communications chemistry

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42004-025-01708-7

A large language model for deriving spectral embeddings for accurate compound identification in mass spectrometry

Check for updates

Yang Xu ✉, Yixiao Ma, Weijie Xu, Zuliang Yang & Kai Ming Ting ✉

Identifying chemical components in complex mixtures is a crucial task across many scientific disciplines. Mass spectrometry serves as a key analytical tool for this purpose, yet the accurate identification of compounds from their spectra remains a major bottleneck. Here we introduce LLM4MS, a method that leverages the latent expert knowledge within large language models to generate discriminative spectral embeddings for improved compound identification. LLM4MS is designed to incorporate potential chemical expert knowledge, enabling accurate matching. Evaluated against a million-scale open-source in-silico library using the NIST23 library as a test set, LLM4MS achieves a Recall@1 accuracy of 66.3% (and a Recall@10 accuracy of 92.7%), representing a 13.7% improvement over the state-of-the-art Spec2Vec. Furthermore, LLM4MS enables ultra-fast mass spectra matching, achieving a speed of nearly 15,000 queries per second. Thus, LLM4MS opens up avenues to significantly enhance compound identification in mass spectrometry and accelerate chemical discovery.

Large Language Models (LLMs) have demonstrated a remarkable capacity to acquire and utilize domain-specific knowledge from extensive training corpora of specialized scientific databases. This inherent knowledge has shown significant potential in accelerating scientific discovery across diverse fields, including medicine[1], proteomics analysis[2], climate science[3], and chemistry[4,5], among others.

Mass spectrometry (MS), particularly tandem mass spectrometry (MS/MS or MS^2), is a cornerstone analytical technique celebrated for its sensitivity and versatility in the analysis of complex chemical mixtures. MS plays a vital role in various scientific domains, including metabolomics, proteomics, and synthetic organic chemistry[6]. A key challenge in these fields is the efficient and accurate identification of compounds in complex samples. Mass spectrometry facilitates the generation of large spectral datasets, often containing thousands to millions of individual spectra[7]. Matching experimental spectra against curated libraries of known molecular masses and fragmentation patterns is fundamental for high-throughput compound identification. This process is essential for elucidating the composition of complex samples and driving discoveries across scientific disciplines.

Accurate matching of query mass spectra against extensive mass spectral libraries remains a critical bottleneck in compound identification. The reliability of this process depends on the chosen metric's ability to accurately reflect the underlying structural relationships between query and reference spectra. Traditionally, weighted cosine similarity (WCS) has been a prevalent method in mass spectrometry[8]. Although effective in many scenarios, WCS and other metrics such as weighted average ratio matching[9], probability-based matching[10], and neutral loss matching[11] often exhibit limitations in resolving subtle structural variations. More recently, machine learning techniques have emerged as powerful tools, offering opportunities to enhance spectral matching accuracy. Methods such as SIRIUS[12], CANOPUS[13], MS2LDA[14], MS2Deepscore[15], and Spec2Vec[16] have shown improved performance by leveraging features such as precursor ion neutral losses. Notably, Spec2Vec, which uses word embedding techniques, has shown significant promise in capturing intrinsic structural similarities and enabling effective and efficient matching against millions-scale mass spectral libraries[17,18].

Despite the progress in spectral matching techniques, current methods often struggle to accurately resolve fine-grained structural dissimilarities, occasionally yielding high similarity scores for spectra of structurally distinct compounds. This limitation can result in erroneous identifications, especially when relying on metrics that prioritize global spectral overlap without

State Key Laboratory for Novel Software Technology & School of Artificial Intelligence, Nanjing University, Nanjing, China. ✉e-mail: xuyang@lamda.nju.edu.cn; tingkm@nju.edu.cn

considering fundamental chemical rules. For example, the base peak (typically the most intense ion) in an electron ionization mass spectrum often represents a structurally significant fragment or the molecular ion, serving as a crucial indicator of molecular identity[19,20]. A pronounced mismatch in base peaks between two spectra generally signifies a low probability of shared structural motifs. Additionally, focusing on fragment spectra corresponding to fragment ion peaks also aids in mass spectra matching[21–26]. However, traditional similarity metrics fail to adequately weigh such essential chemical information, potentially assigning high similarity scores to spectra with disparate base peaks. We performed a query spectrum **q** (corresponding to the compound C=CCCCCCCCCC[N+](=O)[O-]) from the NIST23 library[7] against a million-scale publicly available in-silico EI-MS library[17]. As illustrated in Fig. 1A, traditional methods like WCS and more advanced approaches such as Spec2Vec focus on the overall intensity distribution only, lacking an inherent capacity to apply the underlying chemical principles and domain-specific knowledge. In the depicted scenario, both WCS and Spec2Vec incorrectly assign a higher similarity score to spectrum **a** (corresponding to the compound C=CCCCCCCCCC#N), overlooking the critical base peak difference that strongly suggests a closer relationship between the query and spectrum **b** (corresponding to the compound C=CCCCCCCCCC[N+](=O)[O-]). This discrepancy underscores the limitations of these metrics in capturing the intricate relationship between chemical structure and fragmentation patterns.

In contrast, LLMs, pre-trained on vast and diverse text corpora, have demonstrated a remarkable emergent capacity to apply chemical principles during data interpretation (Fig. 1B). This capability likely arises from the sheer volume of scientific content available across the public internet, which, even when encountered during pre-training as fragments from literature, textbook excerpts, or scientific commentaries, collectively allows for the synthesis of complex chemical knowledge[27]. As we will illustrate, this allows them to "reason" about fragmentation patterns and their influence on mass spectral similarity, a capability that traditional metrics inherently lack. As illustrated in Fig. 1B, LLM-based inferences, exemplified by prominent LLMs models DeepSeek-R1[28] and GPT-4o[29], demonstrate a strong focus on chemically significant features. Instead of merely comparing overall intensity distributions, the LLMs prioritize the alignment of diagnostically important peaks. For example, both LLMs immediately identify the critical mismatch in the base peak for spectrum **a**. DeepSeek notes that **a**'s "Highest peak at m/z 41 (unrelated to **q**'s base peak)", while GPT states, "In **a**, the base peak is at m/z 41 (intensity = 1.0), which strongly differs from **q**". Conversely, for spectrum **b**, DeepSeek observes that it "matches **q**'s base peak (m/z 55, intensity 1.0)", and GPT concurs, stating, "In **b**, the base peak is also at m/z 55 (intensity = 1.0), matching **q** exactly in both position and intensity". Furthermore, both the LLMs consider the presence or absence of high-mass ions potentially indicative of the molecular weight. DeepSeek points out that spectrum **a** "lacks m/z 182, with its highest ion at m/z 166," whereas spectrum **b** "includes m/z 182 (low intensity), supporting a possible match with **q**'s molecular ion". Similarly, GPT highlights that spectrum **a** "does not have a peak at m/z 182," while "**b** includes m/z 182 similar to the small intensity observed in **q**". By learning and applying such chemical heuristics - focusing on base peak alignment and the presence of key high-mass fragments - LLMs can discern subtle structural differences often missed by traditional metrics, leading to more accurate and robust mass spectra matching. More LLMs question-and-answer examples can be found in Supplementary Note 1.

The remarkable success of LLMs stems significantly from their ability to transform complex information, including textual data, into high-dimensional vector representations, or embeddings, where semantic relationships are effectively encoded as spatial proximity. Capitalizing on this inherent capability, we introduce LLM4MS, a method designed to harness the power of LLMs for enhancing electron ionization mass spectral matching via LLM embeddings. Distinct from traditional similarity metrics and earlier machine learning approaches like Spec2Vec (Fig. 2A), which often rely primarily on explicit spectral features and predefined rules learned from spectral data context, LLM4MS operates on a different paradigm (Fig. 2B). It leverages the vast chemical and scientific knowledge acquired by the LLM during its extensive pre-training on diverse corpora. By textualizing mass spectra and processing these textual representations through a purpose-fine-tuned LLM, LLM4MS generates richer, more chemically informed embeddings. This approach produces a more nuanced representation of mass spectra, effectively capturing subtle structural information reflected in fragmentation patterns that leads to a more accurate compound identification. This work pioneers the direct application of LLM-derived embeddings to large-scale spectral library searching, offering a promising direction for the field.

Results

Superior performance and efficiency of LLM4MS in large-scale spectral matching

The performance of LLM4MS for compound identification was evaluated using a well-established spectral library and a test set. The reference database employed for spectral matching was the publicly available, million-scale in-silico EI-MS library developed by Yang et al.[17]. This library represents a significant advancement designed to address the critical limitation of coverage in traditional experimental spectral libraries, containing over 2.1 million predicted EI-MS spectra. For querying against this reference library, we constructed a high-quality test set derived from NIST23's Small Molecule High Resolution Accurate Mass MS/MS Library (mainlib), selecting 9921 spectra corresponding to compounds also present within the in-silico reference library. To ensure reproducibility and provide a benchmark for future studies, the complete list of these 9921 query compounds is provided in our publicly available code repository. This purposeful overlap ensures that our evaluation directly assesses the capability of LLM4MS to correctly identify known compounds in this extensive, million-scale library. This purposeful overlap ensures that our evaluation directly assesses the capability of LLM4MS to correctly identify known compounds in this extensive, million-scale library.

