COMPILING SCIENCE INTO GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific generation requires outputs to satisfy deterministic physical and structural constraints, yet generative models typically treat such constraints as losses to be learned, which can bias sampling toward validity but **cannot forbid violation**. Published molecular-generation systems report chemical validity of 50%–98%, underscoring that chemical validity remains an **empirical property** rather than a **constraint compiled into the generation process**. To address this issue, we propose **Scientific Compilation** and instantiate this principle in **Mass-ASM**, an executable graph-construction language that **makes illegal structures unrepresentable** as successful outputs under explicitly encoded contracts. Mass-ASM provides a common executable representation for neural navigation and spectrum-based candidate evaluation. Machine audits confirm **100% compilation fidelity** and verify that the compiled system preserves its invariants, rejects adversarially invalid constructions, and exposes boundary cases as structured failures rather than invalid outputs. A random-weight control shows that legality alone does not provide effective navigation: an untrained model reaches 0% valid halt closure. The pretrained generator SpectraForge effectively navigates the compiled space, reaching 25.8% Top-1 on MassSpecGym after deterministic physics-based reranking. In this paradigm, deterministic rules define the boundaries of the feasible space, neural models navigate it, and mass-spectrometry physics selects. Scientific knowledge thus ceases to be merely a criterion for judging outputs but **becomes part of the computational substrate** that defines the space of possible generation.

1 INTRODUCTION

Tandem mass spectrometry molecular elucidation requires inferring molecular structure from a set of peaks defined by mass-to-charge ratios and intensities (Hong et al., 2025). Even when the molecular formula is known, a single elemental composition can correspond to multiple connectivity patterns, and bond cleavage, rearrangement, and neutral loss further render the mapping between structure and spectrum ambiguous (Dührkop et al., 2019). A generative system must therefore distinguish two separate questions: whether a candidate satisfies chemical and physical constraints, and whether a candidate is the true structure of the measured molecule. The former concerns execution legality, whereas the latter concerns identification capability. A legal structure does not necessarily explain the observed spectrum (Stravs et al., 2022).

Existing methods approach this task through fingerprint prediction, candidate generation, and spectral matching. CSI:FingerID (Dührkop et al., 2015) and MIST (Goldman et al., 2023) extract structural fingerprints from spectra, MetFrag (Ruttkies et al., 2016) establishes structure-spectrum relationships through candidate fragmentation and peak matching, and MSNovelist (Stravs et al., 2022), DiffMS (Bohde et al., 2025), and MS-GPT (Zhao et al., 2026) explore *de novo* candidate generation. Closest to our setting, MS-BART (Han et al., 2025), MolForge (Neo et al., 2025), and MS-GPT (Zhao et al., 2026) condition generative decoding on MIST-predicted fingerprints over standard string tokenizations — SELFIES, SMILES, and fragment-based SAFE, respectively — none of which enforces mass, molecular-formula, or DBE closure during decoding. Collectively, these works raise a systemic question: which properties should be guaranteed directly by deterministic rules, and which capabilities should still be learned by models.

Grammar VAE (Kusner et al., 2017), Syntax-Directed VAE (Dai et al., 2018), and SELFIES (Krenn et al., 2020) embed syntactic or chemical validity into generation through grammars, semantic at-

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.

