

Article

Enhancing Scrap Steel Yield Identification Precision by Community Division of Knowledge Graph

Yuqing Li ¹, Haotian Xu ^{2,3,*}, Dehao Han ¹ and Hongbing Wang ¹¹ School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China² State Key Laboratory of Metallurgical Intelligent Manufacturing System, Beijing 100071, China³ Steel Industry Green and Intelligent Manufacturing Technology Center, China Iron and Steel Research Institute Group, Beijing 100081, China

* Correspondence: xuhaotian0626@163.com

Abstract

Accurately identifying scrap steel yield rates remains challenging due to the diverse types, mixed sources of scrap, and complex furnace working conditions. This paper proposes a mechanism and data joint-driven identification method, and identification precision is enhanced by community division of a knowledge graph. Firstly, a knowledge graph for scrap charging is constructed, and the label propagation algorithm (LPA) is used to divide communities with similar charging patterns. Then, a physics-informed neural network is designed for each community to identify scrap steel yield rates. Finally, the shapley additive explanations approach is employed to assess and quantify the influence of these factors on scrap steel yield rates. Experimental results indicate the following: (1) The proposed model for scrap steel yield rate based on knowledge graph community division achieves the highest identification precision, with a Root Mean Square Error (RMSE) of 3.20 tons and a Mean Absolute Error (MAE) of 2.62 tons. (2) Compared with the baseline joint-driven model without community division, the proposed method reduces the MAE by 26.6% (from 3.57 t to 2.62 t) and significantly improves the hit rate within ± 5 tons by 13.97 percentage points (reaching 87.55%). These improvements validate the effectiveness of the community-based divide-and-conquer strategy in handling complex charging patterns.

Keywords: scrap steel yield rate; converter steelmaking; neural network; community division; knowledge graph

1. Introduction

Steelmaking is a dominant steel production method that utilizes hot metal and scrap steel as primary raw materials, characterized by high efficiency and large-scale production. The primary material inputs include hot metal, scrap steel, slag-forming agents (such as lime and magnesia balls), and oxygen. High-speed oxygen is blown into the converter through an oxygen lance, where high-concentration elements like carbon and silicon in the hot metal are oxidized, and harmful elements such as phosphorus and sulfur are removed, thereby producing molten steel with the desired temperature and composition, as well as generating slag and flue gas [1]. Steel scrap is regarded as one of the key raw materials in steelmaking, and the utilization of scrap steel is of great significance for lowering carbon emissions. For a single type of steel scrap, its yield is defined as the proportion of steel converted into molten steel to its total weight, measuring the efficiency of scrap conversion into molten steel during the smelting process. The scrap steel yield rate is an



Academic Editor: Chun-Chung Liao

Received: 7 June 2026

Revised: 1 July 2026

Accepted: 8 July 2026

Published: 23 July 2026

Copyright: © 2026 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

important indicator of scrap resource utilization efficiency in the steelmaking process [2–5]. The average scrap steel yield rate for different types of scrap is a comprehensive metric, calculated as the weighted average of the individual yields, reflecting the overall scrap utilization efficiency. The identification of scrap steel yield rate plays a significant role in the recycling efficiency of scrap resources and the green and low-carbon transformation of the steel industry. With the advancement of China's strategic research into the high-quality recycling of scrap resources under the "double carbon" background, improving scrap steel yield rates is of great significance for promoting the green and low-carbon development of the steel industry [6].

However, there is limited research specifically focused on the identification of scrap steel yield rate. The yield rate of a single scrap type is defined as the mass ratio of molten steel converted from this scrap to its total feeding mass, and the average scrap yield of one whole heat is the weighted average value of all single scrap yields. Predicting scrap steel yield rate in steelmaking is challenging due to the wide variety of scrap types and numerous process factors. Firstly, the influencing factors are complex and diverse, including scrap charging patterns, hot metal temperature, endpoint temperature, oxygen flow rate, and oxygen blowing volume, as well as the contents of elements like carbon, sulfur, manganese, and phosphorus in hot metal. Identifying scrap steel yield rates through mechanism analysis is difficult due to the high dimensionality and non-linearity of these factors. Secondly, due to the continuity and high efficiency of the steelmaking process, it is impractical for online measurement for each heat. The complex furnace working conditions cannot be simulated in a laboratory, where it is difficult to fully replicate the high temperatures, high pressures, complex chemical reactions, and large-scale material flow in actual furnace working conditions. Additionally, the iron content in scrap cannot be equated to the true yield. The iron content only reflects the proportion of iron in the scrap raw materials, while yield involves the proportion of steel actually converted into steel products during the complex process. Scrap steel yield rates are mainly affected by iron oxidation, scrap melting losses, and splashing losses. Therefore, in the actual smelting process, the scrap steel yield rate cannot be equal to the iron content in the scrap steel.

Existing identification models for scrap steel yield rates are primarily categorized into mechanism-driven and data-driven approaches. Mechanism-driven methods primarily predict the scrap steel yield rate during steelmaking through static and dynamic mechanism models formed over a long period in the metallurgical field, and they possess a certain degree of interpretability. Existing studies mostly focus on mechanism analysis of the scrap melting process in converters. Wang [7] numerically calculated the melting behavior of scrap steel in a full-scale converter under different blowing conditions using a three-dimensional coupling model for turbulent multiphase flow, melting, and solute transport. Singha [8] developed a thermodynamic calculation model to describe the steelmaking process involving scrap and lime dissolution. Deng et al. [9] established a mathematical model to study the melting mechanism of scrap in a dephosphorization converter, considering the coupling effects of the carbon content in scrap steel, the temperature of hot metal, and the temperature of preheating scrap on the scrap melting rate.

In terms of scrap steel yield rate identification methods, mechanism-driven research has mainly focused on laboratory measurement methods. The scrap steel yield rate is measured using an intermediate frequency furnace proposed by Ding Jian [10]. However, this method relies on preprocessed samples with a certain composition. However, the composition of scrap steel is diverse in the actual steel-making process. Moreover, only the weight ratio of molten steel to raw scrap is calculated without considering the loss by smelting, such as chemical oxidation loss, smoke emission, and slag entrainment. The comparative analysis method combining heat balance and material balance was proposed

by He Ruifei [11]. But the establishment of balance equations relies on ideal metallurgical assumptions, making it difficult to quantify the impact of dynamic factors such as uneven molten pool temperature and fluctuating oxygen supply under industrial operating conditions; some key parameters in the equations need to be estimated through experience, and the accumulation of errors will significantly reduce the accuracy of yield rate identification. The adaptive algorithm for scrap steel yield rate (ASY-DCO) proposed by Gan Junjie [12] realizes adaptive adjustment of the yield rate through dynamic cycle optimization based on the yield rate measured in the laboratory. But the initial reference value of the algorithm is derived from laboratory measurement results, and the deviation between laboratory data and industrial reality will be directly transmitted to the optimization process, leading to an inaccurate basis for adaptive adjustment, and the adjustment strategy only focuses on the cyclic correction of yield rate values without adapting to the core influencing factors of scrap steel yield rate, making it difficult to truly reflect the actual utilization efficiency of scrap steel in industrial scenarios.

Mechanism calculation methods utilize long-established mechanisms and expert knowledge in the metallurgical field and have excellent interpretability. However, since mechanism formulas contain a large number of unmeasurable key parameters, the identification of unmeasurable parameters is a key challenge affecting the accuracy of mechanism methods.