Prior to benchmarking, we have validated the representativeness and chemical diversity of the selected test set. The visualization of the LLM4MS embedding space using Uniform Manifold Approximation and Projection (UMAP)[30,31] (with `n_neighbors=15`, `n_components=3`, and `metric='cosine'`) revealed significant overlap and close proximity between the embeddings of the entire NIST23 mainlib (Fig. 3A, blue) and our 9921-spectrum test set (Fig. 3A, orange), indicating that the test set accurately reflects the broader distribution of experimental spectra within the learned embedding space. Furthermore, to ensure our evaluation was not biased towards a narrow range of chemical structures, we characterized the composition of the test set using NPClassifier[32]. This analysis serves to validate the chemical diversity of the queries. The results, illustrated in Fig. 3B, confirmed that the test set is indeed diverse, encompassing a wide variety of compound classes including fatty acyls, fatty esters, alkaloids, and terpenoids. Establishing this diversity is crucial because it provides strong evidence that the high accuracy of LLM4MS is broadly applicable across varied chemical structures and not limited to a few over-represented compound types.

We compared the accuracy of LLM4MS against several established methods: the machine learning-based Spec2Vec, the traditional Weighted Cosine Similarity (WCS), and standard Cosine Similarity. For the embedding-based methods (LLM4MS and Spec2Vec), accuracy was assessed using cosine similarity calculated between the derived embeddings, while for the traditional methods (WCS and standard Cosine Similarity), scores were computed directly on the original spectra as a baseline comparison. The results, presented as Recall@*x* curves (Fig. 3C), demonstrate a marked improvement in compound identification accuracy achieved by LLM4MS across all top-x retrieval scenarios. Notably, LLM4MS achieved a Recall@1 accuracy of 66.3%, substantially outperforming Spec2Vec (58.3%), WCS (56.5%), and Cosine Similarity (28.6%). This indicates a significantly higher probability of retrieving the correct compound as the top match using LLM4MS. The superior performance extends to deeper recall levels, with LLM4MS reaching a Recall@10 of 92.7%, compared to 85.7% for

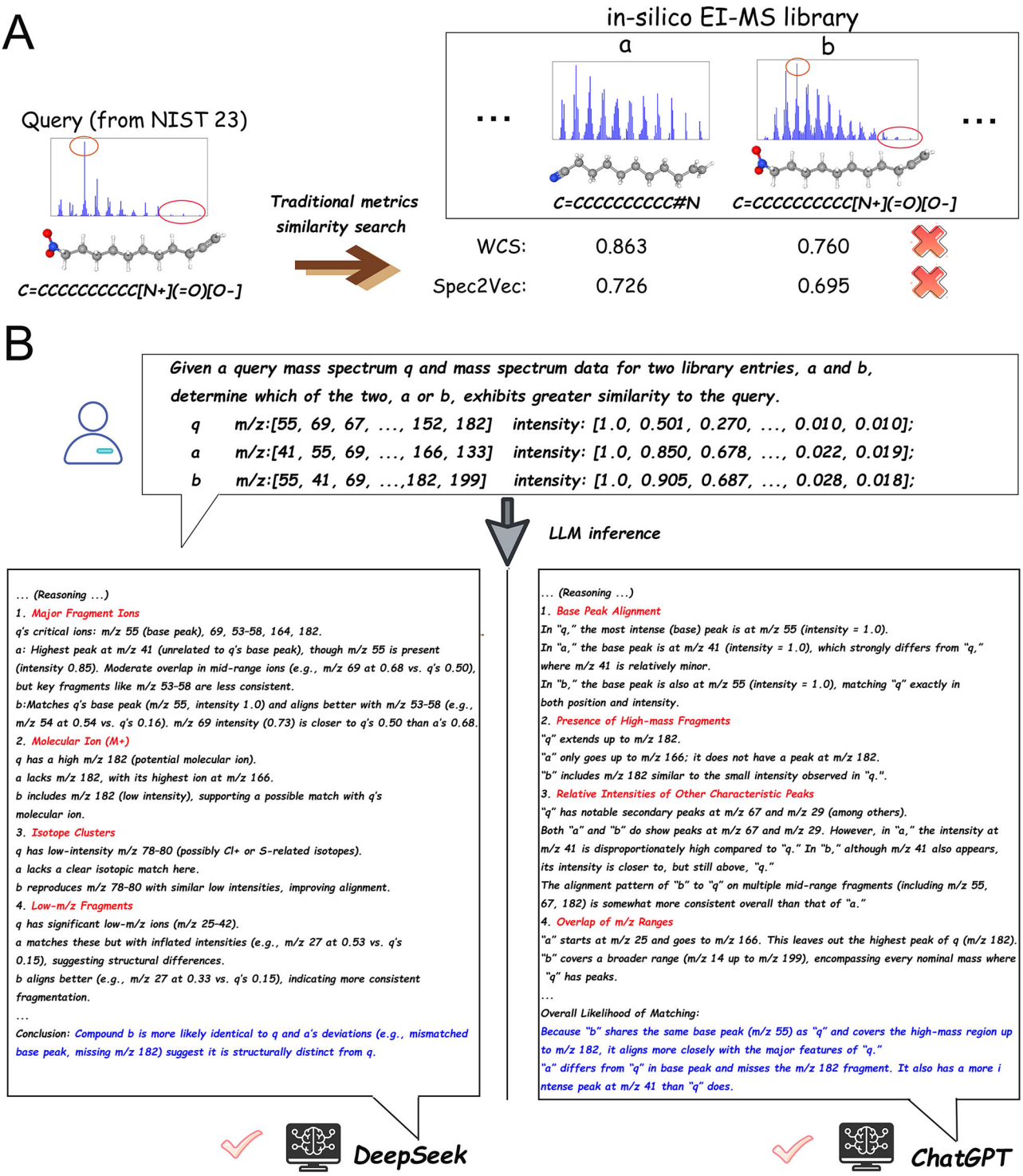


Fig. 1 | Comparison of spectral similarity assessment using traditional metrics and LLM-driven inference. The query mass spectrum **q**, corresponding to the compound C=CCCCCCCCCC[N+](=O)[O-], is from the NIST23 library, and the mass spectra of two compounds, **a** (C=CCCCCCCCCC#N) and **b** (C=CCCCCCCCCC[N+](=O)[O-]), are from the in-silico EI-MS library. **A** Traditional similarity metrics (WCS and Spec2Vec) only consider overall spectral similarity, leading to the wrong retrieval results. **B** LLMs (DeepSeek and GPT) infer similarity based on chemical principles reasoning (e.g., key peak/fragment matching), which improves compound matching for complex structures, leading to the true retrieval result.

Spec2Vec and 84.1% for WCS. This consistent outperformance highlights the effectiveness of the LLM-derived embeddings in capturing structurally relevant spectral information. The complete comparison of the performance metrics can be found in Supplementary Note 2.

Beyond accuracy, computational efficiency is critical for practical applications, especially when searching against million-scale libraries[17,33,34]. We evaluated the search speed of LLM4MS when integrated with the state-of-the-art Approximate Nearest Neighbor Search (ANNS) indexing techniques, including Annoy[35], Faiss[36] (the HNSW[37] implementation), HNSWlib[38], and NMSLIB[39]. Figure 3D illustrates the trade-off between retrieval speed, measured in queries per second (QPS), and Recall@1 accuracy for LLM4MS using these different ANNS techniques. As a

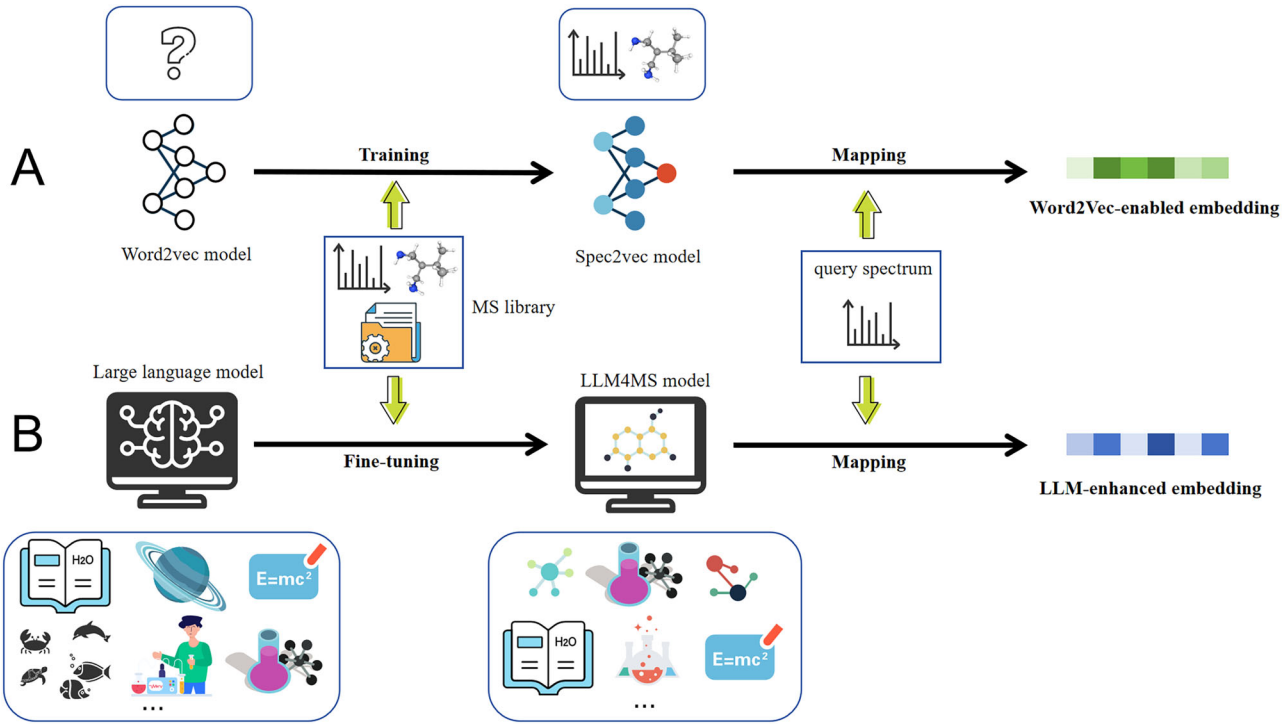


Fig. 2 | Conceptual comparison of spectral embedding generation strategies. **A** Schematic representation of the Spec2Vec method. A Word2vec model is trained using relationships derived from a mass spectral (MS) library to learn the spectral embeddings. **B** Schematic representation of the proposed LLM4MS approach. A large language model, pre-trained on extensive and diverse knowledge domains, undergoes fine-tuning using the MS library. This process leverages the LLM's embedded knowledge to generate chemically informed spectral embeddings for improved mass spectral matching.

reference, a brute-force search, while ensuring maximum possible accuracy (equivalent to 66.3% Recall@1), is computationally intensive, achieving approximately 0.27 QPS. In stark contrast, employing an ANNS technique dramatically accelerates the search process with a small impact on accuracy. For example, using HNSWlib, LLM4MS processed an impressive 14,440 QPS while maintaining a high Recall@1 accuracy of 64.6%. This is an approximately 54,000-fold speedup compared to the brute-force search, with only a minor decrease in top-1 (Recall@1) accuracy. Similar significant speed enhancements were observed with other ANNS techniques. The ability of LLM4MS to couple with ANNS to achieve ultra-fast search speeds while preserving high accuracy underscores its suitability for a high-throughput compound identification workflow. The detailed runtime metrics can be found in Supplementary Note 3.

