Briefings in Bioinformatics, 2026, 27, bbag054
https://doi.org/10.1093/bib/bbag054
Published: 13 February 2026
Case Study

OXFORD

ChemEmbed: a deep learning framework for metabolite identification using enhanced MS/MS data and multidimensional molecular embeddings

Muhammad Faizan-Khan[1], Roger Giné[1,2], Josep M. Badia[1], Maribel Pérez-Ribera[3], Manuel Ruiz-Botella[3], Alexandra Junza[1,2], Jordi Capellades[4], Iván Pérez-López[1], Shipei Xing[5], Abubaker Patan[5], Laura Brugnara[2,6], Anna Novials[2,6], Joan-Marc Servitja[2,6], Maria Vinaixa[1,2,4], Pieter C. Dorrestein[5,7,8,9], Marta Sales-Pardo[3], Roger Guimerà[3,10], Oscar Yanes[1,2,4,*]

[1]Department of Electronic Engineering, Universitat Rovira i Virgili, Tarragona 43007, Spain
[2]CIBER de Diabetes y Enfermedades Metabólicas Asociadas (CIBERDEM), Instituto de Salud Carlos III, Madrid 28029, Spain
[3]Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona 43007, Spain
[4]Metabolomics Platform, Institut de Recerca Biomèdica Catalunya Sud, Hospital Universitari Sant Joan de Reus, Reus 43204, Spain
[5]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA 92093, United States
[6]Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona 08036, Spain
[7]Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA 92093, United States
[8]Department of Pharmacology, University of California San Diego, La Jolla, CA 92093, United States
[9]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA 92093, United States
[10]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain

[*]Corresponding author. Department of Electronic Engineering, Universitat Rovira i Virgili, Tarragona 43007, Spain. E-mail: oscar.yanes@urv.cat

Abstract

Machine learning offers a promising path to annotating the large number of unidentified MS/MS spectra in metabolomics, addressing the limited coverage of current reference spectral libraries. However, existing methods often struggle with the high dimensionality and sparsity of MS/MS spectra and metabolite structures. ChemEmbed tackles these challenges by integrating multidimensional, continuous vector representations of chemical structures with enhanced MS/MS spectra. This enhancement is achieved by merging spectra across multiple collision energies and incorporating calculated neutral losses from 38 472 distinct compounds, providing richer input for a convolutional neural network (CNN). ChemEmbed ranks the correct candidate first in over 42% of cases and within the top five in more than 76% of cases. In external benchmarks such as CASMI 2016 and 2022, ChemEmbed outperforms SIRIUS 6, the current state-of-the-art in computational metabolomics. We applied ChemEmbed to predict structures in the Annotated Recurrent Unidentified Spectra (ARUS) dataset and confirmed 25 previously unidentified compounds. These findings demonstrate ChemEmbed's potential as a robust, scalable tool for accelerating metabolite identification in untargeted mass spectrometry workflows.

Keywords metabolite identification; untargeted metabolomics; mass spectrometry; deep learning; molecular embeddings

Introduction

The interpretation of tandem mass spectrometry (MS/MS) data is crucial for the identification of metabolites, which in turn is particularly relevant for the growing applications of metabolomics in biomedicine, nutritional and environmental sciences [1].

The predominant strategy for metabolite identification involves comparing experimental spectra with pre-recorded MS/MS spectra of known compounds to find matching fragments. However, due to the limited size, quality, and diversity of available reference spectral libraries, a significant portion of MS/MS spectra generated in metabolomic experiments remain unidentified [2].

In silico fragmentation tools have emerged as a solution to address this challenge. These tools combine chemical rules to capture known fragmentation events [3], probabilistic models to assign probabilities to potential fragmentations [3, 4], and/or machine learning

Received: June 4, 2025. **Revised:** November 7, 2025. **Accepted:** January 6, 2026

algorithms to learn the complex relationships between molecular structures and their corresponding fragmentation patterns by MS [5–7].

To enable efficient computation and learning by machine learning algorithms, both experimental MS/MS spectra and the chemical structures of metabolites must be converted into numerical formats that preserve key information. MS/MS spectra, typically represented as lists of m/z values and intensities, are encoded into fixed-length vectors using techniques such as feature hashing [8, 9], binning [7, 10–12], chemical formula representations [5], and spectral fingerprinting [7]. Chemical structures are typically encoded as molecular fingerprints [13–15], though graph-based representations are also used [16, 17].

However, encoding techniques are severely affected by the sparsity and high dimensionality of metabolomics data, which can lead to many issues including model interpretation, complexity and overfitting. For example, molecular fingerprints are binary or count vectors that represent the presence, absence, or the number of specific chemical features within a molecule, such as atom types, functional groups, or structural motifs. While the dimensionality of these fingerprints is fixed by the number of features they are designed to capture, the vast chemical diversity of metabolites leads to inherent sparsity—most features remain inactive or set to zero for most compounds.

Encoded m/z values and intensities from MS/MS spectra face a similar sparsity issue when represented as fixed-length vectors. Given that spectra only contain a few peaks within a broad m/z range, most bins are left empty, with only a small number of non-zero bins corresponding to specific fragment ions.

To overcome these limitations, we propose ChemEmbed, a new tool for metabolite identification from MS/MS spectra. Our tool is based on two complementary strategies designed to enrich the representations of both chemical structures and MS/MS spectra. First, we employ 300-dimensional embeddings generated by Mol2vec [18], an unsupervised machine learning approach that captures complex structural properties of molecules with greater depth than traditional molecular fingerprints. Second, to address the sparsity in spectral data, we use merged spectra that combine multiple collision energies and calculated neutral losses from reference libraries. This expanded spectral representation broadens the range and diversity of binned spectra, providing a richer input for a convolutional neural network (CNN). The CNN is trained to predict 300-dimensional Mol2vec embeddings from these enhanced spectra, and these embeddings are then compared and ranked against a reference database of millions of Mol2vec embeddings.

We consider a dataset of 38 472 distinct compounds with reference MS/MS data, which we split into training, validation, and test sets. In the test set, ChemEmbed accurately annotated the top-ranked candidate (Tanimoto $\geq$0.95) in over 42% of cases and identified the correct candidate within the top five in more than 76% of cases. When challenged with external datasets such as CASMI 2016 and 2022, ChemEmbed consistently outperformed the latest version of SIRIUS [7], the current state-of-the-art method in computational metabolomics. Furthermore, in a validation experiment using the Annotated Recurrent Unidentified Spectra (ARUS) from the NIST Mass Spectrometry Data Center, our tool successfully identified 25 previously unannotated compounds, demonstrating its broad applicability and superior performance across diverse datasets.

Methods

Dataset and data preprocessing

For our spectra-to-molecule CNN model, we used a dataset of 38 472 unique compounds sourced from the NIST20, MSDIAL, GNPS, and Agilent METLIN databases. Initially, the NIST20 dataset provided only InChIKey information, so we leveraged the PubChem API to retrieve the corresponding SMILES strings. The Agilent METLIN dataset also contained missing SMILES and InChIKey data, which we supplemented using the RDKit tool to convert available SMILES to molecular information, followed by PubChem API queries to obtain the remaining SMILES for specific InChIKeys.

After gathering the SMILES information for all datasets, we removed any spectra with null values. Since stereochemistry was not considered in this work, we used only the first 14 characters of the unique InChIKey identifier.

Spectral data preprocessing

To prepare the MS/MS spectra for model training, several preprocessing steps were implemented. Only $[M+H]^+$ adducts in positive ionization and $[M\text{-}H]^-$ adducts in negative ionization were used. Peaks that exceeded the precursor mass by more than 0.5 Da were removed. To reduce noise, fragments with intensities below 1% of the highest peak were filtered out, and the remaining intensities were binarized. Each MS/MS spectrum was encoded as a vector with a bin size of 0.01. Pareto analysis revealed that 20% of the unique m/z bins accounted for 83% of the spectral values (Supplementary Fig. 8). Consequently, we limited the m/z axis to $\leq$700, reducing the vector length to 70 000 bins. This adjustment eased computational demands while retaining 80% of the spectral data.

Merged and merged+neutral loss spectral vectors

For each compound, we first discretized the *m/z* axis into 0.01 Da bins. Fragment peaks from all available MS/MS spectra (across collision energies) were mapped to these bins, and we then took the union across spectra: when the same fragment appeared at multiple collision energies (i.e. mapped to the same bin), it was counted once. A binary vector was then generated by assigning a value of 1 to bins containing a fragment and 0 otherwise. This merged spectral vector provides a standardized, fixed-length representation of the compound's MS/MS spectra.

