MSAgent: An Evidence Grounded Agentic Framework for Autonomous Scientific Exploration in Mass Spectrometry-based Metabolomics

Anonymous Author(s)

Abstract

Mass spectrometry (MS) is a cornerstone high-throughput technology for molecular discovery, yet the reliable elucidation of chemical structures remains a formidable, expert-dependent bottleneck. Here we present MSAgent, an autonomous agentic framework that bridges computational automation and expert intuition to empower an end-to-end trajectory from raw spectra to autonomous scientific discovery. By orchestrating a MSToolbox of over 50 domain-specific tools via Large Language Models (LLMs) and integrating *Metabolisminsight*—a large-scale knowledge graph mapping over 3.8 million metabolite-context associations mined from public studies and literature—MSAgent dynamically unifies the analytical pipeline into an evidence-grounded workflow with intent-aware planning, cross-resource synthesis, and traceable analytical reports. We evaluated MSAgent on `MSAgent-Bench` and a real-world clinical cohort. MSAgent achieves over 10% improvement in identification accuracy across distinct tool-dataset configurations while maintaining a 98.35% non-degradation rate. It also demonstrates strong epistemic calibration, with confidence better aligned to structural fidelity (Pearson $r = 0.438$ vs. -0.219 for GPT-4), reducing hallucinations. Moving beyond automated annotation, MSAgent re-analyzed a LUAD cohort and uncovered a tumor-microenvironment program centered on gamma-aminobutyric acid (GABA), cross-validated using the study's transcriptomic supplement. It further formulated three testable hypotheses spanning neuroactive tumor chemistry, lipid-redox remodeling, and nucleotide-driven immune-genome regulation. By transforming raw spectral features into concrete experimental designs, our work shifts metabolomics from descriptive profiling to hypothesis-driven scientific exploration.

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Metabolomics, Automated Structure Elucidation, Autonomous Scientific Discovery, LLM Agent, Tumor Microenvironment

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, Washington, DC, USA
© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM Reference Format:
Anonymous Author(s). 2026. MSAgent: An Evidence Grounded Agentic Framework for Autonomous Scientific Exploration in Mass Spectrometry-based Metabolomics. In . ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

High-throughput and high-resolution mass spectrometry has become indispensable for modern chemical biology. By rapidly generating large volumes of accurate molecular information from environmental, dietary, or pathological samples[1–12], MS enables researchers to identify biologically relevant small molecules and to infer structural features of previously unknown compounds[13–16]. Over the past decades, the widespread adoption of MS technologies together with the strong global open-source community has produced massive and heterogeneous spectral repositories such as GNPS[17], MassIVE[18], and MONA[19]. This rich landscape of publicly accessible resources has in turn catalyzed the development of numerous computational tools that extract structure-related patterns from fragmentation peaks and associated metadata to elucidate the composition, biochemical roles, and mechanisms of small molecules.

Mass spectrometry (MS) analysis integrates established chemical priors—such as neutral losses and McLafferty rearrangements[20] ,to infer molecular structures or explain the chemical mechanisms underlying peak generation through computation modeling. These methods evolved from combinatorial optimization and shortest-path algorithms to more recent machine learning paradigms, including autoencoders[21], diffusion models[22], and transformer-based architectures[23].

These domain-specific tools perform exceptionally well within targeted scopes and have become indispensable components of modern MS workflows. However, their strengths also expose inherent limitations when applied to real-world molecular identification tasks. Reliable structure elucidation from a single mass spectrum typically requires coordinated evidence from multiple tools, including molecular formula prediction to determine elemental composition, database search to generate candidate structures, and spectrum simulation for cross-validation. In practice, this process remains highly human-driven. Researchers must carefully operate each tool, interpret heterogeneous outputs and reason across multiple lines of evidence to reach a final conclusion. This is because different tools adopt self-defined data formats and evaluation protocols, leading to divergent data requirements and scoring schemes. The resulting analytical evidence often consists of heterogeneous information that must be manually synthesized, imposing substantial cognitive and operational burdens on researchers. Moreover, substantial technical barriers also exist. Each system typically requires specialized installation environments including various package requirements

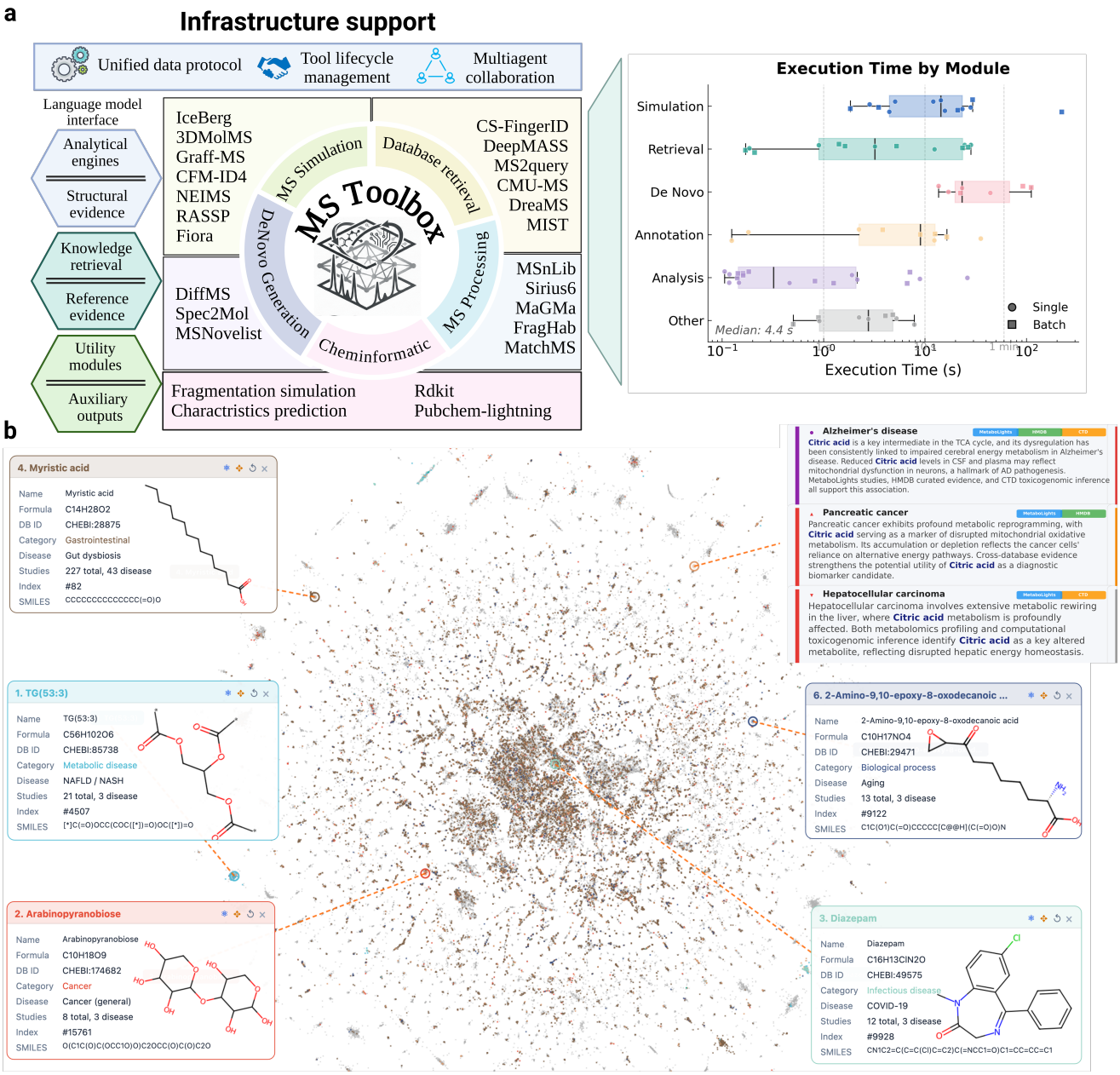

Figure 1: The foundational architectural infrastructure of MSAgent. (a) MSToolbox declarative tool orchestration, illustrating the unification of over 50 heterogeneous mass spectrometry utilities, chemical informatics modules, and service-style database search engines under a cohesive data protocol. (b) Metabolisminsight cross-scale knowledge graph, delineating the multi-source integration of public metabolomics studies, molecular identity catalogs, and real-time literature evidence tracking.

and complex dependencies under strict version requirements. Additionally, tools are also optimized for specific scenarios, such as particular instrument types, ionization modes, chemical classes or data modalities, making them individually insufficient for comprehensive real-world MS analysis. Consequently, effective use of the rich ecosystem of MS analysis tools often requires substantial expertise and poses a steep learning curve, limiting accessibility for non-computational researchers and hindering reproducible, scalable deployment of advanced MS workflows. This situation also slows the translation of newly developed MS methods into practical analytical workflows.

Large Language ModelsLLMs offer a compelling path toward addressing these challenges through enabling autonomous analysis. By leveraging vast pre-trained knowledge and compositional linguistic structures, LLMs exhibit unparalleled capabilities in semantic reasoning, knowledge synthesis, and task orchestration[24–28]. When provided with sufficient support evidence, they are capable of evaluating complex scenarios, selecting optimal analytical recipes from multiple facts, and delivering high-level decision support that rivals or even exceeds human expertise. However, a critical domain gap persists: the highly structured and dimensional numerical format of mass spectrometry data remains fundamentally misaligned with the text-centric tokenization of standard LLMs. This lack of native spectral perception prevents LLMs from directly interpreting raw MS data, often constraining them to shallow, surface-level pattern matching, leading to severe hallucination and unreliable predictions in downstream MS analysis.

Recently, the emergence of LLM-powered AI agents has shown great promise in scientific research, driven by the demand for autonomous workflow construction, multimodal reasoning, and iterative information aggregation. These AI assistants demonstrate remarkable effectiveness in unlocking the true potential of different domains via the combination of domain-specific tools and knowledge-enriched repositories guided by human instructions. However, despite these advances, agentic frameworks remain underexplored in the mass spectrometry domain, where complex data structures and tool ecosystems pose unique challenges. Inspired by these pioneer studies, we developed MSAgent, an LLM-driven MS analysis framework that leverages MS domain knowledge by autonomously orchestrating various MS tools into its workflow. It is designed as an intelligent MS interpretation assistant capable of understanding user intent, constructing and executing analysis pipeline plans, dynamically invoking domain tools, gathering and refining evidence from multiple resources, and ultimately delivering human-readable reports.

2 Architectural Infrastructure of MSAgent for Metabolism Research

2.1 Declarative Tool Orchestration and Heterogeneous Mass Spectral Unification

Mass spectrometry analysis operates over high-dimensional, structured spectral data with heterogeneous metadata and loosely defined tool input–output conventions, which cannot be reliably or transparently mediated through language alone. Without an explicit abstraction layer to ground tool interactions in structured data dependencies, critical analytical information cannot be faithfully propagated across tools or preserved during agent reasoning.

To address this challenge, we developed MSToolbox(Figure 1a), the first tool integration framework and foundational infrastructure specifically designed for LLM agents in the field of metabolomics and mass spectrometry analysis. MSToolbox encompasses over 50 professional domain tools, covering the latest deep learning models as well as classical, well-established tools that complement each other across diverse stages of MS analysis. In addition to integrating these existing tools, we also developed specific toolkits tailored for mass spectrometry analysis, including chemical informatics modules, statistical and visualization modules, and a high-speed database search engine. By explicitly declaring data dependencies, outputs, and execution semantics, MSToolbox unifies these heterogeneous tools—spanning spectral preprocessing, database search, *de novo* generation, and *in silico* spectrum generation—under a common data protocol and coherent execution flow. Substantial engineering optimizations were applied to transform standalone scripts into persistent, service-style tools in MSToolbox (e.g., GPU-resident models), bringing the execution latency of most tools to the sub-minute level (Fig. 1a). This improvement significantly reduces invocation overhead, enabling robust and efficient iterative tool usage within the workflow.