Data-driven scrap steel yield rate identification methods, centered on machine learning algorithms, have attracted certain attention in recent years. Zhang Chaojie [13] proposed a method for predicting the deviation coefficient of scrap steel yield rate in steelmaking based on the Random Forest (RF) algorithm and the multi-layer neural network (MLNN), realizing the identification of scrap steel yield rate by leveraging the fitting and learning capabilities of the algorithms. Erik Sandberg et al. [14] evaluated scrap steel properties using data from the Electric Arc Furnace (EAF) steelmaking process and constructed identification models for molten steel chemical composition, power consumption, and scrap steel yield rate using Partial Least Squares (PLS). However, these data-driven methods face significant limitations. Neural networks, often regarded as “black boxes”, lack interpretability regarding the relationship between input factors and identifications, limiting their acceptance in industrial operations. The generalization ability of some algorithms relies on a large amount of high-quality data, resulting in the inability to reproduce the theoretical hit rate of the model in actual on-site applications, thereby leading to low hit rates and poor scalability.

While existing studies have made some progress in scrap steel yield rate identification, critical limitations persist. The scrap steel yield rate in steelmaking is inherently affected by multi-dimensional and non-linear factors. The above methods fail to integrate metallurgical mechanisms with neural networks, resulting in models lacking interpretability and the capability for online identification.

Considering the severe heterogeneity of scrap feeding patterns across different heats, a single global model cannot fit diversified furnace conditions accurately. A knowledge graph can quantitatively describe the feeding similarity between heats through topologically weighted edges, providing an unsupervised clustering carrier for subsequent batch community division. Therefore, this paper constructs a scrap charging knowledge graph and adopts the LPA algorithm to divide heats with analogous feeding features into independent sub-communities for separate modeling. Thus, there is a critical need for the development of an efficient and online scrap steel yield rate identification method, and this method must be highly adaptable to the complex and changing furnace working conditions. In this paper, a mechanism and data joint-driven model for scrap steel yield rate identification is designed, and identification precision is further enhanced by community division of the knowledge graph. In Section 2, a knowledge graph for scrap charging patterns and the

label propagation algorithm (LPA) is used to identify communities for simulating the heats with similar material input characteristics. The mechanism formula for iron mass balance is incorporated into a multi-layer neural network as the loss function so that the difference between input and output iron mass can be minimized. SHAP is employed to evaluate and quantify the influence of various factors on scrap steel yield rates. In Section 3, the experimental design and results are presented. Finally, Section 4 is for the presentation of the experimental conclusions and the future outlook of this study.

2. Mechanism and Data Joint-Driven Identification Model for Scrap Steel Yield Rate

2.1. Data Item Description

Input variables of the model are preliminarily selected based on existing literature. Spearman correlation coefficient analysis is adopted for variable screening to statistically verify the correlation between each input feature and scrap steel yield rate and confirm the validity of selected factors. Meanwhile, all parameters are verified for rationality together with on-site technical experts from steel plants. The input factors for our identification model are listed in Table 1, including hot metal amount, total scrap amount, oxygen consumption, endpoint temperature, and elemental contents (e.g., carbon, silicon, phosphorus) in hot metal, as well as the quantities of various scrap types and auxiliary materials (e.g., lime).

Table 1. Input factors and descriptions.

No.	Item	Unit	Reference Range	No.	Item	Unit	Reference Range
X ₁	Hot metal amount	t	265.3	X ₁₂	Lime amount	t	6.2
X ₂	Oxygen consumption	Nm ³	14,980	X ₁₃	Scrap01 amount	t	2.26
X ₃	Endpoint temperature	°C	1695	X ₁₄	Scrap02 amount	t	6.27
X ₄	[C] content in hot metal	%	4.652	X ₁₅	Scrap03 amount	t	4.65
X ₅	[Mn] content in hot metal	%	0.2336	X ₁₆	Scrap04 amount	t	1.12
X ₆	[P] content in hot metal	%	0.1134	X ₁₇	Scrap05 amount	t	5.6
X ₇	[S] content in hot metal	%	0.003526	X ₁₈	Scrap06 amount	t	0.56
X ₈	[Si] content in hot metal	%	0.3613	X ₁₉	Scrap07 amount	t	1.79
X ₉	Hot metal temperature	°C	1368	X ₂₀	Scrap08 amount	t	3.62
X ₁₀	molten steel amount	t	299.9	X ₂₁	Scrap09 amount	t	2.15
X ₁₁	alloy amount	t	0.43	X ₂₂	Scrap10 amount	t	14.16

2.2. Proposed Model

In recent years, neural networks have been increasingly applied in industrial fields [15,16]. Neural networks excel at analyzing large-scale historical data, particularly in handling multi-dimensional and non-linear problems by automatically extracting complex relationships between input and output. With the rapid development of deep learning, more studies have explored the integration of physical mechanisms and data-driven models to solve complex industrial problems. Raissi [17] proposed physics-informed neural networks (PINNs) to address forward and inverse problems involving non-linear Partial Differential Equations (PDE). By embedding physical constraints directly into the neural network's loss function, the model adheres to physical laws during training, improving identification accuracy and generalization ability. The concept of physics-aware neurons is reflected in embedding physical constraints directly into the neural network architecture. Jia [18] proposed a Physics-Guided Recurrent Neural Network (PGRNN) that introduces physical factors as intermediate factors into the network architecture. Xia [19] proposed a hybrid data-driven and mechanism-based method (dmPINNs) for predicting the endpoint carbon content in converters. This method integrates decarburization mechanism formu-

las into the neural network and identifies unmeasurable parameters in the mechanism formulas by the neural network.

Physics-Informed Neural Networks have advantages such as improving identification accuracy and ensuring interpretability guided by physical mechanisms. Therefore, this paper adopts an error correction method, utilizing the iron mass balance mechanism formula to construct a physical baseline (for calculating the theoretical iron output). A multi-layer neural network is employed to learn the non-linear residual between the actual molten steel amount and this physical baseline. Finally, the mechanism-based baseline is combined with data-driven residual correction to minimize the difference between the input and output iron mass. Moreover, since scrap charging has specific patterns for different heats, community division is introduced to find the charging pattern. The scrap steel yield rate identification model is then constructed for each identified pattern or community.

This paper proposes an innovative framework realizing the full-process closed loop of “pattern classification—separate modeling—result interpretation”. As shown in Figure 1, the construction of this model primarily involves four steps: (1) dataset is processed by methods such as outlier detection; (2) a knowledge graph is built from data to visually represent the similarities and differences in scrap charging patterns for different heats; (3) the graph is partitioned into communities by using LPA algorithm, so that nodes with similar scrap charging pattern can be grouped; and (4) physics-informed neural networks for different communities are trained respectively.

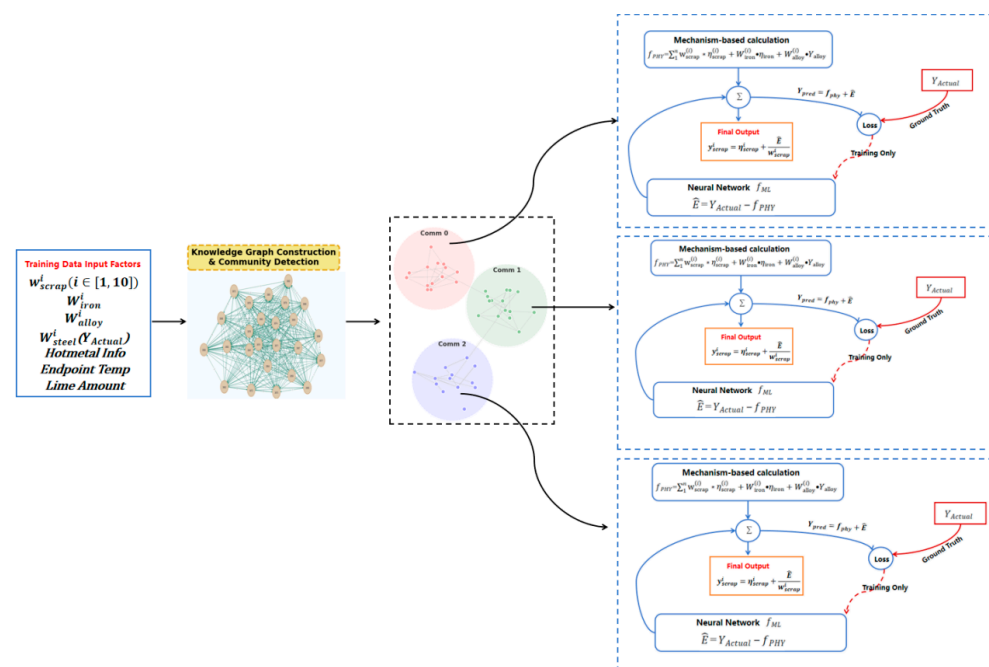


Figure 1. Mechanism and data joint-driven scrap steel yield rate identification model.

