Chemical Space Localization for Unknown Metabolite Annotation via Semantic Similarity of Mass Spectral Language

Hongchao Ji[1,2], Ran Du[1,2], Qinliang Dai[1], Meifeng Su[1], Yaqing Lyu[1], Jianbin Yan[1, *]

1. Shenzhen Branch, Guangdong Laboratory for Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Genome Analysis Laboratory, Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen, 518120, China PR.

2. These authors contributed equally.

*Correspondence: Jianbin Yan (jianbinlab@caas.cn)

ABSTRACT

Untargeted metabolomics, employing liquid chromatography–mass spectrometry (LC-MS), allows quantification of known and unidentified metabolites within biological systems. However, in practical metabolomic studies, the majority of metabolites often remain unidentified. Here, we developed a novel deep learning-based metabolite annotation approach via semantic similarity analysis of mass spectral language. This approach enables the prediction of structurally related metabolites for unknown compounds. By considering the chemical space, these structurally related metabolites provide valuable information about the potential location of the unknown metabolites and assist in ranking candidates obtained from molecular structure databases. Validated with benchmark datasets, our method has consistently demonstrated superior performance compared to existing methods for metabolite annotation. In a case study involving the *Qianxi* cultivar cherry tomato, our approach reaffirmed well-established biomarkers of ripening processes and identified a set of promising and rational new biomarker metabolites. Overall, presented method exhibits significant potential in annotating metabolites, particularly in revealing the "dark matter" in untargeted metabolomics.

Mass spectrometry (MS) plays a crucial role in the analysis of the diverse small-molecule content of complex systems, particularly metabolic systems. It contributes significantly to chemobiological research, including research on genetics, evolution, and biosynthesis[1]. The main challenge in metabolite annotation is the limited coverage and accuracy[2]. The molecular fragment information provided by tandem mass spectrometry (MS/MS) serves as the foundation for metabolite structural identification. The most commonly used strategy for structural identification is to match the detected metabolite MS/MS spectra with standard spectral libraries such as GNPS and NIST. Improvements in library-searching algorithms, such as MS-Entropy and MS2Query enhanced the efficacy and efficiency of molecular annotation[3–5].

However, this strategy is limited by two aspects: firstly, variations in instrument types, conditions, and parameters can lead to different MS/MS spectra for the same metabolite, which reduces the universality of standard spectral libraries and consequently decreases the number of identified known metabolites in practical research[6]. Secondly, the coverage of standard spectral libraries is far smaller than the potential number of metabolites, and the absence of standard spectra makes it nearly impossible to identify novel metabolites through library matching[7]. Therefore, the "dark matter" problem remains unaddressed, resulting in over 90% of chemical signatures remaining uncharacterized in untargeted plant metabolomics studies[8]. Molecular networking can reveal the chemical neighbors of unknown metabolites and estimate their class but cannot predict their specific structures automatically[9–11].

Therefore, employing computational methods predicting chemical structures from the spectra excluding from spectral database is the most promising approach for improving coverage[12,13]. Currently, the published methods can be mainly divided into in silico annotation algorithms[7,14–19] and known-to-unknown extension through pathways[20–22]. The in-silicon annotation algorithms commonly include two steps. First, they find candidates from structural database[23] based on the chemical formula. The chemical formula can be inferred from the precursor masses, ionization types and isotopic pattern, or through specialized toolkits such as BUDDY[24] or MIST-CF[25]. Second, they assess and prioritize potential candidates based on specific rules. For example, SIRIUS scores candidates by comparing the predicted fingerprints of an unknown compound with the calculated fingerprints of potential candidates[15]. MS-Finder scores candidates by considering the

frequency of hydrogen rearrangement inferred from the fragment ions[16]. CFM-ID evaluates candidate compounds based on the accessibility of the inferred substructural motif from the fragment ion[17].

Here, we introduce DeepMASS v2, an enhanced iteration of our previous work. This in-silico metabolite annotation GUI software features a novel candidate assessment strategy (still referred to as DeepMASS in the following sections for brevity). DeepMASS can be seen as both a competitor and a complement to existing tools, such as SIRIUS, MS-Finder, and CFM-ID, with different strategies. These tools share a common objective in plant metabolomics studies and can be utilized together by researchers for comparison, supplementation, or reference purposes. This iteration is characterized by predicting structurally related compounds by assessing the semantic similarity in their "mass spectral language" with deep learning and by pinpointing their location within the chemical space. By comparing the predicted chemical spatial position with that of potential candidates, it assigns scores and ranks these candidates.

This paper primarily focuses on the methodological advancement of DeepMASS, while also encompassing validation and application aspects. It is structured as follows: First, we introduce the workflow and principles underlying DeepMASS. Second, we evaluate and compare its annotation performance with rival tools using benchmark datasets. Third, we apply DeepMASS in a specific context, demonstrating its utility in assisting plant metabolomics studies. Finally, we provide an overview of the software's graphical interface and detail its instructions.

RESULTS

Workflow and principles of DeepMASS

DeepMASS is situated within the metabolomics process following the data preprocessing of raw files. At this stage, metabolomic features along with their corresponding mass spectra are exported by data preprocessing software such as XCMS[26], MZMine[27], OpenMS[28], KIPC2[29] or MS-DIAL[30]. DeepMASS initiates by accepting the spectrum of features and proceeds to retrieve structural candidates of the unknown metabolites from a comprehensive bio database[18]. If the chemical formula has already been assigned by BUDDY[24] or MIST-CF[25], the retrieval process is based on the chemical formula. If the formula is unavailable, it alternatively retrieves candidates based on the monoisotopic mass derived from precursor masses and ionization types. All candidates with a monoisotopic mass difference within a specific threshold (10 ppm as default) are considered.

Then, DeepMASS utilizes a trained Word2Vec-based semantic similarity model embedding the unknown spectrum into a numeric vector. A rapid hierarchical navigable small-world graph (HNSW)[31]-based search is conducted with the reference database, which comprises standard spectra embedded with the same model. This searching enables to find the neighbor spectra, which aims to find the structurally related neighbor metabolites of the unknown. All the neighbor metabolites encompass the potential chemical space of the unknown metabolite. Previously, the Word2Vec-based semantic similarity model was employed by Spec2Vec to generate a spectral similarity score, drawing inspiration from natural language processing algorithms. It has been demonstrated that Spec2Vec scores exhibit a stronger correlation between spectral similarity and structural similarity compared to cosine-based scores. However, Spec2Vec is tailored for library matching and molecular networking, and it lacks the capability to identify unknown metabolites that are not present in the spectral database[32,33].

Subsequently, the candidates retrieved from the bio database are compared with the predicted neighbor metabolites in the chemical space. Their spatial positions in the chemical space are determined using Morgan molecular fingerprints, and their relative spatial distances are measured based on the Dice similarity between the fingerprint vectors. It is expected that the correct candidate for the unknown metabolite will be surrounded by several predicted neighbor metabolites. Based on this assumption, a similarity scoring function outlined in the Method section is employed to rank the candidates, aiding in the annotation of the unknown metabolite (**Fig. 1**).

Performance evaluation of DeepMASS

We assembled a benchmarking dataset comprising 499 spectra associated with 478 unique metabolites obtained from the Critical Assessment of Small Molecule Identification (CASMI) for the years 2014, 2016, and 2022. Notably, CASMI 2017 was excluded due to its utilization of MS^E mode, as all the spectra of the reference dataset are from DDA mode. Within this set, 189 spectra correspond to metabolites present in the reference dataset. However, it's important to note that, in order to prevent data leakage, we have excluded all spectra from the training data that exhibit a cosine similarity exceeding 0.95 when compared to any spectrum in this benchmarking dataset. Statistically, 401 out of 499 spectra (80.3%)

in the dataset, along with their corresponding metabolites, are present in the bio database. This is the theoretical upper limit of the annotation performance.

Firstly, we evaluated the annotation accuracy of DeepMASS by directly comparing with the state-of-art in silico metabolite annotation software tools SIRIUS[15] and MS-Finder[18], along with the cosine-based and Spec2Vec-based library matching aided by MatchMS[34]. The reference spectral database for MatchMS was the same as that for DeepMASS. As shown in **Fig. 2A**, among all the test spectra, DeepMASS attained a Top 1 accuracy of 50.7%, signifying that it correctly placed 248 true annotations in the top position, in contrast to 44.7% for SIRIUS, 40.1% for MS-Finder, 32.3% for cosine-based library matching, 36.5% for Spec2Vec-based library matching. Upon expanding the evaluation to the Top 10 annotations, DeepMASS, SIRIUS, and MS-Finder exhibited improved performance, with annotation rates of 69.9%, 62.1%, and 63.1%, respectively. However, the improvement in the performance of library matching methods were not significant because a large portion of the metabolites corresponding to the test spectra were not included in its reference database.

