition information of each steel grade in the initial dataset is shown in Supplementary Figure 1.



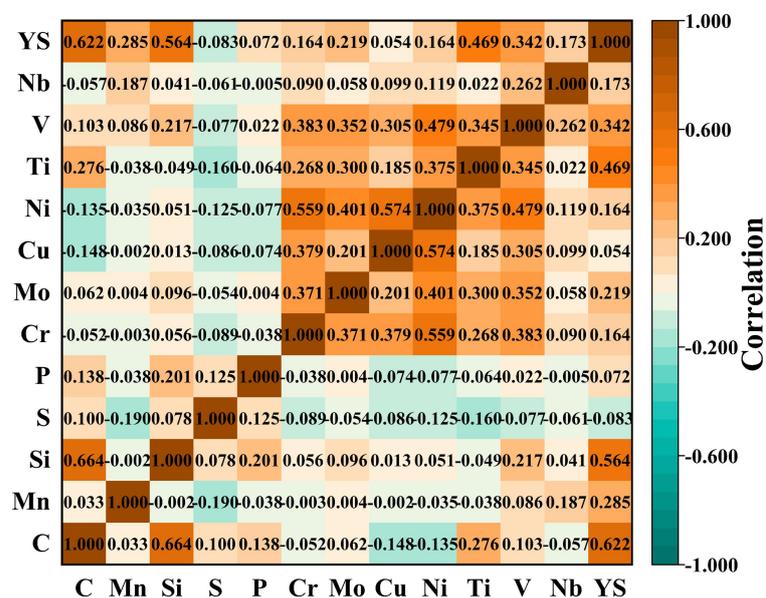
Supplementary Figure 1. Mean composition values of each steel.

Calculated the Pearson correlation coefficients between composition features and output features in each subdataset, as shown in Supplementary Figure 2. In terms of composition features, C, Mn, and Si were considered the basic elements. For other alloying elements, Pearson correlation coefficients were calculated for each subdataset, and alloying element features were selected based on feature importance analysis results. First, S and P both showed a low correlation with yield strength across all four subdatasets. Additionally, in Cluster 1, Cr, Mo, and Ti exhibited a low correlation with yield strength. In Cluster 2, Mo and V had a low correlation with yield strength. For Cluster 3, Nb showed a low correlation with yield strength, while in Cluster 4, other alloying elements demonstrated a high correlation with yield strength.



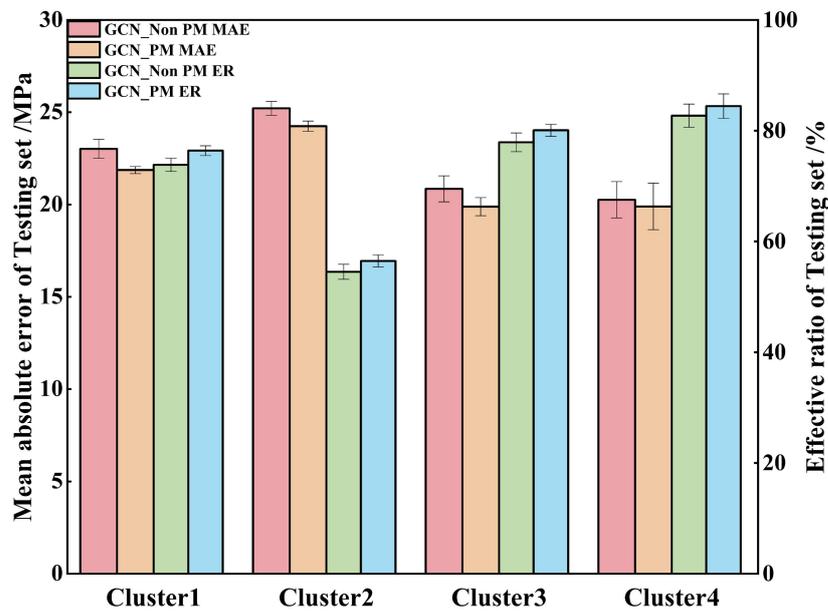
Supplementary Figure 2. Heat map of Pearson correlation coefficient of composition features. A: Cluster1; B: Cluster2; C: Cluster3; D: Cluster4.

