MARS: REFLECTIVE TREE SEARCH FOR MOLECULAR STRUCTURE ELUCIDATION FROM MASS SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Mass spectrometry (MS) is a fundamental tool for molecular identification, yet automatically elucidating molecular structures from MS spectra remains a long-standing challenge. Existing approaches either rely on retrieval-based methods constrained by database coverage, or on one-shot *de novo* generation, which often yields chemically implausible structures due to the lack of verification mechanisms. Inspired by the iterative reasoning workflow of human experts, we propose **MARS**, a **M**olecular **A**gent framework for **R**eflective Tree **S**earch based on MS spectra. MARS reformulates one-shot molecular prediction as a structured, reflection-driven search process powered by LLMs, following a three-stage expert workflow of intuitive hypothesis, logical reflection, and local refinement. It first retrieves structurally similar references from a large-scale external knowledge base, and an Intuitive-Prior Chain-of-Thought (IP-CoT) synthesizes them with spectral evidence into high-quality initial candidates. A Spectral Diagnostic Chain-of-Thought (SD-CoT) then translates spectrum-level discrepancies into verbalized correction signals, which guide a Monte Carlo Tree Search (MCTS) to iteratively refine candidates within the chemical space. MARS achieves competitive performance on NPLIB1 and MolPuzzle datasets and generalizes well to real wastewater spectra under distribution shifts, highlighting the value of combining prior-guided hypotheses with spectrum-grounded structural refinement. Our code can be found at https://anonymous.4open.science.

1 INTRODUCTION

Mass spectrometry (MS) is a core analytical tool for molecular identification in modern analytical chemistry, drug discovery, and natural product research Kind et al. (2018); Wen & Zhu (2015); Wang et al. (2016). By capturing precursor masses and fragmentation patterns, MS provides rich structural constraints for molecular identification. In recent years, deep learning-based approaches have achieved remarkable progress in mass spectral interpretation, spanning tasks from functional group identification to full molecular structure elucidation, substantially reducing manual effort and enabling automated, high-throughput molecular analysis Wang et al. (2025); Qin et al. (2026); Bohde et al. (2025).

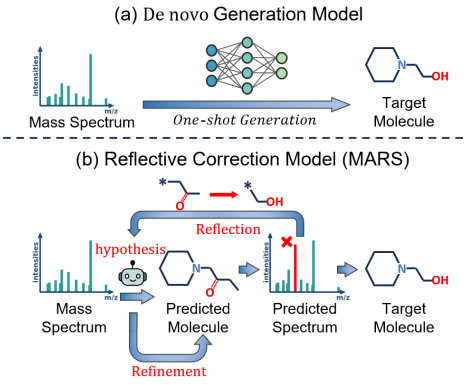

Figure 1: Comparison of two paradigms for MS-based molecular structure elucidation.

Overall, MS-based molecular prediction methods can be broadly divided into retrieval-based and *de novo* generation approaches. Retrieval-based methods identify molecules by matching query spectra against reference libraries, achieving high accuracy but remaining limited by library coverage. In contrast, *de novo* methods generate structures directly from spectra and thereby reduce dependence on reference libraries Elser et al. (2023); Litsa et al. (2023); Skinnider et al. (2021). Many supervised *de novo* generators, however, lack an explicit feedback loop for verifying and correcting candidate structures against the observed spectrum. Without such verification, these methods may produce chemically implausible structures Su et al. (2022);

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Verissimo et al. (2022). Adapting these parameterized models to shifted spectral conditions can also require additional task-specific data. Human experts address this limitation through an iterative hypothesis–verification process, refining candidate structures against spectral evidence Feng et al. (2025); Elyashberg & Williams (2015); however, this approach requires extensive expertise and substantial manual effort Madika et al. (2025); van den Broek et al. (2025).

In recent years, LLM agents have demonstrated strong capabilities in multi-step reasoning and autonomous decision making across complex tasks Ferrag et al. (2025); Wei et al. (2026). Notably, the two core elements underlying expert spectral interpretation—reflective correction and iterative search—closely align with two key emerging capabilities in LLM agent research. LLMs can iteratively refine their own predictions through verbalized self-reflection, yielding substantial gains on complex reasoning tasks Madaan et al. (2023); Shinn et al. (2023), while agents built upon Monte Carlo Tree Search (MCTS) enable efficient structured exploration over vast solution spaces in scientific domains such as chemical analysis Zhuang et al. (2025); Sprueill et al. (2023). The integration of these two capabilities opens up the possibility of automatically reproducing the expert interpretation workflow in an agentic manner, offering a new reasoning paradigm for MS-driven molecular structure elucidation.

In this work, we propose **MARS**, a **M**olecular **A**gent framework for **R**eflective Tree **S**earch based on MS spectra. As illustrated in Figure 1, MARS follows a three-stage cognitive cycle inspired by human experts: intuitive hypothesis—logical reflection—local refinement. It reformulates one-shot molecular prediction as a structured, reflection-driven search process, progressively converging to the target molecule under explicit verification. Rather than adapting model parameters, MARS adapts candidate structures in the search space, using explicit spectral feedback to guide inference-time correction. The two CoT mechanisms make the LLM's role explicit: IP-CoT turns chemical priors and spectral evidence into structural hypotheses, while SD-CoT turns discrepancies into proposed edits that can be checked against molecular and spectral constraints. In **intuitive hypothesis** stage, to mitigate the drift toward chemically implausible regions in *de novo* generation, MARS builds a large-scale external molecular knowledge base as long-term memory and retrieves structurally similar references via fingerprint prediction as rich priors for initialization. An Intuitive-Prior Chain-of-Thought (IP-CoT) then guides the agent to integrate the target formula and spectral evidence into initial candidates, offering a high-quality warm start. In **logical reflection** stage, to prevent unanchored self-reflection from spiraling into self-reinforcing hallucinations, we introduce a Spectral Diagnostic Chain-of-Thought (SD-CoT). By comparing the predicted and target spectra, SD-CoT localizes substructures responsible for matched and missing peaks and produces verbalized correction suggestions, anchoring refinement to external spectral evidence. In **local refinement** stage, to overcome local optima where a single reflection trajectory often oscillates, MARS adopts Monte Carlo Tree Search (MCTS) as the global planner, organizing candidates as tree nodes and invoking SD-CoT at each expansion step, with spectral matching as the reward. Through the iterative loop, MARS efficiently explores the chemical space and converges to structures aligned with the target spectrum. Our contributions can be summarized as follows:

- We propose **MARS**, an molecular structure elucidation framework that reformulates one-shot molecular prediction as a structured, reflection-driven search process.
- We introduce spectral-aware reasoning mechanisms that enable LLM agents to perform prior-guided initialization and evidence-driven refinement. IP-CoT integrates retrieved molecular knowledge with spectral constraints to generate reliable candidates, while SD-CoT converts spectral discrepancies into actionable structural edits. Combined with MCTS, these mechanisms enable systematic exploration of the molecular space.
- Experiments on NPLIB1 and MolPuzzle, together with case studies on real wastewater spectra, show that MARS remains competitive on standard benchmarks and transfers to shifted settings without task-specific fine-tuning.

2 METHODOLOGY

2.1 PERCEPTION AND LONG-TERM MEMORY RETRIEVAL

MARS adopts a modular perception–memory design that decouples the knowledge base, the perception module, and the retrieval mechanism, allowing each component to be independently improved or

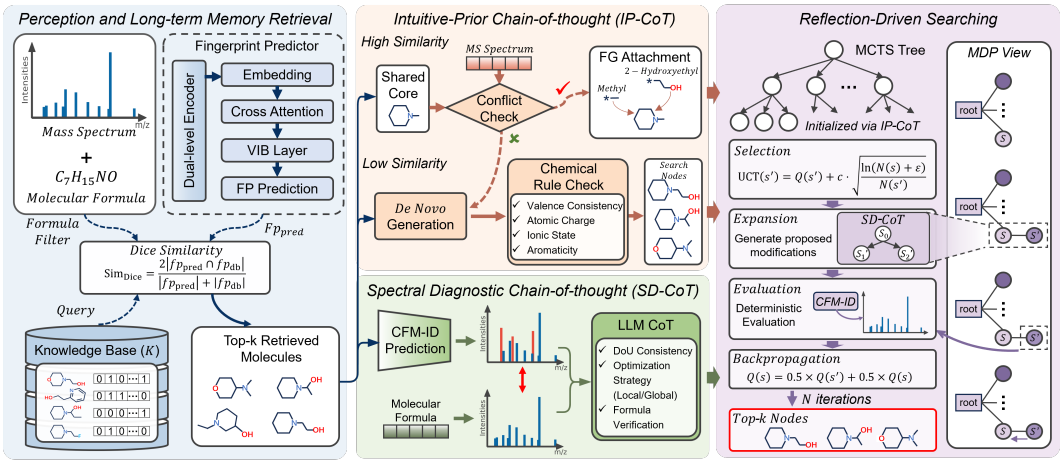

Figure 2: Overview of the MARS framework. Given a target spectrum, MARS encodes it into molecular fingerprint and retrieves the top-k similar references from a knowledge base $\mathcal{K}$ as long-term memory. The IP-CoT then generates initial candidates via scaffold-based attachment under high similarity or *de novo* generation under low similarity. The SD-CoT produces a verbalized diagnostic report that guides the LLM in proposing chemically meaningful edits. These candidates initialize an MCTS tree, in which MARS iteratively refines them to return the top-k predictions.

replaced. This design provides the agent with broad chemical priors to anchor its reasoning, while bridging the modality gap between raw spectra and the structural representation used for retrieval.