Moreover, we separately assessed the annotation performance of the spectra when matched metabolites were present inside and outside the reference data. Out of the 189 spectra that included metabolites from the reference dataset, cosine-based and Spec2Vec-based library matching demonstrated superior performance over the in-silico annotation software tools, achieving a Top 1 accuracy of 85.2% and 89.9%. SIRIUS and MS-Finder attained Top 1 accuracy rates of 75.1% and 70.3%, respectively, which are also higher compared with the compounds absent from the reference dataset. This can be attributed to the fact that these metabolites are more prevalent and easily identifiable by their inherent characteristics. Excitingly, DeepMASS achieved even greater accuracy than cosine-based library matching, boasting Top 1 and Top 10 accuracy rates of 88.4% and 97.9%, respectively.

For the metabolites not included in the reference dataset, library matching was unable to provide correct annotations for any of them. Furthermore, the performance of SIRIUS and MS-Finder also experienced a substantial decline, yielding Top 1 accuracy rates of 26.1% and 21.6% and Top 10 accuracy rates of 41.2% and 45.8%, respectively. In contrast, DeepMASS achieved Top 1 and Top 10 accuracy of 27.7 % and 52.9%. Both of them are higher than either SIRIUS or MS-Finder. Compared with rival tools, DeepMASS offers higher accuracy and clear interpretability. Furthermore, it operates without requiring any prior information unrelated to spectral measurement, such as referred frequency in research papers and bio databases. It reduces the potential bias toward commonly annotated metabolites[35].

Secondly, we demonstrated the interpretability of the approach through two example of annotation results. They can illustrate how DeepMASS arrives at its annotation decisions (**Fig. 2B)**. We employ UMAP dimensionality reduction[36] on the Morgan fingerprints of both candidate and neighbor compounds to visualize their relationships within the chemical space. In the scatter plots, at least three of the Top 10 neighbor structures are situated near the correct annotation. These neighbor structures play a significant role in locating the chemical space of the spectrum to be annotated, thereby assisting DeepMASS in making the decision. These neighbor structures are subsequently compared to the correct structure, and their shared substructures are highlighted with a red background color.

In the case of the 6-bromoquinolin-2(1H)-one spectrum in the positive mode, the first neighbor spectrum corresponds to 6-chloroquinolin-2-ol, which exhibits losses similar to those of 6-bromoquinolin-2(1H)-one. The other two neighbor spectra correspond to 2,6-dibromoquinoline and 1,4-dibromoisoquinoline. These compounds share a highly similar carbon skeleton with 6-bromoquinolin-2(1H)-one, leading to the presence of similar spectral fragments. In the case of the N-acetyl-5-aminosalicylic acid spectrum in negative mode, the top two references originate from spectra with identical structures, resulting in duplicate scatter points with the true annotation. Nevertheless, additional references serve to further validate the position in the chemical space, specifically acetaminophen and paracetamol. Both of these references share a significant portion of the same substructure as N-acetyl-5-aminosalicylic acid. These additional references show why DeepMASS performs even better than spectrum matching when annotating spectra present in the database. Obviously, the decision-making process of DeepMASS is comprehensible and aligns consistently with theoretical expectations. The utilization of neighbor compounds predicted by the semantic similarity of mass spectral language is effective in positioning the chemical space of unannotated metabolites.

Finally, DeepMASS is also supposed to have the potential to identify unknown compounds which have never discovered before. The restricted condition is the possible structures (candidates) are predicted and have been added to the database. A promising application identifying natural product derivatives. As a forward-looking example, three compounds beyond current libraries were synthesized. These derivatives of indole were synthesized strictly according to the protocol of the literature in which they were published[37]. Deservedly, their spectra are not included in spectral

databases. Therefore, their structures cannot be identified by any spectral matching methods. Hence, we measured their mass spectra with ESI-FTMS instrument, and input the results to DeepMASS. The identification results are shown in **Fig. S1**, accompanied with the Top 3 neighbors and number of candidates. Both corrected structure of C1 and C3 are ranked at the first position, and C2 is ranked at the third position.

Case study for tomato ripening processes

To explore the practical utility of DeepMASS, we conducted a metabolomic analysis of the *Qianxi* cultivar cherry tomato during ripening processes (**Fig. 3A**). For the DeepMASS pipelines, data preprocessing and MS/MS extraction were conducted through MS-DIAL[30]. In total, 21,601 features assigned with MS/MS were obtained for the positive ion mode, and 18,846 features were obtained for the negative ion mode. DeepMASS successfully annotated 20,140 features from the positive mode data and 17,776 features from the negative mode data, accounting for 93.2% and 94.3% of the total, respectively. For comparison, TraceFinder commercial software was employed in the contrasting pipeline. It managed to acquire 7,087 metabolite features with assigned chemical formulas and MS/MS peaks for the positive ion mode and 6,943 metabolite features for the negative ion mode. However, it could only annotate 21.7% of the features in the positive ion mode and 14.0% in the negative ion mode with library matching.

Furthermore, the number of unique metabolites annotated by DeepMASS amounted to 8,704 from positive mode data and 8,824 from negative mode data, showcasing a marked increase compared with TraceFinder, the corresponding values from which were 1,531 and 970, respectively. These findings demonstrate that DeepMASS has the capability to significantly enhance the identification of MS features and the annotation of compounds in metabolomic analyses (**Fig. 3B**).

Then, we conducted a statistical analysis to compare the changes in the metabolomes from different ripening stages of *Qianxi* cherry tomatoes using the annotations obtained from DeepMASS as the basis of comparison alongside results obtained from the commercial software. To assess metabolite differences among various ripening stages, we first employed principal component analysis (PCA). As visualized using a score plot in **Fig. 3C**, based on the DeepMASS results, it was easy to distinguish fruit samples from different ripening stages completely, whereas it is less easy to effectively distinguish samples from different ripening stages based on the annotations from the commercial software. Furthermore, a continuous change trend was observed in the principal component space of the scatters along with fruit ripening. This suggests that the extra metabolites annotated by DeepMASS were valuable in elucidating the metabolic changes occurring during the ripening process of the *Qianxi* cultivar cherry tomato.

Subsequently, we performed partial least squares discriminant analysis (PLS-DA) to identify metabolites within the DeepMASS annotations, focusing on those with variable importance (VIP) scores surpassing 2. Notably, many of these metabolites have previously been linked to attributes such as resistance and fruit quality[38–40]. We then examined how these metabolites exhibited variation trends across ripening stages. Notably, we observed a gradual enrichment in the abundance of metabolites related to flavor and nutrients, including glucose, norfuraneol, glutamine, citric acid, and aspartic acid (**Fig. 4A**). These trends align with the known variations in flavor and nutrient-related metabolites during the ripening process. We noticed a decline in the abundance of metabolites annotated as related to resistance but considered undesirable for consumers, such as solasodine, tomatidine, chlorogenic acid-like compounds, and hydroxycoumarin-like compounds, particularly in the later ripening stages, which was consistent with the observed patterns of these compounds in the cherry tomato cultivar. To validate these findings, we compared the detected spectra with the spectra obtained using standard substances. Additionally, we extracted the XIC of these metabolites and confirmed the trends with their chromatographic peaks (**Fig. S2**). These findings underscore the advantages of the DeepMASS approach in obtaining a more comprehensive and accurate metabolome compared to commercial software methods. This enhanced capability significantly improves the discovery potential for metabolites of interest.

Finally, to investigate the potential capability for the discovery of new biomarker metabolites related to ripening, we examined the change patterns of metabolites annotated by DeepMASS across the five ripening stages. Through application of the K-means clustering method, we obtained 37 clusters in the negative mode results and 40 clusters in the positive mode results. As shown in **Fig. 4B**, the change trends of annotated metabolites across the five ripening stages exhibited significant differences, with the MG and Br3 stages showing similar trends, while the Br7, Br10, and Br15 stages displayed notable variations. Furthermore, a substantial number of metabolite clusters exhibited significant enrichment during ripening, suggesting a close association between these metabolite clusters and the development of flavor and commercial qualities in *Qianxi* cherry tomatoes. We then selected clusters with a gradual increase in trends

during the early ripening process from the top five clusters in both positive and negative modes. These clusters included cluster 32 in the positive mode results, as well as clusters 2 and 6 in the negative mode results (**Fig. 4C**).

