Pak. J. Bot., 58(6): 1352-1363, June 2026.

ARTICLE DOI: http://dx.doi.org/10.30848/PJB2026-6(5)

LOW-ABUNDANCE ION-ENHANCED MS ENTROPY SIMILARITY MODEL BASED ON MLP: ADVANCING BIOLOGICAL MASS SPECTROMETRY ANALYSIS

JIAYAO PAN[1], RUIYANG LI[2] AND YONG LI[1*]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China
[2]Faculty of Humanities, Arts & Social Sciences, the University of Queensland, Brisbane, 4072, Australia

*Corresponding author's email: leon@kust.edu.cn

Abstract

Traditional similarity methods in small-molecule mass spectrometry are severely hindered by a three-order-of-magnitude signal disparity between high-abundance backbone ions (relative intensity >10%) and low-abundance characteristic ions (relative intensity <1%). To address this limitation, a low-abundance ion–enhanced mass spectrometry entropy (MSE) similarity calculation model based on a multi-layer perceptron (MLP) is proposed. The approach involves four-layer db4 wavelet decomposition, soft-threshold denoising, intensity normalization, and calculation of *MSE* and statistical features. An *MSE*-constrained nonlinear function and dual-channel *MLP* establish spectral peak intensity-dynamic parameter mapping, with backpropagation optimizing parameters to enhance low-abundance ion contribution and suppress high-abundance interference. Validation using the *MassBank.us* and *KUST-MS* datasets demonstrate statistically significant performance improvements, with 81.18% (KUST-MS) and 77.27% (MassBank.us) of sample groups achieving t-values greater than 2, and over 50% exhibiting p-values below 0.05.The overall *Cohen's d* was 0.879, with 88.0% large effect sizes (*Cohen's d* of 0.8 or higher), and 29.4% extremely large effect (*Cohen's d* of 1.5 or higher), confirming dynamic weighting significantly enhances the capability to discriminate structural differences in low-spectral-entropy scenarios. The findings of this research are expected to significantly enhance the accuracy of complex mass spectrometry data analysis, thereby providing efficient technical solutions for applications such as metabolomics biomarker screening, environmental trace pollutant analysis, which is important not only for human health but also for biodiversity conservation and drug impurity identification.

Key words: *MSE*; Low-abundance ion recognition; Nonlinear function; *MLP* model

Introduction

Renowned for its precise analysis of substance composition and structural characteristics, mass spectrometry technology has become an indispensable tool in modern analytical science (Zhang *et al.*, 2023). It plays a crucial role across diverse fields, including metabolomics, biomedical research, and environmental monitoring (Jones, 2020; Fang *et al.*, 2024; Thomas *et al.*, 2022; Xu *et al.*, 2023; Wenk *et al.*, 2024). As a fundamental data carrier, mass spectrometry data encompasses ionic strength information that reflects the molecular structural attributes of substances. However, the concentrated distribution of high-abundance backbone ions often obscures the subtle structural differences conveyed by low-abundance ions in traditional methods for similarity retrieval. In recent years, similarity calculation methods based on *MSE* have introduced novel approaches to address these challenges. *MSE*, defined as a measure quantifying the uniformity of ion intensity distribution, has been demonstrated to effectively characterize the complexity of mass spectra. In low-entropy spectra, where high-abundance ions dominate, low-abundance ions possess particularly high structural indication value. Although existing studies have employed *MSE* to model ion distribution characteristics for enhancing low-abundance ion contributions, critical challenges persist (Li *et al.*, 2021; Li & Fiehn, 2023). The fixed characterization of ion distribution uniformity via *MSE* limits its ability to dynamically adapt to the nonlinear variations in ion intensity across different entropy intervals. As a result, low-abundance ion information is often underestimated in similarity computations, and an effective collaborative optimization mechanism for high- and low-abundance signals has yet to be established.

To overcome these challenges, this study proposes a low-abundance ion-enhanced *MSE* similarity model based on a dual-channel *MLP*. The proposed method aims to address the issue of diminished low-abundance feature representation in the identification of small molecule compounds. Through a series of data preprocessing steps, multi-dimensional features are systematically constructed by integrating *MSE* and statistical properties of ion intensity, thereby providing a robust foundation for subsequent feature-to-contribution mapping. A nonlinear contribution function is then formulated with *MSE* serving as the core constraint. By leveraging the dual-channel *MLP* architecture, an end-to-end mapping between ion characteristics and weight coefficients is established, enabling differentiated signal enhancement strategies across various entropy intervals. This approach effectively amplifies the contributions of low-abundance ions during similarity calculations. The findings of this research are expected to significantly enhance the accuracy of complex mass spectrometry data analysis, thereby providing efficient technical solutions for applications such as metabolomics biomarker screening, environmental trace pollutant analysis, which is important not only for human health but also for biodiversity conservation and drug impurity identification.

Material and Methods

General process of the model: A low-abundance ion-enhanced *MSE* similarity calculation model based on *MLP* was developed in this study. As illustrated in Fig. 1, the proposed model consists of four main modules: data preprocessing, extraction of ion distribution features, dynamic weighted decision-making, and similarity calculation.

Received: 30-09-2025 Revised: 14-01-2026 Accepted: 20-01-2026 Online: 03-02-2026

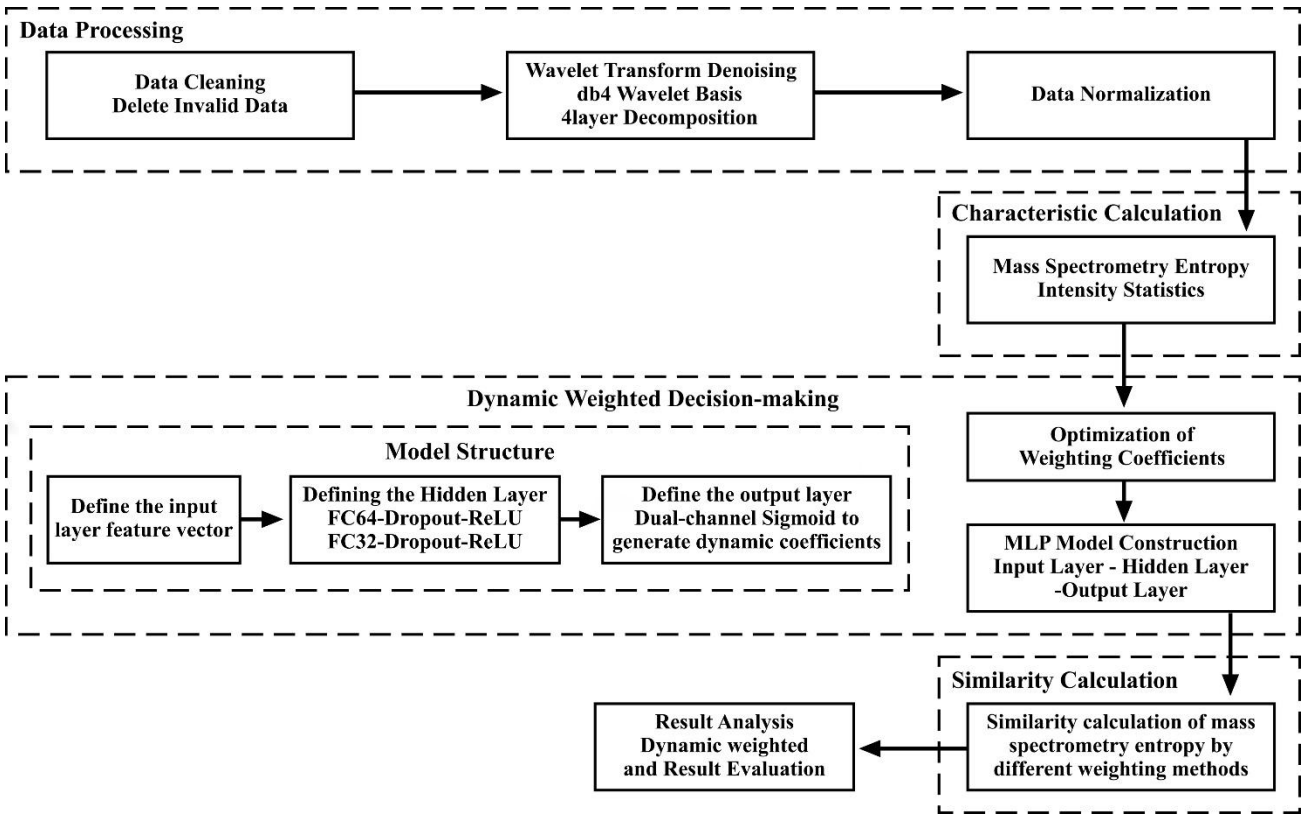

Fig. 1. Flow chart of the low-abundance ion-enhanced mass spectra similarity calculation model.