Calculated the Pearson correlation coefficient between the composition features and output features in the initial dataset, as shown in Supplementary Figure 4.



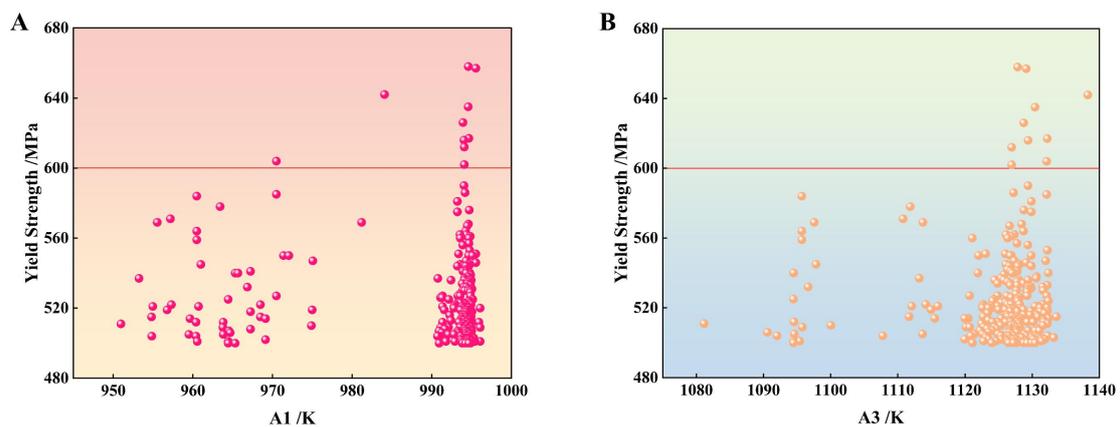
Supplementary Figure 4. Heat map of Pearson correlation coefficient of composition features of Non-Classification.

To highlight the importance of PM parameters, a GCN model was developed using a dataset that excluded PM parameters to predict yield strength. The prediction results of this model were then compared with those of the previously established GCN model guided by PM information, as illustrated in Supplementary Figure 5. The results demonstrate that the GCN model incorporating PM parameters outperforms the model without PM parameters in terms of performance prediction. This indicates that the inclusion of PM parameters enhances the predictive capability of the GCN model while also improving its interpretability.



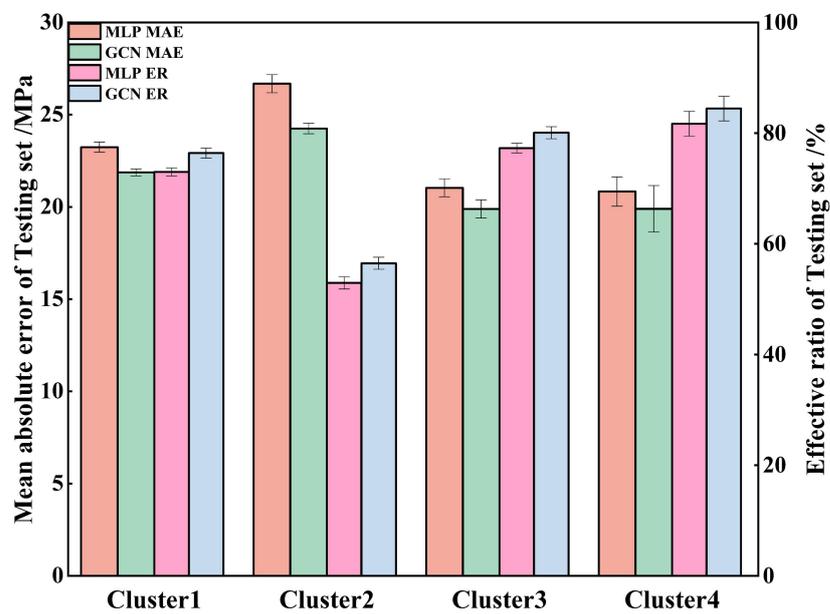
Supplementary Figure 5. Comparison of prediction results for GCN_PM and GCN_Non PM.

