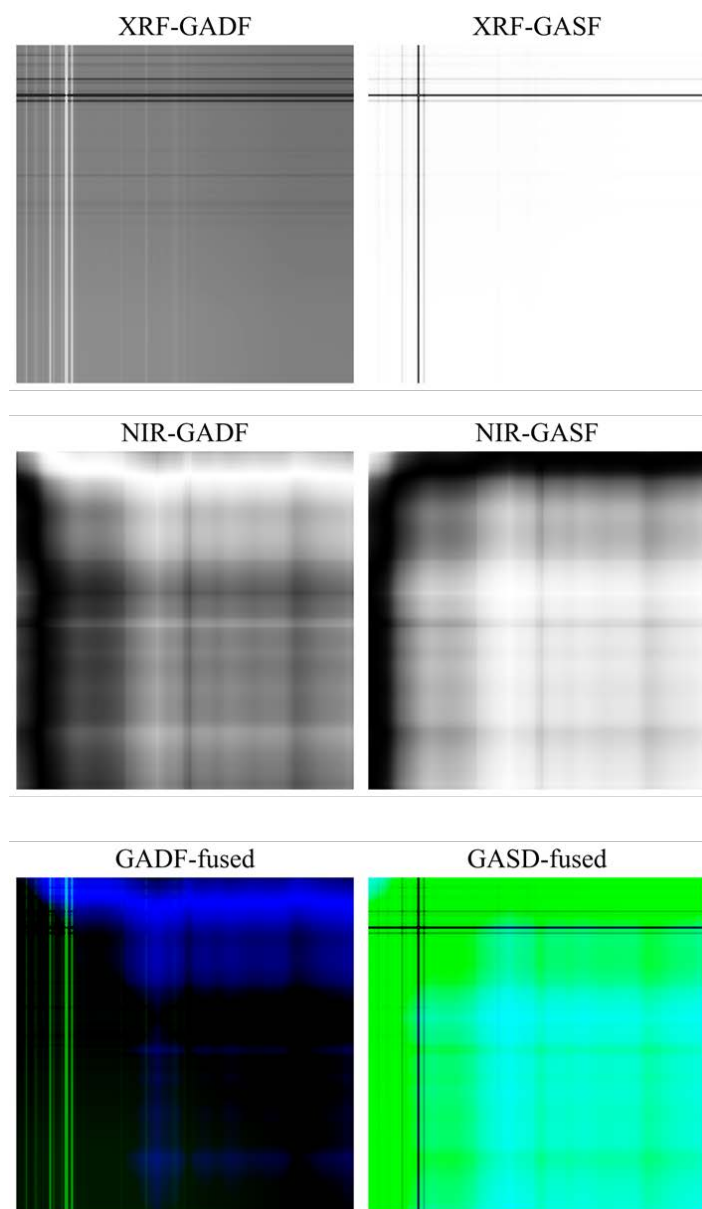