In the data preprocessing stage, invalid data were first removed through data cleaning. Subsequently, wavelet transform denoising using a *db4* wavelet basis and four-layer decomposition was applied to suppress high-frequency noise while preserving the core ion peak features. Instrument response variations were normalized to ensure comparability across spectra. Finally, statistical features including *MSE*, mean ion intensity, standard deviation, and skewness, were extracted to quantify information value of low-abundance ion. Based on these preprocessed features, a nonlinear function constrained by *MSE* was constructed, followed by the design of a dual-channel *MLP* network for adaptive ion contribution parameters. The input layer of the network fused the extracted statistical features. The hidden layer consisted of two fully connected layers with dimensions of 64 and 32 dimensions, respectively. Nonlinearity was introduced by the ReLU activation function, and overfitting was mitigated through the combination of *Dropout* and *B*atch *N*ormalization (*BN*). The output layer generated dynamic parameters through independent dual-channel Sigmoid functions, enabling end-to-end regulation of weight distribution. The *Adam* optimizer minimized a loss function combining mean square error and *L2* regularization, promoting the adaptive learning of ion feature importance. Finally, *MSE* similarity calculations were performed to assess the dynamic weighting results, confirming that this approach significantly enhances mass spectrometry matching accuracy and robustness while maintaining data feature integrity.

Data Preprocessing

Data cleaning, wavelet transform denoising, and normalization: Invalid records with missing key information, such as mass-to-charge ratio (m/z) or ion intensity, were first removed during data cleaning. To suppress high-frequency noise while preserving characteristic fragment ion peaks, a discrete wavelet transform (DWT) de-noising strategy was employed using a db4 wavelet basis with four-layer decomposition. The db4 wavelet was selected due to its favorable time–frequency localization capability and smoothness, which provides a balanced performance between noise suppression and peak preservation for mass spectrometry signals. Compared with shorter wavelets (e.g., sym4) that may insufficiently suppress baseline noise, and longer wavelets (e.g., coif4) that may oversmooth weak ion peaks, db4 has been widely adopted in spectral denoising tasks for preserving low-abundance ion characteristics. A four-layer decomposition was chosen to ensure effective noise removal without compromising ion peak integrity; shallower decompositions (e.g., three layers) were found to retain residual noise, while deeper decompositions (e.g., five layers) may attenuate weak characteristic ions. High-frequency coefficients were processed using a soft-threshold contraction algorithm to reduce random noise and baseline fluctuations. Subsequently, ion intensities were normalized to eliminate instrument response variations and ensure comparability across datasets, according to:

$$\hat{I}_i = \frac{I_i}{\sum_{i=1}^{n} I_i} \quad (1)$$

where I_i represents the original intensity of the i^{th} ion, and $\hat{I}_i$ represents the normalized intensity.

Feature calculation: Feature extraction was performed to characterize ion distribution properties and quantify the informational value of low-abundance ions.

Mass spectral entropy calculation: For each mass spectrometry dataset containing n ion peaks, where the relative intensity of each ion peak can be regarded as a random variable I_i. The formula for calculating H is:

$$H = -\sum_{i=1}^{n} \hat{I}_i * \log_2(\hat{I}_i) \quad (2)$$

This metric quantifies the uniformity of ion intensity distributions. Low-entropy spectra (H<3) are typically dominated by a small number of high-abundance backbone ions, often masking subtle difference signals from low-abundance ions. In contrast, high-entropy spectra (H≥3), exhibit relatively more uniform ion intensity distributions with low information redundancy,

Statistical characteristics of ion intensities: The mean, standard deviation, and skewness of ion intensities were calculated to describe the central tendency, dispersion, and asymmetry of ion distribution. These statistical descriptors complement MSE by capturing nonlinear variations in ion intensity patterns.

Mean: Reflects the average ion intensity and overall signal strength trend.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \hat{I}_i \quad (3)$$

Standard deviation: Measures intensity dispersion, higher values indicate more pronounced differences between high- and low-abundance ion.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{I}_i - \mu\right)^2} \quad (4)$$

Skewness: Describes distribution symmetry: negative skewness indicates a higher proportion of low-abundance ions, while positive skewness reflects dominance of high-abundance ions.

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{I}_i - \mu\right)^3}{\sigma^3} \quad (5)$$

Dynamic weighting strategy design: The proposed dynamic contribution adjustment strategy aims to enhance the representation of low-abundance ions under low-entropy spectral conditions by adaptively regulating ion intensity weights through a dual-channel MLP network.

The detailed methodology is described as follows:

(1) Definition of *MSE*-driven nonlinear function: To address the issue where high-abundance skeleton ions obscure structural difference signals in low-entropy spectral (H<3) scenarios, leading to insufficient discriminative power of low-abundance ions, a nonlinear function is defined. A dual-channel *MLP* network is constructed based on *MSE* and statistical features of ion intensity distribution to enable dynamic adjustment of contribution parameters, with exponential weighting applied to ion intensity. The function is expressed as:

$$I' = \begin{cases} I & (H \geq 3) \\ I^w, w = \alpha_1 + \alpha_2 H & (H < 3) \end{cases} \quad (6)$$

where H represents the *MSE*, I represents the original ion intensity, I' signifies the dynamically weighted ion intensity, α_1 and α_2 are dynamic parameters output by the *MLP*. Physically, α1 primarily controls the baseline enhancement of low-abundance ions based on statistical distribution features, while α2 modulates the sensitivity of intensity adjustment according to spectral entropy H, enabling entropy-dependent adaptive weighting. The initial values of α1 and α2 were set to **0.25**, following the fixed-weight parameter benchmark commonly used in traditional MSE-based similarity models (Li *et al*., 2021). This initialization provides a neutral prior that ensures stable network convergence while allowing sufficient flexibility for adaptive optimization during training.

When H<3, the mass spectrometry is categorized into a low-entropy scenario, which is indicative of a high potential for information mining due to the probable presence of abundant low-abundance ions. Under such conditions, the dynamic contribution adjustment process is initiated. This design employs a dual-channel *MLP* network to optimize weight parameters, thereby enhancing the contribution of low-abundance ions while maintaining computational efficiency (Hart *et al.*, 2024).

(2) *MLP* network construction: A dual channel *MLP* neural network comprising an input layer, two hidden layers, and a dual-channel output layer was implemented using the PyTorch framework. The core architecture is shown in Fig. 2.

Input layer: The input feature vector was a 6-dimensional dataset: $x = [H,\mu,\sigma,S,\alpha_1,\alpha_2]^T$.

where H is the mass spectral entropy, μ, σ, and S represent the mean, standard deviation, and skewness of ion intensities, respectively. Here α1, α2 are the initial contribution coefficients.

***MS* Entropy:** Quantifies the uniformity of ion intensity distribution and reflects the potential information value of low-abundance ions.

Statistical characteristics of ion intensity distribution (mean, standard deviation and skewness): Characterize the central tendency, dispersion, and symmetry of ion intensities, respectively.

Initial weight coefficient α_1 and α_2: Provide a priori parameter baselines to facilitate network convergence.