As shown in Supplementary Figure 6, the correlation between PM parameters and yield strength was further investigated for samples with higher yield strength in the dataset. The results indicate that higher yield strength samples tend to have higher A1 and A3 temperatures, suggesting that an increase in A1 and A3 temperatures can enhance yield strength to some extent. In metallic materials, an increase in A1 and A3 temperatures enhances atomic activity, leading to more complete austenitization. This improves the diffusion of alloying elements, reduces compositional segregation in the microstructure, and results in a more uniform and stable structure. A uniform and stable structure effectively hinders dislocation motion, enhancing the material's resistance to plastic deformation and thereby increasing its yield strength. Consequently, samples with high yield strength are typically associated with higher A1 and A3 temperatures.



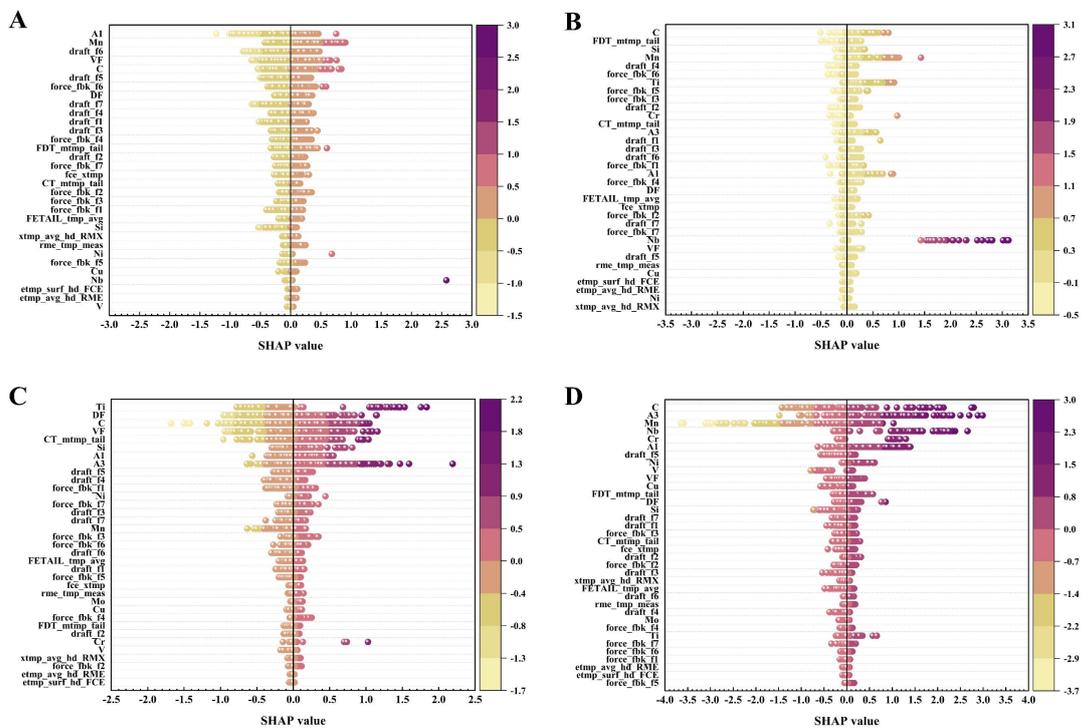
Supplementary Figure 6. The relationship between A1 and A3 and yield strength in high-strength samples.

To further validate the superiority of the GCN model, an MLP-based performance prediction model was developed for each of the four subdatasets to predict yield strength. The MLP and GCN models were compared by calculating the MAE and ER values on the test sets. The attribute prediction results are shown in Supplementary Figure 7. The results indicate that the GCN model consistently outperforms the MLP model in terms of both MAE and ER across all four subdatasets. This demonstrates that the GCN model achieves higher prediction accuracy, further confirming that, compared to non-graph-based models (such as the MLP model), the GCN model improves prediction accuracy and reliability, highlighting its superiority.