For the Merged+Neutral Loss representation, neutral losses were additionally calculated for each fragment as:

$$\Delta \mathrm{m}/z = \mathrm{m}/\mathrm{z}_{\mathrm{precursor}} - \mathrm{m}/\mathrm{z}_{\mathrm{fragment}}.$$

Both the fragment *m/z* values and their corresponding neutral losses were discretized into the same 0.01 Da bins. The two sets were concatenated and converted into a binary vector following the same procedure as above. This combined representation captures both fragment and neutral loss information in a standardized, fixed-length format suitable for downstream machine learning models.

Convolutional neural networks

We employed a CNN to predict 300-dimensional Mol2vec embeddings from input MS/MS spectra. The models were trained and tested using an NVIDIA Tesla T4 GPU, with experiments conducted within the PyTorch framework. For development and implementation, we utilized the PyCharm integrated development environment (IDE).

- Individual spectra dataset: This set included 527 236 spectra (80%) for training, 64 774 spectra (10%) for validation, and 66 843 spectra (10%) for testing. In this configuration, the same compound can appear multiple times, as MS/MS spectra were acquired under different collision energies. Each MS/MS spectrum is treated as an independent input, but all are paired with the same Mol2vec embedding corresponding to the compound's structure. This setup, therefore, produces multiple training examples per compound, each linking a unique spectrum to a shared molecular representation.

- **Merged spectra dataset**: this set was divided into 30 775 spectra (80%) for training, 3847 spectra (10%) for validation, and 3850 spectra (10%) for testing.

Network architecture

Six convolutional layers capture spatial relationships between mass-to-charge (m/z) values across the spectrum. Six max pooling layers reduce the dimensionality of the feature space. The first fully connected layer flattens the spatial data, and the final fully connected layer outputs the 300-dimensional Mol2vec embedding. The input to the network is a vector of 70 000 bins, representing the m/z values of the spectra.

Hyperparameters and training

We used the Adam optimizer with a learning rate of 0.0001 for backpropagation. Early stopping based on validation loss was implemented to prevent overfitting. The merged spectra network was trained for 70 epochs, while the individual spectra network was trained for 15 epochs, both using a mini-batch size of 32. Mean squared error was used as the cost function.

Activation function and regularization

The Rectified Linear Unit (ReLU) was chosen as the activation function due to its effectiveness in mitigating the vanishing gradient problem in deep networks. Dropout regularization was applied to prevent overfitting. No activation function was applied to the final output layer, allowing the predicted embeddings to span any value within the Mol2vec embedding space.

Analysis with SIRIUS 6.0

To validate our tool against SIRIUS+CSI FingerID, we used the graphical interface of SIRIUS v6.0.0 (June 3, 2024). First, we imported a custom reference database of 5.52 million entries into the software using SMILES information to ensure consistency across both tools. Next, we removed chemical structures from CASMI 2022 that were part of the SIRIUS 6.0 training set by downloading the InChI identifiers for positive and negative ion mode from the respective training structure links: https://csi.bright-giant.com/v3.0/api/fingerid/trainingstructures?predictor=1 and https://csi.bright-giant.com/v3.0/api/fingerid/trainingstructures?predictor=2. We then converted these InChI identifiers to SMILES format using RDKit.

We imported the MS/MS spectra from the CASMI 2022 dataset and initiated the analysis by selecting the **'**Compute**'** option, which opened a parameter configuration window. For the SIRIUS tool, we configured the instrument as ORBITRAP, MS2 mass accuracy to 5 ppm, the adduct to $[M+H]^+$ and $[M\text{-}H]^-$ for positive and negative ion mode, respectively, and set the molecular formula generation method to De novo + bottom-up. Following this, we enabled CSI:FingerID for property prediction with default threshold score settings. For the structure database search, we selected our custom database of 5.52 million structures and disabled PubChem as a fallback. Upon running the analysis, SIRIUS returned a list of candidate molecules for each MS/MS spectrum, which we sorted in ascending order by the CSI:FingerID score, a key feature used to assess the quality of candidate molecules.

In the final step, we identified the top five candidates from both tools and compared their performance by evaluating which tool returned a higher number of correct candidates in the top 5 for each MS/MS spectrum. This comparison enabled us to assess the relative accuracy and effectiveness of the two tools in generating valid candidate molecules.

Molecular fingerprints

Molecular fingerprints, commonly used to represent molecular structures, encode predefined molecular features, such as specific substructures, into bit vectors. To replace Mol2vec embeddings, we used all five fingerprint types defined in the SIRIUS CSI framework [19]: (i) CDK Substructure Fingerprints: Capture 307 molecular properties using predefined substructure patterns (Chemistry Development Kit, version 1.5.8); (ii) PubChem (CACTVS) Fingerprints: Represent 881 properties, based on PubChem specifications; (iii) Klekota–Roth Fingerprints: cover 4860 properties, providing detailed structural and functional group information; (iv) FP3 Fingerprints: represent 55 properties derived from SMARTS patterns (Open Babel, version 2.3.2); and (v) MACCS Fingerprints: Encode 166 SMARTS-based properties (Open Babel, version 2.3.2). Fingerprint calculations utilized the Chemistry Development Kit (CDK) version 1.5.8 and PyFingerprint (https://pypi.org/project/pyfingerprint/).

Data preprocessing

The initial combined fingerprint set encompassed 6269 molecular properties. Constant and redundant features were removed from the training dataset, reducing the feature set to 4237 properties. These features were split into training (80%), validation (10%), and testing (10%) subsets, maintaining the same distribution used for Mol2vec-based training.

Model development and training

Initial attempts to train the CNN architecture designed for Mol2vec embeddings with molecular fingerprints failed, as training and validation losses remained constant. To address this, a DNN comprising four fully connected layers with ReLU activation functions was implemented. Dropout layers were added for regularization, and binary cross-entropy served as the loss function to handle the multi-label classification task. During training, outputs were binarized using a threshold of 0.50, assigning a label of 1 to predictions ≥0.50 and 0 otherwise. The DNN demonstrated learning, with steadily decreasing training and validation losses.

Molecular ranking and evaluation

The trained model was evaluated using the same molecular ranking method as for Mol2vec-based predictions. A dataset of 0.52 million molecules was used, substituting the refined molecular fingerprints for Mol2vec embeddings. Model predictions were compared to ground-truth embeddings by computing pairwise distances (cosine similarity and Euclidean distance), and performance was quantified using Top-1 and Top-5 accuracy.

We used Top-k accuracy instead of AUROC, F1-score, sensitivity, or specificity because ChemEmbed formulates metabolite annotation as a regression task in the embedding space rather than a binary or multiclass classification problem. In this setting, Top-k metrics are the most informative, as they quantify how often the correct compound appears among the top-ranked candidates—closely mirroring standard metabolomics practice, where candidate lists are generated and refined with additional orthogonal evidence.

ChemBERTa-2 embeddings

ChemBERTa-2 is a Transformer-based language model for molecules, developed by DeepChem and trained on ∼77 million SMILES strings. To generate embeddings, we used the Hugging Face transformers library with the pre-trained checkpoint DeepChem/ChemBERTa-77 M-MLM. SMILES strings were tokenized with AutoTokenizer and passed through AutoModel to obtain fixed-length representations. Each molecule was encoded as a 384-dimensional embedding vector. These embeddings were then compiled into a reference database comprising 0.52 million molecules with their corresponding ChemBERTa-2 representations (Supplementary File 7). The trained model was evaluated using the same molecular ranking method as for Mol2vec- and molecular fingerprints-based predictions.

Exercise training dataset

Serum samples were collected from 46 sedentary but healthy individuals (35 women and 11 men) before and after a 3-week exercise training program consisting of nine sessions. The participants had a mean body mass index (BMI) of 23.8 ± 3.9 kg/m^2 and a mean age of 33.1 ± 7.0 years (mean ± standard deviation). The dataset includes multiple exercise modalities, namely HIIT, MICT, and SSST. Samples were analyzed using an Acquity UPLC BEH HILIC column (2.1 × 150 mm, 1.7 μm; Waters) coupled to a Thermo Scientific™ Orbitrap IDX Tribrid mass spectrometer equipped with a HESI interface operating in both positive and negative ionization modes. MS and MS/MS data acquisition were performed as previously described in Giné et al. [20], except that the normalized collision energy (HCD cell) was applied in stepped mode at 10, 20, 30, and 40%. Only features detected in more than 80% of samples were retained for statistical analysis. Significant differences were assessed using a paired univariate *t*-test ($P < .05$).

