BMC Bioinformatics

RESEARCH

Open Access

FeSseqdb: a curated sequence-level database and interpretable machine learning framework for identifying iron–sulfur proteins

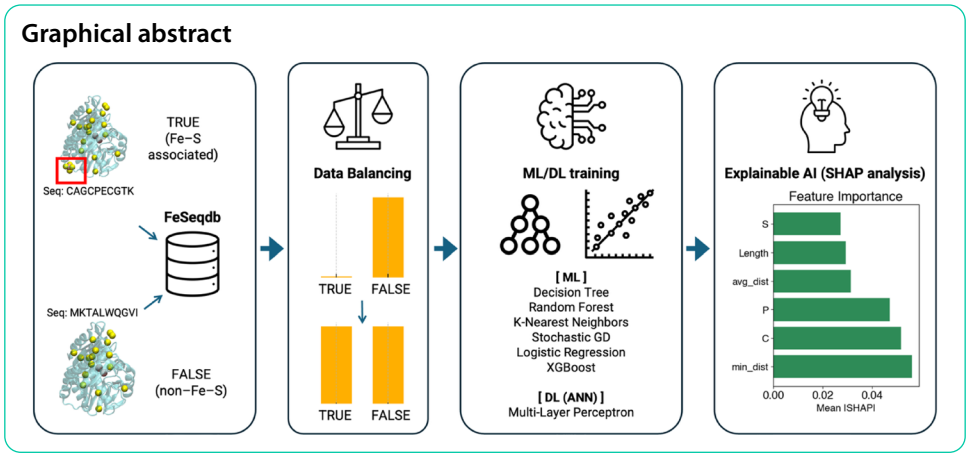
Jiyeon Min[1,2*], Bernard R. Brooks[1] and Muhamed Amin[1*]

*Correspondence:
Jiyeon Min
jmin7@umd.edu
Muhamed Amin
muhamed.amin@nih.gov
[1]Laboratory of Computational Biology, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD 20892, USA
[2]Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA

Abstract

Iron–sulfur (Fe–S) clusters are ubiquitous cofactors in metalloproteins, supporting essential biological functions such as electron transfer, enzymatic catalysis, and metabolic regulation. Despite their importance, large-scale identification of Fe–S proteins remains challenging due to limitations of experimental methods and inconsistent annotations in existing databases. To address this, we introduced FeSseqdb, a curated sequence-level database derived from the Protein Data Bank (PDB), in which Fe–S cluster-containing chains are systematically verified using atomic coordinates. By standardizing diverse ligand annotations, FeSseqdb provides a unified and reliable resource for Fe–S protein research. Building on this foundation, a machine learning framework was developed to predict Fe–S proteins using only sequence-derived features, including amino acid composition, cysteine-related metrics, and sequence length. Among the classifiers evaluated, the random forest model trained on data resampled with SVMSMOTE achieved the highest predictive performance, highlighting the discriminative power of simple sequence features. To elucidate the biological relevance of these features, explainable AI methods were applied to identify key sequence characteristics associated with Fe–S proteins. Cysteine frequency and spatial distribution, along with proline content, emerged as primary contributors, which is consistent with their known structural roles in cluster coordination. Additionally, serine, glutamic acid, and arginine were identified as secondary determinants, in line with their reported roles in redox and electrostatic environments surrounding metal cofactors. The inclusion of these biologically relevant features demonstrates the potential of sequence-based models not only for accurate prediction but also for uncovering functional insights that align with known biochemical principles. This approach provides a foundation for large-scale, sequence-based discovery of Fe–S proteins and supports future investigations into their functional diversity across proteomes.

Keywords Metalloproteins, Iron-sulfur proteins, Explainable AI, Protein Data Bank, Sequence-based prediction



Introduction

Iron–sulfur (Fe–S) clusters are indispensable prosthetic groups in metalloproteins [1] and play essential roles in electron transfer, enzyme catalysis, and metabolic regulation [2–4]. These clusters, composed of iron, sulfur, and occasionally nitrogen atoms in specific geometric arrangements, are tightly integrated into proteins, supporting both structural integrity and functional stability [5]. Their stabilization relies on precise interactions with key amino acid residues, particularly cysteines [6], within the broader context of the surrounding protein environment. Despite their biological importance, the sequence-derived features that govern Fe–S cluster incorporation remain insufficiently understood, complicating the identification and analysis of Fe–S proteins across diverse proteomes. This knowledge gap has significant implications for health and disease [7, 8]. Disruptions in Fe–S cluster assembly or function can cause redox imbalances and protein dysfunction, driving conditions such as Friedreich's ataxia, mitochondrial disorders, and neurodegenerative diseases such as Parkinson's disease [9–11]. Given the essential roles that Fe–S proteins play in these diseases, reliable methods to detect and characterize them are urgently needed. Beyond scientific curiosity, understanding these clusters is fundamental to advancing targeted diagnostics and therapeutics, offering hope for combating Fe–S-related diseases [12].

Traditional experimental techniques for Fe–S protein identification, such as X-ray crystallography, paramagnetic NMR, and EPR spectroscopy, remain the gold standard for confirming the presence of Fe–S clusters [13–16]. These methods provide high-resolution structural insights and definitive confirmation, but they are costly, labor-intensive, and unsuitable for large-scale screening. Consequently, the number of experimentally resolved Fe–S protein conformations is limited. To address this scarcity, computational approaches have been developed, including conventional bioinformatics methods that detect Fe–S clusters by searching for conserved cysteine motifs or matching local coordination geometries to those in known metalloproteins [17–19], and deep learning methods that process entire sequences or three-dimensional structures to capture more complex and noncanonical binding patterns [20–22].

However, despite these advances, each approach has inherent limitations [23, 24]. Bioinformatics methods are inherently local in scope, focusing on short sequence motifs or specific coordination geometries and often overlooking the broader structural context essential for Fe–S cluster stability [25]. As a result, they frequently miss noncanonical

Fe–S proteins that lack well-defined motifs and those with atypical cluster geometries that deviate from established coordination patterns. In contrast, deep learning methods can integrate information from entire sequences or structures, enabling the recognition of more complex and noncanonical binding patterns [21, 22]. Nevertheless, their accuracy still depends heavily on the characteristics of the training data, most of which contain predominantly canonical Fe–S cluster geometries, thereby limiting the ability to reproduce atypical arrangements.

A notable example of these limitations is the recent application of structure-prediction tools such as AlphaFold and RoseTTAFold to identify Fe–S binding sites from predicted protein models, which has significantly expanded the apparent metalloproteome [20, 22]. However, these models often fail to fully capture the complex ion–protein interactions required for Fe–S cluster binding, and because their detections are based on predicted coordinates, they are susceptible to false positives if the underlying models do not reproduce the precise geometry of Fe–S cluster coordination [26].

In light of these challenges, there is a clear need for datasets built exclusively from experimentally determined structures, prioritizing reliability over the indiscriminate expansion of databases. By incorporating consistent geometric and chemical definitions to verify the presence of Fe–S clusters and adopting a chain-level perspective, such resources can capture finer structural detail, retain both canonical and atypical clusters present in the structural record, and provide a more accurate foundation for large-scale Fe–S protein characterization.

To address the need for more efficient and scalable solutions for Fe–S protein identification, this paper introduces FeSseqdb, a curated database specifically designed for chain-level analysis of Fe–S proteins. Unlike resources that depend on existing database annotations, which can be missing or inconsistently represented, or on predicted structures, FeSseqdb systematically verifies the presence of Fe–S clusters using atomic coordinates from experimentally resolved structures in the Protein Data Bank (PDB). By focusing on individual protein chains rather than entire protein entries, FeSseqdb provides finer granularity, as proteins often contain multiple Fe–S clusters associated with distinct chains. This coordinate-based verification minimizes the impact of incomplete annotations, enabling more accurate classification of Fe–S proteins. Each chain is labeled according to its proximity to Fe–S clusters, resulting in a robust and reliable dataset optimized for machine learning applications.