2.3. Construction of the Knowledge Graph for Scrap Charging Pattern

To reveal the complex relationships between heats and the added scrap, a knowledge graph was constructed with heats and the added scrap as core entity nodes. Specifically, heat numbers and scrap types were extracted from smelting data as key entities, and directed edges representing the “heat-includes-scrap” relationship were established to form a multi-dimensional network structure reflecting scrap charging patterns. For example, heat 29 is connected to scrap types SCRAP01, SCRAP02, and SCRAP03 through “includes” relationships if these three types of scrap are added into this heat. By using the graphic modeling functions of the py2neo library, the above structure is stored in the Neo4j graph

database for the intuitive comparison of charging patterns among different heats. Figure 2 illustrates that a single heat node connects to multiple scrap type nodes. The graph clearly shows the differences in scrap types in each heat, providing a basis for subsequent community detection. By this structured representation, the knowledge graph can intuitively show the similarities and differences in scrap charging patterns across different heats.

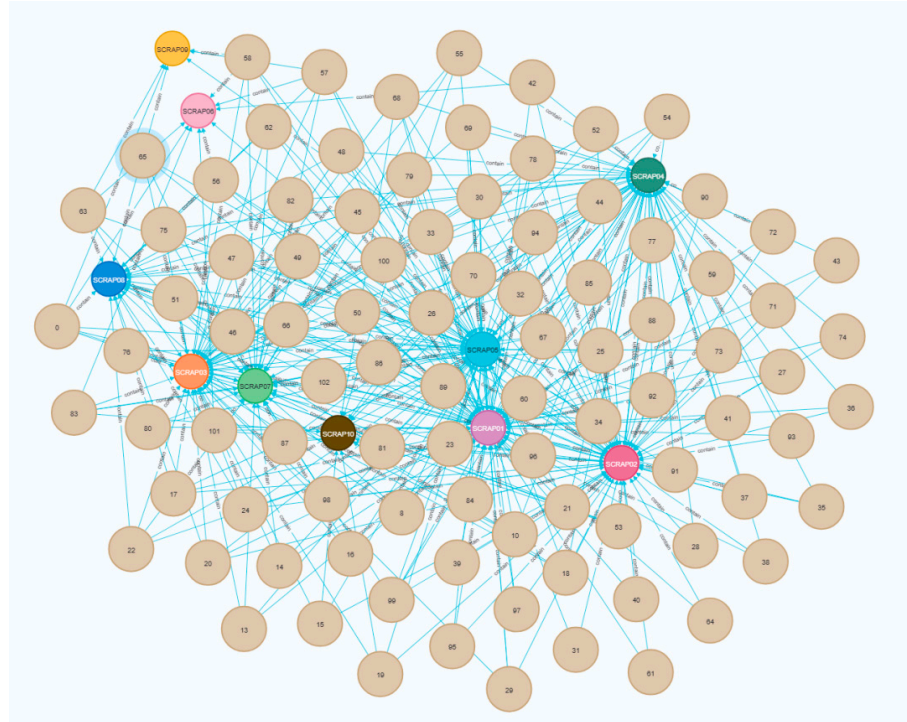


Figure 2. Knowledge graph for scrap charging containing heats and added scraps.

The constructed graph structure is stored in the Neo4j graph database. The graph structure provides an efficient topology for subsequent community detection algorithms. To further facilitate the application of the label propagation algorithm for community division, it is necessary to construct a knowledge graph for scrap charging patterns where each heat is defined as a node, and the direct relationships between heats are represented as edges. Associative edges between heats are constructed based on the similarity of the scrap charging structure, considering both scrap types and their specific weights. To eliminate the interference of trivial additions, a scrap type is considered a valid feature only if its charging amount exceeds a specific threshold (>1 ton). Consequently, each heat is represented as a high-dimensional scrap weight vector $V = [w_1, w_2, \dots, w_n]$. The edge weight between two heats is determined by the cosine similarity of their vectors rather than a simple count of overlapping types. For example, if Heat 80 and Heat 83 both utilize SCRAP03 and SCRAP05 with highly consistent weight ratios, the edge weight will approach 1, thereby accurately reflecting the similarity of their furnace working conditions. This edge weight directly reflects the similarity of scrap charging patterns among heats. Figure 3 illustrates the projected knowledge graph of heats, where nodes represent individual heats, and edges denote the similarity in scrap charging structures and their proportional distributions among heats. The weight of each edge is determined by the cosine similarity of the scrap weight vectors between two heats, enabling accurate reflection of the similarity in furnace conditions across different heats.

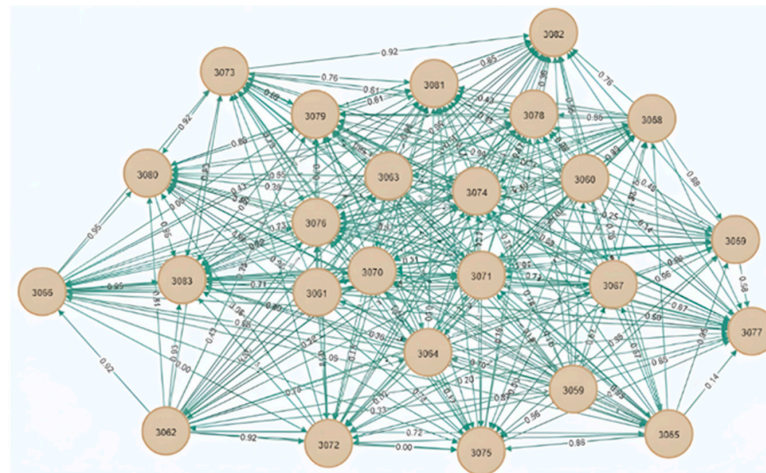


Figure 3. Knowledge graph for scrap charging containing heats.

2.4. Community Detection of the Scrap Charging Pattern Knowledge Graph

The LPA [20,21] is used to divide the knowledge graph for scrap charging pattern into communities. The process involves initializing each heat node's label, iteratively updating each node's label to the most frequent label among its neighbors, and stopping when all node labels match their neighbors' most frequent labels, forming several heat communities. The complete grouping rule is detailed as follows:

1. **Node Initialization:** Each heat node is assigned an exclusive initial unique label according to its heat number.
2. **Edge Weight Calculation:** The connection weight between two heat nodes is calculated by cosine similarity of normalized scrap feeding weight vectors, which quantifies the similarity of scrap types and feeding mass ratios.
3. **Iterative Label Update:** In each iteration, every heat node updates its label to the label with the maximum cumulative weight of adjacent connected nodes.
4. **Convergence Termination:** The iteration loop stops once no node label changes in a full round. All heats sharing highly consistent scrap feeding vector features will be automatically clustered into one community.

As visualized in the figure, the heats spontaneously form distinct communities under the force of the topological connections. These communities represent groups of heats with highly similar scrap charging patterns. This topological clustering provides the physical basis for training scrap steel yield rate identification models for different working conditions.