Results

Construction of spectral and structural databases

We generated SMILES identifiers for 38 472 unique compounds sourced from the NIST20 [21], MSDIAL [22], GNPS [23], and Agilent METLIN metabolomics libraries (see Methods) (Supplementary File 1), comprising 683 664 MS/MS spectra in positive ionization ($[M + H]^+$) and 158 382 MS/MS spectra in negative ionization ($[M\text{-}H]^-$), as detailed in Supplementary Tables 1 and 2, respectively. Since many compounds were annotated across different reference libraries, we calculated the overlap by matching the first 14 characters of their InChIKey identifiers (Supplementary Fig. 1). To resolve duplicates, we prioritized the libraries in the following order: NIST20 > Agilent METLIN > GNPS > MSDIAL (Supplementary Table 3 and 4). For example, if a compound was found in both NIST20 and Agilent METLIN, we discarded the MS/MS spectra from Agilent METLIN, and so on.

These spectra were pre-processed and curated to remove background signals and noise, as described in the Methods section. This collection formed what we termed the individual spectra dataset, capturing the diversity and redundancy of compounds fragmented at multiple collision energies.

Next, we created a second collection of MS/MS spectra by merging all available spectra across different collision energies for each compound, producing a single consensus MS/MS spectrum per compound (see Methods). This merging process resulted in 38 472 spectra in positive ionization mode and 14,168 spectra in negative ionization mode. The merged spectra dataset broadened the range and diversity of encoded bins, providing a more comprehensive view of each compound's fragmentation behavior by combining information from various collision energies (Supplementary Fig. 2).

To further enhance spectral representation, we generated a third collection of MS/MS spectra by incorporating neutral losses (NL) into the merged spectra. For both positive and negative ionization modes, we calculated the mass difference between precursor and fragment ions, and then binarized their intensities, rather than retaining the original fragment ion intensities, as done in previous studies [24]. The resulting neutral loss values were added to each merged spectrum (see Methods and Supplementary Fig. 2), resulting in an enriched dataset that combines fragment ions from multiple collision energies with their corresponding neutral losses, all in binarized format.

Mol2vec embeddings reference database

Mol2vec is an unsupervised deep learning model, trained on 19.9 million compounds from the ZINC [25] and ChEMBL [26] databases, which converts molecular structures from their SMILES into 300-dimensional numerical vector embeddings. Since a single InChIKey can produce multiple redundant SMILES, and our goal was to use Mol2vec embeddings as labels for training the CNN model, we needed to confirm that Mol2vec was insensitive to different SMILES representations. To verify this, we randomly selected 100 compounds from our database, generated 10 compatible SMILES for each, and compared the resulting 300-dimensional embeddings. The Euclidean distance between the embeddings was nearly zero (data not shown), confirming that Mol2vec embeddings are consistent across SMILES versions and can serve as reliable labels for CNN model training.

Following this validation, we generated a comprehensive reference database of 0.52 million molecules, each with its corresponding Mol2vec embedding, sourced from COCONUT [27], HMDB [2], NIST20, GNPS, Agilent METLIN, and MSDIAL. These embeddings serve as the 'ground truth' or reference embeddings for subsequent comparisons in the CNN model (Supplementary File 2).

CNN model for molecular embedding prediction from MS/MS data

We developed a CNN (Supplementary Fig. 3) that accepts MS/MS spectra as input and generates 300-dimensional embeddings as output. The CNN's primary objective is to map the input spectra into a molecular embedding space that replicates the Mol2vec representations. For each spectrum in the test dataset, the predicted embeddings were then compared to reference database of 0.52 million molecules, each

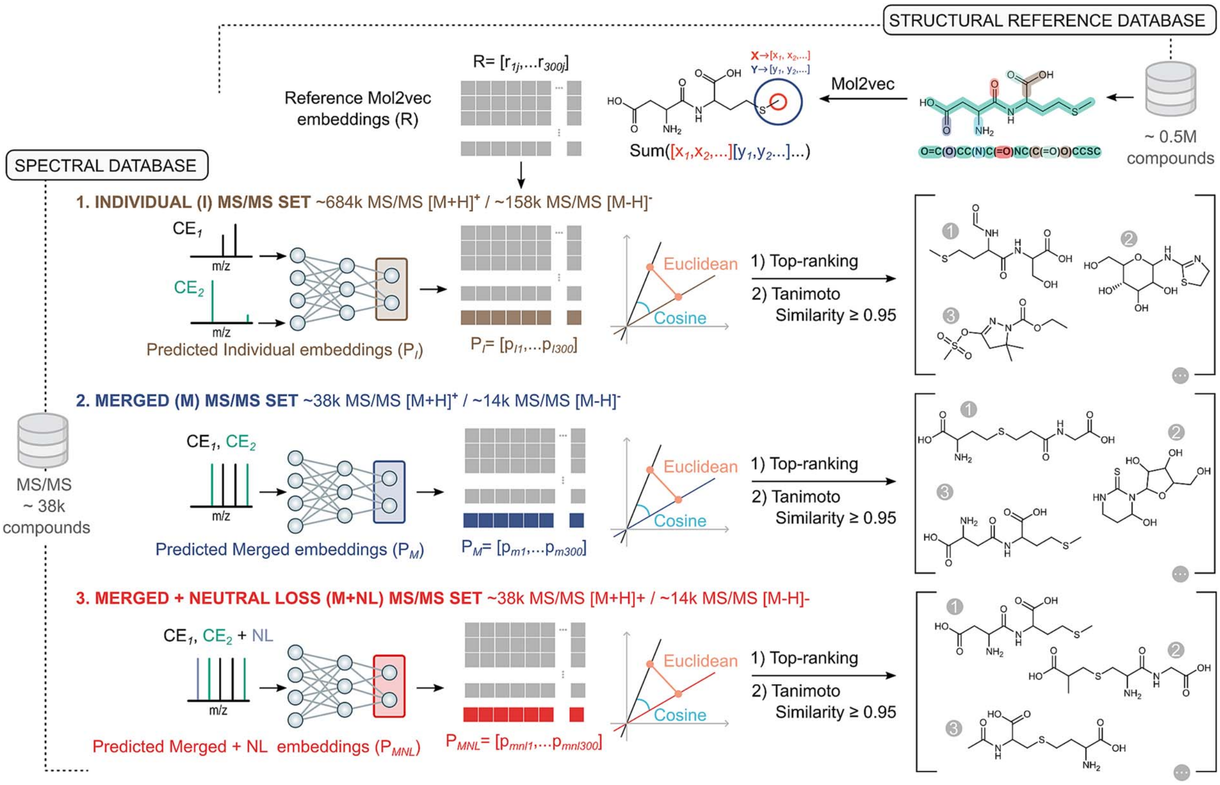
STRUCTURAL REFERENCE DATABASE
~ 0.5M compounds
Mol2vec
O=C(O)CC(N)C(=O)NC(C(=O)O)CCSC
X→[x1, x2,...]
Y→[y1, y2,...]
Sum([x1,x2,...][y1,y2...]...)
R= [r1j,...r300j]
Reference Mol2vec embeddings (R)
SPECTRAL DATABASE
MS/MS ~ 38k compounds
1. INDIVIDUAL (I) MS/MS SET ~684k MS/MS [M+H]+ / ~158k MS/MS [M-H]-
CE1
CE2
m/z
Predicted Individual embeddings (PI)
PI= [pI1,...pI300]
Euclidean
Cosine
1) Top-ranking
2) Tanimoto Similarity ≥ 0.95
2. MERGED (M) MS/MS SET ~38k MS/MS [M+H]+ / ~14k MS/MS [M-H]-
CE1, CE2
m/z
Predicted Merged embeddings (PM)
PM= [pm1,...pm300]
Euclidean
Cosine
1) Top-ranking
2) Tanimoto Similarity ≥ 0.95
3. MERGED + NEUTRAL LOSS (M+NL) MS/MS SET ~38k MS/MS [M+H]+ / ~14k MS/MS [M-H]-
CE1, CE2 + NL
m/z
Predicted Merged + NL embeddings (PMNL)
PMNL= [pmnl1,...pmnl300]
Euclidean
Cosine
1) Top-ranking
2) Tanimoto Similarity ≥ 0.95
1
2
3

Figure 1 Computational workflow for molecular identification using CNN-generated embeddings. MS/MS spectra are the input to a convolutional neural network (CNN) trained to generate 300-dimensional embeddings aligned with Mol2vec representations. Predicted embeddings are compared to a reference database in a multi-step process: (i) precursor mass filtering: Candidate molecules with precursor ion masses outside a ± 0.001 Da error tolerance are excluded; (ii) similarity calculations: Euclidean distance and cosine similarity are computed between the predicted embeddings and reference Mol2vec embeddings, ranking molecules by similarity; (iii) structural validation: The top-ranked candidates are further evaluated using the Tanimoto score, ensuring structural alignment between predicted and reference molecules. A Tanimoto score threshold of ≥0.95 is applied to confirm molecular identity.

represented by its Mol2vec embedding. To ensure accurate matching between the predicted embeddings and known molecules, we employed a multi-step filtering and ranking strategy (Fig. 1).