Long-term Memory Construction. To equip the agent with reliable chemical priors, we construct a large-scale external molecular knowledge base $\mathcal{K}$ as its long-term memory. The library consists of $\sim$4M molecules drawn from MassSpecGym Bushuiev et al. (2024). Each molecule in $\mathcal{K}$ is indexed by its Morgan fingerprint computed with RDKit, enabling efficient retrieval during inference.

Spectrum-to-Fingerprint Perception. A central challenge in retrieval is bridging the modality gap between the input mass spectrum x_{ms} and the fingerprint-indexed knowledge base. To this end, MARS includes a fingerprint perception module that maps a tandem mass spectrum into a probabilistic fingerprint vector $fp_{\text{pred}} \in [0,1]^d$, where each dimension estimates the probability that the corresponding substructure bit is present in the underlying molecule. The predicted fingerprint is binarized with threshold $\varepsilon = 0.2$ and used as the query representation. We instantiate this module with MIST Goldman et al. (2023) as a representative implementation and treat the perception output as an inherently noisy estimate that downstream reflection-driven search must remain robust to.

Memory Retrieval. Given the predicted fingerprint fp_{pred}, MARS retrieves structural priors from $\mathcal{K}$ in two steps. First, a hard filter restricts the search space to molecules whose heteroatom composition matches the target formula x_{formula}. Within this filtered subspace, MARS ranks candidates by the Dice similarity between fp_{pred} and each candidate fingerprint fp_{db}:

$$\mathrm{Sim}_{\mathrm{Dice}}(fp_{\text{pred}}, fp_{\text{db}}) = \frac{2 \cdot |fp_{\text{pred}} \cap fp_{\text{db}}|}{|fp_{\text{pred}}| + |fp_{\text{db}}|}. \tag{1}$$

The top-k molecules with the highest similarity scores are retrieved as structural context, providing the agent with high-confidence set of analogs to ground its initial hypothesis.

2.2 Chain-of-Thought based Spectral Reasoning Mechanism

MARS decomposes spectral reasoning into two stages: a *prior-driven* mode that exploits accumulated chemical experience to propose a plausible structure at the start of analysis, and an *evidence-driven* mode that iteratively reconciles the candidate structure with the observed spectrum. To formalize the two modes within MARS, we design two specialized modules: an Intuitive-Prior CoT (IP-CoT) for generating prior-grounded initial candidates, and a Spectral Diagnostic CoT (SD-CoT) for producing verbalized correction signals during search.

2.2.1 INTUITIVE-PRIOR CHAIN-OF-THOUGHT (IP-COT)

The IP-CoT formalizes the prior-driven mode by guiding LLM to integrate the target spectrum, the molecular formula, and the SMILES representations of the top-k retrieved references into an initial structural hypothesis. Since retrieved priors vary widely in informativeness, IP-CoT applies a similarity-adaptive strategy to balance references against spectral evidence.

High-Similarity Branch. For high-similarity references, the LLM extracts their shared core scaffold as a structural anchor and performs formula- and spectrum-consistency checks before scaffold-based attachment. Rejected branches fall back to the low-similarity case. Otherwise, the formula and Degree of Unsaturation (DoU) gap between the target molecule and the scaffold is computed, and the LLM attaches functional groups or side chains to the scaffold accordingly.

Low-Similarity Branch. When references provide only weak structural cues, IP-CoT switches to a spectrum-driven mode in which the LLM reasons primarily over the molecular formula, DoU, characteristic fragment ions, and common neutral losses, treating retrieved fragments only as auxiliary hints rather than as a structural template.

Validation and Initialization. Generated candidates are filtered by chemical validity checks (e.g., valence consistency) and then instantiated as the first-layer nodes of the search tree. Through this similarity-adaptive design, IP-CoT actively integrates retrieved priors with spectral evidence rather than passively interpolating between them, providing the subsequent search with high-quality, chemically grounded starting points.

2.2.2 SPECTRAL DIAGNOSTIC CHAIN-OF-THOUGHT (SD-COT)

While IP-CoT supplies a strong starting point, refining a candidate against the target spectrum requires evidence-driven reasoning that translates spectral mismatches into actionable structural edits. SD-CoT formalizes this evidence-driven mode and serves as the agent's reflection mechanism, invoked at every expansion step of the search.

Diagnostic Report Generation. For a selected candidate, MARS uses a forward spectrum simulator (instantiated with CFM-ID Wang et al. (2021) in our standalone experiments) to obtain a predicted spectrum, and contrasts it with the target spectrum to produce a verbalized diagnostic report. SD-CoT categorizes discrepancies into *MATCHED*, *EXTRA*, and *MISSING* peaks, converting spectral mismatches into structural editing suggestions. This coarse-grained presence/absence categorization is intended to reduce dependence on precise simulated intensities.

Reflective Reasoning. During this process, the diagnostic report, together with the target molecular formula, the SMILES of the selected node, and structural hints aggregated from retrieved references, is passed to the LLM as the input for reflection. The LLM then performs a structured CoT that re-evaluates DoU consistency, decides between local optimization and global reconstruction depending on the severity of the mismatch, and verifies that the proposed edit conforms to the molecular formula. By turning spectral discrepancies into explicit, verbalized correction plans, SD-CoT supplies the search with directly actionable, evidence-grounded guidance, which is the key signal that drives the reflection-driven MCTS.

2.3 REFLECTION-DRIVEN SEARCHING

While SD-CoT supplies single-step reflective edits, the molecular chemical space is combinatorially vast, and a single trajectory of local edits offers no guarantee of reaching a globally consistent structure. To enable structured exploration, we cast molecular structure refinement as a Markov Decision Process (MDP) over the chemical space, formalized as a tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, R), \tag{2}$$

where $\mathcal{S}$ is the state space of candidate structures encoded as SMILES strings; $\mathcal{A}$ is the action space of semantic structural edits sampled by the LLM under SD-CoT guidance; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the deterministic transition that applies an edit a to state s to yield s'; and $R : \mathcal{S} \to \mathbb{R}$ is the reward measuring the consistency between a candidate and the target spectrum.

The search tree is initialized in a warm-start manner: the k candidate structures generated by IP-CoT are attached as first-layer nodes under a virtual root. As a fallback, when all IP-CoT candidates

are invalid or violate heavy-atom constraints, the agent instead promotes the retrieved references themselves as initial nodes, ensuring that the search remains operable under low-quality priors.

Selection. Starting from the root s_0, the algorithm recursively descends the tree by selecting child nodes according to the Upper Confidence Bound for Trees (UCT) criterion:

$$\mathrm{UCT}(s') = Q(s') + c \cdot \sqrt{\frac{\ln(N(s) + \varepsilon)}{N(s')}}, \tag{3}$$

where $Q(s')$ is the value estimate of node s', $N(\cdot)$ is its visit count, and c and ε are exploration hyperparameters. Nodes are selected using UCT to balance exploitation and exploration.

Expansion. Unlike generic LLM search, MARS uses SD-CoT-guided edits, where spectral mismatches directly determine structural modifications. Conditioned on this report, the LLM proposes a chemically meaningful edit, yielding a child node s'. This grounded expansion is the central mechanism by which spectral evidence enters the search loop, distinguishing reflection-driven MCTS from generic LLM tree search. An automatic retry mechanism guarantees 100% chemical validity of all proposed edits, as detailed in Appendix E.

Evaluation. Each generated node is scored by a reward that combines three complementary signals:

$$R(s) = \lambda_1 S_{\text{Spec}}(s) + \lambda_2 S_{\text{Form}}(s) + \lambda_3 S_{\text{FP}}(s), \tag{4}$$

where $S_{\text{Spec}}(s)$ is the cosine similarity between the simulated and target spectra, capturing spectrum-level alignment; $S_{\text{Form}}(s)$ measures the consistency between the candidate and the target molecular formula, with stricter penalties on heavy-atom mismatches (e.g., N, O) to enforce hard chemical constraints; and $S_{\text{FP}}(s)$ is the Tanimoto similarity between the candidate's Morgan fingerprint and the fingerprint produced by the perception module, capturing structural alignment in the perception space. The weighting coefficients $\lambda_1, \lambda_2, \lambda_3$ are empirically set to 0.5, 0.2, and 0.3, respectively, to balance the contributions of each signal. By fusing three signals derived from different sources, the hybrid reward avoids over-reliance on any single estimate and remains informative even when one source is noisy or partially miscalibrated. The resulting score initializes $Q(s)$, with $N(s) = 1$.