Machine learning models trained using FeSseqdb utilize only sequence-derived features, such as amino acid composition, sequence length, and cysteine distribution, to classify sequences based on their association with Fe–S clusters. Explainable AI [27, 28] further enhances interpretability by pinpointing the sequence features that are most critical for classifying a chain as Fe–S adjacent, offering additional biological insights into residues that promote cluster coordination and stability. This chain-level perspective captures the interplay between the entire polypeptide and Fe–S cofactors, offering new insights into how protein sequence and metal cofactor biology are tightly linked [29, 30].

In summary, this paper presents FeSseqdb, combined with machine learning and explainable AI, as an accurate, scalable framework for Fe–S protein identification. Moreover, by integrating interpretability, this study reveals the key sequence-level features that govern Fe–S cluster coordination and stability, particularly the roles of cysteine and

other underexamined residues such as proline. This interpretability not only improves confidence in predictive models but also offers deeper biological insights, underscoring how local and global sequence patterns contribute to Fe–S cluster stability and function. Ultimately, our framework aims to advance Fe–S protein research, paving the way for transformative applications, such as identifying novel therapeutic targets for Fe–S protein-associated diseases or improving diagnostics for Fe–S-related disorders.

Methods

Iron–sulfur sequence database (FeSseqdb)

FeSseqdb is a curated database of protein chain sequences associated with iron–sulfur (Fe–S) clusters, built from all existing protein data in the RCSB PDB (RCSB.org) [31]. Rather than focusing on the clusters themselves, FeSseqdb focuses on sequences adjacent to Fe–S clusters, enabling detailed analysis of their unique characteristics and interactions. The overall two-stage workflow, including database construction and downstream machine learning analysis, is illustrated in Fig. 1. The database covers a wide variety of Fe–S clusters, including common types such as [2Fe–2S], [3Fe–4S], and [4Fe–4S], as well as rarer clusters, such as those shown in Fig. 2. By leveraging reliable experimental data, FeSseqdb supports machine learning models for accurately identifying Fe–S proteins and discovering key sequence-level features, such as residues involved

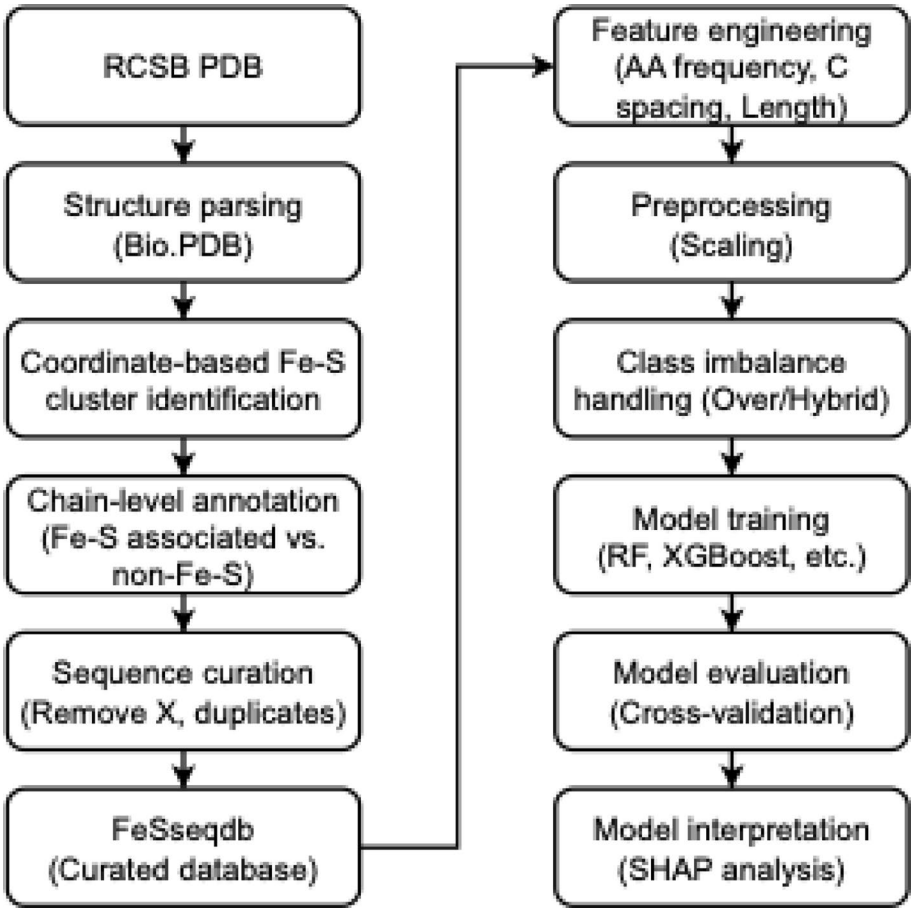


Fig. 1 Overview of the two-stage workflow: (i) construction of FeSseqdb, a curated Fe–S protein dataset, and (ii) machine learning–based Fe–S protein classification and model interpretation

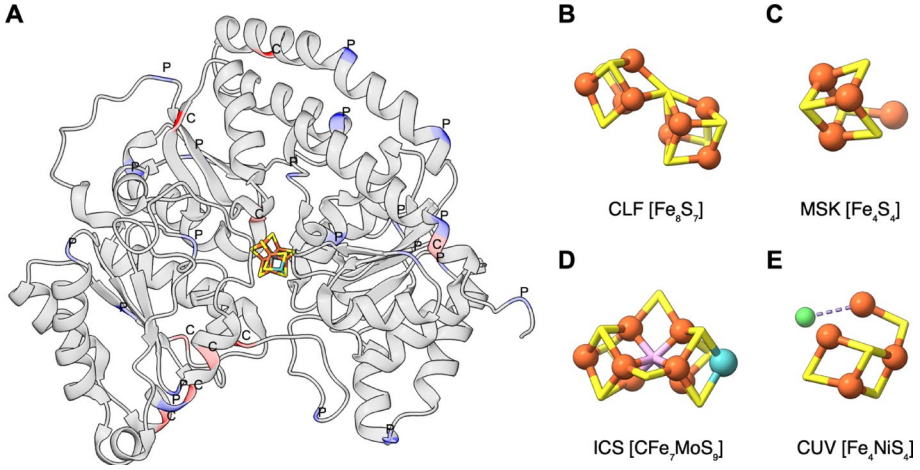


Fig. 2 Representative structures of Fe–S clusters and their coordinating environments. Fe atoms are shown in orange, S atoms in yellow, Mo atom in cyan, and Ni atom in green. **A** ICS ([CFe_7MoS_9], PDB: 8BTS) shown within the protein chain, with cysteine residues (C, red) and proline residues (P, blue) highlighted. **B** CLF ([Fe_8S_7], PDB: 7MCI). **C** MSK (degraded [Fe_4S_4] cluster, PDB: 6T7J). **D** ICS ([CFe_7MoS_9], PDB: 6OP2). **E** CUV ([Fe_4NiS_4], PDB: 6B6W)

in Fe–S cluster coordination and stability. At the time of analysis, the PDB contained 224,905 structures comprising 926,025 sequences. After removal of 53,153 nucleic acid sequences, 872,872 protein chain sequences were retained. These sequences were classified into Fe–S–associated and non–Fe–S classes, yielding 12,668 and 860,204 sequences, respectively. The 12,668 Fe–S–associated sequences corresponded to 5304 Fe–S proteins, and the distribution of Fe–S cluster types among these proteins is summarized in Supplementary Table S1. Duplicate sequences were removed independently within each class. To prevent overlap between classes, 216 identical sequences present in both datasets were excluded from both. In addition, 376 structures containing undefined amino acid residues (denoted as 'X') were removed. The final curated dataset comprised 160,247 protein sequences, including 2261 Fe–S sequences and 157,986 non–Fe–S sequences. The database is publicly available in CSV format via Zenodo (https://doi.org/10.5281/zenodo.14911572) and served as fixed input for all downstream analyses. No machine learning was involved in dataset curation itself.