First, we applied a precursor ion mass filter to exclude any precursor ions outside a mass error tolerance of ±0.001 Da, focusing the comparison on candidate molecules with closely matching mass. Following this, we calculated both the Euclidean distance and cosine similarity between the predicted 300-dimensional vector representation and the 0.52 million reference Mol2vec embeddings. Molecules were then ranked by their similarity, with Euclidean distance providing a measure of closeness between two points in the embedding space, and cosine similarity capturing the directional alignment between the predicted and reference vectors. Together, they provide a more comprehensive assessment of similarity, capturing both spatial proximity and pattern alignment, enhancing the reliability of the molecular ranking process.

Finally, to assess the structural similarity between the predicted and reference molecules, we calculated the Tanimoto score for the top-ranked molecules based on their Euclidean distance and cosine similarity. The Tanimoto score ranges from 0 to 1, where 0 indicates no similarity and 1 indicates perfect similarity. In cheminformatics, thresholds above 0.85 are often regarded as indicative of significant structural similarity [28–30]. Here, we set a threshold of 0.95 for the Tanimoto score to determine molecular identity. This threshold provided an additional layer of validation, ensuring that the predicted embedding not only closely matched the reference embedding but also corresponded to a structurally very similar (e.g. stereoisomers) or identical molecule in terms of their SMILES representations.

Performance of ChemEmbed in a test dataset

The three collections of MS/MS spectra—individual, merged, and merged with neutral losses—were used to train and test three CNN models to predict molecular embeddings, with 80% of the spectral data used for training, 10% for validation, and 10% for testing (see Methods).

Figure 2A displays the distributions of Euclidean distances and cosine similarities between the predicted 300-dimensional embeddings and the reference Mol2vec embeddings for the individual spectra dataset in positive ionization. The model trained on this dataset showed a mean Euclidean distance of 40 and a mean cosine similarity of 0.93. In comparison, the model trained on the merged spectra dataset without neutral losses exhibited a significantly lower mean Euclidean distance of 24.5 and a higher mean cosine similarity of 0.96. Incorporating neutral losses provided only a slight improvement, reducing the Euclidean distance to 23.1 and increasing the cosine similarity to 0.97. These findings demonstrate that as the richness of spectral information increases—by merging spectra and including neutral losses—the CNN's predicted embeddings more closely align

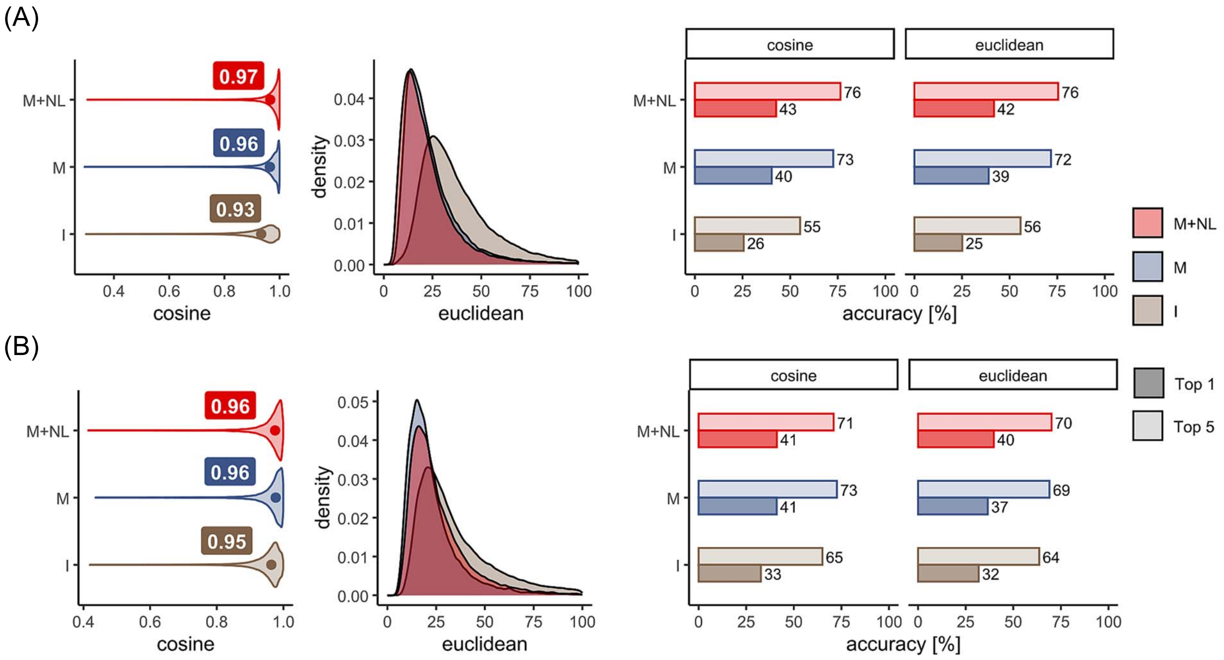

Figure 2 Alignment of CNN-predicted embeddings with reference Mol2vec embeddings. Violin and density plots illustrate the distributions of cosine similarities and Euclidean distances between the CNN-predicted 300-dimensional embeddings and the reference Mol2vec embeddings. Results are shown for positive ionization (A) and negative ionization (B) datasets. Three model configurations are compared: Trained with individual spectra (I), trained with merged spectra (M), and trained with merged spectra incorporating neutral losses (M + NL). Accuracy metrics are represented as bars, indicating the percentage of correctly identified molecules ranked in the top 1 (dark color) and top 5 (light color) positions for both ionization modes. The plots highlight the performance improvements associated with merging spectra and incorporating neutral losses.

with the ground truth Mol2vec embeddings. Similar improvements were observed in the negative ionization mode (Fig. 2B).

This improvement in embedding accuracy directly impacted the model's ability to rank correct compound annotations. When trained on the individual spectra dataset, the CNN correctly annotated the top-ranked candidate (Tanimoto score ≥ 0.95) in over 26% of cases and identified the correct compound within the top five ranked candidates in over 56% of cases. In contrast, the model trained on the merged spectra dataset significantly improved annotation performance, with the correct top-ranked candidate identified in over 40% of cases and the correct compound found within the top five in over 73% of cases. The inclusion of neutral losses further enhanced CNN's performance, with correct top-ranked candidates identified in over 43% of cases and correct annotations within the top five in over 76% of cases.

Performance in the negative ionization mode showed consistent and reproducible behavior, with only a slightly slower ranking performance compared to positive ionization (Fig. 2B). We attribute this difference to the typically lower number of fragment ions in negative ionization, which may provide less informative data for the CNN to generate accurate molecular embeddings.

As expected, neither the combination of individual mass spectra with neutral losses (59.8% for top 5 annotations) nor neutral losses alone (69% for top 5 annotations, calculated from merged spectra) outperformed the full combination of merged spectra with neutral losses (Supplementary Table 5).

An additional advantage of using merged spectra is that the processing time for the test dataset, which included 3850 compounds, was substantially reduced with respect to considering individual spectra. On a basic laptop CPU (Mac 2020, M1 Core, 16 Gb RAM), the full pipeline—covering MS/MS preprocessing, CNN inference, and candidate ranking—took less than 20 min for the merged spectra with neutral losses. In contrast, processing the individual spectra dataset required nearly 6 h, highlighting the efficiency and scalability of the merged spectra approach.

To further demonstrate the value of combining enhanced MS/MS data with multidimensional molecular embeddings, we evaluated two alternative approaches for encoding chemical structures: molecular fingerprints and transformer-based embeddings. Specifically, we tested (i) molecular fingerprints capturing 4237 chemical properties, similar to those used in SIRIUS-CSI:FingerID [19], and (ii) ChemBERTa2 embeddings [31], 384-dimensional vectors derived from SMILES using a pre-trained transformer model (see Methods).