Backpropagation. The reward of the new node is propagated upward to update its ancestors via a soft averaging rule:

$$Q(s) = 0.5 \cdot Q(s') + 0.5 \cdot Q(s), \tag{5}$$

applied recursively until the first-layer nodes are reached. This update balances the most recent reflective edit against the accumulated history of the parent path, allowing newly discovered high-reward edits to influence selection without overwriting prior evidence along the trajectory.

The four steps are executed iteratively for N rounds, after which MARS returns the k nodes with the highest reward as the final predictions. Through this loop, each expansion is grounded in spectral evidence via SD-CoT, each evaluation aggregates complementary signals, and each backup propagates this grounding throughout the tree, realizing the reflection-driven structured search that lets MARS converge to chemically and spectrally consistent structures within the vast molecular space.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate MARS on two complementary benchmarks: NPLIB1 Dührkop et al. (2021) for standard performance on diverse natural products, and MolPuzzle Guo et al. (2024) to assess robustness under upstream distribution shift. MassSpecGym Bushuiev et al. (2024) additionally supports a matched-initialization refinement comparison. To prevent data leakage, all evaluation molecules are strictly removed from the external knowledge base prior to retrieval. Further details are provided in Appendix C.1.

Baselines. We compare MARS against three supervised *de novo* generation methods, namely MADGEN Wang et al. (2025), DiffMS Bohde et al. (2025), and MS-BART Han et al. (2025), as well as two LLM-agent baselines sharing MARS's backbone: Chat+RAG, which uses retrieved molecules as in-context examples, and Reflexion Shinn et al. (2023), a verbalized self-reflection framework without structured search. The supervised methods are evaluated on MolPuzzle using weights trained

Table 1: Performance comparison on NPLIB1 and MolPuzzle datasets. All experiments are repeated five times, and the mean results are reported. Results marked with * are reproduced from MS-BART. **Bold** and underline denote the top two results within each row, with the best performance being statistically significant under a t-test at the 95% confidence level.

Backbone	Model	NPLIB1						MolPuzzle					
		Top-1			Top-10			Top-1			Top-10		
		Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
Non-LLM	Retrieval	0.00	8.02	0.46	0.00	**5.18**	**0.58**	0.00	8.00	0.18	0.00	4.38	0.31
	MADGEN*	2.10	20.56	0.22	2.39	12.69	0.27	-	-	-	-	-	-
	DiffMS*	8.34	11.95	0.35	15.44	9.23	0.47	9.52	16.33	0.22	16.40	4.30	0.33
	MS-BART*	7.45	9.66	0.44	10.99	8.31	0.51	3.90	7.39	0.22	5.85	6.43	0.30
Qwen3-14B	Simple CoT	0.00	25.03	0.13	0.12	24.32	0.14	4.88	5.98	0.24	11.71	4.97	0.33
	Thinking	0.24	23.48	0.13	0.37	22.80	0.14	8.78	6.33	0.27	12.68	5.58	0.34
	MARS	2.09	**7.38**	0.46	3.31	6.28	0.52	15.61	4.27	0.33	30.24	2.50	0.52
DeepSeek-V3.2	Simple CoT	1.71	14.92	0.24	2.81	13.11	0.29	15.12	4.01	0.34	30.24	2.49	0.52
	Thinking	5.01	12.71	0.28	9.40	10.94	0.36	21.46	3.10	0.39	60.49	1.41	0.73
	Chat+RAG	2.34	13.85	0.27	3.92	11.78	0.34	15.12	4.47	0.32	29.27	2.99	0.49
	Reflexion	4.13	13.20	0.29	7.65	11.42	0.37	20.98	3.62	0.38	34.15	2.39	0.54
	MARS	**8.91**	7.43	**0.47**	**15.87**	5.64	**0.58**	**25.86**	**3.01**	**0.42**	**66.35**	**1.17**	**0.76**

on MassSpecGym. Since DiffMS does not support halogen-containing molecules, we report its results on a halogen-free subset. We further compare MARS with FRIGID Bohde et al. (2026), which uses fragment-consistency-guided remasking and denoising, and an K-MSE Zhuang et al. (2025), which combines knowledge-enhanced self-reflection with tree search, using FRIGID-base as the initial candidate generator.

Backbones. We adopt DeepSeek-V3.2 Liu et al. (2025) and Qwen3-14B Yang et al. (2025) as open-source backbones. For direct LLM baselines, we test both Chat mode with Simple CoT and Thinking modes. These baselines receive only the formula and peak list, providing a brief reasoning process and k candidate SMILES.

Evaluation Metrics. We adopt three complementary metrics to verify the quality of generated molecules: Top-k Accuracy, Top-k Maximum Tanimoto Similarity, and Top-k Minimum MCES (Maximum Common Edge Subgraph).

3.2 Main Results

As shown in Table 1, MARS achieves the best performance on nearly all metrics across both datasets, validating the reflection-driven tree search paradigm for molecular structure elucidation. On NPLIB1, MARS surpasses all baselines built on the same LLM backbone, leading by 3.90% and 6.47% on Top-1 and Top-10 accuracy, demonstrating the benefit of the complete framework beyond direct use of the same backbone. Despite being training-free, MARS remains competitive against supervised retrieval and generative baselines, attaining 8.91% Top-1 accuracy and surpassing the strongest supervised baseline DiffMS (8.34%), a particularly encouraging result given that supervised methods are trained on large-scale MS data.

To assess robustness under upstream distribution shift, we evaluate MARS on MolPuzzle, where spectra are dominated by $[M]^+$ rather than the $[M+H]^+$ ions used in training. Here LLM-based methods show substantially larger relative gains than supervised baselines, suggesting the latter rely heavily on the training distribution while the former generalize more gracefully. MARS reaches 66.35% Top-10 accuracy, 5.86 points above the strongest LLM-agent baseline and roughly four times the strongest supervised baseline under this shift.

3.3 Structural Recovery and Candidate Refinement on MassSpecGym

We next investigate the more challenging setting of larger molecules by stratifying the MassSpecGym test set by molecular weight (Figure 3(a)). On the 800–1000 Da subset, MARS achieves a Top-10 MCES of 15.54, outperforming DiffMS (19.37) and MS-BART (22.69), as shown in Table 2. Despite this advantage in structural recovery, exact identification remains difficult, with Top-10 accuracy

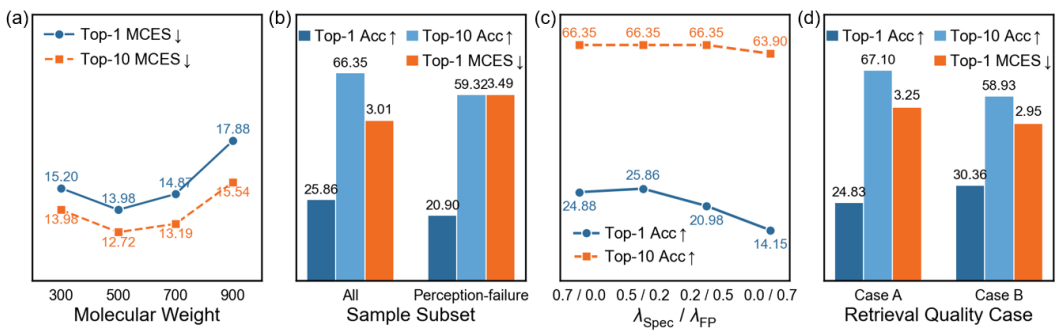

Figure 3: (a) MCES results across different molecular weight ranges. (b) Performance on the perception-failure subset of MolPuzzle. (c) Impact of varying reward weights. (d) Performance across varying retrieval qualities.

Table 2: Performance comparison on the 800–1000 Da subset of MassSpecGym.

	Top-1			Top-10		
Model	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
DiffMS	0.00	38.54	0.24	0.00	19.37	0.31
MS-BART	0.00	24.48	0.31	0.00	22.69	0.38
MARS	**0.00**	**17.88**	**0.37**	**2.00**	**15.54**	**0.58**

Table 3: Candidate refinement performance comparison on MassSpecGym.

Generator	Refinement	Top-1 Acc.	Top-10 Acc.
FRIGID-base	None	16.09	18.19
FRIGID-base	K-MSE	17.81	20.94
FRIGID-base	FRIGID	18.29	22.00
FRIGID-base	MARS	19.81	24.02

reaching only 2.00%. A potential bottleneck lies in constructing sufficiently informative initial candidates: as molecular complexity increases, the expanding space of possible structures makes it harder to establish promising starting points for subsequent refinement. This motivates investigating whether specialized generators can supply stronger structural hypotheses for MARS to refine.

Motivated by this potential limitation, we conduct a separate candidate-refinement evaluation on MassSpecGym using FRIGID-base as the initial generator. We compare MARS with FRIGID's fragment-consistency-guided refinement and K-MSE's knowledge-enhanced self-reflection and tree search. As shown in Table 3, MARS achieves the highest Top-1 and Top-10 accuracy, reaching 19.81% and 24.02%, respectively. These correspond to improvements of 3.72% and 5.83% over the unrefined candidates, and 1.52% and 2.02% over FRIGID refinement. These results demonstrate that MARS can improve strong candidates supplied by a pretrained molecular generator and outperform the other evaluated refinement strategies, supporting the complementary roles of specialized molecular generation and spectral diagnostic reasoning.

