Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Annotate this: A comparative benchmark of MS/MS annotation algorithms across reference compounds and complex samples

Marta Costa

costa@metasysx.com

metaSysX GmbH

Álvaro Cuadros-Inostroza

metaSysX GmbH

Sylwia Kierszniowska

metaSysX GmbH

Research Article

Keywords: metabolomics, mass spectrometry, benchmarking, annotation, MS2, metabolite identification

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11509200/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. The authors are employees of metaSysX GmbH, a company providing metabolomics and lipidomics analytical services. The authors declare no other competing interests.

Annotate this: A comparative benchmark of MS/MS annotation algorithms across reference compounds and complex samples

Marta Costa[1,2*], Álvaro Cuadros-Inostroza[1] and Sylwia Kierszniowska[1]

[1*]metaSysX, Potsdam-Golm, 14476, Germany.
[2] Freie Universität Berlin, Department of Biology, Chemistry, Pharmacy, Arnimallee 22, 14195, Berlin, Germany.

*Corresponding author(s). E-mail(s): costa@metasysx.com;
Contributing authors: inostroza@metasysx.com;
kierszniowska@metasysx.com;

Abstract

Introduction Metabolite annotation from LC-MS/MS data remains a bottleneck in untargeted metabolomics. A growing number of computational tools employing diverse strategies, from spectral library matching and *in silico* fragmentation to machine learning and *de novo* structure generation, have been developed. However, the diversity of tools makes it difficult to determine which approaches provide reliable annotations. Direct comparisons under consistent, realistic conditions are needed to guide tool selection and improve annotation confidence.

Objectives We benchmarked five popular annotation tools: MS2Query, SIRIUS, MSNovelist, MetFrag and ChemDistiller, under default parameters and without prior sample knowledge.

Methods Tools were evaluated on two datasets: 357 in-house acquired MS/MS spectra from reference compounds with unambiguous ground truth, and 194 expert-annotated spectra from NIST SRM 1950 plasma. Performance was assessed across formula and structure accuracy, coverage, precision, Tanimoto similarity to ground truth, and score reliability.

Results MS2Query and SIRIUS emerged as top performers with complementary profiles: MS2Query achieved the highest annotation coverage (100% on reference compounds) while SIRIUS yielded the highest conditional accuracy (78%). MSNovelist performed comparably to SIRIUS. MetFrag performance

was database-dependent, with PubChem yielding near-complete annotation failure and KEGG enabling moderate precision. ChemDistiller excelled at formula annotation but underperformed structurally.
Conclusion No single tool dominates across all metrics. Tool selection should be guided by balance between coverage and precision, with score-based filtering and multi-tool consensus strategies recommended to improve annotation confidence.

Keywords: metabolomics, mass spectrometry, benchmarking, annotation, MS2, metabolite identification

1 Introduction

Untargeted metabolomics enables the detection of thousands of molecular features from complex biological samples, however, the majority of these features remain structurally uncharacterized (Novoa-del Toro & Witting, 2024). Metabolite annotation is thus one of the major bottlenecks in metabolomics experiments, limiting biological interpretation and cross-study comparability. Tandem mass spectrometry (MS/MS) provides additional structural information through fragmentation patterns and has gained a substantial role in metabolite annotation workflows. Nonetheless, the chemical diversity of metabolites, the prevalence of isomers, and the incomplete coverage of spectral databases make confident structural annotation a persistent challenge.

Several computational approaches have been developed to facilitate MS/MS-based metabolite annotation. Traditional spectral library matching relies on comparison with experimentally acquired reference spectra, while *in silico* fragmentation methods rank candidate structures according to model-predicted fragmentation patterns. More recently, Machine Learning (ML) and deep learning (DL) approaches have emerged, enabling the prediction of molecular properties and structures directly from spectral data. Additional strategies include analogue search methods, which identify structurally related compounds, and hybrid approaches that integrate multiple sources of evidence to improve annotation confidence. These strategies are reviewed in depth by Nguyen et al. (2019). The methods can also be categorized according to the direction of information flow as either MS-to-compound (MS2C) approaches, which infer molecular identity from experimental spectra, or compound-to-MS (C2MS) approaches, which predict spectra from candidate chemical entities and subsequently evaluate their agreement with the observed data (Wang et al., 2021).

Despite the number of available tools, their comparative performance remains difficult to assess. Recent recommendations for benchmarking computational metabolomics annotation tools have highlighted the lack of standardized datasets, evaluation metrics, and reporting practices, which limits direct comparison between methods (de Jonge et al., 2022). Performance is frequently assessed using reference standards alone, or a small number of features (Hupatz et al., 2025) which may not reflect the complexity of real biological samples. Updated evaluations are needed given the rapid emergence of ML-based approaches and substantial updates to existing tools since the most recent completed Critical Assessment of Small Molecule Identification

(CASMI) challenge (CASMI, 2022) was held in 2022, with the next edition scheduled for December 2026 (CASMI, 2026). Hoffmann et al. (2023) demonstrated that benchmark design and hidden information leakage can inflate performance estimates, highlighting the importance of independent evaluations. Using spectra acquired independently of the tools' training data, with known identities, enables a more objective assessment while reducing benchmark-specific bias and spectral data leakage.

In this study, we benchmarked five annotation workflows representing distinct annotation strategies: MS2Query (de Jonge et al., 2023), MetFrag (Ruttkies et al., 2016), ChemDistiller (Laponogov et al., 2018), SIRIUS (Dührkop et al., 2019), and MSNovelist (Stravs et al., 2022). Performance was evaluated using a reference compound mixture with known identities and a complex National Institute of Standards and Technology (NIST) plasma sample. Plasma was chosen as a complex yet tractable matrix widely used in metabolomics and clinical biomarker research (Gowda et al., 2008). By comparing annotation yield, accuracy and precision, chemical similarity between predictions and ground truth, chemical-space coverage, as well as computational efficiency, accessibility and ease of use, we aimed to assess how different MS2-based annotation approaches perform under controlled and biologically relevant conditions.

2 Methods

2.1 LC-MS/MS Data Acquisition and Annotation

Reference compound standards were prepared individually from stock solutions and extracted using a modified MTBE:MeOH:H_2O biphasic protocol, with the polar phase collected for measurement. The NIST Standard Reference Material 1950 (SRM 1950) (Metabolites in Frozen Human Plasma) (Phinney et al., 2013) was prepared and measured in-house under identical conditions. Both samples were analyzed on a Waters ACQUITY RP-UPLC system coupled to a Thermo Fisher Scientific QExactive Orbitrap mass spectrometer, acquiring full scan and data-dependent MS/MS spectra in both positive and negative ionization modes using HCD fragmentation (normalized collision energy 25, 100–1500 m/z). Full acquisition parameters are provided in Online Resource 2.