Generalization performance on compounds unseen during model training

It is worth noting that the in-silico EI-MS library was generated by the NEIMS model, which was trained on publicly available data derived from the NIST17 library[40]. To rigorously assess the generalization capability of our method, we conducted an additional evaluation on a dataset of only unseen compounds. Specifically, we removed all compounds which are present in its original training data from our initial 9921 NIST23-derived queries. This process resulted in a benchmark containing 2618 "unseen" compounds.

The performance evaluation of all methods on this benchmark is detailed in Table 1. As anticipated, the absolute recall values for all methods are lower on the benchmark of "unseen" compounds compared to the results presented in Fig. 3C. This is attributable to the fact that the predicted spectra for compounds unseen by the NEIMS model are "harder" compared to those compounds known to the model. This makes the matching task more challenging for any algorithm. LLM4MS consistently outperforms all baseline methods across all recall levels even when matching against these lower-quality predicted spectra. Notably, it achieves a Recall@1 of 41.9%, representing a significant performance margin over the next best method, Spec2Vec (36.7%). This result demonstrates the superior generalization capability and robustness of LLM4MS in identifying new compounds.

Evaluation of heuristic-based baselines

The LLMs-driven inference, shown in Fig. 1B, suggests that LLMs could utilize some heuristics, such as base peak matching and the overlap of m/z ranges, as the criteria to judge whether different mass spectra originate from the same compound. This raises a question: could the performance of existing baseline methods be improved by explicitly augmenting them with such heuristics? The answer would ascertain the source of the power of using LLM4MS, i.e., it stems from a holistic learned model or simply from the application of a few dominant, easily replicable rules. To this end, we test the impact of a heuristic, that is based on base peak matching, on the performance of all baseline methods. For a comprehensive evaluation, we adopt two distinct strategies in applying this heuristic for this evaluation. The first is a direct filter strategy, where for each query, the reference library was pre-filtered to retain only candidates with a matching base peak, after which the similarity was calculated based on this reduced subset. The second is a weighted strategy, where the similarity score derived from a baseline method was linearly combined with a binary base peak matching score to produce the final similarity score (see the Methods section for the specific formula).

The results of applying both strategies are shown in Table 2. A clear and consistent trend emerged for all baseline methods (Spec2Vec, WCS, and CS). As highlighted by the row-wise best results (in boldface), the original "Baseline" performance was consistently the best. Both the "Filter" and "Weighted" strategies invariably led to a degradation in accuracy. For example, Spec2Vec's Recall@1 dropped from 58.3% to 49.4% under the filter strategy. This significant drop is unsurprising, as our analysis revealed that only 69% of the total number of pairs of query and ground truth are correct, sharing an identical base peak. This imposes a performance ceiling on the filter strategy that is not only lower than any of the Recall@1 scores of the baseline methods but is also substantially lower than the Recall@10

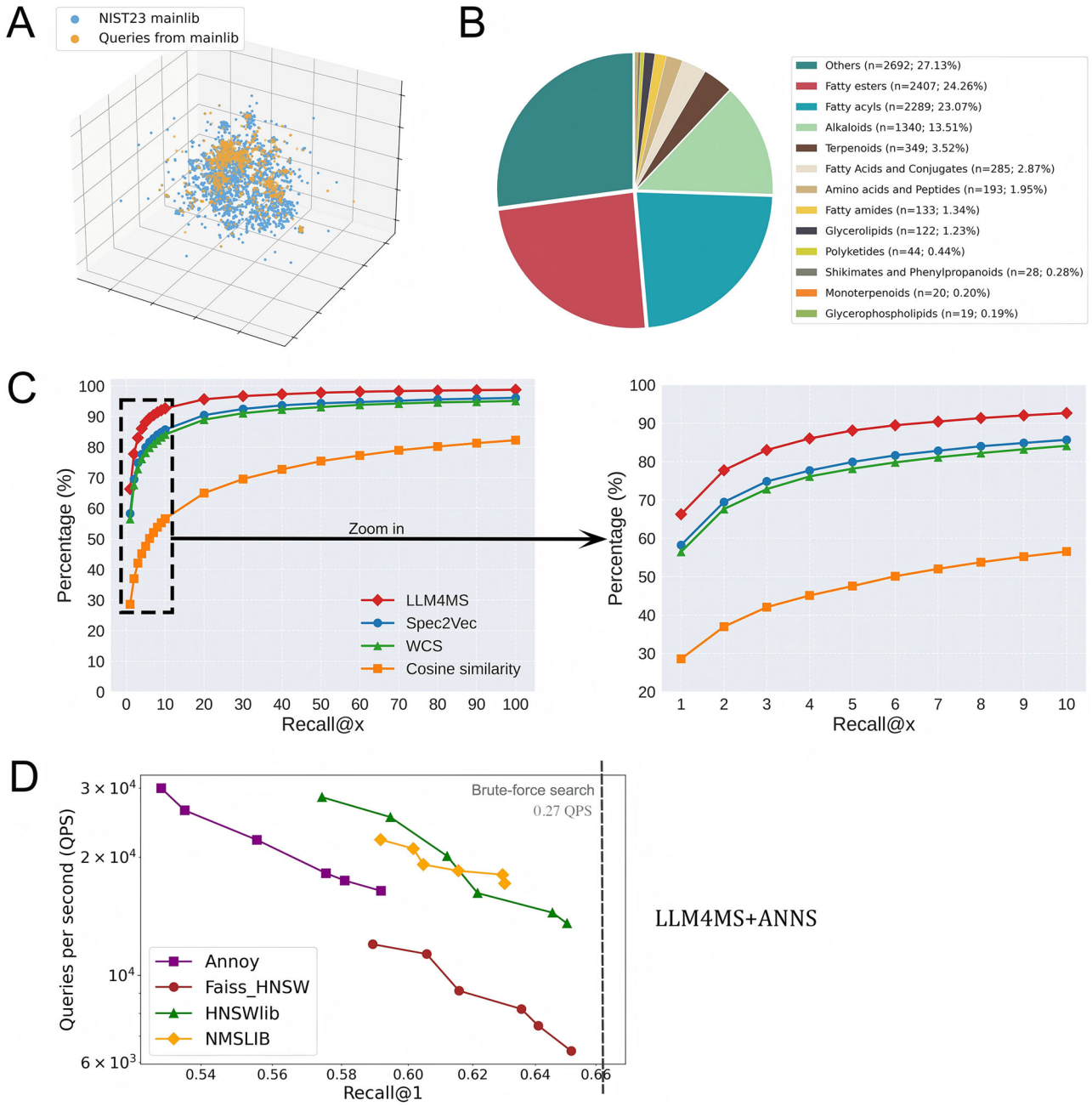


Fig. 3 | Information of datasets and performance of LLM4MS. A Three-dimensional UMAP visualization of the NIST23 mainlib (blue) and a test set of the NIST23 mainlib that overlaps with the in-silico EI-MS library, used as queries (orange). **B** Distribution of the compound classes within the NIST23 test set. The number of compounds and their respective percentages for each class are indicated in the legend. **C** Recall@*x* performance of LLM4MS, Spec2Vec, WCS, and cosine similarity on the NIST23 test set against the million-scale in-silico EI-MS library. The inset provides a zoomed-in view of the Recall@*x* performance for the top 10 retrieved candidates. **D** Comparison of the query speed (queries per second) as a function of Recall@1 accuracy for LLM4MS using different approximate nearest neighbor search (ANNS) machine learning indexing techniques (Annoy, Faiss HNSW, HNSWlib, and NMSLIB). Each point on a curve represents the performance achieved with a specific set of hyperparameters for the corresponding ANNS method (e.g., varying parameters like *ef_construction* and *M* for HNSWlib). The performance of a brute-force search is also indicated as a baseline for accuracy, with its corresponding query speed highlighted.

performance of the baseline methods (e.g., 92.7% for LLM4MS, 85.7% for Spec2Vec, and 84.1% for WCS). Moreover, we observed the same pattern for our own method. The LLM4MS model achieves the best performance (66.3% Recall@1), and its accuracy also decreases when either the filter (58.0%) or weighted (62.7% for $\alpha = 0.9$) strategy is applied. This suggests that the rich, data-driven model learned by LLM4MS has already found a more holistic balance of chemical evidence. Superimposing an external heuristic disrupts this balance rather than enhancing it. This finding indicates that the superiority of LLM4MS arises not from an easily engineered heuristic, but from a complex learned model of holistic chemical knowledge that is far more robust and effective.