2.2 Cross-Scale Biological Evidence Grounding via Metabolisminsight

To connect molecular annotations with interpretable biological context, MSAgent construsted a knowledge-graph Metabolisminsight(Figure 1b). It is a knowledge base derived primarily built from public MetaboLights studies by design and utilizing data mining scripts and refined by using gpt-4 to extract inforation and output structured data. We further expand it by HMDB, ChEBI, CTD, and KEGG, providing 599,635 compound identity records and 2,924,596 name aliases for robust molecule naming and literature search. The current database contains 2,685 detailed metabolic studies reference, 599,635 unique metabolite catalog records, and 3,864,945 metabolite molecule–disease/context association rows. In addition, Metabolisminsight internally equipped with connection to pubmed, this provided millions of literature support and supplement the information of current database . This knowledge graph thereby linking molecules to diseases, biological processes, or phenotypic contexts through public metabolomics studies.

3 Autonomous scientific discovery in untargeted metabolomics via knowledge-grounded AI agents

As illustrated in Figure 2a, MSAgent transforms raw liquid chromatography tandem mass spectrometry cohorts into traceable scientific proposals through a two-stage framework consisting of LLM-driven molecular identification and hierarchical scientific reporting. The end-to-end scientific discovery workflow is conceptualized in Figure 2b. In Step 1, faced with complex biological samples containing thousands of heterogeneous features such as lipids, metabolites, or exogenous chemical variants (Figure 2c), MSAgent extracts case-control abundance contrasts and utilizes a knowledge-graph-guided dynamic attention mechanism to prioritize features most likely driving phenotypic variance. In Step 2, MSAgent instantiates two distinct operating modes tailored for either study-specific validation or open-ended discovery, which subsequently guides the autonomous, parallel molecular identification loop in Step 3 via ReAct-guided tool orchestration. Finally, in Step 4, these validated molecular profiles are mapped back to macro biological contexts to synthesize multi-molecule hypotheses, ultimately producing a structured, three-level scientific report that strictly bounds evidentiary certainty and discovery-grade insights (Figure 2d).

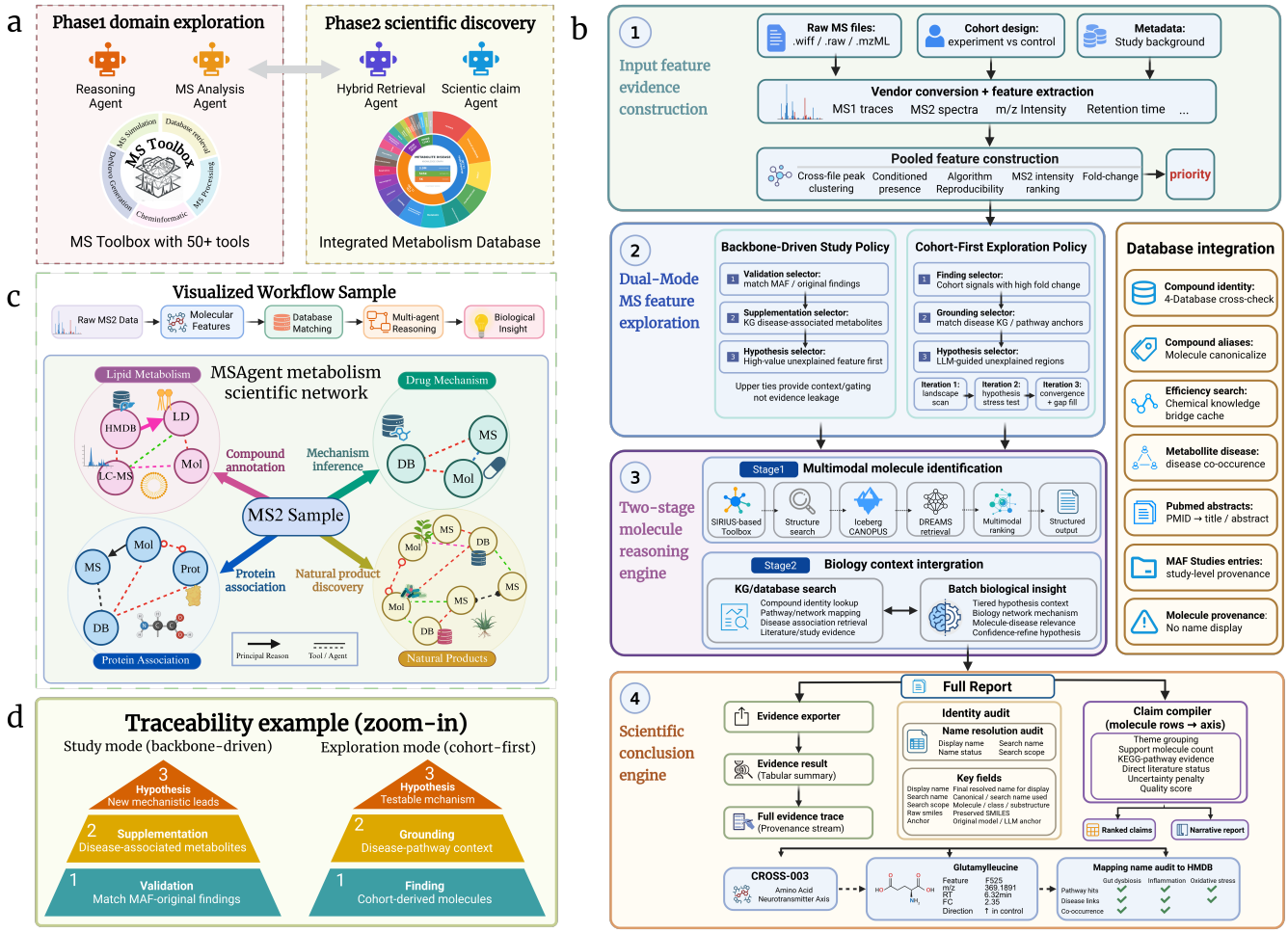

Figure 2: Overview of the MSAgent framework for autonomous mass spectrometry-based scientific discovery. (a) Two-stage architecture, highlighting the computational flow from LLM-driven autonomous molecule identification to evidence-bound hierarchical scientific reporting. (b) End-to-end operational workflow, detailing the sequential execution trace from raw cohort input (Step 1) through contextual mode selection (Step 2), parallelized identification loops (Step 3), to macro-level biological hypothesis synthesis (Step 4). (c) Knowledge-graph-guided scientific attention allocation, demonstrating the filtering of thousands of multi-class features (including lipids, metabolites, and exogenous variants) to isolate phenotypically significant candidates based on contrast statistics and prior knowledge. (d) Hierarchical scientific report generation, illustrating the progressive layering of text synthesis across distinct epistemic boundaries, moving from exact empirical validation to bounded discovery-grade hypotheses.

4 Robust molecular identification and uncertainty calibration via multi-tool synergy

The operational architecture of MSAgent unfolds across two sequential phases—Molecular Identification and Scientific Discovery—where high-resolution compound discrimination establishes the empirical bedrock for downstream mechanistic elucidation. To rigorously assess MSAgent's capacity for autonomous compound annotation from tandem mass spectrometry (MS/MS) data, we introduce `MSAgent-Bench`, a comprehensive evaluation suite comprising 500 high-quality LC-MS/MS spectra curated from the MassBank RIKEN spectral library.

As illustrated in Figure 3a, MSAgent delivers a substantial leap in identification accuracy compared to standalone baselines. Across six tool-dataset configurations spanning Fiora, Graff-MS, and Iceberg on the CASMI16 and CASMI22 benchmarks, MSAgent improves performance by over 10% in five of the six scenarios. It also yields superior or identical identification outcomes in 418 out of 425 trials, corresponding to a 98.35% non-degradation rate .

Figure 3b shows the effect of selectively integrating chemical and mass spectrometry (MS) knowledge modules. Chemical knowledge alone yields a 20% absolute improvement in Top-1 accuracy over vanilla GPT-4, while isolated MS knowledge causes a 10%

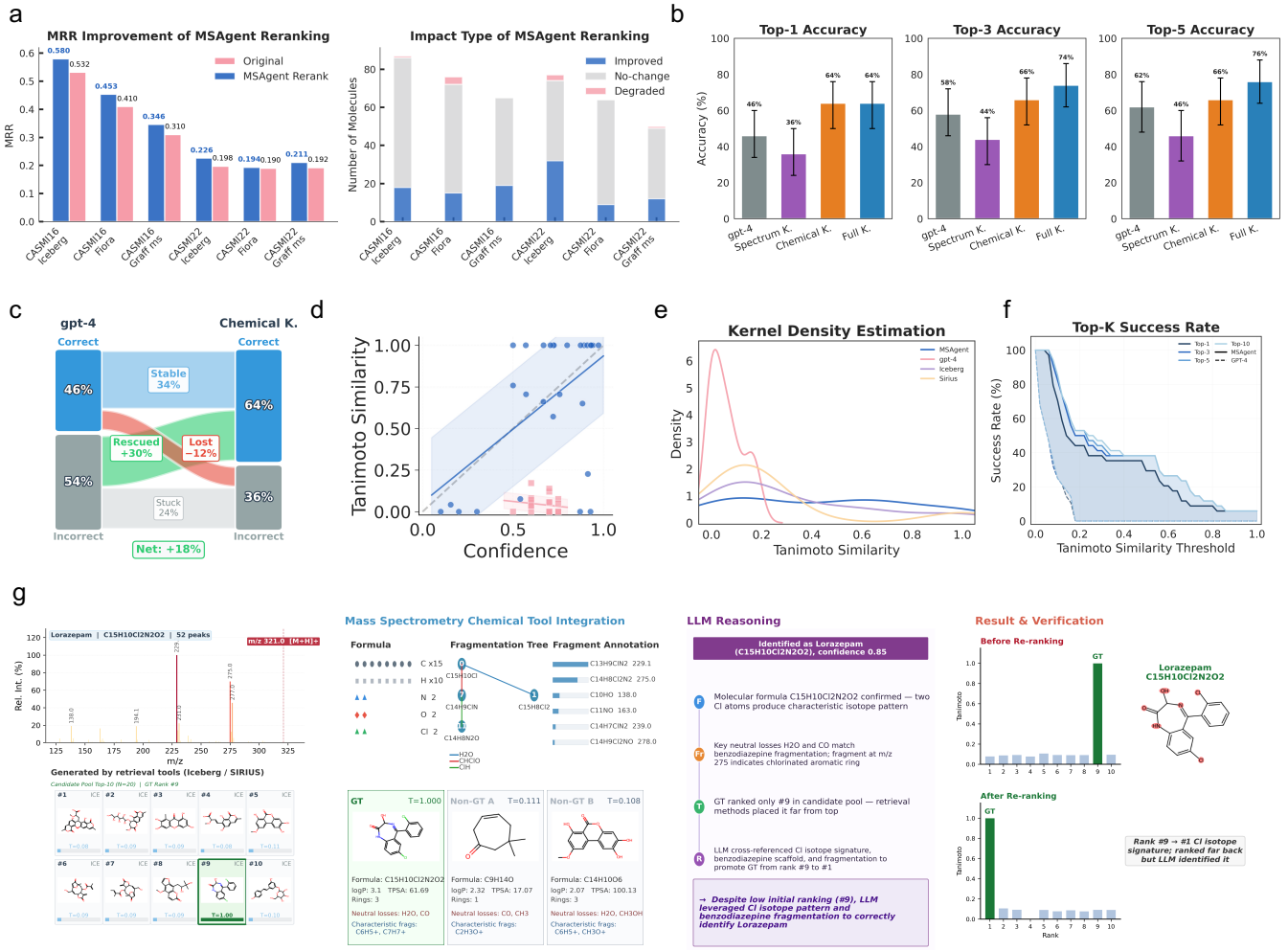

Figure 3: Quantitative evaluation and autonomous reasoning traces of MSAgent on MSAgent-Bench. (a) Baseline identification accuracy and reliability, benchmarking MSAgent against standalone algorithms across six tool-dataset configurations and demonstrating its robust non-degradation rate. (b) Knowledge dimension performance dynamics, comparing the identification accuracy of MSAgent when selectively integrating multi-dimensional chemical and mass spectrometry knowledge modules. (c) Fine-grained sample transition analysis, tracking the operational prediction shifts and net positive advancements of the core LLM upon the infusion of chemical heuristics. (d) Uncertainty calibration and hallucination assessment, evaluating the correlation between model-estimated confidence scores and empirical Tanimoto similarity to the ground truth. (e) Kernel Density Estimation (KDE) profiles, illustrating the distribution shift of Top-1 predicted molecular similarity toward high-fidelity regimes. (f) Empirical candidate retention curves, showing the stable candidate prioritization and decay trajectories across a continuum of progressive Tanimoto similarity thresholds. (g) Step-by-step autonomous execution and reasoning trace, delineating the multi-tool synergy and expert-level cross-referencing heuristics during a challenging molecular identification task for Lorazepam, successfully correcting a flawed initial retrieval from Rank #9 to Rank #1.