Data collection

Experimental methods for detecting Fe–S clusters can disrupt their bonding or alter their geometric arrangement. Therefore, more permissive criteria were applied for identification. In the RCSB PDB, Fe–S clusters are recorded under a variety of ligand IDs. The most common types include SF4 (Fe_4S_4), FES (Fe_2S_2), and F3S (Fe_3S_4). However, to ensure comprehensive coverage, less common clusters were also considered, including ICS ([CFe_7MoS_9], PDB: 8BTS, Fig. 2A), CLF ([Fe_8S_7], PDB: 7MCI, Fig. 2B), MSK (degraded [Fe_4S_4], PDB: 6T7J, Fig. 2C), ICS ([CFe_7MoS_9], PDB: 6OP2, Fig. 2D), and CUV ([Fe_4NiS_4], PDB: 6B6W, Fig. 2E). Previous studies have reported that Fe–S proteins are sometimes misannotated in databases [32]. This issue is partly due to the inconsistent use of ligand IDs across the PDB, which makes it difficult to retrieve all relevant entries without accounting for every possible ID.

The data collection process for identifying Fe–S proteins involves two main steps: the identification of Fe–S clusters from structural data and the extraction of sequences from protein chains adjacent to these clusters. Fe–S clusters were identified by analyzing

three-dimensional atomic coordinate data from protein structures in PDB or CIF formats, depending on availability. For each protein, structure files were parsed using the Bio.PDB module [33] to locate Fe atoms. Atoms within a 2.5 Å radius of these Fe atoms were mapped, with a specific focus on sulfur atoms originating from cysteine residues or inorganic sulfides, as well as other potential coordinating atoms (O and N). Clusters where iron was exclusively connected to methionine sulfur were excluded, as these clusters are characteristic of heme-binding proteins rather than Fe–S proteins. In contrast, clusters containing methionine sulfur were retained only when additional non-methionine coordinating atoms, particularly sulfur ligands, were present within this cutoff. This filtering process refined the dataset to include only genuine Fe–S clusters. Protein sequences were then extracted from chains containing the nearest non-heteroatom residue to the identified Fe–S clusters, classifying them as "true" Fe–S proteins. For clusters located at the interface of multiple chains, each sequence from all adjacent chains was included. This approach ensured that the collection accurately captured sequences directly associated with Fe–S clusters, generating a refined resource suitable for studying Fe–S proteins.

Data preprocessing for training

To prepare the FeSseqdb dataset for effective machine learning training, preprocessing techniques were applied to reduce sequence redundancy and annotation ambiguity. Ambiguous entries that could compromise the model's accuracy were carefully reviewed and removed, allowing the focus to remain on accurately distinguishing Fe–S proteins. For example, sequences with unresolved residues, represented as "X" (e.g., "XXXXXXX"), were excluded, as these residues indicate undetermined amino acids. Duplicate sequences were identified and removed both within and across Fe–S and non–Fe–S classes. For instance, the sequences for 5LMC (a protein containing an Fe–S cluster) and 4D02 (a protein without an Fe–S cluster) were found to be identical; therefore, both entries were removed to prevent conflicting annotations from affecting model training. These preprocessing steps resulted in a curated dataset with reduced ambiguity and redundancy, suitable for accurate Fe–S protein classification.

Feature selection

To classify Fe–S proteins effectively while minimizing the risk of overfitting, simple yet biologically relevant sequence-based features were selected. This approach avoids reliance on structural features derived from three-dimensional coordinates, enhancing generalization and reducing computational complexity. The features included the length of the amino acid sequence, which serves as a straightforward metric to represent the overall size of the protein. Additionally, the frequency of all 20 standard amino acids (e.g., C, P, S, etc.) was calculated to assess the impact of individual residues on classification performance. Recognizing the critical role of cysteine residues in stabilizing Fe–S clusters through metal–protein interactions, features related to cysteine distribution were prioritized. Specifically, the minimum (min_dist) and average sequence distances (avg_dist) between cysteine residues were computed, where "sequence distance" refers to the number of amino acid positions between cysteine residues in the sequence. These sequence-based features were selected to ensure biological relevance while maintaining

computational efficiency. By relying solely on sequence data, the method avoids the need for structural information, making it broadly applicable across diverse datasets.

Addressing data imbalance

The Fe–S protein prediction task requires detecting a rare class among a large number of non–Fe–S proteins, resulting in a strongly skewed class distribution. Without explicit imbalance handling, models tend to favor majority-class patterns and show limited sensitivity to Fe–S sequences. Therefore, addressing class imbalance was an essential step in the training procedure. Nine commonly used resampling methods were evaluated, spanning three categories: Baseline (no resampling), undersampling methods including Random Undersampling (RandomUnder), TomekLinks, and Edited Nearest Neighbors (ENN); oversampling methods including Random Oversampling (RandomOver), Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Support Vector Machine–based SMOTE (SVMSMOTE); and hybrid approaches, namely SMOTE–Tomek and SMOTE–ENN. All model–resampling combinations (n = 80) were benchmarked using the imbalanced-learn library in Python [34].

Outlier removal and data standardization

Classification performance was compared under different outlier handling and feature scaling strategies. Outliers were handled using the Interquartile Range (IQR) method. For each numerical feature, values outside the range $Q1 - k \times IQR$ to $Q3 + k \times IQR$ (with $k \in \{1.5, 2.0, 3.0\}$) were clipped to the corresponding boundary values (winsorization), preserving all samples. Baseline models without outlier clipping were also evaluated for comparison. After outlier processing, class imbalance was addressed using SVMSMOTE resampling applied to the training data only. Numerical features were then scaled using either Robust scaling or Min–Max scaling. Robust scaling standardizes features using the median and IQR, making it less sensitive to extreme values, while Min–Max scaling maps features to the [0, 1] range. To determine the optimal preprocessing strategy, the preprocessed data were subsequently used for Random Forest model training. Model performance was evaluated using mean F1-score, area under the receiver operating characteristic (ROC) curve (AUC), and average precision (AP) based on fivefold stratified cross-validation (see Figure S1 for the comparison results).

Classification models

The models described below were trained using the finalized FeSseqdb dataset and were not incorporated into the database as built-in predictive tools. A wide range of machine learning and deep learning algorithms, ranging from simple to complex, were applied to classify Fe–S clusters. Conventional machine learning models included Decision Tree (DT), K-Nearest Neighbors (KNN), Logistic Regression (LR), Random Forest (RF), Stochastic Gradient Descent (SGD), Support Vector Classifier (SVC), and eXtreme Gradient Boosting (XGBoost). Additionally, a multilayer perceptron (MLP) was used as a representative deep learning model. All models were implemented using Python libraries, Scikit-learn [35], and XGBoost [36]. Key hyperparameters for all models are summarized in Supplementary Table S2. Initially, the dataset was split into training and testing sets with an 80/20 ratio, and model performance was evaluated over 80 independent runs using different resampling methods. After selecting the most promising resampling

method, the selected resampling strategy was fixed, and a more robust evaluation was conducted for all 8 models using tenfold cross-validation. In this process, the dataset was divided into 10 equal parts (folds). In each iteration, one-fold served as the test set, whereas the remaining nine folds were used for training.