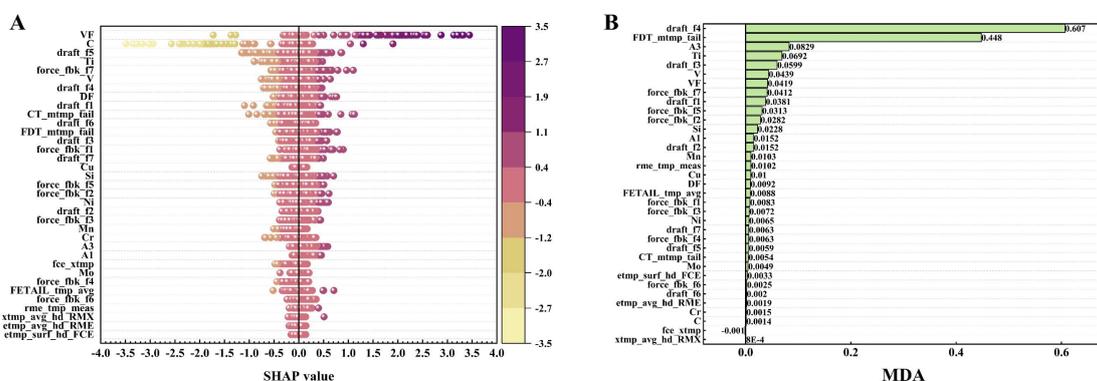
Supplementary Figure 7. Comparison of prediction results for GCN and MLP.

A detailed analysis of input features was performed using the SHAP method. Initially, models were trained based on the GCN algorithm using four subdatasets. Subsequently, the SHAP method was employed to investigate the influence of each input feature on the optimal training model, as illustrated in Supplementary Figure 8. The findings reveal that Mn and C in the alloy composition of cluster 1 make a significant contribution, while A1 and VF in the PM parameters also have substantial contribution. Within cluster 2, C and Si in the alloy composition have relatively high contributions, while the contributions of PM parameters are relatively moderate. For cluster 3, Ti and C in the alloy composition show substantial contribution, along with both PM parameters. Lastly, in cluster 4, C and Mn in the alloy composition have considerable contribution, along with the PM parameters.



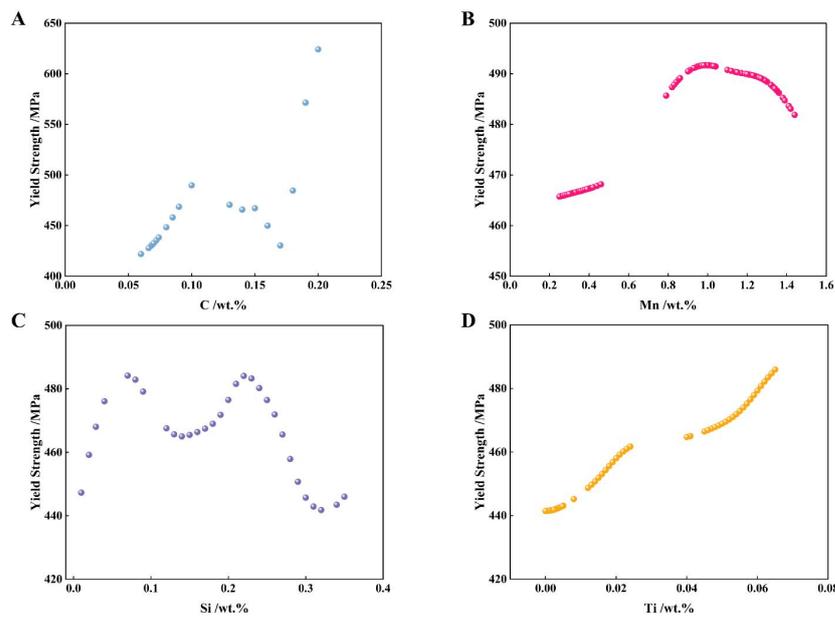
Supplementary Figure 8. Analyzing the importance of input features in four subdatasets using the SHAP method. A: Cluster1; B: Cluster2; C: Cluster3; D: Cluster4.