degradation. MSAgent resolves this conflict by combining both domains, preserving Top-1 parity with the chemical-only setting while improving Top-3 accuracy to 74% and Top-5 accuracy to 76%. These trajectories show that MSAgent does not merely accumulate data, but actively contextualizes ambiguous spectral features under chemical constraints.

To further dissect this interplay, Figure 3c tracks the fine-grained sample transitions of GPT-4 after chemical tool infusion. Tool integration alters predictions in 30% of evaluation instances. Although 12% of previously correct samples are mis-ranked, the net result is an 18% positive advancement. This suggests that domain-specific knowledge improves decision boundaries even when it is not uniformly beneficial at the single-sample level.

To systematically investigate hallucination and model self-awareness, we evaluate the alignment between model-estimated confidence and empirical prediction correctness (Figure 3d). MSAgent demonstrates a robust correlation between confidence and Tanimoto similarity to groundtruth (Pearson $r = 0.438$), whereas the LLM baseline (`GPT-4`) exhibits a negative correlation ($r = -0.219$). In extreme uncertainty regimes (Tanimoto < 0.1), MSAgent downscales its confidence (< 0.3), while `GPT-4` remains overconfident (> 0.6) for highly erroneous structures. This indicates that structured evidence improves uncertainty calibration and reduces hallucinations.

Finally, Figure 3e–f show that MSAgent shifts Top-1 predictions toward higher structural similarity and maintains a retention probability above 10% across settings even in the ultra-hard crystalline tier (90%–100% similarity). Compared with `GPT-4`, MSAgent provides more faithful structural ranking and supports downstream biomedical mechanism discovery.

4.1 Overcoming Retrieval Limitations via Expert-Level Cross-Referencing

To provide a granular, mechanistic understanding of MSAgent's autonomous tool orchestration and reasoning under ambiguity, we trace the automated identification workflow for Lorazepam ($C_{15}H_{10}Cl_2N_2O_2$) (Figure 3g). Upon receiving the raw MS/MS spectrum, MSAgent autonomously dispatches peak annotation and database search utilities to generate an initial candidate pool of 10 structurally analogous molecules. Crucially, due to standard spectral alignment limitations, the true ground-truth molecule is severely mis-prioritized, languishing at an unfavorable initial tier of **Rank #9**. Recognizing this structural ambiguity, MSAgent counters the retrieval limitation by leveraging intrinsic chemical information tools within `MSToolbox` to simulate expected fragmentations. By explicitly modeling theoretical neutral losses and diagnostic fragments for the candidates, the agent constructs a rich, multi-dimensional evidence matrix that serves as the foundation for downstream deliberative reasoning.

4.2 From data recovery to autonomous hypothesis generation in LUAD progression

To evaluate whether MSAgent could recover and extend a published biological signal, we re-analyzed the positive-ion MTBLS13565 cohort in study mode, using solid lung adenocarcinoma (LUAD) and pure ground-glass opacity (GGO) as the two comparison groups Figure 4a. The original study reported subtype-associated metabolic separation, with glycerophospholipid and linoleic-acid metabolism, phospholipase A2 (PLA2) activity, and phosphatidylcholine 32:0 [PC(32:0)]-linked intervention as central biological findings. MSAgent therefore provided a stringent case study: a successful rerun should not only rediscover this lipid-remodeling axis, but also preserve the molecular evidence needed to explain how each downstream claim was derived. MSAgent converted 3555 positive-ion mass spectrometry (MS) features into 174 molecule-level evidence packets after identity gating, cohort-signal annotation, knowledge-graph lookup, and PubMed passage retrieval (Figure 4b).

Consistent with the original report, the validation layer recovered a lipid/carnitine remodeling signature aligned with glycerophospholipid and linoleic-acid metabolism (Figure 4d). Representative evidence rows included depleted linoleyl carnitine and O-palmitoylcarnitine, together with L-palmitoylcarnitine and ethyl oleate as direct or mechanism-supported lipid-context molecules.

Having established recovery of the published lipid-remodeling axis, we next asked whether the same evidence packets could be organized into mechanistic neighborhoods rather than isolated molecule annotations. MSAgent grouped the validated evidence into a mechanism-bundle map connecting lipid remodeling to mitochondrial energy handling, redox stress, inflammatory signaling, nucleotide-immune biology, and a gamma-aminobutyric acid (GABA)-focused follow-up branch (Figure 4e).

Figure 4f resolves this topology at the molecule level. Nucleotide-immune signaling was represented by differential adenine, thymine, and cytosine signals, linking the lung adenocarcinoma (LUAD)-ground-glass opacity (GGO) comparison to DNA-repair, epigenetic, and immune-sensing contexts. The GABA branch was displayed separately as a follow-up module: GABA was retained as the core candidate, whereas L-theanine and L-asparagine were used as amino-acid context molecules that connect neuroactive chemistry, oxidative stress, and immune follow-up biology without being treated as direct GABA evidence.

The most informative follow-up example was gamma-aminobutyric acid (GABA), because it illustrates how MSAgent can recover a hypothesis that was not foregrounded in the original metabolomics narrative. The original article did not explicitly discuss GABA in the main text, and the metabolite supplement contained Gabazine rather than GABA itself. In the positive-ion (POS) rerun, however, GABA-related molecular evidence was retained and direct PubMed passages connected GABA to lung adenocarcinoma biology. One study supported an anti-proliferative gamma-aminobutyric acid type B receptor (GABABR)-cyclic adenosine monophosphate response element-binding protein (CREB)/extracellular signal-regulated kinase (ERK) signaling context, whereas another reported tumor-derived GABA effects on tumor-associated macrophage (TAM) polarization and neovascularization; a third passage provided broader anti-inflammatory and antioxidant mechanism context (Figure 4g). These literature directions are not collapsed into a single abundance interpretation. They define GABA as a tumor-microenvironment follow-up candidate with explicit evidence boundaries instead.

The key orthogonal support came from re-examining the original study's own transcriptome supplement. Although GABA was not highlighted as a metabolite in the published main narrative, GAD1, GABRB2, and GABRQ were all upregulated in solid lung adenocarcinoma (LUAD) relative to ground-glass opacity (GGO) in the supplementary expression table. This independent transcriptomic signal connects the MSAgent-derived GABA hypothesis back to the original dataset rather than to an external narrative alone. We therefore used the final hypothesis panel to state three testable directions: a GABA/GAD1 axis linking positive-ion (POS) metabolomics to solid-LUAD transcriptome signals, a lipid-carnitine/fatty-acid oxidation (FAO)-redox extension of the original lipid axis, and a nucleotide-base layer nominating DNA-repair and epigenetic immune-sensing follow-up experiments (Figure 4h).

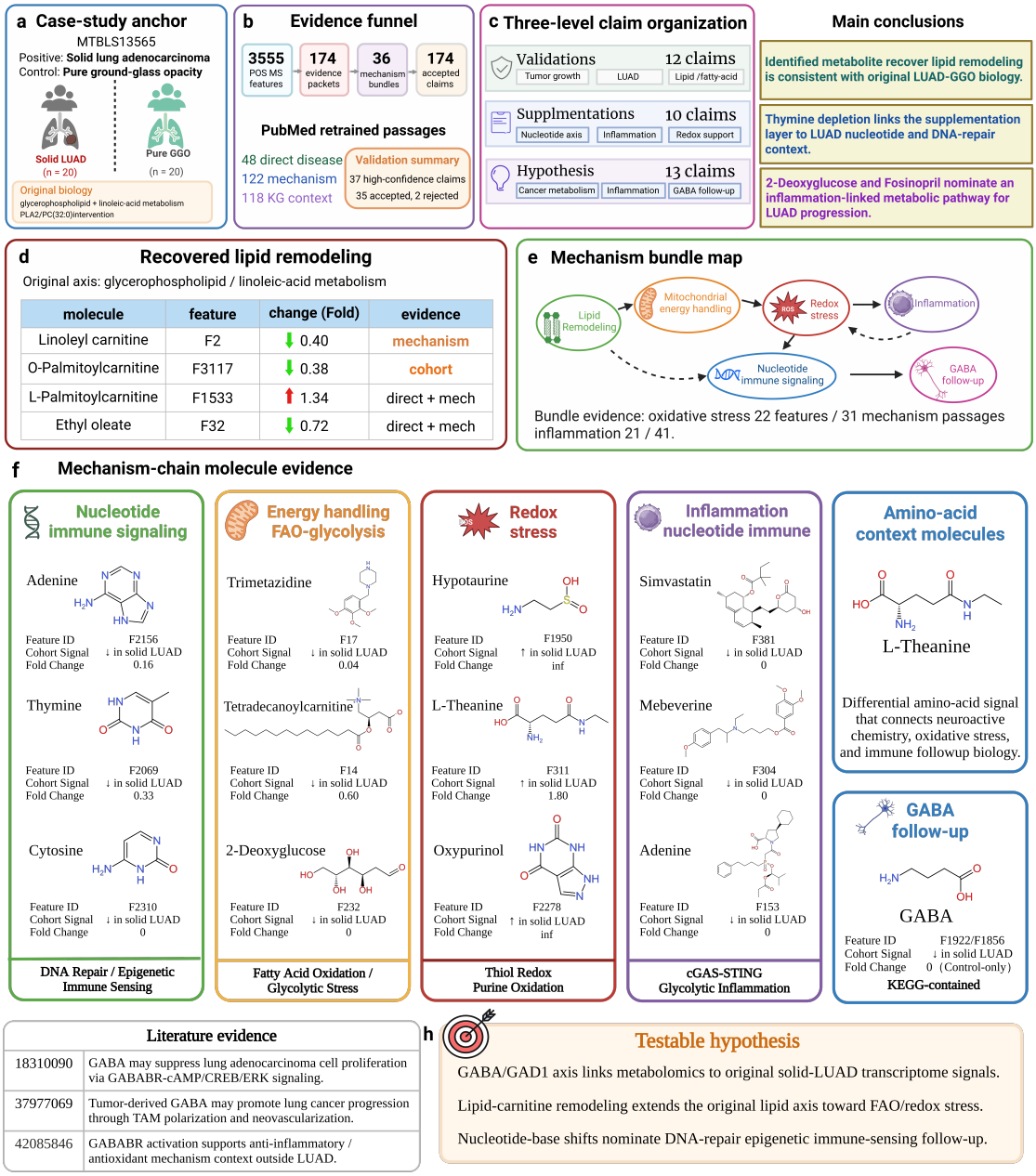

Figure 4: MSAgent-driven autonomous biological signal validation and scientific discovery on the LUAD cohort. (a) Cohort comparison design, illustrating the re-analysis setup of the MTBLS13565 dataset comparing solid LUAD and pure GGO. (b) Evidence extraction pipeline, showing the high-throughput distillation of raw MS features into structured evidence packets, bundles, and validated claims. (c) Hierarchical reporting tiers, detailing the conservative report organization into validation, supplementation, and hypothesis layers. (d) Published signal recovery, displaying the evidence-grounded reconstruction of the original lipid and carnitine remodeling axis. (e) Mechanistic neighborhood topology, mapping validated molecule clusters to interconnected functional pathways like redox and inflammation. (f) Molecule-specific resolution, highlighting individual key metabolites driving energy-handling, nucleotide-immune, and redox contexts. (g) Hidden hypothesis extraction, demonstrating the literature-mined recovery and multi-passage boundary tracking for the GABA pathway. (h) Testable experimental directions, presenting three concrete, multi-omic verification strategies to model aggressive LUAD progression.