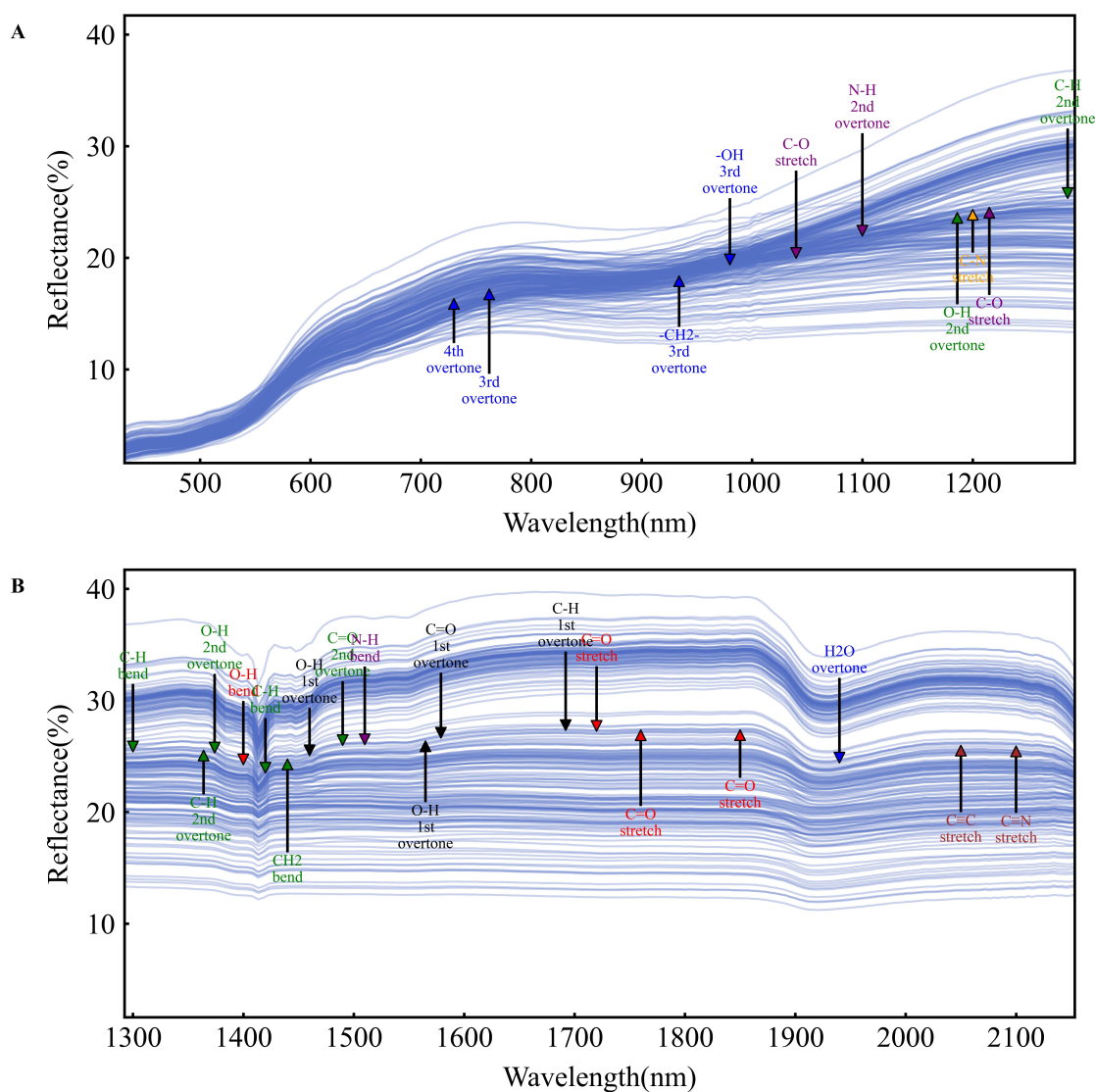