Feature importance analysis

To enhance the interpretability of the model, SHapley Additive exPlanations (SHAP) was applied. While traditional statistical models, such as linear methods like elastic net, can offer some degree of interpretability through feature selection and coefficient estimation, their linear nature limits their ability to capture non-linear relationships or interactions between features. To address these constraints, SHAP was utilized to interpret the results of non-linear models, particularly tree-based algorithms. SHAP, grounded in game theory, assigns each feature a contribution score, the SHAP value, indicating how much a given feature positively or negatively influences a single prediction. By aggregating SHAP values across a broad dataset, features that consistently influence predictions can be identified, and the varying impact of individual features across different data points can be analyzed. To provide a detailed analysis of feature contributions, TreeExplainer [37] was employed for tree-based models. This study focuses on the analysis of a RF model, which exhibited the highest classification accuracy.

Model performance evaluation metrics

Model performance was evaluated using precision, recall, F1-score, ROC-AUC and AP. These metrics provide a comprehensive assessment of classification performance, particularly in the presence of class imbalance. The metrics are defined as follows:

- *Precision*: The proportion of correctly predicted positive instances among all instances predicted as positive.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- *Recall (sensitivity)*: The proportion of correctly predicted positive instances among all actual positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- *F1-score*: The harmonic mean of precision and recall, providing a single metric to balance both.

$$\text{F1} - \text{Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- *Area under the ROC curve (AUC)*: AUC represents the area under the ROC curve, which plots the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. It measures the model's ability to discriminate between positive and negative classes.

$$\mathrm{AUC} = \int_0^1 TPR\,(FPR)\,d\,(FPR)$$

- *Average precision (AP)*: AP summarizes the precision–recall curve by computing the weighted mean of precision values achieved at different recall levels, with the increase in recall used as the weight.

$$\mathrm{AP} = \sum_n (R_n - R_{n-1}) \cdot P_n$$

where P_n and R_n denote the precision and recall at the n-th threshold, respectively. Together, these metrics offer a robust evaluation of model performance by capturing classification accuracy, discrimination capability, and performance across varying decision thresholds, making them particularly suitable for imbalanced classification tasks.

Results and discussion

The primary objective of this study is to gain biological insights into Fe–S proteins through feature importance analysis, enabled by a machine learning framework that identifies these proteins using only their amino acid sequences. To achieve this goal, we constructed a set of biologically meaningful features derived solely from the sequence, including the frequency of each amino acid and the overall sequence length (see Methods for details). In this section, we begin with a detailed description of the data preprocessing and feature engineering steps, followed by the evaluation of multiple classification models. We then perform a comprehensive analysis of feature importance to identify the most discriminative characteristics that distinguish Fe–S proteins from non–Fe–S counterparts. Finally, we interpret our findings in the context of established biochemical and structural properties of Fe–S proteins, highlighting how our model aligns with known mechanisms of Fe–S cluster stabilization and function.

Outlier removal and data scaling

To evaluate the impact of different outlier handling and feature scaling strategies, we systematically compared multiple preprocessing configurations. Supplementary Figure S1 summarizes their effects on classification performance. Overall, models without explicit outlier clipping achieved equal or superior performance compared to those applying IQR-based clipping, regardless of the clipping threshold. In particular, aggressive clipping (k = 1.5 and 2.0) consistently led to reduced F1-score and AP, indicating potential loss of informative extreme values. Across all configurations, Robust scaling showed stable performance and was less affected by the presence of extreme values than Min–Max scaling. Taken together, these results indicate that explicit outlier removal is unnecessary for this task. The combination of no outlier clipping and Robust scaling consistently yielded strong performance across all evaluation metrics, balancing robustness with the preservation of informative feature variation. Based on these findings, Robust scaling without outlier clipping was selected for the final modeling pipeline.

Impact of resampling methods

The original dataset exhibited severe class imbalance, with 2261 Fe–S sequences and 157,986 non–Fe–S sequences (Fe–S ratio = 0.014). The training subset (80% split:

Table 1 Effect of resampling strategies on training set composition. Numbers show post-resampling counts for the 80% training split (n = 128,197 baseline)

Balancing type	Name	Total	Fe–S	non–Fe–S	Fe–S ratio
None	Baseline	128,197	1809	126,388	0.014
Under	RandomUnder	3618	1809	1809	0.500
Under	TomekLinks	128,070	1809	126,261	0.014
Under	ENN	126,441	1809	124,632	0.014
Over	RandomOver	252,776	126,388	126,388	0.500
Over	SMOTE	252,776	126,388	126,388	0.500
Over	ADASYN	252,559	126,171	126,388	0.500
Over	SVMSMOTE	252,776	126,388	126,388	0.500
Hybrid	SMOTE-Tomek	252,772	126,386	126,386	0.500
Hybrid	SMOTE-ENN	247,590	126,376	121,214	0.510

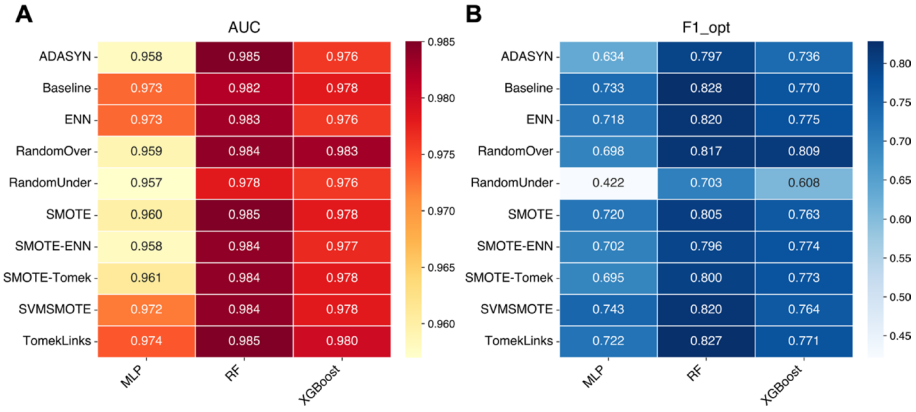


Fig. 3 Performance of MLP, RF, and XGBoost across different resampling strategies. **A** AUC and **B** optimized F1-score. Oversampling and hybrid methods generally improved F1-score while maintaining high AUC (> 0.95). In contrast, undersampling approaches, particularly RandomUnder, preserved AUC but led to reduced F1-score, likely due to information loss caused by discarding majority-class samples

128,197 samples) retained this imbalance under baseline conditions. Table 1 quantifies the effect of each resampling strategy on dataset size and class distribution, while Fig. 3 illustrates performance impacts on top models (MLP, RF, and XGBoost); the full version across all 8 models is in Supplementary Fig. S2. Undersampling methods showed markedly different behaviors. RandomUnder generated a perfectly balanced dataset with 1809 samples per class; however, this led to a drastic reduction of the total sample size to 3618. In contrast, TomekLinks and ENN removed only a small subset of majority-class samples, leaving the overall class distribution nearly unchanged compared to the baseline. All oversampling approaches, including RandomOver, SMOTE, ADASYN, and SVMSMOTE, markedly increased the number of minority-class samples, producing nearly balanced datasets (Fe–S ratio ≈ 0.500). Hybrid methods (SMOTE-Tomek and SMOTE-ENN) also achieved near-balanced distributions. Overall, oversampling and hybrid resampling methods effectively mitigated class imbalance while preserving the majority of the original dataset size.

Model performance

The dataset was divided into 80:20 train/test splits, where 80% of the data was used to train the machine learning models, and 20% was reserved for evaluating their performance. Model performance was evaluated using AUC and F1-score, as shown in Fig. 3.