These hypotheses highlight the out-of-boundary capability of MSAgent because they convert evidence recovery into concrete experimental directions for LUAD progression. The first hypothesis, the GABA/GAD1 axis, suggests that solid LUAD may carry a GABAergic tumor-microenvironment program that was not visible

in the original metabolomics headline. Its practical value is clear: it nominates GAD1 and GABA receptor subunits as markers for distinguishing solid LUAD from GGO-like lesions, and it points to macrophage polarization and angiogenesis as TME readouts that could explain more aggressive solid-tumor behavior. The second hypothesis, lipid-carnitine to FAO/redox remodeling, gives a mechanistic interpretation of the recovered lipid axis. Its practical value is clear: it suggests that the lipid signature may report a shift in mitochondrial fuel handling and oxidative stress during LUAD solidification, making acylcarnitine profiling, fatty-acid oxidation activity, mitochondrial stress, and glutathione/redox measurements directly relevant to tumor metabolic progression. The third hypothesis, nucleotide-base immune-genome remodeling, expands the study toward DNA-repair, epigenetic regulation, and immune-sensing biology. Its practical value is clear: it nominates thymine DNA glycosylase (TDG)/ten-eleven translocation (TET) activity, cyclic GMP-AMP synthase-stimulator of interferon genes (cGAS-STING) signaling, and tumor immune-state measurements as readouts for testing whether solid LUAD has acquired genome-stress or immune-sensing programs beyond the lipid phenotype. Thus, Panel H does not merely list plausible mechanisms; it prioritizes testable experiments that could convert a metabolomics case study into a broader model of LUAD progression from GGO-like lesions to solid tumors.

5 Conclusion

MSAgent establishes a substantial advance in high-resolution compound identification. On our newly curated benchmark, MSAgent-Bench, the system surpasses standalone computational tools by an accuracy margin exceeding 10% across diverse configuration settings, while maintaining a 98.35% non-degradation rate. This gain arises from tool orchestration rather than unconstrained model generation: by coordinating complementary computational modules under explicit chemical constraints, MSAgent reduces epistemic noise and improves uncertainty calibration. In particular, model confidence becomes positively aligned with structural fidelity (Pearson $r = 0.438$), in contrast to the negative confidence–fidelity relationship observed for generalist baselines such as GPT-4 (Pearson $r = -0.219$).

Beyond spectral annotation, MSAgent further enables evidence-grounded scientific discovery. In a clinical lung adenocarcinoma (LUAD) case study, MSAgent re-analyzed a published solid-LUAD versus ground-glass opacity cohort and recovered the original lipid-remodeling axis centered on glycerophospholipid and linoleic-acid metabolism. More importantly, the system extended beyond the published metabolomics narrative by nominating a GABA/GAD1 tumor-microenvironment axis. This hypothesis was not foregrounded in the original main text and was not supported by a GABA metabolite entry in the supplementary metabolite table; instead, MSAgent linked POS metabolomics evidence, LUAD-associated PubMed literature, and an independent transcriptomic supplement in which GAD1 and GABA receptor subunits were upregulated in solid LUAD. In the same analysis, MSAgent further converted these evidence chains into three disease-relevant, testable hypotheses. First, the GABA/GAD1 axis nominates a underexplored neuroactive tumor-microenvironment program in solid LUAD, suggesting that GAD1 and GABA receptor subunits may serve as markers of the transition from GGO-like lesions to more aggressive solid tumors and motivating follow-up assays of macrophage polarization and angiogenesis. Second, the lipid-carnitine fatty-acid oxidation/redox hypothesis provides a mechanistic interpretation of the recovered lipid-remodeling signature, suggesting that solidification may be accompanied by altered mitochondrial fuel handling, oxidative stress, and glutathione-linked redox adaptation. Third, the nucleotide-based immune-genome hypothesis extends the analysis beyond metabolism alone, nominating DNA-repair, epigenetic regulation, and cGAS-STING-related immune sensing as candidate programs acquired during LUAD progression. Together, these hypotheses convert metabolomics annotation into experimentally actionable models of tumor evolution, linking molecular features to testable mechanisms of solid LUAD emergence from ground-glass-like lesions.

In a broader context, \texttt{MSAgent} pioneeringly opens the door to full-pipeline autonomous scientific discovery within metabolomics, charting a direct trajectory from raw, unannotated spectral features to verifiable biological insights. By introducing the essential agentic infrastructure—seamlessly integrating domain-specific computational toolboxes, large-scale semantic knowledge graphs, and intent-aware cognitive orchestration—this framework bridges the historical chasm between specialized chemical data and high-level mechanistic reasoning. Looking forward, we envision such agentic architectures becoming indispensable cognitive co-pilots across the life sciences. As mass spectrometry foundation models and biomedical knowledge bases continue to co-evolve, this paradigm lays the structural groundwork for future closed-loop autonomous laboratories capable of real-time data interpretation, dynamic experimental design, and fully automated, hypothesis-driven biochemical exploration.

References

[1] Albert T Lebedev. 2013. Environmental mass spectrometry. *Annual review of analytical chemistry* 6, 1 (2013), 163–189.

[2] Timothy R Croley, Richard J Hughes, Brenda G Koenig, Chris D Metcalfe, and Raymond E March. 2000. Mass spectrometry applied to the analysis of estrogens in the environment. *Rapid Communications in Mass Spectrometry* 14, 13 (2000), 1087–1093.

[3] Susan D Richardson. 2001. Mass spectrometry in environmental sciences. *Chemical reviews* 101, 2 (2001), 211–254.

[4] Martin Kussmann, Michael Affolter, Kornél Nagy, Birgit Holst, and Laurent B Fay. 2007. Mass spectrometry in nutrition: understanding dietary health effects at the molecular level. *Mass spectrometry reviews* 26, 6 (2007), 727–750.

[5] Heidi Goenaga Infante, Ruth Hearn, and Tim Catterick. 2005. Current mass spectrometry strategies for selenium speciation in dietary sources of high-selenium. *Analytical and bioanalytical chemistry* 382, 4 (2005), 957–967.

[6] Christof Van Poucke, Christel Detavernier, R Van Cauwenberghe, and Carlos Van Peteghem. 2007. Determination of anabolic steroids in dietary supplements by liquid chromatography–tandem mass spectrometry. *Analytica chimica acta* 586, 1-2 (2007), 35–42.

[7] Jean LJM Scheijen, Egbert Clevers, Lian Engelen, Pieter C Dagnelie, Fred Brouns, Coen DA Stehouwer, and Casper G Schalkwijk. 2016. Analysis of advanced glycation endproducts in selected food items by ultra-performance liquid chromatography tandem mass spectrometry: Presentation of a dietary AGE database. *Food chemistry* 190 (2016), 1145–1150.

[8] Peter Eneroth, Kjell Hellström, and Ragnar Ryhage. 1964. Identification and quantification of neutral fecal steroids by gas–liquid chromatography and mass spectrometry: studies of human excretion during two dietary regimens. *Journal of Lipid Research* 5, 2 (1964), 245–262.

[9] Angela WS Fung, Vijithan Sugumar, Annie He Ren, and Vathany Kulasingam. 2020. Emerging role of clinical mass spectrometry in pathology. *Journal of clinical pathology* 73, 2 (2020), 61–69.

[10] Michaela Aichler and Axel Walch. 2015. MALDI Imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Laboratory investigation* 95, 4 (2015), 422–431.

[11] Jeremy L Norris and Richard M Caprioli. 2013. Imaging mass spectrometry: a new tool for pathology in a molecular age. *PROTEOMICS–Clinical Applications* 7, 11-12 (2013), 733–738.

[12] Nathan Heath Patterson, Balqis Alabdulkarim, Anthoula Lazaris, Aurélien Thomas, Mieczyslaw M Marcinkiewicz, Zu-hua Gao, Peter B Vermeulen, Pierre Chaurand, and Peter Metrakos. 2016. Assessment of pathological response to therapy using lipid mass spectrometry imaging. *Scientific reports* 6, 1 (2016), 36814.

[13] Tobias Kind and Oliver Fiehn. 2010. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical reviews* 2, 1 (2010), 23–60.

[14] Emma L Schymanski, Junho Jeon, Rebekka Gulde, Kathrin Fenner, Matthias Ruff, Heinz P Singer, and Juliane Hollender. 2014. Identifying small molecules via high resolution mass spectrometry: communicating confidence.

[15] Cris Lapthorn, Frank Pullen, and Babur Z Chowdhry. 2013. Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: Separating and assigning structures to ions. *Mass spectrometry reviews* 32, 1 (2013), 43–71.

[16] Hantao Qiang, Fei Wang, Wenyun Lu, Xi Xing, Hahn Kim, Sandrine AM Mérette, Lucas B Ayres, Eponine Oler, Jenna E AbuSalim, Asael Roichman, et al. 2026. Language model-guided anticipation and discovery of mammalian metabolites. *Nature* (2026), 1–10.

[17] Allegra T Aron, Emily C Gentry, Kerry L McPhail, Louis-Félix Nothias, Mélissa Nothias-Esposito, Amina Bouslimani, Daniel Petras, Julia M Gauglitz, Nicole Sikora, Fernando Vargas, et al. 2020. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nature protocols* 15, 6 (2020), 1954–1991.

[18] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets.* Cambridge university press.

[19] Steffen Neumann, René Meier, Michael Wenk, Anjana Elapavalore, Takaaki Nishioka, Tobias Schulze, Michael Stravs, Hiroshi Tsugawa, Fumio Matsuda, and Emma L Schymanski. 2025. MassBank: an open and FAIR mass spectral data resource. *Nucleic Acids Research* 54, D1 (11 2025), D601–D606. arXiv:https://academic.oup.com/nar/article-pdf/54/D1/D601/65272223/gkaf1193.pdf doi:10.1093/nar/gkaf1193

[20] Fred W McLafferty. 1959. Mass spectrometric analysis. Molecular rearrangements. *Analytical chemistry* 31, 1 (1959), 82–87.

[21] Dor Bank, Noam Koenigstein, and Raja Giryes. 2023. Autoencoders. *Machine learning for data science handbook: data mining and knowledge discovery handbook* (2023), 353–374.