## A Support materials

### A.1 Figures



**Figure S1** The conversion process of GADF and GASF and the fusion process of NIR-XRF



**Figure S2** Characteristics of near-infrared spectral lines and chemical bonds

This figure displays the characteristic absorption bands of the near-infrared (NIR) spectra identified in the studied samples. Each annotated label corresponds to a specific vibrational mode or overtone of key chemical bonds, such as O–H, C–H, N–H, and C=O, as indicated at their respective wavelengths. The wavelength positions and bond assignments are based on established NIR spectral interpretation principles, allowing direct visualization of the functional groups and molecular structures present in the samples. These assignments provide an important chemical basis for subsequent spectral analyses and model interpretation.

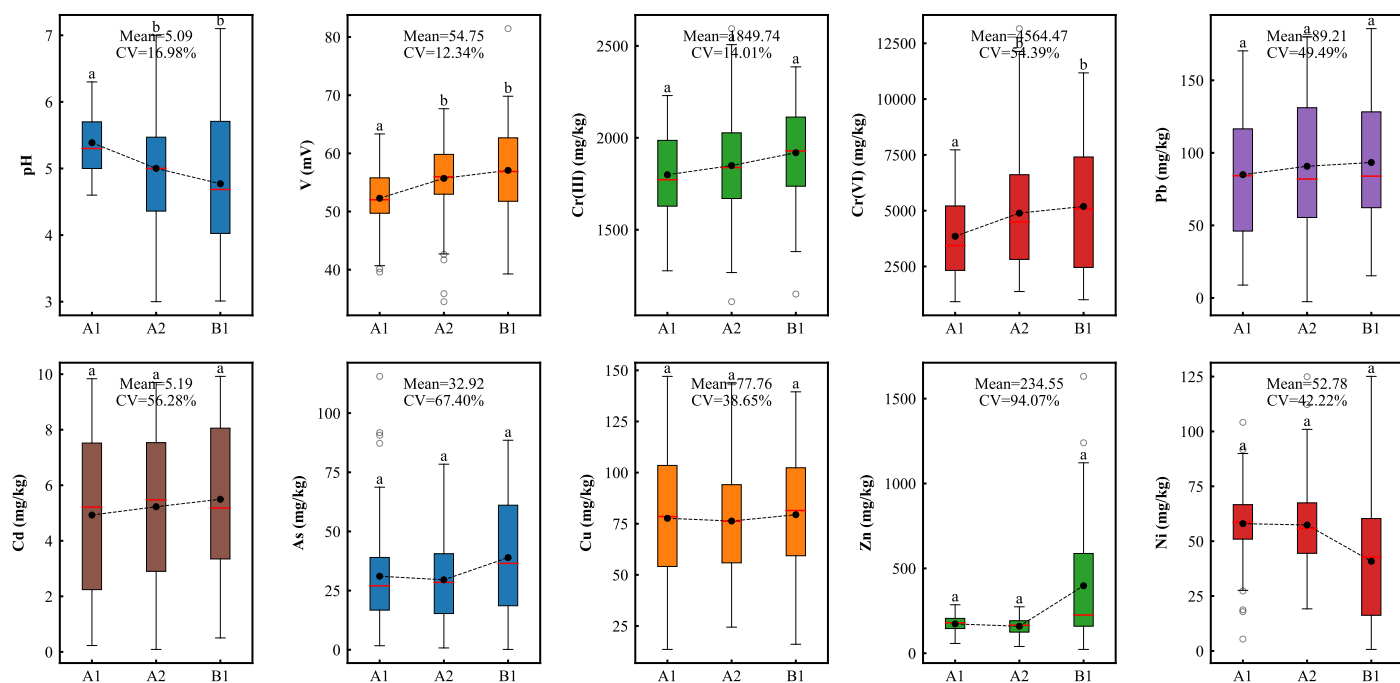
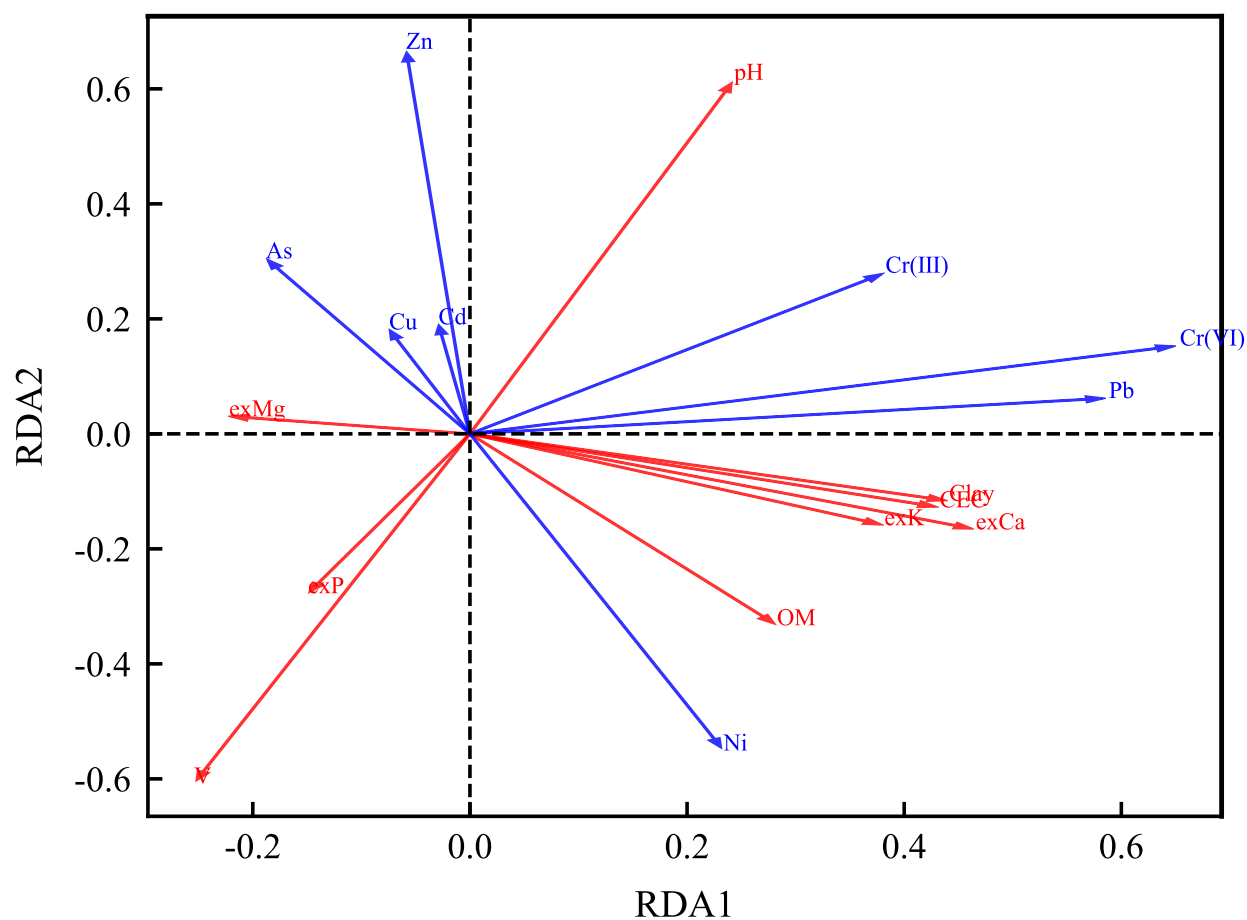