We first retrained our CNN architecture using either fingerprints or ChemBERTa2 embeddings, keeping all architectural components and hyperparameters identical to the Mol2Vec-based setup to ensure fair comparisons. When replacing Mol2Vec with fingerprints, the CNN failed to demonstrate meaningful learning. Similarly, while ChemBERTa2 embeddings outperformed a random baseline, they did not match the performance of the CNN + Mol2Vec model. We explored several alternative CNN configurations—including lower learning rates, fewer convolutional layers, and feature reduction—but none yielded significant improvements, with only 27% of correct top-ranked candidates and 60% within the top five using ChemBERTa2 embeddings (see Supplementary Table 6).

We then evaluated the performance of fingerprints using a fully connected deep neural network (DNN) consisting of four dense layers. This architecture improved performance compared to the CNN, yielding 39% correct top-ranked predictions and 73% within the top five (for positive ionization mode). Next, we tested ChemBERTa2 embeddings using DNNs with hyperparameter tuning and early stopping. This combination significantly outperformed the CNN + ChemBERTa2 setup and achieved results comparable to CNN + Mol2Vec, with 44% correct top-ranked candidates and 78% within the top five. For completeness, we also trained DNN models with Mol2Vec embeddings, which performed slightly below the ChemBERTa2 + DNN configuration (see Supplementary Table 6). These findings underscore the interplay between embedding type and model architecture: Mol2Vec embeddings are best suited for convolutional architectures, whereas ChemBERTa2 embeddings perform more effectively with deep fully connected networks.

To better understand the source of these performance differences, we conducted two complementary analyses: (i) Principal Component Analysis (PCA) of 3850 randomly selected embeddings showed that Mol2Vec captured substantially more variance in the first two principal components (PC1 = 35.06%, PC2 = 19.12%; total = 54.18%) than ChemBERTa2 (PC1 = 13.14%, PC2 = 10.07%; total = 23.21%) (Supplementary Fig. 4). Visual inspection of the PCA plots suggest that Mol2Vec encodes molecules into a low-dimensional, smooth latent space. This continuous and spatially coherent structure aligns well with CNNs, which exploit local relationships and hierarchical patterns. In contrast, ChemBERTa2 produced high-dimensional, more fragmented embeddings in which a smaller portion of the variance is explained by the first two PCs. This may indicate a more granular and distributed representation of chemical space, likely capturing subtle differences in local structure and SMILES-specific syntax. (ii) SMILES robustness analysis further emphasized these differences in embedding behavior. For 10 structurally distinct compounds, we generated 10 alternative SMILES strings per compound—each representing the same molecule (Tanimoto similarity = 1)—and compared their corresponding embeddings. As previously observed, Mol2Vec produced nearly identical embeddings across all SMILES variants (cosine similarity ≈ 1.0), demonstrating strong robustness to SMILES encoding. In contrast, ChemBERTa2 embeddings exhibited greater variability in both cosine similarity and Euclidean distance, reflecting higher sensitivity to input SMILES syntax (Supplementary Fig. 5). DNNs, which do not assume spatial locality, might be better equipped to work with ChemBERTa2's fine-grained and potentially more fragmented representations.

Validation with non-annotated spectra

To evaluate the tool with datasets of non-annotated metabolites, we used three publicly available resources: the Critical Assessment of Small Molecule Identification (CASMI) challenges [32] from 2016 and 2022, as well as the Annotated Recurrent Unidentified Spectra (ARUS) database [33] from the NIST Mass Spectrometry Data Center. To expand the pool of potential candidate structures, we created a reference database containing Mol2vec embeddings for 5.52 million molecules, which included 5 million random compounds from PubChem.

The CASMI 2016 challenge dataset contained 208 MS/MS spectra from 188 unique structures, with 127 spectra acquired in positive ion mode and 81 in negative ion mode. After removing any molecules from CASMI 2016 that were present in our CNN training dataset, we retained 27 unique structures in positive mode and 30 in negative mode. For CASMI 2022, which initially included 177 structures in positive ion mode and 108 in negative ion mode, we applied the same filtering process. This resulted in 149 unique compounds in positive mode and 97 in negative mode.

The ARUS dataset contains MS/MS spectra frequently observed in human samples, but which remain unannotated. We selected ARUS spectra with putative molecular formulas assigned using the BUDDY software [34] (see Supplementary File 3). This dataset includes 25 801 spectra from plasma and 68,478 spectra from urine.

For each unknown molecule, we pre-processed its associated spectrum as previously described (see Methods). That is, peaks exceeding the precursor mass by more than 0.5 Da were removed, and fragments with intensities below 1% of the most intense peak were filtered out. The mass differences between precursor and fragment ions were calculated, and the resulting neutral loss values were incorporated into each spectrum. All intensity values were binarized and each spectrum was then encoded as a vector with a bin size of 0.01 Da, with the m/z axis restricted to ≤700 Da.

ChemEmbed generated a list of candidate annotations for each spectrum. These candidates were ranked based on both the Euclidean distance and cosine similarity between the predicted 300-dimensional embeddings and a reference set of 5.52 million Mol2vec embeddings. The top five candidates were selected for further analysis.

In the CASMI 2016 challenge, our model successfully ranked the correct molecule within the top five candidates in over 60% of cases, for both positive and negative ionization spectra (Supplementary Table 7).

We then compared the performance of ChemEmbed against the latest version of SIRIUS (version 6.0.0) [7] (see Methods) using both the CASMI 2022 and ARUS datasets. Both tools used the same reference dataset of 5.52 million molecules as the pool of potential candidate structures, matching either molecular fingerprints (SIRIUS) or 300-dimensional Mol2vec embeddings (ChemEmbed). To ensure a fair comparison, we removed compounds from the CASMI 2022 dataset that had been included in the training of the latest version of SIRIUS (see Methods), resulting in 107 unknown compounds for both tools in positive ionization mode and 64 in negative ionization mode. ChemEmbed ranked the correct compound within the top five candidates in 33% of cases for positive ionization spectra and 30% for negative ionization spectra. In contrast, SIRIUS achieved success rates of 28% for positive ionization and 17% for negative ionization when considering the top five candidates (Supplementary Table 7).

Finally, we applied ChemEmbed to process 25 801 MS/MS spectra from plasma and 68 478 spectra from urine in the ARUS dataset. ChemEmbed generated potential annotations for 23.8% of positive ionization spectra (Supplementary File 4) and 19.7% of negative ionization spectra (Supplementary File 5), based on criteria of cosine similarity scores above 0.95 and Euclidean distances below 25 (Supplementary Table 8). Among these high-confidence matches, we identified seven compounds by spectral matching against an emerging repository of synthesized compounds from UCSD [35]—four from plasma and three from urine (Supplementary Fig. 6). Additionally, to further assess annotation performance, we randomly selected 40 spectra for manual review. An expert in analytical chemistry evaluated the top five candidates for each spectrum, identifying 25 compounds in total (Fig. 3, Supplementary File 6 and Supplementary Fig. 7).

Attempts to replicate this analysis using SIRIUS 6.0 were unsuccessful, as the tool either failed to complete the processing of the large dataset or caused the server to crash. Consequently, we restricted direct comparisons to the 25 manually reviewed compounds. Among these, SIRIUS correctly ranked 16 compounds within the top five

