Bioinformatics, 2025, **41(7)**, btaf354
https://doi.org/10.1093/bioinformatics/btaf354
Advance Access Publication Date: 17 June 2025
Original Paper

OXFORD

Systems biology

JESTR: Joint Embedding Space Technique for Ranking candidate molecules for the annotation of untargeted metabolomics data

Apurva Kalia[1], Yan Zhou Chen[1], Dilip Krishnan[2], Soha Hassoun[1,3,*]

[1]Department of Computer Science, Tufts University, Medford, MA 02155, United States
[2]Google DeepMind, Mountain View, CA 94043, United States
[3]Department of Chemical and Biological Engineering, Tufts University, Medford, MA 02155, United States

*Corresponding author: Department of Computer Science, Tufts University, Medford, MA 01255, United States. E-mail: soha.hassoun@tufts.edu.

Associate Editor: Macha Nikolski

Abstract

Motivation: A major challenge in metabolomics is annotation: assigning molecular structures to mass spectral fragmentation patterns. Despite recent advances in molecule-to-spectra and in spectra-to-molecular fingerprint (FP) prediction, annotation rates remain low.

Results: We introduce in this article a novel tool (JESTR) for annotation. Unlike prior approaches that "explicitly" construct molecular FPs or spectra, JESTR leverages the insight that molecules and their corresponding spectra are views of the same data and effectively embeds their representations in a joint space. Candidate structures are ranked based on cosine similarity between the embeddings of query spectrum and each candidate. We evaluate JESTR against mol-to-spec, spec-to-FP, and spec-mol matching annotation tools on four datasets. On average, for rank@[1–20], JESTR outperforms other tools by 55.5%–302.6%. We further demonstrate the strong value of regularization with candidate molecules during training, boosting rank@1 performance by 5.72% across all datasets and enhancing the model's ability to discern between target and candidate molecules. When comparing JESTR's performance against that of publicly available pretrained models of SIRIUS and CFM-ID on appropriate subsets of MassSpecGym dataset, JESTR outperforms these tools by 31% and 238%, respectively. Through JESTR, we offer a novel promising avenue toward accurate annotation, therefore unlocking valuable insights into the metabolome.

Availability and implementation: Code and dataset available at https://github.com/HassounLab/JESTR1/.

1 Introduction

Analysing biological samples using "untargeted metabolomics," where masses of thousands of metabolites within a biological sample are detected, presents unprecedented opportunities to characterize the metabolome. Annotation, the process of assigning chemical structures to metabolomics measurements, however is riddled with uncertainty. Naïvely, one can presume that the measured mass could be used to determine a metabolite's molecular structure. However, a particular molecular mass can map to possibly thousands of candidate molecular structures that share the same chemical formula. For example, there are 44 374 known molecular structures in PubChem that are associated with chemical formula $C_{12}H_{18}N_2O_2$. Advanced computational methods promise solutions to address the challenges of annotation uncertainty. IPA (Del Carratore *et al.* 2019) and ipaPy2 (Del Carratore *et al.* 2023) enhance confidence in metabolite identification by integrating additional data such as isotope patterns and adduct formation, to provide statistically rigorous probability estimates for each annotation. Mummichog (Li *et al.* 2013) shifts the focus from individual metabolites to entire metabolic pathways, leveraging peak co-occurrence to infer biological relevance even when exact structures remain unknown. PUMA (Hosseini *et al.* 2020) further advances annotation by applying probabilistic modeling, predicting pathway activity, and assigning chemical identities based on generative inference. Together, these tools improve annotation accuracy, enhance biological interpretation, and enable a more comprehensive understanding of the metabolome. Despite these advances based on analyzing the ionized metabolites, annotation rates remain low.

With advances in instrumentation, it is now possible to not only measure the mass of ionized molecules, but to also measure masses of ionized molecular fragments. Combining liquid chromatography (LC) with mass spectrometry (MS) or combining two such analysis steps (tandem MS/MS) have now become dominant in metabolomics. The measured mass spectrum is a collection of peaks (Fig. 1A). Each peak is represented by its mass-to-charge (m/z) ratio, where the charge is known and is often $+1$ or -1, and a relative intensity. Even for an experienced analytical chemist, assigning a chemical structure to LC–MS or MS/MS spectra is an unsolved problem as the spectrum provides a partial view on the measured molecule. Indeed, only ions that are formed by the loss or a gain of a charge can be detected by MS.

Several techniques address the spectra annotation problem. The go-to technique is "spec-to-spec" comparison (Fig. 1B) of the measured query spectra against spectra that are cataloged in

Received: 21 November 2024; **Revised:** 5 June 2025; **Editorial Decision:** 10 June 2025; **Accepted:** 13 June 2025