Table 2 Performance metrics of classification models evaluated using tenfold cross-validation on SVMSMOTE-balanced data (mean ± 1 standard deviation). The metrics include precision, recall, F1-score, AUC, and AP. The following classifiers were assessed: DT, RF, KNN, SGD, LR, XGBoost, SVC, and ANN (MLP)

	DT	RF	KNN	SGD	LR	XGBoost	SVC	ANN (MLP)
Precision	0.73 ± 0.02	0.95 ± 0.01	0.67 ± 0.01	0.53 ± 0.00	0.54 ± 0.00	0.76 ± 0.01	0.49 ± 0.00	0.82 ± 0.03
Recall	0.86 ± 0.01	0.87 ± 0.01	0.92 ± 0.01	0.79 ± 0.02	0.74 ± 0.02	0.91 ± 0.01	0.40 ± 0.02	0.87 ± 0.02
F1-score	0.78 ± 0.01	0.90 ± 0.01	0.73 ± 0.01	0.51 ± 0.01	0.55 ± 0.00	0.82 ± 0.01	0.22 ± 0.01	0.84 ± 0.01
AUC	0.86 ± 0.01	0.99 ± 0.00	0.95 ± 0.01	0.87 ± 0.02	0.87 ± 0.02	0.98 ± 0.01	0.40 ± 0.03	0.97 ± 0.01
AP	0.35 ± 0.03	0.85 ± 0.02	0.54 ± 0.03	0.08 ± 0.01	0.08 ± 0.01	0.80 ± 0.02	0.08 ± 0.04	0.71 ± 0.02

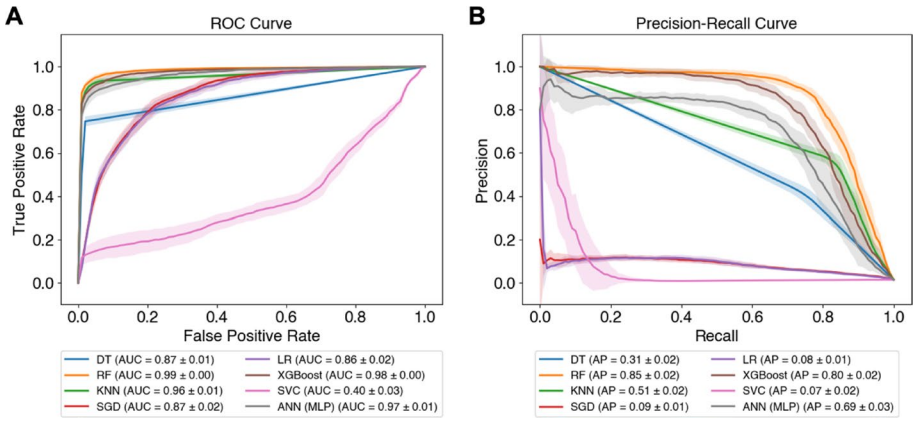


Fig. 4 Performance comparison of classification models using tenfold cross-validation (mean ± 1 standard deviation). **A** ROC curves with AUC values. **B** Precision-recall curves with AP scores. Models evaluated include Decision Tree (DT), Random Forest (RF), K-Nearest Neighbors (KNN), Stochastic Gradient Descent (SGD), Logistic Regression (LR), eXtreme Gradient Boosting (XGBoost), Support Vector Classifier (SVC), and Artificial Neural Network (ANN; implemented as a Multi-Layer Perceptron (MLP))

RF consistently achieved strong performance across resampling methods. Although TomekLinks showed comparable results, it was excluded due to its undersampling nature and potential removal of informative borderline samples. In contrast, SVMSMOTE preserves minority class information by generating synthetic samples near the decision boundary. Consequently, RF with SVMSMOTE, which achieved a high F1-score alongside a strong AUC, was selected for subsequent analyses.

To ensure reliability, each model was validated using tenfold cross-validation on SVMSMOTE-balanced data, and the means and standard deviations for precision, recall, F1-score, AUC, and AP metrics are presented in Table 2. Figure 4 shows the mean ROC curves and Precision-Recall curves with ± 1 standard deviation. The ROC curve assesses the trade-off between the true positive rate and false positive rate, with the Area Under the Curve (AUC) quantifying the area under this curve. A higher AUC indicates better performance in separating positive and negative classes. Similarly, the Precision-Recall curve captures the trade-off between precision and recall, with the AP representing the area under this curve. A higher AP reflects stronger performance, particularly in identifying minority classes within imbalanced datasets. Among all methods, the RF classifier achieved the highest AUC (0.99 ± 0.00) and AP (0.85 ± 0.02), confirming its superior performance across these metrics. While XGBoost and MLP showed competitive AUC values of 0.98 ± 0.01 and 0.97 ± 0.01, respectively, their AP scores were slightly lower. These

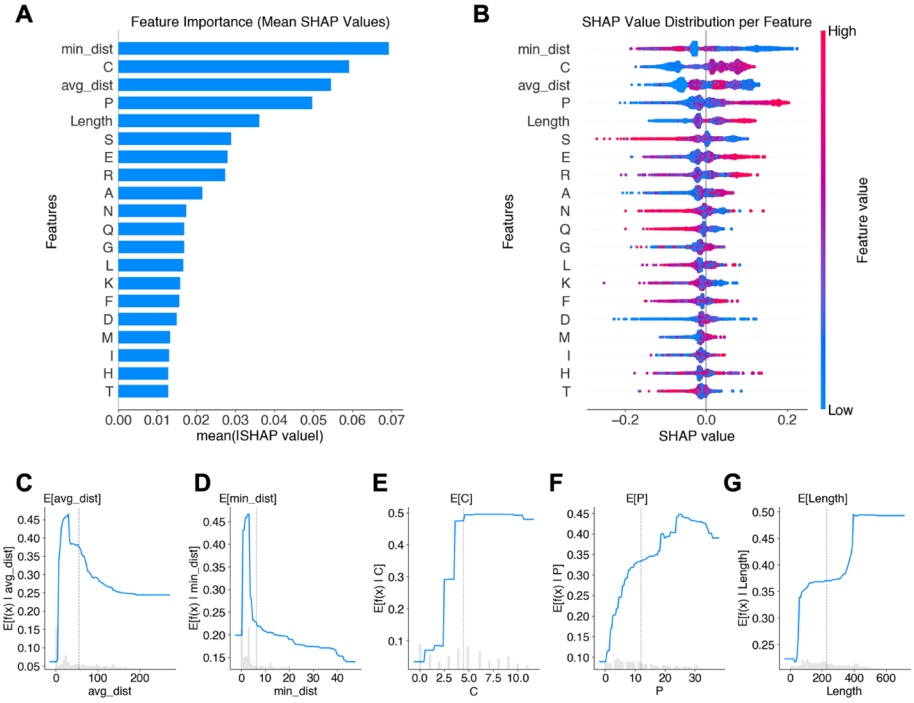


Fig. 5 SHAP analysis results for Fe–S identification: **A** Bar chart of mean SHAP values showing overall feature importance, with avg_dist, min_dist, C, P, and Length identified as the most influential features. **B** SHAP summary plot illustrating feature contributions to Fe–S classification. Each dot represents a sequence, with red/blue indicating high/low feature values and the x-axis representing SHAP values. **C**–**G** SHAP partial dependence plots for the top five features, demonstrating how changes in each feature influence the probability of Fe–S protein classification. The gray histograms represent feature distributions, and the dashed lines indicate the mean feature values

results demonstrate the effectiveness of the RF model, particularly when combined with SVMSMOTE resampling, in delivering robust and reliable classification performance.

Explainable AI: SHAP analysis

SHAP analysis was employed to interpret the model's predictions, focusing on the RF classifier identified in Sect. "Model performance" as the best-performing model. To make SHAP value computation computationally feasible for the tree-based RF model, SHAP values and related plots were computed using a randomly selected subset of 10,000 samples. SHAP-based feature importance results for additional classifiers (XGBoost, DT, and LR) are provided in Supplementary Fig. S3; in these cases, SHAP values were computed using the full training dataset (128,197 samples).