The results indicate that within these three selected clusters, we not only identified plant hormones related to fruit ripening, such as gibberellic acids (interlinked regulatory loops of ABA catabolism and biosynthesis that coordinate fruit growth and ripening in woodland strawberry) and traumatic acid (involved in the effect of lipoxygenase and jasmonic acid on ethylene biosynthesis during tomato fruit ripening) but also flavor and nutrient metabolites, such as sugars and furaneols. Moreover, beyond well-known ripening-related compounds, we also identified flavor compounds or precursors that have been less frequently reported in tomato fruit, such as 1-phenyldodecan-1-one-like[41], 2-furancarboxaldehyde-like[42], hydroxypropionic acid-like[43], and 1-octen-3-yl glucoside-like[44] metabolites. These compounds may play significant roles in flavor formation in ripe fruits.

Additionally, the results revealed that these three clusters also contained a series of lipids, including phospholipids such as phosphatidic acid (PA), phosphatidylcholine (PC), phosphatidylethanolamine (PE), diacylglycerol (DG), and lipid hydroperoxide (LPO). Given the importance of lipids in texture-related metabolites, these compounds could serve as biomarkers and key contributors to the formation of essential commercial qualities.

GUI and software instruction

We developed DeepMASS as a cross-platform GUI software tool, with the front end implemented by the QT framework and the back end implemented by the Python programming language. The software has been tested on Windows 11 and MacOS. As depicted in **Fig. 5**, the user interface is straightforward and user-friendly. For the sake of convenience and broad applicability, upon opening the software, the semantic model and reference dataset are automatically loaded. The reference dataset in the public version exclusively encompasses spectra from GNPS, but we offer scripts to assist users in incorporating their own in-house data and NIST data if they possess the necessary permissions.

Unidentified spectra are uploaded in batches as *mgf* format files. DeepMASS autonomously recognizes and obtains the metadata for each spectrum. For instance, the ion mode dictates which model and reference dataset should be used for annotation. If the formula is included in the metadata, DeepMASS retrieves candidates with that specific formula; otherwise, it retrieves candidates based on the parent mass. The annotation results can be saved as an Excel file by clicking the *save* button.

The display panels consist of three sections: *spectrum*, *structure*, and *information*. Within the *spectrum* panel, a visual comparison is presented between the unidentified spectrum and the reference spectrum. Users have the option to choose the reference spectrum for comparison by clicking the corresponding item in the reference spectrum list widget. This feature helps users observe how DeepMASS reaches its annotation decisions and assesses the reliability of the annotations. Within the *structure* panel, a comparison is made between the structure of the reference compound and the candidate annotation. Through highlighting shared substructures, users can access additional explanatory information about the annotation. The *information* panel is used to show the meta information of the input spectrum. In general, DeepMASS is a user-friendly and accessible software tool for metabolite annotation that is designed for users who may not have programming skills.

DISCUSSION

In this study, we introduce DeepMASS as an enhanced iteration of our previous work[7]. The primary advancement is the strategy changing for finding the neighbor structures. The original version, which depended on constructing metabolite pairs and utilizing a linear fully connected neural network for predicting structural similarity based on their binned spectra, can only handle hundreds of spectra in the case of combination explosion. In this version, with the incorporation of semantic similarity of the mass spectral language, metabolite pairs are not essential any more. Therefore, the number of spectra for training increase from 752 to approximate 780,000, which is a three orders of magnitude improvement. The enhanced DeepMASS boasts the capability exhibits strong scalability, thus expanding its potential applications to fields such as plant and microbial metabolomics.

Our positioning of DeepMASS is a competitive and collaborative program of SIRIUS and MS-Finder in MS/MS-based compound identification. They share common objective of structural annotation for metabolites, irrespective of the availability of corresponding spectra. This strategic integration addresses the limitations in coverage observed in standard spectral databases, ensuring a more comprehensive approach to metabolite identification. This goal makes them

distinguished from enhanced spectral matching software such as MatchMS or MS-Entropy, as well as from spectral embedding algorithms for numerical representation and spectral network establishment, like Spec2Vec or MS2DeepScore.

With the benchmark datasets collected from CASMI 2014, 2016 and 2022, DeepMASS achieves higher annotation accuracy than the competitive software. Considering that some of the test spectra have corresponding reference spectra in the spectral database, we evaluate the performance separately for spectra that are within the database and those that are not. As expected, both DeepMASS and the competitive software exhibit significantly better performance when using spectra already present in the database. However, only DeepMASS demonstrates comparable, and in some cases, even superior, performance when compared to the spectral matching strategy. This superiority is significant because, in practical applications, we often do not have prior knowledge of whether the spectra to be annotated are present in the database or not. When researchers use spectral matching and in silico annotation independently, the outcomes can diverge due to their distinct coverage of the chemical space. While spectral matching might excel in accuracy for common metabolites, it can introduce bias and hinder the discovery of new metabolites. DeepMASS circumvents this predicament by more effectively leveraging the reference database compared with the other in silico annotation software. When the metabolite is already present in the database, it serves as the optimal choice for localizing the chemical space, and the other neighbors can complement and enhance its performance. This approach yields results that surpass those obtained by using reference spectra in isolation.

On the other hand, in cases where the spectra lack corresponding metabolites within the reference database, the predicted neighbors based on spectral similarity with the unknown spectra play a pivotal role in capturing the chemical space and evaluating potential candidates. Since spectral similarity is a reflection of structural similarity, this prediction is logically justified. Moreover, given that diverse functional metabolites typically stem from a restricted set of initial metabolites through biochemical reactions, there exists a reasonable probability that reference neighbors might lie upstream or downstream of the unknown metabolite within metabolic pathways. This circumstance enhances the likelihood of accurate annotations.

To evaluate the application value in an actual metabolomics study, we conducted a metabolomics analysis of the *Qianxi* cultivar cherry tomato during ripening processes and found that DeepMASS achieved 5.69 to 9.10 times more annotated metabolites than the traditional method based on commercial software from raw MS profiles, which significantly increased the utilization rate of raw data and effectively reduced data bias in subsequent analysis. In addition, from the annotated results, we also found a series of metabolites that might contribute to tomato fruit flavor and commercial quality trait formation during ripening.

In summary, metabolomic analysis of these tomato fruit samples confirmed that DeepMASS not only substantially enhanced the extraction of features and the annotation of metabolites from raw mass spectral data but also did so with high reliability. This underscores its utility as a potent tool for the discovery of new metabolite biomarkers in various biological contexts.

Looking ahead, DeepMASS shows promise in effectively identifying undiscovered compounds, on condition that the possible structures are predicted and added into the database. A promising showcase of identifying natural product derivatives is presented in the supplementary information.

The limitation of the current version of DeepMASS is its strong dependence on both the quantity and quality of spectra within the reference database. This reliance is crucial for identifying reasonable and accurate neighbor metabolites that assist in pinpointing the chemical space of the unknown metabolite. The expansion of spectral data further enhances the precision of the connection between structural similarity and semantic similarity within the mass spectral language.

Overall, DeepMASS demonstrates commendable annotation performance in both benchmark evaluations and real-world data analysis. It offers meaningful and valuable metabolite annotations, features user-friendly and extensible software, and holds promise for even greater potential with the incorporation of more reliable data in the future.

METHODS

Construction of the bio database

The bio database is employed within DeepMASS's annotation process to retrieve structural isomers for a specified formula. Our bio database is constructed upon the foundation of metabolome databases also utilized by MS-Finder. This comprehensive database comprises sixteen different sources, which are ChEBI, NIST, KNApSAcK, GNPS, HMDB, DrugBank, BloodExp, ECMDB, FooDB, NANPDB, SMPDB, STOFF, T3DB, YMDB, PlantCyc and UNPD. As a result, 573,444 unique compounds are included.

Semantic model training

Three distinct datasets were utilized to train the word2vec-based semantic model. Initially, the open-source GNPS dataset, acquired as of June 1st, 2022, comprised 495,810 MS/MS spectra. Second, the commercial NIST 20 software contributed 1,026,506 MS/MS spectra. Last, our in-house dataset added an additional 23,610 MS/MS spectra. A publicly available version excluding the NIST 20 training set is provided due to commercial licensing constraints. To avoid data leakage, we eliminate all spectra with a cosine similarity greater than 0.95 compared to any spectrum in the benchmarking data.