3.4 Robustness Analysis

Robustness to perception errors. To examine resilience against uninformative fingerprint predictions, we isolate a perception-failure subset of MolPuzzle where the predicted fingerprint similarity to the ground truth falls below 0.4. As shown in Figure 3(b), MARS still achieves 59.32% Top-10 accuracy, only modestly below its overall 66.35%, with Top-1 metrics remaining stable. This indicates that even when the perception output is unreliable, SD-CoT's spectrum-grounded reflection effectively corrects flawed initial priors to steer the search accurately.

Robustness to reward composition. We sweep the relative weights of S_{Spec} and S_{FP} to test reliance on specific reward components. As shown in Figure 3(c), Top-10 accuracy remains flat at 66.35% even when $\lambda_{\text{FP}} = 0$. Top-1 accuracy shows a milder dependency, degrading only when the spectral term itself is fully suppressed. Thus, MARS avoids over-reliance on auxiliary signals and maintains spectral evidence as the dominant source of guidance, as intended by SD-CoT.

Robustness to retrieval quality. To test whether the framework over-relies on high-similarity references, we stratify the test set by retrieval quality and compare performance between Case A (high-similarity) and Case B (low-similarity) as defined in Section 2.2.1. Figure 3(d) shows that MARS performs comparably on both cases, with Top-1 accuracy even higher for Case B (30.36% vs.

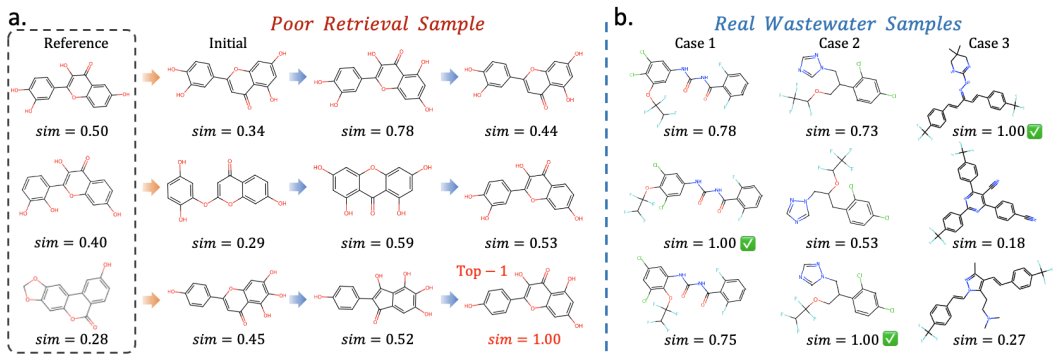

Figure 4: (a) Iterative refinement of candidate structures under challenging retrieval conditions. (b) Training-free structure prediction for unseen molecules in real wastewater samples.

24.83%). This reflects the adaptive design of IP-CoT, which switches to spectrum-driven *de novo* reasoning when retrieval quality is low, preventing inappropriate priors from misleading the search.

3.5 Disentangling Reflection from External Evaluation

To assess the contribution of SD-CoT, we compare different proposal mechanisms during MCTS expansion. Specifically, the **Reward-only** baseline generates SMILES edits relying solely on the target molecular formula, while **No-reflection** omits the SD-CoT diagnosis, prompting the LLM to propose edits directly from the target spectrum and current SMILES. As shown in Table 4, the three settings exhibit a sharp gradient in performance. Despite sharing the same reward and search framework, the Reward-only search exhibits a severe performance gap compared to MARS (22.00% vs 27.00% in Top-1 accuracy). This indicates that even with a spectrum-aware evaluator rescoring candidates, an uninformed proposal mechanism cannot effectively drive the search. Moreover, while allowing the LLM to directly propose edits (No-reflection) improves Top-10 accuracy to 54.00%, it still falls short of MARS. This gap reveals that directly mapping raw spectral discrepancies to structural edits relies on implicit, error-prone reasoning. By disentangling the search process into external evaluation and explicit diagnosis, SD-CoT translates opaque spectral mismatches into localized structural constraints, which significantly restricts the combinatorial action space and drives the search process.

Table 4: Performance impact of disentangling explicit SD-CoT reflection from external evaluation.

	Top-1			Top-10		
Setting	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
Reward-only	22.00	3.52	0.38	41.00	2.00	0.59
No-reflection	24.00	3.28	0.41	54.00	1.54	0.67
MARS	**27.00**	**3.24**	**0.41**	**60.00**	**1.31**	**0.71**

3.6 Error Analysis and Case Study

As illustrated in Figure 4(a), we highlight a challenging sample with poor retrieval (max similarity=0.5). Here, IP-CoT bypasses unreliable scaffolds, switching to the *de novo* branch to generate initial candidates directly from spectrum. SD-CoT then iteratively refines this candidate; node similarities increase from 0.34 across multiple search paths until achieving a perfect Top-1 match. This demonstrates that even with weak external priors, SD-CoT's reflection still effectively steers the search toward chemically consistent solutions.

Furthermore, we categorize prediction failures into four types: **retrieval** (no similar reference), **perception** (inaccurate fingerprint), **evaluation** (reward misranking), and **search** (MCTS budget exhausted). Figure 5(a-b) reveals distinct error distributions: DeepSeek-V3.2 errors concentrate in retrieval (48%) and evaluation (50%), indicating its reasoning fully exploits external signals, shifting the bottleneck to surrounding modules. In contrast, Qwen3-14B's retrieval-dominated errors (64%) reflect constraints in internal reasoning. Ultimately, MARS's bottleneck shifts with LLM capability: stronger backbones expose external module limits, whereas weaker ones are restricted by internal reasoning. To demonstrate MARS's training-free applicability, we evaluate it on real wastewater mass spectra of polyfluorinated compounds Wang et al. (2018), where the target molecules exhibit structural characteristics beyond the model's training distribution. As shown in Figure 4(b), MARS successfully

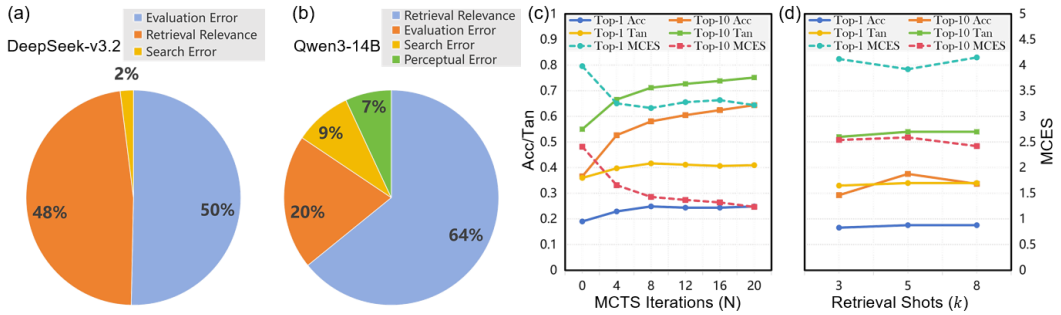

Figure 5: (a-b) Error distributions of MARS with different backbones. (c) Performance impact of MCTS iteration budget N. (d) Performance impact of retrieved reference count k.

recovers the correct structures or highly similar candidates for diverse samples, demonstrating that spectrum-grounded reflection enables robust structure elucidation without additional fine-tuning on real-world data.

3.7 Ablation Study and Sensitivity Analysis

Ablation Study. We evaluate the contribution of each component in MARS on the MolPuzzle dataset. As shown in Table 5, removing any single module degrades performance, highlighting the complementary roles of different components. Among them, MCTS has the largest impact, with Top-10 accuracy decreasing from 66.35% to 36.59% after removal, confirming that without reflection-driven search, the agent collapses into one-shot generation. Removing SD-CoT further reduces Top-10 accuracy to 54.15%, confirming that spectral diagnostic feedback is crucial for guiding effective structural refinement. IP-CoT and RAG provide additional improvements by supplying chemically informed initialization. Removing all components reduces performance close to the LLM baseline, indicating that the overall gains originate from the integrated reflective search framework rather than individual modules alone.

Table 5: Module ablations on MolPuzzle dataset.

Variant	Top-1			Top-10		
	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
w/o RAG	24.88	3.43	0.42	61.95	1.70	0.73
w/o IP-CoT	24.39	3.16	0.36	60.98	1.34	0.73
w/o SD-CoT	21.95	3.40	0.39	54.15	1.70	0.67
w/o MCTS	19.02	3.98	0.36	36.59	2.41	0.55
w/o all	4.88	4.50	0.25	15.12	3.30	0.40
MARS (full)	**25.86**	**3.01**	**0.42**	**66.35**	**1.17**	**0.76**

Sensitivity Analysis. We investigate the effects of the MCTS iteration budget N and the number of retrieved references k (Figure 5(c-d)). Increasing N consistently improves performance before reaching saturation, suggesting that MARS benefits from iterative exploration without requiring excessive search depth. For retrieval size k, performance initially improves and then declines, indicating that overly large reference sets may introduce noisy priors. The sequential inference cost is 372.0s and 20.6k tokens per molecule; with 30-way parallelization, the wall-clock time is reduced to 12.4s, demonstrating the scalability of MARS for high-throughput molecular prediction.