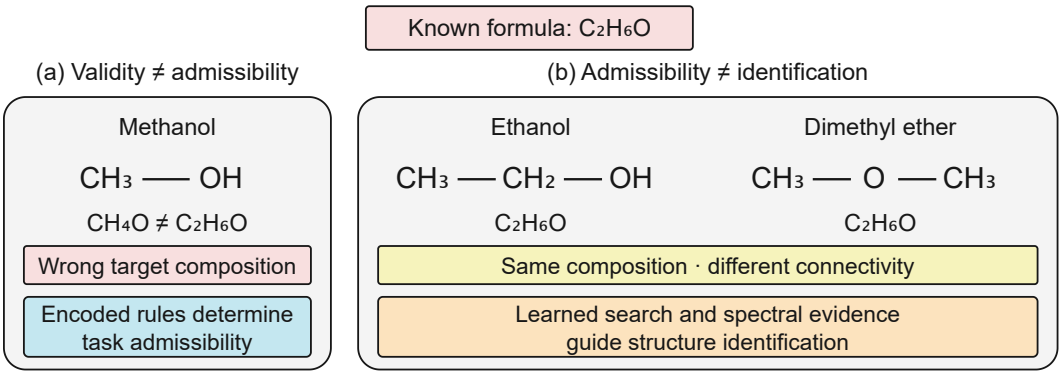


Figure 1: Chemical validity and molecular identification are distinct requirements. For the formula C_2H_6O, methanol is chemically valid but compositionally inconsistent, whereas ethanol and dimethyl ether satisfy the same formula and require spectral evidence to distinguish their connectivity.

tributes, and robust molecular representations, respectively. Tool calling provides access to external capabilities, while constrained decoding enforces specified action restrictions at runtime (Schick et al., 2023; Scholak et al., 2021). We treat molecular generation as autoregressive program construction over a domain-specific instruction set architecture (ISA). The ISA specifies architectural state, instruction semantics, and acceptance conditions. Its deterministic executor enforces state transitions and conservation rules, while the Transformer model selects among admissible instructions. This separates learned instruction selection from rule-governed execution.

We call this design principle Scientific Compilation: scientific rules define the execution semantics, while the model learns to navigate the resulting admissible space. We instantiate this principle for mass spectrometry–based structure elucidation through two core components. Mass-ASM (Mass Assembly Language) defines molecular construction instructions, and its deterministic executor enforces chemical and physical constraints during construction and acceptance. SpectraForge is a Transformer-based molecular language model that generates candidate structures by composing Mass-ASM programs. Beam scores are then combined with MassProof and MetFrag scores to rank these candidates. The resulting system achieves 25.8% Top-1 accuracy on MassSpecGym. By making scientific rules part of the execution semantics of generation, this approach integrates deterministic constraints, learned search, and physical evidence for de novo metabolite structure elucidation.

2 Scientific Compilation and Mass-ASM

2.1 Executable Molecular Representation

For a molecular graph G in its representation domain, Mass-ASM encodes construction as an instruction program. With fixed atom ordering and traversal settings, depth-first compilation produces a deterministic reference program that can be replayed and decompiled.

Each instruction couples a graph operation to execution guards and state updates (Table 1).

For target formula f, the executor maintains

$$s_t = \big(\text{element quota}, \text{valence budget}, \text{topology state}, \text{mass and unsaturation ledger}\big).$$

Element quotas track composition consumption; valence budgets track available bonding capacity. The topology state stores the cursor, branch stack, and open ring labels. Mass and double-bond-equivalent (DBE) ledgers track consistency with the target formula. Supplementary Information D1–D4 specifies the vocabulary, state fields, and hydrogen accounting.