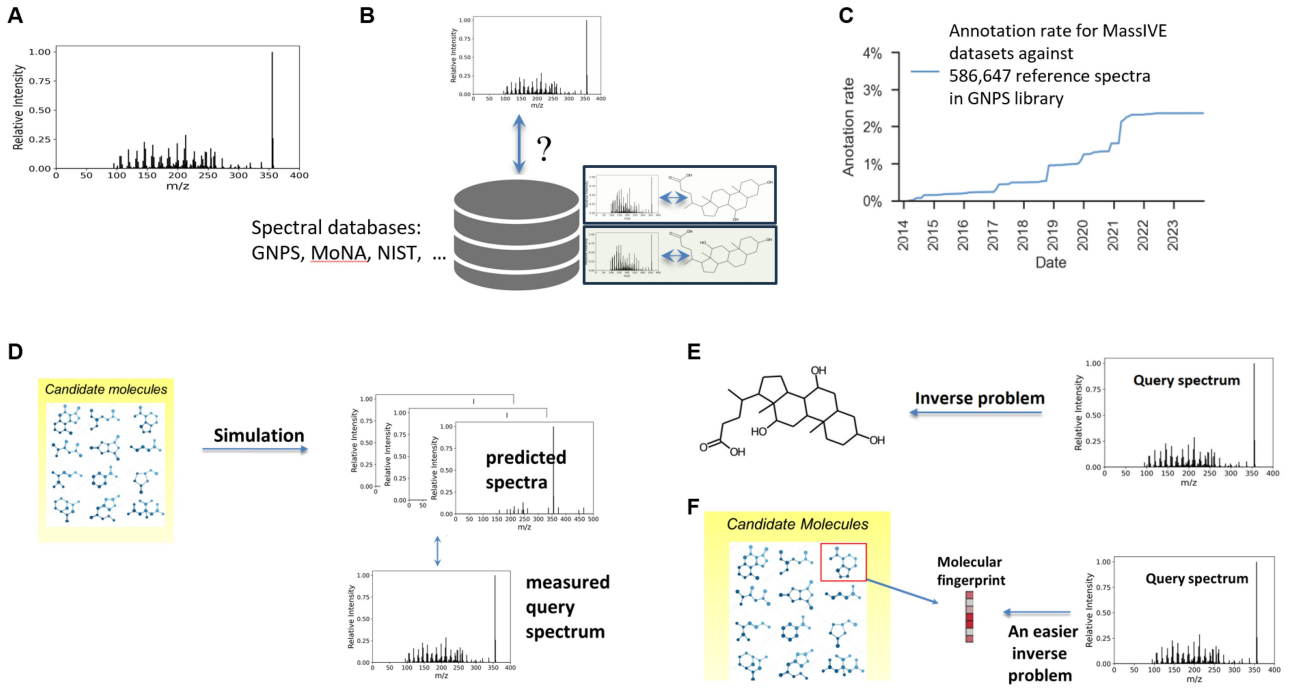

Figure 1. Current annotation workflows. (A) Example spectrum measured using LC–MS or MS/MS, where the *x*-axis represents the mass-to-charge ratio and the *y*-axis represents the relative intensity of each peak. (B) Reference library search using spec-to-spec comparison. (C) Current annotation rates using state-of-the-art library search are low despite growth in reference databases. (D) Mol-to-spec predictive approach mimics the mass spectrometry fragmentation process. (E) Spec-to-mol involves *de novo* molecular generation from spectra and forms an inverse problem. (F) Spec-to-FP approach predicts a molecular fingerprint and identifies the candidate structure that most likely matches the predicted fingerprint.

spectral reference libraries (Kind *et al.* 2018). Searching spectral libraries using learned embeddings, e.g. Spec2Vec (Huber *et al.* 2021), MSBERT (Zhang *et al.* 2024), DreaMS (Bushuiev *et al.* 2024a), can improve search performance. However, despite embedding advances and growth in spectral libraries, e.g. GNPS (Wang *et al.* 2016), NIST (https://chemdata.nist.gov/), MoNA (https://mona.fiehnlab.ucdavis.edu/), annotation rates remain extremely low due to the limited coverage of spectral libraries in comparison to the space of all potential molecules. In addition, measured spectra vary tremendously under differing instrument settings, e.g. ionization energy, solvent type, and adduct formation (additional functional groups attached or removed from the ionized molecule). A molecule therefore may have many corresponding spectra, which further limits library coverage. A recent search of spectra within 15 327 datasets (Wang *et al.* 2016) deposited in the MassIVE (Mass Spectrometry Interactive Virtual Environment) database against 586 647 reference spectra cataloged in the GNPS reference library yielded a positive identification rate of 2.3% (Martin *et al.* 2024) (Fig. 1C).

Two types of supervised predictive annotation techniques have emerged. "Mol-to-spec" techniques (Fig. 1D) utilize combinatorial fragmentation approaches, e.g. MetFrag (Wolf *et al.* 2010, Ruttkies *et al.* 2016) and CFM-ID (Wang *et al.* 2021), MLPs (Wei *et al.* 2019), or GNNs (Zhu *et al.* 2020, Young *et al.* 2021, Li *et al.* 2024), to translate a molecular structure into a predicted spectrum. Candidate molecular structures are retrieved by either chemical formula, if available, or molecular mass from large molecular databases such as PubChem, or more biologically relevant, smaller, databases. The candidate with the most similar spectrum to the query spectrum is ranked highest and used as the annotation. In contrast, "spec-to-mol" techniques (Fig. 1E) aim to generate *de novo* molecular candidates that potentially match the query spectrum, e.g. MSNovelist (Stravs *et al.* 2022), Spec2Mol (Litsa *et al.* 2023), MS2Mol (Butler *et al.* 2023). For example, MS2Mol uses sequence-to-sequence transformers to translate spectra into *de novo* molecular structures in the form of SMILES strings. Due to their current limited capabilities, *de novo* generation is presently of limited use in the metabolomics community. An alternative and earlier approach is "spec-to-FP" (Fig. 1F), where a molecular fingerprint (FP) vector is predicted for the query spectrum, e.g. Sirius (Dührkop *et al.* 2019), MIST (Goldman *et al.* 2023). Here, the predicted FP is compared against those of the candidate molecular structures, and the best match, via Tanimoto or cosine similarity, is declared the annotation result. Despite recent advances in all such techniques, annotation rates remain low as the *reconstruction* of spectrum, FP, or molecular structure is a difficult task.

We address in this article the problem of assigning chemical structures from a candidate set to spectral data. The novelty in our approach lies in avoiding the explicit generation of molecular FPs and spectra (Fig. 2A), "and" in considering a molecule and its spectra as views of the same object (Fig. 2B). One way of avoiding the explicit construction was recently suggested by ChemEmbed (Faizan-Khan *et al.* 2025), which trains a spectra encoder to predict the Mol2Vec molecular embeddings (Jaeger *et al.* 2018), effectively defining the embeddings based on the Mol2Vec embedding space. The molecule-spectra multi-modal view insight allows us to embed molecules and their matching spectra close in the molecule–spectrum joint-embedding space. Our approach avoids the need for any kind of reconstruction to intermediate forms such as FPs or spectra and therefore removes any reconstruction loss that would invariably creep into the ranking pipeline with any reconstruction based approach. The ranking of candidate structures can be attained by comparing their embeddings against that of the query spectrum and selecting the candidate with the highest cosine similarity. The idea of

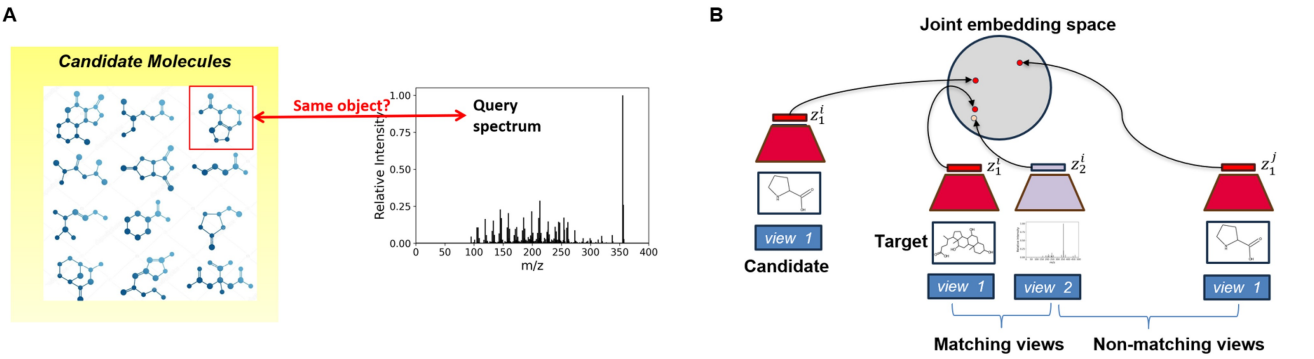

Figure 2. Novelty of the JESTR annotation approach. (A) JESTR avoids the explicit generation of spectra, molecules, and fingerprints, and ranks the candidate molecules against the query spectrum based on their joint-space embeddings. (B) JESTR learns to place representations of matching molecule–spectrum pairs close in the joint-embedding space relative to non-matching pairs. Further, JESTR utilizes additional molecules beyond those in the training set to learn to distinguish target molecules in the training dataset from candidate molecules (those with similar molecular formulas). During ranking, the candidate whose embeddings have the maximum cosine similarity to spectrum embedding is chosen as the highest ranked molecule for that spectrum.

