Mass Spectra Pairwise Comparison with Siamese Networks for Forensic Applications

Diego Brito

Parisa Azizian

M.Sc. in Applied Modelling & Quantitative Methods

Trent University

April 2025

Table of Contents

Table of Figures

Abstract

This study presents a novel approach to forensic drug identification using Siamese Neural Networks (SNNs) to compare mass spectrometry data. Traditional spectral matching techniques often struggle with instrument variability, incomplete reference libraries, and newly emerging compounds. By leveraging similarity learning, this work introduces a flexible and scalable solution capable of determining whether two mass spectra belong to the same compound without relying on fixed-class classification. The model was trained and evaluated on a curated dataset of synthetic opioids, cathinones, and cannabinoids, achieving 99.01% accuracy in pairwise identification and 100% accuracy in compound family classification. Key preprocessing steps—including max normalization and rational transformation—enhanced spectral resolution and low-intensity signal detection. Embedding representations generated by the Siamese network demonstrated strong discriminatory power, validated through both quantitative metrics and dimensionality reduction visualizations. These findings highlight the potential of deep learning-based similarity frameworks to advance forensic mass spectrometry by improving accuracy, adaptability, and interpretability in real-world applications.

Introduction

Drug identification remains a cornerstone of forensic toxicology, law enforcement, and pharmaceutical safety, demanding analytical techniques capable of detecting and distinguishing compounds with high precision. Among these techniques, mass spectrometry (MS) has become the gold standard due to its unparalleled sensitivity, specificity, and ability to generate reproducible molecular fingerprints [2], [4]. However, traditional MS workflows, which often rely on manual interpretation, spectral libraries, and rule-based classification, are increasingly inadequate in the face of novel synthetic drugs and complex sample matrices [1]. As a result, forensic science is turning toward advanced computational solutions, particularly machine learning (ML), to enhance the robustness, scalability, and automation of spectral analysis [12].

This project explores the application of Siamese Neural Networks (SNNs) for forensic mass spectrometry [6]. Unlike conventional classifiers that map spectra to predefined classes, SNNs learn to evaluate similarity between spectral pairs, making them ideal for tasks involving novel or unregistered compounds [7]. By focusing on pairwise similarity, this method promises generalization beyond the training set, reducing reliance on exhaustive libraries and manual thresholds. Our study uses a curated dataset of forensic spectra, leveraging preprocessing, feature transformation, and embedding learning to develop a scalable and precise system for drug identification. The following literature review contextualizes this work by tracing the evolution of mass spectrometry in forensic science

[7],[10], the limitations of conventional approaches, and the growing role of AI in spectral analysis [12].

Literature Review

Mass spectrometry has long been recognized as a central tool in analytical chemistry due to its ability to accurately determine the mass-to-charge ratio of ions, generating a distinctive spectrum for each compound. In forensic contexts, MS is favored for its high sensitivity and selectivity, which enable the identification of trace compounds within complex matrices. However, as Wallace and Moorthy point out, the effectiveness of MS-based identification often hinges on the quality and completeness of reference libraries, such as those maintained by the NIST Mass Spectrometry Data Center [3]. Their further analysis highlights NIST's role in standardizing spectral interpretation across labs [9], though challenges remain in covering the ever-expanding spectrum of emerging drugs. While these libraries support automated matching and reproducibility, they struggle to keep pace with the emergence of novel psychoactive substances (NPS), designer drugs, and structural isomers—substances that frequently appear in forensic investigations yet are often absent from existing reference databases.

To address the limitations of library-dependent methods, researchers have proposed algorithmic approaches for spectral comparison. Moorthy and Sisco introduced the Min-Max Test as a statistical alternative to subjective thresholding in spectral discrimination, offering a more objective framework for determining whether two spectra represent the same compound [7]. Their method laid the groundwork for similarity-based learning in spectral analysis by showing that fixed thresholds and manual interpretation could be replaced by data-driven metrics. Our work builds directly on this foundation by extending the principle of similarity to a machine learning paradigm, using SNNs to learn a continuous embedding space in which spectral similarity corresponds to spatial proximity.

Instrument variability, a longstanding issue in mass spectrometry, further complicates compound identification. Mehnert et al. tackled this by training a binary classifier on mass spectra collected from different instruments, demonstrating that machine learning can generalize across platforms to reduce false positives and negatives in forensic drug detection [8]. This is particularly relevant to our work, as the use of SNNs also mitigates inter-instrument variation by focusing on relational rather than absolute features. By comparing spectra in pairs, SNNs reduce dependence on the exact spectral profile and instead emphasize comparative consistency—an approach more resilient to instrument- or condition-induced noise.

The challenge of distinguishing structurally similar compounds—especially positional isomers—has also drawn attention. Bonetti employed multivariate statistical tools like Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to highlight subtle spectral differences often overlooked in traditional matching methods [10]. While effective, these approaches still rely on feature extraction and fixed classifiers. Our method enhances this line of inquiry by learning from raw or minimally preprocessed spectral

vectors and capturing both local and sequential features using a hybrid CNN-LSTM architecture. This allows the network to model fine-grained variations that are critical in distinguishing closely related substances.

Parallel advancements in spectral data processing have also informed this study. Loahavilai et al. demonstrated how chemometric techniques such as PCA and Partial Least Squares Regression (PLSR) can differentiate and quantify components in mixtures, particularly in pharmaceutical and food safety contexts [16]. Similarly, Krier et al. developed mutual information-based feature selection methods to improve classification accuracy while minimizing computational cost [18]. In metabolomics, Fiehn [5] emphasized combining targeted and untargeted GC-MS profiling to uncover both known and novel compounds, reinforcing the need for adaptable spectral workflows. However, despite their robustness, these methods often rely on well-defined features and high-quality data. In contrast, our approach uses learned features from binned and transformed spectra, enabling flexible adaptation to noisy or complex datasets.

The influence of machine learning on mass spectrometry continues to grow. Barea-Sepúlveda et al. evaluated preprocessing and classification strategies for gasoline samples using Support Vector Machines (SVMs) and Random Forests (RFs), underscoring the importance of feature engineering and normalization in spectral workflows [11]. Their work emphasizes how data preprocessing—such as baseline correction, peak alignment, and normalization—can significantly impact classification performance. Our use of max normalization and rational transformation reflects similar motivations: to reduce data variance and amplify informative features.

To move beyond traditional models, Beck et al. conducted a comprehensive review of deep learning in mass spectrometry, detailing how architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and autoencoders are being adopted to automate and enhance spectral interpretation [12]. They argue that deep learning not only improves classification accuracy but also facilitates the discovery of previously unrecognized patterns in spectral data. Our integration of CNN and LSTM layers echoes this insight, as these components capture both local peak structures and long-range dependencies across the mass-to-charge axis.

In high-dimensional applications such as proteomics and metabolomics, ensemble methods and decision-tree algorithms have proven especially useful. Geurts et al. applied Random Forests and Boosted Decision Trees to proteomic mass spectrometry data, demonstrating improved classification and robustness in noisy environments [13]. Similarly, Streun et al. developed ANN-based systems for handling high-resolution spectral data, emphasizing their advantages over linear statistical techniques in capturing nonlinear relationships [15]. These findings reinforce the broader trend toward using neural network ensembles and deep architectures to extract chemically meaningful features from complex spectra.

The transition from laboratory settings to real-time, field-deployable systems also informs the relevance of our project. Wang et al. explored the use of CNNs for single-particle mass

spectrometry (SPMS), automating the identification of airborne particles in environmental monitoring applications [14]. Their study highlights the potential of deep learning in handling high-throughput, low-latency spectral analysis—capabilities essential to forensic workflows that require rapid decision-making.

Advances in hardware have supported these algorithmic developments. Bristow and Webb conducted an intercomparison of high-resolution instruments, such as time-of-flight (TOF), Fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap MS systems, showing that while resolution varies, the choice of instrument significantly impacts spectral accuracy and reproducibility [20]. Jennifer Colby's work on HRMS data analysis further stressed the need for standardized parameter optimization across labs to ensure consistency [21]. These challenges underscore the importance of software models—like our SNN framework—that can abstract away some of the instrumental variability and deliver consistent, data-driven decisions across different conditions.

Beyond instrument-focused studies, practical applications in forensic science remain a driving force. Mauer et al. used infrared spectroscopy to detect melamine contamination in infant formula, a case that underscores the public health implications of rapid and accurate chemical screening [19]. This aligns with our project's goal of automating drug identification using a data-driven framework capable of operating under forensic constraints.

Our literature review would be incomplete without considering the importance of model interpretability and decision transparency in forensic science. Unlike commercial or exploratory applications, forensic analyses often support legal proceedings—demanding explanations that are both understandable and defensible in court. This requirement has led researchers like Jennifer Colby and Bonetti to emphasize reproducibility and analytical clarity in their methodologies [21, 10]. In line with these priorities, our work strives to balance the predictive power of deep learning with transparency, using embedding visualizations and confusion matrices to maintain interpretability and build trust. This approach also resonates with the work of Gullo et al. [17], who applied spectral clustering to time-series mass spectrometry data, demonstrating how meaningful structural patterns can emerge even without explicit supervision.