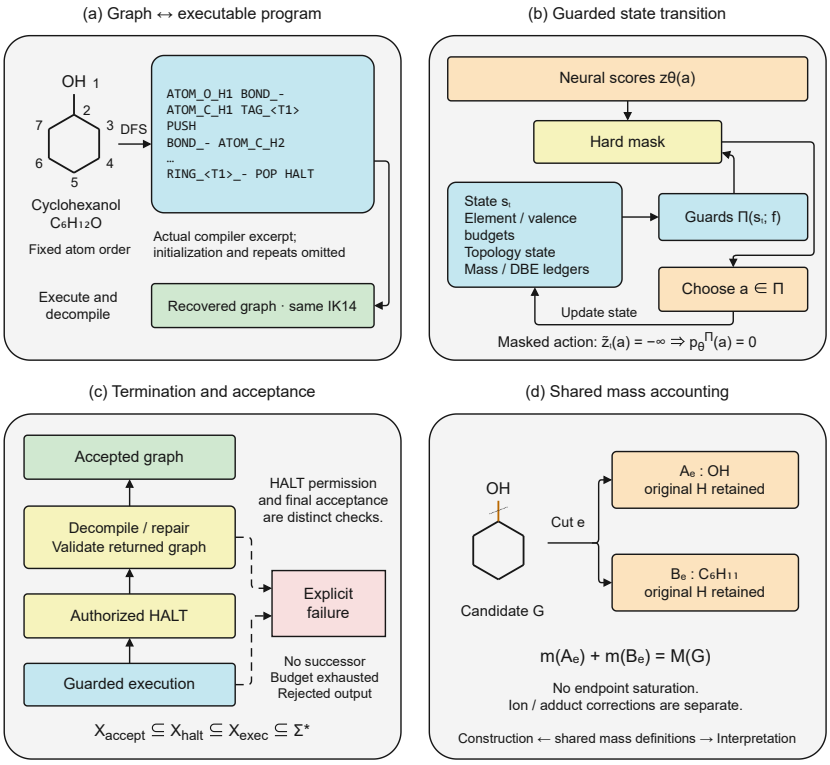


Figure 2: Scientific Compilation turns chemical rules into an executable generation contract. Mass-ASM compiles a molecular graph into a deterministic instruction program, maintains composition, valence, topology, mass, and unsaturation state, and masks inadmissible actions during neural search. Authorized `HALT` states are decompiled and checked for structural acceptance; the same conservation semantics then support spectrum-based physical ranking.

Table 1: Mass-ASM instructions, guards, and state updates

Instruction	Operation	Guards	State update
`ATOM`	Introduce an atom and its attributes	Context, quotas, connectivity	Composition and local state
`BOND`	Set the pending bond type	Context, valence	Pending bond and ledgers
`PUSH/POP`	Save or restore branch context	Stack state	Stack and cursor
`TAG/RING`	Register an anchor or close a ring	Anchor, connectivity, valence	Ring labels and connectivity
`HALT`	End construction	Composition, hydrogen, mass, DBE, closure	Proceed to structural acceptance

2.2 Hard Constraints on Generation

Let Σ denote the Mass-ASM vocabulary and $g_j(s_t, a; f)$ the deterministic guard for rule j. The admissible actions at state s_t are

$$\Pi(s_t; f) = \{a \in \Sigma : \; g_j(s_t, a; f) = 1, \; \forall j\} .$$

The GPU executor maintains a separate state for each sample and search branch and evaluates its guards in parallel. Given neural logits $z_\theta(a)$, it masks inadmissible actions before normalization:

$$\widetilde{z}_t(a) = \begin{cases} z_\theta(a), & a \in \Pi(s_t; f), \\ -\infty, & a \notin \Pi(s_t; f). \end{cases}$$

When $\Pi(s_t; f)$ is nonempty and normalization is well defined, prohibited actions have zero probability. With optional conditioning c, the masked distribution satisfies

$$\operatorname{supp} p_\theta^\Pi(\cdot \mid a_{<t}, f, c) \subseteq \Pi(s_t; f) \subseteq \Sigma.$$

For fixed s_t, f, and execution settings, conditioning changes preferences within this support; selected actions determine the subsequent state. The executor defines which actions are admissible, while the model determines their relative probabilities.

2.3 Execution and Acceptance

For target formula f, initialize the ledgers from f with an empty construction, branch stack, and ring registry. Let $\mathcal{X}_{\mathrm{exec}}(f)$ denote the instruction prefixes admitted by the guards, $\mathcal{X}_{\mathrm{halt}}(f)$ the subset ending in an authorized `HALT`, and $\mathcal{X}_{\mathrm{accept}}(f)$ the subset of halted programs whose decompiled structures pass structural and formula checks. Thus,

$$\mathcal{X}_{\mathrm{accept}}(f) \subseteq \mathcal{X}_{\mathrm{halt}}(f) \subseteq \mathcal{X}_{\mathrm{exec}}(f) \subseteq \Sigma^*.$$

`HALT` requires exhausted heavy-atom quotas, a closed hydrogen ledger, mass and DBE balances within the configured tolerances, and no unfinished branches, ring labels, or pending bonds. Acceptance is evaluated on the reconstructed graph returned by the executor.

With model prefixes and executor states synchronized, retaining only admitted extensions preserves the support restriction at every step. Neural scores therefore cannot promote a prohibited action into an accepted trajectory. If search ends without an accepted candidate, the system reports failure. Supplementary Information D5 specifies initialization and verification conditions.