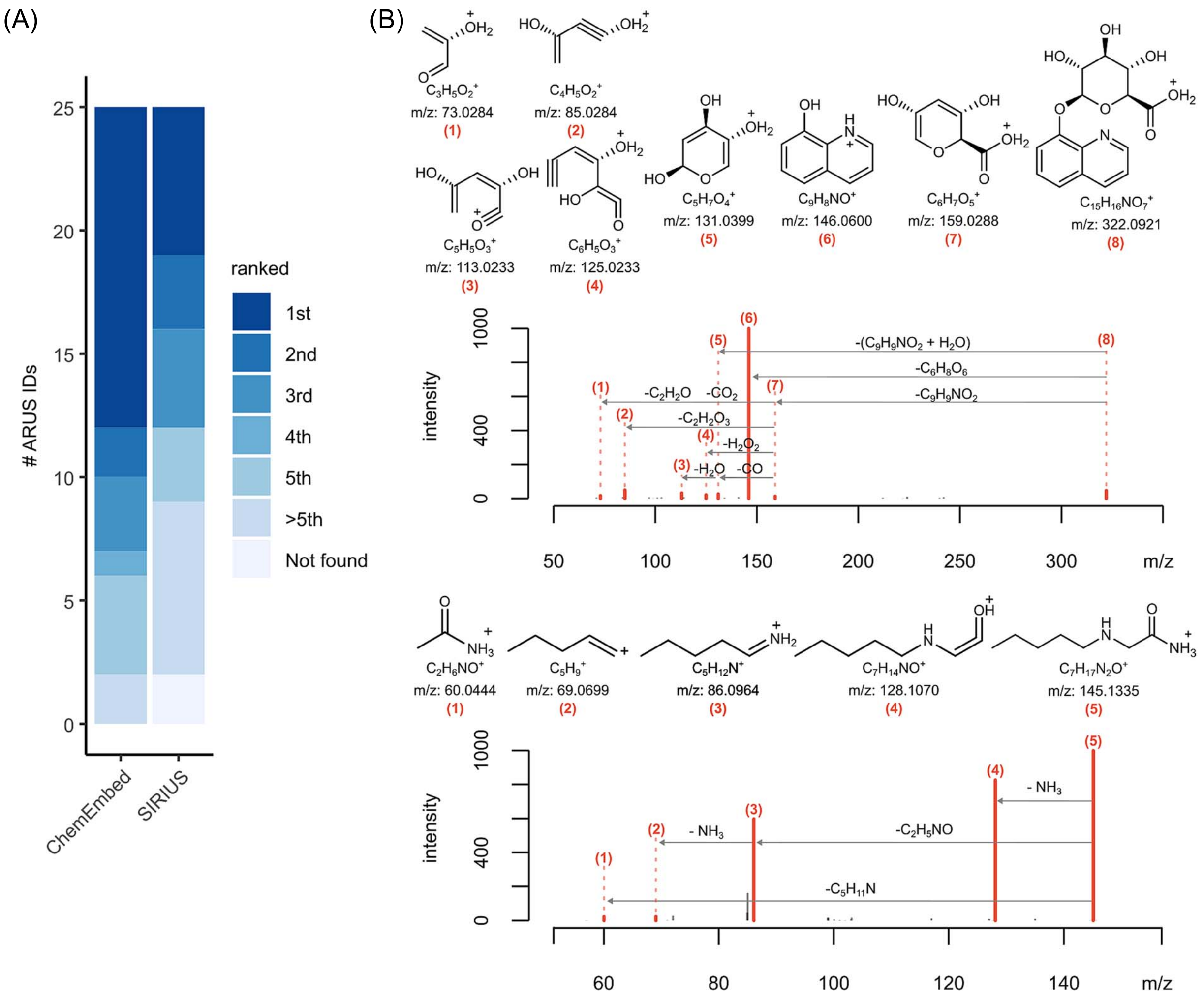

Figure 3 ChemEmbed applied to the ARUS dataset. (A) Stacked bar chart showing the ranking positions of 25 compounds identified in positive ionization mode across plasma and urine by ChemEmbed, compared to SIRIUS. (B) Annotated MS/MS spectra for 8-hydroxyquinoline glucuronide (top) and Milacemide (bottom) from the ARUS urine spectral dataset, showing tentative fragment ion assignments and inferred neutral losses, calculated as the mass differences between the precursor ion and each observed product ion.

candidates (Fig. 3), with five compounds ranked higher by SIRIUS than by ChemEmbed, highlighting the complementarity of the two tools (Supplementary File 6). Therefore, when maximum annotation coverage is desired, combining both computational approaches may be beneficial.

Metabolite annotation in response to physical activity

Finally, we analyzed serum samples from a cohort of 46 healthy sedentary individuals collected before and after 3 weeks of exercise training, using both positive and negative ionization modes. The dataset included three exercise modalities—high-intensity interval training (HIIT), moderate-intensity continuous training (MICT), and super-slow strength training (SSST)—which were treated collectively to identify overall metabolic changes induced by the 3-week training program.

A paired *t*-test identified 166 statistically significant features ($P <$ .05; 56 in positive mode and 110 in negative mode), which were subsequently annotated using both conventional spectral matching against common reference databases (NIST23, GNPS, MS-DIAL, and METLIN) and ChemEmbed. Spectral matching identified 6 and 11 compounds (cosine similarity >0.80) in positive and negative ionization modes, respectively, whereas ChemEmbed identified 10 and 44 compounds (cosine similarity >0.90) (Fig. 4), highlighting ChemEmbed's ability to extend metabolite coverage beyond existing spectral resources. Several metabolites, such as the dipeptides Phe-Trp and Phe-Phe, O-acetylcarnitine, LPE(16:0/0:0), PI(18:0/20:4), an isoform of dihydroxybenzoic acid, and phenyllactic acid, were consistently annotated by both spectral matching and ChemEmbed (Supplementary File 8). In contrast, compounds such as inosine were detected exclusively by spectral matching, as ChemEmbed did not annotate sodium-adducted species ($[M+Na]^+$), which were not included in its training set.

Discussion

Despite the growing number of reference MS/MS spectra, public and commercial databases still cover less than 15% of the known small molecule chemical space [36, 37]. Even focusing on genome-scale metabolic networks—excluding the vast chemical diversity introduced by microbiota, environmental exposures, diet, and

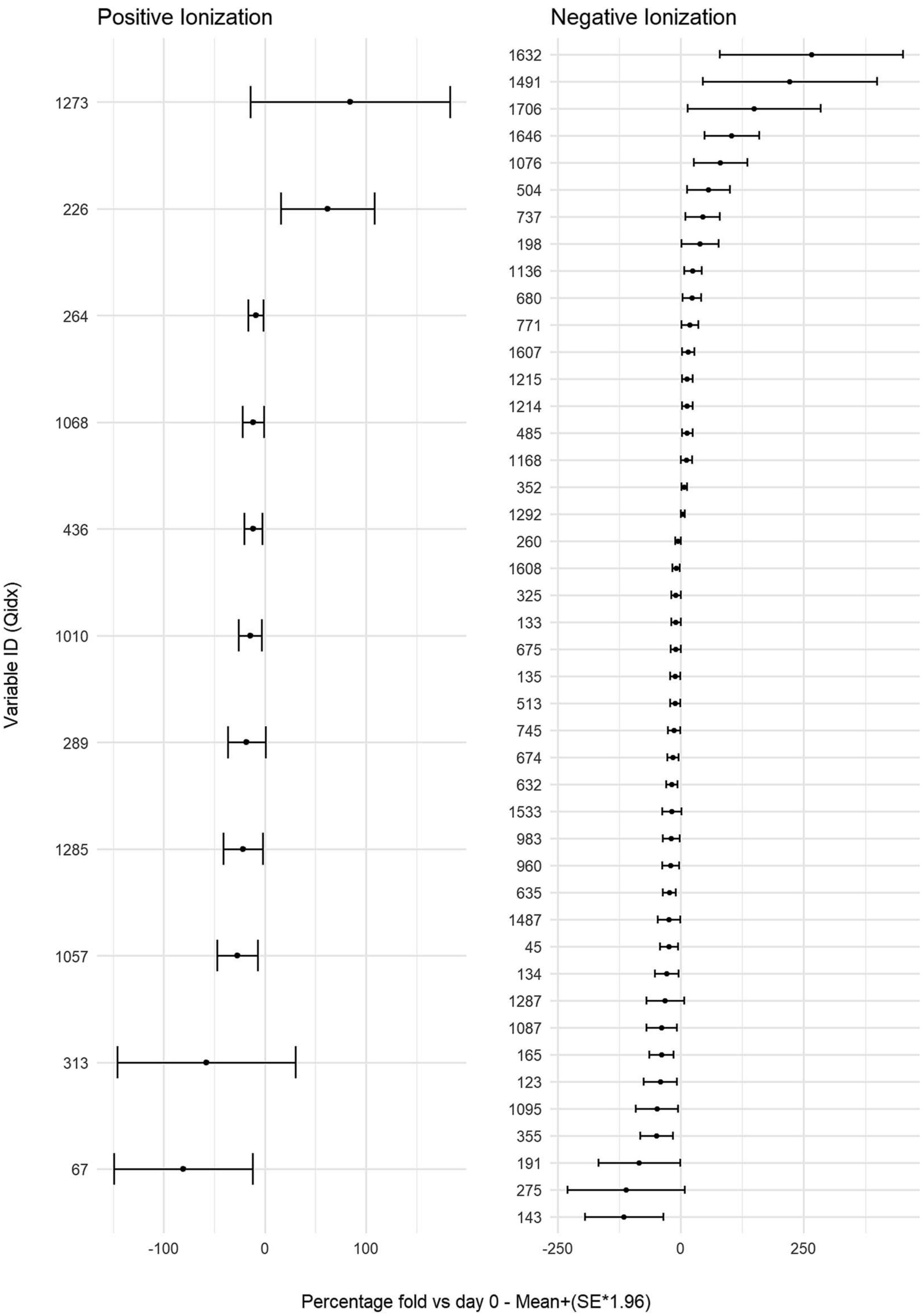

Figure 4 Differentially up- and down-regulated metabolites after 3 weeks of physical activity, analyzed in positive (left) and negative (right) ionization modes. The x-axis represents the percentage change relative to baseline (day 0), showing the mean percentage ± standard error (×1.96) corresponding to a 95% confidence interval. The y-axis lists the variable identifiers of compounds annotated by spectral matching and/or ChemEmbed (see supplementary file 8 for details).

contaminants—only ∼40% of eukaryotic metabolic networks can be mapped using available spectral standards [38].