Notably, our study extends the work of Moorthy and Sisco by embedding their insights on statistical discrimination into a deep learning context. While their Min-Max Test defined a new standard for objective spectral comparison, it still required manually chosen metrics and thresholds. Our SNN-based model learns a flexible, non-linear embedding where the notion of "similarity" is learned directly from the data. Furthermore, we augment this by dynamically calibrating distance thresholds and applying classifiers to predict compound identity based on embedding distances, providing a more nuanced and scalable decision-making system.

The interpretability of the learned embeddings is also significant. By training the SNN to cluster spectra from the same compound closely in the embedding space, we enable subsequent classifiers to operate with high accuracy and low complexity. This structure is not only useful for pairwise comparison but also for higher-level classification tasks, such

as identifying the compound family. Our successful use of a secondary classifier to distinguish synthetic opioids, cathinones, and cannabinoids validates the chemical relevance of the learned representations and offers a path toward more integrated multi-task systems.

Dimensionality reduction techniques such as PCA, t-SNE, and UMAP provide visual confirmation of the model's ability to extract meaningful patterns. Clear clustering of compound families in these projections supports our hypothesis that deep similarity learning captures both fine and coarse-grained relationships between chemical entities. These visualizations are not merely illustrative; they play a crucial role in model interpretability and user trust—key considerations in forensic applications where transparency and reproducibility are paramount.

In conclusion, the reviewed literature provides a comprehensive backdrop for our study, highlighting the persistent challenges in spectral variability, compound novelty, and workflow automation. It also showcases the rapid progression toward AI-powered solutions in analytical chemistry. By leveraging a Siamese Neural Network, this project addresses many of the limitations identified in past research, offering a system that is both accurate and adaptable. Our work contributes to the growing body of evidence supporting similarity learning as a transformative approach for forensic mass spectrometry, with implications extending to pharmaceuticals, environmental science, and beyond. The methodology developed here has the potential to not only improve the identification of known substances but also provide a foundation for the detection of emerging and structurally ambiguous compounds, which continue to challenge the forensic community.

Dataset Overview

This study employs a curated mass spectrometry dataset introduced in *The Min-Max Test: An Objective Method for Discriminating Mass Spectra* by Moorthy and Sisco [7]. The dataset includes spectra from 136 chemical compounds relevant to forensic applications, organized into three chemically distinct families: synthetic opioids (CM1), synthetic cathinones (CM2), and synthetic cannabinoids (CM3). Each compound is represented by 10 replicates, providing a robust basis for evaluating intra-class consistency and enabling meaningful inter-class comparisons.

Each replicate is stored in a separate CSV file containing three columns: the mass-to-charge ratio (m/z), the corresponding ion intensity (counts), and a point index. These spectra represent one-dimensional chemical fingerprints obtained through mass spectrometric analysis. An accompanying metadata file provides information on compound identity, category assignment, molecular formula, and substance ID, facilitating structured analysis and label-driven learning.

The dataset encompasses approximately 445,500 data points across all replicates, with each file containing 330 measurements. This uniform structure supports efficient

preprocessing, binning, and modeling operations. The consistent number of replicates per compound also ensures statistical balance and enables stratified data splitting for training, validation, and testing.

Initial exploratory data analysis revealed systematic variations in m/z distributions and intensity patterns across compound families, highlighting chemically meaningful differences. These spectral differences are critical to the study's objective of using similarity learning to detect subtle structural and compositional distinctions. Distinctive peak locations and intensity profiles across compound families indicate that the dataset contains features suitable for both pairwise comparison and supervised classification tasks.

Further analysis of representative spectra from each category demonstrates unique fragmentation patterns, reinforcing the chemical diversity present in the dataset and supporting its use in learning discriminative representations. The results of preprocessing steps and spectral visualizations are discussed in detail in the Results section.

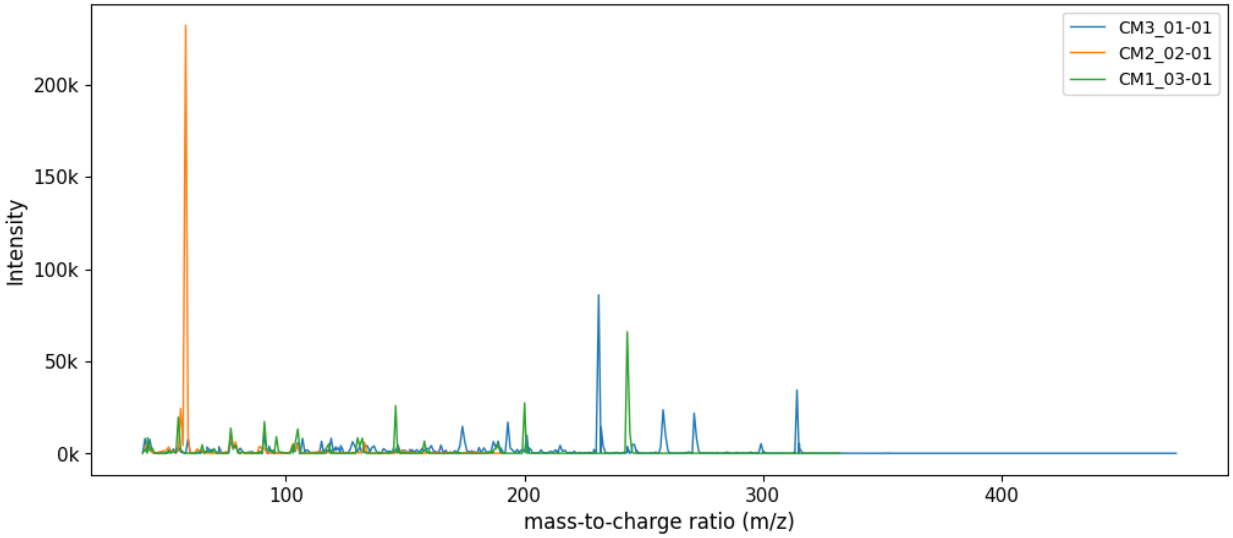

Figure 1. Representative Mass Spectra from CM1, CM2, and CM3 Compound Families

Methodology

This study investigates the application of a Siamese Neural Network (SNN) for determining whether two mass spectrometry samples correspond to the same chemical compound. The methodology was designed with a strong emphasis on robustness, precision, and the interpretability of learned representations. It involves several interconnected stages including data acquisition, preprocessing, spectral transformation, pair construction, network training, classification strategies, and an auxiliary evaluation of embedding quality.

The dataset used in this work comprises mass spectra from 136 compounds, grouped into three categories: synthetic opioids (CM1), cathinones (CM2), and synthetic cannabinoids (CM3). Each compound is represented by ten replicates to support consistent training and validation. The spectral data for each compound is stored in individual CSV files that include an index column, a mass-to-charge ratio (m/z) column, and an intensity column. These files are systematically organized in three separate directories corresponding to the compound categories. A custom data loading script was developed to ingest this data efficiently and convert each file into a structured Data Frame. These Data Frames are stored in a Python dictionary using the filenames as keys, allowing for dynamic access and manipulation. This organization facilitates streamlined preprocessing, pair construction, and subsequent embedding.

To ensure comparability across samples, two preprocessing steps were applied to the spectral data. The first step was max normalization, where all intensity values in a spectrum are divided by its maximum intensity. This process ensures that the most prominent peak in each spectrum is scaled to 1, while the relative structure of the spectrum is preserved. This approach was favored over sum normalization, which compressed peak values too much, and Z-score normalization, which introduced negative values and distorted peak interpretation. The second step involved applying a rational transformation to enhance low-intensity peaks that may hold important chemical information. The selected transformation, defined as $T(x) = 3x / (2x + 1)$, was specifically chosen after testing alternatives like logarithmic scaling, which was found to overly distort lower peaks and flatten the spectral distribution. This rational function retains the upper limit of 1 for normalized values, selectively amplifies smaller values, and maintains the overall spectral shape, improving the visibility of chemically relevant features.

Given that mass spectra vary in length and resolution, it was necessary to convert them into uniform fixed-length vectors suitable for input into a neural network. This was achieved through a binning process using a bin width of one m/z unit. Within each bin, the maximum intensity was recorded, creating a feature vector that captures the most significant peaks in each range while filtering out minor fluctuations and noise. This vectorization process ensures that all input samples share the same dimensionality and are thus compatible with deep learning architectures.

The core training dataset consists of positive and negative pairs. Positive pairs were created by matching different replicates of the same compound. Negative pairs were generated by randomly selecting spectra from different compounds. To prevent model bias due to the vast number of potential negative pairs, an equal number of positive and negative pairs was used. Importantly, data splitting was performed at the compound level, meaning that all replicates of a compound were confined to a single subset—training, validation, or testing—to avoid information leakage. This approach promotes fair evaluation and generalization to unseen compounds. The final dataset used for training included 7,290 balanced pairs, while validation and test sets each contained 2,430 pairs.