4 Conclusion

We propose **MARS**, an agent framework that recasts molecular structure elucidation from MS spectra. Mirroring the iterative workflow of human experts, MARS unifies chemical priors, spectrum-grounded reflection, and Monte Carlo tree search in a single loop, allowing an LLM agent to hypothesize, reflect, and refine candidate structures under explicit spectral verification. Two complementary CoT mechanisms realize this loop: IP-CoT grounds initial hypotheses in retrieved priors and spectral evidence, while SD-CoT refines them by translating spectral discrepancies into verbalized correction signals. Experiments demonstrate strong performance across two benchmarks and robustness under distribution shift, with real wastewater samples supporting its applicability without task-specific fine-tuning. MARS also improves candidates from existing generators, extending its use to candidate refinement. We hope this work establishes reflection-driven search as a viable paradigm for scientific reasoning over structured evidence and vast solution spaces.

AI USE STATEMENT

The authors used generative AI tools to help refine the manuscript's language, presentation, and structure. All AI-assisted revisions were carefully reviewed and proofread by the authors.

REFERENCES

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

Montgomery Bohde, Hongxuan Liu, Mrunali Manjrekar, Magdalena Lederbauer, Shuiwang Ji, Runzhong Wang, and Connor W Coley. Frigid: Scaling diffusion-based molecular generation from mass spectra at training and inference time. *arXiv preprint arXiv:2604.16648*, 2026.

Roman Bushuiev, Anton Bushuiev, Niek de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

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. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature biotechnology*, 39(4):462–471, 2021.

David Elser, Florian Huber, and Emmanuel Gaquerel. Mass2smiles: deep learning based fast prediction of structures and functional groups directly from high-resolution ms/ms spectra. *bioRxiv*, pp. 2023–07, 2023.

Mikhail Elyashberg and Antony J Williams. *Computer-based structure elucidation from spectral data*, volume 89. Springer, 2015.

Shuo Feng, Meng Huang, Yanbo Li, Aoran Cai, Xiaoyu Yue, Song Wang, Linjiang Chen, Jun Jiang, and Yi Luo. Intelligent understanding of spectra: from structural elucidation to property design. *Chemical Society Reviews*, 54(18):8243–8286, 2025.

Mohamed Amine Ferrag, Norbert Tihanyi, and Merouane Debbah. From llm reasoning to autonomous ai agents: A comprehensive review. *arXiv preprint arXiv:2504.19678*, 2025.

Samuel Goldman, Jeremy Wohlwend, Martin Stražar, Guy Haroush, Ramnik J Xavier, and Connor W Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5(9):965–979, 2023.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688, 2023.

Yang Han, Pengyu Wang, Kai Yu, Xin Chen, and Lu Chen. Ms-bart: Unified modeling of mass spectra and molecules for structure elucidation. *arXiv preprint arXiv:2510.20615*, 2025.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, et al. Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714, 2010.

Florian Huber, Lars Ridder, Stefan Verhoeven, Jurriaan H Spaaks, Faruk Diblen, Simon Rogers, and Justin JJ Van Der Hooft. Spec2vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLoS computational biology*, 17(2):e1008724, 2021.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital discovery*, 2(5):1233–1250, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S Mehta, Gert Wohlgemuth, Dinesh Kumar Barupal, Megan R Showalter, Masanori Arita, et al. Identification of small molecules using accurate mass ms/ms search. *Mass spectrometry reviews*, 37(4):513–532, 2018.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7871–7880, 2020.

Eleni E Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia E Kavraki. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry*, 6(1): 132, 2023.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational drug editing using retrieval and domain feedback. In *The twelfth international conference on learning representations*, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024.

Qinyu Ma, Yuhao Zhou, and Jianfeng Li. Automated retrosynthesis planning of macromolecules using large language models and knowledge graphs. *Macromolecular Rapid Communications*, pp. 2500065, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Benediktus Madika, Aditi Saha, Chaeyul Kang, Batzorig Buyantogtokh, Joshua Agar, Chris M Wolverton, Peter Voorhees, Peter Littlewood, Sergei Kalinin, and Seungbum Hong. Artificial intelligence for materials discovery, development, and optimization. *ACS nano*, 19(30):27116–27158, 2025.

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

Heewoong Noh, Namkyeong Lee, Gyoung S Na, Kibum Kim, and Chanyoung Park. Ir-agent: Expert-inspired llm agents for structure elucidation from infrared spectra. *arXiv preprint arXiv:2508.16112*, 2025.

Xiaohan Qin, Chao Wang, Zhengyang Zhou, Linjiang Chen, Wenjie Du, and Yang Wang. Msanchor: De novo molecular generation from mass spectrometry data with anchor-extended molecular scaffolds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 953–961, 2026.

Nicholas T Runcie, Charlotte M Deane, and Fergus Imrie. Assessing the chemical intelligence of large language models. *Journal of Chemical Information and Modeling*, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Michael A Skinnider, Fei Wang, Daniel Pasin, Russell Greiner, Leonard J Foster, Petur W Dalsgaard, and David S Wishart. A deep generative model enables automated structure elucidation of novel psychoactive substances. *Nature Machine Intelligence*, 3(11):973–984, 2021.

Henry Sprueill, Carl Edwards, Mariefel Olarte, Udishnu Sanyal, Heng Ji, and Sutanay Choudhury. Monte carlo thought search: Large language model querying for complex scientific reasoning in catalyst design. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8348–8365, 2023.

An Su, Xinqiao Wang, Ling Wang, Chengyun Zhang, Yejian Wu, Xinyi Wu, Qingjie Zhao, and Hongliang Duan. Reproducing the invention of a named reaction: zero-shot prediction of unseen chemical reactions. *Physical Chemistry Chemical Physics*, 24(17):10280–10291, 2022.

Remco L van den Broek, Shivam Patel, Gerard JP van Westen, Willem Jespers, and Woody Sherman. In search of beautiful molecules: a perspective on generative modeling for drug design. *Journal of chemical information and modeling*, 65(18):9383–9397, 2025.

Gabriel C Verissimo, Mateus Sa M Serafim, Thales Kronenberger, Rafaela S Ferreira, Kathia M Honorio, and Vinicius G Maltarollo. Designing drugs when there is low data availability: one-shot learning and other approaches to face the issues of a long-term concern. *Expert Opinion on Drug Discovery*, 17(9):929–947, 2022.

Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S Wishart. Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analytical chemistry*, 93(34):11692–11700, 2021.

Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016.

Yi Wang, Nanyang Yu, Xiaobin Zhu, Huiwei Guo, Jianguo Jiang, Xuebing Wang, Wei Shi, Jichun Wu, Hongxia Yu, and Si Wei. Suspect and nontarget screening of per-and polyfluoroalkyl substances in wastewater from a fluorochemical manufacturing park. *Environmental science & technology*, 52 (19):11007–11016, 2018.

Yinkai Wang, Xiaohui Chen, Liping Liu, and Soha Hassoun. Madgen: Mass-spec attends to de novo molecular generation. *arXiv preprint arXiv:2501.01950*, 2025.

Tianxin Wei, Ting-Wei Li, Zhining Liu, Xuying Ning, Ze Yang, Jiaru Zou, Zhichen Zeng, Ruizhong Qiu, Xiao Lin, Dongqi Fu, et al. Agentic reasoning for large language models. *arXiv preprint arXiv:2601.12538*, 2026.

David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

Bo Wen and Mingshe Zhu. Applications of mass spectrometry in drug metabolism: 50 years of progress. *Drug metabolism reviews*, 47(1):71–87, 2015.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hailiang Zhang, Qiong Yang, Ting Xie, Yue Wang, Zhimin Zhang, and Hongmei Lu. Msbert: Embedding tandem mass spectra into chemically rational space by mask learning and contrastive learning. *Analytical Chemistry*, 96(42):16599–16608, 2024.

Zhiling Zheng, Nakul Rampal, Theo Jaffrelot Inizan, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Large language models for reticular chemistry. *Nature Reviews Materials*, 10(5):369–381, 2025.

Xiang Zhuang, Keyan Ding, Tianwen Lyu, Yinuo Jiang, Xiaotong Li, Zhuoyi Xiang, Zeyuan Wang, Ming Qin, Kehua Feng, Jike Wang, et al. Instructbiomol: Advancing biomolecule understanding and design following human instructions. *arXiv preprint arXiv:2410.07919*, 2024.

Xiang Zhuang, Bin Wu, Jiyu Cui, Kehua Feng, Xiaotong Li, Huabin Xing, Keyan Ding, Qiang Zhang, and Huajun Chen. Boosting llm's molecular structure elucidation with knowledge enhanced tree search reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22561–22576, 2025.