Machine learning tools are helping to bridge this gap, but they face challenges in learning meaningful patterns due to the high dimensionality and sparsity of MS/MS spectra and metabolite structures. A key issue stems from training models on reference MS/MS spectra acquired at fixed collision energies, as seen in databases like METLIN, MassBank, and GNPS. Our study demonstrates that training neural networks with individual collision energy spectra per molecule underperforms compared to using a single merged spectrum. The higher sparsity of individual spectra complicates learning and increases the risk of overfitting.

Using merged spectra also addresses the computational complexity of fixed-energy inputs. For instance, training one epoch with individual spectra took over 95 min, compared to just 5.5 min with merged spectra. This efficiency extends to inference: processing 25 801 MS/MS spectra from plasma and 68 478 from urine in the ARUS dataset required less than 1 h and 2 h on a basic laptop CPU, respectively. These results highlight the scalability and practicality of our approach.

Modern qTOF and Orbitrap instruments now support ever growing faster scan rates, improved duty cycles, and advanced collision energy options, such as full-CE ramps and stepped collision energies. These advancements provide more comprehensive fragmentation patterns, enhancing metabolite identification accuracy. However, leveraging such rich spectral information requires careful preprocessing. ChemEmbed merges data from multiple collision energies and binarizes the intensities, simplifying the input and reducing overfitting risks by focusing on the presence of fragment ions rather than their intensities. Fragment intensities can vary widely due to factors such as fragmentation technique, collision energy, and instrument-specific differences [39, 40] adding unnecessary complexity and potential noise if treated as model inputs.

Our results also highlight the critical interplay between molecular embeddings and machine learning architectures. Although both Mol2Vec and ChemBERTa2 encode chemically meaningful information, their latent structures differ in ways that directly affect performance. Mol2Vec produces smooth, low-dimensional spaces that are highly invariant to SMILES syntax, properties that align well with convolutional architectures designed to exploit local continuity and hierarchical patterns. In contrast, ChemBERTa2 embeddings, trained on a much larger corpus (77 M versus 19.9 M molecules for Mol2Vec), are more sensitive to SMILES representation and distribute information across higher-dimensional, fragmented spaces. These characteristics make them less compatible with CNNs but better suited to fully connected DNNs, which can leverage fine-grained, distributed representations without assuming spatial coherence. Future work should extend this analysis to other embedding strategies, such as graph neural networks (GNNs) and alternative Transformer-based models, to systematically investigate how molecular representations and neural architectures jointly shape performance in metabolite identification tasks.

Similarly, the poor performance of fingerprints with CNNs can be explained by the mismatch between the representation and the inductive biases of the architecture. Fingerprints are sparse and binary vectors in which each bit encodes the presence of a predefined substructure, but the bit positions are arbitrary and lack intrinsic ordering. As a result, adjacent features do not carry related chemical information, making convolutional filters ineffective since they are designed to exploit local continuity and hierarchical structure. In contrast, fully connected DNNs do not assume spatial locality and instead evaluate each feature independently, which is more consistent with the discrete, unordered nature of fingerprints. Nevertheless, their performance remained inferior to Mol2Vec or ChemBERTa2 embeddings, likely because handcrafted fingerprints provide a limited and lossy representation of molecular structure compared to the continuous, chemically contextualized latent spaces produced by learned embeddings.

Finally, both the limited availability of reference spectra [41] and the quality of MS/MS data in both public (community-contributed) and commercial libraries remain critical factors for improving computational annotation. At an early stage, we considered incorporating low-resolution spectral data (e.g. from triple quadrupole instruments) to increase coverage. However, these datasets were ultimately excluded because our spectral binning strategy relies on a fixed bin width of 0.01 Da. This resolution was chosen to enhance specificity while keeping computational costs manageable. Low-resolution spectra are incompatible with this approach, as their broader peak widths would be distributed across multiple bins, leading to information loss and reduced discriminative power.

ChemEmbed is currently limited to annotating MS/MS spectra of compounds whose structures are present in the embedding space. However, expanding the reference database indiscriminately does not necessarily improve performance. In the ARUS dataset, we tested the effect of adding ∼5 million randomly selected molecules from PubChem to broaden chemical coverage. Although this substantially increased the database size, it also introduced many compounds of limited biological relevance (e.g. purely synthetic or drug-like molecules) along with large numbers of isomeric variants—predominantly stereoisomers and tautomers that will not likely be distinguished by MS/MS alone. When test spectra were compared against this expanded 5.5 million–compound database rather than the original 0.52 million Mol2Vec embeddings, ranking performance declined (Top-1: 43% → 26%; Top-5: 76% → 55%): top candidate positions became dominated by isomeric forms, adding noise rather than biologically meaningful alternatives. These findings suggest that, in metabolomics applications, using a smaller, context-specific reference library enriched in chemically and biologically relevant molecules may be more effective than expanding the search space with large, unspecific databases. For example, HMDB and ChEBI are well suited for biomedical studies, COCONUT and LOTUS for natural products in plants or microbial metabolites, and the NORMAN database for environmental contaminants. Such targeted libraries should reduce chemical redundancy and improve the precision and interpretability of metabolite annotation results.

Moreover, ChemEmbed currently provides structural predictions without associated confidence estimates. Addressing this limitation, along with the challenges outlined above, points to several promising directions for future development. These include integrating confidence measures to improve the reliability and interpretability of predictions [42], strategically expanding chemical space to incorporate newly discovered metabolites [43, 44], and programmatically generating plausible structural variants—such as structural isoforms or common phase I/II metabolic derivatives of biologically relevant molecules—with precomputed Mol2Vec embeddings.

In summary, ChemEmbed aligns with the capabilities of modern mass spectrometry instrumentation by balancing performance metrics like accuracy, computational cost, and scalability. Unlike many prior tools, it addresses real-world usability, making it a robust solution for metabolite identification in high-throughput biomedical applications and re-annotating large-scale clinical datasets to uncover previously unrecognized metabolites associated with disease, diet, or microbiome-related pathways.

Key Points

- ChemEmbed is a machine learning tool that improves the identification of unknown small molecules from mass spectrometry data.
- It combines merged MS/MS spectra from multiple collision energies with predicted neutral losses from over 38 000 compounds to enhance CNN input.
- ChemEmbed ranks the correct molecule first in 42% of cases and in the top five in over 76%.
- It outperforms SIRIUS 6 on CASMI 2016/2022 benchmarks and the ARUS dataset.
- Scalable to large datasets, ChemEmbed accelerates accurate metabolite identification in medicine, nutrition, and environmental research.

Author contributions

Conceptualization: MS-P, RGu, OY. Methodology: MS-P, RGu, OY, LB, AN, J-MS. Software: MFK. Investigation: MFK, MP-R, MR-B, IPL, RGi, SX, AP, JMB, SJ, LB, AN, J-MS. Data curation: MFK, MP-R, MR-B, IPL, RGi, SX, AP, JMB. Formal analysis: MFK, PD, MS-P, RGu, OY. Validation: SJ, LB, AN, J-MS. Visualization: MFK, MV. Writing—original draft, review and editing: MFK, OY. Supervision: MS-P, RGu, OY.

Conflict of interest

P.C.D. is an advisor and holds equity in Cybele, BileOmix, Sirenas and a scientific co-founder, advisor, holds equity and/or received income from Ometa, Enveda, and Arome with prior approval by UC San Diego. P.C.D. also consulted for DSM animal health in 2023. The rest of the authors declare no conflict of interest.

Funding

This research was funded by projects PID2022-136226OB-I00 (OY and MV) and PID2022-142600NB-I00 (MS-P and RGu) from MCIN/AEI/10.13039/501100011033 FEDER, UE, by projects 2021SGR842 (OY and MV) and 2021SGR-633 (MS-P and RGu) from the Government of Catalonia, and by the European Union NextGenerationEU/PRTR, and by projects PI20-00658 and PI23/00514 (AN and J-MS) by the Instituto de Salud Carlos III (ISCIII) co-funded by the European Union. MFK has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 945413 and from Universitat Rovira i Virgili.

Data availability

All Supplementary Files and raw MS files, including MS1 (mzML) and MS2 (mzML and MGF) data, are available on the Zenodo repository at: https://zenodo.org/records/17534670.