Additionally, to explore the reasons for the relatively lower performance prediction results of the GCN model for Steel V, a separate performance prediction model was developed using the GCN algorithm to predict the yield strength of Steel V. Subsequently, the SHAP method was applied to the trained GCN model to reveal the importance of input features, as illustrated in Supplementary Figure 9A. The results demonstrate that C, Ti, and V in the alloy composition features exhibit high SHAP values, indicating their significant influence on the model's predictions. Moreover, it was observed that PM parameters have a notable impact on the prediction of yield strength, with variables such as VF and DF also showing substantial contributions. From the classification results (Table 3), it is evident that Steel V predominantly resides in Cluster 3. By comparing the SHAP results of Steel V with those of Cluster 3, it is observed that C, Ti, DF, and VF in Cluster 3 also exhibit high SHAP values. Additionally, the SHAP values of A1 and A3 in Cluster 3 are higher than those in Steel V, indicating a more significant contribution to performance prediction. Furthermore, the Mean Decrease Accuracy (MDA) method combined with a random forest model was employed to analyze the feature importance of Steel V samples and investigate the correlation between each input feature and yield strength. As shown in Supplementary Figure 9B, the MDA scores for each input feature are displayed on the horizontal axis. The feature importance analysis results indicate that Ti in the alloy composition features and A3 in the powder metallurgy parameters exhibit strong correlations with yield strength. This suggests that the A3 temperature and Ti content have the most significant impact on the model's predictions. Unlike the MDA results of Steel V, the MDA score of Ti in Cluster 3 is much higher than that of other features. This may be attributed to the low contribution degree of Ti caused by outliers in Steel V.



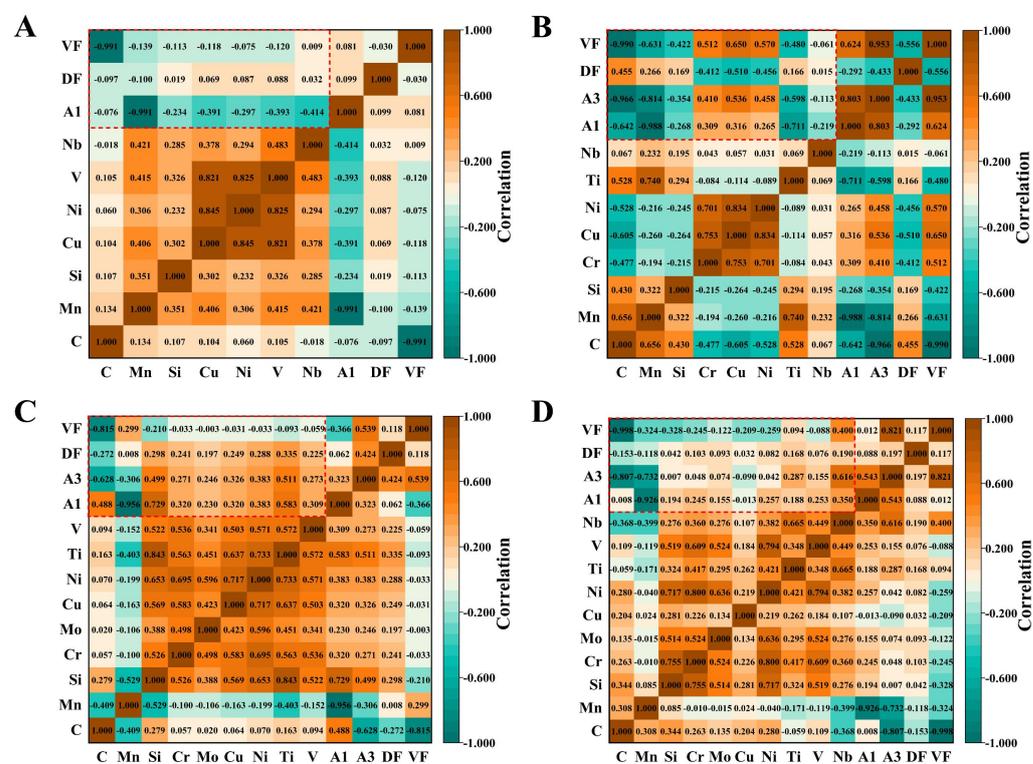
Supplementary Figure 9. Feature importance analysis based on Steel V. A: SHAP values; B: mean decrease accuracy.