2.5. Design of the Physics-Informed Neural Network for Identifying Scrap Steel Yield Rate

Based on community detection, a physics-informed neural network is constructed to train the found communities by LPA, respectively. As shown in Figure 4, the neural network incorporates a hybrid loss function that integrates the iron mass balance mechanism and data-driven residual constraints, ensuring consistency between the predicted iron content and the actual molten steel amount while minimizing the deviation between the input and output iron mass. By feeding into the neural network using the factors affecting the scrap steel yield rate—such as oxygen consumption; endpoint temperature; the contents of elements in hot metal; the amounts of fluxes, including lime, magnesia balls, and dolomite; the amount of hot metal; the amount of molten steel; the amounts of various types of scrap steel added; and the amount of alloy—the complex non-linear relationships between these factors and the scrap steel yield rate can be learned. The scrap steel yield rates of various types and the hot metal rate for each heat are output. The model's accuracy is verified by comparing the iron mass balance errors of the test heats.

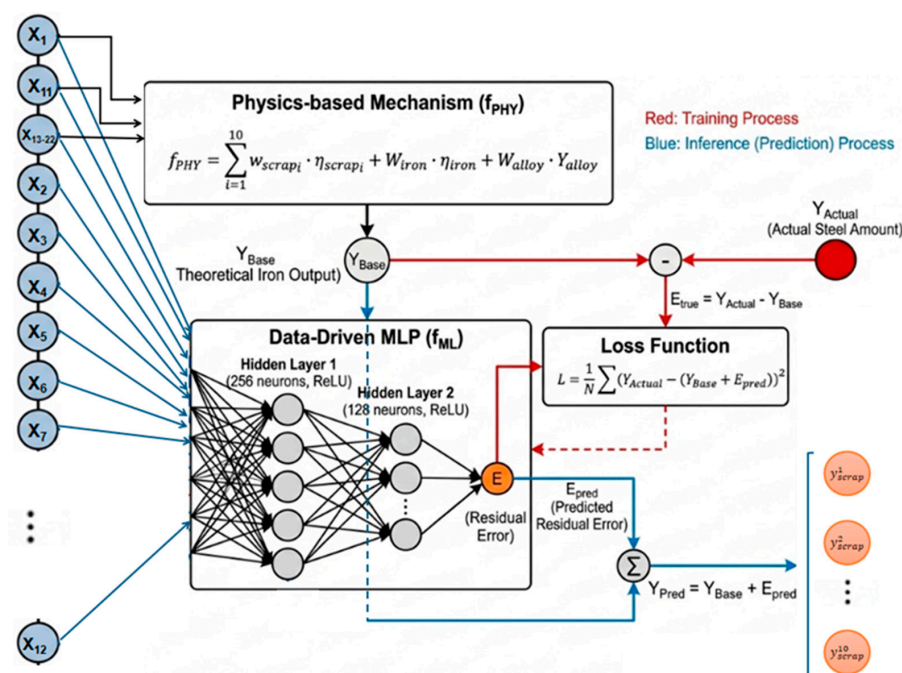


Figure 4. Scrap steel yield rate identification model for each community.

According to the iron balance formula:

$$W_{\text{steel}}^{(i)} = \sum_1^n w_{\text{scrap}}^{(i)} \times y_{\text{scrap}}^{(i)} + W_{\text{iron}}^{(i)} \cdot Y_{\text{iron}} + W_{\text{alloy}}^{(i)} \cdot Y_{\text{alloy}} \quad (1)$$

where $y_{\text{scrap}}^{(i)}$ is the scrap steel yield rate of the i -th heat, $W_{\text{steel}}^{(i)}$ is the amount of molten steel, $W_{\text{iron}}^{(i)}$ is the amount of hot metal, $W_{\text{alloy}}^{(i)}$ is the amount of alloy added, $w_{\text{scrap}}^{(i)}$ is the amount of scrap steel added, Y_{iron} is the hot metal yield rate, and Y_{alloy} is the alloy yield rate.

Since the scrap steel yield rate of each individual scrap type varies with furnace conditions, ground truth values are unavailable. Relying on mechanism formulas allows only for the calculation of the total scrap steel yield rate, failing to identify the scrap steel yield rate of specific scrap types individually. Furthermore, current data-driven studies typically use empirical values as ground truth for training, which presents significant limitations. Therefore, a mechanism and data joint-driven identification model for scrap steel yield rate is proposed. Based on the principle of iron balance, the theoretical iron output (f_{PHY}) is established as the physical baseline, calculated as:

$$f_{\text{PHY}} = \sum_1^n w_{\text{scrap}}^{(i)} \times \eta_{\text{scrap}}^{(i)} + W_{\text{iron}}^{(i)} \cdot \eta_{\text{iron}} + W_{\text{alloy}}^{(i)} \cdot Y_{\text{alloy}} \quad (2)$$

where $\eta_{\text{scrap}}^{(i)}$ is the empirical scrap steel yield rate of the i -th heat, and η_{iron} is the empirical hot metal yield rate.

To accurately identify the scrap steel yield rates of the 10 selected scrap types and account for complex process variations, a residual modeling strategy is adopted. In the training phase, the difference between the actual molten steel amount (Y_{Actual} , representing the true iron content) and the physical baseline (f_{PHY}) is calculated as the residual (Error = $Y_{\text{Actual}} - f_{\text{PHY}}$). A neural network (f_{ML}) is then trained to learn this non-linear deviation. Consequently, the final prediction is obtained by superimposing the neural network's estimated residual onto the physical baseline, thereby effectively combining the interpretability of the physical iron balance with the corrective precision of data-driven modeling. To ensure consistency between the predicted iron content and the actual molten

steel amount, a hybrid loss function integrating both mechanism and data-driven components is formulated. This function minimizes the Mean Squared Error (MSE) between the ground truth and the total predicted value. Specifically, the total prediction is defined as the sum of the mechanism-based baseline and the data-driven residual correction. The mathematical expression is as follows:

$$L = \frac{1}{N} \sum_{i=1}^N \left(Y_{\text{Actual}}^{(i)} - \left(\underbrace{Y_{\text{Base}}^{(i)}}_{\text{Mechanism}} + \underbrace{\hat{E}^{(i)}}_{\text{Data-Driven}} \right) \right)^2 \quad (3)$$

where N is the batch size. $Y_{\text{Actual}}^{(i)}$ denotes the measured molten steel amount for the i -th heat. $Y_{\text{Base}}^{(i)}$ represents the baseline iron content calculated by the iron mass balance mechanism, while $\hat{E}^{(i)}$ is the learnable residual correction term generated by the neural network to capture the complex non-linear deviations that the static mechanism cannot explain.

In this study, the data-driven component (f_{ML}) is designed as a Multi-Layer Perceptron (MLP) consisting of an input layer, two hidden layers, and an output layer. Unlike traditional methods that incorporate physical laws merely as penalty constraints within the loss function, the proposed residual modeling framework explicitly decouples the linear physical relationships from complex non-linear process variations. Specifically, the loss function is formulated to minimize the difference between the network's predicted residual and the actual residual derived from the iron balance. By training the MLP to focus on the discrepancies that the fixed yield rate formula cannot explain, this architecture prevents the model from overfitting to basic physical laws and allows it to capture the unmodeled dynamic fluctuations, thereby enhancing the overall accuracy and robustness of the scrap steel yield rate identification.

2.6. Scrap Steel Yield Rate Identification for New Heats by Voting Algorithm

The new heats in the test set were divided into corresponding communities through a similarity-based voting algorithm. A weighted community assignment algorithm for test heats based on vector dot product is proposed in this paper. The community assignment algorithm for a new heat to be predicted is implemented through the following process. The voting mechanism process is described in Algorithm 1.