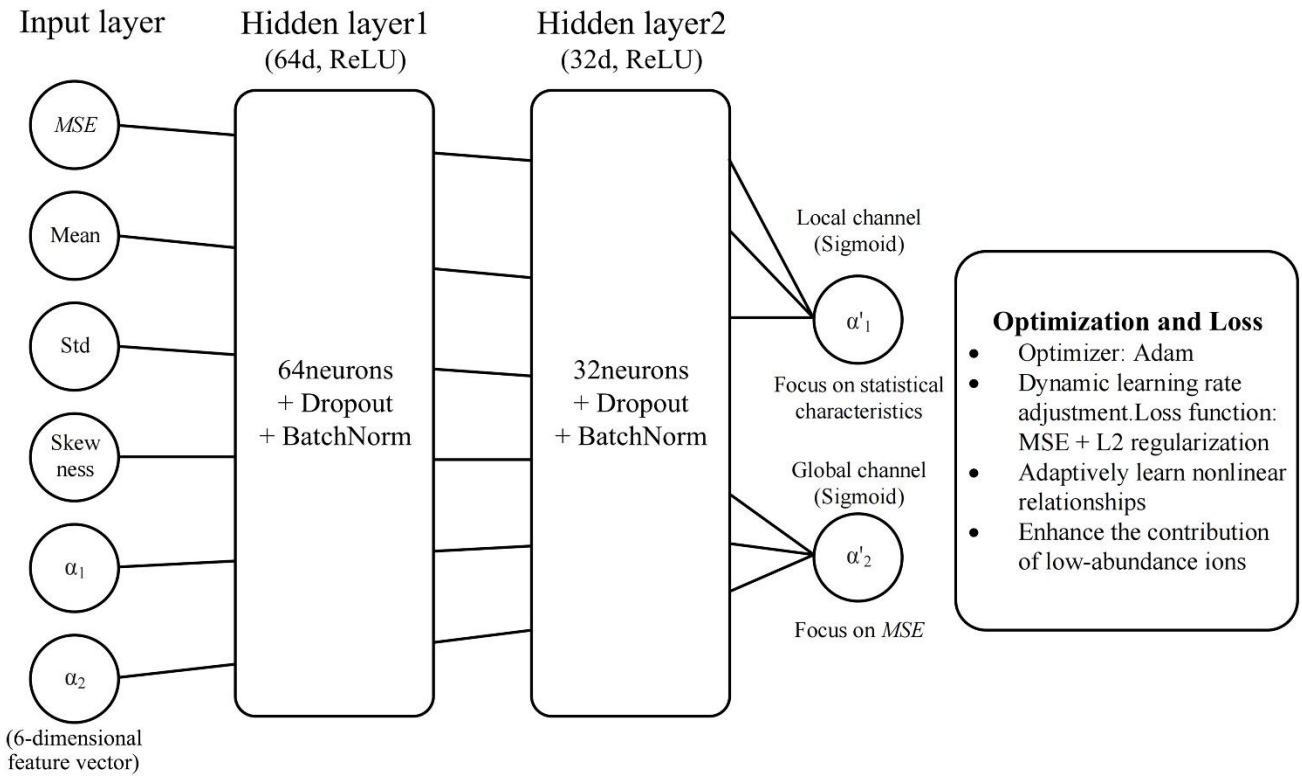

Fig. 2. Architecture diagram of dual-channel MLP neural network model based on PyTorch.

Hidden layer: Two fully connected layers with 64 and 32 nodes, respectively, are included to extract high-order features. Nonlinearity is introduced by the ReLU activation function, and overfitting is suppressed through the combination of *Dropout* and *BN*. The hidden layer computations are:

$$z_1 = W_1 x + b_1 \ , \ h_1 = \mathrm{Re}\,LU\left(BatchNorm\left(z_1\right)\right) \quad (7)$$

$$z_2 = W_2 h_1 + b_2, h_2 = \mathrm{Re}\,LU\left(BatchNorm\left(z_2\right)\right) \quad (8)$$

where $W_1 \in R^{64\times 6}$ represents the connection weight from the input layer to the first hidden layer, which is responsible for mapping the original features to high-dimensional local features, $b_1 \in R^{64}$ corresponds to the bias vector of the first hidden layer, which is used to adjust the activation threshold of the first hidden layer and enhance the model's fitting ability for nonlinear features. $W_2 \in R^{32\times 64}$ represent the weight matrices, $b_2 \in R^{32}$ corresponds to the bias vector of the second hidden layer, which is used to correct the information loss during the feature dimensionality reduction process.

Output layer: A dual-channel architecture was designed to independently output dynamic parameters α_1' and α_2'. Feature-wise differentiation was achieved by applying the sigmoid function to constrain parameters within [0,1]:

The local channel output α_1' prioritizes statistical features (H,μ,σ), the weight matrix $W_{\alpha 1} \in R^{1\times 32}$ assigns higher weights to hidden-layer outputs related to ion intensity distribution characteristics.

The global channel output α_2' focuses on *MSE*, the weight matrix $W_{\alpha 2} \in R^{1\times 32}$ reinforces the mapping of H -related features in the hidden-layer outputs.

The basis for setting the initial value to 0.25: Referring to the parameter setting benchmark of (Li *et al*., 2021). in the fixed-weight MSE model, combined with the pre-experiment results of this study.

The parameter update equations are shown in Equation (9) and Equation (10).

$$\alpha_1' = f\left(W_{\alpha_1} \cdot h_2 + b_{\alpha_1}\right) \quad (9)$$

$$\alpha_2' = f\left(W_{\alpha_2} \cdot h_2 + b_{\alpha_2}\right) \quad (10)$$

where f represents the sigmoid activation function. The weight matrix $W_{\alpha 1}$ prioritizes statistical feature-related latent representations, enhancing low-abundance ion sensitivity, while $W_{\alpha 2}$ emphasizes entropy-related features to regulate global intensity distribution adaptively.

Loss function: The loss function adopted the mean square error as the primary metric to measure the difference between the predicted and true similarity scores, with an *L2* regularization term (with intensity (λ) added:

$$Loss = \frac{1}{n}\sum_{i=1}^{n}(SI_i - SI_i')^2 + \lambda\left(\left\|W_1\right\|_2^2 + \left\|W_2\right\|_2^2\right) \quad (11)$$

Here, SI_i represents the dynamically weighted similarity prediction SI_i' represents the true similarity score, λ represents the regularization parameter, and n represents the sample size.

(3) Weight coefficient optimization: Weight coefficient optimization was implemented via the *Adam* optimizer with dynamic learning rate adjustment. During backpropagation, gradients of the weight coefficients with respect to the loss function were computed to iteratively update the *MLP* network's weights and bias parameters independently. The loss function, combining mean square error and *L2* regularization, quantified differences between predicted and true similarity scores as described in Equation (11). This optimization process facilitated adaptive learning of nonlinear relationships between *MSE* and ion intensity distribution features, enhancing the contribution of low-abundance ions in similarity assessments.

Calculation of entropy similarity in mass spectrometry: The calculation of entropy similarity in mass spectrometry is based on the information entropy theory (Tian *et al.*, 2023). Which evaluates the similarity of chemical structures by quantifying the divergence in information distribution between two mass spectra. When combined with the dynamic contribution adjustment strategy described in this study, the enhanced similarity calculation for two mass spectra *X* and *Y* is given by Equation (12):

$$Similarity(X,Y)=\frac{1}{2}\sum_{i,j}\begin{cases}0 & m/z_{X,i}\neq m/z_{Y,j}\\ f(I'_{X,i}+I'_{Y,j})-f(I'_{X,i})-f(I'_{Y,i}) & m/z_{X,i}=m/z_{Y,j}\end{cases} \quad (12)$$

where $f(x)=x\log_2 x$, I' represents the intensity after dynamic contribution adjustment and $\sum I'_{X,i}=1$, $\sum I'_{Y,i}=1$. When mass-to-charge ratios *m/z* between two spectra do not match, their contribution to the score is zero. For matching *m/z* values, calculating the contribution of this ion to the entropy similarity of mass spectrometry. A similarity score closer to 1 indicates a higher degree of chemical structure similarity between the two mass spectrometry datasets, while a value approaching 0 signifies greater dissimilarity. The similarity calculation process is shown in the Fig. 3.