It is crucial to design compositions and processing parameters based on the trained model to achieve desirable properties. Therefore, a performance prediction model was established using the GCN algorithm and high yield strength grade steel to predict yield strength and investigate the influence of alloying elements on mechanical properties. This lays the foundation for achieving superior performance in the future, as illustrated in Supplementary Figure 10. The prediction results reveal that the yield strength initially increases with increasing carbon content, then decreases, and subsequently rises again. The highest yield strength is observed at a C content of 0.2 wt.%, although higher C content tends to reduce the ductility of the steel. When the Mn content reaches 1 wt.%, the steel exhibits the maximum yield strength. Regarding silicon content, the yield strength peaks at 0.07 wt.% or 0.22 wt.%. Increasing Ti content leads to a gradual improvement in yield strength; however, excessively high Ti content significantly escalates production costs. In summary, the above analysis provides optimization guidelines for achieving exceptional mechanical properties.



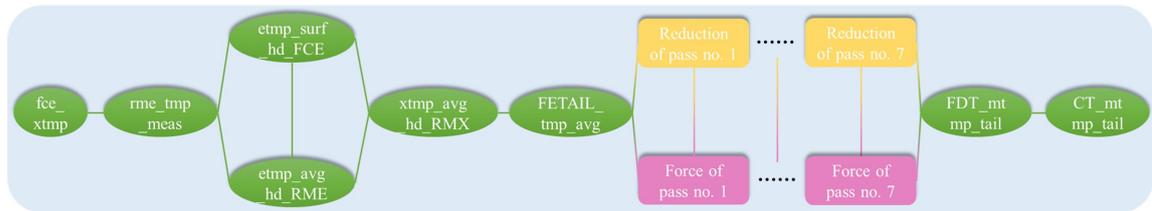
Supplementary Figure 10. The effect of alloy element changes on yield strength. A: C element; B: Mn element; C: Si element; D: Ti element.

Calculated the Pearson correlation coefficient between composition features and PM variables in the subset, as shown in Supplementary Figure 11.



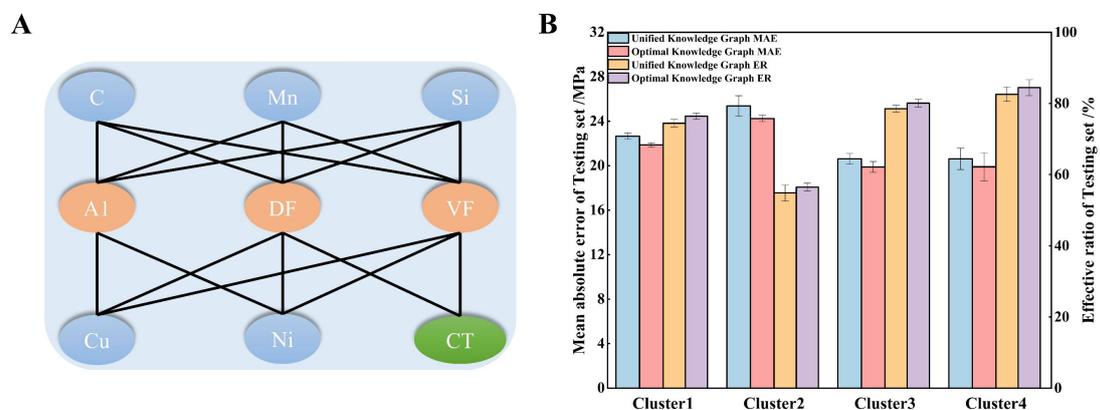
Supplementary Figure 11. Pearson correlation coefficient heatmap of composition features and PM variables for each cluster. A: Cluster1; B: Cluster2; C: Cluster3; D: Cluster4.

The process parameters in the input features of each subdataset were consistent and connected sequentially according to the process flow to obtain a knowledge graph of process features, as shown in Supplementary Figure 12. Combined with PM Graph1-4, performance prediction was achieved.



Supplementary Figure 12. Knowledge graph of process features.

To further demonstrate the superiority of the optimal knowledge graph selected in each subdataset in this study, a unified knowledge graph was constructed for the four subdatasets for performance prediction, as shown in Supplementary Figure 13.



Supplementary Figure 13. Property prediction based on a unified knowledge graph.

A: unified knowledge graph; B: comparison of property prediction results.