learning joint-embedding spaces from multiple views of the data dates back to the seminal work on Siamese Networks (Chopra *et al.* 2005). More recently, CLIP (Contrastive Language-Image Pre-training) was trained to create a shared embedding space for both images and text, enabling the model to match relevant images and captions without the need for direct labeling or supervised training on specific dataset (Radford *et al.* 2021). As we embed molecules and spectra in a joint-embedding space, our method is termed Joint Embedding Space Technique for Ranking candidate molecules, JESTR. We use CMC, contrastive multiview coding (Tian *et al.* 2020), to learn view-invariant information across different views of the data and produce embeddings in a joint-embedding space. CMC utilizes Normalized Temperature-scaled Cross-Entropy (InfoNCE) Loss, which incentivizes the model to distinguish true positive pairs from all possible negative pairs in a batch by maximizing the similarity between positive pairs while minimizing similarity with negatives. To assess the closeness of a given pair, CMC relies on a temperature-scaled cosine similarity as a discriminating function. The temperature is a model hyper-parameter that controls the sharpness of similarity scores, where lower values sharpens the distinction between positive and negative pairs. In contrast, the recent contrastive-learning based joint-embedding model CMSSP (Chen *et al.* 2024) applies cross-entropy (CE) loss to the dot product of spectral and molecular embeddings after softmax, aiming for high embedding similarity for positives and low embedding similarity for negatives. Broadly speaking, in all contrastive-learning approaches (e.g. Chen *et al.* 2020, Tian *et al.* 2020, Khosla *et al.* 2020) the key question is how to generate the paired views (either appearing naturally or generated via data augmentation); and how to ensure that paired views end up close together in the joint-embedding space.

Another novelty of our approach lies in utilizing regularization on additional data consisting of millions of molecules with the same chemical formulas as those in the training dataset. While this form of data augmentation does not contribute directly to additional labeled training data (Tian and Zhang 2022), the additional data are utilized to distinguish target molecules from their candidates. Here, regularization is used as a fine-tuning strategy toward the end of training. When combined with contrastive loss, regularization with additional data provides two key benefits: it improves model generalization by training on a larger, more diverse set of molecules, and it enhances representation learning for both molecule and spectra embeddings by using non-congruent pairs as additional data during training.

We conduct experiments and analysis to answer the following research questions. (Q1) Does JESTR's implicit annotation method (with and without regularization) outperform prior explicit methods (mol-to-spec and spec-to-fp)? (Q2) Is learning the molecule–spectrum joint-embedding space effective for distinguishing target molecules among their respective candidate sets? We compare JESTR against state-of-the-art mol-to-spec technique, ESP (Li *et al.* 2024), spec-to-FP technique, MIST (Goldman *et al.* 2023), and the recent mol-spec matching technique, CMSSP (Chen *et al.* 2024). We conduct the evaluation using four datasets: the NPLIB1 dataset that was previously released with the CANOPUS tool (Dührkop *et al.* 2021), the well-curated, available-for-purchase NIST2020 dataset, and user-deposited data from MassBank of North America (MoNA), and the recently released benchmarking datasets, MassSpecGym dataset (Bushuiev *et al.* 2024b). Additionally, to ensure a fair evaluation against SIRIUS and CFM-ID, we identify subsets of the MassSpecGym test set that are disjoint from their pre-training data and use these subsets to assess the performance of JESTR relative to these tools.

The contributions of this article are:

- Novel implicit formulation of the annotation problem to avoid explicit prediction of spectra and FPs that has dominated the field since earliest attempts in solving the problem (Heinonen *et al.* 2012). Our formulation is grounded in the novel insight that molecules and spectra are views of the same object, similar to recent advances in linking text/image data.
- Demonstrating that contrastive learning is effective in creating a molecular-spectra joint-embedding space, and that cosine similarity of the embeddings is sufficient for ranking candidate molecules. That is, there is no need for an explicit (learnable) downstream ranking task.
- Fine-tuning the implicit model via regularization using the candidate sets of the training molecules improves the rank@1 performance in the range of 0.51%–37.05% when compared to a baseline that does not utilize regularization.
- Demonstrating that JESTR outperforms ESP and MIST on all ranking metrics and all datasets with the exception of rank@1 for the MoNA dataset. For rank@[1–20], JESTR outperforms ESP by 55.45%, MIST by 56.6%,

and CMSSP by 302.56% across four datasets. On carefully selected subsets of the MassSpecGym dataset, JESTR outperforms SIRIUS and CFM-ID on rank@1 by 31% and 238%, respectively. These remarkable improvements are achieved even though JESTR does not utilize the additional data in the form of chemical formulae labels for spectra peaks that are currently used by MIST.

2 Methods

The JESTR model architecture (Fig. 3) consists of a molecular encoder and a spectral encoder. They are trained to create embeddings in a molecule–spectrum joint-embedding space. To place views of the same object close to each other in the embedding space, we use the CMC contrastive-learning loss (Tian *et al.* 2020). To improve performance, we utilize regularization. At inference, when provided a candidate set for the query spectrum, the cosine similarity is computed between each candidate and the query spectrum. The candidates are then ranked based on their cosine similarities. Ranking results are reported using rank@k, which is defined as the percentage of target molecules that are ranked at rank k or better.

2.1 Encoders

The molecular encoder is implemented using a multi-layer graph neural network (GNN) encoder. Molecular structures are encoded as graphs, where node features include atom type, atomic mass, valence, if the atom is in a ring, formal charge, radical electrons, chirality, degree, number of hydrogens, and aromaticity. Edge features are the bond type, whether the bond is part of a ring, conjugacy, and one hot encoding of the stereo configuration of the bond. The encoder consists of graph convolutional networks (GCNs) (Kipf and Welling 2016) that aggregate information at each node. The GCNs are followed by a pooling layer and two fully connected layers to generate the final molecular embeddings, z_{mol}, for a given molecule graph c:

$$z_{mol} = \text{MLP}_{\times 2}(\text{MAXPOOL}(\text{GCN}(c))) \tag{1}$$

To prepare the spectrum for its encoder, peak m/z values of the spectra are discretized into bins that are 1 Da wide. Peaks with m/z values larger than 1000 Da were dropped. The intensity are normalized to a max value of 999—a common practice in normalizing spectral data (e.g. for the NIST datasets). For multiple peaks falling within the same bin, peak intensities within each bin are summed to generate the overall intensity value for that bin. A 1000-dimension binned vector therefore encodes the spectrum. A log10/3 transformation is applied to this binned vector to ensure that a few peaks and/or a long tail do not dominate the embedding vector. This 1000-dimension encoded vector was passed through a three-layer MLP to obtain the final spectral embedding, z_{spec}:

$$z_s = \frac{1}{3}\log_{10}(\{\Sigma\ I_i, \forall\, i,\ n < (mz)_i < (n+1), \text{ for } n \text{ in } 0..999\}) \tag{2}$$

where I_i is the intensity of the ith peak and mz_i is the m/z value of the ith peak.