Analysis of key features in Fe–S identification

SHAP analysis identified avg_dist, min_dist, C, P, and Length as the most influential features in Fe–S protein classification (Fig. 5A, B). The SHAP plots, including the summary and partial dependence plots, provide a comprehensive understanding of feature importance and behavior in predictions. This study interprets SHAP results to explain how cysteine proximity, frequency, proline content, and sequence length influence Fe–S protein classification, aligning with known structural and functional requirements. avg_dist and min_dist are complementary indicators of cysteine distribution, capturing the overall clustering tendency (avg_dist) and the local proximity (min_dist) of cysteine residues, respectively. Both measures indirectly reflect the three-dimensional structural

arrangement of cysteines, which is crucial for stabilizing Fe–S clusters. A shorter avg_dist indicates tighter clustering along the sequence, whereas a smaller min_dist reflects greater local proximity, both of which increase the likelihood of Fe–S protein classification. In contrast, larger avg_dist and min_dist values imply a more dispersed distribution that weakens the structural and sequence-based clustering, leading to a decrease in the predicted probability (Fig. 5C, D). These observations align with the biological necessity of closely spaced cysteine residues for forming and stabilizing Fe–S clusters.

Cysteine frequency (C) has also emerged as a critical factor. Proteins with cysteine frequencies greater than approximately 5 showed consistently high prediction probabilities, emphasizing the importance of sufficient cysteine residues to serve as binding sites for Fe–S clusters. Beyond a certain threshold (approximately 5), the effect plateaued, indicating diminishing returns at very high cysteine frequencies (Fig. 5E). Proline frequency (P) contributed in a distinct but complementary manner. Proline had a gradual positive impact on the predictions, with prediction probabilities peaking when proline frequencies were between approximately 20 and 30, increasing the likelihood of classification as Fe–S proteins. This finding suggests that proline supports cysteine clustering, and that the structural context is required for Fe–S cluster formation. At higher proline frequencies, the effect slightly diminished or stabilized, but it remained beneficial overall (Fig. 5F).

Sequence length (Length) displayed a threshold effect. Shorter sequences were associated with lower probabilities of Fe–S protein classification, whereas sequences longer than approximately 400 residues showed a sharp increase in prediction probabilities (Fig. 5G). This reflects the need for adequate sequence length to accommodate the cysteine clustering and structural motifs necessary for Fe–S cluster formation. Although sequence length has a secondary role compared to cysteine-related features, its influence was still notable in providing the structural framework for Fe–S proteins.

Together, these features avg_dist, min_dist, C, P, and Length capture essential sequence properties that indirectly reflect structural characteristics critical for Fe–S protein identification. Cysteine proximity and frequency emerged as the primary drivers of predictions, emphasizing their direct role in coordinating Fe–S clusters. Meanwhile, proline frequency and sequence length acted as complementary features, suggesting an indirect role in Fe–S clustering or a potential correlation with the structural features required for protein conformation. These findings highlight the strength of sequence-based approaches in predicting Fe–S proteins, leveraging sequence-derived features as proxies for underlying structural properties. As a representative case study, we examined MitoNEET, a human [2Fe–2S] protein with non-canonical coordination by three cysteines and one histidine. Single-sequence SHAP analysis showed that cysteine residue spacing features (min_dist and avg_dist) and proline were the dominant contributors, highlighting the importance of local structural geometry beyond cysteine frequency alone (Supplementary Fig. S4).

In addition to partial dependence analysis, dependence analysis (Supplementary Fig. S5) was conducted to explore interactions between the top five features. This analysis underscores the importance of proper scaling during feature preprocessing. Without scaling, inter-feature dependencies were obscured, as demonstrated in Supplementary Fig. S6, emphasizing preprocessing as a prerequisite for robust and interpretable SHAP analysis. Through SHAP analysis, partial dependence analysis effectively highlighted

each feature's individual impact on predictions, whereas dependence analysis provided additional insights into inter-feature interactions and their combined effects on the model's outcomes.

Importance of cysteine and proline in Fe–S proteins

Cysteine (C) and proline (P) residues play crucial roles in both the identification and functionality of Fe–S proteins. SHAP analysis underscored cysteine-related features, such as avg_dist, min_dist, and frequency, and proline frequency as the most influential predictors. These results can also be discussed in the context of the different structural characteristics of cysteine and proline in protein three-dimensional structures.

This finding aligns with the well-known role of cysteine in Fe–S cluster coordination. The thiol (-SH) group of cysteine forms strong coordination bonds with iron atoms, stabilizing Fe–S clusters and ensuring their biological activity. In three-dimensional protein structures, cysteine is typically positioned to directly coordinate metal centers, making its spatial proximity and arrangement critical for Fe–S cluster formation. The strong contribution of cysteine-related features reflects their direct role in stabilizing and coordinating Fe–S clusters, which are essential for protein functionality. Interestingly, proline also emerged as a significant factor in Fe–S protein classification, ranking fourth in importance. Although its role in Fe–S proteins is known [38–40], it has been explored less extensively compared to cysteine.

Unlike cysteine, which directly binds Fe–S clusters through electron-donating functional groups, proline does not participate in direct cluster binding. Instead, its contribution is structural, making a substantial impact on Fe–S cluster stability. Proline is often referred to as a "helix breaker" due to its pyrrolidine ring, which restricts backbone torsion angles and disrupts regular secondary structures [41]. This conformational property leads to local structural distortions, such as bends or kinks, that influence protein backbone geometry without directly engaging in metal coordination. In the context of Fe–S proteins, these distortions introduce well-positioned kinks or loops that optimize the spatial arrangement of critical ligating residues, such as cysteines, around the Fe–S cluster. Thus, proline may influence the local structural environment surrounding Fe–S clusters through its general conformational effects on protein structure. The structural contribution of proline enhances the stability of the cluster-binding region and improves the electron transfer efficiency by orienting redox-active cofactors in a structurally favorable manner. While proline does not directly bind Fe–S clusters, its indirect role in shaping the protein's local environment reinforces Fe–S cluster stability and functionality.

The machine learning model's ability to recognize these residues as key features validates its biochemical relevance, even though it was trained solely on sequence data. By capturing the sequence-based determinants of Fe–S proteins, the model highlights cysteine's direct coordination of Fe–S clusters and proline's essential structural support. Importantly, this interpretation does not assume a causal relationship between specific three-dimensional mechanisms and model predictions, but rather provides a structural context for understanding why these residues may be informative at the sequence level. These findings align with the established roles of these residues in fine-tuning the three-dimensional architecture of Fe–S proteins, providing new insights into their contributions to Fe–S protein identification and classification.

Importance of other amino acids in Fe–S proteins

While cysteine and proline are unequivocally central to Fe–S protein function and identification, our SHAP analysis also illuminated the significant contributions of several other amino acid residues, notably serine (S), glutamic acid (E), and arginine (R). These residues, although typically not direct cluster ligands, play crucial supporting roles that are vital for the structural integrity and catalytic efficiency of Fe–S proteins. The machine learning model's ability to detect these subtle yet critical sequence features highlights its effectiveness in capturing biologically meaningful patterns beyond direct coordination.

Serine (S) has emerged as an influential feature in Fe–S protein classification. While it does not directly bind the Fe–S cluster, serine's capacity for forming hydrogen bonds is particularly noteworthy. These hydrogen bonds can significantly influence the local environment around a cluster, impacting its stability, dynamics, and redox state [5]. In some instances, the oxygen of a nearby serine residue can even participate in iron ligation upon two-electron oxidation, providing a mechanism for coupling proton and electron transfer [42, 43].

The acidic residue, glutamic acid (E), was also a significant factor. Its importance aligns with the known role of charged residues in creating favorable electrostatic environments for metal-cluster binding and stability. Notably, glutamic acid's side chain can directly ligate various Fe–S clusters, including [2Fe–2S] and [4Fe–4S] types [44]. Its enriched content has also been observed in some Fe–S containing enzymes, such as *Escherichia coli* glutamate synthase [45].