The Siamese Neural Network implemented in this study consists of two identical subnetworks with shared weights. Each subnetwork processes a binned spectrum independently and produces an embedding in a learned latent space. The architecture combines convolutional layers for extracting local peak features, LSTM layers for modeling sequential dependencies along the m/z axis, and dense layers for embedding generation. Although mass spectra are not temporal sequences in the conventional sense, the ordered nature of m/z values lends itself well to sequence modeling techniques. The network is trained using a contrastive loss function, which penalizes large distances between embeddings of similar spectra and small distances between embeddings of dissimilar ones. This encourages the formation of an embedding space in which similar compounds cluster closely and dissimilar ones are well-separated. Early stopping and validation monitoring are employed during training to avoid overfitting and ensure the model generalizes well to new data. The best-performing model is saved using a checkpointing system.

Following training, predictions are made by computing the Euclidean distance between embeddings of two input spectra. A default threshold of 0.5 is initially used to classify whether two spectra originate from the same compound. To optimize this decision boundary, we evaluated several thresholds on the validation set and selected the one that produced the highest F1 score. A threshold of 0.45 was found to offer the best trade-off between precision and recall. However, to improve upon this threshold-based approach, we implemented a shallow neural network classifier that receives the Euclidean distance as input and learns a nonlinear mapping to predict binary labels. This classifier not only improved accuracy but also provided a learned probability threshold that adapts more flexibly to the data distribution. The selected threshold (around 0.8149) improved the model's ability to distinguish borderline cases.

To further assess the quality of the learned embeddings, we introduced an auxiliary classification task focused on identifying the compound family of a given spectrum. Using embeddings generated by the Siamese network, we trained a separate neural network classifier to predict whether a spectrum belongs to CM1, CM2, or CM3. The classifier achieved 100% accuracy, indicating that the learned representations not only distinguish individual compounds but also capture higher-level chemical patterns relevant to family classification. This auxiliary task reinforces the conclusion that the embeddings are chemically meaningful and structurally informative.

Finally, dimensionality reduction techniques were applied to visualize the embedding space in two dimensions. We used Principal Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP) to project the high-dimensional embeddings into a 2D space. Among these, PCA provided the clearest separation between compound families, though some overlap was observed with t-SNE and UMAP projections. These visualizations further confirm the structural coherence of the learned embedding space and its ability to reflect real chemical relationships.

This methodology combines carefully selected preprocessing techniques, rigorous pair balancing, a hybrid CNN-LSTM architecture, and flexible classification strategies to build a powerful and interpretable forensic analysis system. The resulting Siamese Neural Network demonstrates strong potential for both fine-grained compound identification and broader family-level classification, offering a scalable and generalizable solution to forensic mass spectrometry challenges.

Results

This section presents the experimental outcomes of the proposed Siamese neural network framework, which was designed to compare mass spectra and classify compounds into their chemical families. Evaluation metrics include classification accuracy, F1-score, precision, recall, confusion matrices, and visualizations using PCA and performance curves. Results are presented for both the pairwise comparison task (same or different compound) and the compound family classification task.

Effect of Preprocessing on Spectral Shape

To evaluate the impact of preprocessing steps, we visualized mass spectra across three stages: the raw spectrum, the max-normalized spectrum, and the rationally transformed spectrum. The raw spectrum exhibited wide fluctuations in intensity, with high peaks often overshadowing subtler signals. Following max normalization, each spectrum was scaled by its maximum intensity value, preserving the relative shape while enabling comparison across different samples. However, this method alone could result in the loss of potentially informative low-intensity features.

Recognizing this limitation, we implemented a rational transformation defined as $T(x) = 3x / (2x + 1)$, This function was selected for its ability to enhance the visibility of smaller peaks while maintaining the relative order of intensities. The transformation improved the representation of low-abundance signals without distorting the broader spectral structure, making the data more informative and machine-learning-ready.

To better demonstrate the progression through these preprocessing steps, we included visualizations for each stage. The first plot shows the raw mass spectrum produced by the loading function, with the x-axis representing the mass-to-charge ratio (m/z) and the y-axis indicating ion intensity. In this unprocessed view, peaks vary widely in height, and while low signals might appear as background noise, they may carry significant chemical insights—especially in spectra without dominant peaks.

A clearer picture emerges after max normalization. Here, the spectrum retains its shape, but intensities are scaled to set the highest peak to 1.0. This standardization minimizes discrepancies across different datasets, allowing analysts to focus on pattern recognition rather than raw signal magnitude. Still, less prominent peaks remain visually subdued, which could limit interpretability.

Finally, the rational transformation yields a more balanced view. Although the tallest peaks remain capped at 1.0, the transformation redistributes the range in a way that emphasizes smaller features. As a result, previously faint but important signals become easier to detect and analyze. This enhancement is especially useful for distinguishing between spectra with subtle variations, strengthening both interpretability and classification accuracy.

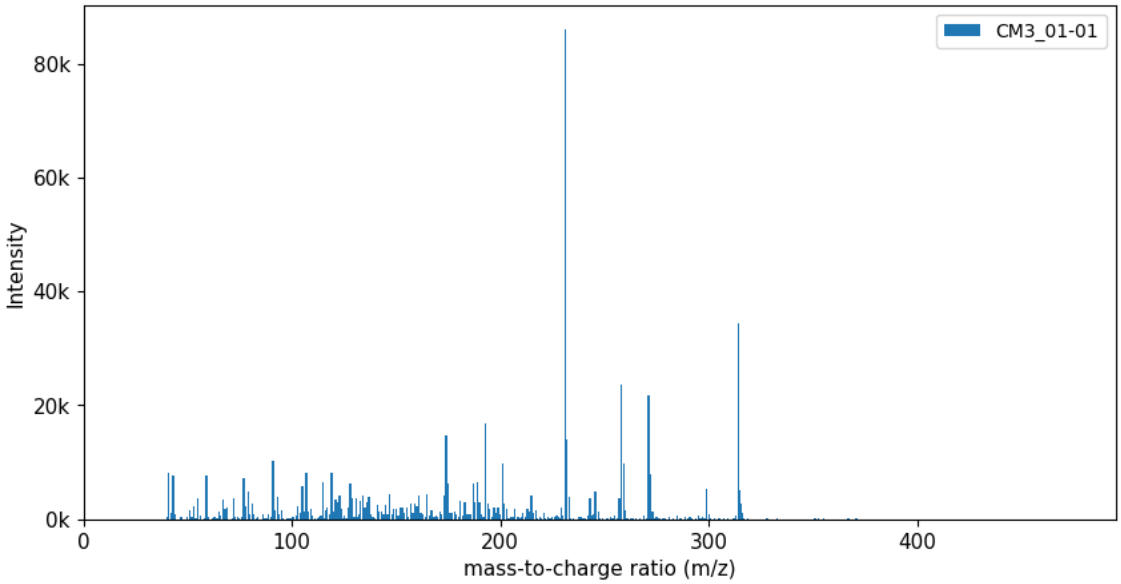

Figure 2. Raw spectrum showing high variance and extreme peaks.

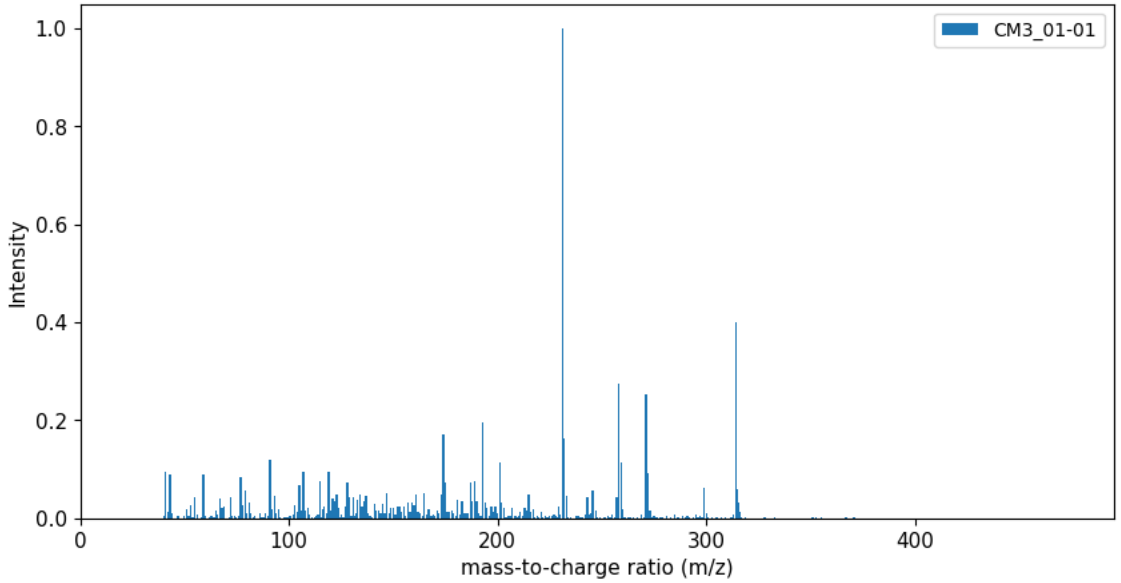

Figure 3. Spectrum after max normalization, where the highest peak is scaled to 1 and the relative shape is preserved.