Figure S3 Distribution of sample-related parameters



**Figure S4** RDA analysis of sample content

The Redundancy Analysis (RDA) was performed to investigate the relationships between physicochemical parameters and heavy metal concentrations in the tailings samples. The RDA biplot (Figure S4) reveals several significant patterns and correlations:

**(1) Primary Correlations:**

- pH and Eh exhibit strong positive correlations with both Cr(III) and Cr(VI), as indicated by the acute angles between their respective vectors. This confirms the fundamental role of these parameters in controlling chromium speciation.
- The length of the Cr(VI) vector is notably longer than that of Cr(III), indicating greater variability in Cr(VI) concentrations across samples.

**(2) Physicochemical Parameters:**

- CEC (Cation Exchange Capacity) shows a moderate positive correlation with heavy metal retention, particularly for Cr(III), suggesting its importance in metal binding.
- Organic Matter (OM) displays an inverse relationship with Cr(VI) concentrations, consistent with its known role in reducing Cr(VI) to Cr(III).
- Clay content demonstrates positive correlations with both Cr species, indicating its significance in chromium retention.

**(3) Inter-metal Relationships:**

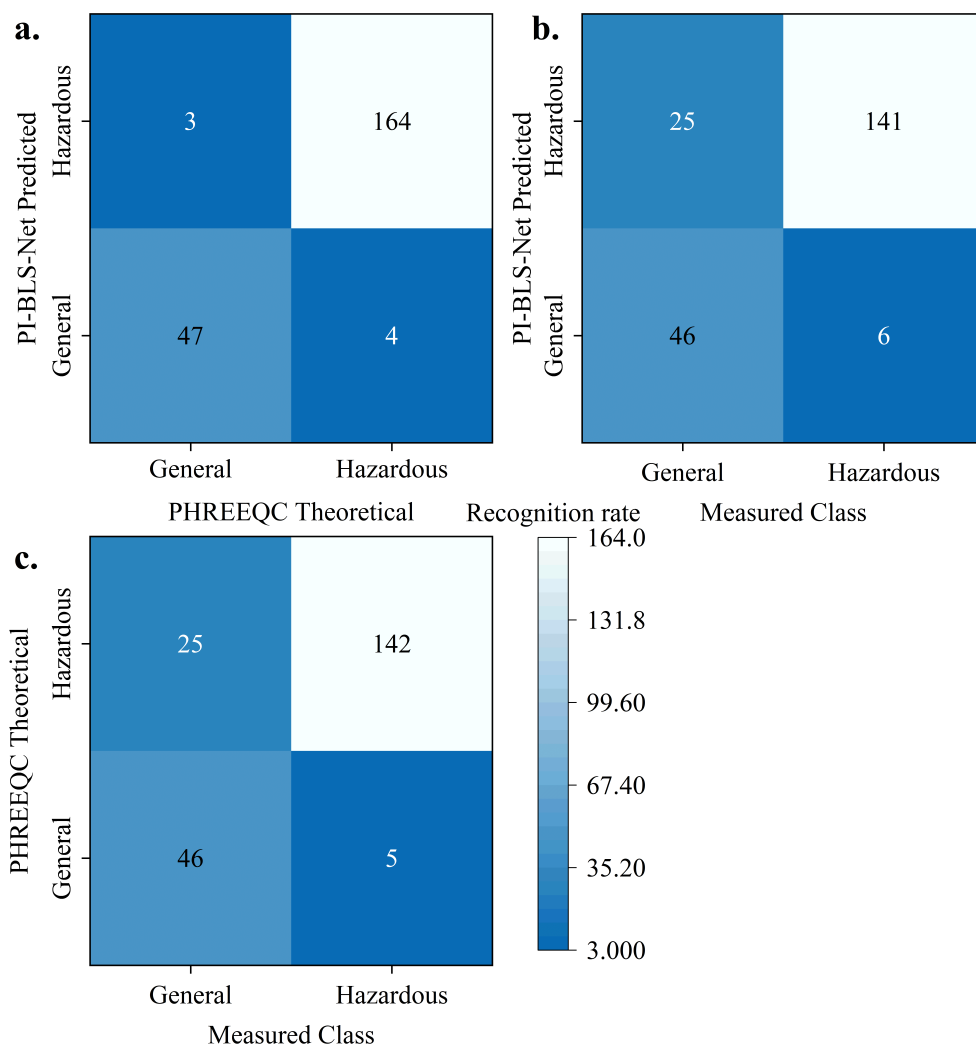
- Cr(VI) shows strong associations with As and Cu, suggesting similar geochemical behavior or common sources.
- Cr(III) clusters more closely with Ni and Zn, reflecting similar retention mechanisms in the soil matrix.
- Pb and Cd exhibit relatively independent behavior, as shown by their distinct vector directions.

**(4) Exchangeable Cations:**

a. Exchangeable Ca (exCa) and Mg (exMg) show moderate correlations with heavy metal distribution, particularly with Cr(III).

b. Exchangeable K (exK) and P (exP) display weaker correlations, suggesting less influence on chromium speciation.

The first two RDA axes explain 73.2% of the total variance (RDA1: 46.8%, RDA2: 26.4%), indicating that the analysis captures the majority of the variation in the dataset. These relationships underscore the complex interplay between soil physicochemical properties and chromium speciation, supporting the incorporation of multiple parameters in the PI-BLS-Net model for improved prediction accuracy.



**Figure S5** Comparative confusion matrices for tailings hazard classification.

**a.** Comparison between the PI-BLS-Net predicted classification and the PHREEQC theoretical classification to assess chemical consistency.

**b.** Comparison between the PHREEQC theoretical classification and the laboratory-measured actual classification (Measured Class) to establish a geochemical model baseline.

**c.** Comparison between the PI-BLS-Net predicted classification and the Measured Class to evaluate the model's practical predictive performance. In all matrices, the rows represent the "true" or "reference" class, and the columns represent the "predicted" class. The color intensity corresponds to the number of samples in each cell, as indicated by the color bar labeled "Recognition rate".

## A.2 PCANet Hyperparameter Optimization Details

This appendix provides the detailed procedure and results for optimizing the hyperparameters of the PCANet feature extraction module using Bayesian optimization. The objective was to find the optimal set of PCANet hyperparameters that maximizes the recognition rate of the overall pipeline (PCANet feature extraction followed by the BLS classifier) on a dedicated validation set, while also considering the computational efficiency.