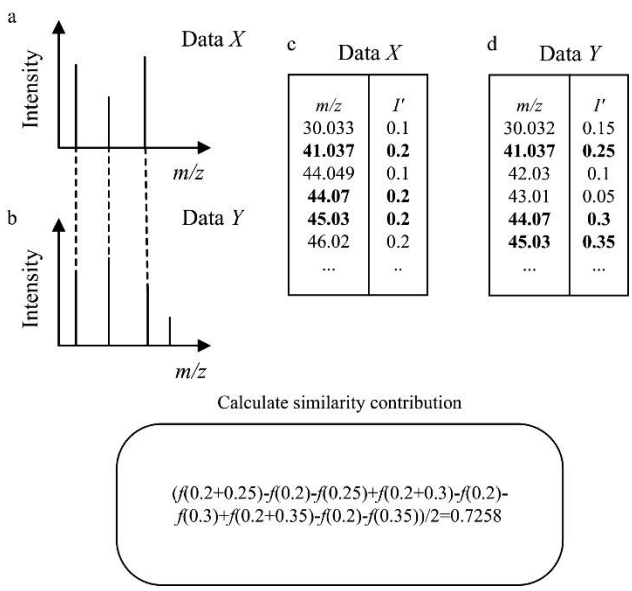

Fig. 3. Workflow for mass spectrometry similarity calculation incorporating dynamic weighting.

Figure 3a and Figure 3b present the mass spectra of mass spectrometry data *X* and *Y*, where the solid lines represent the intensity values, and the dashed lines indicate the correspondences of the matching ions. Figure 3c and Figure 3d present the mass spectrometry data after dynamic weight adjustment, where *I'* denotes the enhanced ionic intensity and the bolded segments highlighting the matching ions. The procedure is as follows: Firstly, mass spectrometry data *X* and *Y* are preprocessed (including denoising and normalization), and their mass spectral entropies are calculated. Since the mass spectral entropies of both *X* and *Y* are less than 3, a dynamic weighting mechanism is automatically triggered to enhance the weights of low-abundance ions via a nonlinear function. Finally, the entropy similarity between the two datasets is calculated using Equation (12).

Results

Dataset and evaluation metrics: All experiments in this study were conducted on a Windows 10 system equipped with a 13th Gen Intel® Core™ i5-13600KF 3.50 GHz CPU and an NVIDIA GeForce RTX 4070 GPU, with data processing completed using the PyCharm 2023 development environment; the specific software environment for model training includes Python 3.9.18, PyTorch 2.0.0 ,CUDA 11.8, and key dependency libraries such as numpy 1.24.3, pandas 2.1.3, scikit-learn 1.3.2 , matplotlib 3.8.2.Data were sourced from the *MassBank.us* public dataset and the Department of Chemistry Fundamentals at Kunming University of Science and Technology. The *MassBank.us* dataset (official public access link: https://massbank.us/) includes LC-MS/GC-MS data for over 20,000 compounds, covering detection results from six major instrument types (e.g., Thermo Q-Exactive, Bruker maXis) and encompassing diverse compound structures and instrument noise characteristics. The *KUST-MS* dataset contains 12,628 positive ion mass spectra of 1,203 compounds under seven voltage conditions (10 to 70 *V*), with differentiated ion intensity distributions generated through bombardment at varying energies, thus provide multidimensional data support for low-abundance ion feature analysis.

For comparative assessment, three groups were established: the control group employed the *MSE* similarity algorithm and its fixed-weight version (with a coefficient of 0.25), the experimental group utilized the low-abundance ion-enhanced model described in this study and the classic method group (including cosine similarity and Pearson correlation coefficient).

Evaluation metrics included *C*ohen's effect size, paired *t*-tests, *Average* similarity score and the standard deviation of the similarity score. *Cohen's d* was used to measure the practical significance of intergroup mean differences, with values of d of 0.8 or higher defined as large effects. Paired *t*-tests were performed to assess

differences in average similarity scores between the experimental group and control group, with a significance threshold set at *p*-value below 0.05. The mean similarity score (SimMean) reflects the central tendency of similarity calculation. The higher the value, the better the capture effect on the structural similarity between samples. The standard deviation of the similarity score (SimStd) quantifies the degree of dispersion of the similarity score. A larger value indicates a better degree of dispersion of the similarity score, and the model can better distinguish different compounds, especially those with highly similar structures.

Data Preprocessing Effect: In the data preprocessing stage, the *db4* wavelet basis was used to decompose the data into four layers, and the soft-threshold denoising algorithm was applied for signal preprocessing. The typical original mass spectrometry signals and their wavelet denoising results are presented in Fig. 4.

As clearly shown in the Fig. 4, wavelet denoising achieved remarkable results in removing the original signal noise. After denoising, the high-frequency noise spikes in the original signal were significantly reduced, the ion peak contours became clearer, and the signal baseline became more stable.

Implementation effect of dynamic weighting strategy: This section evaluates the effectiveness and robustness of the proposed dynamic weighting strategy through convergence analysis, statistical comparison, and entropy-aware performance assessment using the MassBank.us and KUST-MS datasets.

Model convergence behavior: Model training was conducted using the Adam optimizer with a learning rate of 0.001. The ReLU activation function was employed to introduce nonlinearity. *Dropout* (0.3) and *Batch Normalization* were applied to suppress overfitting. As illustrated in Fig. 5, the loss value decreased rapidly during the initial training phase due to effective gradient-based optimization and gradually stabilized after approximately 100 iterations, converging to a final value of 0.012.

In the early training stage, due to the random initialization of weights and biases in the network, there was a large deviation from the optimal solution. Through parameter updates using the small-batch gradient descent method, the model was able to quickly capture data patterns, resulting in a rapid decrease in the loss value. As training progressed, the model gradually approached the optimal solution, the parameter update amplitude decreased, and the rate of loss reduction also gradually slowed down. Eventually, the loss value stabilized at 0.012, indicating that the model successfully learned the data characteristics and achieved a good fitting effect on the training set.

Comparison of *MSE* similarity between dynamic and fixed - coefficient strategies: Due to generally low similarity scores among substances, 1,000 groups were constructed from the MassBank.us dataset (each group corresponds to one isomer, containing 5–6 mass spectrometry samples of the same isomer under different collision energies: 10–60 eV) and 300 groups from the KUST-MS dataset (each group corresponds to one compound, containing 7 mass spectrometry samples of the same compound under gradient voltage conditions: 10 V, 20 V, 30 V, 40 V, 50 V, 60 V, 70 V) for in-depth analysis. Average similarity score calculation, paired - sample *t* - test, and correlation analysis were performed on the *MSE* similarity scores under the dynamic and fixed - initial - coefficient strategies. During the research, the dynamic coefficient weight adjustment strategy and the fixed initial coefficient strategy were established, and the calculation results of *MSE* similarity after low-abundance ion weight optimization under the two strategies were systematically compared (Figs. 6 and 7).