Enhanced identification of structurally similar compounds by LLM4MS

During its training process on extensive databases, an LLM acquires substantial domain-specific chemical knowledge, enabling the LLM to leverage chemical properties for the analysis of mass spectral similarity. We observed that similarity scores derived from LLM-enhanced embeddings produce

Table 1 | Recall@*x* performance comparison of all methods on the benchmark containing 2618 "unseen" compounds

	LLM4MS	Spec2Vec	WCS	Cosine Similarity
Recall@1	**41.9**	36.7	35.2	12.1
Recall@2	**55.2**	49.7	47.5	18.3
Recall@5	**70.6**	66.1	64.0	28.1
Recall@10	**79.7**	76.0	73.6	36.8
Recall@20	**86.9**	83.8	81.2	46.1
Recall@50	**92.9**	90.9	88.2	58.6
Recall@100	**96.1**	94.3	91.5	67.3

The best results are highlighted in **bold**.

Table 2 | Detailed performance comparisons of individual baseline methods against those with either of the two strategies of applying a heuristic

Method	Recall@1 (%)			Recall@10 (%)		
	Baseline	Filter	Weighted (α = 0.9/0.5)	Baseline	Filter	Weighted (α = 0.9/0.5)
Spec2Vec	**58.3**	49.4	53.4 / 43.7	**85.7**	61.8	81.9 / 62.6
WCS	**56.5**	45.6	50.3 / 43.5	**84.1**	58.1	78.8 / 62.5
CS	**28.6**	27.3	23.1 / 20.8	**56.6**	49.9	48.1 / 41.0
LLM4MS	**66.3**	58.0	62.7 / 53.1	**92.7**	65.6	89.5 / 76.2

"Baseline" refers to the baseline methods. The "Filter" applies the base-peak matching heuristic before searching using a baseline method. The "Weighted" combines the score of a baseline method with that obtained from the heuristic using different weighting parameters (α). The best row-wise results are highlighted in **bold**.

more effective structural similarity than existing methodologies. To investigate this ability, we analyzed the structural similarity of the top-ranked retrieved candidates by examining the relationship between the spectral embeddings and the corresponding chemical structures. We employed the Tanimoto coefficient, a widely used metric to quantify the structural similarity between two molecules based on their chemical fingerprints[41]. The Tanimoto coefficient ranges from 0 to 1, where 1 indicates that both vectors have the identical structure; and 0 indicates no shared structural features.

Our analysis focused on two key aspects: the retrieval of exact matches (Recall@1) and the retrieval of structurally analogous compounds within the top 10 candidates. Figure 4A presents the fraction of exact matches achieved by LLM4MS, Spec2Vec, and WCS as a function of the Tanimoto coefficient between the query compound and the top-ranked retrieved candidate. As expected, all methods predominantly retrieve exact matches with a Tanimoto score of 1.0. Notably, LLM4MS exhibits a higher fraction of exact matches compared to Spec2Vec and WCS across the higher Tanimoto score bins (0.8-1.0), suggesting a greater ability to identify the correct compound as the top hit when it is present in the library.

To evaluate the ability to retrieve structurally analogous compounds, we examined the distribution of the maximum Tanimoto coefficient within the top 10 retrieved candidates for each query, excluding any exact matches (Fig. 4B). This analysis provides insights into whether the methods can identify compounds with significant structural overlap, even if they are not identical to the query. The histogram reveals that LLM4MS tends to find top-10 candidates with higher maximum Tanimoto scores compared to Spec2Vec and WCS. This indicates that the LLM-derived embeddings are more effective at ranking structurally similar compounds higher in the retrieval list. For a substantial fraction of queries, LLM4MS retrieves at least one compound within the top 10 with a Tanimoto score exceeding 0.4, a commonly used threshold to define structural analogues. This suggests that LLM4MS is better at capturing and leveraging structural relationships in the spectral embedding space.

Figure 4C provides illustrative examples of the query spectra and their top 10 candidates retrieved by each method, along with the corresponding Tanimoto coefficients. The left panel showcases a scenario where LLM4MS retrieves top candidates with high structural similarity to the query, reflected by high Tanimoto scores. In contrast, Spec2Vec and WCS retrieve compounds with lower structural similarity. The right panel presents a case where the retrieved candidates exhibit more structural diversity across all methods, highlighting the challenges in retrieving highly similar analogues for certain query compounds. The average Tanimoto coefficient for the top 10 candidates further quantifies the overall structural similarity of the retrieval results for each method. In both examples, LLM4MS demonstrates a tendency to retrieve candidates with, on average, higher structural similarity to the query compared to those of the other methods. These findings collectively underscore the ability of LLM4MS to generate spectral embeddings that effectively encode the structural information, leading to improved retrieval of both exact matches and structurally related compounds. More results can be found in Supplementary Note 4.

LLM4MS software

To facilitate the practical application of LLM4MS for compound identification, we developed a user-friendly software tool based on Python programming languages. The graphical user interface (GUI) of LLM4MS, as depicted in Fig. 5, is designed for intuitive operation, allowing researchers to easily leverage the power of LLM-derived spectral embeddings for mass spectrometry data analysis. The LLM4MS software runs on Windows 7, 10, and 11 operating systems.

Upon launching the LLM4MS software, the pre-computed embedding vectors for the million-scale in-silico EI-MS library are loaded, as indicated by the status messages in the interface. This pre-loading ensures that the system is ready for rapid querying without requiring users to wait for computationally intensive embedding generation. The software provides a straightforward interface for users to input their query mass spectra. As shown in the "Input M/Z" and "Input Intensity" fields, users can directly paste or load their experimental spectral data in a standard m/z-intensity pair format. Once the query spectrum is provided, clicking the "Search" button initiates the nearest neighbor search within the pre-computed LLM4MS embedding space. The software efficiently retrieves the top-ranked candidate compounds based on the proximity of their embeddings to the query spectrum's embedding. The progress of the search is indicated in the status window. The search results are then displayed in a clear and organized manner, as illustrated in the lower portion of Fig. 5. For each query spectrum, the software presents a list of the top-ranked candidate compounds, along with their corresponding chemical structures. This visual representation allows users to quickly assess the plausibility of the identification results. By clicking on a specific candidate compound, users can access further information, such as the predicted spectrum from the in-silico library (if available) and potentially a side-by-side comparison with the query spectrum.

The LLM4MS software also includes functionalities for clearing the input fields ("Clear" button) and loading pre-existing models or data ("Load Model" button), offering flexibility in how users interact with the tool. The intuitive design and efficient search capabilities of the LLM4MS software make it a valuable resource for researchers seeking to rapidly and accurately identify compounds from their mass spectrometry data using the benefits of LLM-derived spectral embeddings. Our LLM4MS software is available at https://doi.org/10.5281/zenodo.17036712. The current implementation of LLM4MS software, using the in-silico library and the Faiss indexing technique, allows efficient and accurate mass spectrometry queries.

Discussion

Mass spectral matching, the comparison of experimentally acquired mass spectra against curated libraries of known compounds, is a fundamental task for compound identification in MS-based analyses. Accurate identification is crucial across diverse scientific domains, including metabolomics, proteomics, and synthetic chemistry, enabling the elucidation of complex