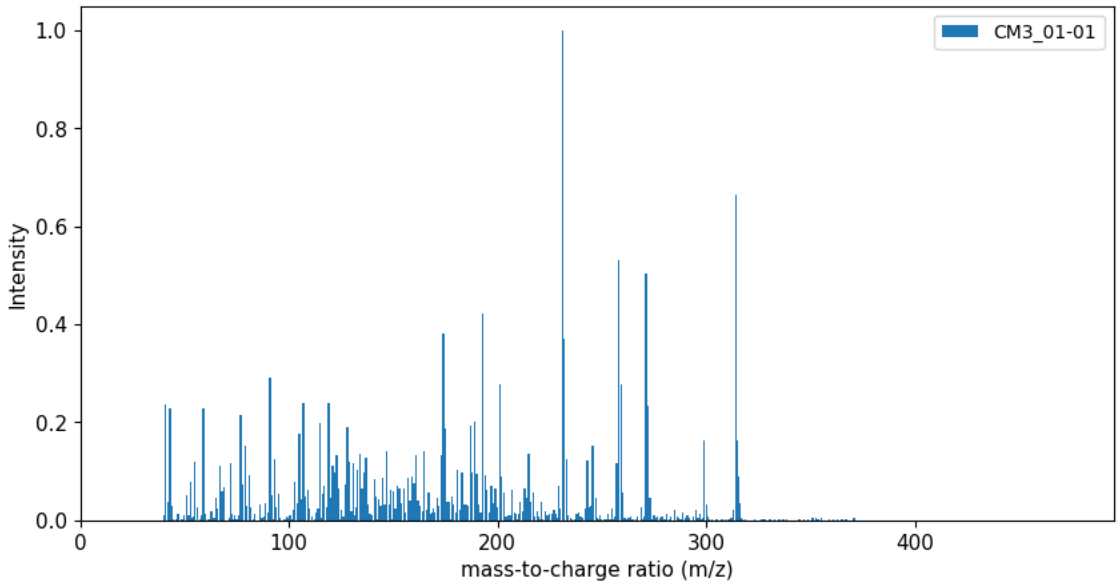

Figure 4. Spectrum after rational transformation, emphasizing smaller peaks and maintaining spectrum shape.

Dataset Statistics and Pair Construction

The dataset was divided by compound into training, validation, and test sets to prevent data leakage between replicates. The training set contained 810 compounds, yielding 3,645 positive and 324,000 potential negative pairs. After random under-sampling, the number of negative pairs was reduced to match the positive pairs, resulting in a balanced dataset of 7,290 training pairs. Similarly, both the validation and test sets comprised 270 compounds each, yielding 1,215 positive and 1,215 sampled negative pairs for a total of 2,430 pairs per set.

This balancing strategy was crucial for preventing model bias toward the dominant negative class. It ensured that the Siamese network was equally exposed to examples of both same and different compound pairs during training and evaluation, enhancing its ability to generalize and maintain performance consistency across classes.

Table1. Dataset Split and Pair Counts for Training, Validation, and Testing

Subset	# of compounds	Positive pairs	Negative Pairs (Before Sampling)	Negative Pairs (Sampled)
Training	810	3,645	324,000	3,645
Validation	270	1,215	35,100	1,215
Test	270	1,215	35,100	1,215

Model Selection / Hyperparameter Tuning

A hyperparameter tuning experiment was conducted to optimize the performance of the Siamese Neural Network for compound classification using mass spectrometry data. Various configurations were tested, differing in preprocessing techniques, CNN architecture, LSTM hidden sizes, fully connected layers, and training parameters such as batch size, learning rate, and distance metrics. Each configuration was evaluated using contrastive loss, with metrics including training time, training and validation loss, and classification accuracy. Observational notes were also documented to aid in interpreting the behavior and stability of each model across runs. A summary of these configurations and their outcomes is provided in the corresponding table.

Table 2. Hyperparameter Configurations and Model Performance Metrics

Preprocessing	CNN filters	CNN kernel	LSTM hidden	FC layers	Epochs	Val Loss	Val Accuracy
transformed	32	5	64	[128, 64]	20	0.0455	0.9393
transformed	32	5	-	[128, 64]	20	0.0312	0.9666
raw	32	5	-	[128, 64]	20	0.0736	0.8978
raw	32	5	64	[128, 64]	20	0.0466	0.9105
transformed	**64**	**8**	**64**	**[128, 64]**	**30**	**0.0015**	**1**

The best results were obtained using transformed input data, contrastive loss, and Euclidean distance, with a CNN consisting of 64 filters and a kernel size of 8, padding set to 2, pooling enabled, and an LSTM hidden size of 64. This configuration (row 6) achieved perfect validation accuracy (1.0) with a validation loss of just 0.0015 in 20 epochs, and training time of 119 seconds. Notably, this setup maintained stable convergence without overfitting. Similar strong performance was observed in rows 9, 10, and 11 with slight architectural variations and 30 training epochs. In contrast, configurations using raw data (rows 3 and 4) or without LSTM/fully connected layers (rows 1, 2) showed lower validation accuracy and higher losses. One setup without pooling (row 8) failed to reach meaningful accuracy levels, never exceeding the 0.5 threshold, highlighting the importance of pooling for effective feature extraction. Cosine distance (row 12) also failed to produce effective learning under the same architecture. Triplet loss was not tested, as it would require a different architecture based on triplet input format. These results helped select a configuration that balances accuracy, efficiency, and stability for the final model.

Model Training and Loss Progression

The Siamese Neural Network was trained for 30 epochs, with loss monitored on both training and validation sets. As shown in Figures 5 and 6, the model quickly converged within the first few epochs, with minimal overfitting and stable validation accuracy across training.

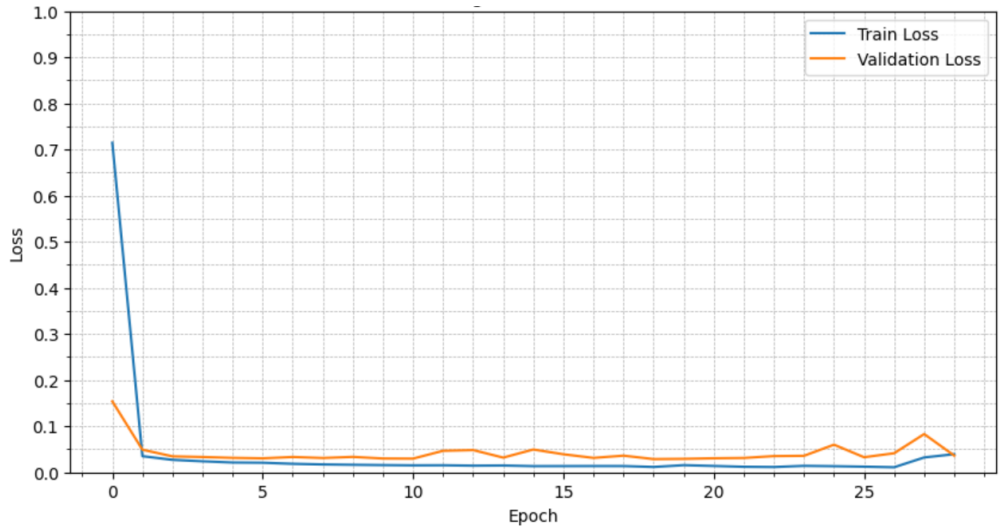

Figure 5. Training and validation loss over epochs.

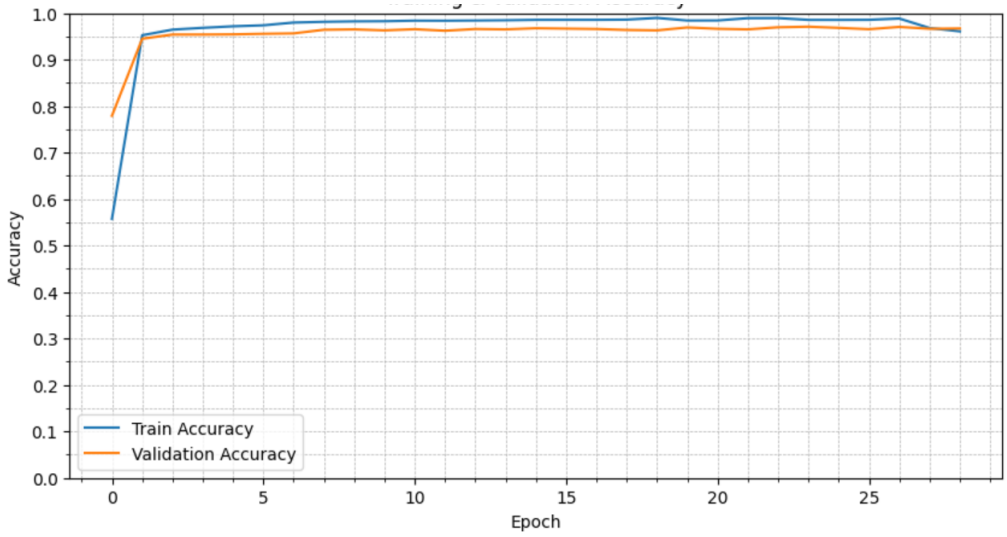

Figure 6. Training and validation accuracy over epochs.