[22] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. 2023. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence* 45, 9 (2023), 10850–10869.

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[25] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 2011, 15 pages.

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1800, 14 pages.

[27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL] https://arxiv.org/abs/2302.13971

[28] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] doi:10.48550/arXiv.2501.12948

[29] Jennifer N Wei, David Belanger, Ryan P Adams, and D Sculley. 2019. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science* 5, 4 (2019), 700–708.

[30] Richard Licheng Zhu and Eric Jonas. 2023. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical chemistry* 95, 5 (2023), 2653–2663.

[31] Yuhui Hong, Sujun Li, Christopher J Welch, Shane Tichy, Yuzhen Ye, and Haixu Tang. 2023. 3DMolMS: prediction of tandem mass spectra from 3D molecular conformations. *Bioinformatics* 39, 6 (2023), btad354.

[32] Runzhong Wang, Mrunali Manjrekar, Babak Mahjour, Julian Avila-Pacheco, Joules Provenzano, Erin Reynolds, Magdalena Lederbauer, Eivgeni Mashin, Samuel Goldman, Mingxun Wang, et al. 2025. Neural Spectral Prediction for Structure Elucidation with Tandem Mass Spectrometry. *BioRxiv* (2025).

[33] Michael Murphy, Stefanie Jegelka, Ernest Fraenkel, Tobias Kind, David Healey, and Thomas Butler. 2023. Efficiently predicting high resolution mass spectra with graph neural networks. In *International Conference on Machine Learning*. PMLR, 25549–25562.

[34] Yannek Nowatzky, Francesco Friedrich Russo, Jan Lisec, Alexander Kister, Knut Reinert, Thilo Muth, and Philipp Benner. 2025. FIORA: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nature Communications* 16, 1 (2025), 2298.

[35] Niek F de Jonge, Joris JR Louwen, Elena Chekmeneva, Stephane Camuzeaux, Femke J Vermeir, Robert S Jansen, Florian Huber, and Justin JJ van der Hooft. 2023. MS2Query: reliable and scalable MS2 mass spectra-based analogue search. *Nature Communications* 14, 1 (2023), 1752.

[36] Florian Huber, Sven van der Burg, Justin JJ van der Hooft, and Lars Ridder. 2021. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of cheminformatics* 13, 1 (2021), 84.

[37] Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. 2025. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nature Biotechnology* (2025), 1–11.

[38] Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W Coley. 2025. Diffms: Diffusion generation of molecules conditioned on mass spectra. *arXiv preprint arXiv:2502.09571* (2025).

[39] Eleni Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia Kavraki. 2021. Spec2Mol: An end-to-end deep learning framework for translating MS/MS Spectra to de-novo molecules. (2021).

[40] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. 2022. MSNovelist: de novo structure generation from mass spectra. *Nature Methods* 19, 7 (2022), 865–870.

[41] Samuel Goldman, Jiayi Xin, Joules Provenzano, and Connor W Coley. 2023. Mist-cf: Chemical formula inference from tandem mass spectra. *Journal of chemical information and modeling* 64, 7 (2023), 2421–2431.

[42] Kai Dührkop, Markus Fleischauer, Marcus Ludwig, Alexander A Aksenov, Alexey V Melnik, Marvin Meusel, Pieter C Dorrestein, Juho Rousu, and Sebastian Böcker. 2019. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods* 16, 4 (2019), 299–302.
[43] Kai Dührkop, Louis-Félix Nothias, Markus Fleischauer, Raphael Reher, Marcus Ludwig, Martin A Hoffmann, Daniel Petras, William H Gerwick, Juho Rousu, Pieter C Dorrestein, et al. 2021. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology* 39, 4 (2021), 462–471.
[44] Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. 2015. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. *Proceedings of the National Academy of Sciences* 112, 41 (2015), 12580–12585.
[45] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. 2023. PubChem 2023 update. *Nucleic acids research* 51, D1 (2023), D1373–D1380.
[46] David S Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L Lee, et al. 2022. HMDB 5.0: the human metabolome database for 2022. *Nucleic acids research* 50, D1 (2022), D622–D631.

Appendix

A MSToolbox

The core design principle of MSToolbox is to decouple tool implementation from agent integration. Given a function f that implements specific analytical logic and a configuration specification C declaring its data requirements, MSToolbox registers the function as an agent-accessible tool T, that is,

$$\text{MSToolbox}(f, C) = T \tag{1}$$

where the configuration $C = (D, O, R, P, \phi)$ specifies the input dependencies set D (data keys that must exist in the session), output keys O (data fields the tool will populate), reflection keys R (outputs to present to users), execution priority P, and an optional LLM interpretation prompt ϕ. This formulation enables tool authors to focus exclusively on analytical logic while the system automatically handles dependency resolution, execution scheduling, and result integration.

Upon registration, MSToolbox constructs a directed dependency graph $G = (V, E)$ where nodes $V = V_{tools} \cup V_{data}$ represent both tools and data fields, and edges E capture the producer-consumer relationships between them. Specifically, for each tool T_i with dependencies D_i, the system creates edges $e(D_i, T_i)$ if D_i is included in the dependencies set$D$. This graph serves as the foundation for the Agent's reachability analysis, enabling automatic determination of tool availability based on current session state $\mathcal{S}$, where tools become accessible when all their dependencies are satisfied. Once a tool configuration is saved through MSToolbox, the Agent automatically recognizes the new capability and can invoke it in subsequent analysis sessions without any code changes to the agent system. Furthermore, the dependency graph enables the agent to reason about tool reachability, identifying prerequisite tool chains that can make currently unavailable tools accessible. For instance, if a user requests a tool T_{target} whose dependencies are not yet satisfied, the system traverses the graph to find producer tools that can generate the missing data, constructing an executable plan that chains multiple tools to achieve the desired analysis.

A.1 MSToolbox Architecture

A.1.1 Computational Mass Spectrometry Simulation.

NEIMS. NEIMS (Neural Electron Ionization Mass Spectrometry) takes a molecule's SMILES string as input. NEIMS utilizes a lightweight multi-layer perceptron architecture to directly predict electron ionization mass spectra[29]. Our implementation utilizes its official implementation and its github open weight. We also design a uvicorn-based client-server architecture to perform batch predictions through a dedicated api.

RASSP. RASSP (Rapid approximate subset-based spectra prediction) takes a molecule's SMILES representation as input. RASSP employs graph-based neural networks to capture structural features and predict spectral peak intensities efficiently[30]. The framework allows for specific model selections: FormulaNet, which predicts probability distributions over chemical subformulae, or SubsetNet, which utilizes GNNs to predict vertex subsets of the molecular graph for better generalization. Our implementation leverages a standalone service to handle high-throughput spectral simulations.

3DMolMS. 3DMolMS (3D Molecular Network for Mass Spectra Prediction) takes a molecule's SMILES string, precursor type and collision energy as input in our implementation. 3DMolMS integrates 3D molecular conformations and geometric features to simulate fragmentation patterns across different instrument types[31]. Our implementation employs a service-based client that includes rigorous molecular filtering and 3D coordinate generation for high-fidelity predictions.

Iceberg. Iceberg takes a molecule's SMILES string, precursor type and precursor m/z as input in our implementation. Iceberg utilizes a DAG-based model architecture[32], supporting both CANOPUS and MassSpecGym model versions. It simulates fragment ion generation and intensity prediction, allowing for instrument-specific simulations. Our implementation delegates the workload to a standalone prediction script that leverages pre-trained checkpoints for intensity and generation models.

Graff-MS. Graff-MS takes a molecule's SMILES representation, adduct type and collision energy as input in our implementation. Graff-MS featurizes molecules into graphs by computing graph laplacians and eigenfeatures to capture complex structural information[33]. It supports detailed metadata such as instrument type and isotope presence. Our implementation utilizes the official PyTorch Lightning inference script to generate high-fidelity spectral data.

Fiora. Fiora (Fragment Ion Reconstruction Algorithm) takes a molecule's SMILES string, precursor mode and collision energy as input in our implementation[34]. Fiora employs a GNN model that integrates atom, bond, and covariate feature encoders given molecular weight and instrument type. It constructs molecular structure graphs and fragmentation trees to simulate compound fragmentation patterns, specifically optimized for HCD instruments. Our implementation includes its official GPU-accelerated inference.

A.1.2 Mass Spectrometry–Driven Molecular Retrieval.

Simulation-based Database Retrieval. This function is implemented on Iceberg framework, matching experimental or Iceberg predicted spectra against a reference SMILES list or a comprehensive library of reference spectra. The retrieval process employs the cosine similarity algorithm to calculate the match between query peaks and

library spectra, accounting for fragment ion alignment within a specified mass tolerance. Our implementation optimizes performance through a singleton cache manager that handles the loading of normalized spectral data and metadata, enabling rapid candidate selection and optional LLM-based refinement of the results.

MS2Query. MS2Query is a search engine that utilizes deep learning-based embeddings to identify both exact matches and chemical analogues in large spectral libraries[35]. It first employs MS2Deepscore to calculate the structural similarity between the query spectrum and library entries[36], bypassing the need for strict precursor m/z preselection. The retrieval is further optimized by a random forest re-ranking model that integrates five distinct features to predict a reliability score between 0 and 1. Our implementation execute the official MS2Query inference script, processing input data and filtering results based on ion mode and minimum score thresholds to ensure high-fidelity matches.

DreaMS. DreaMS takes a raw MS/MS spectrum, precursor m/z and precursor ionization mode as input in our implementation. DreaMS (Deep Representations of Any Mass Spectrum) is designed for spectral retrieval and similarity searching using a transformer-based foundational embedding model[37]. It maps mass spectra into a high-dimensional latent space where proximity reflects chemical and structural similarity. Our implementation implements DreaMS official inference script that leverages pre-trained DreaMS models to calculate similarity scores and retrieve the top-k most relevant SMILES strings within MassSpecGym dataset.

A.1.3 De Novo Molecular Structure Elucidation from MS Data.

DiffMS. DiffMS takes a precursor molecular formula, ionization type, and extracted fragment subformulas as input. DiffMS leverages a generative diffusion model to predict molecular structures by learning to generate valid molecular graphs conditioned on the input formula and spectral subformulas[38]. The model ensures generated structures are chemically consistent with the observed fragmentation pattern. Our implementation utilizes the official DiffMS inference script but modifies the graph structure source to rely on the encoder's fingerprint predictions.

Spec2Mol. Spec2Mol takes a raw MS/MS spectrum and precursor ionization mode as input. Spec2Mol utilizes a sequence-to-sequence translation approach with an Encoder-Decoder neural network to map spectral data directly to molecular strings[39]. It employs an encoder to transform the mass spectrum into a high-dimensional latent embedding, which a transformer-based decoder uses to autoregressively generate SMILES sequences. Our implementation manages data flow through Spec2Mol github, delegating the spectral encoding and SMILES decoding tasks to specialized scripts for efficient processing.

MSNovelist. MSNovelist takes a raw MS/MS spectrum and precursor ionization mode as input. MSNovelist represents a true de novo identification tool, generating molecular structures solely from MS/MS spectra without querying a structure database[40]. It combines fingerprint prediction using CSI:FingerID with an encoder-decoder neural network. The encoder transforms the predicted molecular fingerprint and formula into a latent vector, which the LSTM-based decoder then uses to generate SMILES sequences.

A.1.4 Chemistry-Aware Analytical Tools.