Similarly, arginine (R) emerged as a noteworthy feature. Its guanidinium group mediates electrostatic interactions and hydrogen bonds, properties extensively studied in its interactions with phosphates [46, 47]. Crucially, arginine can directly ligate [2Fe–2S] clusters, as observed in biotin synthase [48], and its ligation to Fe–S clusters has been demonstrated in various cases [6, 49]. This direct structural contribution helps optimize Fe–S cluster positioning and electron transfer.

Collectively, these findings corroborate and extend existing biochemical and structural insights, suggesting a broader spectrum of amino acids involved in Fe–S protein characterization than traditionally emphasized. While cysteine remains the primary direct coordinator of Fe–S clusters, these auxiliary amino acids appear to play indispensable indirect roles in cluster stabilization, functional integrity, and overall protein architecture, roles that our sequence-based model successfully identifies as distinctive features. Overall, this work demonstrates the effectiveness of sequence-based approaches in uncovering biologically meaningful patterns.

Conclusion

This paper presents a robust and interpretable machine learning framework for the identification of Fe–S proteins, based entirely on sequence-derived features and supported by our newly developed database, FeSseqdb. In contrast to prior studies that depend heavily on structural templates or protein annotations, both of which are prone to error, our approach minimizes annotation errors by systematically verifying Fe–S clusters using atomic coordinate data from the PDB. To overcome annotation inconsistency in the PDB, FeSseqdb was constructed using an annotation-independent, structure-based identification strategy, rather than relying on conventional annotation-driven keyword searches. By doing so, all PDB entries could be systematically examined, ensuring

comprehensive coverage independent of existing annotations. Central to this framework is FeSseqdb, a curated chain-level database that emphasizes data granularity and reliability.

The classification model leverages biologically meaningful sequence features, including amino acid composition, cysteine spacing, and chain length. To ensure data integrity, the preprocessing pipeline includes Robust scaling and addresses the severe class imbalance using SVMSMOTE oversampling. When trained on these features, a Random Forest classifier achieves high performance (precision = 0.95 ± 0.01, F1-score = 0.90 ± 0.01), demonstrating that purely sequence-derived features can yield excellent predictive accuracy.

To improve transparency, explainable AI techniques were applied to elucidate feature importance, revealing cysteine and proline as the most influential predictors. These results align well with biochemical knowledge: cysteine is a direct ligand for Fe–S clusters via thiolate coordination, whereas proline plays an indirect but critical role in shaping the local structure of the binding site. Other residues can also contribute, sometimes even serving as direct ligands; in particular, serine, glutamic acid, and arginine play important roles in stabilizing the cluster environment and modulating electron transfer. This interpretable, sequence-focused strategy effectively captures both direct and indirect contributors to Fe–S cluster stability, complementing existing structural approaches. While the current model relies on sequence-derived features, future extensions could incorporate structure-aware descriptors to better represent Fe–S coordination. Recent advances in protein language models also offer promising opportunities for complementary evaluation and integration. Beyond binary classification, the framework establishes a foundation for more detailed tasks, such as distinguishing between Fe–S cluster types or analyzing other metalloprotein families. By integrating high-quality data with simple yet powerful learning models, this work provides a scalable, accurate, and interpretable solution for Fe–S protein discovery, advancing both fundamental protein research and bioinformatics applications.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06448-6.

Supplementary Material 1.

Acknowledgements

The authors acknowledge the National Institutes of Health high-performance computing cluster Biowulf (https://hpc.nih.gov) and LoBoS cluster (NIH/NHLBI) for providing the computational time and resources for this project.

Author contributions

J.M. led tool development, data analysis, and manuscript writing. B.R.B. and M.A. contributed to conceptualization, design, interpretation, and supervised the research. All authors reviewed and approved the manuscript.

Funding

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH). The contributions of the NIH authors were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered works of the United States Government. However, the findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

Data availability

The dataset **FeSseqdb** supporting the conclusions of this article is available in the Zenodo repository, https://doi.org/10.5281/zenodo.14911572. The dataset includes protein IDs, amino acid compositions (frequencies of the 20 standard residues), sequence length, and cysteine-related features (min_dist, avg_dist), which represent the minimum and average sequence distances between cysteine residues. It also contains full amino acid sequences per chain and binary labels (FeS), indicating the presence or absence of iron–sulfur clusters.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 September 2025 / Accepted: 10 April 2026
Published online: 23 June 2026

References

1. Liu J, Chakraborty S, Hosseinzadeh P, Yu Y, Tian S, Petrik I, et al. Metalloproteins containing cytochrome, iron-sulfur, or copper redox centers. Chem Rev. 2014;114(8):4366–469. https://doi.org/10.1021/cr400479b. (PubMed PMID: 24758379; PubMed Central PMCID: PMC4002152).
2. Beinert H, Holm RH, Münck E. Iron-sulfur clusters: nature's modular, multipurpose structures. Science. 1997;277(5326):653–9. https://doi.org/10.1126/science.277.5326.653. (PubMed PMID: 9235882).
3. Beinert H. Iron-sulfur proteins: ancient structures, still full of surprises. J Biol Inorg Chem. 2000;5(1):2–15. https://doi.org/10.1007/s007750050002. (PubMed PMID: 10766431).
4. Read AD, Bentley RET, Archer SL, Dunham-Snary KJ. Mitochondrial iron–sulfur clusters: structure, function, and an emerging role in vascular biology. Redox Biol. 2021;47:102164. https://doi.org/10.1016/j.redox.2021.102164.
5. Johnson DC, Dean DR, Smith AD, Johnson MK. Structure, function, and formation of biological iron-sulfur clusters. Annu Rev Biochem. 2005;74(1):247–81. https://doi.org/10.1146/annurev.biochem.74.082803.133518.
6. Meyer J. Iron-sulfur protein folds, iron-sulfur chemistry, and evolution. J Biol Inorg Chem. 2008;13(2):157–70. https://doi.org/10.1007/s00775-007-0318-7. (Epub 20071109. PubMed PMID: 17992543).
7. Sheftel A, Stehling O, Lill R. Iron-sulfur proteins in health and disease. Trends Endocrinol Metab. 2010;21(5):302–14. https://doi.org/10.1016/j.tem.2009.12.006. (Epub 20100108. PubMed PMID: 20060739).
8. Wachnowsky C, Fidai I, Cowan JA. Iron-sulfur cluster biosynthesis and trafficking - impact on human disease conditions. Metallomics. 2018;10(1):9–29. https://doi.org/10.1039/c7mt00180k. (PubMed PMID: 29019354; PubMed Central PMCID: PMC5783746).
9. Gervason S, Larkem D, Mansour AB, Botzanowski T, Müller CS, Pecqueur L, et al. Physiologically relevant reconstitution of iron-sulfur cluster biosynthesis uncovers persulfide-processing functions of ferredoxin-2 and frataxin. Nat Commun. 2019;10(1):3566. https://doi.org/10.1038/s41467-019-11470-9.
10. Liang LP, Patel M. Iron-sulfur enzyme mediated mitochondrial superoxide toxicity in experimental Parkinson's disease. J Neurochem. 2004;90(5):1076–84. https://doi.org/10.1111/j.1471-4159.2004.02567.x. (PubMed PMID: 15312163).
11. Isaya G. Mitochondrial iron-sulfur cluster dysfunction in neurodegenerative disease. Front Pharmacol. 2014;5:29. https://doi.org/10.3389/fphar.2014.00029. (Epub 20140303. PubMed PMID: 24624085; PubMed Central PMCID: PMC3939683).
12. Rouault TA, Tong WH. Iron-sulfur cluster biogenesis and human disease. Trends Genet. 2008;24(8):398–407. https://doi.org/10.1016/j.tig.2008.05.008. (Epub 20080705. PubMed PMID: 18606475; PubMed Central PMCID: PMC2574672.).
13. Tsibris JCM, Woody RW. Structural studies of iron-sulfur proteins. Coord Chem Rev. 1970;5(4):417–58. https://doi.org/10.1016/S0010-8545(00)80100-9.
14. Shulman RG, Eisenberger P, Blumberg WE, Stombaugh NA. Determination of the iron-sulfur distances in rubredoxin by X-ray absorption spectroscopy. Proc Natl Acad Sci USA. 1975;72(10):4003–7. https://doi.org/10.1073/pnas.72.10.4003. (PubMed PMID: 1060082; PubMed Central PMCID: PMC433126).
15. Machonkin TE, Westler WM, Markley JL. Paramagnetic NMR spectroscopy and density functional calculations in the analysis of the geometric and electronic structures of iron– sulfur proteins. Inorg Chem. 2005;44(4):779–97.
16. Shi W, Punta M, Bohon J, Sauder JM, D'Mello R, Sullivan M, et al. Characterization of metalloproteins by high-throughput X-ray absorption spectroscopy. Genome Res. 2011;21(6):898–907. https://doi.org/10.1101/gr.115097.110.
17. Zhang Y, Zheng J. Bioinformatics of metalloproteins and metalloproteomes. Molecules. 2020. https://doi.org/10.3390/molecules25153366.
18. Valasatava Y, Rosato A, Banci L, Andreini C. MetalPredator: a web server to predict iron-sulfur cluster binding proteomes. Bioinformatics. 2016;32(18):2850–2. https://doi.org/10.1093/bioinformatics/btw238. (Epub 20160606. PubMed PMID: 27273670).
19. Passerini A, Lippi M, Frasconi P. MetalDetector v2.0: predicting the geometry of metal binding sites from protein sequence. Nucleic Acids Res. 2011;39(Web Server issue):W288–92. https://doi.org/10.1093/nar/gkr365. (Epub 20110516. PubMed PMID: 21576237; PubMed Central PMCID: PMC3125771).
20. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2.
21. Golinelli-Pimpaneau B. Prediction of the Iron–Sulfur binding sites in proteins using the highly accurate three-dimensional models calculated by AlphaFold and RoseTTAFold. Inorganics. 2022;10(1):2. https://doi.org/10.3390/inorganics10010002.
22. Wehrspan ZJ, McDonnell RT, Elcock AH. Identification of Iron-Sulfur (Fe–S) cluster and Zinc (Zn) binding sites within proteomes predicted by DeepMind's AlphaFold2 program dramatically expands the metalloproteome. J Mol Biol. 2022;434(2):167377. https://doi.org/10.1016/j.jmb.2021.167377. (Epub 20211124).
23. Shi W, Chance MR. Metalloproteomics: forward and reverse approaches in metalloprotein structural and functional characterization. Curr Opin Chem Biol. 2011;15(1):144–8.
24. Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5(12):e1000605. https://doi.org/10.1371/journal.pcbi.1000605. (Epub 20091211).