To further verify the model's reliability, we conducted 5-fold cross-validation, training the model on different data partitions and observing performance consistency. Figures 7 and 8 show validation accuracy and loss per epoch across the five folds.

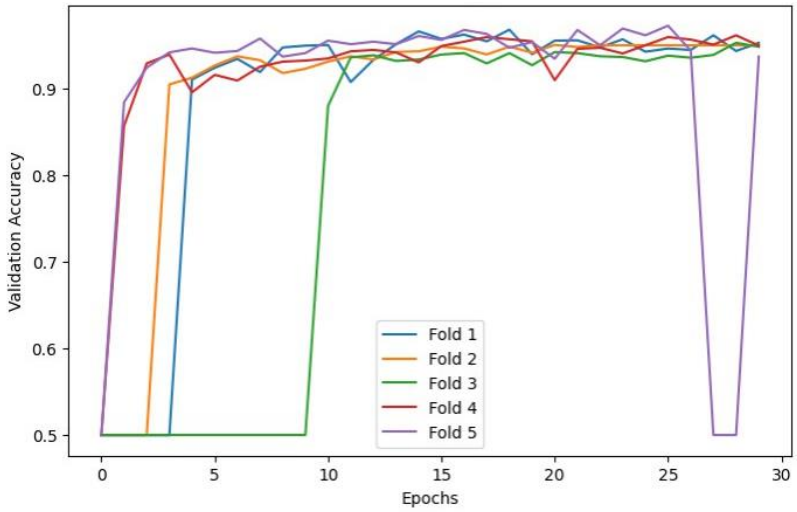

Figure 7. Validation accuracy per epoch across 5 folds.

All folds rapidly converge to above 90% accuracy, demonstrating stable generalization performance.

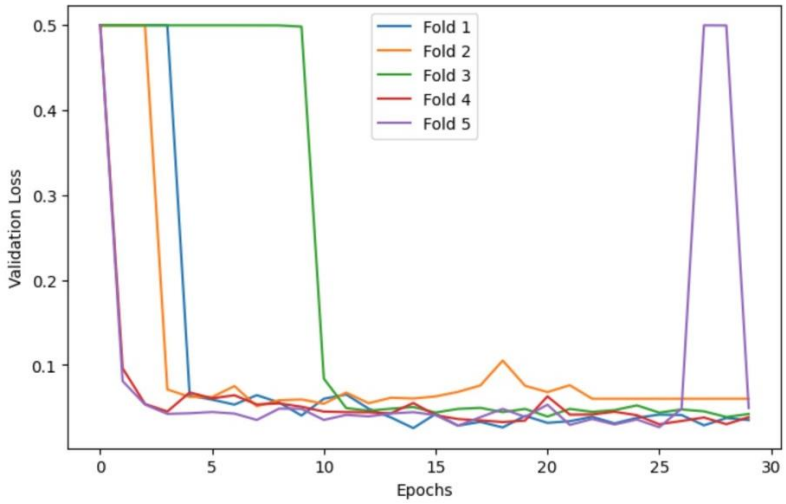

Figure 8. Validation loss per epoch across 5 folds.

Loss drops significantly within the first few epochs for each fold, indicating fast convergence. Minor fluctuations are observed but remain within a low range.

The combination of full-training curves (Figures 5 & 6) and cross-validation results (Figures 7 & 8) confirms that the model is not only well-trained on the complete dataset but also generalizes effectively across unseen splits.

Performance Using Fixed Thresholding

To evaluate the Siamese Neural Network's ability to differentiate compound pairs, we first tested the trained model on the unseen test set using a fixed threshold of 0.5. This threshold was applied to the Euclidean distances between embeddings to determine whether a given pair belonged to the *same compound* or not. The model correctly classified 2399 out of 2430 test pairs, resulting in a test accuracy of 98.72%.

To gain further insight into this result, we examined the distribution of distance scores for positive (same compound) and negative (different compound) test pairs (Figure 9). For visualization clarity, distances greater than 1 were clipped, as most scores ranged between 1 and 5. A vertical line at 0.5 represents the initial decision boundary used in this test.

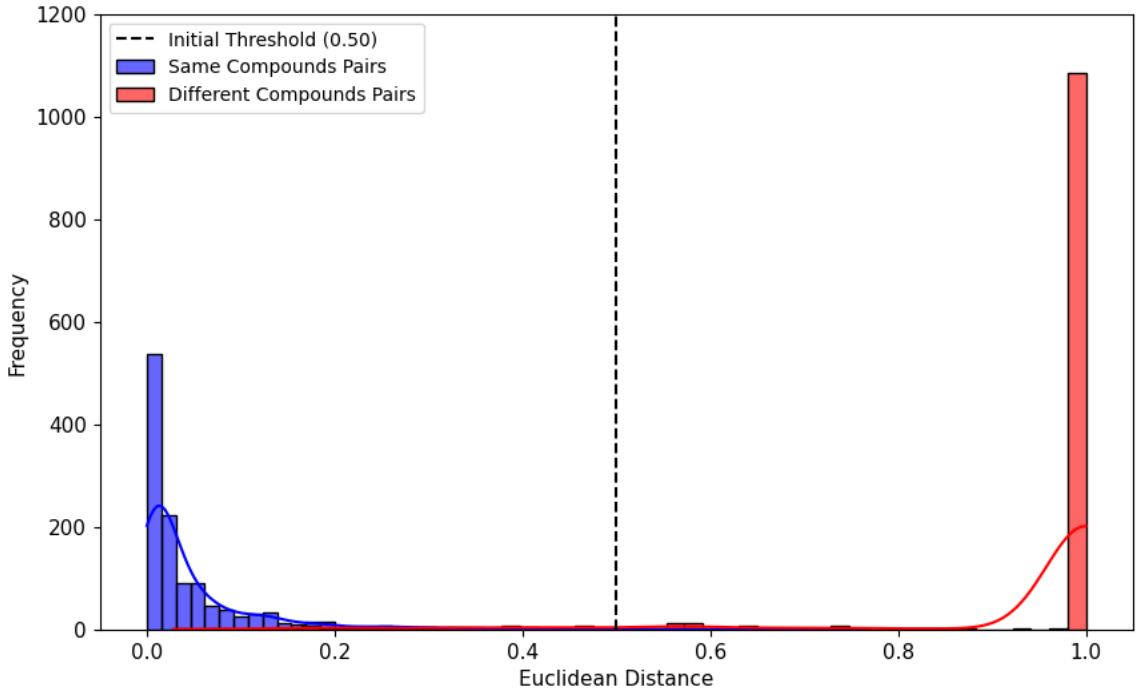

Figure 9. Distance score distributions for positive and negative pairs.

While the selected threshold demonstrated strong performance, the distribution of scores indicated that alternative thresholds might offer improved class separation. Given the observed overlap and variability in distance values, several candidate thresholds (0.35, 0.40, 0.45, and 0.50) were subsequently assessed using F1 score and accuracy as evaluation metrics.

Table 3. Classification Accuracy and F1 Score for Different Distance Thresholds

Threshold	F1 Score	Accuracy
0.35	0.984	97.77%
0.4	0.9821	98.23%
0.45	0.9825	98.23%
0.5	0.9805	98.19%

Threshold 0.45 achieved the highest F1 score while preserving the same accuracy as 0.40. In contrast, a lower threshold like 0.35, although it maintained a good F1, produced a lower accuracy, indicating more false negatives due to overly conservative classification.

To visualize threshold performance more broadly, we plotted the ROC curve across the range of values (Figure 10). The model showed excellent discriminative ability, with a curve that closely approached the top-left corner, and a high area under the curve (AUC), confirming strong sensitivity and specificity. The confusion matrix and corresponding classification metrics for the optimal threshold of 0.45 are summarized in the following analysis.