Efficient fragmentation simulation. This tool takes a molecule's SMILES string and adduct type as input. Efficient fragmentation function utilizes a rule-based expert system to predict theoretical fragmentation patterns, identifying potential neutral losses, fragile bonds, and characteristic fragments based on chemical substructures. The framework allows for the identification of core fragmentation pathways and interpretable spectral features without reliance on black-box models. Our implementation leverages a lightweight RDKit-based engine that processes configurable fragmentation rules, enabling rapid batch predictions for retrieved molecular candidates.

Chemical information identification. This tool takes a list of molecule SMILES strings as input. Chemical Information Identification function employs the RDKit cheminformatics suite to compute essential physicochemical descriptors and molecular properties. This function acts as a foundational characterization step and enriched molecules with detailed chemical metadata to support downstream spectral validation and filtering.

Enthalpy prediction. This tool takes candidate SMILES strings, the experimental MS2 spectrum, and the precursor type as input. It utilizes a thermodynamic analysis framework to predict enthalpy changes of fragmentation pathways, providing a physical basis for spectral annotation. Specifically, it matches experimental peaks to theoretical fragments generated by CFM-ID and computes their approximate formation enthalpies using RDKit molecular force fields to assess the structural likelihood of candidates.

MISTCF Formula Prediction. This tool takes the experimental MS2 spectrum and the precursor mass-to-charge ratio (m/z) as input. MISTCF Formula Prediction function determines the most probable molecular formula for the precursor ion by evaluating fragment distributions and intensity patterns[41]. Our implementation integrates the MIST-CF framework to prune the candidate formula space and select the optimal elemental composition.

MISTCF Subformula Annotation. This tool takes the experimental MS2 spectrum and the assigned precursor molecular formula as input. MISTCF Subformula Annotation function identifies the elemental composition of individual fragment peaks, effectively mapping product ions to their corresponding chemical sub-structures[41]. Our implementation employs a combinatorial decomposition strategy that matches observed fragment masses against valid subsets of the parent formula.

A.1.5 Sirius-series Analysis Tools.

Sirius Peak Annotation. This tool takes an experimental MS2 spectrum and precursor information as input. Sirius peak annotation tool systematically analyzes mass spectral data to calculate molecular formulas for the precursor ion and annotate fragment peaks, concurrently generating fragmentation trees that map parent-product ion relationships[42].

Sirius-CANOPUS Classification. This tool takes processed fragmentation trees from Sirius as input. It predicts compound classes and chemical ontologies directly from MS2 spectra without relying on spectral library matches, providing hierarchical classification

consistent with the ClassyFire taxonomy. Our implementation integrates CANOPUS (Class Assignment and Ontology Prediction Using Mass Spectrometry) into the Sirius workflow[42, 43], extracting probable chemical classes based on a configurable probability threshold to offer semantic context for unknown metabolites.

Sirius CSI:FingerID Fingerprint Prediction. This tool takes an experimental MS2 spectrum and precursor information as input. The CSI:FingerID Fingerprint Prediction function calculates molecular fingerprints that represent the existence of specific chemical substructures within the compound[44].

Sirius Structure Search. This tool takes an experimental MS2 spectrum and precursor information as input. Sirius Structure Search function identifies potential molecular candidates from extensive chemical databases by matching spectral features against predicted molecular fingerprints. Our implementation interfaces with the Sirius suite to predict molecular fingerprints from fragmentation trees and search them against biological and chemical structure databases. The tool retrieves ranked molecular candidates with their corresponding SMILES strings and structural names, using Tanimoto similarity scores to quantify the confidence of the matches.

A.2 LLM-series intelligence tools

A.2.1 Base Brain Model. The Base Brain Model functioning as the central dispatch agent is responsible for planning and orchestrating the execution of tools based on user intentions. The agent formats direct and multi-step tool options and recent session into a structured format. To determine the appropriate next action, the brain agent dynamically evaluates the availability of various analytical tools within the system. It intelligently distinguishes between directly available tools and arrivable tools in following steps.

A.2.2 Mass Spectrometry Data Processing.

Raw Spectrum Ingestion. This tool takes mass spectrometry files in MGF or mzML formats as input. Its primary function is to parse heterogeneous data sources, extracting critical spectral information such as precursor m/z, retention time, ion mode, and fragmentation peaks, and facilitating the standardization of this data into a unified internal format.

Feature-based MS2 Association. This tool takes LC-MS/MS data files, specifically in mzML format, as input. Its function is to perform chromatographic feature detection at the MS1 level to identify distinct metabolic peaks and accurately associate them with their generated MS2 fragmentation spectra, thereby resolving precursor ambiguity and enhancing spectral purity. It addresses the challenge of linking fragmentation events to their parent ions in complex LC-MS runs.

A.2.3 Database support.

PubChem Query. This tool takes a list of SMILES strings, PubChem CIDs, or molecular formulas as input. The PubChem query function serves as a primary reference for chemical structure validation and candidate generation, enabling both exact structure verification and formula-based isomer expansion[45]. It retrieves comprehensive metadata, including InChI strings and molecular identifiers, to contextualize spectral features. Our implementation utilizes a locally hosted dataset to perform high-efficiency filtering and matching, allowing for rapid retrieval of compound information without external API latency.

HMDB Query. This tool takes a list of SMILES strings as input. The HMDB query function maps molecular structures to the Human Metabolome Database to retrieve biological context, providing essential annotations such as metabolite names and descriptions[46]. It bridges chemical structure identification with biological interpretation.

A.3 Online platform

An interactive web-based demonstration of MSAgent is available, The platform allows users to interact with MSAgent through natural language instructions, inspect intermediate tool outputs, and observe how multiple domain-specific mass spectrometry tools are dynamically orchestrated into coherent analytical workflows. Both interactive (chat-style) and structured analysis modes are supported, reflecting common real-world MS analysis scenarios.The website is provided to facilitate understanding of the system design and to encourage feedback from the mass spectrometry and AI communities. The platform is intended for academic use only.

A.4 Benchmark Datasets and Evaluation Protocol

To evaluate the performance of MSAgent under realistic compound identification scenarios, we benchmarked the system on the CASMI 2016 and 2022 challenge series. The CASMI datasets provide large-scale testbeds designed to simulate untargeted metabolomics workflows, including raw MS/MS data acquired in both positive and negative modes, along with accurate precursor *m/z* values and retention times. The CASMI 2016 query set contains Orbitrap spectra for 188 unique compounds, with 124 [M+H]+ queries used for evaluation. An associated candidate list accompanies each spectrum. The CASMI 2022 query set contains 500 queries, with 229 [M+H]+ or [M+Na]+ queries used for evaluation. Associated candidate list is constructed by sampling compounds from PubChem with molecular masses within a specified tolerance of the true query mass for each challenge, followed the MassFormer setting to ensure there is no *a priori* information about the query molecule.

In addition, we expanded our evaluation using the Canopus (NPLIB1) dataset. Adopting the data split introduced by MIST, we selected 819 test samples representing 714 unique compounds. All of these samples were acquired in the positive ionization mode as [M+H]+ adducts. Finally, for every method evaluated across these contest settings, MSAgent automatically filtered out unsupported candidates that contained incompatible atoms.

To further assess the effect of multi-tool orchestration, domain knowledge integration, and comparison against LLM-only baselines, We constructed MSAgent-Bench, a comprehensive evaluation suite for assessing LLM-driven compound identification from tandem mass spectrometry (MS/MS) data. The benchmark is derived from 500 high-quality LC-MS/MS spectra curated from the MassBank RIKEN spectral library, each with complete structural annotations (SMILES, molecular formula, InChIKey) and experimental metadata (precursor m/z, adduct type, ion mode, instrument type). MSAgent-Bench comprises four complementary subsets designed

to probe distinct aspects of the identification pipeline: (1) MSAgent-Bench-Open evaluates end-to-end identification performance in an unconstrained setting where no candidate pool is provided, testing the system's ability to retrieve and rank candidate structures from scratch; (2) MSAgent-Bench-Pool assesses candidate discrimination under controlled difficulty, using pre-constructed candidate pools stratified into four levels—from easily distinguishable mass decoys to near-identical isomers—to systematically characterize how structural similarity among candidates affects identification accuracy; (3) MSAgent-Bench-MultiTool quantifies the benefit of integrating complementary computational tools (Iceberg and Sirius) by evaluating performance while no single tool dominates; and (4) MSAgent-Bench-Knowledge isolates the contribution of domain knowledge through ablation, comparing identification performance under four conditions—no knowledge, spectral knowledge only (Sirius peak annotations and theoretical fragmentation), chemical knowledge only (molecular descriptors), and their combination—to disentangle the roles of spectral reasoning and chemical property matching in tool-augmented identification. Together, these four subsets provide a multi-dimensional evaluation framework spanning open-domain retrieval, fine-grained structural discrimination, multi-tool synergy, and knowledge ablation, enabling rigorous and mechanistic comparison between tool-augmented agentic framework and standalone LLM baselines. The detailed construction process is provided in supplementary materials.

B method

Inputs and outputs definition. Given raw spectra $\mathcal{R}$, background information B, experimental and control data $(G_{\text{exp}}, G_{\text{ctrl}})$ when available, knowledge sources $\mathcal{K}$, a core large language model L,and an optional iteration budget T, MSAgent returns a set of validated claims $\mathcal{C}$, claim clusters $\mathcal{Z}$, system prompt s, mode-specific scientific reports $\mathcal{R}^{(1:3)}$, a narrative report $\mathcal{R}_{\text{report}}$, and an audit trace $\mathcal{T}$ linking each claim to feature ids, molecule identity decisions, KG rows, and cohort abundance statistics.

B.1 KG-Guided Scientific Attention allocation

The bottleneck in untargeted metabolomics is attention: a cohort may contain thousands of MS features, which assosiated with a unidentified molecule, but only a small subset are no trivial or can be deeply identified, searched, and interpreted. MSAgent treats feature selection as a knowledge-conditioned attention problem. Study background enrichment first constructs a context state by retrieve related information related disease or phenotype context, literature keywords, pathway priors, and sample-arm semantics.

$$S = \text{Enrich}(B; \mathcal{K}),$$

which contains the research question, This context is used both for experimental/control arm resolution and for downstream feature selection. For each feature f_i, we compute a feature descriptor

$$x_i = (\text{mz}_i, \text{rt}_i, \text{MS2}_i, q_i, \Delta_i, a_i)\,,$$

where q_i summarizes spectral quality, Δ_i is fold-change or contrast ratio, and a_i records the feature's source arm. A policy then assigns knowledge-conditioned attention scores based on the content a predifined scoring rubric

$$\alpha_{i,\ell}^{(t)} = L(x_i, S) \tag{2}$$

The selected feature set is the budgeted top-k subset

$$S_{t,\ell} = \text{Topk}\left(\{\alpha_{i,\ell}^{(t)}\}\right). \tag{3}$$

This attention mechanism is to convert background knowledge and cohort statistics into a resource-allocation decision: which features deserve expensive molecular identification and downstream biological reasoning.

B.2 Autonomous Molecule Identification via ReAct-Guided Tool Orchestration

Before synthesizing high-level biological claims, MSAgent must resolve the structural identity of each mass spectrometry feature $i \in F$. Given the raw spectral data $\mathcal{R}_i \in \mathcal{R}$ and background knowledge $\mathcal{K}$, the objective is to determine the definitive molecule identity decision $\xi_i^{\text{id}} \in \Xi_i$. Because mass spectrometry data is high-dimensional and heterogeneous, MSAgent achieves this by executing a multi-step Reasoning-and-Acting (ReAct) loop that iteratively orchestrates tools registered within MSToolbox.