### A.2.1 Hyperparameters Tuned and Search Space

The performance of the PCANet model is influenced by several hyperparameters that govern its architecture and feature processing. The following key hyperparameters were selected for optimization:

(1) **Patch Size ( $k$ ):** The dimension of the square image patches ( $k \times k$ ) extracted in each convolutional stage. This parameter influences the scale of local features captured by the initial filters. Search space: Candidate values from  $\{3, 5, 7\}$ .

(2) **Number of Filters per Stage ( $\{L_s\}_{s=1}^S$ ):** The number of principal components (filters) learned and applied in each of the  $S$  stages. This dictates the dimensionality of the feature maps at each level. Assuming a 2-stage PCANet ( $S = 2$ ), we optimized the number of filters for the first stage ( $L_1$ ) and the second stage ( $L_2$ ) independently. Search space: Integer values ranging from 4 to 16 for both  $L_1$  and  $L_2$ .

(3) **Histogram Block Size ( $H$ ):** The dimension of the square blocks ( $H \times H$ ) used for histogram computation in the final pooling layer. Smaller blocks capture more localized feature patterns. Search space: Candidate values from  $\{8, 16\}$ .

(4) **Block Overlap Ratio:** The ratio determining the overlap between adjacent histogram blocks during pooling. Overlapping blocks can help capture spatial context more robustly. Search space: Candidate values from  $\{0, 0.25, 0.5\}$ .

A 2-stage PCANet architecture was fixed ( $S = 2$ ) for this optimization process.

### A.2.2 Optimization Methodology

Bayesian optimization was utilized to efficiently explore the defined hyperparameter space. This method builds a probabilistic model of the objective function and uses it to select the most promising hyperparameters to evaluate in each iteration. The optimization process was conducted on a dedicated validation set, strictly separated from the training and final test sets.

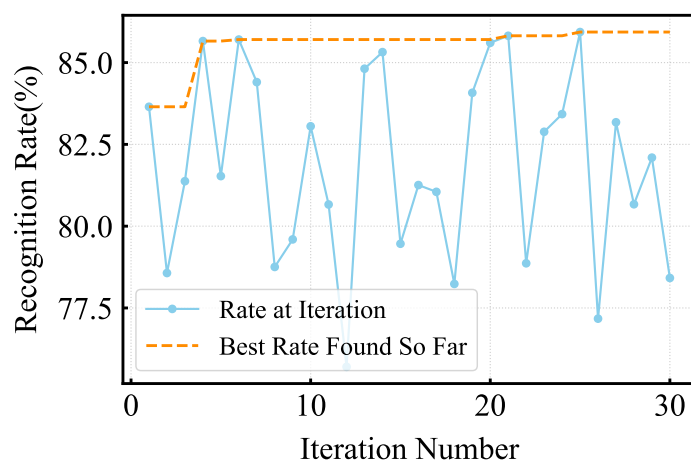
The objective function to be maximized was the recognition rate (classification accuracy) of the complete PCANet+BLS pipeline on the validation set. For each evaluation of the objective function, the following steps were performed for a given set of PCANet hyperparameters:

- (1) The PCANet filters were learned using the specified hyperparameters on the training set images.
- (2) PCANet features were extracted from both the training and validation set images using the learned filters.
- (3) The extracted PCANet features were concatenated with the corresponding physicochemical parameters (pH, Eh).
- (4) A BLS classification model was trained using the combined features from the training set.
- (5) The trained BLS model was evaluated on the combined features from the validation set, and the recognition rate was computed.

The Bayesian optimization was performed for a total of 30 iterations, aiming to converge towards the hyperparameter set yielding the highest validation recognition rate.