A Related Work

Structure Elucidation from Mass Spectra. Currently, mass spectral interpretation mainly relies on two mainstream paradigms: database matching and *de novo* generation. Database matching typically defines mass spectral interpretation as an information retrieval problem. With the continuous expansion of public spectral libraries such as GNPS Wang et al. (2016) and MassBank Horai et al. (2010), this paradigm exhibits high stability in the annotation tasks of known compounds. Some studies (such as Spec2Vec Huber et al. (2021) and MSBERT Zhang et al. (2024)) introduce spectral embeddings to mitigate the impact of experimental noise, or combine simulated fragmentation prediction Wang et al. (2021); Murphy et al. (2023) with molecular fingerprint prediction strategies Goldman et al. (2023) to perform candidate screening in large-scale structural libraries. Nevertheless, the performance of this method is limited when encountering unknown compounds not covered by databases, making it difficult to effectively address problems of the "metabolic dark matter" type.

***De Novo* Generation Methods.** In contrast, *de novo* generation methods generate molecular structures directly from spectra. Madgen Wang et al. (2025) adopts a two-stage strategy, first retrieving potential molecular scaffolds, and then generating complete molecular structures combined with mass spectral information under the constraints of the retrieved scaffolds. DiffMS Bohde et al. (2025) models molecular generation as a conditional discrete diffusion process, progressively constructing molecular graphs under mass spectral embeddings and molecular formula constraints. MS-BART Han et al. (2025) treats spectra and molecular structures as the same modality, training BART Lewis et al. (2020) through a pre-training-fine-tuning-alignment paradigm, achieving end-to-end generation from mass spectra to molecular structures. Recent methods include MBGen for many-body molecular diffusion and FRIGID for forward-model-guided structural refinement. These developments motivate examining how spectral feedback is converted into a structural correction, in addition to measuring generation accuracy. Based on observations of the above research, MARS introduces an explicit reflection-correction mechanism. It does not rely on the model's one-time generation but guides the large model to perform multi-round iterative refinement through semantic feedback converted from spectra, thereby enhancing the stability and reliability of the interpretation process.

LLMs in Molecular Structure Recognition. LLMs have demonstrated immense potential in fields related to molecular structure recognition Jablonka et al. (2023); Zhuang et al. (2024); Zheng et al. (2025). Multiple studies Runcie et al. (2025); Guo et al. (2023) indicate that LLMs possess significant potential in the understanding of chemical languages (such as SMILES Weininger (1988)) and structural information. However, the inherent hallucination problem of LLMs Ji et al. (2023) severely limits their reliability in molecular recognition-related tasks. To mitigate this issue, ChemCrow M. Bran et al. (2024) introduced external tools to assist LLMs in completing downstream tasks; ChatDrug Liu et al. (2024) incorporated domain knowledge into iterative reasoning and feedback loops to adapt to complex molecular editing tasks. Meanwhile, IR-Agent Noh et al. (2025) utilizes external tools to extract local functional groups and retrieve similar structures, thereby completing infrared spectrum interpretation tasks. At the same time, studies have also enhanced the precision of chemical tasks by combining traditional search and planning algorithms. Existing research Ma et al. (2025) proposed a new Multi-Branch Reaction Pathway Search algorithm (MBRPS) to help identify all valid multi-branch reaction pathways, thereby realizing retrosynthetic planning for macromolecules; K-MSE Zhuang et al. (2025) validated the effectiveness of combining an external molecular substructure knowledge base and MCTS in handling molecular structure elucidation tasks. Although the aforementioned works have significantly expanded the application of LLMs in molecular tasks, most methods are limited to the optimization of a single strategy, meaning there exists a clear boundary between external tools and search algorithms. Based on this observation, we introduced external tools CFM-ID and MIST in MARS to improve task precision and reliability, while utilizing MCTS to expand the search of chemical space, thereby achieving efficient processing of MS structure elucidation tasks while ensuring precision.

B Broader Impacts and Limitations

Broader Impacts. MARS provides a training-free framework for MS-based structure elucidation, deployable on top of existing LLMs without task-specific tuning. This lowers the ML entry barrier in analytical workflows, accelerating downstream applications like drug discovery and metabolomics. Furthermore, SD-CoT's verbalized reports enhance interpretability by making the reasoning process

transparent. As a plug-and-play module, MARS seamlessly integrates into current pipelines and scales naturally with future LLM advancements.

Limitations. Despite its strengths, MARS still has several limitations. First, the iterative nature of MCTS-based reflection introduces higher inference cost than one-shot generative baselines, although molecule-level parallelization and standard search optimizations can substantially reduce this overhead. Second, MARS inherits the noise floor of the perception module and forward simulator used in its pipeline; while the modular design allows these components to be replaced by stronger alternatives without altering the framework, the absolute performance of MARS still benefits from improvements in these modules. Third, as a plug-in framework MARS depends on the reasoning capabilities of the underlying LLM, and we expect that fine-tuning LLMs on MS-specific reasoning tasks will further unlock the framework's potential in future work.

C Supplementary Experimental Setup

C.1 Datasets

MolPuzzle is a recently proposed multimodal benchmark designed to assess the logical reasoning capabilities of LLMs in the chemical domain. It contains molecular instances associated with various spectral modalities. In this work, we focus on the subset involving MS data. Since MS spectra in MolPuzzle are provided as image-based plots, we manually annotated and transcribed these spectra into structured text peak lists to enable direct integration with our text-based reasoning and retrieval framework.

NPLIB1 is a widely used benchmark in previous studiesWang et al. (2025); Bohde et al. (2025); Han et al. (2025), originally curated from the GNPS library for training the CANOPUS model. It consists predominantly of natural products with high structural diversity and complexity, making it well-suited for simulating real-world scenarios involving previously unseen compounds.

To ensure a fair and leakage-free evaluation, we strictly removed all molecules appearing in the evaluation datasets from the external knowledge base prior to retrieval. This guarantees that MARS cannot access ground-truth structures during inference, and all predictions are generated in a true open-world setting.

C.2 Evaluation Metrics

Top-k Accuracy. We measure exact structure identification accuracy by converting generated molecules and ground-truth molecules into their corresponding International Chemical Identifier Keys (InChIKeys). Since tandem MS is largely insensitive to 3D stereochemistry, we only compare the first block of the InChIKey, which encodes 2D molecular connectivity. If the 2D InChIKey of any top-k generated candidate matches the reference molecule, the prediction is considered correct.

Top-k Maximum Tanimoto Similarity. This metric quantifies structural similarity based on molecular fingerprints. We use RDKit to calculate Morgan fingerprints with a radius of 2 and a bit length of 2048, and report the maximum Tanimoto similarity between the ground-truth molecule and the top-k predicted candidates.

Top-k Minimum MCES (Maximum Common Edge Subgraph). The MCES metric measures graph edit distance by identifying the largest shared substructure between two molecular graphs. We report the minimum MCES distance between the ground-truth molecule and the top-k predictions, providing a fine-grained evaluation of substructure-level correctness.

D Detailed Results on MassSpecGym

Table 6 reports the performance of MARS across four molecular weight intervals on MassSpecGym. The absolute Top-k accuracy remains low across all intervals, which is primarily attributable to two characteristics of MassSpecGym. First, MassSpecGym adopts a strict scaffold-based train-test split, ensuring that the test set contains structurally novel scaffolds rather than near-duplicates of training molecules. This makes the dataset substantially more challenging than benchmarks such as NPLIB1. Second, the spectra in MassSpecGym are obtained from real-world experimental measurements,

Molecular Weight	Top-1			Top-10		
	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
200–400	0.00	15.20	0.19	2.00	13.98	0.25
400–600	0.00	13.98	0.24	0.00	12.72	0.30
600–800	8.00	14.87	0.36	8.00	13.19	0.61
800–1000	0.00	17.88	0.37	2.00	15.54	0.58
All	2.00	15.48	0.29	3.00	13.86	0.35

Table 6: Performance of MARS on MassSpecGym across different molecular weight ranges. The backbone is DeepSeek-V3.2.

introducing additional noise, calibration variations, and fragmentation patterns that further widen the gap between simulated and observed spectra. Despite these difficulties, MARS maintains stable structural similarity across all weight intervals, with Top-10 Tanimoto reaching 0.61 in the 600-800 Da range and remaining at 0.58 in the most challenging 800-1000 Da subset. This stability indicates that, although exact identification is intrinsically difficult under such evaluation conditions, MARS consistently recovers candidates that are structurally close to the ground truth. Combined with the comparison against supervised baselines reported in the main text, these results confirm that the reflection-driven design of MARS provides robust structural recovery capability even in the most challenging real-world MS scenarios.

E Implementation Details of LLM Edits

To ensure the validity of the structural edits proposed by the LLM, we evaluated the base valid edit rate. On the MolPuzzle dataset using DeepSeek, the initial valid edit rate of the LLM is 93.8%. To handle invalid edits, we implemented an automatic validation and retry mechanism: molecules that cannot be parsed by RDKit are discarded, and a maximum of three retries with gradually increasing temperature ($0.0 \rightarrow 0.3 \rightarrow 0.6$) are applied. This ensures generation diversity and avoids repeated error patterns. After this process, the validity rate of all final generated molecules reaches 100%.

F Detailed Robustness Analysis

This section provides the full numerical results corresponding to the three robustness analyses in Section 3.4. All experiments use DeepSeek-V3.2 as the backbone and are conducted on MolPuzzle.