Algorithm 1. Voting Algorithm for Community Assignment Algorithm

- (1) Feature Vector Generation: The scrap vector $\vec{W}$ of a new heat is constructed by normalized weights of scraps with usage ≥ 1 t, with dimensions equal to the total number of scrap types and weights of non-effective scraps set to 0.
 - (2) Community Feature Representation: The feature vector $\vec{\mu}_k$ of each community C_k is the vector from its training heats, where V_{ki} takes 1 at positions of effective scraps and 0 otherwise.
 - (3) Weighted Similarity Calculation: The dot product $S_k = \vec{W} \cdot \vec{\mu}_k$ is computed between $\vec{W}$ and the feature vector $\vec{\mu}_k$ of each community, characterizing the weight overlap degree.
 - (4) Weighted Voting: The dot product results are normalized to obtain voting weights w_k , and heat is assigned to the community with the maximum weight. Ties are broken by prioritizing the community with a larger dot product of high-weight components in $\vec{W}$.
-

As illustrated in Figure 5, the identification process for new heats adopts a community-based “divide-and-conquer” strategy to handle complex charging patterns. The workflow proceeds as follows: First, the test data, comprising multi-dimensional input factors such as scrap weights, hot metal information, and endpoint temperature, is ingested into the system. A Voting Mechanism based on the Knowledge Graph (KG) and label propagation algorithm (LPA) results is then employed to evaluate the similarity between the new heat and the pre-identified communities. The heat is automatically assigned to the community with the highest matching degree, effectively isolating similar process characteristics. Subsequently, the data is routed to the specific Mechanism and Data Joint-Driven Sub-model trained for that community. Within this module, the iron mass balance mechanism computes the theoretical physical baseline, while the neural network predicts the non-linear residual specific to the current furnace condition. Finally, the specific yield rate for each scrap type is dynamically identified by decoupling the learned residual and superimposing it onto the empirical value, realizing precise online identification.

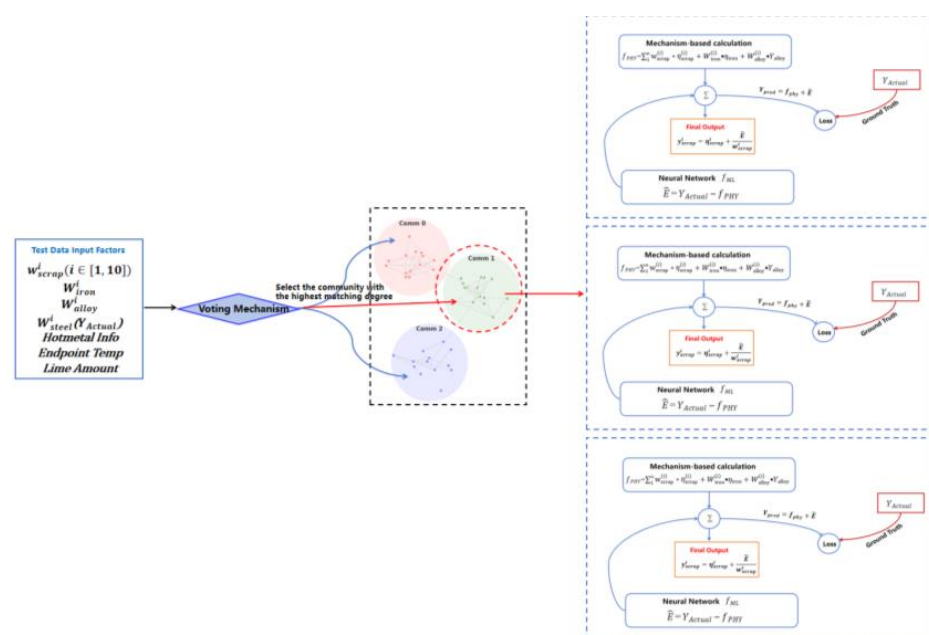


Figure 5. Scrap yield identification for new heats by voting algorithm.

3. Experiments

3.1. Data Preprocessing

All input industrial data adopted in this paper are historical converter smelting records collected from a large Chinese integrated iron and steel enterprise. Based on real data from a steelmaking plant, comprehensive data quality improvement work was carried out, such as missing value removal. For instance, if the value of carbon content in the hot metal is void, this heat will not be taken into account.

3.2. Experimental Setup

The hyperparameter settings of the model are presented in Table 2.

Table 2. Hyperparameters in the neural network.

Hyperparameter	Details
Input Layer	20 input features
Number of Hidden Layers	2

Table 2. *Cont.*

Hyperparameter	Details
Number of Neurons in the First Hidden Layer	256
Number of Neurons in the Second Hidden Layer	128
Activation Function	ReLU function is used in the hidden layers
Number of Training Epochs	2000 epochs, and the early-stopping method is applied
Output Layer	11 neurons, representing the ten types of scrap steel yield rates and the hot metal yield rate
Optimization Algorithm	Adam optimizer
Loss Function	$L = \frac{1}{N} \sum_{i=1}^N \left(Y_{\text{Actual}}^{(i)} - \left(\underbrace{Y_{\text{Base}}^{(i)}}_{\text{Mechanism}} + \underbrace{\hat{E}^{(i)}}_{\text{Data-Driven}} \right) \right)^2$
Batch Size	32
Learning Rate	The initial learning rate is 0.001, and a cosine annealing learning rate adjustment strategy is applied

3.3. Experimental Results

3.3.1. Accuracy of Our Proposed Network

The dataset used in this paper consists of a total of 2650 heats, which are stratified sampled by community distribution into three subsets: 2125 training heats, 260 validation heats, and 265 test heats. Only the 2125 heats from the training subset are utilized to build the scrap charging knowledge graph and execute LPA community detection; validation set and test set samples never participate in graph topology construction or label propagation clustering. During prediction inference, unseen test heats only match existing pre-trained communities via a weighted voting algorithm, which thoroughly avoids data leakage between the training and test samples. The detection communities are shown in Table 3.

Table 3. Community detection in the knowledge graph for scrap charging pattern.

Community Name	Number of Members (Train)	Number of Members (Test)
Comm0	420	74
Comm1	765	108
Comm2	940	83

The test set was divided into 74 heats in Comm0, 108 heats in Comm1, and 83 heats in Comm2 by the voting algorithm. The model without community detection is a single model as a baseline, and the models for community detection are Comm0, Comm1, and Comm2 models. In addition, based on empirical values of the scrap steel yield rate in the plant, a mechanistic model and a data-driven model were designed, respectively. The mechanistic model works by substituting the empirical values of scrap steel yield rate into the iron balance mechanism formula and then calculating the error against the actual molten steel weight. The data-driven model takes the empirical values of the scrap steel yield rate as the ground truth and integrates the furnace condition factors that affect scrap steel yield rate as the input factors.

To ensure the fairness and comparability of the experimental results, all models involved in the comparison (mechanistic model, data-driven model, baseline joint-driven model without community detection, and the proposed joint-driven model with KG community detection) adhere to strict consistency in key experimental settings, except for the

core differences in modeling frameworks and whether community detection is adopted. Specifically, the data-driven approach based on empirical values shares the same model architecture as the proposed model: it employs a Multi-Layer Perceptron (MLP) with two hidden layers. All models use the identical input feature set. For training parameters, consistent hyperparameters are adopted across all models.