$$z_{spec} = \text{MLP}_{\times 3}(z_s) \tag{3}$$

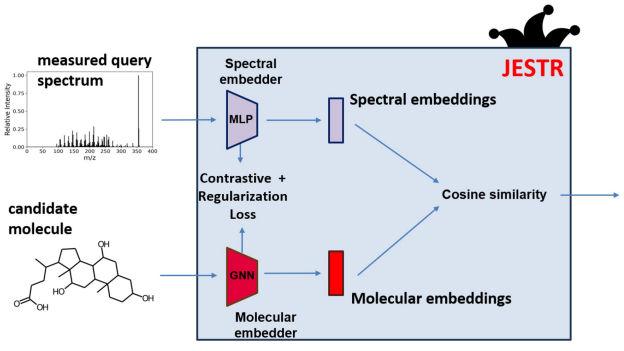

Figure 3. Overview of the JESTR model architecture. The model is trained to minimize the contrastive and regularization losses. The embeddings produced by the encoders are used to compute the cosine similarity in the joint-embedding space between a molecule and a spectrum.

2.2 Contrastive learning of spectral and molecular views

We consider two views of each data item: a molecular and a spectral view. Each data item will have one molecular view but may have multiple spectral views as measurements may be collected under different MS instrumentation conditions. Matching molecule–spectrum views arise from a molecule and its spectrum, while non-matching views arise between a molecule and any of its non-matching spectra. The objective of contrastive learning on multi-views (Chopra *et al.* 2005, Chen *et al.* 2020, Khosla *et al.* 2020, Tian *et al.* 2020) is to learn embeddings that separate samples from matching and non-matching distributions, and to ensure that paired views are close in the joint-embedding space.

As in CMC (Tian *et al.* 2020), we use a discriminator function, h, to measure the closeness of spectral and molecular embeddings using their cosine similarity, modulated by a temperature parameter τ. Thus, given a embedding for spectrum n, and an embedding for molecule m, we can define h as:

$$h(z^n_{spec}, z^m_{mol}) = \exp\left(\frac{z^n_{spec} \cdot z^m_{mol}}{||z^n_{spec}|| \cdot ||z^m_{mol}||} \cdot \frac{1}{\tau}\right) \tag{4}$$

The hyper-parameter τ controls the importance of non-matching pairs in pushing the embeddings apart in the joint-embedding space. To ensure that the discriminator assigns high values for matching pairs and low values for non-matching pairs, we define a contrastive loss, $L_{\text{contrastive}}$, over a batch of size k as:

$$L_{\text{contrastive}} = \frac{1}{k}\sum_{n=1}^{k}\left[-\mathbb{E}\left[\log\frac{h_\theta(z^n_{spec}, z^n_{mol})}{\sum_{m=1}^{k} h_\theta(z^n_{spec}, z^m_{mol})}\right]\right] \tag{5}$$

This loss effectively ensures that the cosine similarity between each matching pair, (z^n_{spec}, z^n_{mol}), is highest among all possible pairings, (z^n_{spec}, z^m_{mol}), within the batch.

2.3 Regularization

As candidate molecules typically have the same molecular formula as the target molecule, we fetch such candidates

from the PubChem database, and utilize regularization to train the model to better distinguish between target molecules and their candidates. Our regularization objective is therefore to push candidates away from the spectra, and hence from the corresponding target, in the joint-embedding space. Training using regularization is implemented by introducing an additional loss to minimize the cosine similarity between the embeddings of each spectrum and the candidate molecules of the corresponding target. Candidate sets are sorted by their Tanimoto similarity to their respective target molecule. For each spectrum within a training batch, we chose a set of candidates given by the batch parameter, k_{aug}. Candidates selected for regularization in each batch are therefore the k_{aug} most similar candidates, and are taken sequentially in each training epoch. Figure 1, available as supplementary data at *Bioinformatics* online, demonstrates the batching process for computing the regularization and contrastive losses. We then explored when and how to incorporate the regularization loss with our contrastive loss. Since our final ranking predictions are made using the molecular-spectral similarity in the joint space, the regularization attempts to push the most similar candidates away from the target molecule by minimizing a regularization loss function in addition to the contrastive loss. The regularization loss function minimizes the cosine similarity between the most similar candidates and the associated spectra—and hence the associated target molecules. The regularization loss, $L_{\text{regularization}}$, is defined as:

$$L_{\text{regularization}} = \frac{1}{k}\sum_{n=1}^{k}\frac{1}{k_{aug}}\sum_{m=1}^{k_{aug}}(cosine\_sim(z^{n}_{spec}, z^{m}_{cand})) \quad (6)$$

The total training loss, L_{total}, is the sum of the two losses weighted by hyper-parameters α and β:

$$L_{\text{total}} = \alpha * L_{\text{contrastive}} + \beta * L_{\text{regularization}} \quad (7)$$

We explored values for the α and β parameters, and we observed that utilizing regularization as a fine-tuning strategy toward the end of the training provided the best performance. Regularization was turned on for the last 3% of the training epochs. The weight given to regularization loss was 10% to ensure that the matching pairs are not pushed too far apart during regularization. Therefore, $\alpha = 1.0$, $\beta = 0.0$ for first 97% epochs and $\alpha = 0.9$, $\beta = 0.1$ for last 3% of epochs. Implementation details and various hyper-parameters are provided in Section 2, available as supplementary data at *Bioinformatics* online.

3 Results

3.1 Datasets

Four datasets were used to evaluate JESTR. The NPLIB1 dataset, or the CANOPUS dataset, was first utilized by the CANOPUS tool to predict compound classes, e.g. benzenoids, phenol ethers, and others, from spectra, thus providing partial annotation on spectra in cases when spec-to-spec comparisons in reference spectral databases yield unsatisfactory matches (Dührkop *et al.* 2021). This dataset was created by selecting spectra from the NIST2020, GNPS (Wang *et al.* 2016), and MoNA databases. The selection ensured a desired distribution of compound classes. This dataset was recently renamed to NPLIB1 (Goldman *et al.* 2023) to distinguish it from the CANOPUS tool. We utilized the NPLIB1 data as assembled by MIST (Goldman *et al.* 2023). This dataset comprised 8030 spectra measured under positive mode (positively charged, with an H adduct, [M+H]+) belonging to 7131 unique target molecules. We utilize the same data split as proposed by MIST, where the split was structure-disjoint such that a molecule with the same InChiKeys did not appear both in the training and test sets. Therefore, 714 molecules and their 819 spectra were utilized for testing. Two additional datasets were utilized to explore training JESTR on larger datasets. The NIST2020 dataset is well-curated spectral database released by the National Institute of Standards and Technology. NIST2020 comprises a variety of molecules from human, bacteria, environmental, plant, and food samples. A variety of instruments and settings are used to measure spectra for each compound. The measurements are repeated and a consensus spectrum is created for each measurement. The NIST datasets are available under a commercial license, and we had access to the NIST2020 version of this dataset. The MassBank of North America (MoNA) is a collaborative database, with contributions by users. Both experimental and *in silico* spectra are accepted. Here, we only retrieved the experimental dataset. Statistics for the three datasets is provided in Table 1. The number of unique molecules is largest in NIST2020, while the number of spectra per molecule is the lowest in NPLIB1. The splits for the NIST2020 and MoNA were created ensuring that no molecules overlapped between the training and test sets.