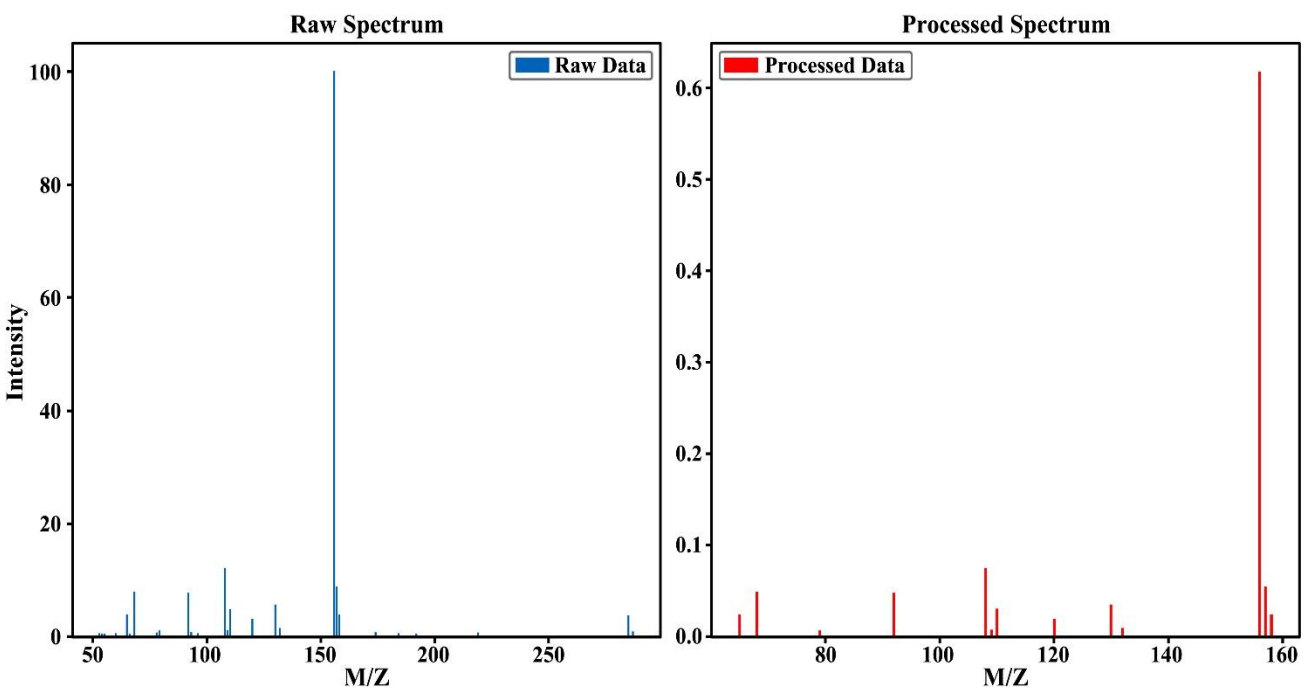

Fig. 4. Comparison of mass spectral signals before and after wavelet denoising.

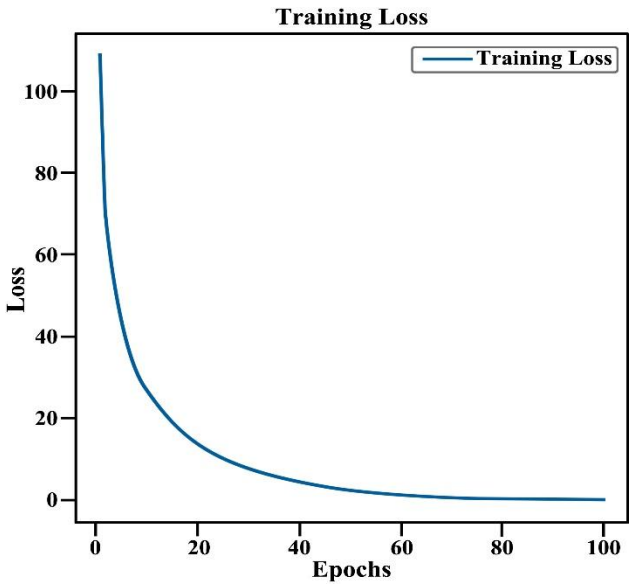

Fig. 5. *MLP* training loss curve.

Figure 6 and Figure 7 show that the overall distribution of the dynamic similarity score (orange points) was significantly higher than that of the fixed similarity score (blue points), and the trends of the two datasets of data were highly consistent. This indicates that in the similarity calculation of isomers from the *MassBank.us* dataset and 10 - 70*V* voltage mass spectrometry data in the *KUST-MS* dataset, the dynamic contribution adjustment strategy successfully achieved adaptive adjustment adaptive adjustment through the feature-to-contribution mapping of the *MLP* network for low-abundance ions, effectively amplifying the contribution of this type of ion information in similarity measurement. The similarity evaluation after dynamic enhancement exhibited higher values in the dataset, suggesting that dynamic weighting could more sensitively capture the similar features between samples and significantly improve the overall level of similarity evaluation.

Statistically, a *p*-value less than 0.05 were observed in approximately 54.12% of the groups in the *KUST-MS* data and 50.97% of the groups in the *MassBank.us* data. *P*-value less than 0.05 serve as a critical criterion for determining statistically significant differences in mean values between two groups of data. This indicates that significant differences in mean values between dynamic similarity and fixed similarity were identified in more than half of the groups. Effect size analysis provided additional insight into the practical significance of the observed effects: Among these samples, 88.0% exhibited large effect sizes, defined as a *Cohen' d* value of 0.8 or higher, and 29.4% were characterized by extremely large effect sizes, corresponding to a *Cohen' d* value of 1.5 or higher. Collectively, these results, from a statistical standpoint, demonstrate the substantial impact of the low-abundance ion enhancement strategy on similarity scores and provide additional validation for the effectiveness and superiority of this strategy in *MSE* similarity calculations.

In the correlation analysis, the correlation coefficients of each group were generally high, mostly close to 1. This indicates a strong positive correlation between dynamic similarity and fixed similarity, implying that the dynamic strategy, through nonlinear feature importance modeling, introduced new information. Especially in mass spectrometry data dominated by low-abundance ions, its ability to capture feature differences showed obvious advantages, fully verifying the differentiated enhancement effect of the adaptive contribution mechanism on low-abundance ion information. Compared with the fixed initial coefficient strategy, the proposed model achieved remarkable results in improving the similarity score, providing a more accurate and effective approach for mass spectrometry data similarity calculation.

Comparison with classic similarity methods: To fully demonstrate the comprehensive advantage of the proposed model in similarity calculation accuracy and stability, a horizontal comparison was conducted with cosine similarity and Pearson correlation coefficient using the same dataset. The performance metrics (Mean, Std) of each method are shown in Table 1.

As shown in Table 1, the proposed model outperformed the other three methods in both datasets in terms of comprehensive performance: The proposed model's average similarity score is 0.0413 higher than the fixed-weight mass spectrometry entropy algorithm, 0.0842 higher than cosine similarity, and 0.0927 higher than the Pearson correlation coefficient, fully demonstrating superior structural similarity capture. This advantage stems from the model's ability to balance "structural feature capture" and "differentiation of similar substances": the dynamic weighting of low-abundance ions improves the accuracy of capturing structural similarities (higher SimMean), while the adaptive adjustment mechanism increases the dispersion of similarity scores (higher SimStd), enabling more precise discrimination between highly similar isomers.

Visual comparative analysis of heat maps: Four groups of substances from the *KUST-MS* dataset (No.29, No.35, No.65, and No.70) under 10 - 70*V* voltage bombardment and four groups of isomers ($C_{12}H_{14}N_4O_2S$, $C_{25}H_{38}O_5$, $C_{31}H_{36}N_2O_{11}$, $C_{10}H_{10}N_4O_2S$) from the *MassBank.us* dataset under different collision conditions were randomly selected for comparative analysis. The *MSE* data are shown in Table 2 and Table 3. Subsequently, heat maps were drawn to more intuitively observe and analyze the distribution of similarity coefficients under different weighting strategies, some heat maps were presented in Fig. 8.

The analysis of heat map similarity calculation results indicated that the dynamic weighting strategy remarkably enhanced the accuracy of *MSE* similarity calculation. For example, when No.29 was considered, with initial coefficient weighting, the red outside the diagonal gradually fades while the blue gradually deepens. The similarity scores in the low - voltage (10 - 50*V*) range were between 0.3 - 0.8, showing an underestimation. The optimized weighting darkens the color in the low-voltage region, increased the scores to 0.4 - 0.9, and raised the average score from 0.5885 to 0.5927. However, in the 50 - 70*V* high - voltage, low - entropy scenario, due to the model's limited adaptability to extreme parameters, occasional score decreases were observed. The analysis of isomers No.65, No.70 and *MassBank.us* shows that this strategy presents a similar action pattern: in the low collision energy range (low-entropy environment), the similarity scores generally increase, while in the high collision energy (high-entropy) range, the improvement effect is relatively limited.

Statistical significance verification: To quantitatively evaluate the effect of the dynamic weighting strategy, the paired sample *t*-test is adopted in the study and the effect size (*Cohen's d*) is calculated (Stratton *et al*., 2019; Di Leo & Sardanelli, 2020). The analysis results are shown in Table 4.