The spectra underwent a cleansing and correction process facilitated by the MatchMS package[34], including the following procedures: (1) extract structural information (SMILES, InChI, InChIKey) from the metadata, and discard entries for which this information could not be derived; (2) extract information related to precursor type (adduct, charge) from the metadata, retaining only the spectra with adduct type of M+H, M+Na, M+K, and $M+NH_4$ for positive mode and M-H and M+Cl for negative mode; (3) exclude spectra with fewer than five fragment ions; (4) normalize peak intensities to a range of 0 to 1; and (5) to avoid any overlap in data, eliminate all spectra with cosine similarity greater than 0.95 compared to any spectrum in the CASMI datasets. Consequently, 600,289 positive spectra derived from 43,653 unique compounds and 182,323 distinct negative spectra from 20,883 unique compounds were retained.

After processing, spectra were transformed into documents. For this, each peak was represented by a word containing its position up to a specified decimal precision ("peak@xxx.xx"). In all presented outcomes, a binning approach involving two decimal places was applied, meaning that a peak at m/z 200.445 was represented as the word "peak@200.45". Alongside all the peaks within a spectrum, neutral losses ranging from 5.0 to 200.0 Da were also included as "loss@xxx.xx". The calculation of neutral losses involved the precursor m/z subtracted from the peak m/z. The compilation of all the generated peak and loss words constitutes what is referred to here as a document. Two semantic models were separately trained using the Spec2Vec package[32] for spectra in both positive and negative modes. The hyperparameters were configured as follows: (1) word vector dimension = 300, (2) negative sampling (negative = 5), (3) initial learning rate = 0.025, (4) learning rate decay per epoch = 0.00025, and (5) number of epochs = 30. During the parameter optimization process, 10% of the dataset was separated from the training set to form the validation set. However, when training the final model, the entire training set was utilized. After training, the semantic information is excepted to be captured through the Word2Vec model.

Neighbor compound searching

Using our trained semantic model, we embedded all spectra from the training set into vectors. This process resulted in two matrices: a 600,289 × 300 matrix for positive mode and a 182,323 × 300 matrix for negative mode. Each row in these matrices represents an embedded vector, and the similarity between vectors correlates with the structural similarity of the corresponding molecules. When presented with an unknown spectrum, it is embedded into a vector with the same model. Previous studies on Spec2Vec have demonstrated that the distance between the semantic model embedded vectors effectively reflects the structural similarity of corresponding metabolites[32]. Here, our objective is to efficiently compare the embedded vector of the query spectrum with pre-built matrices.

Therefore, we employed Hierarchical Navigable Small-World (HNSW) graphs. HNSW is an Approximate Nearest Neighbor Search (ANNS) method based on hierarchical graph structures and graph traversal techniques. HNSW indices do not require training with data. The constructional parameters are $M = 64$ and *ef_construction* = 800, where M defines the maximum number of outgoing connections in the graph, and *ef_construction* defines a construction time/accuracy trade-off. Essentially, the HNSW index serves as an efficient data structure, organizing hierarchically multi-layer graphs to store, organize, and calculate distances between vectors. This approach significantly improves the speed of ANNS. ANNS method is applied for rapid retrieval of the Top 300 vectors with the least distance. The structures associated with the retrieved vectors are considered “neighbor compounds”, indicating that they likely share similar structures with the unknown.

Candidate retrieval and ranking

A typical metabolomic data processing workflow utilizing comprehensive software like XCMS[26], MZMine[27], OpenMS[28], KPIC2[29] or MS-DIAL[30] provides quantitative features along with corresponding information such as precursor m/z, ion type, charge, and isotopic pattern for unknowns. This information enables the users annotate the molecular formulas using external software tools like BUDDY[24] or MIST-CF[25]. Therefore, DeepMASS encourages users to input unknown spectra along with assigned formulas, allowing it to retrieve candidates from the structural bio database based on the provided formula. In cases where uploaded spectra lack this information, DeepMASS requires precursor m/z, ion type, charge, and isotopic pattern. It then infers the monoisotopic mass based on these parameters and retrieves and evaluates potential candidates with a monoisotopic mass difference within a set threshold (defaulting to 10 ppm).

DeepMASS evaluates and ranks candidates by considering the relative location between the encompassed region of "neighbor compounds" and the candidates within the chemical space. The position of compounds in the chemical space is estimated using Morgan fingerprints, and the distance between paired compounds is determined through the *Dice* distance calculated from their fingerprint vectors.

To elaborate, for each candidate, DeepMASS scans the "neighbor compounds", identifying the nearest 20 compounds. The algorithm then calculates the cumulative distance between these selected "neighbor compounds" and the candidate, using this summed distance as the ranking score. The mathematical expression for this is:

$$\text{Score} = \sum_{i=0}^{20}\left\{1 - Dice(FP_c, FP_{n(i)})\right\} \qquad (1)$$

where FP_c represents the fingerprints of the candidate compound, $FP_{n(i)}$ represents the fingerprints of the i_{th} "neighbor compounds", and $Dice$ refers to the Dice distance. From the equation, it can be seen that the candidates with higher scores are closer to the predicted structurally similar neighbors.

According to the equation, candidates which are closer to the predict neighbor compounds will receive higher scores. In fact, it's not necessary for all predicted compounds to be highly related to the unknown substance. A subset of neighbor compounds showing structural similarities with the unknown substance is adequate to map the chemical space, thereby allowing the correct structure to receive a significantly higher score than others. This feature also means that the prediction of neighbors does not have to be exceptionally accurate. A few false positive predictions do not significantly impact the identification results.

Benchmarking data

We utilized CASMI 2014, 2016, and 2022 as benchmarking data for our evaluation. MS/MS peak lists for CASMI 2014 and 2016 were readily available on the official website. In the case of CASMI 2022, raw files of *mzML* format are provided, we directly extracted MS/MS data from the raw files using provided retention time and precursor m/z information. We assumed that correct molecular formulas were annotated and could be directly employed for candidate retrieval in the software utilized for comparison.

Competitive software

To assess the performance of DeepMASS, a comparative analysis was conducted using two renowned state-of-the-art in silico metabolite annotation software tools: MS-Finder (version 3.52) and SIRIUS (version 5.5.7). The parameter settings were as follows: SIRIUS, *MS2 mass accuracy* of 10 ppm, *Database* of 'Bio-Database', *Use heuristic above m/z* of *300*; MF-Finder, *MS1 mass tolerance* of 0.001 Da, *MS2 mass tolerance* of 0.01 Da, *Tree depth* of 2, *Database* of 'all local databases'.

SIRIUS utilized query spectra provided in *mgf* format, while MS-Finder employed spectra in *msp* format. In both DeepMASS and the comparative software, the chemical formulas were considered known. Default parameters were maintained for the competitive software. SIRIUS applied the CSI:FingerID algorithm, while MS-Finder utilized in silico fragmentation algorithm to prioritize candidate rankings. Since, the bio database construction of DeepMASS mirrors the methodology described for MS-Finder, and from the description, it indicates that SIRIUS also employs a nearly identical constitution for bio database construction as MS-Finder. It suggests that the candidates retrieved by the compared software tools expected to be very similar.

Due excessive time consumption by MS-Finder when handling complex structures, it is unable to process all instances. Therefore, only instances with an exact mass lower than 750 Da were considered for analysis using MS-Finder. This

selective approach could prove advantageous for MS-Finder, as annotating smaller compounds tends to be less challenging in nature. Another noteworthy aspect is that both SIRIUS and MS-Finder have already integrated all GNPS data into their training sets. Consequently, a portion of the CASMI data could be considered exposed to their models.

Experimental protocol for the tomato study

Cherry tomato (cultivar *Qianxi*) was used for the transgenic experiments. The plant materials were planted in the greenhouse at the AGIS experimental station in Shenzhen, China (114°30'3.63" E, 22°36'4.66" N). All seedlings were grown in a commercial nursery for 30-40 days at 28°C and then transplanted to fields. Three to five fruits per plant were picked at selected ripe stages and immediately frozen in liquid nitrogen. Then, the frozen tissues were ground to fine powders with an IKA A11 machine (IKA Works, Inc., NC, USA) and stored in a -80 °C freezer for metabolomic analysis.