F.1 Robustness to Perception Module Errors

Subset	Top-1			Top-10		
	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
All	25.86	3.01	0.42	66.35	1.17	0.76
Perception-failure	20.90	3.49	0.37	59.32	1.43	0.71

Table 7: Performance of MARS on the perception-failure subset of MolPuzzle, defined as samples for which the predicted fingerprint similarity to the ground truth is below 0.4.

Table 7 reports the full set of metrics on the perception-failure subset compared with the overall MolPuzzle results. Across all six metrics, the performance gap between the two settings is consistently small. Top-1 accuracy drops by 4.96% and Top-10 accuracy by 7.03%, while structural similarity metrics show only minor changes. Considering that the perception-failure subset contains samples on which the perception module is essentially uninformative, this level of degradation is markedly modest, indicating that MARS does not catastrophically fail when the upstream perception is unreliable. The robustness mainly comes from SD-CoT's reflection step, which compares the simulated spectrum of each candidate against the target spectrum and corrects errors that originate from a misleading initial

fingerprint, allowing the search to recover even when the perception-driven retrieval starts from a poor prior.

F.2 Robustness to Reward Composition

λ_{Spec}	λ_{FP}	Top-1			Top-10		
		Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
0.7	0.0	24.88	2.99	0.41	66.35	1.16	0.76
0.5	**0.2**	**25.86**	**3.01**	**0.42**	**66.35**	**1.17**	**0.76**
0.2	0.5	20.98	3.66	0.37	66.35	1.20	0.76
0.0	0.7	14.15	3.98	0.32	63.90	1.31	0.74

Table 8: Sensitivity of MARS to the relative weights of S_{Spec} and S_{FP} in the reward function. $\lambda_{\text{Form}} = 0.2$ is fixed across all configurations. The default configuration used in all other experiments is highlighted.

Table 8 reports the full sensitivity analysis under four reward configurations sweeping the relative weights of S_{Spec} and S_{FP}. Top-10 accuracy remains essentially flat across configurations, holding at 66.35% even when $\lambda_{\text{FP}} = 0$ and only dropping slightly to 63.90% when $\lambda_{\text{Spec}} = 0$. Top-1 metrics show a clearer asymmetric pattern, with performance peaking at the default setting and degrading more noticeably as λ_{Spec} approaches 0 (Top-1 accuracy 14.15%) than as λ_{FP} approaches 0 (Top-1 accuracy 24.88%). This asymmetry indicates that the spectral matching term is the dominant guiding signal in the reward, while the fingerprint term contributes only marginally to the final ranking. Importantly, removing the fingerprint term entirely does not compromise overall performance, confirming that MARS does not implicitly rely on the perception-derived fingerprint signal in the reward function.

F.3 Robustness to Retrieval Quality

Table 9 reports performance stratified by retrieval quality. The two cases yield broadly comparable performance, but with a notable asymmetry between Top-1 and Top-10. On Top-1 metrics, Case B actually outperforms Case A, with Top-1 accuracy rising from 24.83% to 30.36% and Top-1 MCES improving from 3.25 to 2.95. On Top-10 metrics, Case A holds an advantage, reaching 67.10% Top-10 accuracy compared with 58.93% on Case B. This pattern reflects the complementary behaviors of the two IP-CoT branches. Under high-similarity retrieval, the scaffold-based branch produces a tightly clustered set of candidates around the retrieved scaffold, which improves Top-10 coverage but may bias the very top-ranked candidate when the retrieved scaffold is not perfectly aligned with the target. Under low-similarity retrieval, the *de novo* branch reasons more directly from spectral evidence and produces fewer but more focused candidates, leading to a higher Top-1 hit rate at the cost of reduced Top-10 diversity. Together, these results indicate that MARS is not dependent on retrieval quality in any one-sided manner, and the IP-CoT design effectively leverages both retrieval regimes.

To further demonstrate that the agent can recover from a low-quality but high-confidence retrieval, we analyzed all Top-1 hit samples in MolPuzzle. As shown in Table 10, the average retrieval similarity of initial references is significantly higher than their actual Tanimoto similarity to the ground truth. Yet, MARS progressively corrects this via search and reflection, eventually achieving exact Top-1 matches ($sim = 1.00$). This confirms that the MCTS search space is not rigidly constrained by potentially biased initializations.

G Case Study

To evaluate the effectiveness and stability of MARS in interpreting unknown mass spectra under real-world application scenarios, we further conducted tests on a mass spectrometry dataset of polyfluorinated compounds derived from real wastewater samples from a fluorination industrial park Wang et al. (2018). The data were acquired by a Triple TOF 5600 (SCIEX) in information-dependent acquisition (IDA) mode, operating with a negative electrospray ionization (ESI) source, which represents

Case	Top-1			Top-10		
	Acc↑	MCES↓	Tan↑	Acc↑	MCES↓	Tan↑
A (high)	24.83	3.25	0.42	67.10	1.12	0.76
B (low)	30.36	2.95	0.42	58.93	1.37	0.71

Table 9: Performance of MARS on MolPuzzle under different retrieval-quality regimes. High- and low-similarity retrieval trigger the scaffold-based and *de novo* branches of IP-CoT, respectively.

Search Type	**Retrieval Similarity**	**Tanimoto Similarity**
Top-1	0.66	0.38
Top-5	0.59	0.28
Average	**0.62**	**0.32**

Table 10: Analysis of retrieval quality versus structural similarity for Top-1 hit samples on MolPuzzle.

the complex scenarios closer to practical applications in non-targeted mass spectrometry analysis. Figure 6 illustrates the prediction performance of MARS on three polyfluorinated compounds randomly selected from this dataset. For each sample, we present the ground truth molecular structure along with the Top-3 candidate structures generated by MARS. It can be observed that in Case 1 and Case 2, even if the Top-1 candidate did not match the target, the system provided molecules highly similar to the true molecule in terms of the overall molecular skeleton and the composition of key functional groups, which can assist human experts in mass spectrometry interpretation. In Case 3, although candidate molecules significantly different from the true molecular structure were generated at one point during the search process due to the influence of noise and other factors, MARS still possessed the capability to gradually correct the search direction through reflection, thereby accurately identifying the true molecular structure. The above results indicate that the reflection-based search framework maintains stable performance in real mass spectrometry interpretation tasks, validating the feasibility of MARS in actual application scenarios.

H Computational Cost

Table 11 compares the computational cost of MARS with two supervised generative baselines on MolPuzzle. Sequentially, MARS is substantially slower than supervised baselines, reflecting the iterative nature of MCTS-based reflection. However, since each molecule is processed independently, MARS naturally supports molecule-level parallelization. With 30-way concurrency on standard infrastructure, the wall-clock time drops to 12.4 s per molecule, comparable to DiffMS and well within the budget of typical high-throughput MS analysis pipelines. Further reductions are possible through standard optimizations such as MCTS early stopping or reducing the iteration budget, which we leave to future engineering effort. Regarding the retrieval phase, utilizing FAISS for fast similarity search across our 4M-molecule knowledge base is highly efficient. The average query time is 251 ms on MolPuzzle and 271 ms on NPLIB1, which adds negligible overhead to the overall pipeline.

I Details of Prompts

In this section, we provide detailed information on the main experimental prompts to clarify the input-output specifications of the task:

IP-CoT Prompt: In the input, {`Context`} refers to the retrieved top-k reference molecules, including their molecular formulas, SMILES, and shared structures with the target molecule calculated via fingerprint bits. {`Target Formula`} is the molecular formula of the target molecule. {`Target_Peaks`} represents the spectrum of the target molecule. The output consists of {`Top-K`} SMILES sequences.

SD-CoT Prompt: In the input, {`Target Formula`} is the molecular formula of the target molecule. {`Current_SMILES`} is the SMILES of the current molecule. {`Diagnostic Report`} is the spectral diagnostic report constituted by the target spectrum and the predicted

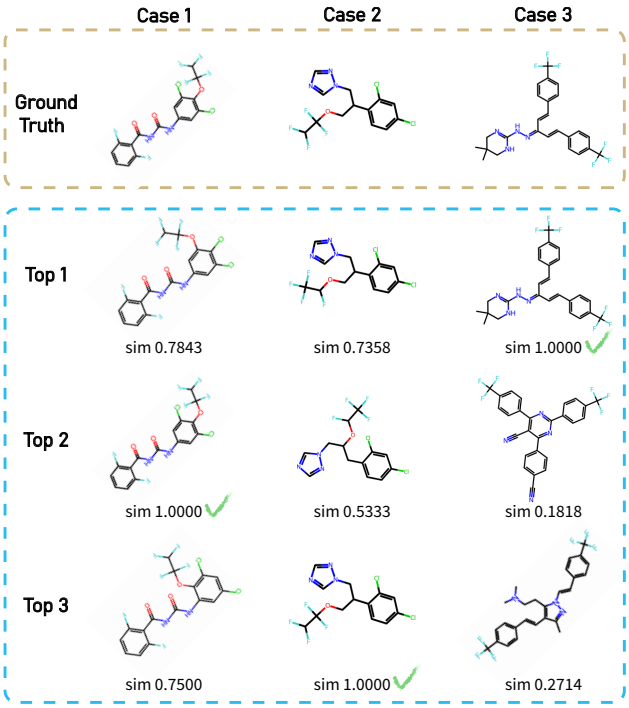

Figure 6: Case study of MARS on a dataset of polyfluorinated compounds derived from real wastewater samples. The first row displays the ground truth structures, while the subsequent rows show the top-3 generated results.