Code availability

A Python implementation of ChemEmbed, including support for Mol2Vec and ChemBERTa-2 embeddings, along with the trained ChemEmbed models, is available at https://github.com/massspecdl/ChemEmbed.

References

1. Giera M, Yanes O, Siuzdak G. Metabolite discovery: Biochemistry's scientific driver. *Cell Metab* 2022;**34**:21–34.
2. Wishart DS, Guo A, Oler E *et al.* HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;**50**:D622–31. https://doi.org/10.1093/nar/gkab1062.
3. Ruttkies C, Neumann S, Posch S. Improving MetFrag with statistical learning of fragment annotations. *BMC Bioinformatics* 2019;**20**: 1–14.
4. Tsugawa H, Kind T, Nakabayashi R *et al.* Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal Chem* 2016;**88**:7946–58. https://doi.org/10.1021/acs.analchem.6b00770.
5. Goldman S, Wohlwend J, Stražar M *et al.* Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nat Mach Intell* 2023;**5**:965–79.
6. Young A, Wang F, Wishart DS. *et al. FraGNNet: A Deep Probabilistic Model for Mass Spectrum Prediction*. arXiv:2404.02360v2 2024. https://doi.org/10.48550/arXiv.2404.02360.
7. Dührkop K, Fleischauer M, Ludwig M *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**:299–302. https://doi.org/10.1038/s41592-019-0344-8.
8. Wolf S, Schmidt S, Müller-Hannemann M *et al.* In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 2010;**11**:1–12.
9. Bittremieux W, Laukens K, Noble WS. Extremely fast and accurate open modification spectral library searching of high-resolution mass spectra using feature hashing and graphics processing units. *J Proteome Res* 2019;**18**:3792–9. https://doi.org/10.1021/acs.jproteome.9b00291.
10. Van Der Hooft JJJ, Wandy J, Barrett MP *et al.* Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci USA* 2016;**113**:13738–43.
11. Fan Z, Alley A, Ghaffari K *et al.* MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. *Metabolomics* 2020;**16**. https://doi.org/10.1007/s11306-020-01726-7.
12. Smith CA, Want EJ, O'Maille G *et al.* XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 2006;**78**: 779–87.
13. Baygi SF, Barupal DK. IDSL_MINT: a deep learning framework to predict molecular fingerprints from mass spectra. *J Chem* 2024;**16**:1–8.
14. Laponogov I, Sadawi N, Galea D *et al.* ChemDistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics* 2018;**34**:2096–102.
15. Dührkop K. Deep kernel learning improves molecular fingerprint prediction from tandem mass spectra. *Bioinformatics* 2022;**38**:i342–i349. https://doi.org/10.1093/bioinformatics/btac260.
16. Li S, Liu Y, Chen D *et al.* Encoding the atomic structure for machine learning in materials science. *Wiley Interdiscip Rev Comput Mol Sci* 2022;**12**:e1558. https://doi.org/10.1002/wcms.1558.
17. Huber F, van der Burg S, van der Hooft JJJ *et al.* MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J Chem* 2021;**13**:1–14.
18. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 2018;**58**:27–35. https://doi.org/10.1021/acs.jcim.7b00616.
19. Dührkop K, Shen H, Meusel M *et al.* Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc*

Natl Acad Sci USA 2015;**112**:12580–5. https://doi.org/10.1073/pnas.1509788112.

20. Giné R, Capellades J, Badia JM *et al.* HERMES: a molecular-formula-oriented method to target the metabolome. *Nat Methods* 2021;**18**:1370–6. https://doi.org/10.1038/s41592-021-01307-z.
21. Wallace WE, Moorthy AS. NIST mass spectrometry data Center standard reference libraries and software tools: application to seized drug analysis. *J Forensic Sci* 2023;**68**:1484–93. https://doi.org/10.1111/1556-4029.15284.
22. Tsugawa H, Cajka T, Kind T *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat Methods* 2015;**12**:523–6. https://doi.org/10.1038/nmeth.3393.
23. Wang M, Carver JJ, Phelan VV *et al.* Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* 2016;**34**:828–37. https://doi.org/10.1038/nbt.3597.
24. Aisporna A, Benton HP, Chen A *et al.* Neutral loss mass spectral data enhances molecular similarity analysis in METLIN. *J Am Soc Mass Spectrom* 2022;**33**:530–4. https://doi.org/10.1021/jasms.1c00343.
25. Irwin JJ, Tang KG, Young J *et al.* ZINC20 - a free Ultralarge-scale chemical database for ligand discovery. *J Chem Inf Model* 2020;**60**:6065–73. https://doi.org/10.1021/acs.jcim.0c00675.
26. Zdrazil B, Felix E, Hunter F *et al.* The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* 2024;**52**:D1180–92. https://doi.org/10.1093/nar/gkad1004.
27. Sorokina M, Merseburger P, Rajan K *et al.* COCONUT online: collection of open natural products database. *J Chem* 2021;**13**:1–13.
28. Zulfiqar M, Stettin D, Schmidt S *et al.* Untargeted metabolomics to expand the chemical space of the marine diatom Skeletonema marinoi. *Front Microbiol* 2023;**14**. https://doi.org/10.3389/fmicb.2023.1295994.
29. Dunkel M. SuperNatural: a searchable database of available natural compounds. *Nucleic Acids Res* 2006;**34**:D678–83. https://doi.org/10.1093/nar/gkj132.
30. Nuzzo A, Saha S, Berg E *et al.* Expanding the drug discovery space with predicted metabolite–target interactions. *Commun Biol* 2021;**4**:288. https://doi.org/10.1038/s42003-021-01822-x.
31. Ahmad W, Simon E, Chithrananda S *et al.* ChemBERTa-2: towards chemical foundation models. arXiv:2209.01712v1. https://doi.org/10.48550/arXiv.2209.01712
32. McEachran AD, Chao A, al-Ghoul H *et al.* Revisiting five years of CASMI contests with EPA identification tools. *Meta* 2020;**10**:260. https://doi.org/10.3390/metabo10060260.
33. Simón-Manso Y, Marupaka R, Yan X *et al.* Mass spectrometry fingerprints of small-molecule metabolites in biofluids: building a spectral library of recurrent spectra for urine analysis. *Anal Chem* 2019;**91**:12021–9. https://doi.org/10.1021/acs.analchem.9b02977.
34. Xing S, Shen S, Xu B *et al.* BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nat Methods* 2023;**20**:881–90.
35. Gentry EC, Collins SL, Panitchpakdi M *et al.* Reverse metabolomics for the discovery of chemical structures from humans. *Nature* 2023;**626**:419–26.
36. Vinaixa M, Schymanski EL, Neumann S *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. *TrAC Trends Anal Chem* 2016;**78**:23–35.
37. de Jonge NF, Louwen JJR, Chekmeneva E *et al.* MS2Query: reliable and scalable MS2 mass spectra-based analogue search. *Nat Commun* 2023;**14**:1752. https://doi.org/10.1038/s41467-023-37446-4.
38. Frainay C, Schymanski EL, Neumann S *et al.* Mind the gap: mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites* 2018;**8**. https://doi.org/10.3390/metabo8030051
39. Hoang C, Uritboonthai W, Hoang L *et al.* Tandem mass spectrometry across platforms. *Anal Chem* 2024;**96**:5478–88. https://doi.org/10.1021/acs.analchem.3c05576.
40. Kind T, Tsugawa H, Cajka T *et al.* Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom Rev* 2017;**37**:513–32. https://doi.org/10.1002/mas.21535.
41. Böcker S. Searching molecular structure databases using tandem MS data: are we there yet? *Curr Opin Chem Biol* 2017;**36**:1–6.
42. Hoffmann MA, Nothias LF, Ludwig M *et al.* High-confidence structural annotation of metabolites absent from spectral libraries. *Nat Biotechnol* 2022;**40**:411–21. https://doi.org/10.1038/s41587-021-01045-9.
43. Mohanty I, Mannochio-Russo H, Schweer JV *et al.* The underappreciated diversity of bile acid modifications. *Cell* 2024;**187**:1801–1818.e20. https://doi.org/10.1016/j.cell.2024.02.019.
44. Mannochio-Russo H, Charron-Lamoureux V, van Faassen M *et al.* The microbiome diversifies N-acyl lipid pools - including short-chain fatty acid-derived compounds. *bioRxiv* 2024.10.31.621412; https://doi.org/10.1101/2024.10.31.621412.

Briefings in Bioinformatics, 2026, **27**, bbag054
https://doi.org/10.1093/bib/bbag054
Case Study