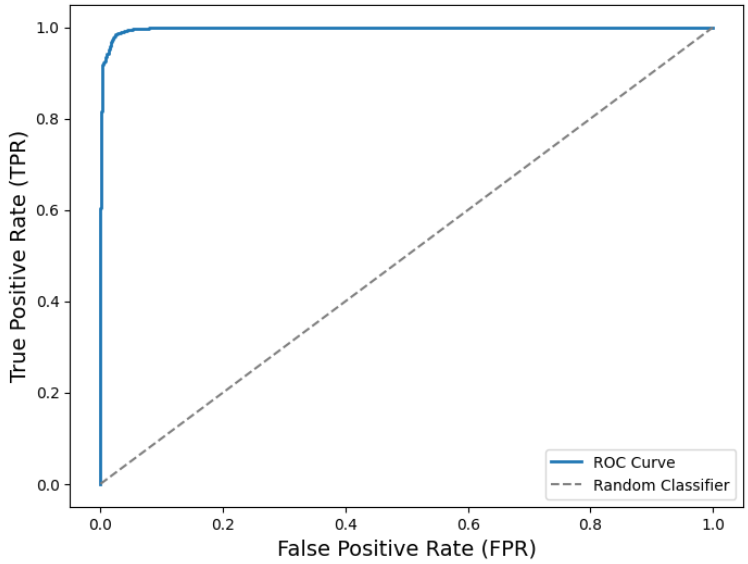

Figure 10. ROC Curve with evaluated thresholds.

Table 4. Confusion Matrix and Classification Metrics for Calibrated Threshold (0.45)

	Predicted: Same	Predicted: Different
Actual: Same	1178	37
Actual: Different	6	1209

Table 5.Classification Report

Class	Precision	Recall	F1-Score
Class 0	0.9949	0.9695	0.9821
Class 1	0.9703	0.9951	0.9825
Accuracy	0.9823		

These results confirm that calibrating the threshold using the distance distribution can slightly improve classification by balancing precision and recall, particularly in edge cases near the decision boundary.

Small Neural Network Classifier for Pairwise Prediction

To further improve classification, we trained a shallow neural network to predict the similarity label based on the embedding distance. Unlike the fixed threshold approach, this model learns a non-linear decision boundary and internally determines an optimal probability threshold (in this case, 0.8149) for class separation. The classifier achieved excellent performance, with an F1 score of 0.9902 and an overall accuracy of 99.01%.

Table 6. Performance Metrics for Neural Network Classifier Based on Embedding Distance

	Predicted: Same	Predicted: Different
Actual: Same	1198	17
Actual: Different	7	1208

Table 7. Classification Report

Class	Precision	Recall	F1-Score
Class 0	0.9942	0.986	0.9901
Class 1	0.9861	0.9942	0.9902

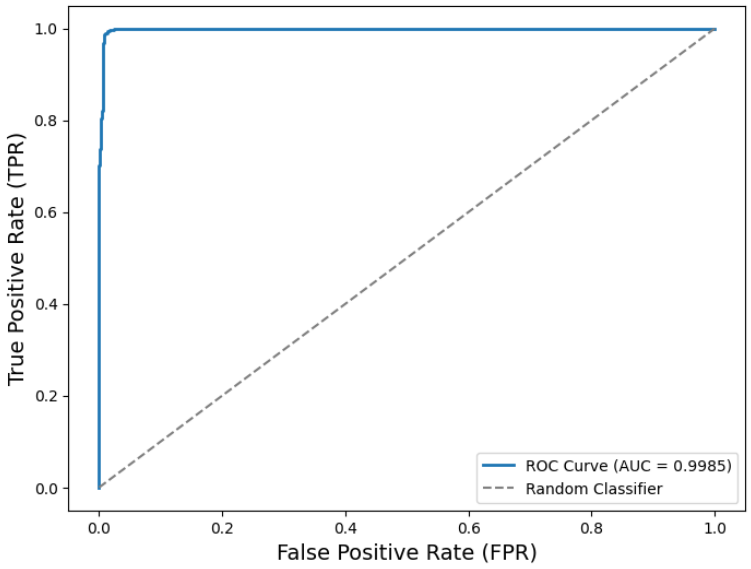

Figure 11. Final ROC Curve

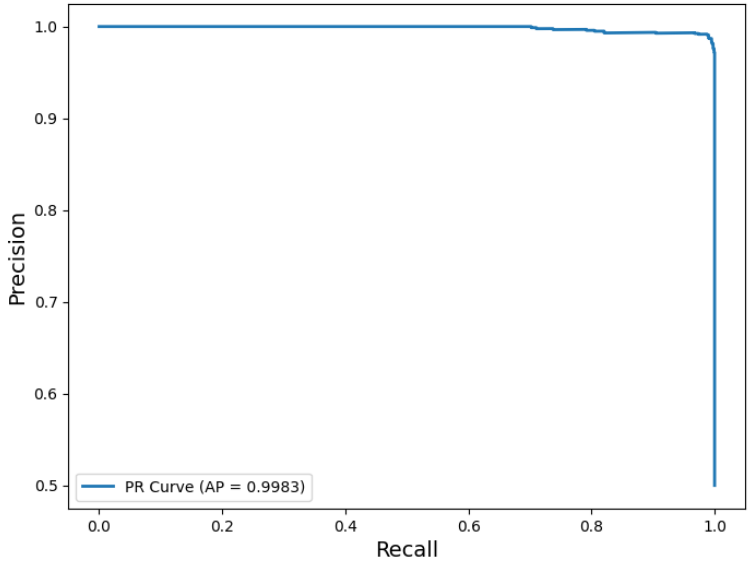

Figure 12. Precision-Recall Curve for small neural network

These results show that the shallow classifier, even with minimal architecture, outperforms fixed-threshold methods by adapting to data distribution. Its use of a learned probability threshold makes it especially effective in borderline cases and demonstrates the strength of using learned decision functions over heuristic ones.

Compound Family Classification

To provide additional insight into the structure of the learned embeddings, we trained a classifier to predict compound families based on the embeddings generated by the Siamese network. Although not part of the main objective, this step serves as a complementary evaluation to assess whether the embeddings capture chemically meaningful groupings.

Among the classifiers evaluated, the Random Forest model demonstrated the strongest performance, achieving perfect classification with 100% accuracy and an F1 score of 1.00 across all classes.

Complete agreement between the true and predicted labels for all three compound families is evident in the confusion matrix (Figure 13), where each class (CM1, CM2, CM3) was correctly classified without error. This outcome suggests that the learned embeddings effectively capture distinct and separable information at the family level.

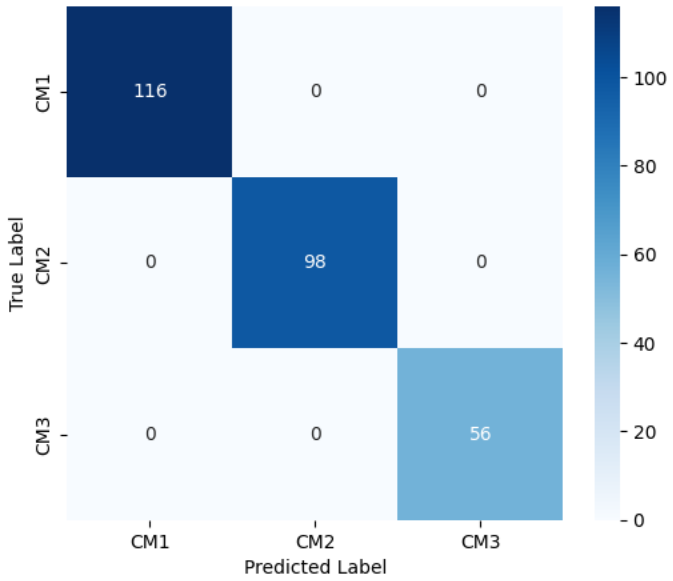

Figure 13. Confusion matrix for compound family classification

This outcome confirms that the embedding space not only supports accurate pairwise differentiation but also reflects higher-level chemical structure, such as family groupings. This reinforces the utility of the Siamese network in capturing both fine-grained and coarse-grained spectral relationships, enhancing interpretability and validating the quality of the learned representations.

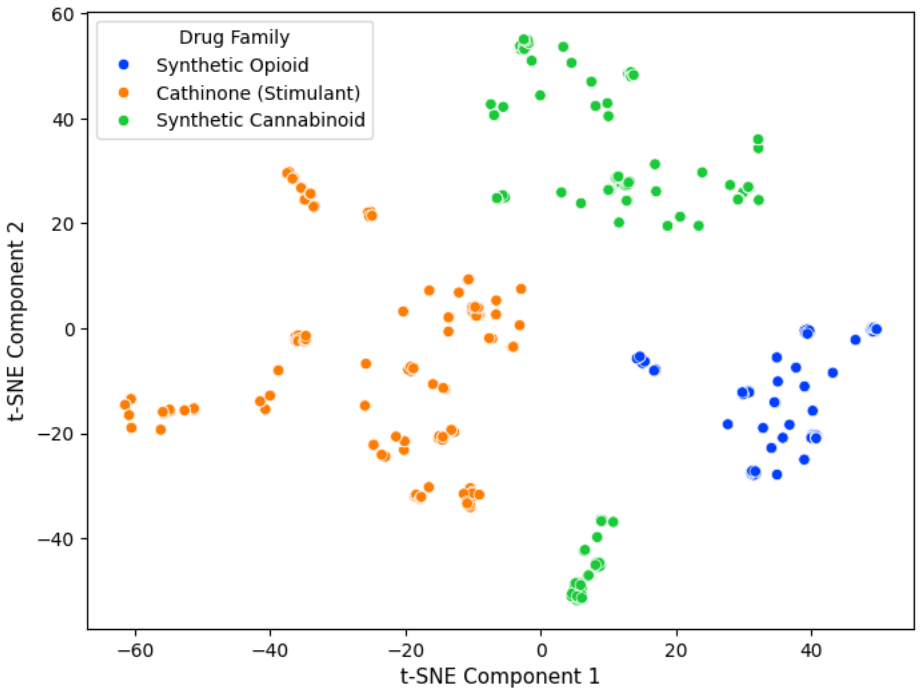

Figure 14. t-SNE visualization of compound families

A t-SNE visualization based on the raw spectral data is shown in Figure 14, illustrating that some degree of clustering by compound family is already present prior to neural network processing. This pattern implies the existence of intrinsic differences among the spectra, even before the application of learned feature representations.