2.4 Shared Conservation Semantics

Construction and spectral interpretation use the same elemental mass functions and monoisotopic mass table. For a connected candidate graph G, removing a bridge e yields disjoint components A_e and B_e. Retaining the original atoms and their hydrogen assignments gives

$$m(A_e) + m(B_e) = M(G),$$

where $M(G)$ is the candidate's neutral mass and $m(A_e)$ and $m(B_e)$ are the composition masses of its complementary components. Charge, adduct, and hydrogen-transfer corrections are applied when mapping these neutral masses to observed ions (Supplementary Information D6).

Generation enforces target composition; physical ranking tests whether candidate fragmentation and complementary masses explain the spectrum. This shared accounting connects accepted constructions to spectral evidence.

3 Neural Search and Physical Reranking

3.1 Molecular Prior and Fingerprint Conditioning

SpectraForge generates Mass-ASM programs $x = (a_1, \ldots, a_T)$ conditioned on molecular formula f and fingerprint prefix c:

$$p_\theta(x \mid f, c) = \prod_{t=1}^{T} p_\theta(a_t \mid a_{<t}, f, c).$$

Pretraining uses molecular formula f alone. Adaptation first trains the ECFP embedding, then jointly updates it and the query/value LoRA parameters while keeping the base weights frozen. Training structures supply oracle ECFP; at inference, MIST predicts ECFP probabilities from spectrum S.

The 4096-dimensional fingerprint is thresholded, capped at the 256 highest-probability active bits, and sorted by bit index. Dedicated embeddings encode this prefix; the output head predicts Mass-ASM instructions. Each decoding step applies the hard mask from Section 2 to obtain p_θ^Π.

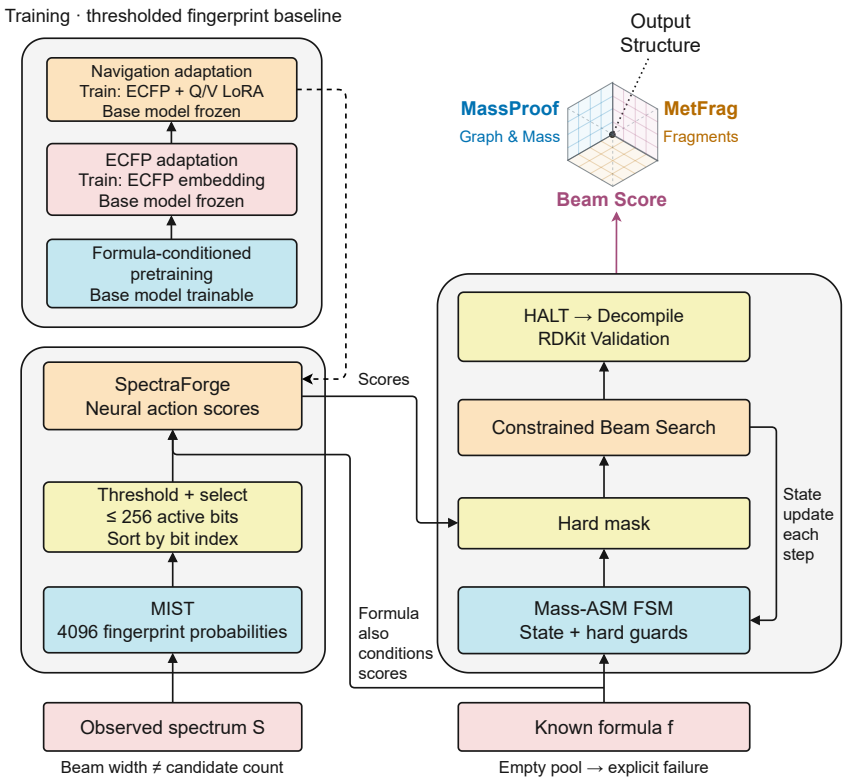


Figure 3: SpectraForge couples fingerprint-conditioned search with physical selection. MIST predicts fingerprint conditions; Mass-ASM constrains each construction step. Accepted structures form a fixed pool ranked by MassProof, MetFrag, and beam scores. The training inset shows the thresholded fingerprint baseline; the oracle branch supplies ground-truth conditions for diagnosis.