For nontarget metabolomic analysis, 100 mg of tomato powder for each sample was extracted with 1 mL of extraction buffer (methanol:water = 80:20, v/v, containing 0.1% formic acid, with 100 ng/mL of d_5-coumarin, d_5-L-phenylalanine, and d_3-homovanillic acid as internal standards (Cambridge Isotope Laboratories, Inc., MA, USA)), followed by a 30 min ultrasonic treatment. The supernatants were pipetted and dried in a LABCONCO CentriVap vacuum centrifugal concentrator (We Brothers Scientific (Pvt.) Ltd., MO, USA) and resuspended in 100 µL of reconstitution buffer (methanol:water = 80:20, v/v). A 10 µL aliquot of the supernatant from each sample was pipetted and mixed as QC samples.

Then, extracts were analyzed by a U3000 UHPLC system coupled with a Q Exactive orbitrap mass spectrometer (Thermo Fisher Scientific Ltd., CA, USA) equipped with a HESI probe under both positive and negative modes. Extracts were separated by an ACQUITY UPLC™ BEH C18 column (Waters Ltd., MA, USA) (150 × 2.1 mm, 2.6 µm). The injection volume was set at 1 µL, and the flow rate was set at 300 µL/min. Solvents for the mobile phase were 0.05% formic acid in acetonitrile (A) and 0.1% formic acid in water (B). The elution gradient was as follows: 0-25 min, linear gradient 40%-100% A, followed by column washing with 100% B and 100% A. Data with mass ranges of m/z 80-1200 and m/z 80-1200 were acquired in both positive ion mode and negative ion mode separately, with data-dependent MS/MS acquisition. The full scan and fragment spectra were collected with resolutions of 70,000 and 17,500, respectively.

For manual data analysis, TraceFinder analysis software (version 4.1) was used to analyze the acquired MS/MS data. The data processing parameter settings were as follows: *peak width* from 10 to 45 s, *mzwid* of 0.015, *minfrac* of 0.5, *bw* of 5, and signal/noise threshold of 6. Then, compounds were annotated based on our in-house database via high-resolution MS-associated methods. Accurate high-resolution m/z values and MS/MS fragmentation patterns were used to identify metabolites, with a mass accuracy of 2 ppm and RT window of 20 s. Then, the annotated metabolites were checked and confirmed by comparing retention times, fragment patterns and isotope peaks manually. The library matching score threshold was set as 30.

For the DeepMASS pipelines, data preprocessing and MS/MS extraction were conducted through MS-DIAL[30]. The data processing parameter settings were as follows: *peak height* of 20000, *mass slice* of 0.1 Da, *alignment retention time tolerance* of 0.05 min. Then, MS/MS files were exported and fed to DeepMASS software. Subsequently, the MS/MS files were exported and inputted into the DeepMASS software. The sole parameter setting adjusted was the candidate retrieving mass tolerance, which was set to 10 ppm. The retrieval range defaulted to the entire database, while the structural annotation processing is non-parametric.

The metabolomic data were first normalized by sum and scaled with the Pareto scaling method (mean-centered and divided by the square root of the standard deviation of each variable). Principal component analysis (PCA) and partial least squares discrimination analysis (PLS-DA) were used for difference analysis between groups, and a heatmap was generated using relative areas of all annotated metabolites, with Euclidean and Ward used for distance measurement and cluster analysis between samples. Change trends in metabolite abundance among different ripening stages were analyzed by the OPTICS clustering algorithm from the scikit-learn package in Python.

DATA AVAILABILITY

Spectral data from GNPS are available at the official website (https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp); testing data are available at the CASMI website (http://casmi-contest.org); and spectral data from NIST 20 are commercially available and can be purchased from multiple vendors.

CODE AVAILABILITY

The software and source code for data cleaning, model training and performance evaluation of DeepMASS have been released under the AGPL 3.0 license at https://github.com/hcji/DeepMASS2_GUI.

ACKNOWLEDGMENTS

This work is financially supported by Shenzhen Science and Technology Program (Grant No. RCBS20210609103157070), the National Natural Science Foundation of China (Grant No. 22103036) and National Key R&D Program of China (grant. 2023YFA0915800). It was also supported by Innovation Program of Chinese Academy of Agricultural Sciences and the Elite Young Scientists Program of CAAS.

AUTHOR CONTRIBUTIONS

Conceptualization, H.J.; methodology, H.J.; formal analysis: H.J.; investigation: H.J., R.D.; resources: R.D.; Y.L.; J.Y.; experiment: S.F.; data curation: H.J., Q.D.; statistics: R.D.; writing original draft preparation: H.J., R.D.; review and editing, H.J., J.Y.; visualization: H.J., R.D.; supervision: J.Y.; project administration, J.Y.

DECLARATION OF INTERESTS.

The authors declare no conflict of interest.

REFERNCES

1. Shen, S., Zhan, C., Yang, C., Fernie, A. R. & Luo, J. Metabolomics-centered mining of plant metabolic diversity and function: Past decade and future perspectives. *Mol. Plant* **16**, 43–63 (2023).
2. Alseekh, S. *et al.* Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. *Nat. Methods* **18**, 747–756 (2021).
3. Li, Y. *et al.* Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. *Nat. Methods* **18**, 1524–1531 (2021).
4. De Jonge, N. F. *et al.* MS2Query: reliable and scalable MS2 mass spectra-based analogue search. *Nat. Commun.* **14**, 1752 (2023).
5. Yang, Q. *et al.* Ultra-fast and accurate electron ionization mass spectrum matching for compound identification with million-scale in-silico library. *Nat. Commun.* **14**, 3722 (2023).
6. Bazsó, F. L. *et al.* Quantitative Comparison of Tandem Mass Spectra Obtained on Various Instruments. *J. Am. Soc. Mass Spectrom.* **27**, 1357–1365 (2016).
7. Ji, H., Xu, Y., Lu, H. & Zhang, Z. Deep MS/MS-Aided Structural-Similarity Scoring for Unknown Metabolite Identification. *Anal. Chem.* **91**, 5629–5637 (2019).
8. Da Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl. Acad. Sci.* **112**, 12549–12550 (2015).
9. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
10. van der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl. Acad. Sci.* **113**, 13738–13743 (2016).

11. Morehouse, N. J. *et al.* Annotation of natural product compound families using molecular networking topology and structural similarity fingerprinting. *Nat. Commun.* **14**, 308 (2023).

12. Cai, Y., Zhou, Z. & Zhu, Z.-J. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. *TrAC Trends Anal. Chem.* **158**, 116903 (2023).

13. Hu, G. & Qiu, M. Machine learning-assisted structure annotation of natural products based on MS and NMR data. *Nat. Prod. Rep.* **40**, 1735–1753 (2023).

14. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci.* **112**, 12580–12585 (2015).

15. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).

16. Tsugawa, H. *et al.* Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.* **88**, 7946–7958 (2016).

17. Wang, F. *et al.* CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **93**, 11692–11700 (2021).

18. Lai, Z. *et al.* Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* **15**, 53–56 (2018).

19. Ji, H., Deng, H., Lu, H. & Zhang, Z. Predicting a Molecular Fingerprint from an Electron Ionization Mass Spectrum with Deep Neural Networks. *Anal. Chem.* **92**, 8649–8653 (2020).

20. Wang, L. *et al.* Metabolic Pathway Extension Approach for Metabolomic Biomarker Identification. *Anal. Chem.* **89**, 1229–1237 (2017).

21. Shen, X. *et al.* Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat. Commun.* **10**, 1516 (2019).

22. Zhou, Z. *et al.* Metabolite annotation from knowns to unknowns through knowledge-guided multi-layer metabolic networking. *Nat. Commun.* **13**, 6656 (2022).

23. Kim, S., Thiessen, P. A., Bolton, E. E. & Bryant, S. H. PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem. *Nucleic Acids Res.* **43**, W605–W611 (2015).

24. Xing, S., Shen, S., Xu, B., Li, X. & Huan, T. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nat. Methods* **20**, 881–890 (2023).

25. Goldman, S., Xin, J., Provenzano, J. & Coley, C. W. MIST-CF: Chemical Formula Inference from Tandem Mass Spectra. *J. Chem. Inf. Model.* (2023) doi:10.1021/acs.jcim.3c01082.