A random subset of MS/MS files was selected from the metaSysX in-house MS/MS database for ground-truth annotation. Detected MS/MS spectra, comprising precursor m/z, retention time (RT) and fragment ions, were matched against the expected values for each standard compound recorded in the database, using tolerances of 5 ppm and 0.05 min for m/z and RT, respectively, with fragment ion matching providing additional confirmation. This yielded 357 clean, unambiguously annotated spectra (178 positive, 179 negative mode), each corresponding to the compound's main MS1 adduct (MSI level 1; Summer et al. (2007)). Only these 357 spectra were provided as input to the algorithms for benchmarking. For the NIST plasma sample, MS1-level annotation was first performed, after which MS1-annotated features were matched to MS/MS precursor ions by m/z and RT, yielding 194 spectra with confident in-house annotations (114 positive, 80 negative mode) at MSI level 1 except for 16 spectra matching several isomeric standards, that served as the evaluable ground truth for the

complex sample benchmarking stage. The full NIST plasma MS/MS dataset, exceeding 9,000 spectra per ionization mode, was submitted to all benchmarking algorithms without prior filtering (except as noted in Section 2.3).

2.2 Algorithm Selection

Five annotation tools were selected for benchmarking based on their use in the metabolomics literature, current operability, accessibility, batch-processing capability, and ability to accept standardized MS/MS spectra and return ranked candidate annotations. The selected tools represent complementary annotation strategies, including spectral library searching (MS2Query v1.5.4), molecular formula and structure prediction (SIRIUS v6.3.3), *de novo* structure generation (MSNovelist), *in silico* fragmentation (MetFrag v2.6.8), and a combination of spectral/structural and fragmentation-based approaches (ChemDistiller v0.1). Notably, MS2Query is primarily an analogue-search approach, identifying structurally related compounds rather than requiring exact spectral matches; its scoring is additionally intended to distinguish likely true matches from structural analogues. MSNovelist is integrated with SIRIUS and uses intermediate outputs from CSI:FingerID (Dührkop et al., 2015; Shen et al., 2014), making its performance partly dependent on the SIRIUS workflow. The main differences between the algorithms in candidate generation, ranking and adduct handling are summarized in Table 1.

2.3 Annotation Parameters and Experimental Design

All algorithms were run with default or recommended parameters and, as in real-world use, without prior knowledge of sample composition. A mass tolerance of 5 ppm was applied across all tools. Data were submitted to each algorithm separately for each ionization mode (positive and negative). For the SIRIUS suite, all available structure databases were included in the search space to ensure no candidate structures were excluded *a priori*. Because SIRIUS and MSNovelist run substantially longer on high-mass spectra, precursors with $m/z > 850$ were excluded for these methods, which affected only the complex sample dataset. ChemDistiller was similarly run with all databases in its default folder and the available linear kernel function. MetFrag, which requires a single database per run, was run with PubChem and KEGG (Kanehisa et al., 2025) independently for the reference compound dataset; based on results from that dataset only the KEGG database was used for the complex sample analysis.

2.4 Data Processing and Evaluation

Given the scale of the comparison and in line with recommendations by de Jonge et al. (2022), the primary analysis focuses on the top-ranked prediction from each method. Compound identifiers were harmonized across methods using PubChem (queried during the first half of 2026) to resolve missing or mismatched identifiers. Stereoisomers were treated as correct annotations. Chemical similarity between predicted and reference structures was quantified using the Tanimoto coefficient (Bajusz et al., 2015), calculated from SMILES strings using Morgan fingerprints following canonicalization and tautomer normalization, as implemented in RDKit (version 2025.09.1) (RDKit

Table 1 Comparison of the annotation algorithms benchmarked in this study and their main methodological characteristics.

Algorithm	Strategy	Candidate generation	Candidate ranking	Adduct handling
MS2Query	MS2C; spectral analogue search	Spectral library; no precursor-m/z pre-filter	MS2DeepScore[1] retrieves top library spectra, followed by a random-forest model combining five features to re-rank candidates	Ionization mode is specified as positive or negative; precursor adduct is not explicitly used as a candidate-generation constraint
SIRIUS	MS2C	Molecular formulas from precursor mass, followed by structure database search using CSI:FingerID	SIRIUS/ZODIAC[2] scores molecular formulas; CSI:FingerID ranks structures using predicted molecular fingerprints	Adduct/ionization type is explicitly provided or detected and influences formula annotation and downstream structure annotation; multiple possible adducts can be considered
MSNovelist	MS2C; *de novo*	Structures generated directly from the molecular fingerprint predicted by CSI:FingerID; no structure database retrieval required	RNN-based *de novo* generation and ranking of candidate structures	Adduct handling is inherited from the preceding SIRIUS workflow, which provides the molecular formula/fingerprint input
MetFrag	C2MS	Structures retrieved from databases by precursor mass	*In silico* fragmentation score based on explained fragments and bond dissociation information, with optional orthogonal evidence	Precursor ion/adduct type is explicitly specified and used to calculate theoretical fragment masses and perform candidate scoring
ChemDistiller	MS2C + C2MS	Candidates retrieved by precursor m/z from large compound databases and optionally filtered by formula/elemental composition	Combined fingerprint- and fragmentation-based scores (TotalScore)	Adduct and isotope information are applied to pre-computed fragmentation patterns to match the experimental ion form

[1]Huber et al. (2021); [2]Ludwig et al. (2020)

contributors, 2025). Compound class assignments were obtained using ClassyFire (Djoumbou Feunang et al., 2016). All downstream analyses and visualizations were performed in Python 3.10.18.

3 Results

3.1 Dataset Description

The benchmark was conducted in two stages. The first stage focuses on a reference compound mixture of 357 spectra (178 positive, 179 negative mode, Figure 1a-b), comprising in-house standards with known chemical identity. The dataset spans a broad chemical space (Figure 1c), with a predominance of organic acids and derivatives (particularly amino acids, peptides and analogues) alongside organic oxygen compounds, including carbohydrates and carbohydrate conjugates. Its m/z distribution, however, is skewed towards lower values, with no precursor exceeding m/z 700 (Figure 1d), a pattern that, while partly reflecting the random selection, is also consistent with the known detection bias in LC-MS-based metabolomics and the compositional biases of reference databases (Broeckling, 2026).