For a specific feature i, the identification process is modeled as an interactive session over discrete execution steps $k \in \{1, \ldots, K\}$. We define the session state at step k as $\mathcal{S}_i^{(k)}$, which tracks the accumulated data keys and resolved molecular attributes. At each step, the core LLM L ingests the current session state, the historical execution trace, and the task-specific system instruction s to generate a reasoning *thought* $h^{(k)}$ and select an analytical *action* $a^{(k)}$:

$$\left(h^{(k)},\ a^{(k)}\right) = L\left(\mathcal{S}_i^{(k-1)},\ h^{(k-1)},\ a^{(k-1)},\ s\right). \tag{4}$$

The action $a^{(k)}$ corresponds to the invocation of a registered tool $T_j \in$ MSToolbox. To guarantee execution safety and determinism, the system constraints $a^{(k)}$ using the directed dependency graph $G = (V, E)$ constructed by MSToolbox. A tool T_j is accessible if and only if its declared input dependency set D_j is fully satisfied by the current session state:

$$a^{(k)} = T_j \quad \text{subject to} \quad D_j \subseteq \mathcal{S}_i^{(k-1)}. \tag{5}$$

If a target tool's dependencies are unsatisfied, the agent performs a reachability analysis over G to traverse antecedent nodes and chain prerequisite tools (e.g., routing from raw spectral preprocessing to *in silico* fragments generation, and finally to database search engine queries).

Once a valid tool T_j is invoked, it executes within the optimized sub-minute latency environment, generating a structured biological *observation* $o^{(k)}$ that populates the tool's declared output keys O_j:

$$o^{(k)} = T_j\left(\mathcal{S}_i^{(k-1)}\right). \tag{6}$$

The session state is then dynamically advanced via an update function that merges the new observations while maintaining complete provenance:

$$\mathcal{S}_i^{(k)} = \mathcal{S}_i^{(k-1)} \cup \{o^{(k)}\}. \tag{7}$$

This ReAct loop terminates when the agent encounters a convergence criteria—such as a definitive database hit with high spectral similarity—or exhausts the local step budget K. The final converged

session state $\mathcal{S}_i^{(K)}$ is compiled to resolve the identity boundary slot within the molecule's context tuple:

$$\xi_i^{\text{id}} = \text{ResolveIdentity}\left(\mathcal{S}_i^{(K)}\right), \tag{8}$$

where ξ_i^{id} is categorized into distinct epistemic tiers (e.g., unresolved feature, formula-only observation, putative structure, or resolved database identity). Concurrently, the entire interactive trajectory, including tool inputs, intermediate spectral scores, and database PMIDs, is logged into $\xi_i^{\text{provenance}}$, establishing an immutable foundation for the downstream deterministic validator V.

B.3 Molecule-level biological context acquisition

Before any claim is written, MSAgent asks what biological context is actually available for each molecule-like feature. The acquired context is represented as

$$\Xi_i = \left(\xi_i^{\text{id}}, \xi_i^{\text{cohort}}, \xi_i^{\text{disease}}, \xi_i^{\text{mechanism}}, \xi_i^{\text{literature}}, \xi_i^{\text{provenance}}\right).$$

Here ξ_i^{id} records the identity boundary, such as unresolved feature, formula-only observation, putative structure, resolved structure, or resolved database identity. ξ_i^{cohort} records the local signal: direction, fold-change or arm contrast, abundance summaries, and the group semantics that define experimental and control arms. ξ_i^{disease} records exact KG disease/context rows, direct molecule-disease PubMed passages, and indirect structural-neighbor disease context when exact support is absent. $\xi_i^{\text{mechanism}}$ records pathway, mechanism, and KG-neighborhood signals. $\xi_i^{\text{literature}}$ records bounded PubMed passages and their retrieval class, rather than only hit counts. Finally, $\xi_i^{\text{provenance}}$ links the feature id, MS-tool outputs, database identifiers, KG rows, PMIDs, retained passages, and full-report locations. This step makes the molecule's evidence state explicit before the LLM sees it.

Structural-similarity KG bridge. Exact molecule evidence is often sparse because many MS candidates have no direct disease or pathway annotations. MSAgent therefore queries the KG through two separate channels:

$$E_i^{\text{KG}} = E_i^{\text{KG,exact}} \cup E_i^{\text{KG,sim}},$$

where $E_i^{\text{KG,exact}}$ contains exact database or alias matches and

$$E_i^{\text{KG,sim}} = \{(j, \rho_{ij}, c_j) : \rho_{ij} \geq \tau\}$$

contains structurally similar KG neighbors j, similarity ρ_{ij}, and neighbor context c_j. In the current implementation, ρ_{ij} is a fingerprint similarity score and τ defaults to 0.5 for retained same-bin or structural-neighbor support. This structural-similarity channel is algorithmically important because it lets the system cross an information barrier: a newly observed or weakly annotated molecule can still inherit hypothesis-grade biological context from chemically similar molecules. MSAgent keeps this channel separate from exact evidence, so it can guide mechanism discovery without being misreported as exact molecule-disease proof.

B.4 Evidence Bundling and Multi-Molecule Claim Synthesis

Raw biological contexts captured in Ξ_i are often redundant across different features. To expose cross-molecule convergence, MSAgent aggregates the fragmented biological facts—referred to as "atoms" $\mathcal{A}_i \subseteq (\xi_i^{\text{disease}} \cup \xi_i^{\text{mechanism}})$—into shared mechanism evidence bundles:

$$\mathcal{B} = \text{Bundle}\left(\bigcup_{i \in F} \mathcal{A}_i\right). \tag{9}$$

These bundles are keyed by evidence family and biological target label. The current families encompass exact/direct disease evidence, mechanism/pathway evidence, and indirect context. Crucially, $\mathcal{B}$ does not create new facts; it acts as a compressed reasoning scaffold that couples cross-feature insights while preserving representative atoms and their intrinsic limitations.

First, the LLM claim synthesizer L receives the compact representations of both individual molecule contexts and shared mechanism bundles to generate candidate atomic claims $\widetilde{C}$:

$$\widetilde{C} = L\left(\text{compact}(\{\Xi_i : i \in F\}),\ \text{compact}(\mathcal{B})\right). \tag{10}$$

The synthesizer is explicitly instructed to generate concrete atomic claims rather than high-level category names, and each claim must cite valid feature identifiers. Subsequently, the deterministic validator V enforces evidence boundaries to yield the finalized admissible claim set C:

$$C = V(\widetilde{C}, \{\Xi_i : i \in F\}). \tag{11}$$

The validator rejects any claim that lacks feature citations, references display-only features, or violates data eligibility.

B.5 Hierarchical Scientific Report Generation

Rather than invoking static, template-driven reporting functions, MSAgent generates comprehensive scientific reports through a secondary, evidence-aware reasoning stage. As specified in the global input definitions, the choice of the active discovery mode $m \in \{\text{study}, \text{explore}\}$ is primarily dictated by the nature of the background information B. Based on m, MSAgent instantiates a mode-dependent semantic configuration vector $\Lambda_m = (\lambda_m^{(1)}, \lambda_m^{(2)}, \lambda_m^{(3)})$, defined as:

$$\Lambda_{\text{study}} = (\text{validation},\ \text{supplementation},\ \text{hypothesis}), \tag{12}$$

$$\Lambda_{\text{explore}} = (\text{finding},\ \text{grounding},\ \text{hypothesis}). \tag{13}$$

For each progressive layer $r \in \{1, 2, 3\}$, the system constructs a layer-specific evidence pool Ω_r to isolate the facts admissible within that layer's epistemic boundaries:

$$\Omega_r = \text{SelectEvidence}\left(C,\ \mathcal{Z},\ \{\Xi_i\}_{i \in F},\ \mathcal{B},\ \mathcal{T},\ \lambda_m^{(r)}\right). \tag{14}$$

The selector filters the validated claims C, claim clusters $\mathcal{Z}$, molecule contexts $\{\Xi_i\}$, mechanism bundles $\mathcal{B}$, and their corresponding audit traces $\mathcal{T}$ to form a bounded context for generation.

To prevent factual drift and unconstrained generation, the core LLM L operates as a dynamic purification process over two cascaded steps for each layer r. First, it acts as a scientific distiller to compress the heterogeneous evidence pool Ω_r into a dense, structural report state D_r:

$$D_r = L\left(\Omega_r,\ \lambda_m^{(r)},\ s\right), \tag{15}$$

where D_r encapsulates representative observations, convergent mechanisms, literature anchors, and explicit uncertainty boundaries. Subsequently, the final scientific report for layer r, denoted as $\mathcal{R}^{(r)}$, is synthesized by conditioning on both the current distilled state D_r and the text compiled in preceding report layers $\mathcal{R}^{(<r)}$:

$$\mathcal{R}^{(r)} = L\left(D_r,\ \mathcal{R}^{(<r)},\ \lambda_m^{(r)},\ s\right). \tag{16}$$

This sequential conditioning is logically imperative: it forces $\mathcal{R}^{(2)}$ to explicitly articulate incremental insights added beyond $\mathcal{R}^{(1)}$, and compels $\mathcal{R}^{(3)}$ to transparently bound speculative mechanisms as hypothesis-grade relative to the validated foundations of $\mathcal{R}^{(<3)}$.

B.6 Report

Diverse discovery modes. MSAgent uses the same evidence engine under two scientific modes, primilarily decided by the type of background information B . In study-driven mode, B indicates the existing research background, which can be the abstract of a specific paper or even the full paper content or some already have a major research report line, Then, scientific reports $\mathcal{R}^{(1:3)}$ generated by MSAgent can be represented as:

$$\mathcal{R}_{\text{study}} = (\text{validation}, \text{supplementation}, \text{hypothesis}),$$

where the validation checks the stated study question, the second searches for additional supporting context like some , and the third proposes testable hypothesis that doesent existes in original study. When B indicates an existing research background (e.g., a specific paper abstract, full text, or a pre-defined major report line), the three-level report $\mathcal{R}^{(1:3)}$ functions as a controlled scientific expansion from an established study backbone.

- **Validation Layer ($\mathcal{R}^{(1)}$):** This layer is not merely a reproduction step; it tests whether the original biological claim is sufficiently solid under an independent, evidence-grounded re-analysis, providing the agent with a reliable epistemic anchor before broader reasoning is allowed.
- **Supplementation Layer ($\mathcal{R}^{(2)}$):** Given that anchor, this layer identifies which metabolite classes, pathways, disease contexts, or mechanism-level signals were under-explored by the original study but are strongly supported by the current experimental cohort, KG, PubMed, and provenance evidence.
- **Hypothesis Layer ($\mathcal{R}^{(3)}$):** This layer deliberately moves beyond the original study boundaries, utilizing the remaining experimental signals and indirect mechanistic support to nominate new, testable biological insights.

This structured progression from robustness assessment, to mechanistic deepening, to disciplined discovery, systematically prevents the system from treating speculative associations as validated conclusions.

In exploration mode, $\mathcal{R}_{\text{report}}$ is:

$$\mathcal{R}_{\text{explore}} = (\text{finding}, \text{grounding}, \text{hypothesis}),$$

In exploration mode, the three-level report is designed for situations where no sufficiently specific prior biological storyline should dominate the analysis. when no prior biological storyline is provided to dominate the analysis, the three-level report is tailored for untargeted open-ended discovery, transforming a high-dimensional feature space into a layered argument.

- **Finding Layer ($\mathcal{R}^{(1)}$):** This layer preserves the absolute autonomy of the cohort data, identifying discriminative molecular signals, abundance directions, and identity confidence without forcing them prematurely into speculative mechanistic narratives.
- **Grounding Layer ($\mathcal{R}^{(2)}$):** This layer performs the critical act of scientific anchoring, systematically connecting raw data observations to known disease biology, pathways, KG neighborhoods, and PubMed passages while explicitly separating exact molecular evidence from weaker contextual support.
- **Hypothesis Layer ($\mathcal{R}^{(3)}$):** Utilizing the grounded evidence landscape, this layer synthesizes candidate mechanisms that extend beyond any single molecule yet remain strictly traceable to concrete feature signals, nominating robust mechanisms that deserve downstream experimental follow-up.