3.2 Constrained Search and Candidate Pools

Beam search retains up to K high-scoring prefixes under formula f and fingerprint condition c. It is a budget-limited, prior-guided search: the model assigns priorities, and Mass-ASM governs admissible extensions. Each branch carries its own Mass-ASM state and expands only admissible actions; branch selection keeps the execution state and model context synchronized. The candidate pool is

$$\mathcal{C} = \mathrm{Dedup}_{\mathrm{IK14}}\left(\mathrm{Accept}\left[\mathrm{BeamSearch}_K\left(p_\theta^{\Pi}(\cdot \mid f, c)\right)\right]\right).$$

Accept retains trajectories that terminate with an authorized `HALT`, decompile, and pass structural and formula validation. Deduplication uses the InChIKey connectivity block (IK14), keeping the highest-scoring representative of each identity. If no trajectory passes acceptance, the pool is empty.

3.3 MassProof, MetFrag, and Beam-Score Fusion

Physical reranking combines MassProof and MetFrag with the generator's beam score. MassProof evaluates spectral coverage, complementary fragment masses, and structure-gated fragmentation events. MetFrag scores matches between in silico fragments and observed peaks. The beam score contributes the retained program's generative log-likelihood. On the frozen candidate pools, both physical scores have near-zero mean within-pool Spearman correlation with beam scores: they provide approximately orthogonal evidence to the learned likelihood and complementary evidence to each other.

For a fixed nonempty pool $\mathcal{C}$, S_{MS} is the weighted sum of beam, MassProof, and MetFrag scores after separate min–max normalization within the pool, with weights $1 : 0.5 : 0.5$. The two-channel

Table 2: Preprocessed spectrum-labeled data

Split	Source	Spectra	Molecules
Training	Spectraverse + NIST 20 - {MCES < 10}	1,050,092	52,637
Validation	MassSpecGym	14,995	3,063
Test	MassSpecGym	13,702	2,903

variant combines beam and MassProof scores with weights $1 : 0.5$. The selected structure satisfies

$$G^* \in \arg\max_{G \in \mathcal{C}} S_{\mathrm{MS}}(G \mid S, \mathcal{C}).$$

Pool-wise normalization accounts for the dependence on $\mathcal{C}$. Scores are sorted in descending order; exact ties preserve input order, and the first candidate is returned. Physical evidence therefore changes selection within an accepted pool.

3.4 Coverage and Selection

For a fixed protocol, AnyHit denotes the event that the pool contains the target IK14 identity, and $\mathrm{Top1}$ that the first-ranked candidate matches it under Section 3.3; empty pools count as misses. For $P(\mathrm{AnyHit}) > 0$,

$$P(\mathrm{Top1}) = P(\mathrm{AnyHit})\, P(\mathrm{Top1} \mid \mathrm{AnyHit}) \leq P(\mathrm{AnyHit}).$$

Conditioning and constrained search determine coverage; reranking a fixed pool changes only conditional selection.

4 Experiments

4.1 Data and Experimental Protocol

Pretraining uses 110,223,235 molecular records (approximately 110M SMILES). We exclude matches to validation or test molecules by InChIKey-14, canonical aromatic SMILES, or canonical Kekulé SMILES. Applying the original MassSpecGym recommendation to exclude neighbors with maximum common edge subgraph (MCES) distance below 2 (Bushuiev et al., 2024) was computationally infeasible within our preprocessing budget. The spectrum-labeled training set combines Spectraverse (Gupta et al., 2026) and NIST 20 (Wallace & Moorthy, 2023), excluding molecules with MCES distance below 10 from validation or test references. Both sources are used under their respective licenses. The NIST 20 library is a commercial resource and cannot be publicly redistributed; Validation and test spectra follow the corresponding MassSpecGym folds (Bushuiev et al., 2024).

Structural identity is determined by the InChIKey-14 connectivity block. Multiple spectra of one identity remain separate observations; empty candidate pools count as misses.

Preprocessing retains parseable, single-component, carbon-containing structures that pass RDKit Kekulization and the configured element whitelist; wildcard and disconnected structures are removed. Pretraining additionally requires a C–C bond and molecular weight in $(0, 2000]$ Da. Spectral preprocessing standardizes precursor, adduct, and collision-energy metadata; removes invalid precursor values, peaks within 0.02 Da of the precursor, and peaks below 1% of the base peak; requires at least five peaks; and retains at most 125 peaks. NIST conversion also removes isotope-labeled structures and spectra whose fragment peaks all exceed the precursor.