The second stage uses data acquired from NIST SRM 1950 human plasma, representing a real-world complex biological matrix. Spectra were minimally processed and no absolute ground truth exists for the full dataset, but 114 positive and 80 negative-mode spectra were expert-annotated in-house. This subset is described in Figure 2, with the main superclasses of compounds represented being organic acids and derivatives (largely amino acids, peptides and analogues), organoheterocyclic compounds, strongly due to purines and purine derivatives, followed by a diverse group of lipids and lipid-like molecules.

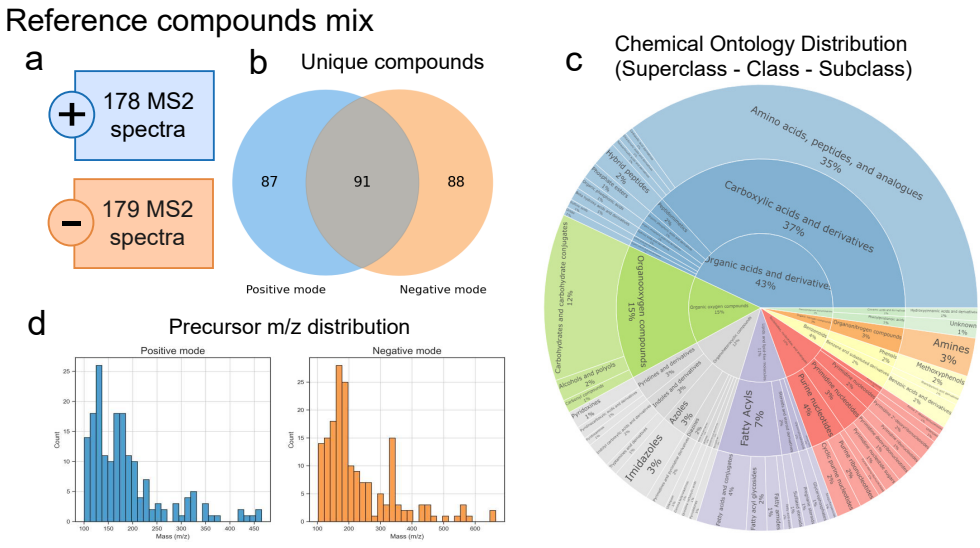


Fig. 1 Descriptors of the reference compounds dataset. a) The reference compound mixture comprises 178 MS/MS spectra acquired in positive mode and 179 in negative mode. **b)** These correspond to 87 unique compounds in positive mode, 88 in negative mode, and 91 in both. **c)** The spectra are predominantly assigned to organic acids and derivatives (mainly amino acids, peptides, and analogues), followed by organooxygen compounds (carbohydrates and conjugates), organoheterocyclic compounds, lipid-like molecules, and nucleosides, nucleotides, and analogues. **d)** Precursor m/z values range from approximately 100 to 700 in both ionization modes.

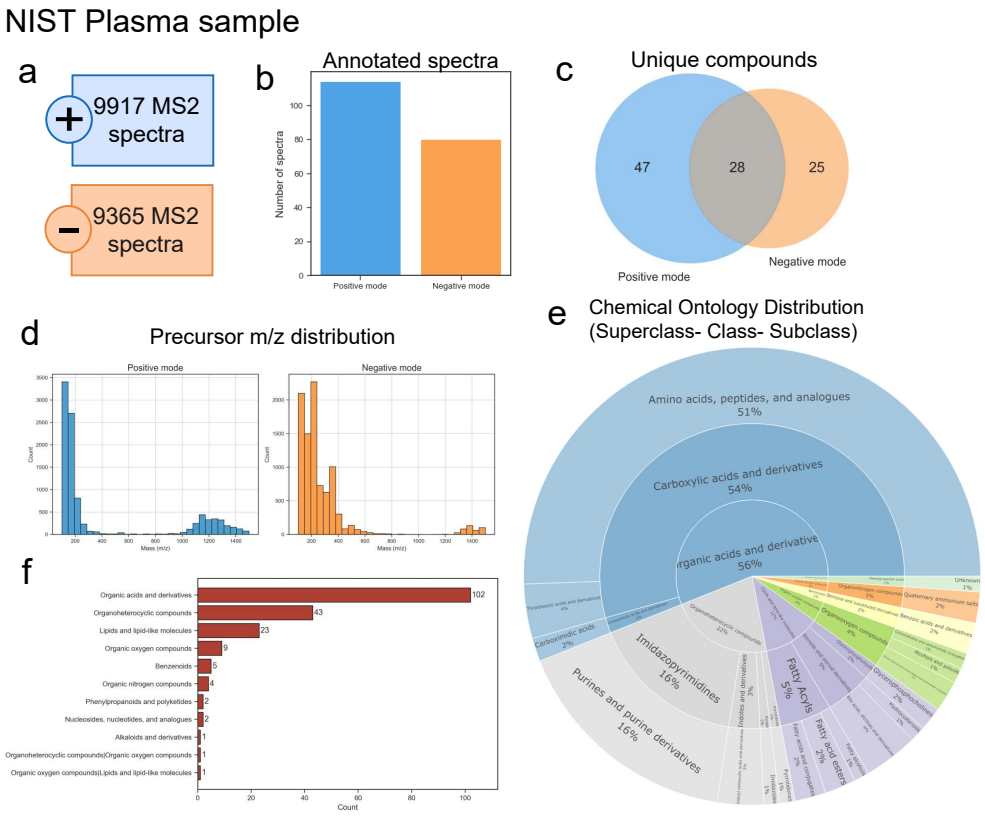


Fig. 2 Descriptors of the NIST datasets. **a)** The NIST blood plasma dataset contains 9,917 spectra in positive mode and 9,365 in negative mode. **b)** Of these, 114 positive-mode and 80 negative-mode spectra were expertly annotated. **c)** These annotations represent 47 unique compounds in positive mode, 25 in negative mode, and 28 shared between both modes. **d)** The precursor m/z distribution spans approximately 100–1500 with most spectra between 100 and 400 m/z and a smaller peak between 1,000 and 1500 m/z. **e–f)** The annotated compounds are mainly organic acids and derivatives (predominantly amino acids, peptides, and analogues), followed by purines and purine derivatives (16%) and lipids and lipid-like molecules (12%), including fatty acyls (5%).

3.2 Benchmarking on Reference Compounds

3.2.1 Formula and Structure Annotation Performance

Formula annotation was generally less challenging than structure annotation, with the top-performing methods (ChemDistiller, MS2Query, SIRIUS and MSNovelist) achieving high formula accuracies (>80%; Online Resource 1, Fig. S1a). Structure annotation, which requires discriminating isomeric candidates, lowered performance for all methods. Here, MS2Query, SIRIUS and MSNovelist emerged as the best-performing methods with comparable overall accuracies (Figure 3a-b). However, distinct annotation strategies resulted in different error profiles: MS2Query prioritized

coverage by returning predictions for nearly all spectra, whereas SIRIUS and MSNovelist produced fewer predictions but achieved higher conditional accuracy. ChemDistiller and MetFrag (KEGG) showed intermediate performance, while MetFrag (PubChem) performed poorly, highlighting the difficulty of ranking correct candidates within very large chemical search spaces. Performance trends were broadly consistent across ionization modes, albeit slightly higher accuracies in positive mode (Online Resource 1, Fig. S1b).