The fourth dataset is the recent MassSpecGym (Bushuiev *et al.* 2024b) dataset, which was developed as a benchmark set for candidate ranking, spectra simulation, and molecular generation tasks. The spectra consisted of a mix of H and Na adducts, with a total of 231 104 spectra for 32 010 molecules of high-quality labeled MS/MS spectra collected from public repositories such as MoNA, MassBank, and GNPS, as well as newly measured in-house data. The dataset includes 231 000 spectra corresponding to 29 000 unique molecular structures, making it the largest publicly available MS/MS benchmark. To ensure rigorous evaluation and prevent data leakage, the dataset is split using Maximum Common Edge Subgraph (MCES) distance, which clusters structurally similar molecules together in the same fold. Unlike InChIKey-based splits, this approach ensures that molecules in the test set have a significant structural difference from those in the training set. For the test split, an MCES distance threshold of >10 was used. Hence, the MassSpecGym test set is more challenging than all other test splits.

For all datasets except MassSpecGym, we select candidates for each target molecule in the training and test data by retrieving molecules from PubChem (Kim *et al.* 2019), by matching formulae of the target molecules. The average candidate sets for the target molecules range from 1322 to 2494 molecules, and for regularization, the average candidate sets for training molecules ranged from a 1390 to 2322 (Table 1). For the MassSpecGym dataset, we utilize the candidate sets provided with the released benchmark. Two types of candidates are provided: either based on mass, or on formula of the precursor ion. The maximum number of candidates for each test spectrum is 256.

3.2 Other annotation tools

We selected three recent annotation tools to provide a comparative evaluation for JESTR: ESP (Li *et al.* 2024), MIST

Table 1. Spectra, molecule, and candidate statistics for the three datasets.

	Total		Train				Test			
Dataset	Spectra	Molecules	Spectra	Molecules	Maximum # of candidates	Average # of candidates	Spectra	Molecules	Maximum # of candidates	Average # of candidates
NPLIB1	8030	7131	7211	6417	44 374	2220	819	714	25 929	2274
NIST2020	291 515	22 001	262 408	19 800	42 542	1390	29 107	2201	42 376	1322
MoNA	35 752	6767	32 216	6090	42 542	2322	3536	677	32 364	2494
MassSpecGym	231 104	32 010	194 119	28 840	48 184	2720	17 556	3170	256	185

(Goldman *et al.* 2023), and CMSSP (Chen *et al.* 2024). ESP and MIST are recent representative methods in the mol-to-spec and spec-to-mol categories, respectively, following the approach of explicit generation of intermediate forms, while CMSSP is the only model other than JESTR that matches spectrum and candidates directly in the embedding space. The ESP model utilizes a GNN-based molecular encoder and an MLP on the molecular FP. ESP is trained to learn a weighting between the molecular and FP representations to predict the spectra. The best ESP performing model on rank@1 was the version that utilized the FP and modeled peak co-dependencies, ESP MLP-PD. MIST first assigns chemical formulas to peaks within each spectrum using SIRIUS (Dührkop *et al.* 2019), and represents a spectrum as a set of chemical formulas. MIST trains a transformer model to learn peak embeddings and to predict FP. MIST also featurizes pairwise neutral losses and predicts substructure fragments as an auxiliary task. CMSSP is similar to JESTR in utilizing contrastive learning in a joint-embedding space, but utilizes a discriminator function based on the dot product.

To streamline the comparison using the same data splits for all datasets, we trained MIST and ESP on all datasets, and confirmed the results with the respective teams. We evaluated CMSSP on all test datasets using CMSSP's pretrained released weights. The released CMSSP was previously trained on MS/MS spectra from two public databases, GNPS and MassBank, and was supplemented by 1906 spectra independently acquired in-house. To ensure that there is no data leakage between the training set and test set molecules, we also publish results by retraining CMSSP on the same training-test split as was used when reporting performance for the other models. ESP was trained for 100 epochs with a learning rate of 0.001 and batch size of 32. Adam was used as the optimizer with L2 norm of 10^{-6}. MIST was trained for 500 epochs with a learning rate of 0.00057 and batch size of 32. For CMSSP, we evaluate the model using the released pretrained weights and also using retrained models on the specific datasets.

3.3 JESTR versus explicit-construction models

Given a query spectrum, the primary task of JESTR is to identify the target molecule among a set of candidates. Candidate ranking was therefore selected as the performance metric. The rank@1, rank@5, and rank@20 indicate the percentage of target molecules that were correctly ranked within the top 1, 5, and 20 candidates, respectively.

JESTR is compared against ESP and MIST (Table 2). For the NPLIB1 dataset, JESTR outperforms ESP and MIST on all reported ranks. For rank@1, JESTR outperforms ESP by 93.2%, and MIST by 64.1%. Further, JESTR achieves 95.8% rank@20, while the maximum performance of ESP and MIST is at 60.2% and 82.1%, respectively. We plot the detailed ranks for the NPLIB1 dataset (Fig. 4A). At all ranks, JESTR provides superior performance to both ESP and MIST. For the NIST2020 dataset, JESTR outperforms all other models. Specifically, JESTR outperforms MIST at rank@1 by 83.8%. Detailed ranks for NIST2020 is provided in Fig. 2A, available as supplementary data at *Bioinformatics* online. For the MoNA dataset, JESTR consistently outperforms ESP. MIST only outperforms JESTR at rank@1 by 17.6%, but not on rank@5, rank@20, or any other rank, as provided in Fig. 3A, available as supplementary data at *Bioinformatics* online. On the MassSpecGym dataset, JESTR achieves 15.62% rank@1 using candidates by mass, a 45.9% improvement over ESP and 6.7% improvement over MIST. Using candidates by formula, JESTR achieves 11.85% rank@1, surpassing ESP by 7.2% and MIST by 23.8%.

Examining the overall performance of JESTR against ESP and MIST, over the four datasets, on average for rank@[1–20], JESTR outperforms ESP by 55.5% and MIST by 56.6%. JESTR's performance was worse on the MoNA dataset when compared to NPLIB1 and NIST2020. We suspect that JESTR's performance on MoNA was low for two reasons. First, MoNA has the fewest number of molecules, thus providing the lowest molecular diversity among the three datasets. Second, MoNA data are uploaded by users and may not have undergone consistent curation efforts. The NIST2020 dataset is well-curated; however, it has the highest ratio of spectra per molecule, an average of 13.25 spectra per molecule, versus 1.13 and 5.28 spectra per molecule for NPLIB1 and MoNA, respectively. As such, we suspect that JESTR finds it hard to place all the spectra embeddings closer to the molecule in the joint space for NIST2020 and MoNA. A combination of high molecule-to-spectra ratio, lower diversity of molecules and inconsistent spectra curation makes JESTR perform the lowest on MoNA. MIST, with additional information in the form of subformulae annotation on peaks, does a better job at distinguishing among various spectra of the same molecule for MoNA, thereby attaining a better rank@1 score on this dataset. However, MIST loses its advantage over JESTR starting with rank@2.