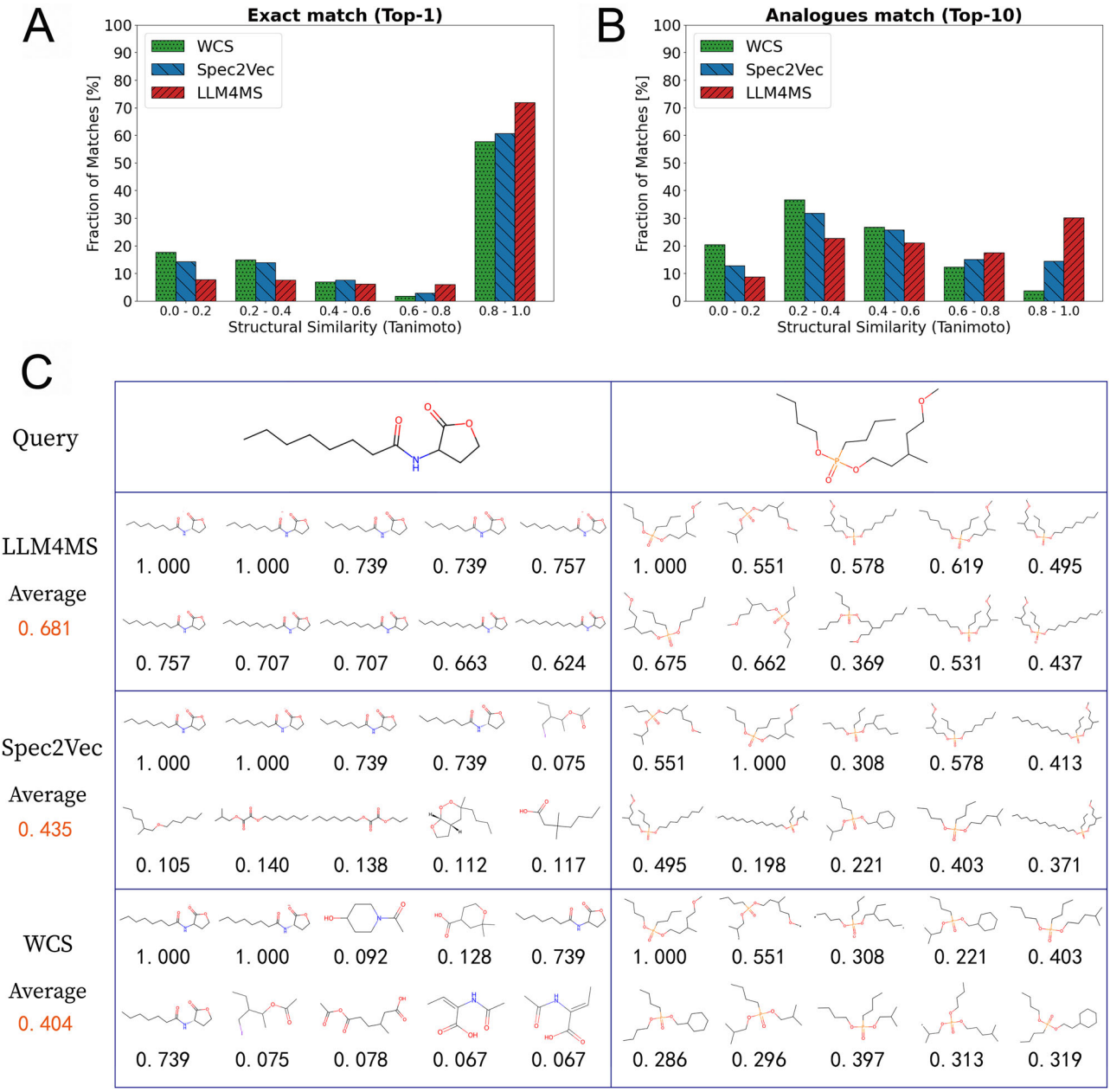


Fig. 4 | Performance of LLM4MS in relation to structural similarity of retrieved candidates. A Fraction of exact matches (Recall@1) for LLM4MS, Spec2Vec, and WCS across different bins of structural similarity (Tanimoto coefficient) between the query spectra and the top-ranked retrieved candidates. **B** Histogram displaying the distribution of the maximum structural similarity score (Tanimoto coefficient) observed within the top 10 retrieved candidates (excluding any exact matches to the query compound) for each query using LLM4MS, Spec2Vec, and WCS. For each query spectrum, the highest Tanimoto score between the query's chemical structure and the structures of the top 10 matched compounds in the library is plotted. **C** Examples of query spectra and the top 10 retrieved candidates by LLM4MS, Spec2Vec, and WCS. The Tanimoto coefficient between the query and each retrieved candidate is displayed below the corresponding structure. The average Tanimoto coefficient for the top 10 candidates for each method is also provided. The left panel shows examples where LLM4MS achieves a high structural similarity for the top matches, while the right panel illustrates a case where the methods retrieve structurally diverse candidates.

sample compositions. The reliability of this process depends on the similarity metric's ability to discern subtle structural differences reflected in fragmentation patterns. While traditional methods like WCS and more recent machine learning approaches such as Spec2Vec have advanced the field, they typically prioritize global spectral features or learned data correlations. As a result, they lack an inherent capacity to interpret spectra based on underlying chemical principles or domain-specific knowledge, often leading to incorrect identifications, especially for structurally distinct compounds exhibiting superficial spectral similarities. Recently, large language models have demonstrated a remarkable ability to acquire and apply domain-specific knowledge across various fields. Their capacity to "reason" about complex data, potentially informed by vast amounts of chemical literature and data absorbed during pre-training, offers a compelling avenue for spectral interpretation. As illustrated by the inference processes of representative LLMs (Fig. 1), these models prioritize chemically significant features, such as base peak alignments and the presence of key high-mass fragments, applying chemical heuristics that traditional metrics often overlook. This suggests that LLMs possess the latent potential to achieve more chemically intuitive and accurate spectral matching.

Our illustration of LLM inference (Fig. 1B) utilized DeepSeek-R1 and GPT-4o as they were representative of the state-of-the-art at the time of our study, demonstrating that the capacity for chemical reasoning is not

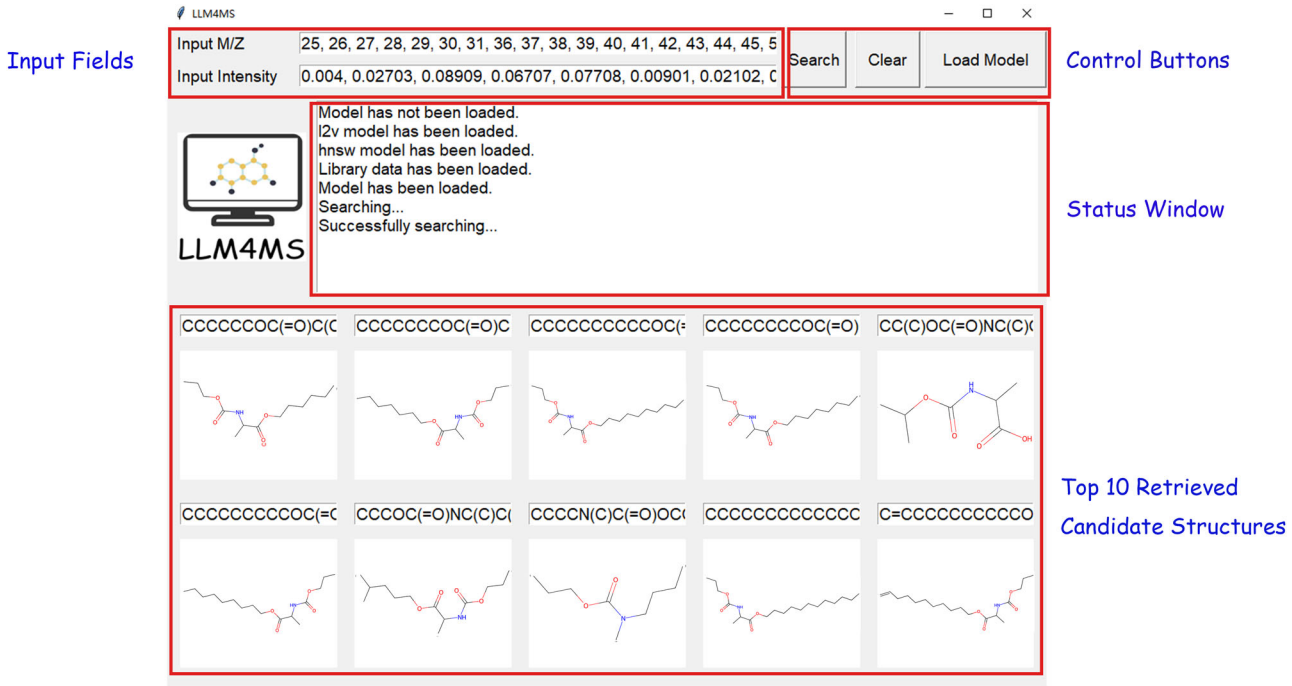


Fig. 5 | LLM4MS software. GUI of the LLM4MS tool designed for rapid compound identification. The interface features: Input Fields where users can paste or load query mass spectra (m/z and intensity pairs); Control Buttons to initiate library searching ("Search"), clear input fields ("Clear"), or load pre-computed models/data ("Load Model"); a Status Window displaying operational feedback such as model loading and search progress; and a results panel showcasing the Top 10 Retrieved Candidate Structures from the in-silico library based on the LLM-derived embedding similarity to the query spectrum.

confined to a single model but is a newly discovered property common to different leading LLMs. We acknowledge that the accessibility and cost of such models are an obstacle for widespread adoptions. The choice of a closed-source model like GPT-4o alongside a more accessible model like DeepSeek-R1 highlights this dichotomy. Indeed, the significant financial cost and practical challenges associated with using a large proprietary model for high-throughput tasks, such as embedding an entire large spectral library, are prohibitive. This challenge was a principal motivation for our work. The core contribution of LLM4MS is not the direct application of a specific large model, but rather the development of a methodology to fine-tune a smaller, accessible, open-weight model (of which Llama 3.1-8B is an example). Through the proposed approach, we distill the powerful reasoning capability of a cost-effective, efficient, and publicly available tool for accurate compound identification in mass spectrometry, ensuring that the benefits of this approach are reproducible and broadly accessible to the scientific community without incurring prohibitive API costs.

Capitalizing on this potential, we introduce LLM4MS, a method designed to harness the power of pre-trained LLMs for enhanced electron ionization mass spectral matching. Unlike methods such as Spec2Vec that typically learn relationships directly from spectral data representations (Fig. 2A), LLM4MS textualizes each mass spectrum and leverages the extensive chemical and scientific knowledge encoded within a purpose-fine-tuned LLM to generate richer, more chemically informed embeddings (Fig. 2B). Our comprehensive evaluations demonstrate the effectiveness of this approach. LLM4MS significantly outperforms existing methods in identification accuracy, achieving substantially higher Recall@x rates against a million-scale in-silico library (Fig. 3C). Furthermore, when coupled with an approximate nearest neighbor search technique, LLM4MS enables ultra-fast query speeds, processing nearly 15,000 queries per second while maintaining high accuracy, making it suitable for high-throughput workflows (Fig. 3D).

A crucial aspect of evaluating advanced spectral similarity methods lies in their ability to accurately reflect molecular structural similarity, commonly benchmarked using metrics like the Tanimoto score derived from molecular fingerprints. This focus stems from the fundamental principle that structurally related molecules often exhibit similar fragmentation behaviors under mass spectrometry conditions, which should ideally translate into high spectral similarity. Although methods like Spec2Vec[16] represented an advancement over traditional cosine scores by learning fragment co-occurrence patterns from spectral data, their capacity to model structural relationships remains inherently limited by the information contained solely within the spectra they are trained on. It is worth noting that while Spec2Vec was originally developed for the tandem mass spectrometry (LC-MS/MS) data[16], a recent work has successfully adapted the method for electron ionization (EI) mass spectrometry in the context of large-scale library searching. Specifically, the FastEI method[17] established this precedent by adjusting the key parameters to suit the EI-MS data characteristics, such as its nominal mass resolution. This involved modifying the precision of peak representation and the binning of m/z values into "words" (*e.g.*, `peak@80`) to accommodate the characteristics of the EI-MS data. To ensure a fair and robust baseline comparison in our study, we have adopted the identical Spec2Vec implementation settings as used in FastEI for processing the EI-MS data[17].