25. Estellon J, Ollagnier de Choudens S, Smadja M, Fontecave M, Vandenbrouck Y. An integrative computational model for large-scale identification of metalloproteins in microbial genomes: a focus on iron–sulfur cluster proteins†. Metallomics. 2014;6(10):1913–30. https://doi.org/10.1039/c4mt00156g.
26. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493–500. https://doi.org/10.1038/s41586-024-07487-w. (Epub 20240508).
27. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st International conference on neural information processing systems. Long Beach, California, USA: Curran Associates Inc.; 2017. pp. 4768–77.
28. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. Entropy. 2021;23(1):18. https://doi.org/10.3390/e23010018.
29. Wilson CJ, Apiyo D, Wittung-Stafshede P. Role of cofactors in metalloprotein folding. Q Rev Biophys. 2004;37(3–4):285–314. https://doi.org/10.1017/S003358350500404X.
30. Wittung-Stafshede P. Role of cofactors in protein folding. Acc Chem Res. 2002;35(4):201–8. https://doi.org/10.1021/ar010106e.
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res. 2000;28(1):235–42.
32. Vallières C, Benoit O, Guittet O, Huang ME, Lepoivre M, Golinelli-Cohen MP, et al. Iron-sulfur protein odyssey: exploring their cluster functional versatility and challenging identification. Metallomics. 2024. https://doi.org/10.1093/mtomcs/mfae025. (PubMed PMID: 38744662; PubMed Central PMCID: PMC11138216).
33. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163. (Epub 20090320).
34. Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res. 2017;18(1):559–63.
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
36. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining. San Francisco, California, USA: Association for Computing Machinery; 2016. pp. 785–94.
37. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9. (Epub 20200117).
38. Gaillard J, Quinkal I, Moulis JM. Effect of replacing conserved proline residues on the EPR and NMR properties of *Clostridium pasteurianum* 2[4Fe–4S] ferredoxin. Biochemistry. 1993;32(38):9881–7. https://doi.org/10.1021/bi00089a002.
39. Sticht H, Rösch P. The structure of iron-sulfur proteins. Prog Biophys Mol Biol. 1998;70(2):95–136. https://doi.org/10.1016/s0079-6107(98)00027-3.
40. Lin J, Zhang L, Lai S, Ye K. Structure and molecular evolution of CDGSH iron-sulfur domains. PLoS ONE. 2011;6(9):e24790. https://doi.org/10.1371/journal.pone.0024790. (Epub 20110916).
41. Richardson JS, Richardson DC. Amino acid preferences for specific locations at the ends of alpha helices. Science. 1988;240(4859):1648–52. https://doi.org/10.1126/science.3381086.
42. Peters JW, Stowell MH, Soltis SM, Finnegan MG, Johnson MK, Rees DC. Redox-dependent structural changes in the nitrogenase P-cluster. Biochemistry. 1997;36(6):1181–7.
43. Lanzilotta WN, Christiansen J, Dean DR, Seefeldt LC. Evidence for coupled electron and proton transfer in the [8Fe–7S] cluster of nitrogenase. Biochemistry. 1998;37(32):11376–84. https://doi.org/10.1021/bi980048d.
44. Ballmann J, Albers A, Demeshko S, Dechert S, Bill E, Bothe E, et al. A synthetic analogue of Rieske-type [2Fe–2S] clusters. Angew Chem Int Ed Engl. 2008;47(49):9537–41. https://doi.org/10.1002/anie.200803418. (PubMed PMID: 18972470).
45. Miller RE. Glutamate synthase from *Escherichia coli*—an iron-sulfur flavoprotein: separation and analysis of non-identical subunits. Biochim Biophys Acta BBA Enzymol. 1974;364(2):243–9. https://doi.org/10.1016/0005-2744(74)90010-2.
46. Min J, Britt M, Brooks BR, Sukharev S, Klauda JB. Thermodynamics of arginine interactions with organic phosphates. Biophys J. 2025;124(13):2176–94. https://doi.org/10.1016/j.bpj.2025.05.019.
47. Woods AS, Ferré S. Amazing stability of the arginine-phosphate electrostatic interaction. J Proteome Res. 2005;4(4):1397–402. https://doi.org/10.1021/pr050077s. (PubMed PMID: 16083292; PubMed Central PMCID: PMC2945258).
48. Broach RB, Jarrett JT. Role of the [2Fe–2S]2+ cluster in Biotin Synthase: mutagenesis of the atypical metal ligand Arginine 260. Biochemistry. 2006;45(47):14166–74. https://doi.org/10.1021/bi061576p.
49. Van V, Brown JB, O'Shea CR, Rosenbach H, Mohamed I, Ejimogu N-E, et al. Iron-sulfur clusters are involved in post-translational arginylation. Nat Commun. 2023;14(1):458. https://doi.org/10.1038/s41467-023-36158-z.

Publisher's Note