As SIRIUS and CFM-ID are widely used within the community, we benchmark JESTR's performance against their publicly available pretrained models. To ensure a fair evaluation, we focus our evaluation on the MassSpecGym benchmark dataset using candidates by formula. Among the datasets, MassSpecGym offers the most equitable basis for comparison, as the distribution of MCES distances between the pretrained molecules in SIRIUS and CFM-ID and the test molecules in MassSpecGym closely matches that of the dataset's train/test split (Figs 6–8, available as supplementary data at *Bioinformatics* online). However, because some test molecules were included in the training of SIRIUS and CFM-ID, these models may benefit from an advantage not available to

Table 2. Ranking results for the NPLIB1, NIST2020, MoNA, and the MassSpecGym datasets.[a]

		Explicit-construction approaches		Implicit/matching approaches			
Dataset	Rank	ESP	MIST	CMSSP (retrained)	CMSSP (pretrained)	JESTR	$JESTR_{NR}$
NPLIB1	@1	23.69	27.96	13.92	**54.09**	45.76	41.06
	@5	44.69	62.39	32.97	67.40	**81.53**	81.12
	@20	60.20	82.05	49.08	74.48	95.77	**96.13**
NIST2020	@1	20.47	20.95	1.90	2.45	**38.62**	36.38
	@5	30.12	57.36	3.95	4.01	**59.64**	56.45
	@20	48.76	78.83	5.34	4.86	**81.03**	77.94
MoNA	@1	19.37	**32.30**	3.44	7.64	26.56	19.38
	@5	38.37	54.93	7.13	11.49	**64.21**	52.82
	@20	53.04	75.28	10.03	12.82	**91.16**	83.88
MassSpecGym candidates by mass	@1	10.71	14.64	2.60	3.85	15.13	**15.62**
	@5	24.84	34.87	5.50	5.44	36.75	**37.47**
	@20	42.66	59.15	8.70	7.71	60.32	**60.55**
MassSpecGym candidates by formula	@1	11.05	9.57	2.30	3.61	**11.85**	11.82
	@5	27.42	22.11	5.10	5.66	32.95	**33.48**
	@20	52.20	41.10	8.70	8.53	**61.46**	**61.46**

[a] Ranking performance of ESP, MIST, CMSSP, JESTR, and $JESTR_{NR}$, where regularization is removed. The numbers in bold are the best performance for that row.

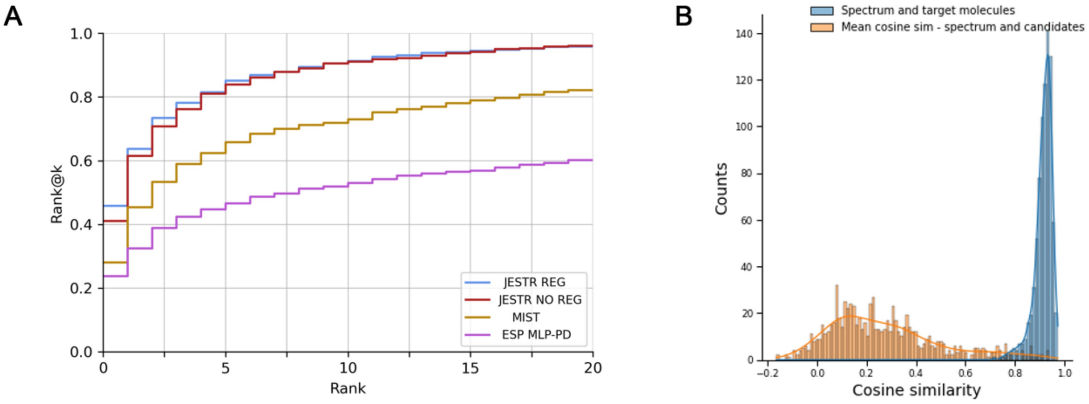

Figure 4. Results on NPLIB1. (A) Rank@k results for JESTR, with and without regularization, ESP MLP-PD, and MIST. (B) Distribution of cosine similarities of query spectra and target/candidate molecules with contrastive learning using JESTR.

JESTR. To address this, we report results on both the full test set and a filtered subset excluding any molecules seen during baseline model pretraining.

For SIRIUS, we create a custom database containing all candidates in the test dataset. We use the GUI version 6.1.0 and ensure the correct adduct and instrument setting for each spectrum is used. Out of 17 556 test spectra, SIRIUS assigned at least one structure to 7766 spectra, with 3342 of those with the target molecule ranked in position 1, i.e. a 19.04% rank@1 performance compared to JESTR at 11.85%. Excluding test spectra whose molecules were in SIRIUS's training set yields 1281 spectra correctly annotated out of 12 314 spectra, a 10.41% rank@1 versus 13.65% for JESTR. This result translates to a 31.12% improvement over SIRIUS, indicating that JESTR generalizes better than SIRIUS on unseen molecules.

For CFM-ID, per the authors' recommendation, we combine three spectra of low (<20 eV), medium (≥20 eV and <40 eV), and high (≥40 eV) collision energies per test molecule. Spectra are chosen randomly within the threshold, and we skip a spectrum of a specific collision energy threshold if there is not a spectrum whose collision energy falls within the threshold. Since JESTR is not designed to rank based on a set of spectra, we report the averages for best (lowest rank) and median performance across the chosen spectra. Out of 3170 merged spectra, CFM-ID correctly annotates 62 spectra, achieving 1.96% rank@1 versus 14.10% rank@1 based on the best performance and 6.56% rank@1 based on the median rank for JESTR. Removing test spectra whose molecules were in the training set yields 60 spectra correctly annotated out of 3004 spectra, a 2.00% rank@1 versus 14.61% rank@1 based on the best rank and 6.76% rank@1 based on the median rank for JESTR. Based on the median rank, JESTR provides a 238.00% improvement over CMF-ID.

The MassSpecGym benchmark dataset is notably more difficult than other datasets for all models, as molecules in the test set drastically differ from those in the training set. Nevertheless, JESTR's superior performance against explicit-construction models highlights the effectiveness of an implicit model using the joint-embedding space for ranking candidate molecules.

3.4 JESTR versus implicit models

Similar to JESTR, CMSSP learns a joint-embedding space for spectra annotation using contrastive learning. We report CMSSP's performance when training on individual datasets and using the released model weights, as training CMSSP on individual datasets shows low performance. JESTR

Table 3. Ranking result on the NPLIB1 dataset, contrasting the CMSSP architecture against the JESTR architecture with varying losses: cross-entropy (CE) and InfoNCE losses, and ranking with dot product (DP) and cosine similarity.