The only differences between the models lie in two aspects: (1) Modeling logic: the mechanistic model relies on empirical scrap steel yield rates substituted into the iron balance formula; the data-driven model takes empirical yield rates as ground truth for supervised training; and the joint-driven models (baseline and proposed) integrate iron balance mechanism constraints into the loss function. (2) Community detection: the proposed model adopts KG-based community detection and separate modeling for each community, while the other three models do not involve community partitioning. This consistent experimental design eliminates the interference of factors such as model complexity, input feature discrepancy, and data distribution difference, ensuring that the superior performance of the proposed model originates from the innovative integration of mechanism-data joint driving and community detection.

To evaluate the prediction performance of the proposed model, two internationally recognized metrics, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), are supplemented. These metrics quantify the model's prediction deviation, average precision, and data fitting ability from multiple dimensions. Their definitions and calculation formulas are as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i)^2} \quad (4)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{y}_i - y_i| \quad (5)$$

where N denotes the number of test samples, $\hat{y}_i$ is the predicted iron content of the i -th heat, and y_i is the true iron content from the molten steel amount. Table 4 and Figures 6–9 summarize the extended evaluation metrics results of the mechanistic model, the data-driven model, the baseline single model (without community detection), and the proposed model (with community detection). It can be observed that the proposed model achieves a lower RMSE and MAE compared with the baseline model, indicating smaller overall prediction deviations and higher average precision. These results collectively validate the robustness and superiority of the proposed model from multiple evaluation perspectives.

Table 4. Comparison of RMSE, MAE, and hit rate metrics across different models.

Model	RMSE	MAE	Hit Rate Within ±1 Ton	Hit Rate Within ±3 Ton	Hit Rate Within ±5 Ton
Mechanistic calculation based on empirical values of scrap steel yield rate	12.70	9.84	12.08%	32.08%	50.57%
Data-driven approach based on empirical values of scrap steel yield rate	8.83	6.79	9.81%	41.13%	58.49%
Mechanism and data joint-driven identification model for scrap steel yield rate (baseline)	4.47	3.57	17.74%	49.81%	73.58%
Mechanism and data joint-driven identification model for scrap steel yield rate (with KG community detection)	3.20	2.62	24.15%	64.15%	87.55%

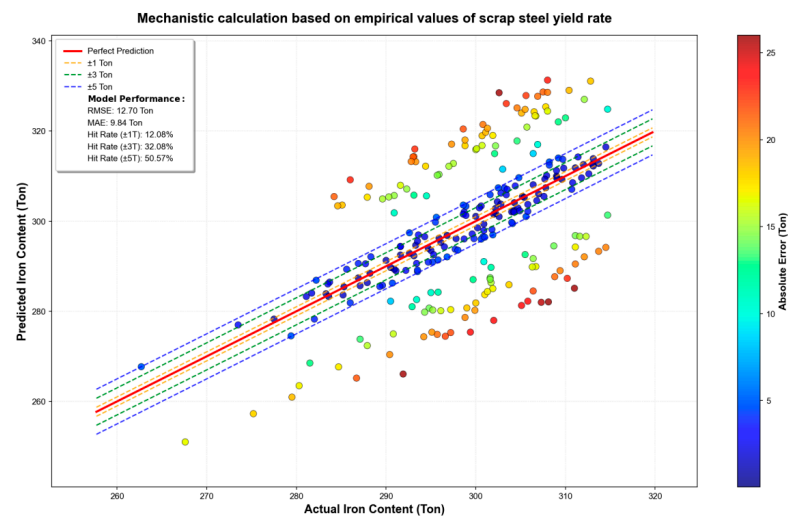


Figure 6. Comparison of actual and predicted iron content based on mechanistic calculation (empirical values of scrap steel yield rate).

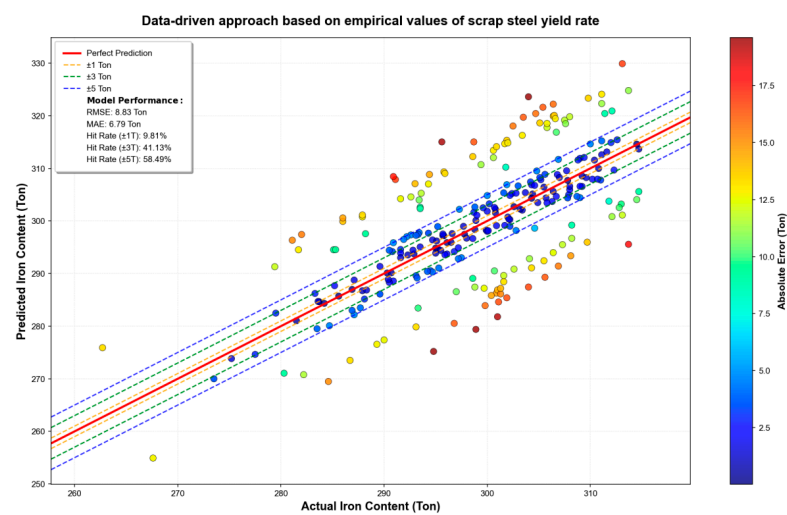


Figure 7. Comparison of actual and predicted iron content based on data-driven approach (empirical values of scrap steel yield rate).

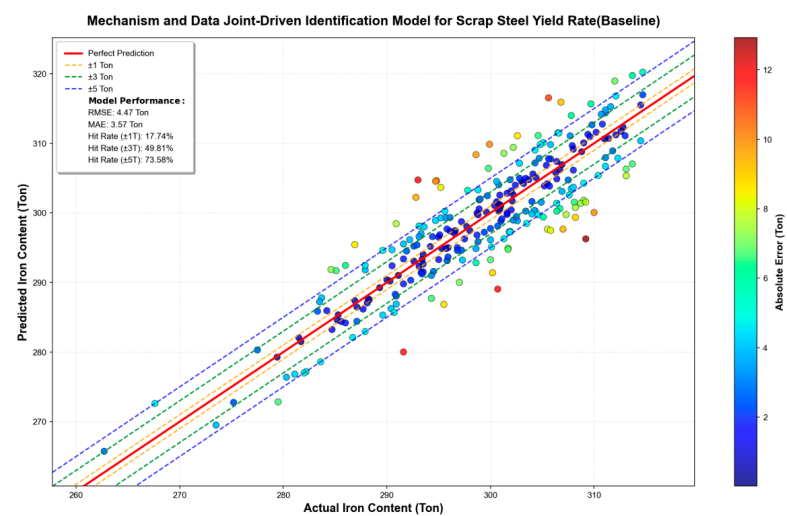


Figure 8. Comparison of actual and predicted iron content without community detection.

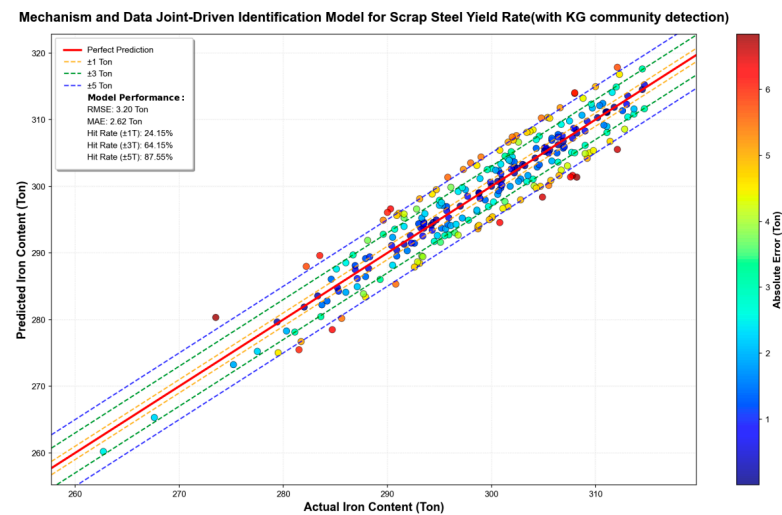


Figure 9. Comparison of actual and predicted iron content based on community detection.