The dynamic weighting strategy demonstrated significant universality and differential efficacy in improving *MSE* similarity scores. In the *KUST-MS* and *MassBank.us* datasets, for substances with a high proportion of low - abundance ions (such as No.29, No.35, $C_{25}H_{38}O_5$), the optimized method showed the *Cohen's d* effect size generally exceeding 0.8 compared to the unweighted strategy, with some cases reaching 1.5. This indicates the notable impact of dynamic weighting on enhancing low - abundance signals.

In summary, the accuracy of mass spectrometry similarity calculation was improved by the dynamic weighting strategy through the synergistic action of noise suppression and adaptive weight allocation. Its advantages were mainly reflected in the enhanced differentiation of low - abundance ion information. However, the model's adaptability to extreme scenarios needed to be further optimized through iterative improvements.

Table 1. Comprehensive Performance Comparison of Different Similarity Calculation Methods

Dataset	Evaluation index	Proposed model	Fixed-weight MSE	Cosine similarity	Pearson correlation coefficient
MassBank.us (1000 isomers)	SimMean	0.7458	0.7045	0.6616	0.6531
	SimStd	0.7045	0.1979	0.1957	0.1812
KUST-MS (300compounds)	SimMean	0.7648	0.7080	0.6647	0.6751
	SimStd	0.2638	0.1837	0.1987	0.1926

Table 2. *KUST-MS* spectral entropy values under different voltages.

Substance number	10*V*	20*V*	30*V*	40*V*	50*V*	60*V*	70*V*
No.29	1.73378	2.28578	2.211543	2.598199	2.54129	2.547287	2.441941
No.35	1.452412	1.653632	1.850494	1.932888	2.112099	2.11008	2.048027
No.65	2.17084	2.070352	2.09143	1.793943	1.531797	1.05647	0.969591
No.70	1.257836	1.855121	2.208465	2.376669	2.289121	2.282412	2.225369

Table 3. *MassBank.us* spectral entropy values under different collision energies.

Chemical formula	10*eV*	20 *eV*	30 *eV*	40 *eV*	50 *eV*	60 *eV*
$C_{12}H_{14}N_4O_2S$	1.717148	1.863052	1.644509	1.78759	2.44829	2.44829
$C_{25}H_{38}O_5$	1.86519	2.184712	2.096748	2.0568	2.422318	
$C_{31}H_{36}N_2O_{11}$	1.820714	1.244158	1.262042	1.085071	3.712289	2.062902
$C_{10}H_{10}N_4O_2S$	0.615503	1.06403	1.142038	1.78759	2.44829	

Table 4. The comparison results of different substances under different contribution adjustment strategies.

Substance number	Comparison group	Sample size	Mean Change	*t*-tests	*p* value	*Cohen's d*
No.29	Before vs after	7	0.5885 → 0.5927	4.9078	<0.0001	1.0710▲
No.29	Unweighted vs after	7	0.5553 → 0.5927	5.4664	<0.0001	1.1929▲
No.29	Unweighted vs before	7	0.5553 → 0.5885	5.3695	<0.0001	1.1717▲
No.65	Before vs after	7	0.5859 → 0.5869	2.0589	0.053	0.4493
No.65	Unweighted vs after	7	0.5682 → 0.5869	3.9655	0.0008	0.8654▲
No.65	Unweighted vs before	7	0.5682 → 0.5859	4.0693	0.0006	0.8880▲
No.35	Before vs after	7	0.5501 → 0.5536	5.8149	<0.0001	1.2689▲
No.35	Unweighted vs after	7	0.4933 → 0.5536	7.2587	<0.000001	1.5840■
No.35	Unweighted vs before	7	0.4933 → 0.5501	7.3283	<0.000001	1.5992■
No.70	Before vs after	7	0.6320 → 0.6437	2.2404	0.037	0.4889
No.70	Unweighted vs after	7	0.5541 → 0.6437	6.9768	<0.000001	1.5225■
No.70	Unweighted vs before	7	0.5541 → 0.6320	5.4756	<0.0001	1.1949▲
$C_{12}H_{14}N_4O_2S$	Before vs after	5	0.7091 → 0.7248	1.9576	0.0820	0.6191
$C_{12}H_{14}N_4O_2S$	Unweighted vs after	5	0.7058 → 0.7248	2.6787	0.0253	0.8471▲
$C_{12}H_{14}N_4O_2S$	Unweighted vs before	5	0.7058 → 0.7091	0.3238	0.7535	0.1024
$C_{25}H_{38}O_5$	Before vs after	5	0.4901 → 0.5066	3.2473	0.0100	1.0269▲
$C_{25}H_{38}O_5$	Unweighted vs after	5	0.4872 → 0.5066	4.0887	0.0027	1.2930▲
$C_{25}H_{38}O_5$	Unweighted vs before	5	0.4872 → 0.4901	2.8771	0.0183	0.9098▲
$C_{31}H_{36}N_2O_{11}$	Before vs after	6	0.6617 → 0.6678	0.3910	0.7017	0.1009▲
$C_{31}H_{36}N_2O_{11}$	Unweighted vs after	6	0.6095 → 0.6678	3.8441	0.0018	0.9925▲
$C_{31}H_{36}N_2O_{11}$	Unweighted vs before	6	0.6095 → 0.6617	3.9890	0.0013	1.0299▲
$C_{10}H_{10}N_4O_2S$	Before vs after	4	0.5689 → 0.5874	2.5763	0.0497	1.0518▲
$C_{10}H_{10}N_4O_2S$	Unweighted vs after	4	0.5008 → 0.5874	4.1458	0.0089	1.6925■
$C_{10}H_{10}N_4O_2S$	Unweighted vs before	4	0.5008 → 0.5689	2.4604	0.0572	1.0045▲

Significance:*p<0.05，**p<0.01，***p<0.001(double-tail test)
Effect size:▲*Cohen's d*≥0.8 (large effect)，■*Cohen's d*≥1.5 (extremely large effect)