Embedding Space Visualization

To better understand the structure of the learned representations, we applied three dimensionality reduction techniques—Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP)—to visualize the high-dimensional embeddings in two dimensions.

Among the three, PCA (Figure 15) provided the clearest and most structured separation between compound families. In contrast, t-SNE (Figure 16) and UMAP (Figure 17) exhibited more overlap in 2D, though this does not imply that the model failed to differentiate between compounds. These techniques project high-dimensional relationships into two dimensions and may not preserve all aspects of class separability in the original space. The model's strong pairwise classification performance (99.01% accuracy) demonstrates that the embeddings are indeed highly discriminative, even when not fully visible in 2D.

Furthermore, the compound family classification task (100% accuracy) reinforces the validity of the learned embeddings. While not the central objective, this task provides valuable context: if spectra embeddings are highly effective for distinguishing individual compounds, they should also reveal family-level structure—which is exactly what we

observed. The success of the family classifier serves as an indirect but meaningful confirmation that the Siamese network is capturing chemically relevant features.

These visualizations, taken together with the quantitative results, confirm that the model's learned embedding space effectively encodes the spectral differences necessary for both compound identification and higher-level chemical categorization.

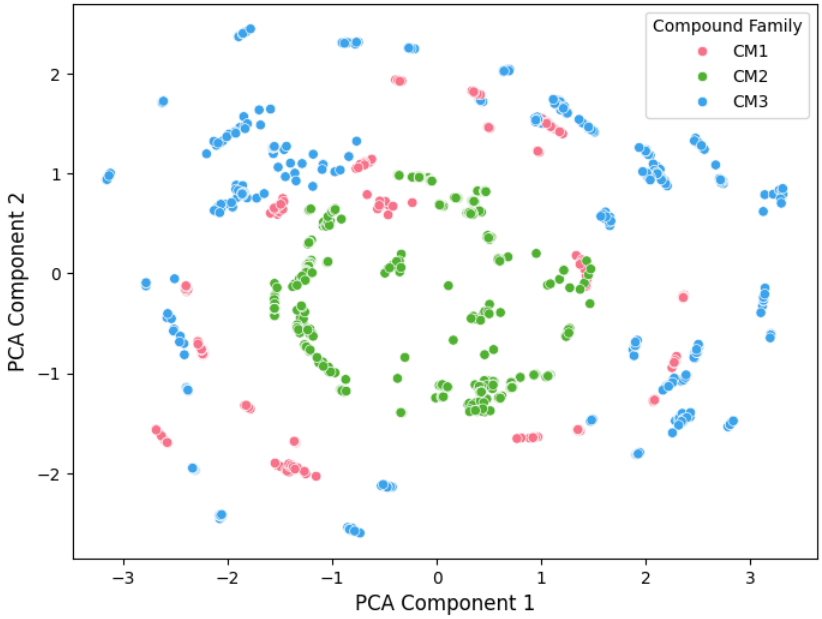

Figure 15. PCA projection (Family Map – PCA)

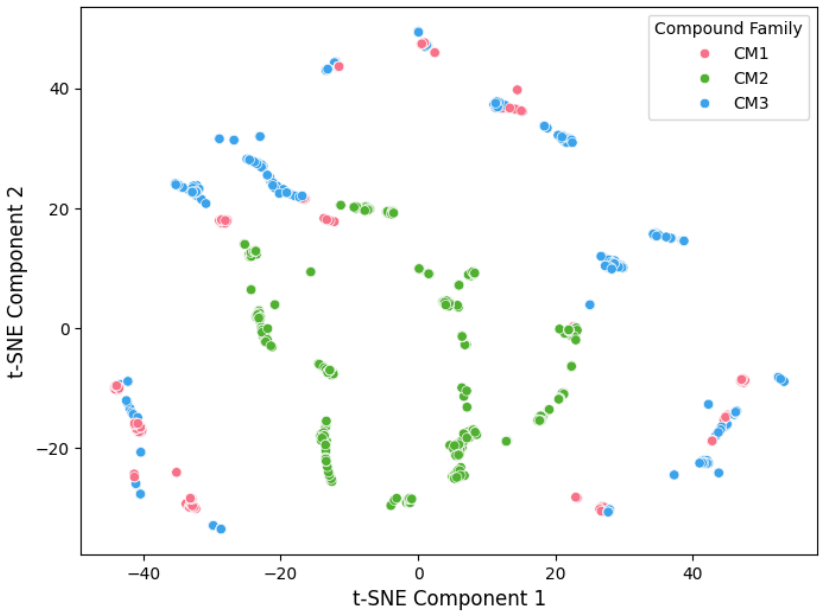

Figure 16. t-SNE projection (Family Map – t-SNE)

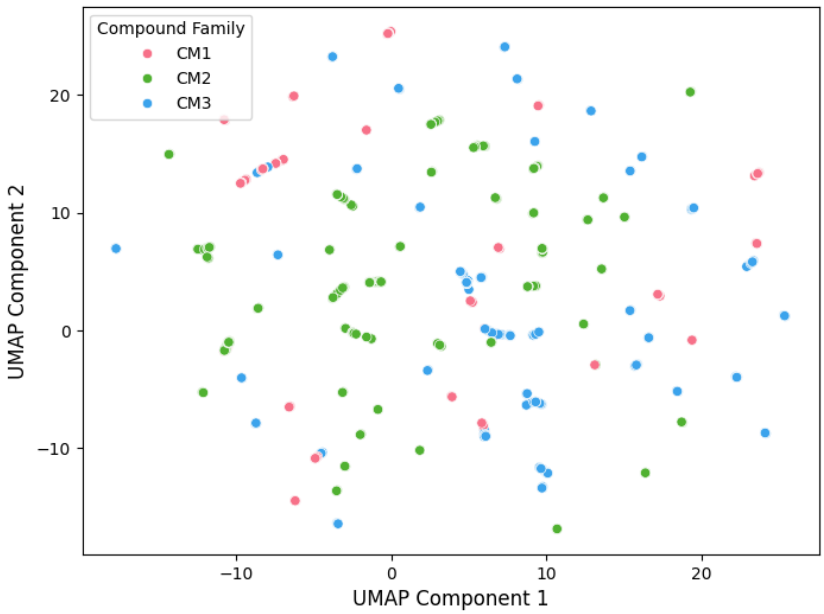

Figure 17. UMAP projection (Family Map – UMAP)

Discussion

The findings of this study validate the effectiveness of Siamese Neural Networks (SNNs) in forensic mass spectrometry and align closely with previous research efforts in spectral similarity learning. The high accuracy achieved in both pairwise comparison (99.01%) and compound family classification (100%) underscore the robustness and generalizability of the proposed approach. These results are consistent with Bonetti, who demonstrated that similarity-based methods outperform traditional classifiers in distinguishing structurally similar compounds such as positional isomers. Our study extends this finding by applying SNNs to a broader range of compounds—including synthetic opioids, cathinones, and cannabinoids—and demonstrating performance improvements in both binary and multi-class settings [10].

The Min-Max Test introduced by Moorthy & Sisco was a key reference for our dataset and inspired the need for objective, reproducible metrics in mass spectrum comparison. While their method provided a statistically sound thresholding mechanism, our SNN-based pipeline offers a more flexible and scalable alternative by learning embeddings and decision boundaries directly from the data. This allows the system to adapt to new substances without requiring manual tuning or static thresholds [7].

Moreover, our use of rational transformation and max normalization to preprocess spectra enhances the visibility of low-intensity yet chemically informative peaks—an approach that aligns with preprocessing strategies employed by Barea-Sepúlveda et al. in their work on spectral classification of gasoline samples. Like their findings, our results confirm that

carefully designed preprocessing steps can significantly influence model performance, especially when dealing with noisy or imbalanced mass spectrometry data [11].

Our findings also corroborate the results of Streun et al. and Beck et al., who advocated for the use of deep learning architectures, including CNNs and ANNs, for handling high-resolution mass spectrometry (HRMS) data. By integrating CNNs and LSTM layers in our Siamese network, we capture both local spectral features and sequential patterns along the m/z axis—mirroring their insight that hybrid models provide superior interpretability and classification power in large-scale spectral datasets [12], [15].