All comparative models are independently trained and tested for 10 repeated experiments under identical hyperparameter configuration. The RMSE and MAE values shown in Table 4 are the average results of 10 repeated runs. The mean and standard deviation of each model's error metrics are summarized in Table 5 to prove the stability of model performance.

Table 5. Mean and standard deviation of RMSE and MAE over 10 repeated experiments.

Model	RMSE	RMSE Std (t)	MAE Mean (t)	MAE Std (t)
Mechanistic calculation based on empirical values of scrap steel yield rate	12.70	0.41	9.84	0.33
Data-driven approach based on empirical values of scrap steel yield rate	8.83	0.35	6.79	0.28
Mechanism and data joint-driven identification model for scrap steel yield rate (baseline)	4.47	0.22	3.57	0.17
Mechanism and data joint-driven identification model for scrap steel yield rate (with KG community detection)	3.20	0.13	2.62	0.11

To further verify the effectiveness of the community partition strategy, we separately calculate the RMSE and MAE of three independent community sub-models, and all community models achieve lower prediction error than the undivided baseline single model. Detailed metrics of each community are listed in Table 6.

Table 6. Community detection in the knowledge graph for scrap charging pattern.

Community Name	Test Heat Quantity	RMSE (t)	MAE (t)
Comm0	74	3.36	2.75
Comm1	108	3.12	2.54
Comm2	83	3.25	2.66
Undivided baseline model	265	4.47	3.57

To further verify the application capability of the proposed method, we compare the single-heat average inference latency of scrap yield prediction models from recent literature and our community-divided joint-driven model. All tests are conducted on a

laptop equipped with an Intel i7-12700H CPU and 16 GB RAM under identical computing environments. The statistical time consumption results are listed in Table 7.

Table 7. Average single-heat inference time of different prediction models.

Model	Average Single Heat Inference Time (ms)
Mechanistic calculation based on empirical values of scrap steel yield rate	0.32
Data-driven approach based on empirical values of scrap steel yield rate	16.82
Mechanism and data joint-driven identification model for scrap steel yield rate (baseline)	22.37
Mechanism and data joint-driven identification model for scrap steel yield rate (with KG community detection)	18.15

Compared with the traditional undivided PINN model, our approach only adds negligible vector cosine similarity matching overhead during community voting matching, with an extra delay of merely 5.78 ms per heat. The total single-heat inference time is controlled within 0.02 s, which fully satisfies the real-time online calculation demands of converter continuous production.

3.3.2. SHAP Analysis with Mechanism and Data Joint-Driven Identification Model for Scrap Steel Yield Rate

The SHAP value, based on the principle of fair distribution in game theory, quantifies the marginal contribution of each input factor to the model identification results, thereby revealing the influence direction and degree of key process parameters on the scrap steel yield rate.

According to the calculation formula for the average scrap steel yield rate of each heat, $\bar{y}_{\text{scrap}} = \frac{\sum_1^n w_{\text{scrap}}^{(i)} y_{\text{scrap}}^{(i)}}{\sum_1^n w_{\text{scrap}}^{(i)}}$, a SHAP analysis is carried out on the model for the scrap steel yield rate and the factors affecting the scrap steel yield rate, as shown in Figure 10. The results show that the factors affecting the scrap steel yield rate can be divided into two categories: positive correlation and negative correlation.

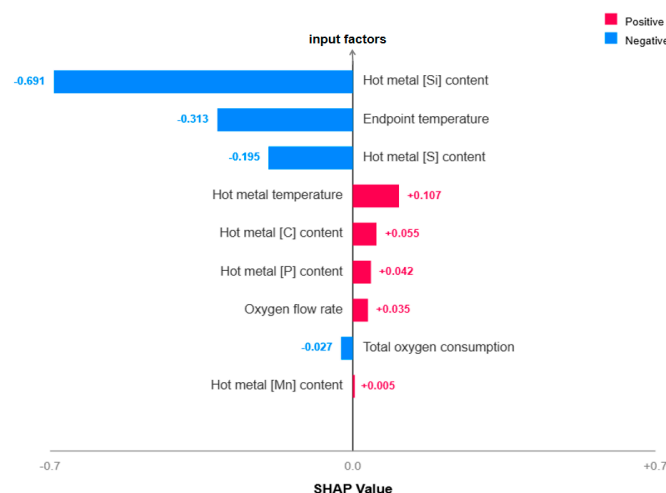


Figure 10. Directions and degrees of influence of process parameters on scrap steel yield rate.

For the temperature of hot metal, the carbon content in hot metal, the phosphorus content in hot metal, the oxygen flow rate, and the manganese content in hot metal, the SHAP values are positive, indicating that these factors have a positive correlation with the scrap steel yield rate. High-temperature hot metal provides sufficient physical heat for con-

verter smelting and significantly accelerates the melting process of scrap steel. On the one hand, sufficient heat supply can reduce the residue of unmelted scrap steel and avoid the scrap steel being wrapped by the slag due to insufficient temperature. On the other hand, high temperature promotes the dissolution of lime and the improvement of slag fluidity, reducing the probability of scrap steel loss due to insufficient slag-steel reaction. The SHAP values show that the temperature of hot metal is one of the core factors for improving the scrap steel yield rate; carbon is an important heat-generating element in the converter. High-carbon hot metal releases a large amount of chemical heat through the oxidation reaction during the oxygen blowing process, indirectly providing energy for the melting of scrap steel. In addition, the CO bubbles generated by the carbon–oxygen reaction can strengthen the stirring of the molten pool and promote the heat and mass transfer between the scrap steel and the hot metal [3]. From the perspective of metallurgical thermodynamics, high-temperature hot metal provides essential physical heat, while the oxidation of carbon and manganese releases critical chemical heat. This enthalpy input significantly accelerates the melting kinetics of scrap steel. Sufficient heat supply reduces the residue of unmelted scrap (skulls) and prevents scrap from being entrapped by viscous slag due to supercooling. Furthermore, high temperatures promote lime dissolution and improve slag fluidity, thereby reducing physical iron loss caused by poor slag-metal separation.

For silicon content in hot metal, endpoint temperature, total oxygen consumption and sulfur content in hot metal, the SHAP values are negative, indicating a negative correlation with the scrap steel yield rate. To improve the scrap steel yield rate, these values should be reduced. Although silicon oxidation is exothermic, the generated SiO_2 requires substantial lime for neutralization, significantly increasing slag volume. High-silicon hot metal tends to form a high amount of slag, impeding full contact between scrap steel and the molten pool and reducing melting efficiency [22]. Excessive total oxygen consumption reflects an over-oxidized state of molten steel, causing iron elements in scrap steel to be oxidized into FeO and enter the slag, resulting in chemical losses [23]. The negative SHAP values suggest that strict control of the oxygen blowing endpoint is necessary to mitigate the adverse effects of oxidative slag on the scrap steel yield rate. Sulfur must be removed by creating high-basicity slag. High-sulfur hot metal requires more lime to form CaS , increasing slag volume and physical losses from slag entrainment. An excessively high endpoint temperature is often accompanied by over-blowing, intensifying steel oxidation and slag thinning, which not only increases iron losses but also enhances slag's encapsulation of scrap steel.

4. Summary and Prospect

This paper developed a method for identifying scrap steel yield rates that integrates a mechanism and data joint-driven approach. By constructing the scrap charging pattern knowledge graph, using the LPA algorithm to find community, and establishing a physics-informed neural network model based on the iron balance mechanism, the identification of the hot metal yield rate and the scrap steel yield rate is achieved. The SHAP analysis method is utilized in this paper to enhance the interpretability of the model from the perspective of feature contribution. The proposed mechanism and data joint-driven identification model for scrap steel yield rates based on knowledge graph community detection achieves the highest identification precision, with an RMSE of 3.20 tons and an MAE of 2.62 tons. Compared with the baseline mechanism and data joint-driven identification model for scrap steel yield rates (without community detection), the proposed method reduces the MAE by 26.6% (from 3.57 t to 2.62 t) and significantly improves the hit rate within ± 5 tons by 13.97 percentage points (reaching 87.55%). These improvements validate the effectiveness of the community-based divide-and-conquer strategy in handling complex charging patterns.