26. Forsberg, E. M. *et al.* Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nat. Protoc.* **13**, 633–651 (2018).

27. Schmid, R. *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nat. Biotechnol.* **41**, 447–449 (2023).

474 28. Pfeuffer, J. *et al.* OpenMS 3 enables reproducible analysis of large-scale mass spectrometry data. *Nat. Methods* 1–3 (2024)
475 doi:10.1038/s41592-024-02197-7.
476 29. Ji, H., Zeng, F., Xu, Y., Lu, H. & Zhang, Z. KPIC2: An Effective Framework for Mass Spectrometry-Based Metabolomics Using Pure
477 Ion Chromatograms. *Anal. Chem.* **89**, 7631–7640 (2017).
478 30. Tsugawa, H. *et al.* MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**,
479 523–526 (2015).
480 31. Malkov, Y. A. & Yashunin, D. A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small
481 World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 824–836 (2020).
482 32. Huber, F. *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Comput. Biol.*
483 **17**, e1008724 (2021).
484 33. Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare
485 tandem mass spectra. *J. Cheminformatics* **13**, 84 (2021).
486 34. Huber, F. *et al.* matchms - processing and similarity evaluation of mass spectrometry data. *J. Open Source Softw.* **5**, 2411 (2020).
487 35. Böcker, S. Searching molecular structure databases using tandem MS data: are we there yet? *Curr. Opin. Chem. Biol.* **36**, 1–6 (2017).
488 36. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
489 37. Gao, J. *et al.* Photoredox/Nickel Dual Catalysis-Enabled Cross-Dehydrogenative C–H Amination of Indoles with Unactivated Amine.
490 *Org. Lett.* **25**, 7716–7720 (2023).
491 38. Zhu, G. *et al.* Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell* **172**, 249-261.e12 (2018).
492 39. Tieman, D. *et al.* A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391–394 (2017).
493 40. Klee, H. J. & Tieman, D. M. The genetics of fruit flavour preferences. *Nat. Rev. Genet.* **19**, 347–356 (2018).
494 41. Li, L. & Herzon, S. B. Temporal separation of catalytic activities allows anti-Markovnikov reductive functionalization of terminal
495 alkynes. *Nat. Chem.* **6**, 22–27 (2014).
496 42. Leunissen, M., Davidson, V. J. & Kakuda, Y. Analysis of Volatile Flavor Components in Roasted Peanuts Using Supercritical Fluid
497 Extraction and Gas Chromatography−Mass Spectrometry. *J. Agric. Food Chem.* **44**, 2694–2699 (1996).
498 43. Fang, X. *et al.* Characteristics of volatile flavor components in traditional fermented yak milk produced in different ecoregions of the
499 Qinghai-Tibetan plateau. *J. Dairy Sci.* **103**, 191–200 (2020).
500 44. Ntoruru, J. M., Ohnishi, T., Katsumata, F., Koeduka, T. & Matsui, K. 1-Octen-3-ol is formed from its primeveroside after mechanical
501 wounding of soybean leaves. *Plant Mol. Biol.* **109**, 551–561 (2022).
502

Fig 1. Workflow of DeepMASS. Spectra from a database are used to train a semantic embedding model through the word2vec algorithm. When an unknown spectrum is provided, MS1 is employed to predict the molecular formula, which is then used to retrieve candidate structures from a chemical library, such as HMDB, FooDB, and DrugBank. MS2 is embedded and used to search for structurally similar neighbors using the trained model. Candidate and reference structures are projected into chemical space using chemical fingerprints, and the correct structure of the unknown spectrum is determined based on the relative location of the chemical space of the reference structures.

Fig 2. Performance of DeepMASS. a. Ratio of correctly identified structures found in the top *k* outputs of the different methods. The three-line plots (from left to right) represent the performance of all spectra, spectra with corresponding metabolites present in the reference database, and spectra with corresponding metabolites absent from the reference database. **b**. Annotation example for 6-bromoquinolin-2(1H)-one (absent from reference dataset) and N-acetyl-5-aminosalicylic acid (included in the reference dataset). The scatter plots are visualized in chemical space by UMAP dimensionality reduction of Morgan fingerprints. In the scatter plots, the top 10 reference structures predicted by DeepMASS's semantic similarity model are represented as blue points, while the retrieved candidate structures are represented as green points. Notably, the correct annotations, tagged with red points, are encompassed by the predicted neighbor structures. The three closest neighbors are displayed, and shared substructures with the correct structures are highlighted. Obviously, their chemical structures are highly relevant. Meanwhile, their corresponding spectra also exhibit a high degree of correlation (red: query spectrum; blue: neighbor spectrum; fragment: the original spectrum; loss: neutral losses calculated from the precursors and fragments).

Fig 3. Enhanced metabolite annotation by DeepMASS in practical applications. a. Assorted ripening stages of tomato samples. **b.** Comparative analysis of annotated feature percentages and the quantity of compounds by DeepMASS and TraceFinder. **c.** PCA scatter plot of samples with annotated compounds by DeepMASS and TraceFinder.

Fig 4. Biomarker identification performance of DeepMASS. a. Metabolites annotated in association with various tomato ripening stages along with their relative abundances. **b.** Heatmap of samples from different stages based on metabolite clusters achieved from the OPTICS clustering algorithm. **c.** Top five clusters from both positive and negative mode results.

Fig 5. Software functionality overview. The software loads spectra via a *mgf* file and offers two modes: spectrum matching using MatchMS and in silico annotation with DeepMASS. It includes user-friendly widgets for displaying spectra, metadata, and predicted results. Additionally, it provides essential decision-making information to assess confidence in obtaining accurate results.

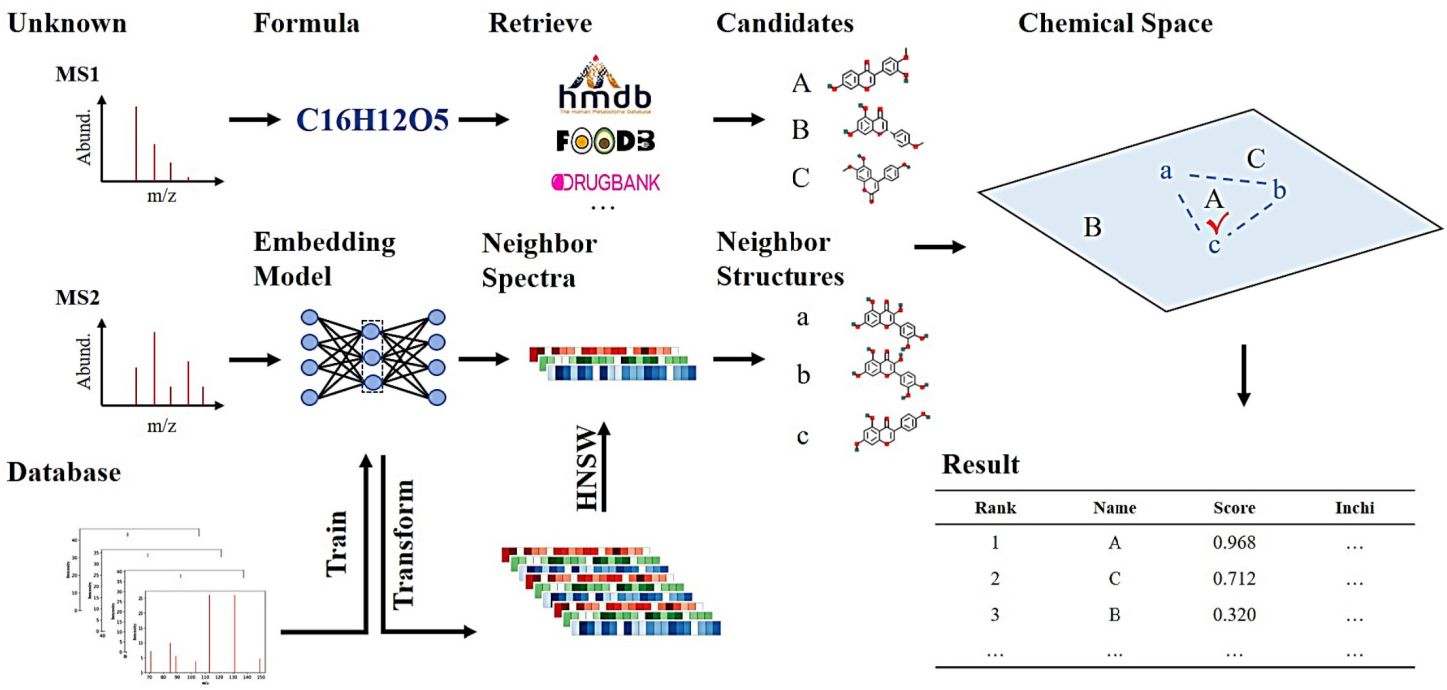

Fig 1. Workflow of DeepMASS. Spectra from a database are used to train a semantic embedding model hrough the word2vec algorithm. When an unknown spectrum is provided, MS1 is employed to predict the molecular formula, which is then used to retrieve candidate structures from a chemical library, such as HMDB, FooDB, and DrugBank. MS2 is embedded and used to search for structurally similar neighbors using the trained model. Candidate and reference structures are projected into chemical space using chemical fingerprints, and the correct structure of the unknown spectrum is determined based on the relative location of the chemical space of the reference structures.