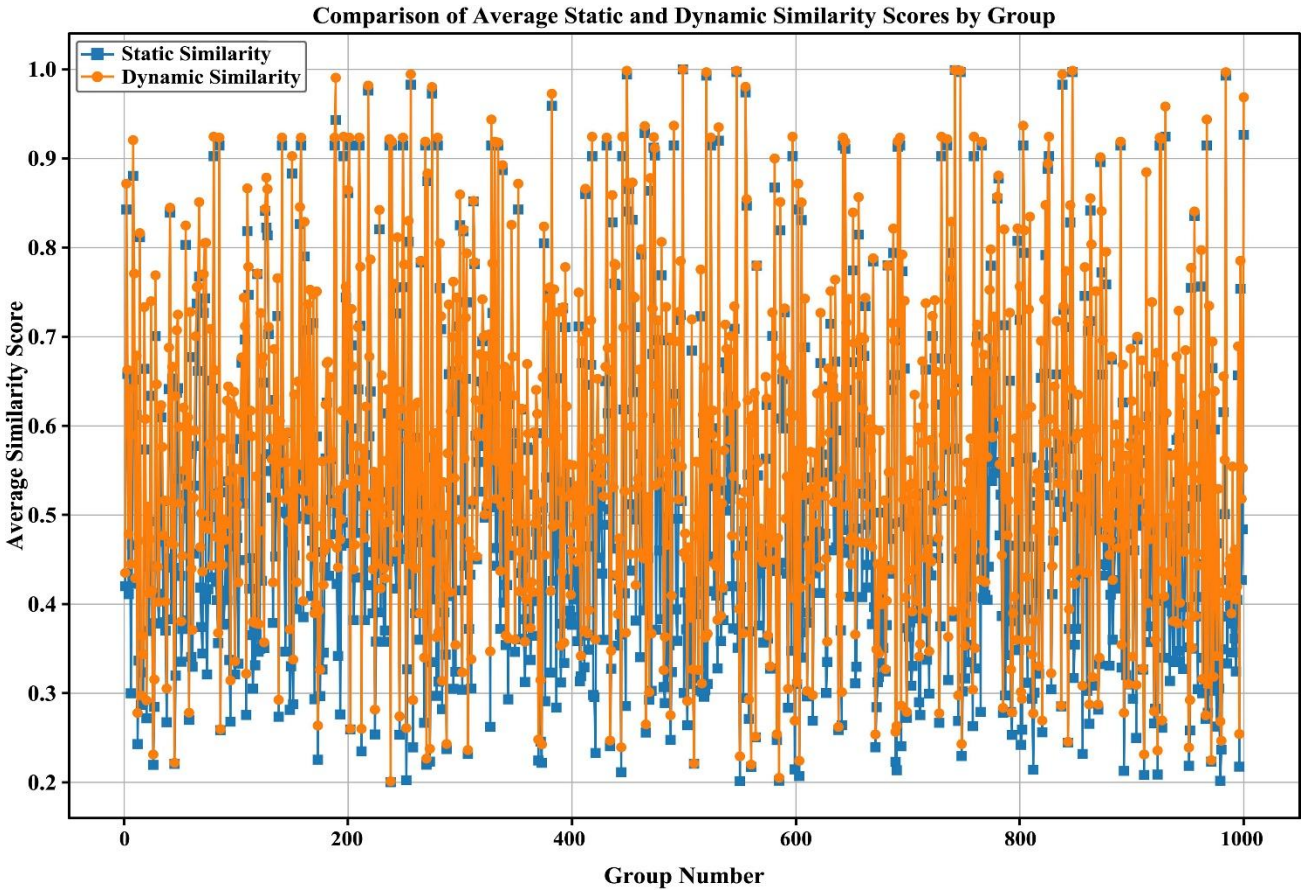

Fig. 6. Dynamic vs. fixed similarity comparison for 1000 isomer pairs in *MassBank.us* dataset.

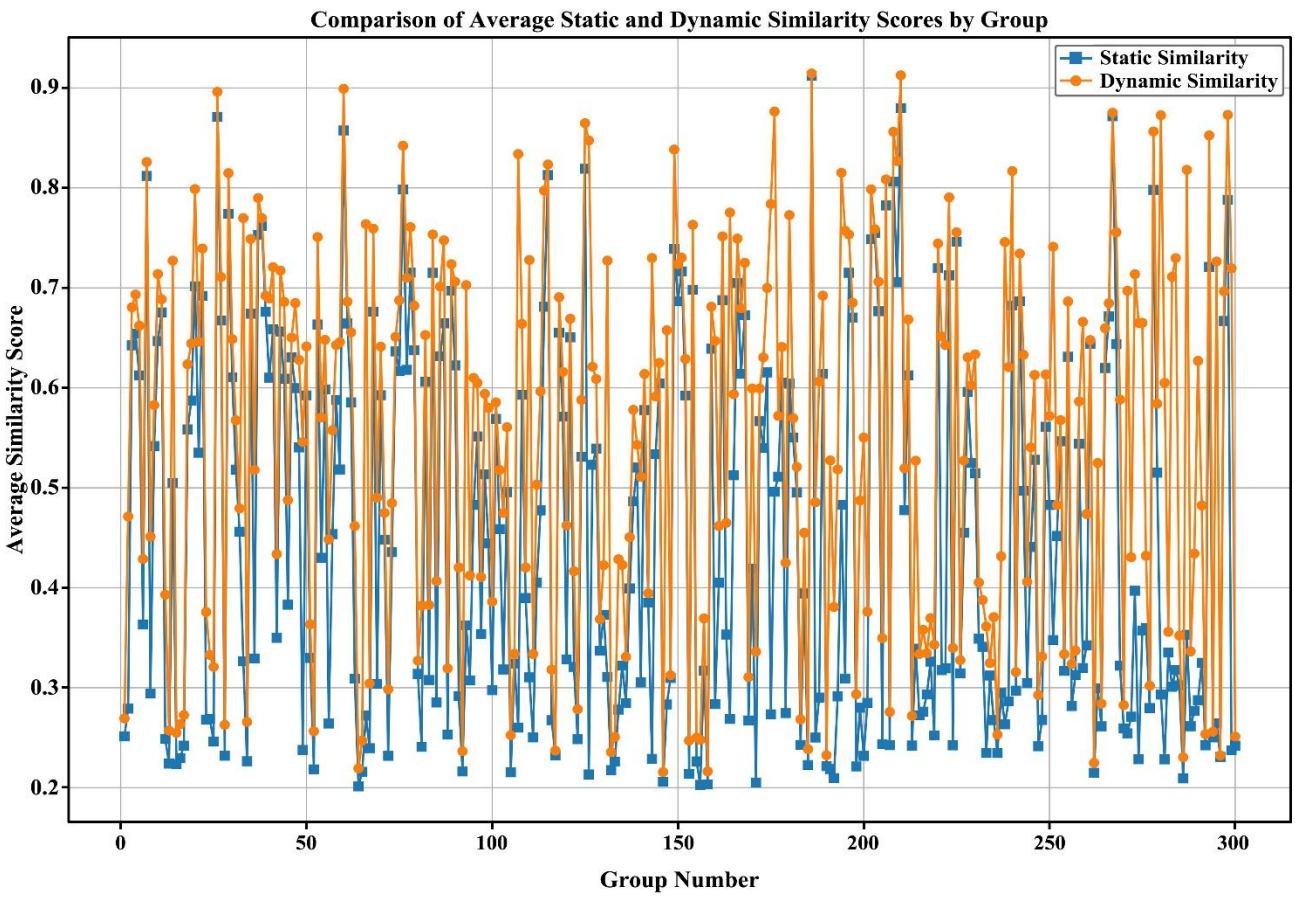

Fig. 7. Dynamic vs. fixed similarity comparison for 300 *KUST-MS* dataset samples under 10–70*V* voltages.