	CMSSP architecture	JESTR architecture		
Rank (%)	CE loss DP ranking	CE loss DP ranking	CE loss Cosine ranking	InfoNCE loss Cosine ranking
rank@1	13.92	3.20	10.84	**45.76**
rank@5	32.97	16.38	40.39	**81.53**
rank@20	49.08	37.68	67.12	**95.77**

The numbers in bold are the best performance in that row.

significantly outperforms CMSSP across all datasets except for rank@1 on the NPLIB1 dataset against the pretrained CMSSP model (Table 2). The strong performance of CMSSP likely results from similarities, or potentially data leakage, between our NPLIB1 test set and CMSSP training data.

We suspect CMSSP performed poorly when it was retrained because of its large model architecture, requiring the training of a large number of parameters. Specifically, CMSSP uses a GNN-based encoder for the molecule together with an MLP for the molecular FP while using a transformer-based encoder with eight attention heads for the spectra. Two other differences between CMSSP and JESTR is the loss function and the use of the dot product versus cosine similarity to evaluate embedding similarities. We evaluate the impact of these differences on the NPLIB1 dataset (Table 3). We first train the JESTR architecture with CMSSP's CE loss and the dot product. Performance drops substantially when swapping the CMSSP architecture for JESTR's architecture. Based on performance, the CMSSP architecture is more suited for using CE loss and the dot product when compared to the JESTR architecture. Maintaining the CE loss, we then evaluate the effect of swapping the cosine ranking to replace the dot product ranking. Using cosine similarity for ranking surpasses using the dot product and improves rank@5 and rank@20 performance over the CMSSP model by 22.5% and 36.8%, respectively, even though it was trained using the dot product. Cosine similarity explicitly normalizes the embedding magnitudes, yielding a competitive edge: similarity comparisons rely purely on embedding direction rather than magnitude. Embeddings with artificially large magnitudes are prevented from dominating similarity scores. InfoNCE with temperature-scaled cosine similarity is more suitable for a contrastive-learning framework, as the normalization from cosine similarity reduces the influence of magnitude differences and the temperature scaling allows for fine-grained control over the distribution of similarities. The combination of a simple architecture and InfoNCE with temperature-scaled cosine similarity provides JESTR a superior performance compared to CMSSP.

3.5 Joint-space embeddings distinguish target molecules from their candidates

The contrastive loss used in training JESTR ensures that the embeddings for matched spectrum–molecule pairs are placed close to each other in the joint-embedding space, while non-matching spectrum–molecule pairs are placed further away. Figure 4B shows the distribution on spectrum–molecule cosine similarities for matching and non-matching pairs in the NPLIB1 test set. The corresponding distributions for NIST2020 and MoNA datasets are shown in Figs 2B and 3B, available as supplementary data at *Bioinformatics* online, respectively. While any molecule other than the target can be considered a non-matching partner for the query spectrum, Fig. 4B only considers candidate molecules (same chemical formula as the target) as the non-matching partner. It is clear that JESTR well discriminates between target and candidate molecules.

3.6 Ablation study—removing regularization

To assess the value of regularization, the model was retrained without. Regularizing the training loss with molecular candidates improves rank@1 by 11.4%, 6.0%, and 37.1% on the NPLIB1, NIST2020, and MoNA datasets, respectively. Improvements using regularization are evident at almost all ranks and all datasets (Table 2). For rank@20 on NPLIB1, the results drop by 0.3% since the rank@20 result for NPLIB1 is high even without regularization, where even a small change in a few targets changes the result slightly. On the MassSpecGym dataset, regularization did not consistently boost model performance. Specifically, we observe a slight decrease in performance across all rank@k's when evaluating candidates by mass, and similarly, a modest reduction in rank@5 performance when using candidates by formula. We suspect that the challenging split within this dataset limits the generalizability benefits provided by regularization.

To further demonstrate the value of regularization, we performed additional analysis on the NPLIB1 dataset. With regularization, the number of target molecules ranked @1 increases significantly, from 301 to 375, causing a ripple effect in improving other rank@k numbers (Fig. 5A). Further, we examined the Tanimoto similarity between molecules in the training set and their candidates. We retrieved 15.86 million candidates from PubChem based on the chemical formulas of the target molecules in the training set. The majority of these candidates show low Tanimoto similarity with the target molecules (Fig. 5B). Hence, our fine-tuning regularization strategy and sorting the candidates by their cosine similarity to the target effectively prioritizes regularization with the most similar candidates. We approximately utilized 7 million candidates for regularization during the last 3% of training epochs. Upon examining the cosine similarity distributions on the embeddings of candidate and target molecules (Fig. 5C), we see that our regularization strategy reduces the average cosine similarity between targets and their candidates. Regularization is therefore effective, enabling the model to discriminate between a target molecule and its candidates. Similar analysis for the NIST2020 and MoNA datasets is shown in Figs 4 and 5, available as supplementary data at *Bioinformatics* online.

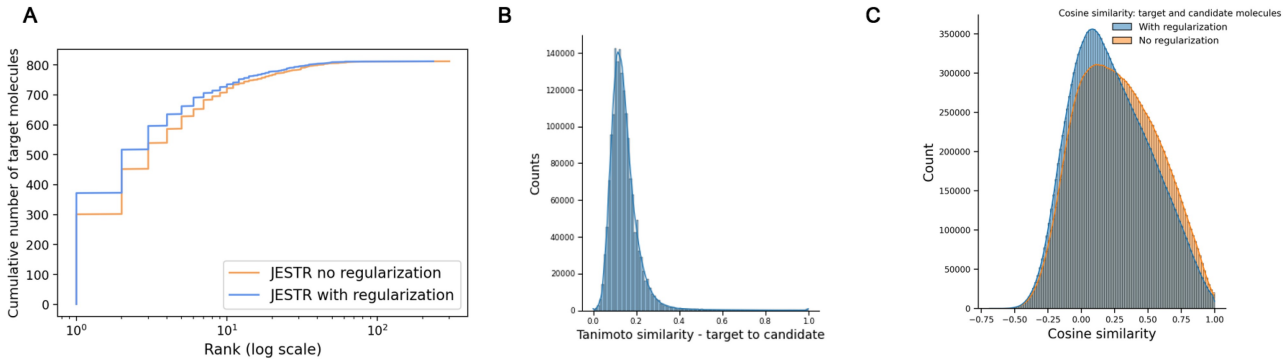

Figure 5. Regularization analysis for JESTR for NPLIB1. (A) Regularization improves rank@k by significantly placing more targets at rank 1. (B) Distribution on Tanimoto similarities on the ECFP fingerprints between target and candidates in the training set. (C) Distribution on cosine similarities, with and without regularization, of the target and candidates within the test set.