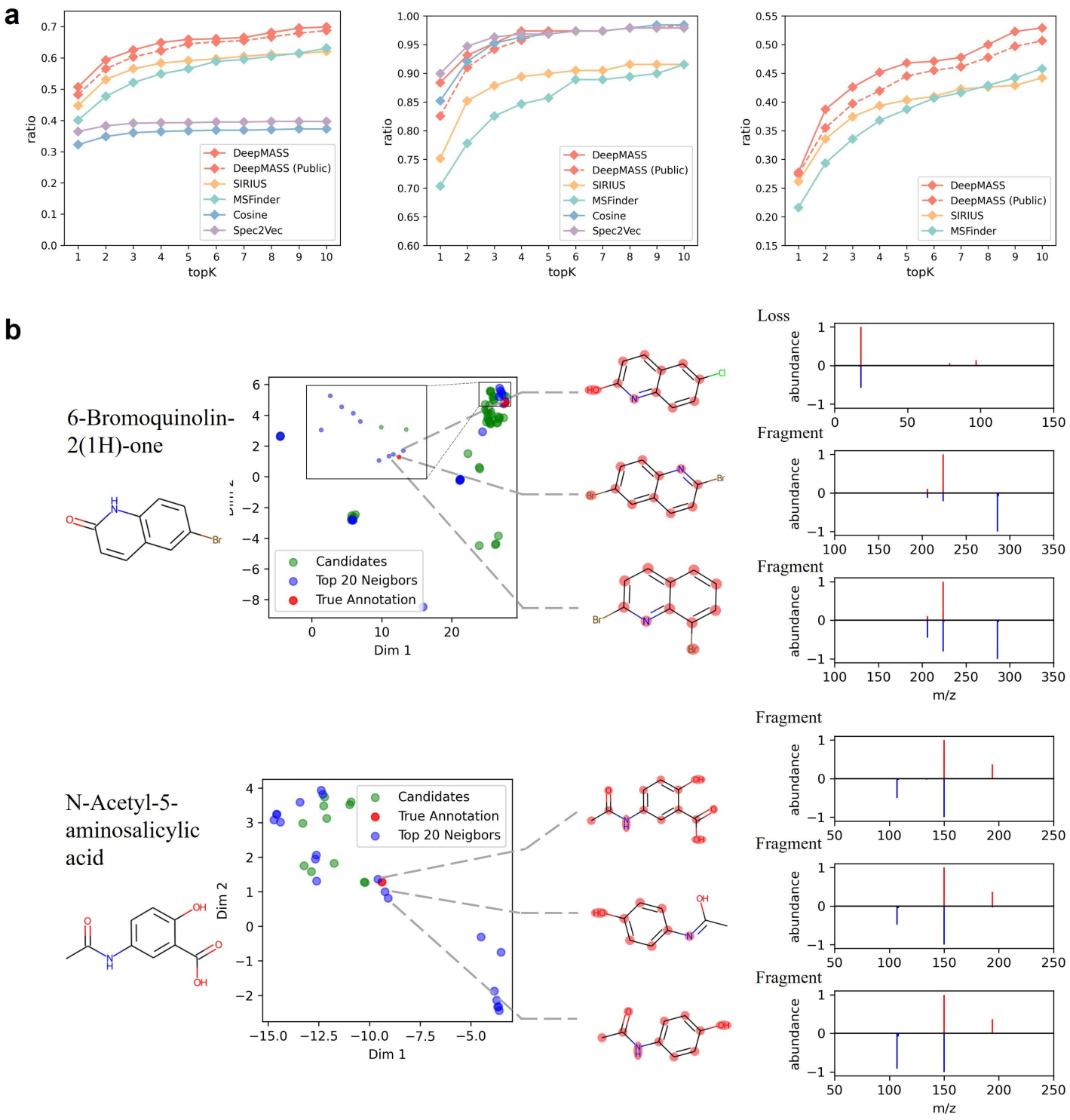

Fig 2. Performance of DeepMASS. a. Ratio of correctly identified structures found in the top k outputs of the different methods. The three-line plots (from left to right) represent the performance of all spectra, spectra with corresponding metabolites present in the reference database, and spectra with corresponding metabolites absent from the reference database. **b.** Annotation example for 6-bromoquinolin-2(1H)-one (absent from reference dataset) and N-acetyl-5-aminosalicylic acid (included in the reference dataset). The scatter plots are visualized in chemical space by UMAP dimensionality reduction of Morgan fingerprints. In the scatter plots, the top 10 reference structures predicted by DeepMASS's semantic similarity model are represented as blue points, while the retrieved candidate structures are represented as green points. Notably, the correct annotations, tagged with red points, are encompassed by the predicted neighbor structures. The three closest neighbors are displayed, and shared substructures with the correct structures are highlighted. Obviously, their chemical structures are highly relevant. Meanwhile, their corresponding spectra also exhibit a high degree of correlation (red: query spectrum; blue: neighbor spectrum; fragment: the original spectrum; loss: neutral losses calculated from the precursors and fragments).

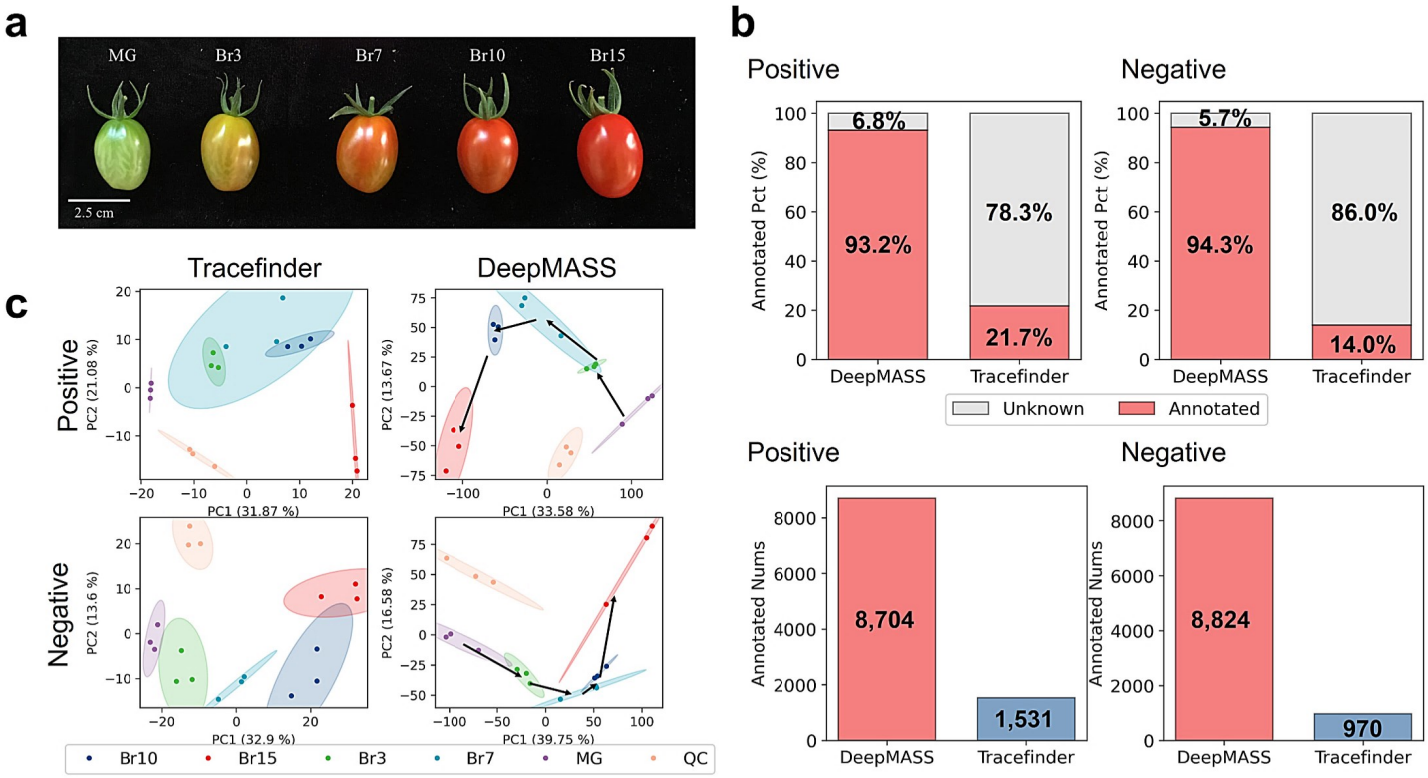

Fig 3. Enhanced metabolite annotation by DeepMASS in practical applications. a. Assorted ripening stages of tomato samples. **b.** Comparative analysis of annotated feature percentages and the quantity of compounds by DeepMASS and TraceFinder. **c.** PCA scatter plot of samples with annotated compounds by DeepMASS and TraceFinder.