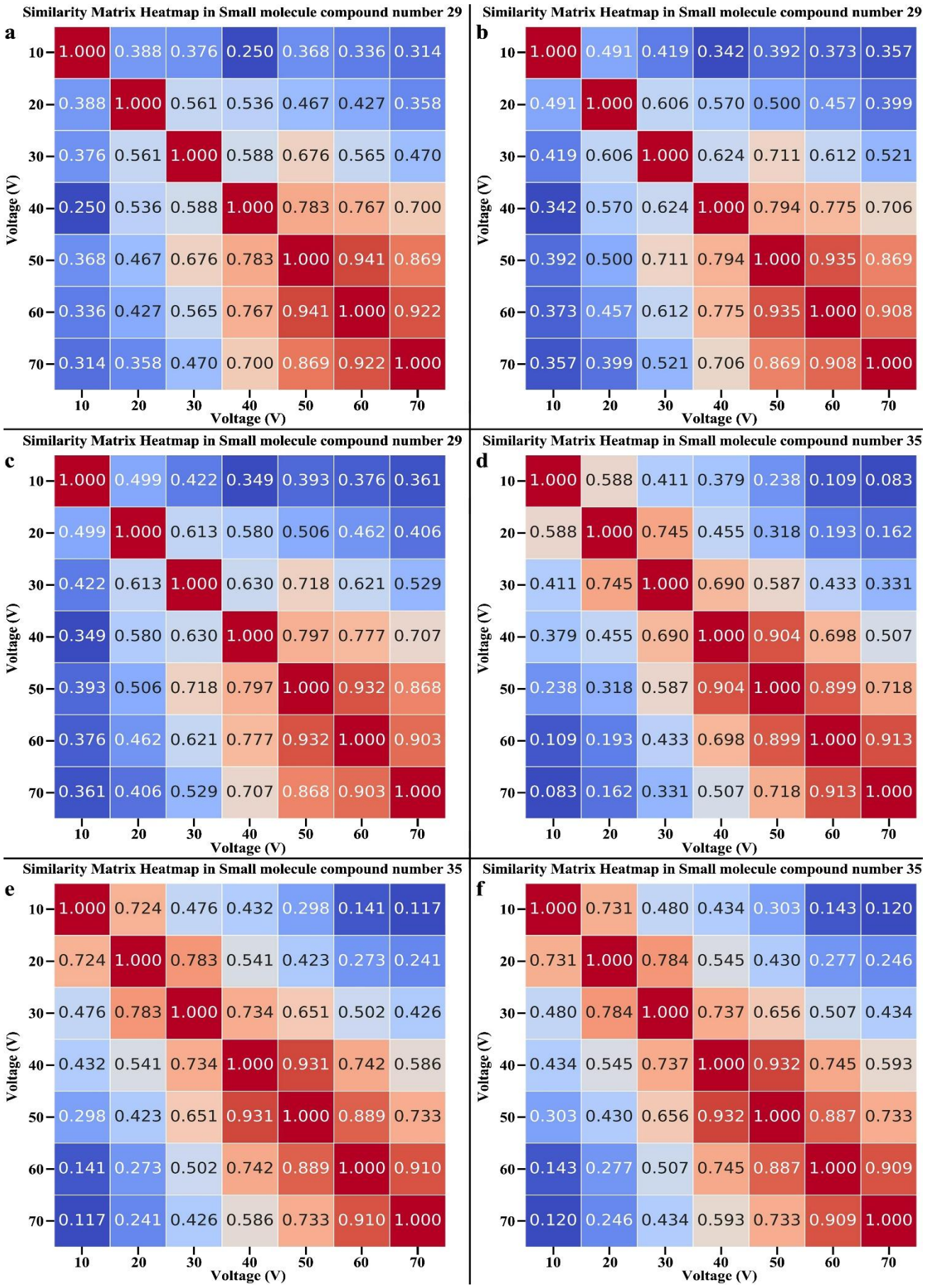

Fig. 8. *MS* Entropy Similarity Coefficient Calculations for Selected Compounds in *KUST-MS* Dataset: a. Unweighted result for Compound No.29; b. Compound No.29 with initial coefficient weighting; c. Compound No.29 with initial coefficient weighting; d. Unweighted result for Compound No.35; e. Compound No.35 with initial coefficient weighting; f. Compound No.35 with optimized coefficient weighting.

Discussion

This study addressed critical challenges in small molecule compound recognition, including the masking of trace structural differences by high-abundance skeleton ions and the limited discriminative power for low-abundance ions, by proposing an *MLP*-based low-abundance ion-enhanced *MSE* similarity model. Data noise was suppressed through four-layer *db4* wavelet *decomposition* and soft-threshold denoising, while differences in instrument responses were eliminated via ion intensity normalization. Features such as *MSE* were extracted to analyze the informational value of low-abundance ions, mean, standard deviation, and skewness were extracted to systematically analyze the informational value of low-abundance ions.

A nonlinear function constrained by MSE was constructed to adaptively respond to ion intensity distribution variations, and a dual-channel MLP network was employed to establish a nonlinear mapping between spectral peak intensities and weights—addresses the fixed-characterization limitation of traditional MSE-based methods (Li *et al*., 2021). This mapping enables collaborative regulation of weight enhancement for low-abundance ions and signal suppression for high-abundance ions. Unlike CNN-LSTM hybrid models that rely on large labeled datasets and suffer from high computational costs, the proposed MLP-based architecture achieves efficient parameter optimization with stable convergence (loss = 0.012 after 100 iterations), balancing performance and practicality (Seddiki *et al*., 2023).

Validation using the MassBank. us and KUST-MS datasets demonstrates that the method achieves favorable results: 50.97%–54.12% of groups exhibit statistically significant differences ($p < 0.05$), 88.0% of samples reach large effect sizes (*C*ohen's d $\geq$ 0.8), and 29.4% achieve extremely large effect sizes (*C*ohen's d $\geq$ 1.5). These outcomes confirm that the model significantly enhances the contribution of low-abundance ions in similarity calculations, effectively addressing the underestimation of characteristic ions in low-entropy spectra ($H < 3$) where high-abundance ions dominate. Compared with classic methods (cosine similarity, Pearson correlation coefficient) and fixed-weight MSE, the proposed model exhibits superior structural difference resolution, as it avoids "equalized" weighting and adaptive adjustment based on spectral characteristics.

The method systematically improves the accuracy and robustness of mass spectrometry similarity calculations, providing reliable technical support for fields such as metabolomics, environmental detection, and drug identification. Future research will focus on adaptive optimization for extreme spectral scenarios, the development of automated parameter selection algorithms, and the design of lightweight neural network architectures to further enhance the analytical efficiency and generalization capability for large-scale complex mass spectrometry data.

Nevertheless, the study also identifies certain limitations. In high-entropy or extreme spectral scenarios, where ion intensity distributions are relatively uniform or highly variable, the improvement effect is less pronounced. Future research will therefore focus on adaptive optimization strategies for extreme entropy conditions, automated parameter selection mechanisms, and the development of lightweight neural network architectures to further enhance scalability and generalization for large-scale mass spectrometry datasets.

Acknowledgements

The authors acknowledge the National Natural Science Foundation of China (Grant: 82160787), the Research on the Quality Standardization of the Chemical Substance Basis of Guizhou Jinshi Oblique and Iron Sheet Stone (Grant: 20025800400).

Conflict of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Authors Contribution: Jiayao Pan: Conceptualization, methodology, software, validation, investigation, data curation, writing original draft, visualization. Yong Li*: Supervision, project administration, funding acquisition, review & editing. RuiYang Li: Methodology, validation, resources, review & editing. *Corresponding author: Yong Li, E-mail: leon@kust.edu.cn

References

Di Leo, G. and F. Sardanelli. 2020. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *Eur. Radiol. Exp.*, 4(1): 18.

Fang, P., S. Yu, X. Ma, L. Hou, T. Li, K. Gao, Y. Wang, Q. Sun, L. Shang, Q. Liu, M. Nie and J. Yang. 2024. Applications of tandem mass spectrometry (MS/MS) in antimicrobial peptides field: Current state and new applications. *Heliyon*, 10: e28484.

Hart, C.E., T. Kind, P.C. Dorrestein, D. Healey and D. Domingo-Fernández. 2024. Weighting low-intensity MS/MS ions and m/z frequency for spectral library annotation. *J. Amer. Soc. Mass Spectrom.*, 35(3): 449-455.

Jones, L.M. 2020. Mass spectrometry-based methods for structural biology on a proteome-wide scale, Biochem. *Soc. Trans.*, 48(3): 945-954.

Li, Y. and O. Fiehn. 2023. Flash entropy search to query all mass spectral libraries in real time, *Nat. Methods*, 20(12): 1475-1478.

Li, Y., T. Kind, J. Folz, A. Vaniya, S.S. Mehta and O. Fiehn. 2021. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification, *Nat. Methods*, 18(12): 1524-1531.

Seddiki, K., F. Precioso, M. Sanabria, M. Salzet, I. Fournier and A. Droit. 2023. Early diagnosis: End-to-End CNN–LSTM models for mass spectrometry data classification. *Anal. Chem.*, 95(36): 13431-13437.

Stratton, K.G., B.J.M. Webb-Robertson, L.A. McCue, B. Stanfill, D. Claborne, I. Godinez, T. Johansen, A.M. Thompson, K.E.

Burnum-Johnson, K.M. Waters and L.M. Bramer. 2019. pmartR: Quality control and statistics for mass spectrometry-based biological data. *J. Proteome Res*., 18(3): 1418-1425.

Thomas, S.N., D. French, P.J. Jannetto, B.A. Rappold and W.A. Clarke. 2022. Liquid chromatography-tandem mass spectrometry for clinical diagnostics. *Nat. Rev. Methods Prim.*, 2: 96.

Tian, D., M. Li, Y. Shen and S. Han. 2023. Intelligent mining of safety hazard information from construction documents using semantic similarity and information entropy, *Eng. Appl. Artif. Intell.*, 119: 105742.

Wenk, D., C. Zuo, T. Kislinger and L. Sepiashvili. 2024. Recent developments in mass-spectrometry-based targeted proteomics of clinical cancer biomarkers. *Clin. Proteomics*, 21: 6.

Xu, Y., Q. Ou, X. Wang, F. Hou, P. Li, J.P. van der Hoek and G. Liu. 2023. Assessing the mass concentration of microplastics and nanoplastics in wastewater treatment plants by pyrolysis gas chromatography–mass spectrometry. *Environ. Sci. Technol*., 57(8): 3114-3123.

Zhang, W., L. Xu and H. Zhang. 2023. Recent advances in mass spectrometry techniques for atmospheric chemistry research on molecular-level. *Mass Spectrom. Rev.*, 43(5): 1091-1134.