LLM4MS, in contrast, produces embeddings derived from a large language model pre-trained on extensive and diverse knowledge domains, potentially encompassing underlying chemical principles and reaction mechanisms beyond simple spectral patterns. This allows LLM4MS to generate spectral representations that are more attuned to chemically significant structural features. Consequently, the similarity scores derived from LLM4MS embeddings exhibit a notably stronger correlation with Tanimoto-based structural similarity compared to baseline methods (Fig. 4). This enhanced alignment with molecular structure is directly advantageous for mass spectral matching; it increases confidence in identifying exact matches and, crucially, significantly improves the reliability of retrieving structurally related analogues from large-scale libraries, thereby facilitating the annotation of unknown compounds and the extraction of meaningful chemical insights from complex datasets.

A robust evaluation framework is crucial for validating different spectral matching methods. Our primary benchmark evaluation follows the precedent established by Yang et al. in their work on the in-silico EI-MS

library[17]. Their main evaluation similarly utilized a "closed-set" approach, where the query spectra (from NIST17) corresponded to compounds whose structures were given to the prediction model's training process. This benchmark design is valuable in assessing an algorithm's ability to bridge the inherent gap between experimental spectra and their in-silico predicted counterparts. Building upon this established framework, our study introduces two key improvements to ensure a more comprehensive and rigorous evaluation. First, we utilized the more recent NIST23 library for our query spectra, in contrast to the NIST17 library used in the original study. The NIST23 library is generally considered to contain higher-quality spectra due to potential advancements in instrumentation and curation standards, ensuring our benchmark reflects the current state-of-the-art in experimental data. Second, we employ an evaluation procedure to assess the generalization capability of an algorithm for unseen compounds. While Yang et al. have a similar test by using ten extra compounds[17], our evaluation is substantially more comprehensive, based on a large set of 2618 compounds that were verified to be absent from the NEIMS model's training data. Together, these enhancements provide a more stringent and reliable assessment of the true performance and generalization capabilities of spectral matching methods.

Approximate nearest neighbor search machine learning techniques are commonly employed to accelerate embedding-based spectral matching against large-scale libraries, overcoming the bottleneck of exhaustive comparisons. Previous work, such as FastEI, has demonstrated the effectiveness of this approach, utilizing the HNSW[37] algorithm with Spec2Vec embeddings to achieve significant speedups over traditional methods, reporting search speeds equivalent to approximately 238 QPS (calculated from the reported 0.0042 seconds per query)[17]. In this study, we extended this investigation by evaluating the performance of our LLM4MS embeddings with a broader range of popular, open-source ANNS techniques: Annoy, Faiss, HNSWlib, and NMSLIB. Our findings confirm the high compatibility of LLM4MS embeddings with these ANNS techniques, enabling ultra-fast search speeds that reached up to approximately 14,440 QPS with HNSWlib, while maintaining high accuracy. As illustrated in Fig. 3D, these ANNS implementations present a clear trade-off between search velocity and Recall@1 accuracy, which users can navigate by selecting specific libraries and optimizing their hyperparameters (e.g., *ef_construction*, *M* for HNSW variants). While other ANNS techniques exist, such as those based on locality-sensitive hashing[42–44] or product quantization[45–47], the libraries tested here represent widely-used, state-of-the-art options covering different indexing paradigms. Our evaluation across these diverse libraries demonstrates the general applicability and effectiveness of ANNS for accelerating LLM4MS-based spectral searches (Fig. 5), providing users with robust and practical options to balance throughput and precision according to their research needs.

The power of LLMs stems from two key aspects relevant to scientific applications: the vast knowledge base acquired during pre-training on diverse, large-scale datasets, and their inherent ability to effectively associate related concepts and represent semantic similarities as proximity in high-dimensional embedding spaces[48,49]. LLM4MS leverages these capabilities by adapting a pre-trained LLM through fine-tuning, enabling it to generate chemically informed spectral embeddings. This process infuses the learned chemical knowledge into the embeddings, allowing them to capture subtle structural nuances and resulting in the superior mass spectral matching performance demonstrated in our benchmarks. While the current LLM4MS framework successfully utilizes these optimized embeddings in conjunction with ANNS for fast and accurate spectral retrieval, it does not yet harness the explicit reasoning and natural language generation capabilities of LLMs to explain the rationale behind specific matches. Integrating the LLM's "thought process", akin to the chemical reasoning observed in the illustrative examples (e.g., Fig. 1B) into the retrieval workflow, presents a compelling avenue for future work. Such explainability could provide users with valuable and interpretable justifications for high-similarity matches, complementing the quantitative scores, and enhancing confidence in compound identification.

Methods

Mass spectra preprocessing and textualization

To convert raw mass spectral data into a format suitable for processing by a large language model (LLM), we implemented a multi-step preprocessing and textualization pipeline (Fig. 6A). For the Tanimoto similarity-guided fine-tuning stage (Fig. 6C), we randomly selected 50,000 mass spectra from the in-silico EI-MS library. This subset size was chosen primarily due to the computational intensity of calculating the pairwise Tanimoto scores required for similarity-guided fine-tuning across the entire million-scale library; while we anticipate that training on more spectra could potentially yield further improvements, the Tanimoto calculation represented a practical bottleneck.

Each selected mass spectrum was then subjected to a peak filtering process. Preliminary experiments indicated that numerous low-intensity peaks could introduce noise and potentially interfere with the model's ability to learn relevant spectral patterns. Therefore, we tested various filtering thresholds (k, representing the number of top peaks retained) and found that retaining up to the 30 peaks with the highest intensities ($k = 30$) provided an effective balance between enriching structurally informative signals and removing potential noise (detailed results in Supplementary Note 5). This intensity-based selection assumes that peaks representing key fragment ions typically exhibit higher intensities compared to the background noise[50,51]. Thus, selecting the top 30 peaks aims to capture the most diagnostic signals essential for structural characterization while filtering out the majority of low-intensity background interfering signals.

Following the peak selection, each spectrum was transformed into a structured textual representation. This textualization step involved assigning specific labels to different types of peaks based on their relative intensity and mass-to-charge (m/z) ratio. Specifically, the peak with the maximum intensity was designated as the `base_peak`, often representing the most stable fragment ion or a primary fragmentation pathway. The peak having the highest m/z value among the selected 30 was labeled as the `max_peak`; this feature is frequently correlated with the intact precursor ion or a high-mass fragment, providing insights into the compound's molecular weight. A curated set comprising the top 3 of the most intense peaks, excluding the `base_peak`, were identified as `key_peaks`. These peaks typically correspond to the major fragment ions from significant fragmentation events. Finally, the remaining peaks in the set of selected top 30 were categorized as `extra_peaks`, capturing further spectral details. This explicit differentiation between peak types within the textual format was designed to enable the subsequent LLM to more effectively discern and leverage the distinct informational value carried by different characteristic features during model training and analysis.

Development of the LLM4MS model

The development pipeline for the specialized LLM4MS commenced with Llama3.1-8B[52] serving as the foundational pre-trained large language model (`Base LLM`) in our primary experiments, as illustrated in the overall process schematic (Fig. 6). To enhance its capability for generating semantically meaningful vector representations (embeddings) relevant to our domain, this `Base LLM` first underwent preliminary unsupervised training adapting the `LLM2Vec` method[53] (detailed in Fig. 6B). This preliminary stage, employing Masked Next Token Prediction `MNTP`) and Simple Contrastive Learning of Sentence Embeddings (`SimCSE`)[54], typically utilizes large-scale text corpora (such as English Wikipedia[55], as used in the original LLM2Vec study) to improve the model's embedding capabilities before domain-specific adaptation. Subsequently, the `LLM2Vec` model served as the basis for a domain-specific fine-tuning process aimed at creating the final LLM4MS model, guided by pairwise spectral similarity (Fig. 6C). This fine-tuning stage leveraged the structured textual representations of the 50,000 preprocessed spectra derived from the in-silico library. The core objective was to adapt `LLM2Vec` to interpret the unique syntax and semantic content of the textualized mass spectral data. To achieve this, the fine-tuning was guided by Tanimoto scores calculated between pairs of mass spectra from the training set; the spectrum yielding the highest Tanimoto score relative to