3.2.2 Chemical Similarity of Predictions

Tanimoto similarity between predicted and reference structures gives a more nuanced view than binary correctness (Figure 3c). All methods except MetFrag with PubChem achieved a median Tanimoto similarity of 1.0, indicating that exact or near-exact matches were frequently recovered. Mean Tanimoto similarities were highest for SIRIUS (0.85), followed by MSNovelist (0.83), MS2Query (0.77), MetFrag KEGG (0.71), ChemDistiller (0.65) and MetFrag PubChem (0.18). Notably, despite relying on analogue search rather than direct structure identification, MS2Query produced chemically similar candidates approaching the performance of SIRIUS and MSNovelist. Pairwise comparisons confirmed that SIRIUS achieved significantly higher similarities than all other methods, while MSNovelist outperformed all except SIRIUS. MS2Query yielded significantly higher similarities than ChemDistiller and both MetFrag configurations (Online Resource 1, Fig. S2).

3.2.3 Performance by Chemical Class

We next evaluated annotation performance at the chemical superclass level (Figure 3d). Nucleosides, nucleotides and analogues were the best-annotated superclass overall, with particularly high class-specific accuracy for MS2Query, MetFrag (KEGG) and ChemDistiller. SIRIUS and MSNovelist showed complementary strengths in organoheterocyclic compounds and phenylpropanoids and polyketides. In contrast, lipids and lipid-like molecules, and organic oxygen compounds consistently proved the most challenging to annotate across methods, despite the latter being the second most abundant superclass in the dataset. Similar trends were observed at finer taxonomic resolution, with class- and subclass-level analyses largely recapitulating the superclass results (Online Resource 1, Figures S1c–d).

3.2.4 Method Score Reliability

Scatter plots of method-assigned scores against Tanimoto similarity (Online Resource 1, Fig. S1e) revealed that MS2Query scores show the strongest linear relationship with chemical similarity ($R^2 = 0.380$), with high-scoring predictions clustering strongly around Tanimoto = 1. In contrast, all other methods showed little linear association ($R^2 \leq 0.03$). This is further supported by the score distribution analysis displayed in Figure 3e: MS2Query shows the most significant score separation between correct and incorrect predictions (Mann-Whitney U, $p < 0.001$), followed by MSNovelist ($p < 0.01$) and MetFrag (KEGG) ($p < 0.05$). No significant score separation was

observed for SIRIUS, ChemDistiller or MetFrag (PubChem), though the latter had only 2 correct predictions, precluding meaningful statistical comparison.

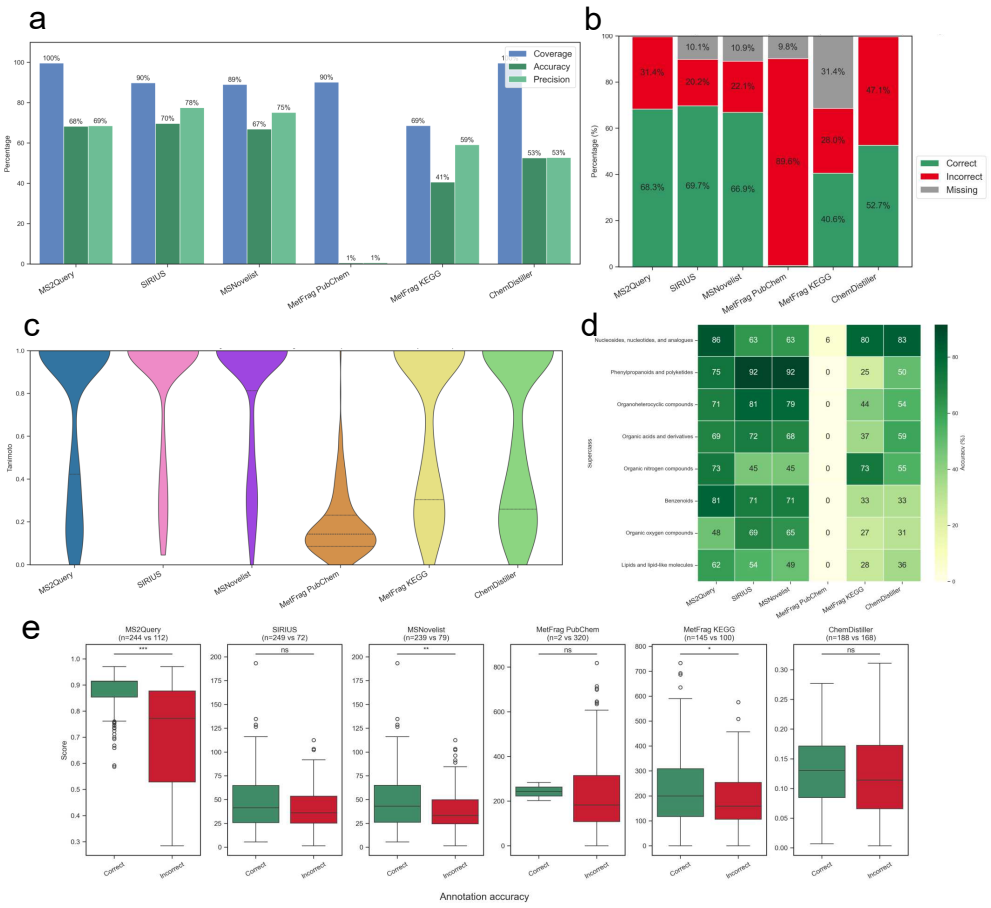


Fig. 3 Reference compound benchmark results. a) MS2Query and ChemDistiller achieved the highest coverage, whereas SIRIUS showed the highest conditional accuracy. **b)** MS2Query, SIRIUS, and MSNovelist correctly annotated approximately 70% of spectra; however, SIRIUS and MSNovelist produced substantially more unannotated spectra (NAs) and fewer incorrect annotations. MetFrag (PubChem) performed poorly overall, while MetFrag (KEGG) produced the highest number of NAs. ChemDistiller yielded approximately equal proportions of correct and incorrect annotations. **c)** These trends are reflected in the Tanimoto similarity distributions, with MS2Query, SIRIUS, and MSNovelist strongly skewed toward 1.0, MetFrag (PubChem) concentrated around 0.1, and MetFrag (KEGG) and ChemDistiller showing similar distributions with a secondary peak near 0.2. **d)** Superclass-specific performance revealed that MS2Query achieved the highest accuracy for benzenoids and nucleosides, nucleotides, and analogues (>80%), whereas SIRIUS and MSNovelist performed best on phenylpropanoids and polyketides and organoheterocyclic compounds. MetFrag (PubChem) underperformed across all superclasses, while MetFrag (KEGG) and ChemDistiller achieved their highest accuracies for nucleosides, nucleotides, and analogues (all superclasses included contained $n \geq 3$ spectra). **e)** Prediction confidence scores were significantly higher for correct than incorrect annotations for MS2Query, MSNovelist, and MetFrag (KEGG) (one-sided Mann-Whitney U test; *** p <0.001, ** p <0.01, * p <0.05, ns: p >0.05).