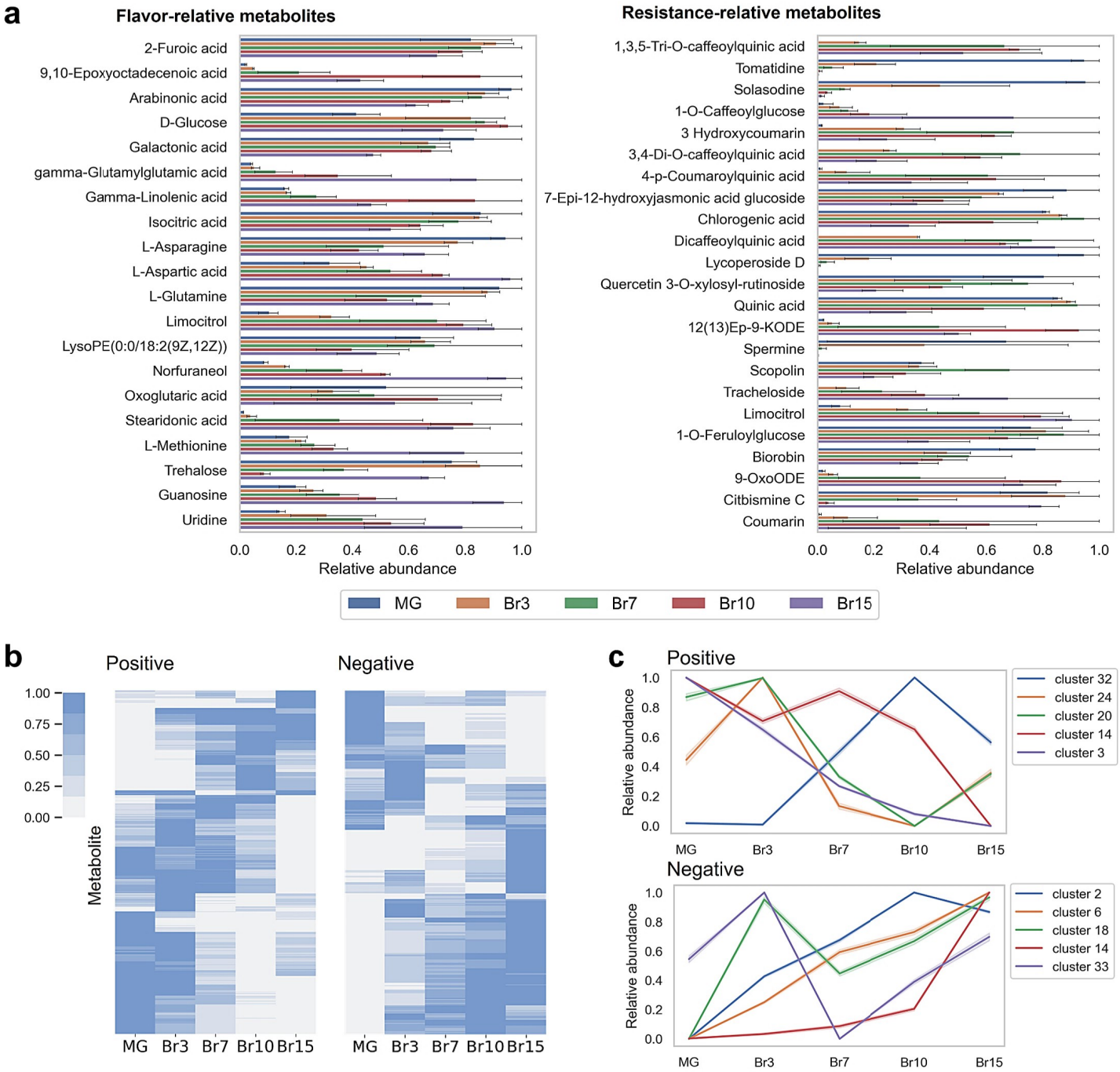

Fig 4. Biomarker identification performance of DeepMASS. a. Metabolites annotated in association with various tomato ripening stages along with their relative abundances. **b.** Heatmap of samples from different stages based on metabolite clusters achieved from the OPTICS clustering algorithm. **c.** Top five clusters from both positive and negative mode results.

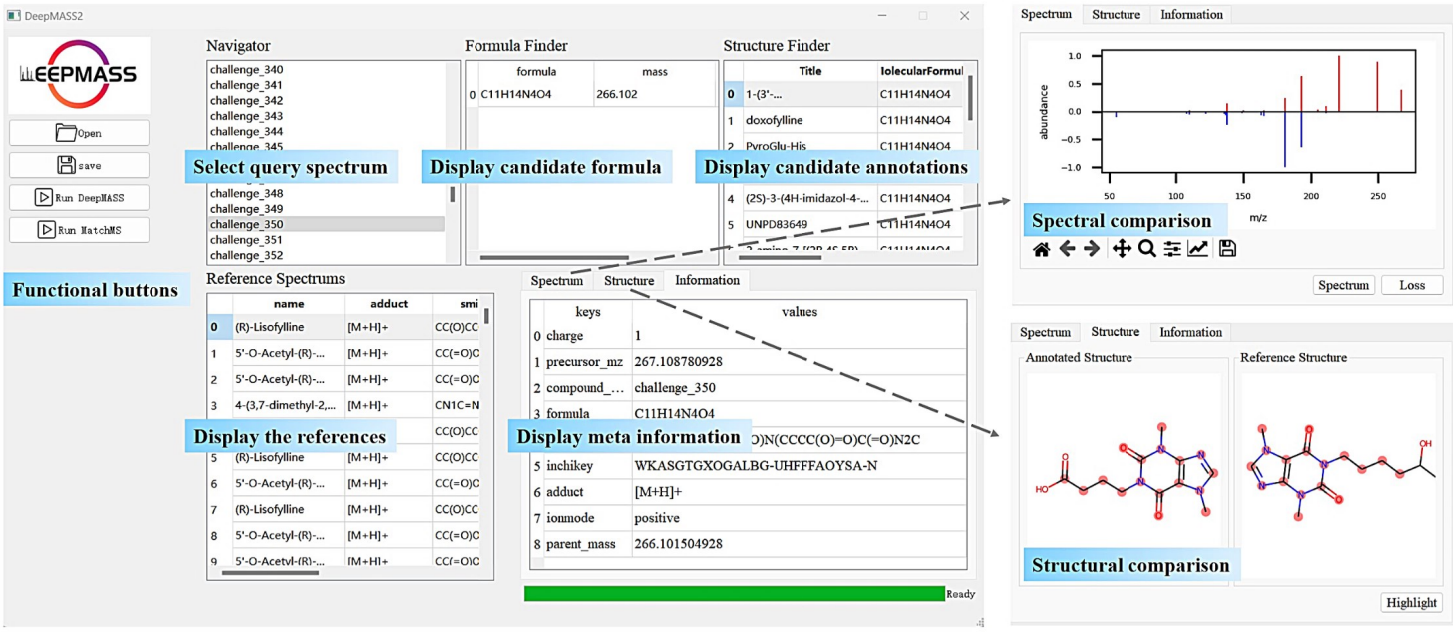

Fig 5. Software functionality overview. The software loads spectra via a mgf file and offers two modes: spectrum matching using MatchMS and in silico annotation with DeepMASS. It includes user-friendly widgets for displaying spectra, metadata, and predicted results. Additionally, it provides essential decision-making information to assess confidence in obtaining accurate results.