### A.2.3 Optimization Results and Visualization

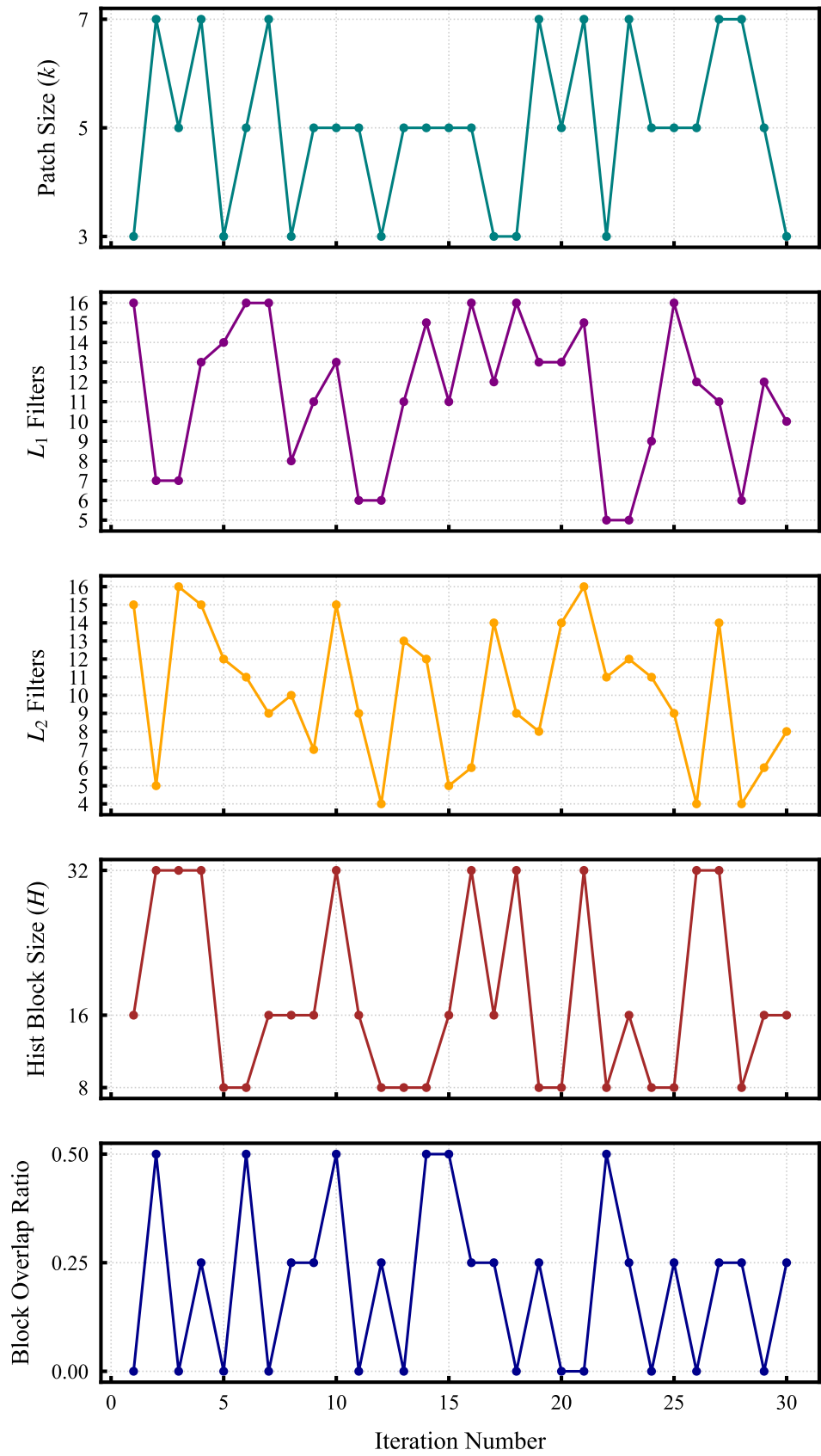
The Bayesian optimization process successfully explored the defined hyperparameter space over 30 iterations. The performance of the PCANet + BLS pipeline for each evaluated hyperparameter combination.



**Figure S6** Validation recognition rate achieved at each iteration of Bayesian optimization, showing the best rate found so far.

Figure S6 illustrates the validation recognition rate achieved at each iteration of the Bayesian optimization process, along with the best recognition rate found up to each iteration. This plot demonstrates the optimization algorithm's progress in finding better hyperparameter configurations over time, indicating the convergence of the optimization search.