3.3 Benchmarking on NIST Plasma Complex Sample

3.3.1 Annotation Coverage and Inter-tool Agreement

In the complex plasma matrix, annotation coverage dropped markedly for all methods (Figure 4a). MS2Query remained the highest-coverage tool, annotating 94% of spectra in positive mode, although this dropped substantially to 58% in negative mode. Most other methods annotated between 26% and 46% of all spectra, with coverage consistently higher in positive mode, the sole exception being ChemDistiller, which showed slightly higher coverage in negative mode. MSNovelist outperformed SIRIUS in negative mode coverage, and MetFrag (KEGG only) achieved the lowest coverage.

Inter-tool agreement was assessed across all submitted spectra (Figure 4b). Of the 19,282 spectra considered across both modes, only 49 received concordant annotations from all five methods. The most frequent pairwise agreement was observed between MetFrag (KEGG) and ChemDistiller, followed closely by SIRIUS and MSNovelist. A further 190 spectra were annotated concordantly by all methods except MetFrag. MS2Query contributed the largest number of unique predictions not corroborated by other tools.

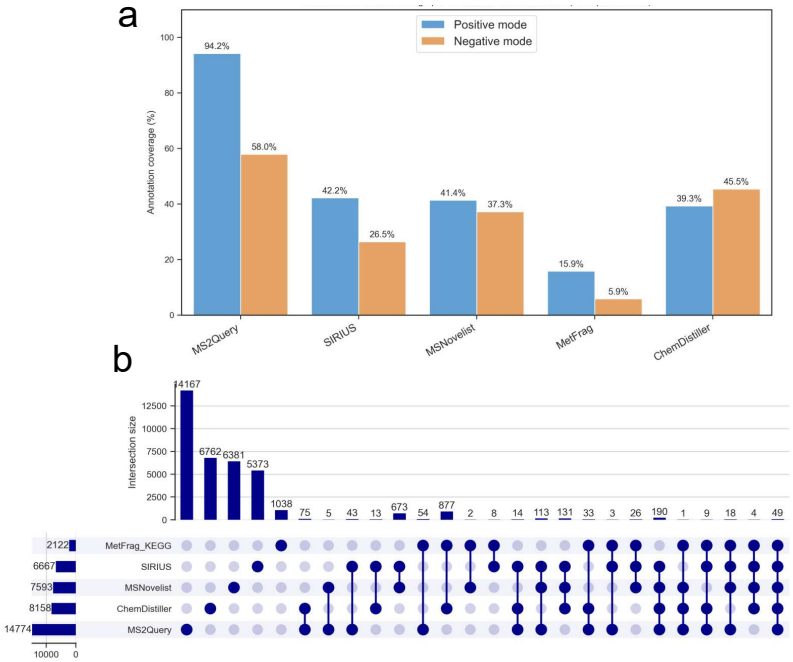


Fig. 4 NIST blood plasma dataset results. a) Considering all spectra, MS2Query achieved the highest coverage, whereas MetFrag (KEGG) achieved the lowest. Overall, annotation coverage was higher in positive than in negative ionization mode. **b)** Regarding inter-tool agreement, MetFrag (KEGG) and ChemDistiller showed the greatest concordance, followed by SIRIUS and MSNovelist. Only 49 spectra received the same annotation from all algorithms, while 190 spectra were assigned the same annotation by all methods except MetFrag.

3.3.2 Performance on Annotated Spectra

Restricting the analysis to the 194 in-house annotated spectra enabled direct comparison with the reference compound benchmark, although interpretation should be tempered by the smaller sample size. As observed for the reference compounds, formula annotation was generally less challenging than structure annotation, with ChemDistiller, MS2Query, SIRIUS and MSNovelist achieving the highest formula accuracies (Online Resource 1, Fig. S3a). Structure annotation accuracy decreased by 8–20 percentage points across all methods relative to the reference benchmark (Figure 5a-b), with MS2Query emerging as the best-performing method, followed by SIRIUS and MSNovelist. As before, MS2Query prioritized coverage and SIRIUS and MSNovelist conditional accuracy. ChemDistiller and MetFrag showed lower overall performance. Performance was consistently lower in negative ionization mode, primarily due to an increase in missing rather than incorrect predictions, while MS2Query, SIRIUS and MSNovelist largely maintained their conditional accuracy across ionization modes (Online Resource 1, Fig. S3b).

3.3.3 Chemical Similarity of Predictions

Tanimoto similarity distributions in this task largely replicated those obtained for the reference compounds dataset (Figure 5c). MS2Query, SIRIUS and MSNovelist retained a median Tanimoto score of 1.0, with mean scores of 0.79, 0.80 and 0.74 respectively. ChemDistiller showed a mean of 0.62 (median 0.70), and MetFrag (KEGG) presented the lowest values in this task, with a mean of 0.59 and median of 0.54, a more pronounced drop relative to its reference compound performance, consistent with the greater structural diversity and complexity of the plasma sample. In pairwise Wilcoxon tests (see Online Resource 1, Fig. S2), SIRIUS again produced significantly higher Tanimoto similarities than most other methods. However, MS2Query did not differ significantly from either SIRIUS (adjusted $p = 0.41$) or MSNovelist (adjusted $p = 0.11$), indicating comparable structural similarity among the top-performing methods in this dataset. Nevertheless, MS2Query, SIRIUS and MSNovelist significantly outperformed ChemDistiller and MetFrag. No significant difference was observed between ChemDistiller and MetFrag (KEGG) (adjusted $p = 0.09$).