Method	Time/molecule	Tokens/molecule
MS-BART	0.5 s	not required
DiffMS	81.5 s	not required
MARS (sequential)	372.0 s	20.6 k
MARS (parallel)	**12.4 s**	**20.6 k**

Table 11: Computational cost per molecule on MolPuzzle. MARS (parallel) refers to molecule-level parallelization with 30-way concurrency. Token usage is reported only for LLM-based methods.

spectrum of the current molecule (see Section 2.2.2). {`Reference Hints`} refers to the SMILES of reference molecules that possess the target peaks missing in the current molecule. The output is the SMILES of the modified molecule.

Role Definition:

You are an expert in mass spectrometry interpretation and chemical structure reconstruction. Your task is to deduce the most probable target molecular structure based on reference molecules (References) and a target molecular formula (Target Formula), and convert it into a SMILES sequence.

I. Core Task:

The system provides reference molecules retrieved based on spectral similarity. You need to adopt different inference strategies to construct the target molecule depending on the Similarity Score of the reference molecules.

II. Task Input

1. Reference Molecules (References):

Context

2. Target Molecular Formula: Target_Formula

3. Target Mass Spectrum Key Peaks (Target Top Peaks m/z), peaks given in the format m/z (relative intensity %):

Target_Peaks

III. Chain-of-Thought (CoT) Guidelines

Please strictly follow the steps below for your deduction **STEP BY STEP**:

Step 1: Reference Evaluation and Strategy Selection

First, check the Similarity scores of all reference molecules in the Input:

[Case A: High-Confidence References Exist (At least one molecule with Similarity $\geq$ 0.6)]

- 1. Observe all reference molecules (References) with Similarity greater than 0.6 to identify their common structural features (e.g., benzene rings, heterocycles, long chains, specific functional groups).
- 2. Check if the Target Top Peaks (m/z) support the core scaffold of the reference molecules. If strong peaks appear in the target spectrum that cannot be explained by the reference scaffold, or if characteristic fragmentation peaks expected from the reference scaffold are missing, determine this as a "Feature Conflict". As a mass spectrometry expert, you must decisively switch to Case B.
- 3. Combining with Target Top Peaks m/z (fragment peaks), if the m/z suggests specific functional groups not present in the reference molecules (such as carbonyl, hydroxyl, halogen), please prioritize marking them.
- 4. Determine the core scaffold that the target molecule very likely possesses based on the SMILES sequences of the reference molecules.
- **Note:** You must compare the common features of similar molecules with the features corresponding to the Target Top Peaks (m/z) for any contradictions. If there is a conflict, you must decisively switch to Case B!!!! Do not be deceived by the high-similarity of the reference molecules!!!

[Case B: Lack of Credible References (All molecules with Similarity $<$ 0.6)]

- 1. Abandon the scaffold: The value of the reference molecules is extremely low; treat their Shared Substructures only as potential fragment hints (not as a fixed scaffold).
- 2. De Novo Inference: You must rely primarily on the Target Formula and Target Top Peaks (m/z) for inference.
- 3. Basis:
  - Calculate the Degree of Unsaturation (DoU) of the Target Formula.
  - Use Target Top Peaks (m/z) for fragment inference: Analyze the correspondence between strong peaks m/z and common fragment ions (e.g., 91→benzyl, 77→phenyl, 43→acetyl/propyl, 57→butyl, 29→ethyl); Check for reasonable Neutral Loss (e.g., M-18 dehydration, M-15 demethylation).
  - Construct several candidate isomers that fit the molecular formula and can explain the main fragments.
- 4. Skip the gap analysis directly and proceed to "Step 3" for structure reconstruction.

Step 2: Atom Gap and Unsaturation Analysis (Gap Analysis) — Only applicable to [Case A]

- 1. Calculate Atom Difference: Target Formula - Core Scaffold Formula = Remaining Atoms.
- 2. Calculate Unsaturation Difference (DoU Diff): Decrease in H (May indicate formation of double-bonds, rings, or introduction of C=O); Increase in H (May indicate double-bond reduction or ring opening).
- 3. Key Decision: Judge the form in which the remaining atoms exist to match the substitution sites of the scaffold based on Shared Substructures and Target Top Peaks.

Step 3: Structure Reconstruction & Validation

- 1. Execute Construction: If [Case A]: Add deduced side chains or functional groups to reasonable positions on the core scaffold. If [Case B]: Combine deduced functional groups and carbon chains to generate a reasonable molecule that fits the Target Formula.
- 2. Chemical Rationality Check: Ensure the constructed molecule is chemically stable under standard conditions; Prioritize stable functional group combinations common in nature or synthetic chemistry.
- 3. Self-Correction — The most important step: Atom Count (Strictly match the element types and quantities of the Target Formula); Syntax Check (Ensure the output SMILES sequence is syntactically closed and valid); Valence Check.

IV. Output Requirements

1. Please output the Reasoning first following the CoT order. In the Reasoning, explicitly state whether you fall under [Case A] or [Case B] and proceed with the deduction accordingly.

2. Output the final result in JSON format. The result list should contain Top_K candidate SMILES, sorted from high to low compliance, with the sequence best fitting the task requirements ranked first.

Reasoning:

Please strictly follow the provided CoT output reasoning process.

[Your CoT here...]

Final Answer (JSON), strictly as a JSON list of strings:

JSON

```
{
 "SMILES_STRING_1",
 "SMILES_STRING_2",
 ...
}
```

Table 12: Prompts for IP-CoT Setting.

Role Definition:

You are a Mass Spectrometry Forensics Expert & Synthetic Chemist. Your goal is to evolve a molecular structure to perfectly match a Target Spectrum, strictly adhering to the Target Formula.

I. Core Task:

The system provides a diagnostic report based on the target spectrum and the predicted spectrum of the current molecule. You need to adopt different inference strategies based on the discrepancies between the two spectra to modify or reconstruct the current molecule.

II. Task Inputs:

1. Target Formula: Target_Formula (The exact atomic composition you must strictly match)

2. Current Structure: Current_SMILES (The molecule you generated in the previous step)

3. Spectral Discrepancy (Diagnosis):

This section compares the Target Spectrum (Ground-Truth) vs. Predicted Spectrum (Current Structure).

- **MATCHED:** Peaks present in both. The SMILES shown is the fragment structure generated by the current molecule. (Action: Preserve these substructures)
- **EXTRA:** Peaks present ONLY in your prediction. The SMILES shown is the fragment you incorrectly generated. (Action: Modify the molecule to REMOVE this fragment.)
- **MISSING:** Peaks present ONLY in the Target. (Action: You need to ADD a substructure that produces this mass.)

[Detailed Report]:

Diagnostic_Report

4. Reference Knowledge Hints (From RAG):

(Structural clues retrieved from external database based on mass similarity)

Reference_Hints

(CRITICAL: If a reference molecule has a peak we are missing, try to borrow its substructure!)

III. Chain-of-Thought (CoT) Guidelines:

Please strictly follow the steps below for your deduction **STEP BY STEP**:

Step 1: Fundamental Analysis

- 1. Calculate Degree of Unsaturation (DoU) for the Formula.
- 2. Check if the Current Structure matches this DoU. If not, ring/double-bond adjustment is mandatory.

Step 2: Strategy Selection (Case Decision)

- Evaluate the severity of the Spectral Discrepancy:
- *[Case A: Optimization]*: The Base Peak (strongest peak) matches, but minor fragments are missing/extra. $\rightarrow$ Strategy: Perform local modifications (e.g., move a functional group, isomerize a side chain, heteroatom swap).
- *[Case B: Reconstruction]*: The Base Peak is MISSING or the fragmentation pattern implies a wrong core scaffold. $\rightarrow$ Strategy: Abandon the current scaffold. Propose a DIFFERENT isomer (e.g., open/close rings, change from linear to branched, change ring size) that explains the missing major fragments.

Step 3: Fragment Mapping & Execution

- 1. Map the "Missing Peaks" (m/z) to specific substructures (e.g., 91$\rightarrow$Benzyl, 77$\rightarrow$Phenyl, 43$\rightarrow$Acetyl).
- 2. Apply the modification determined in Step 2.
- **FINAL CHECK:** Count atoms. The output SMILES MUST have exactly Formula.

IV. Output Requirements:

1. Reasoning: Explicitly state "Decision: Case A" or "Decision: Case B", then explain your chemical logic.

2. Final Answer: A JSON object containing the SINGLE BEST modified SMILES string.

Reasoning:

Please strictly follow the provided CoT output reasoning process.

[Your CoT here...]

Final Answer (JSON), strictly as a JSON list of strings:

```
JSON
{
 "SMILES": "YOUR_MODIFIED_SMILES"
}
```

Table 13: Prompts for SD-CoT Setting.