4 Conclusion

JESTR offers a novel implicit annotation paradigm that avoids the explicit generation of spectra, FPs, or molecular structures. As molecules and spectra are views of the same object, embedding these views in a joint-embedding space using contrastive learning provides a performance advantage. Evaluation across diverse datasets, such as NPLIB1, NIST2020, MoNA, and MassSpecGym, demonstrates clear performance advantages over current explicit annotation techniques, highlighting the robustness of implicit models using the joint-embedding approach. Moreover, investigation on the contrastive loss used by JESTR reveals that InfoNCE with temperature-scaled cosine similarity is advantageous for contrasting views and learning to distinguish matching pairs from non-matching pairs. Analysis of JESTR's performance on these datasets reveals that dataset diversity, quality, and spectra-to-molecule ratios impact performance. Further, our results showed enhanced value in utilizing candidate molecules during training for regularization, improving performance by an average of 5.72% for all datasets. Nonetheless, the minimal improvement in performance with regularization on the MassSpecGym benchmark suggests that the effectiveness of regularization may vary depending on data splits and generalization difficulty between the train and test set. The overall JESTR results are promising and vouch for the potential of implicit annotation approaches. We expect to attain further improvements by utilizing additional knowledge in the form of subformulae annotation on spectral peaks and by utilizing enhanced molecular and spectral encoders.

Author contributions

Apurva Kalia (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Software [equal], Validation [equal], Writing—original draft [equal], Writing—review & editing [equal]), Yan Zhou Chen (Data curation [equal], Software [equal], Validation [equal], Visualization [supporting], Writing—review & editing [lead]), Dilip Krishnan (Conceptualization [supporting], Supervision [supporting], Writing—original draft [supporting]), and Soha Hassoun (Conceptualization [equal], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Supervision [lead], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data is available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM148219. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data availability

Please the see the github repository for this project for data availability.

References

Bushuiev R, Bushuiev A, Samusevich R *et al.* Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat Biotechnol* 2025. https://doi.org/10.1038/s41587-025-02663-3

Bushuiev R, Bushuiev A, de Jonge N *et al.* MassSpecGym: a benchmark for the discovery and identification of molecules. *Adv Neural Inf Process Syst* 2024b;**37**:110010–27.

Butler T, Frandsen A, Lightheart R *et al.* Ms2mol: a transformer model for illuminating dark chemical space from mass spectra. ChemRxiv, https://doi.org/10.26434/chemrxiv-2023-vsmpx-v3, 2023, preprint: not peer reviewed.

Chen L, Xia B, Wang Y *et al.* CMSSP: a contrastive mass spectra-structure pretraining model for metabolite identification. *Anal Chem* 2024;**96**:16871–81.

Chen T, Kornblith S, Norouzi M *et al.* A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*. PMLR. 2020, 1597–607.

Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. **1**. San Diego, CA, USA: IEEE, 2005, 539–46.

Del Carratore F, Schmidt K, Vinaixa M *et al.* Integrated probabilistic annotation: a Bayesian-based annotation method for metabolomic profiles integrating biochemical connections, isotope patterns, and adduct relationships. *Anal Chem* 2019;**91**:12799–807.

Del Carratore F, Eagles W, Borka J *et al.* ipaPy2: integrated probabilistic annotation (IPA) 2.0—an improved Bayesian-based method for the annotation of LC–MS/MS untargeted metabolomics data. *Bioinformatics* 2023;**39**:btad455.

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.

Dührkop K, Nothias L-F, Fleischauer M *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat Biotechnol* 2021;**39**:462–71.

Faizan-Khan M, Giné R, Badia JM *et al.* ChemEmbed: a deep learning framework for metabolite identification using enhanced MS/MS data and multidimensional molecular embeddings. bioRxiv, https://doi.org/10.1101/2025.02.07.637102, 2025–02, 2025, preprint: not peer reviewed.

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.

Heinonen M, Shen H, Zamboni N *et al.* Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* 2012;**28**:2333–41.

Hosseini R, Hassanpour N, Liu L-P *et al.* Pathway-activity likelihood analysis and metabolite annotation for untargeted metabolomics using probabilistic modeling. *Metabolites* 2020;**10**:183.

Huber F, Ridder L, Verhoeven S *et al.* Spec2vec: improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput Biol* 2021;**17**:e1008724.

Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 2018;**58**:27–35.

Khosla P, Teterwak P, Wang C *et al.* Supervised contrastive learning. *Adv Neural Inf Process Syst* 2020;**33**:18661–73.

Kim S, Chen J, Cheng T *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res* 2019;**47**:D1102–9.

Kind T, Tsugawa H, Cajka T *et al.* Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom Rev* 2018;**37**:513–32.

Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv, arXiv:1609.02907, 2016 , preprint: not peer reviewed.

Li S, Park Y, Duraisingham S *et al.* Predicting network activity from high throughput metabolomics. *PLoS Comput Biol* 2013;**9**:e1003123.

Li X, Zhou Chen Y, Kalia A *et al.* An ensemble spectral prediction (ESP) model for metabolite annotation. *Bioinformatics* 2024;**40**:btae490.

Litsa EE, Chenthamarakshan V, Das P *et al.* An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun Chem* 2023;**6**:132.

Martin MR, Bittremieux W, Hassoun S. Molecular structure discovery for untargeted metabolomics using biotransformation rules and global molecular networking. *Anal Chem* 2025;**97**:3213–9. https://doi.org/10.1021/acs.analchem.4c01565

Radford A, Kim JW, Hallacy C *et al.* Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*. PMLR. 2021, 8748–63.

Ruttkies C, Schymanski EL, Wolf S *et al.* MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J Cheminform* 2016;**8**:3–16.

Stravs MA, Dührkop K, Böcker S *et al.* MSNovelist: de novo structure generation from mass spectra. *Nat Methods* 2022;**19**:865–70.

Tian Y, Zhang Y. A comprehensive survey on regularization strategies in machine learning. *Inf Fusion* 2022;**80**:146–66.

Tian Y, Krishnan D, Isola P. Contrastive multiview coding. In: *Proceedings of the 16th European Conference on Computer Vision–ECCV 2020, Glasgow, UK, August 23–28, 2020, Part XI 16*. Berlin, Germany: Springer, 2020, 776–94.

Wang F, Liigand J, Tian S *et al.* CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal Chem* 2021;**93**:11692–700.

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.

Wei JN, Belanger D, Adams RP *et al.* Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Cent Sci* 2019;**5**:700–8.

Wolf S, Schmidt S, Müller-Hannemann M *et al.* In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* 2010;**11**:148–12.

Young A, Röst H, Wang BO. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat Mach Intell* 2024;**6**:404–16. https://doi.org/10.1038/s42256-024-00816-8

Zhang H, Yang Q, Xie T *et al.* MSBERT: embedding tandem mass spectra into chemically rational space by mask learning and contrastive learning. *Anal Chem* 2024;**96**:16599–608.

Zhu H, Liu L, Hassoun S. Using graph neural networks for mass spectrometry prediction. arXiv, arXiv:2010.04661, 2020, preprint: not peer reviewed.

Bioinformatics, 2025, 41, 1–10
https://doi.org/10.1093/bioinformatics/btaf354
Original Paper