3.3.4 Performance by Chemical Class

The annotated NIST dataset covers a distinct region of chemical space, dominated by organic acids and derivatives (56%), followed by organoheterocyclic compounds (22%) and lipids and lipid-like molecules (12%) (Figure 2e–f). Overall, MS2Query showed the most consistent performance across chemical superclasses (Figure 5d), while SIRIUS and MSNovelist also achieved high conditional accuracy for the predominant organic acids and derivatives. Results for less represented superclasses (organic oxygen compounds, benzenoids, organic nitrogen compounds) were more variable, as they rest on few annotated spectra. At finer taxonomic resolution, indoles and derivatives emerged as the most reliably annotated class, whereas steroids and steroid derivatives were not predicted by either SIRIUS or MSNovelist in the plasma dataset (Online Resource 1, Figures S3c–d).

3.3.5 Method Score Reliability

Method-assigned scores showed trends similar to those observed for the reference compounds (Online Resource 1, Fig. S3e), with MS2Query again exhibiting the strongest correlation between score and Tanimoto similarity ($R^2 = 0.24$). For SIRIUS and MSNovelist, predictions with scores above approximately 60 consistently corresponded to exact structural matches (Tanimoto = 1.0), suggesting the potential for meaningful confidence thresholds. Consistently, correct predictions received significantly higher scores than incorrect predictions for all methods (Figure 5e), although based only on the in-house annotated subset.

To evaluate the practical utility of these scores, we examined the trade-off between annotation accuracy and coverage across score thresholds (Online Resource 1, Fig. S4). MS2Query displayed a consistent operating point at a score of approximately 0.8 in both datasets, where accuracy and coverage converged at roughly 75-80%. Comparable intersections were observed for SIRIUS and MSNovelist at scores of approximately 15-25, whereas MetFrag (KEGG) and ChemDistiller reached lower accuracy–coverage equilibria. Overall, these results indicate that method-assigned scores can serve as useful confidence estimates, particularly for MS2Query, SIRIUS, MSNovelist, and MetFrag (with KEGG database) while ChemDistiller scores remain comparatively less discriminative.

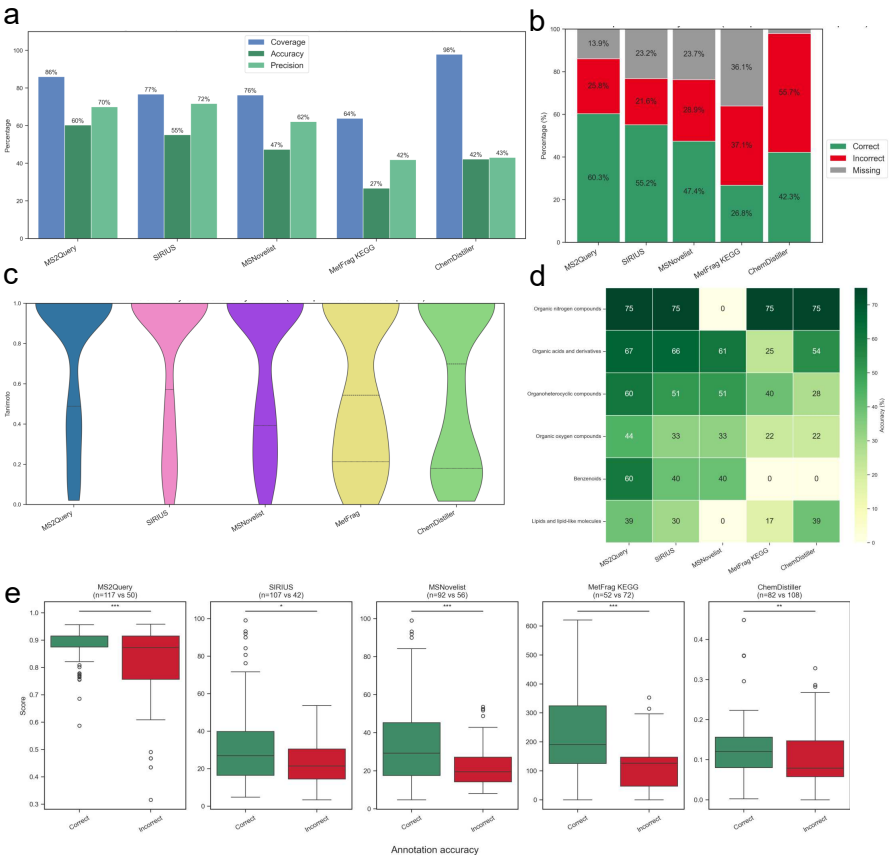


Fig. 5 NIST annotated spectra results. a) For the in-house annotated subset, annotation coverage ranged from 64% to 98%, with the highest coverage achieved by ChemDistiller. SIRIUS again showed the highest conditional accuracy (72%). **b)** Structural annotation performance followed a similar trend to the reference compound benchmark, with MS2Query producing the highest proportion of correct annotations, followed by SIRIUS, MSNovelist, ChemDistiller, and MetFrag. MetFrag generated the largest number of NAs, whereas ChemDistiller generated the fewest. **c)** These trends are, once more, reflected in the Tanimoto similarity distributions, with MS2Query, SIRIUS, and MSNovelist again showing distributions strongly skewed toward 1.0. **d)** Organic nitrogen compounds were the best-annotated superclass for all methods except MSNovelist. Overall, performance across superclasses was more consistent than in the reference compound benchmark, although MetFrag and ChemDistiller showed noticeably higher accuracies for organic nitrogen compounds than for the remaining superclasses (all represented superclasses contained $n \geq 3$ spectra). **e)** Prediction confidence scores were again significantly higher for correct than incorrect annotations for MS2Query, SIRIUS, MSNovelist, and MetFrag. ChemDistiller also showed a significant difference in this set, although the separation between correct and incorrect predictions was less pronounced (one-sided Mann-Whitney U test; *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns: $p > 0.05$).

4 Discussion

This study systematically compared five widely used annotation tools across reference compounds with known ground truth and a complex biological matrix (NIST

SRM 1950 plasma). Together, the evaluated methods represent the major algorithmic paradigms currently used for computational metabolite annotation (Chi et al., 2024; Coler et al., 2024), while also reflecting the growing influence of AI and DL in the field (Pomyen et al., 2020).

MS2Query and SIRIUS consistently emerged as the strongest performers, although with distinct strengths. MS2Query achieved the highest coverage, returning predictions for nearly every queried spectrum through analogue searching, whereas SIRIUS prioritized reliability by abstaining from uncertain predictions and consequently achieved the highest conditional accuracy. MSNovelist closely followed SIRIUS despite its fundamentally different *de novo* structure generation strategy, but offered little practical advantage because its predictions depend on CSI:FingerID fingerprints. Incorrect SIRIUS predictions were generally reproduced by MSNovelist, and the latter rarely generated annotations where SIRIUS failed. The only notable exception was the full NIST plasma negative mode dataset, where MSNovelist achieved approximately achieved higher coverage than SIRIUS (37.3% vs 26.5%), although these additional predictions could not be validated. These findings highlight the fundamental trade-off between annotation coverage and confidence: MS2Query is well suited to exploratory analyses prioritizing broad annotation, whereas SIRIUS is preferable when prediction reliability is the primary objective.