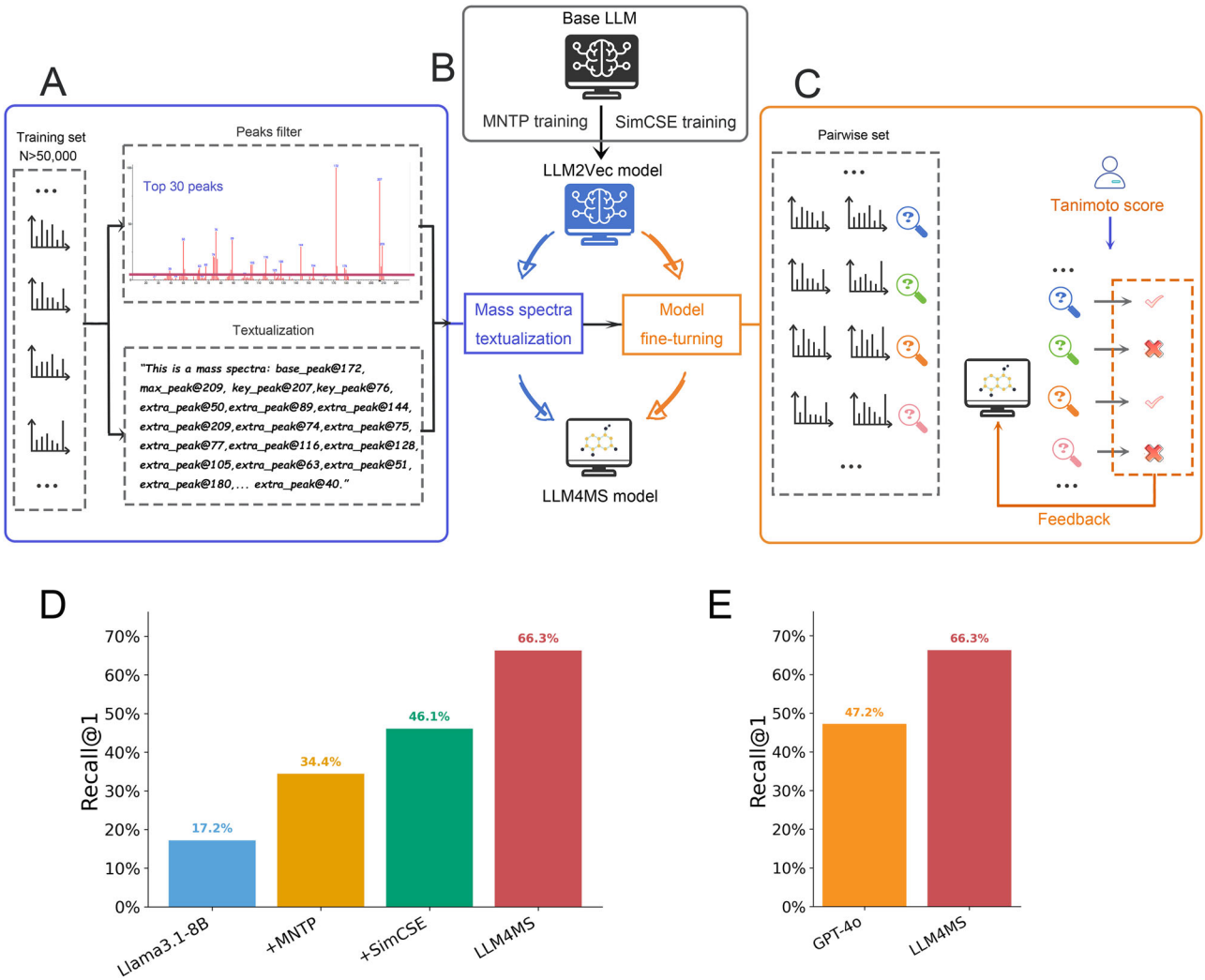


Fig. 6 | Overview of the LLM4MS model fine-tuning process and model evaluation. A Depiction of the mass spectra textualization pipeline, illustrating the conversion of filtered spectral data into a structured text format suitable for language model processing. **B** Schematic of the training phase where a base large language model undergoes MNTP and SimCSE training to be adapted into the LLM2Vec model, specialized for embedding representations. **C** Illustration of the fine-tuning procedure guided by pairwise spectral similarity. Tanimoto scores are calculated for pairs of mass spectra to generate feedback signals, which are subsequently used to refine the LLM2Vec model into the final LLM4MS model. **D** The contribution of each training component to the model performance, measured by Recall@1. Performance is shown for the base Llama 3.1-8B model, after adding MNTP training (+MNTP), after further adding SimCSE training (+SimCSE, resulting in LLM2Vec), and for the final LLM4MS model after Tanimoto-guided fine-tuning. **E** Performance comparison of the specialized LLM4MS model against the general-purpose GPT-4o model on the spectral matching.

an anchor spectrum was designated as the positive pair, while others served as the negative pairs. This Tanimoto-derived similarity information provided the feedback signal for optimizing the model's parameters via a contrastive learning objective. This strategy enabled the final LLM4MS model to effectively learn representations encoding the structural similarity between spectra.

To systematically evaluate the contribution of each stage in our model development process, we performed an ablation study, measuring performance using the Recall@1 metric (Fig. 6D). Recall@1 quantifies the frequency at which the most similar spectrum according to the ground truth (identical compound) is ranked first by embedding similarity. Starting with the foundational Llama 3.1-8B model, which served as our baseline, a Recall@1 of 17.2% was achieved, reflecting its general pre-trained capabilities without any adaptation. The introduction of `MNTP` training led to a notable increase in performance, yielding a Recall@1 of 34.4%. Subsequently incorporating `SimCSE` training to produce the `LLM2Vec` model further enhanced performance, reaching 46.1% Recall@1, indicating the value of contrastive learning for embedding quality. The final LLM4MS model, obtained after the domain-specific fine-tuning guided by the Tanimoto similarity, demonstrated the most substantial improvement, achieving a final Recall@1 of 66.3%. This progressive enhancement underscores the efficacy of each training component, particularly the crucial role of the similarity-guided fine-tuning in tailoring the model for accurate spectral representation and comparison.

We further assessed the efficacy of our specialized LLM4MS model by comparing its performance against GPT-4o, a leading state-of-the-art, general-purpose large language model, on the same spectral similarity task (Fig. 6E). The performance of GPT-4o was evaluated by generating embeddings using its official API, employing the identical textual representation of mass spectra constructed for our LLM4MS model (as detailed for Fig. 6A). Using the Recall@1 metric, the generalist GPT-4o model achieved a score of 47.3%. This score, while serving as a strong baseline for comparison, is considerably higher than that of the base Llama 3.1-8B model (17.2%), likely reflecting GPT-4o's substantially larger model size and the consequently greater amount of implicitly encoded world and potentially scientific knowledge acquired during its pre-training. In the direct

comparison, our LLM4MS model, specifically trained and fine-tuned on mass spectral data using our proposed pipeline, achieved a significantly higher Recall@1 of 66.3%. This result clearly demonstrates the advantages of domain specialization and the effectiveness of our fine-tuning approach, as LLM4MS substantially outperforms a highly capable general-purpose model on this specific scientific task. Furthermore, it is noteworthy that the immense parameter count typical of models like GPT-4o renders further domain-specific fine-tuning for tasks such as mass spectra analysis computationally prohibitive or impractical with standard resources, highlighting the utility of our approach in adapting a moderately sized model effectively. All training and fine-tuning procedures utilized spectra derived solely from the in-silico library, ensuring complete separation from the NIST23-derived test set used for final performance evaluation in the Results section. All experiments are conducted on the same Linux machine: AMD 128-core CPU with each core running at 2 GHz and 1 TB RAM and one RTX A6000 GPU with 48GB VRAM.

Similarity calculation for LLM4MS

Following the generation of high-dimensional vector embeddings (n-dimensional vectors, where $n = 4096$ for Llama 3.1-8B) for each textualized mass spectrum using the LLM4MS model, a quantitative metric is necessary to compare pairs of spectra based on these learned representations. For this purpose, we employed cosine similarity, a standard and widely adopted metric for measuring the similarity between vector embeddings produced by LLMs and other high-dimensional representations in natural language processing and information retrieval[56–58].

The rationale for using cosine similarity stems from the understanding that in high-dimensional spaces typical of LLM embeddings, the direction of the vectors often encodes more critical semantic information than their magnitudes. Vector magnitudes can be influenced by factors such as term frequency or other model-internal characteristics that may not directly correlate with the underlying similarity of the concepts (in this case, mass spectra) being represented. Cosine similarity effectively mitigates the impact of vector length by normalizing the vectors, focusing solely on the cosine of the angle between them. This provides a measure bounded between -1 (indicating vectors pointing in opposite directions) and 1 (indicating vectors pointing in the same direction), where values closer to 1 signify higher similarity in the embedding space. Its computational efficiency and established effectiveness in capturing semantic relatedness make it highly suitable for comparing LLM4MS embeddings.

Given two n-dimensional embedding vectors, $\mathbf{A}$ and $\mathbf{B}$, derived from two distinct mass spectra via the LLM4MS model, the cosine similarity, S_C, is calculated as their normalized dot product:

$$S_C(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\| \mathbf{A} \| \| \mathbf{B} \|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} \mathbf{A}_i^2} \sqrt{\sum_{i=1}^{n} \mathbf{B}_i^2}}$$

Here, $\mathbf{A} \cdot \mathbf{B}$ represents the dot product of the vectors $\mathbf{A}$ and $\mathbf{B}$, while $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ denote their respective Euclidean (L2) norms. The resulting scalar S_C quantifies the similarity captured by the LLM4MS embeddings. Thus, after the complex spectral relationships and chemical knowledge are encoded into the high-dimensional embeddings by the LLM4MS model, the relatively simple and computationally efficient cosine similarity metric proves sufficient for achieving high-performance spectral matching, as demonstrated in our results.

Two strategies in applying the base peak matching heuristic

To evaluate the impact of the base peak matching heuristic on the baseline methods, we implemented a filter strategy and a weighted strategy. The filter strategy aims to reduce the entire search space. For each query spectrum, a subset of the reference library was created, containing only library spectra whose base peak m/z value matched the query's base peak. The standard similarity algorithms (Spec2Vec, WCS, and CS) were then executed exclusively on this pre-filtered subset to identify the top candidates. The weighted strategy, in contrast, modifies the final similarity score without altering the search space. This was achieved by creating a similarity score (S_{new}) that linearly combines the similarity score of the baseline (S_{orig}) and a binary base peak matching score (S_{bp}). The S_{bp} score was set to 1 if the base peaks of the query and library spectra matched, and 0 otherwise. The final score was calculated as:

$$S_{new} = \alpha \cdot S_{orig} + (1 - \alpha) \cdot S_{bp}, \quad (1)$$

where $\alpha \in [0, 1]$ is a parameter that controls the weight of the heuristic. A value of $\alpha = 1$ recovers the original score, while $\alpha = 0$ relies solely on the base peak match.