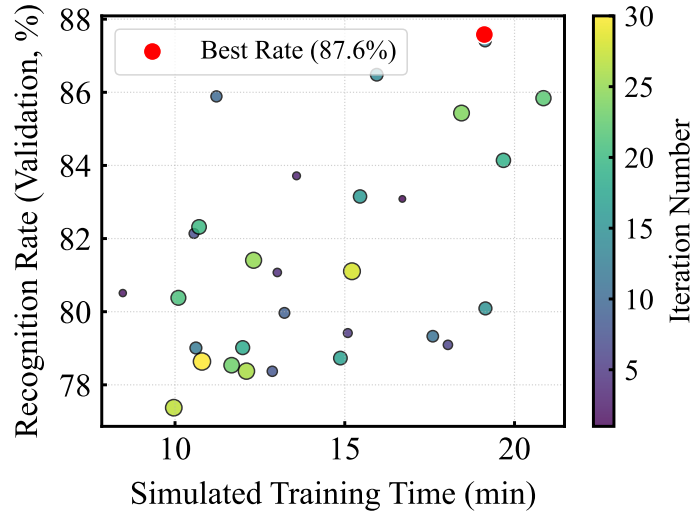
**Figure S7** Values of each hyperparameter explored at each iteration of Bayesian optimization. Each panel corresponds to a specific hyperparameter.

Figure S7 provides a detailed view of how the optimizer explored the values of each tuned hyperparameter across the iterations. Each panel corresponds to a specific hyperparameter, showing its selected value at each step. These plots reveal the search strategy of the Bayesian optimization, illustrating how it samples different combinations in the parameter space and highlighting the range of values explored for each parameter.

The Bayesian optimization process identified the following hyperparameter combination as achieving the highest recognition rate on the validation set:

- a. Patch Size ( $k$ ): 7**
- b. Number of Filters per Stage ( $L_1, L_2$ ):  $L_1 = 16, L_2 = 16$**
- c. Histogram Block Size ( $H$ ): 16**
- d. Block Overlap Ratio: 0.25**

This optimal parameter set achieved a maximum recognition rate of **89.08%** on the validation set. The simulated training time for this configuration was approximately **20.34 minutes**.



**Figure S8** Trade-off between validation recognition rate and simulated training time for each evaluated hyperparameter set. The color of each point indicates the iteration number. The point with the highest overall recognition rate is highlighted.

Figure S8 visualizes the trade-off between validation recognition rate and simulated training time for each evaluated hyperparameter set. Each point represents a single optimization iteration, with the color indicating the iteration number. This plot is crucial for understanding the computational cost associated with achieving different performance levels. It allows for identifying configurations that offer a good balance between recognition rate and training time. The point corresponding to the optimal configuration (highest recognition rate) is specifically highlighted in this plot.

The best hyperparameter set identified through this optimization process was subsequently used to train the final PCANet + BLS model on the combined training and validation datasets, and its performance was rigorously evaluated on the independent test set.

### A.3 Algorithm

---

**Algorithm S1** PCANet Feature Extraction

---

**Input:** Images  $\{\mathcal{I}_i\}$ , Training Set  $\{\mathcal{I}_{\text{train}}\}$ , Params  $S, k, \{L_s\}$ , Block Size  $H$

**1. Learn PCA Filters:**

**for**  $s = 1$  to  $S$  **do**

    Extract patches, compute  $L_s$  principal components as filters  $W_s$

    Apply  $W_s$  to generate next-stage feature maps

**end for**

**2. Extract Features:**

**for** each input image **do**

    Apply  $\{W_s\}$  to obtain final feature maps

    Binarize maps, divide into  $H \times H$  blocks

    Compute histograms, concatenate to form feature vector

**end for**

**Output:** Feature vectors  $\{\mathcal{F}_i\}$

---

---

**Algorithm S2** PI-BLS-Net for Cr Speciation

---

**Input:**  $X_{\text{spec}}, X_{\text{phys}}, Y_{\text{target}}, Cr_{\text{total}}$ , Params

**1. Feature Construction:**

Generate feature nodes from  $X_{\text{spec}}$  and  $X_{\text{phys}}$

Generate enhancement nodes with modulated activation

Concatenate into final input  $A$

**2. Training Loop:**

**for** each iteration **do**

    Predict  $Y_{\text{pred}} = A \cdot W_{\text{output}}$

    Compute losses:  $L_{\text{MSE}}, L_{\text{Redox}}, L_{\text{Mass}}$

    Total loss = weighted sum of above

    Update  $W_{\text{output}}$  and loss weights

**end for**

**Output:** Trained  $W_{\text{output}}$

---