The performance of MetFrag further demonstrated that database selection is a major determinant of annotation quality. When queried against PubChem, MetFrag frequently returned chemically dissimilar top candidates, consistent with the difficulty of ranking among the vast number of structurally similar compounds using *in silico* fragmentation alone (Wolf et al., 2010). Restricting searches to the smaller, curated KEGG database substantially improved structural accuracy, albeit with reduced coverage. ChemDistiller showed the opposite pattern, consistently achieving the highest molecular formula accuracy while performing comparatively poorly at structural identification, suggesting that it often resolved the correct formula but not the correct isomer. This likely reflects both limitations of its scoring strategy and the age of its underlying models and databases, which have not been updated since 2018.

MS2Query scores were the most informative, correlating best with Tanimoto similarity and separating correct from incorrect predictions most clearly. MSNovelist and MetFrag (KEGG) scores also separated correct from incorrect predictions in both datasets, whereas SIRIUS scores did so only in the plasma spectra. Nonetheless, high SIRIUS and MSNovelist scores were informative: in the annotated plasma spectra, all predictions scoring above 60 were exact structural matches. Score threshold analyses further identified method-specific balances between annotation coverage and accuracy, whereas ChemDistiller scores showed limited discriminative value.

Inter-tool agreement across the complete plasma dataset was remarkably low, with only 49 of 19,282 spectra receiving identical top-1 annotations from all five methods. Agreement was highest between MetFrag (KEGG) and ChemDistiller, and between SIRIUS and MSNovelist, reflecting their shared methodological foundations. These findings suggest that consensus-based annotation strategies may substantially improve confidence, although inevitably at the expense of coverage. This limitation of individual annotation tools has also been highlighted by Leonov et al. (2026), who reported

that even combining multiple state-of-the-art tools frequently required expert curation to obtain biologically meaningful annotations from plasma samples.

Consistent with Russo et al. (2024), ease of use is likely to influence tool adoption independently of algorithmic accuracy. SIRIUS, together with its integrated MSNovelist workflow, offered the most accessible interface, while MS2Query required only modest command-line experience and accepted the widely used MGF format. In contrast, MetFrag batch processing and ChemDistiller required more complex setup procedures and less standardized input formats. Regarding computational performance, ChemDistiller showed the shortest runtimes, aided by parallel execution, whereas MetFrag, particularly when searching PubChem, was the most computationally demanding. MS2Query, SIRIUS and MSNovelist exhibited broadly comparable runtimes, with MS2Query's longer plasma processing times offset by its substantially greater annotation coverage.

Several limitations of this study should be acknowledged. Although the reference compounds were measured in-house, overlap with compounds represented in training datasets cannot be completely excluded. Performance evaluation on the plasma sample relied on only 194 annotated spectra, which themselves may be influenced by database bias, limiting generalizability. Furthermore, benchmarking was restricted to a single matrix and instrument platform. The benchmark was also strongly skewed toward protonated and deprotonated precursors, which accounted for 99% and 96% of positive- and negative-mode reference spectra, respectively, and 92% and 90% of the corresponding NIST plasma precursors. Results may therefore not generalize to less common precursor types. Finally, all evaluated tools continue to evolve through database and model updates, with the exception of ChemDistiller, whose static 2018 release should be considered when interpreting its results.

Future work should explore consensus annotation frameworks that combine complementary strengths across tools while minimizing coverage loss. Continued expansion of publicly available annotated MS/MS datasets will be particularly valuable for DL approaches such as MS2Query and MSNovelist, while advances in spectral representation learning (Giné et al., 2026) may further improve annotation performance. Integrating complementary sources of evidence, including molecular networking, substructure annotation and RT prediction, also represents a promising direction for increasing confidence in automated metabolite annotation.

5 Conclusion

Independent evaluation of annotation tools remains critical, as AI and DL approaches have expanded rapidly since the last CASMI contest (2022, at the time of writing). Using clean reference compounds and expert-annotated plasma spectra, we evaluated five tools spanning complementary paradigms and the coverage–accuracy trade-offs arising from tool design and database choice. We also assessed Tanimoto similarity, which is particularly relevant for analogue search tools such as MS2Query, and the reliability of each tool's confidence score. Overall, no single tool dominated across all evaluated metrics: SIRIUS and MS2Query achieved the highest structural accuracies through contrasting coverage strategies, with MSNovelist closely matching SIRIUS

performance despite its *de novo* approach. We advocate for the continued development of user-friendly tools with clear and up-to-date usage guidelines, and for the regular updating of underlying models and databases to reflect the rapidly evolving chemical and spectral landscape. Future studies evaluating annotation performance at larger scale, across specific chemical classes or mass ranges, and on diverse instrument platforms will be essential to further advance the reliability of untargeted metabolomics annotation.

Supplementary information. Supplementary Figures S1–S4 are provided in Online Resource 1. The complete LC-MS/MS acquisition and annotation procedures are provided in Online Resource 2. The data and Python code used to generate all figures are provided to as Online Resource 3 (see code availability statement).

Acknowledgements. The authors gratefully acknowledge Magdalena Musialak-Lange for her valuable comments and constructive feedback on an earlier version of this manuscript.

Declarations

- **Funding** This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101167843 – BICEPS.
- **Conflict of interest** The authors are employees of metaSysX GmbH, a company providing metabolomics and lipidomics analytical services. The authors declare no other competing interests.
- **Ethics approval and consent to participate** This study used commercially available NIST Standard Reference Material 1950 (Metabolites in Frozen Human Plasma); no new human participants were recruited and no identifiable human data were used. Therefore, institutional ethical approval was not required.
- **Data availability** The LC-MS/MS spectral data (MGF format) for the reference compound standards and NIST SRM 1950 plasma, together with the annotation results of all benchmarked algorithms and the in-house reference annotations for both datasets, are available in Zenodo at https://doi.org/10.5281/zenodo.23562374.
- **Code availability** The data and Python code used to generate all figures are provided to reviewers as Online Resource 3 and will be made publicly available upon publication at https://github.com/marta-pereiradacosta/Annotate-This-Benchmark-Analysis.
- **Author contribution** MC conceived the study, performed the analyses, interpreted the results and drafted the manuscript. SK and ACI supervised the study and contributed to study design, interpretation and manuscript revision. All authors read and approved the final manuscript.