Data availability

The data utilized in this study originate from multiple sources. The primary reference dataset, the million-scale in-silico EI-MS library, was generated as described by Yang et al.[17] and is publicly available. The query spectra comprising the test set were selected from the commercially available NIST 2023 Tandem Mass Spectral Library (mainlib). The foundational large language models employed, such as Llama 3.1-8B[52], are publicly accessible through platforms like Hugging Face. Embeddings from the GPT-4o model were obtained using its official API.

Code availability

The LLM4MS model, software and example Python files are available in the Zenodo repository at https://doi.org/10.5281/zenodo.17036712[59]. The repository contains the LLM4MS model, the software GUI, example files, and a README file with instructions for installation and usage. The source code is released under a noncommercial use license.

Received: 6 May 2025; Accepted: 12 September 2025;
Published online: 04 November 2025

References

1. Thirunavukarasu, A. J. et al. Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
2. Liu, W. et al. Drbioright 2.0: an llm-powered bioinformatics chatbot for large-scale cancer functional proteomics analysis. *Nat. Commun.* **16**, 2256 (2025).
3. Vaghefi, S. A. et al. Chatclimate: Grounding conversational ai in climate science. *Commun. Earth Environ.* **4**, 480 (2023).
4. M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
5. Duponchel, L., Rocha de Oliveira, R. & Motto-Ros, V. Large language models (such as chatgpt) as tools for machine learning-based data insights in analytical chemistry. *Anal. Chem.* **97**, 6956–6961 (2025).
6. Boiko, D. A., Kozlov, K. S., Burykina, J. V., Ilyushenkova, V. V. & Ananikov, V. P. Fully automated unconstrained analysis of high-resolution mass spectrometry data with machine learning. *J. Am. Chem. Soc.* **144**, 14590–14606 (2022).
7. Stein, S. E. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem.* **84**, 7274–7282 (2012).
8. Stein, S. E. & Scott, D. R. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass Spectrom.* **5**, 859–866 (1994).
9. Gangnon, R. E. & Clayton, M. K. A weighted average likelihood ratio test for spatial clustering of disease. *Stat. Med.* **20**, 2977–2987 (2001).
10. McLafferty, F., Hertel, R. & Villwock, R. Probability based matching of mass spectra. rapid identification of specific compounds in mixtures. *Org. Mass Spectrom.* **9**, 690–702 (1974).
11. Moorthy, A. S., Wallace, W. E., Kearsley, A. J., Tchekhovskoi, D. V. & Stein, S. E. Combining fragment-ion and neutral-loss matching during mass spectral library searching: a new general purpose algorithm applicable to illicit drug identification. *Anal. Chem.* **89**, 13261–13268 (2017).

12. Dührkop, K. et al. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
13. Dührkop, K. et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nat. Biotechnol.* **39**, 462–471 (2021).
14. van Der Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl Acad. Sci.* **113**, 13738–13743 (2016).
15. Huber, F., van der Burg, S., van der Hooft, J. J. & Ridder, L. Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminformatics* **13**, 84 (2021).
16. Huber, F. et al. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS Comput. Biol.* **17**, e1008724 (2021).
17. 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).
18. de Jonge, N. F. et al. Ms2query: reliable and scalable ms2 mass spectra-based analogue search. *Nat. Commun.* **14**, 1752 (2023).
19. Kearsley, A. J. & Roberts, M. Similarity measures of mass spectra in hilbert spaces. *Technical Note (NIST TN), National Institute of Standards and Technology* (2024).
20. Kapp, E. A. et al. An evaluation, comparison, and accurate benchmarking of several publicly available ms/ms search algorithms: sensitivity and specificity analysis. *Proteomics* **5**, 3475–3490 (2005).
21. Treen, D. G. et al. Simile enables alignment of tandem mass spectra with statistical significance. *Nat. Commun.* **13**, 2510 (2022).
22. Frank, A. M., Pesavento, J. J., Mizzen, C. A., Kelleher, N. L. & Pevzner, P. A. Interpreting top-down mass spectra using spectral alignment. *Anal. Chem.* **80**, 2499–2505 (2008).
23. Cooper, B. T. et al. Hybrid search: a method for identifying metabolites absent from tandem mass spectrometry libraries. *Anal. Chem.* **91**, 13924–13932 (2019).
24. Xing, S. et al. Retrieving and utilizing hypothetical neutral losses from tandem mass spectra for spectral similarity analysis and unknown metabolite annotation. *Anal. Chem.* **92**, 14476–14483 (2020).
25. 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).
26. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci.* **109**, E1743–E1752 (2012).
27. Zheng, Y. et al. Large language models for scientific discovery in molecular property prediction. *Nat. Mach. Intell.* **7**, 437–447 (2025).
28. Guo, D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv:2501.12948 (2025).
29. Achiam, J. et al. Gpt-4 technical report. arXiv:2303.08774 (2023).
30. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using umap. *Nat. Biotechnol.* **37**, 38–44 (2019).
31. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 (2020).
32. Kim, H. W. et al. Npclassifier: a deep neural network-based structural classification tool for natural products. *J. Nat. Products* **84**, 2795–2807 (2021).
33. Bittremieux, W., Meysman, P., Noble, W. S. & Laukens, K. Fast open modification spectral library searching through approximate nearest neighbor indexing. *J. Proteome Res.* **17**, 3463–3474 (2018).
34. Bittremieux, W. et al. Open access repository-scale propagated nearest neighbor suspect spectral library for untargeted metabolomics. *Nat. Commun.* **14**, 8488 (2023).
35. Bernhardsson, E. Annoy (Approximate Nearest Neighbors Oh Yeah). *GitHub*. https://github.com/spotify/annoy (2018).
36. Douze, M. et al. The faiss library. *arXiv:2401.08281* (2024).
37. Malkov, Y. A. & Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 824–836 (2018).
38. Malkov, Y. A. & Yashunin, D. A. hnswlib: A fast approximate nearest neighbors algorithm using hierarchical navigable small world graphs. *GitHub*. https://github.com/nmslib/hnswlib (2018).
39. Naidan, B., Boytsov, L., Malkov, Y., Frederickson, B. & Novak, D. nmslib: Non-metric space library. *Github*. https://github.com/nmslib/nmslib (2018).
40. Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Cent. Sci.* **5**, 700–708 (2019).
41. Godden, J. W., Xue, L. & Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients. *J. Chem. Inf. Computer Sci.* **40**, 163–166 (2000).
42. Kulis, B. & Grauman, K. Kernelized locality-sensitive hashing. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 1092–1104 (2011).
43. Huang, Q., Feng, J., Fang, Q., Ng, W. & Wang, W. Query-aware locality-sensitive hashing scheme for lp norm. *VLDB J.* **26**, 683–708 (2017).
44. Chierichetti, F., Kumar, R., Panconesi, A. & Terolli, E. On the distortion of locality sensitive hashing. *SIAM J. Comput.* **48**, 350–372 (2019).
45. Ge, T., He, K., Ke, Q. & Sun, J. Optimized product quantization. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 744–755 (2013).
46. Xu, D., Tsang, I. W. & Zhang, Y. Online product quantization. *IEEE Trans. Knowl. Data Eng.* **30**, 2185–2198 (2018).
47. Jegou, H., Douze, M. & Schmid, C. Product quantization for nearest neighbor search. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 117–128 (2010).
48. Chang, Y. et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **15**, 1–45 (2024).
49. Zhao, H. et al. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.* **15**, 1–38 (2024).
50. Renard, B. Y. et al. When less can yield more–computational preprocessing of ms/ms spectra for peptide identification. *Proteomics* **9**, 4978–4984 (2009).
51. Reiz, B., Kertész-Farkas, A., Pongor, S. & Myers, M. P. Chemical rule-based filtering of ms/ms spectra. *Bioinformatics* **29**, 925–932 (2013).
52. Meta AI. Llama 3.1 8B. https://huggingface.co/meta-llama/Llama-3.1-8B (2024).
53. BehnamGhader, P. et al. Llm2vec: Large language models are secretly powerful text encoders. arXiv:2404.05961 (2024).
54. Gao, T., Yao, X. & Chen, D. Simcse: Simple contrastive learning of sentence embeddings. arXiv:2104.08821 (2021).
55. Huang, S., Xu, Y., Geng, M., Wan, Y. & Chen, D. Wikipedia in the era of llms: Evolution and risks. arXiv:2503.02879 (2025).
56. Liu, B. et al. How good are llms at out-of-distribution detection? arXiv:2308.10261 (2023).
57. Juvekar, K. & Purwar, A. Cos-mix: cosine similarity and distance fusion for improved information retrieval. arXiv:2406.00638 (2024).
58. Banerjee, D., Singh, P., Avadhanam, A. & Srivastava, S. Benchmarking llm powered chatbots: methods and metrics. arXiv:2308.04624 (2023).
59. Xu, Y., Ma, Y., Xu, W., Yang, Z. & Ting, K. M. LLM4MS: A large language model for deriving spectral embeddings for accurate compound identification in mass spectrometry. *Zenodo*. https://doi.org/10.5281/zenodo.17036712 (2025).

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (NNSFC, grants no. 92470116 and no. 62076120) to K.M.T.

Author contributions

Y.X. and Y.M. conceived the idea and developed the LLM4MS model and software. The conceptualization and experimental design were done by Y.X. Experiments and data analysis were performed by W.X., Y.M. and Z.Y. Y.X., Y.M., K.M.T. and W.X. tested and debugged the program. Y.X., Z.Y. and K.M.T. wrote the original manuscript. Y.X. and K.M.T. contributed to the revision of the manuscript in response to the reviewers' comments. K.M.T. supervised the project. All authors discussed the project and contributed to the writing and preparation of the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42004-025-01708-7.

Correspondence and requests for materials should be addressed to Yang Xu or Kai Ming Ting.

Peer review information *Communications Chemistry* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025