Algorithm 1 Dynamic Multi-Tiered Report Synthesis and Verification

Require: Validated claims C, claim clusters $\mathcal{Z}$, molecule contexts $\{\Xi_i\}_{i\in F}$, mechanism bundles $\mathcal{B}$, active discovery mode m, system instruction s, global audit trace $\mathcal{T}$

Ensure: Multi-tiered scientific reports $\mathcal{R}^{(1:3)}$, cross-layer narrative report $\mathcal{R}_{\text{report}}$, updated audit trace $\mathcal{T}$

Retrieve mode-dependent semantic configuration vector $\Lambda_m \leftarrow (\lambda_m^{(1)}, \lambda_m^{(2)}, \lambda_m^{(3)})$
for $r = 1, 2, 3$ **do**
  $\Omega_r \leftarrow \text{SelectEvidence}\left(C,\ \mathcal{Z},\ \{\Xi_i\}_{i\in F},\ \mathcal{B},\ \mathcal{T},\ \lambda_m^{(r)}\right)$ ▷ Isolate layer-admissible evidence pool
  $D_r \leftarrow L\left(\Omega_r,\ \lambda_m^{(r)},\ s\right)$ ▷ Distill heterogeneous evidence into structural report state
  $\widetilde{\mathcal{R}}^{(r)} \leftarrow L\left(D_r,\ \mathcal{R}^{(<r)},\ \lambda_m^{(r)},\ s\right)$ ▷ Synthesize raw sequential conditional report layer
  $\mathcal{R}^{(r)} \leftarrow \text{VerifyAndCorrect}\left(\widetilde{\mathcal{R}}^{(r)},\ \Omega_r,\ \mathcal{T}\right)$ ▷ Deterministic boundary enforcement & factuality alignment
  $\mathcal{T} \leftarrow \mathcal{T} \cup \text{ExtractLineage}(\mathcal{R}^{(r)})$ ▷ Append current layer lineage to global audit trace
end for
$\widetilde{\mathcal{R}}_{\text{report}} \leftarrow L\left(\mathcal{R}^{(1:3)},\ s\right)$ ▷ Synthesize global cross-layer narrative summary
$\mathcal{R}_{\text{report}} \leftarrow \text{VerifyAndCorrect}\left(\widetilde{\mathcal{R}}_{\text{report}},\ \{\Xi_i\}_{i\in F},\ \mathcal{T}\right)$ ▷ Attach finalized claim, feature, and PMID pointers
return $\mathcal{R}^{(1:3)},\ \mathcal{R}_{\text{report}},\ \mathcal{T}$

Algorithm 2 Evidence-Grounded Full-Cycle Metabolomics Discovery

Require: Raw spectra $\mathcal{R}$, background information B, experimental data $(G_{\text{exp}}, G_{\text{ctrl}})$, knowledge sources $\mathcal{K}$, core LLM L, system instruction s, maximum iteration budget T

Ensure: Validated claims C, claim clusters $\mathcal{Z}$, multi-tiered scientific reports $\mathcal{R}^{(1:3)}$, narrative report $\mathcal{R}_{\text{report}}$, global audit trace $\mathcal{T}$

$\mathcal{B}_{\text{context}} \leftarrow \text{EnrichBackground}(B)$ ▷ Align study metadata, disease landscape, and contrasts
$F \leftarrow \text{ExtractCohortFeatures}(\mathcal{R},\ G_{\text{exp}},\ G_{\text{ctrl}})$ ▷ Quantify alignment, compile cohort statistics and profiles
Initialize global discovery state $\mathcal{H}_0 \leftarrow \emptyset$, global audit trace $\mathcal{T} \leftarrow \emptyset$
Determine active discovery mode $m \in \{\text{study}, \text{explore}\}$ based on B
for $t = 1, \ldots, T$ **do**
  $F_t \leftarrow \Pi(F,\ \mathcal{H}_{t-1},\ \Lambda_m)$ ▷ Select active feature subset via budgeted tier policy Π
  for each active feature $i \in F_t$ **do**
    $\xi_i^{\text{id}} \leftarrow \text{OrchestrateTools}(i,\ \mathcal{R}_i,\ \text{MSToolbox}\ ;\ s)$ ▷ Execute ReAct loop to resolve identity tier
    $\Xi_i \leftarrow \text{AcquireContext}(i,\ \xi_i^{\text{id}},\ \mathcal{K},\ \mathcal{T})$ ▷ Compile disease, mechanism, and literature context
  end for
  $\mathcal{I}_t \leftarrow L\left(\{\Xi_i\}_{i \in F_t},\ \mathcal{H}_{t-1}\ ;\ s\right)$ ▷ Extract cross-feature layered biological insights
  $\mathcal{H}_t \leftarrow \text{Update}\ (\mathcal{H}_{t-1},\ \mathcal{I}_t,\ F_t)$ ▷ Advance accumulated multi-round discovery state
  $\mathcal{T} \leftarrow \mathcal{T} \cup \text{Snapshot}(\mathcal{H}_t)$ ▷ Commit intermediate execution states to global audit trail
  if $\text{StoppingRule}(\mathcal{H}_t)$ **then break**
  end if
end for
Extract relevant assertions $\mathcal{A}_i \subseteq (\xi_i^{\text{disease}} \cup \xi_i^{\text{mechanism}})$ for all evaluated features $i \in F$
$\mathcal{B} \leftarrow \text{Bundle}\ (\bigcup_{i \in F} \mathcal{A}_i)$ ▷ Aggregate cross-molecule atoms into mechanism evidence bundles
$\widetilde{C} \leftarrow L\ (\text{compact}(\{\Xi_i : i \in F\}),\ \text{compact}(\mathcal{B})\ ;\ s)$ ▷ Synthesize multi-molecule candidate claims
$C \leftarrow V\left(\widetilde{C},\ \{\Xi_i : i \in F\}\right)$ ▷ Enforce deterministic biological boundary validation
$\mathcal{Z} \leftarrow \text{Cluster}(C)$ ▷ Group validated atomic claims into coherent biological axes
$\left(\mathcal{R}^{(1:3)},\ \mathcal{R}_{\text{report}},\ \mathcal{T}\right) \leftarrow \text{DynamicSynthesis}\ (C, \mathcal{Z}, \{\Xi_i\}_{i \in F}, \mathcal{B}, m, s, \mathcal{T})$ ▷ Invoke Alg. 1
return $C,\ \mathcal{Z},\ \mathcal{R}^{(1:3)},\ \mathcal{R}_{\text{report}},\ \mathcal{T}$

C Scalable Data Infrastructure and External Database Integration

To bridge the gap between raw mass spectrometry signals and biological insight, MSAgent is supported by a robust data infrastructure. We have integrated two primary databases—HMDB and PubChem—to provide a scalable foundation for chemical discovery.

The HMDB module is designed to contextualize spectral matches within the human metabolome. It features a high-performance lookup system that maps retrieved SMILES strings to unique HMDB identifiers. Once identified, the system automatically pulls detailed metabolite profiles—including physiological concentrations, tissue locations, and biochemical pathways—ensuring that every match is immediately linked to its biological significance.

Complementing this, our PubChem module offers versatile tools for candidate generation and filtering. The PubChem Formula Search allows researchers to rapidly identify potential structures based solely on molecular formulas, effectively narrowing the search space for unknown compounds. Furthermore, the framework supports targeted filtering of compound data by SMILES or Chemical Identification (CID) numbers. Together, these integrations ensure that every structural assignment is backed by comprehensive metadata, transforming raw data into actionable chemical knowledge.

C.1 Dataset Construction

To systematically evaluate MSAgent's compound identification capabilities, we constructed **MSAgent-Bench**, a benchmark suite comprising four complementary evaluation subsets. All subsets are derived from a common pool of 500 high-quality LC-MS/MS spectra curated from the MassBank RIKEN spectral library. Each spectrum is annotated with compound name, SMILES, molecular formula, InChIKey, precursor m/z, adduct type, ion mode, and instrument type. The curation pipeline retained only entries with complete structural annotations and valid MS2 spectra.

C.1.1 MSAgent-Bench-Open. This subset evaluates end-to-end identification in an unconstrained setting where no candidate pool is provided. Thirty spectra were sampled from the curated collection. MSAgent operated with a full tool chain: database structure search, SIRIUS peak annotation, chemical property retrieval, and theoretical fragmentation prediction. The baseline (LLM) received only the MS2 spectrum and metadata without access to any computational tools. Both methods were evaluated on exact structure match, Tanimoto similarity (Morgan fingerprint, radius = 2, 2048 bits), Top-K recall, molecular formula accuracy, and confidence calibration.

C.1.2 MSAgent-Bench-Pool. This subset isolates different methods' ability to discriminate among structurally related candidates under controlled difficulty. Candidate pools were pre-constructed for all 500 spectra by querying PubChem, and 20 spectra per difficulty level were sampled for evaluation. Four pool configurations of increasing difficulty were designed:

- **Cross-Formula**: decoy molecules with different molecular formulae but similar monoisotopic mass (±10 ppm), which can typically be eliminated through formula determination alone.
- **Structural Variants**: molecules sharing the same molecular formula but exhibiting low structural similarity (Tanimoto < 0.5), requiring fragmentation-level reasoning to distinguish.

- **Near-Isomers**: highly similar isomers with the same formula and high structural similarity (Tanimoto ≥ 0.5), representing the most challenging case demanding fine-grained spectral matching.
- **Full Pool**: a mixed pool combining all three categories (3–5 isomers, 3–5 structural variants, 2–3 mass decoys), totaling 8–13 candidates per spectrum.

Each pool included the ground truth molecule, and candidates were randomly shuffled to prevent positional bias. Since candidates were pre-provided, MSAgent tool chain excluded the database search tool and consisted of SIRIUS peak annotation, chemical property retrieval, and theoretical fragmentation prediction, followed by spectral similarity-based re-ranking.

C.1.3 MSAgent-Bench-MultiTool. This subset quantifies the benefit of integrating complementary computational tools. Thirty-five spectra were sampled from the curated collection, covering diverse scenarios where different tools exhibit complementary strengths. For each spectrum, MSAgent constructed its candidate pool (~20 candidates) by combining Top-K predictions from Iceberg (a deep learning-based spectral simulator) and SIRIUS (a combinatorial fragmentation engine), with deduplication by canonical SMILES. Five methods were compared: (1) **MSAgent**—full multi-tool integration; (2) **w/o Domain Tools**—MSAgent without domain-specific analysis tools; (3) **Pure LLM**—no tools and no candidate pool; (4) **Iceberg Top-1**—best single prediction from Iceberg; and (5) **SIRIUS Top-1**—best single prediction from SIRIUS.

C.1.4 MSAgent-Bench-Knowledge. This subset isolates the contribution of domain knowledge through ablation. Fifty spectra were sampled with candidate pools of 10 molecules each. Two types of pre-computed domain knowledge were prepared: (K1) spectral knowledge, comprising SIRIUS peak annotations and theoretical fragmentation patterns for all candidates; and (K2) chemical knowledge, comprising molecular descriptors (exact mass, ring count, LogP, TPSA, H-bond donors/acceptors) for both the ground truth and all candidates. Four conditions were compared: LLM, +K1, +K2, and +K1+K2.

D Online Resources

The MSAgent platform is open, providing an interactive gateway for automated metabolomics interpretation and evidence-grounded biological hypothesis generation.