It has important application value for promoting the green and low-carbon transformation of the iron and steel industry. Although the physics-informed neural network can achieve a high accuracy in identifying scrap steel yield rates, there are still some limitations and challenges: This model is trained based on ten types of scrap steel involved in the historical data of a certain iron and steel plant. Therefore, when applied to identify the scrap steel yield rate of new or previously unencountered scrap steel, transfer learning or retraining of the model may be required. The model does not quantitatively decouple error transmission from different types of measuring instruments, and the prediction fluctuation caused by heterogeneous sensor precision is not independently modeled, which will be optimized in follow-up research.

In future research, we will further optimize the model structure to enhance identification accuracy and stability, so as to better serve the actual steel production. For heats with abnormal conditions (e.g., over-blowing or insufficient oxygen blowing), targeted incremental learning will be conducted in the future after more abnormal heats are collected, given the limited amount of data currently available for such abnormal conditions. At the same time, more production link data will be considered for incorporation into the model to achieve comprehensive optimization and intelligent management and control of the steel production process.

Author Contributions: Conceptualization, Y.L., H.X., and H.W.; methodology, Y.L., D.H., and H.W.; validation, Y.L.; formal analysis, Y.L., D.H., and H.W.; investigation, Y.L.; resources, H.X.; data curation, H.X. and H.W.; writing—original draft, Y.L.; writing—review and editing, Y.L.; supervision, D.H. and H.W.; project administration, D.H. and H.W.; funding acquisition, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Science and Technology Major Project (2022ZD0119201).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections.

Conflicts of Interest: Author Haotian Xu was employed by the company Steel Industry Green and Intelligent Manufacturing Technology Center, China Iron and Steel Research Institute Group. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. The company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Lv, M.; Chen, S.; Yang, L.; Wei, G. Research Progress on Injection Technology in Converter Steelmaking Process. *Metals* **2022**, *12*, 1918. [\[CrossRef\]](#)
2. Hu, X.; Wang, C.; Lim, M.K.; Koh, S.C.L. Characteristics and community evolution patterns of the international scrap metal trade. *J. Clean. Prod.* **2020**, *243*, 118576. [\[CrossRef\]](#)
3. Zong, Y.; Wang, Z.; Liu, X.; Nian, Y.; Pan, J.; Zhang, C.; Wang, Y.; Chu, J.; Zhang, L. Judgment of blast furnace iron-tapping status based on data differential processing and dynamic window analysis algorithm. *Prog. Nat. Sci. Mater. Int.* **2023**, *33*, 450–457. [\[CrossRef\]](#)
4. Colla, V.; Zaccara, A.; Dettori, S.; Laid, L.; Martino, I.; Cateni, S.; Branca, T.A.; Vannini, L. Decreasing the environmental impact of the electric steelmaking route through advanced modelling techniques. *Steel Res. Int.* **2026**, *97*, e202501090. [\[CrossRef\]](#)
5. Mukhlis, R.; Satritama, B.; Brooks, G.; Rhamdhani, M.A. Steel scrap: Strategic commodity for green steel circularity and analyses of tramp elements accumulation. *J. Sustain. Metall.* **2026**, *12*, 2705–2740. [\[CrossRef\]](#)
6. Jaimes, W.; Maroufi, S. Sustainability in steelmaking. *Curr. Opin. Green Sustain. Chem.* **2020**, *24*, 42–47. [\[CrossRef\]](#)
7. Wang, J.; Zhu, W.; Zhang, H.; Wang, J.; Lu, P.; Fang, Q. Effect of Asymmetric Bottom Blowing on Melting Behavior of Steel Scrap in a Converter. *Metall. Mater. Trans. B* **2024**, *55*, 3208–3221. [\[CrossRef\]](#)

8. Singha, P. Scrap dissolution effect in BOF converter process. *Ironmak. Steelmak.* **2023**, *50*, 1434–1442. [[CrossRef](#)]
9. Deng, S.; Xu, A.-J. Steel scrap melting model for a dephosphorisation basic oxygen furnace. *J. Iron Steel Res. Int.* **2020**, *27*, 972–980. [[CrossRef](#)]
10. Ding, J.; Long, F. An automatic measurement system for scrap steel yield based on medium frequency furnace. *Shanxi Metall.* **2024**, *47*, 189–192+195. (In Chinese) [[CrossRef](#)]
11. He, R.; Zhang, Z.; Yang, J.; Lin, X. Accurate measurement and cost-benefit evaluation of scrap steel yield. *Henan Metall.* **2022**, *30*, 35–38. (In Chinese)
12. Gan, J.; Zhang, L.; Long, H.; Xia, J.; Zhang, B.; Zhang, C. Research and practice of converter energy model based on online dynamic scrap steel yield. *J. Iron Steel Res.* **2023**, *35*, 1483–1495. (In Chinese) [[CrossRef](#)]
13. Zhang, C.; Nian, Y.; Zhang, L.; Cheng, J.; Zhang, Z. Steel Scrap Yield Prediction in Basic Oxygen Steelmaking Based on Random Forest and Neural Networks. *Steel Res. Int.* **2025**, *96*, 2400713. [[CrossRef](#)]
14. Sandberg, E.; Lennox, B.; Undvall, P. Scrap management by statistical evaluation of EAF process data. *Control Eng. Pract.* **2007**, *15*, 1063–1075. [[CrossRef](#)]
15. Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [[CrossRef](#)]
16. Zong, S.Y.; Yang, Z.; Li, J.Y. SiMBA-augmented physics-informed neural networks for industrial remaining useful life prediction. *Machines* **2025**, *13*, 452. [[CrossRef](#)]
17. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [[CrossRef](#)]
18. Jia, X.; Willard, J.; Karpatne, A.; Read, J.S.; Zwart, J.A.; Steinbach, M.; Kumar, V. Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles. *ACM/IMS Trans. Data Sci.* **2021**, *2*, 1–26. [[CrossRef](#)]
19. Xia, Y.; Wang, H.; Xu, A. dmPINNs: An Integrated Data-Driven and Mechanism-Based Method for Endpoint Carbon Prediction in BOF. *Metals* **2024**, *14*, 926. [[CrossRef](#)]
20. Wen, N.; Wang, H.; Li, L.; Xu, A. Prediction model of the end-point carbon content in BOF by constructing the graph of heats. *ISIJ Int.* **2025**, *65*, 613–618. [[CrossRef](#)]
21. Su, X.; Xue, S.; Liu, F.; Wu, J.; Yang, J.; Zhou, C.; Hu, W.; Paris, C.; Nepal, S.; Jin, D.; et al. A Comprehensive Survey on Community Detection with Deep Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 4682–4702. [[PubMed](#)]
22. de Souza, R.M.; Andreatta, V.; Santos, I.A.S.; Junca, E.; Grillo, F.F.; de Oliveira, J.R. Influence of process parameters, hot metal silicon content and slag properties on steel dephosphorization. *J. Mater. Res. Technol.* **2021**, *15*, 5307–5315. [[CrossRef](#)]
23. He, Z.; Hu, X.; Chou, K.-C. Oxidative modification of industrial basic oxygen furnace slag for recover iron-containing phase: Study on phase transformation and mineral structure evolution. *Process Saf. Environ. Prot.* **2023**, *171*, 167–175. [[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.