References

Bajusz, D., Rácz, A., Héberger, K. (2015). Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, *7*(1), 20,

Broeckling, C.D. (2026). Creating taxonomically-informed metabolome libraries for any species using the pubchem. bio r package. *Analyst*, *151*(3), 851–859,

CASMI (2022). *Critical assessment of small molecule identification (CASMI) 2022.* Retrieved from http://www.casmi-contest.org/2022/index.shtml (Accessed 21 October 2026)

CASMI (2026). *Critical assessment of small molecule identification (CASMI) 2026.* Retrieved from http://casmi-contest.org/2026/index.shtml (Accessed 21 October 2026)

Chi, J., Shu, J., Li, M., Mudappathi, R., Jin, Y., Lewis, F., ... Gu, H. (2024). Artificial intelligence in metabolomics: a current review. *TrAC Trends in Analytical Chemistry*, *178*, 117852,

Coler, E.A., Chen, W., Melnik, A.V., Morton, J.T., Aksenov, A.A. (2024). Metabolomics in the era of artificial intelligence. *Microbiota and Host*, *2*(1), ,

de Jonge, N.F., Louwen, J.J., Chekmeneva, E., Camuzeaux, S., Vermeir, F.J., Jansen, R.S., ... van der Hooft, J.J. (2023). Ms2query: reliable and scalable ms2 mass spectra-based analogue search. *Nature Communications*, *14*(1), 1752,

de Jonge, N.F., Mildau, K., Meijer, D., Louwen, J.J., Bueschl, C., Huber, F., van der Hooft, J.J. (2022). Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. *Metabolomics*, *18*(12), 103,

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., ... others (2016). Classyfire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics*, *8*(1), 61,

Dührkop, K., Fleischauer, M., Ludwig, M., Aksenov, A.A., Melnik, A.V., Meusel, M., ... Böcker, S. (2019). Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods*, *16*(4), 299–302,

Dührkop, K., Shen, H., Meusel, M., Rousu, J., Böcker, S. (2015). Searching molecular structure databases with tandem mass spectra using csi: Fingerid. *Proceedings of the National Academy of Sciences*, *112*(41), 12580–12585,

Giné, R., Pérez-López, I., Badia, J.M., Capellades, J., Yanes, O. (2026). Benchmarking ms/ms featurization strategies for machine learning-driven metabolite structure annotation. *Journal of the American Society for Mass Spectrometry*, *37*(7), 1550,

Gowda, G.N., Zhang, S., Gu, H., Asiago, V., Shanaiah, N., Raftery, D. (2008). Metabolomics-based methods for early disease diagnostics. *Expert review of molecular diagnostics*, *8*(5), 617–633,

Hoffmann, M.A., Kretschmer, F., Ludwig, M., Böcker, S. (2023). Mad hatter correctly annotates 98% of small molecule tandem mass spectra searching in pubchem. *Metabolites*, *13*(3), 314,

Huber, F., van der Burg, S., van der Hooft, J.J., Ridder, L. (2021). Ms2deepscore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics*, *13*(1), 84,

Hupatz, H., Rahu, I., Wang, W.-C., Peets, P., Palm, E.H., Kruve, A. (2025). Critical review on in silico methods for structural annotation of chemicals detected with lc/hrms non-targeted screening. *Analytical and Bioanalytical Chemistry*, *417*(3), 473–493,

Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y., Ishiguro-Watanabe, M. (2025). Kegg: biological systems database as a model of the real world. *Nucleic acids research*, *53*(D1), D672–D677,

Laponogov, I., Sadawi, N., Galea, D., Mirnezami, R., Veselkov, K.A. (2018). Chemdistiller: an engine for metabolite annotation in mass spectrometry. *Bioinformatics*, *34*(12), 2096–2102,

Leonov, D.A., Mednova, I.A., Chernonosov, A.A. (2026). Evaluation of massfrontier, metfrag, ms-finder, and sirius for metabolite annotation using an experimental lc–hrms dataset. *Biomedicines*, *14*(4), 872,

Ludwig, M., Nothias, L.-F., Dührkop, K., Koester, I., Fleischauer, M., Hoffmann, M.A., ... others (2020). Database-independent molecular formula annotation

using gibbs sampling through zodiac. *Nature Machine Intelligence*, *2*(10), 629–641,

Nguyen, D.H., Nguyen, C.H., Mamitsuka, H. (2019). Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches. *Briefings in bioinformatics*, *20*(6), 2028–2043,

Novoa-del Toro, E.M., & Witting, M. (2024). Navigating common pitfalls in metabolite identification and metabolomics bioinformatics. *Metabolomics*, *20*(5), 103,

Phinney, K.W., Ballihaut, G., Bedner, M., Benford, B.S., Camara, J.E., Christopher, S.J., ... others (2013). Development of a standard reference material for metabolomics research. *Analytical chemistry*, *85*(24), 11732–11738,

Pomyen, Y., Wanichthanarak, K., Poungsombat, P., Fahrmann, J., Grapov, D., Khoomrung, S. (2020). Deep metabolome: Applications of deep learning in metabolomics. *Computational and Structural Biotechnology Journal*, *18*, 2818–2825,

RDKit contributors (2025). *Rdkit: Open-source cheminformatics.* Retrieved from https://www.rdkit.org

Russo, F.F., Nowatzky, Y., Jaeger, C., Parr, M.K., Benner, P., Muth, T., Lisec, J. (2024). Machine learning methods for compound annotation in non-targeted mass spectrometry—a brief overview of fingerprinting, in silico fragmentation and de novo methods. *Rapid communications in mass spectrometry*, *38*(20), e9876,

Ruttkies, C., Schymanski, E.L., Wolf, S., Hollender, J., Neumann, S. (2016). Metfrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of cheminformatics*, *8*(1), 3,

Shen, H., Dührkop, K., Böcker, S., Rousu, J. (2014). Metabolite identification through multiple kernel learning on fragmentation trees. *Bioinformatics*, *30*(12), i157–i164,

Stravs, M.A., Dührkop, K., Böcker, S., Zamboni, N. (2022). Msnovelist: de novo structure generation from mass spectra. *Nature Methods*, *19*(7), 865–870,

Summer, L., Amberg, A., Barrett, D., Beale, M., Beger, R., Daykin, C., ... others (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, *3*(3), 211–221,

Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., Wishart, D.S. (2021). Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analytical chemistry*, *93*(34), 11692–11700,

Wolf, S., Schmidt, S., Müller-Hannemann, M., Neumann, S. (2010). In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, *11*(1), 148,

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource2.pdf
- OnlineResource1.pdf
- OnlineResource3.zip