The perfect separation of compound families in our t-SNE and UMAP visualizations further supports the interpretability and structure of the learned embedding space. This is in line with Geurts et al., who found that ensemble and deep learning models can uncover meaningful feature spaces that correlate well with chemical structure and biological function [13].

However, as also noted by Mehnert et al., instrument variability and emerging novel compounds continue to challenge forensic workflows. While our model handles replicate variability well, additional work is needed to evaluate performance across instruments with different ionization techniques or in real-world forensic scenarios involving degraded or mixed samples [8].

In summary, our study builds upon and extends the current body of literature by offering a highly accurate, scalable, and interpretable SNN-based system for forensic mass spectrometry. It confirms key insights from prior work while introducing innovations in data preprocessing, balanced pair generation, and threshold learning. This positions SNNs as a powerful alternative to traditional rule-based and library-matching approaches, particularly in forensic settings where novel and complex substances are increasingly prevalent.

Recommendations

Based on the outcomes of this study, several recommendations can be made for both future research and practical application in forensic mass spectrometry:

Integration into Forensic Workflows: The Siamese Neural Network (SNN)-based similarity framework should be explored further for integration into existing forensic mass spectrometry systems. Given its adaptability and high accuracy, this approach can enhance routine drug screening, especially when dealing with novel or unregistered compounds.

Expand Dataset Diversity: Future work should focus on incorporating data from a wider variety of compounds, instruments, and ionization techniques. This would test the model's robustness in real-world conditions, including different levels of noise, resolution, and sample complexity.

Cross-Instrument Generalization: To ensure broad applicability, models should be evaluated on spectra generated from different instruments. This aligns with challenges highlighted in previous studies and would improve the generalizability of the system.

Real-Time Implementation and User Interface: Developing a lightweight, interpretable front-end system for forensic analysts—powered by pre-trained Siamese models—would improve accessibility and real-time decision-making in lab environments.

Hybrid Approaches with Traditional Methods: While SNNs are powerful, combining their predictions with traditional spectral matching (e.g., NIST library comparisons) may further boost reliability, especially in legal or regulatory contexts where explainability is essential.

Ongoing Learning and Model Updates: As new psychoactive substances and synthetic analogs continue to emerge, the model should be updated using online learning or incremental training to remain effective without full retraining.

Conclusion

This study successfully demonstrates the effectiveness of a Siamese Neural Network architecture for drug identification using mass spectrometry data. By focusing on pairwise spectral similarity rather than fixed-class classification, the proposed system overcomes key limitations of traditional methods, including dependency on complete reference libraries, poor generalization to novel compounds, and difficulty in handling structurally similar drugs.

The results show outstanding performance in pairwise classification (99.01% accuracy), supported by clear visualization of the learned embeddings and balanced evaluation across various metrics. Additionally, compound family classification achieved 100% accuracy, providing valuable context for interpreting the embedding space and further supporting the chemical relevance of the learned representations.

The use of rational transformation and max normalization significantly improved spectral quality, while the modular pipeline design ensured flexibility for multiple forensic tasks. In relation to existing research, this study confirms the advantages of similarity learning over conventional machine learning classifiers in spectral analysis, particularly in forensic science applications. It advances the field by providing a scalable, interpretable, and accurate system that can handle emerging challenges in drug identification.

Ultimately, this work highlights the potential of integrating deep learning-based similarity frameworks into forensic workflows, paving the way for more reliable, efficient, and automated drug identification systems in high-stakes environments.

References

[1] P. D. Maskell and G. Jackson, “Presumptive drug testing—The importance of considering prior probabilities,” *WIREs Forensic Sci.*, vol. 2, no. 5, e1371, 2020, doi: 10.1002/wfs2.1371.

[2] W. Niessen, Liquid Chromatography–Mass Spectrometry, 3rd ed. Boca Raton, FL: CRC Press, 2006.

[3] National Institute of Standards and Technology (NIST), “NIST Mass Spectral Library,” [Online]. Available: https://chemdata.nist.gov

[4] R. G. Cooks et al., "Ambient mass spectrometry," Science, vol. 311, no. 5767, pp. 1566–1570, 2006.

[5] O. Fiehn, "Metabolomics by Gas Chromatography–Mass Spectrometry: Combined Targeted and Untargeted Profiling," Current Protocols in Molecular Biology, vol. 114, no. 1, pp. 30–34, 2016.

[6] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese Neural Networks for One-shot Image Recognition," in Proc. ICML Deep Learning Workshop, 2015.

[7] A. S. Moorthy and E. Sisco, “The Min-Max Test: An Objective Method for Discriminating Mass Spectra,” *Analytical Chemistry*, vol. 93, no. 39, pp. 13319–13325, 2021.

[8] G. L. Streun, A. E. Steuer, L. C. Ebert, A. Dobay, and T. Kraemer, “Interpretable machine learning model to detect chemically adulterated urine samples analyzed by high resolution mass spectrometry,” *Clin. Chem. Lab. Med.*, vol. 59, no. 8, pp. 1392–1399, 2021.

[9] W. E. Wallace and A. S. Moorthy, “NIST Mass Spectrometry Data Center Standard Reference Data and Software Tools for Seized Drug Analysis,” *J. Forensic Sci.*, vol. 68, no. 5, pp. 1484–1493, 2023.

[10] J. N. Wei, D. Belanger, R. P. Adams, and D. Sculley, “Rapid prediction of electron-ionization mass spectrometry using neural networks,” *arXiv preprint*, arXiv:1811.08545, 2019.

[11] M. Barea-Sepúlveda, M. Ferreiro-González, J. L. Pérez-Calle, and M. Palma, “Comparison of different processing approaches by SVM and RF on HS-MS eNose and NIR spectrometry data for the discrimination of gasoline samples,” *Microchem. J.*, vol. 175, p. 107149, Jan. 2022.

[12] A. G. Beck, M. Muhoberac, C. E. Randolph, C. H. Beveridge, P. R. Wijewardhane, H. I. Kenttämaa, and G. Chopra, “Recent developments in machine learning for mass spectrometry,” *ACS Meas. Sci. Au*, vol. 4, no. 3, pp. 233–246, Feb. 2024, doi: 10.1021/acsmeasuresciau.3c00060.

[13] P. Geurts, M. Fillet, D. de Seny, M. A. Meuwis, M. G. Malaise, M. P. Merville, and L. Wehenkel, "Proteomic mass spectra classification using decision tree-based ensemble methods," *Bioinformatics*, vol. 21, no. 14, pp. 3138–3145, 2005.

[14] G. Wang, H. Ruser, J. Schade, J. Passig, T. Adam, G. Dollinger, and R. Zimmermann, "Machine learning approaches for automatic classification of single-particle mass spectrometry data," *Atmospheric Measurement Techniques*, vol. 17, pp. 299–313, 2024.

[15] G. L. Streun, M. P. Elmiger, A. Dobay, L. Ebert, and T. Kraemer, "A machine learning approach for handling big data produced by high resolution mass spectrometry after data independent acquisition of small molecules – Proof of concept study using an artificial neural network for sample classification," *Drug Testing and Analysis*, vol. 12, no. 6, pp. 836–845, 2020.

[16] P. Loahavilai, S. Datta, and T. Limpanuparb, "Chemometric analysis of a ternary mixture of caffeine, quinic acid, and nicotinic acid by terahertz spectroscopy," *ACS Omega*, vol. 7, no. 40, pp. 35783–35791, 2022.

[17] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, and P. Veltri, “A Time Series Based Approach for Classifying Mass Spectrometry Data,” in *Proc. 20th IEEE Int. Symp. Computer-Based Medical Systems (CBMS)*, Maribor, Slovenia, Jun. 2007, pp. 403–408.

[18] C. Krier, D. François, F. Rossi, and M. Verleysen, “Feature Scoring by Mutual Information for Classification of Mass Spectra,” in *Proc. 7th Int. FLINS Conf. Applied Artificial Intelligence*, Genova, Italy, Aug. 2006, pp. 557–564.

[19] L. J. Mauer, A. A. Chernyshova, A. Hiatt, A. Deering, and R. Davis, “Melamine detection in infant formula powder using near- and mid-infrared spectroscopy,” *J. Agric. Food Chem.*, vol. 57, no. 10, pp. 3974–3980, May 2009.

[20] S. A. Mehnert, J. T. Davidson, A. Adeoye, B. D. Lowe, E. A. Ruiz, J. R. King, and G. P. Jackson, “Expert algorithm for substance identification using mass spectrometry: Application to the identification of cocaine on different instruments using binary classification models,” *J. Am. Soc. Mass Spectrom.*, vol. 34, no. 7, pp. 1235–1247, Jul. 2023.

[21] J. M. Colby, “Optimization and Validation of High-Resolution Mass Spectrometry Data Analysis Parameters,” *J. Anal. Chem.*, vol. 89, no. 4, pp. 2117–2124, 2017.