For interpreter auditing, within-split canonical-SMILES deduplication yields 61,907 structure instances (55,586/3,254/3,067 for training/validation/test). These representation-level counts are distinct from the InChIKey-14 molecule counts in Table 2.

4.2 Mass-ASM Interpreter Audit

Using the 2,048-token coverage-expanded vocabulary and its corresponding interpreter extensions, we audit reference replay, adversarial injection, and dead-state behavior. Every reference trajectory

Table 3: Mass-ASM interpreter audit

Check	Passed / evaluated
Forward compilation	61,907 / 61,907
Structure reconstruction	61,907 / 61,907
Round-trip InChIKey-14 agreement	61,907 / 61,907
Reference-action admissibility	4,657,064 / 4,657,064
Terminal closure	61,907 / 61,907
Adversarial-injection rejection	6,635,476 / 6,635,476

preserves structural identity and reaches terminal closure, with no rejected action or fully masked step (Table 3).

The adversarial suite covers quotas, valence, mass balance, ring anchors, branch state, and execution phase. When masking is bypassed, 3,629 of 360,391 forced continuations still reach an authorized `HALT`, showing why stepwise state enforcement is required. All 1,000 uniform walks over admissible actions enter dead prefixes (mean 4.929 steps). A 16-formula search explores 1,611,317 states, finds no violation of the tested sufficient criterion, and exposes 120 counterexamples to its converse.

4.3 Random-Weight Control

We randomly initialize a 13.15M-parameter SpectraForge model (six layers, hidden dimension 384, six heads), retaining formula and MIST conditioning and the constrained decoder. The control uses 500 deterministically subsampled test spectra representing 432 InChIKey-14 identities, fingerprint threshold 0.5, at most 256 active bits, beam width 8, a 224-token generation limit, length penalty 1.0, and seed 618.

Successful termination, Top-1/5/8 accuracy, and Any-Hit are all zero. Of 3,902 returned trajectories, 3,888 hit the length limit and 14 enter dead prefixes; 498/500 highest-scoring trajectories contain at least 50% `PUSH/POP` actions, with a mean share of 96.92%. The result isolates learned navigation from execution legality.

4.4 Candidate Search, Joint Ranking, and Conditioning

We sweep beam widths 1–256 under Oracle and MIST conditioning. Six arms rank each pool after pool-wise normalization: Beam, Beam + MassProof ($1:0.5$), three-channel fusion ($1:0.5:0.5$), MassProof only, MetFrag only, and MassProof + MetFrag ($1:1$). Beam 256 is the available GPU-memory limit, not an architectural limit.

At beam 256, Any-Hit reaches 47.7% under MIST and 76.7% under Oracle; three-channel Top-1/3/10 reaches 25.8%/34.3%/40.6% and 49.5%/61.5%/70.6%, respectively (Figure 4C–D). From beam 32 to 256, Any-Hit rises by 12.8 and 19.6 percentage points, while MIST Beam Top-1 rises by only 0.7 points and three-channel Top-1/10 by 2.9/6.6 points. The final doubling from 128 to 256 still adds 4.1 and 4.9 points of coverage, respectively.

Physical correction becomes more valuable at wider beams: from beam 32 to 256, three-channel gain over Beam increases from +0.3 to +2.5 percentage points under MIST and from +5.0 to +7.0 under Oracle. At beam 256, MassProof alone, MetFrag alone, and their 1:1 combination reach only 2.4%/2.5%/3.3% Top-1 under MIST and 3.5%/4.9%/6.0% under Oracle. The physical scores provide complementary correction when combined with the model prior. Adding MetFrag to Beam + MassProof contributes another 0.9 and 1.0 points, respectively.

The Oracle–MIST gap is 23.7 percentage points under three-channel ranking (49.5% vs. 25.8%) and 23.6 points under the same two-channel scoring rule (48.5% vs. 24.9%). The gap persists despite training a larger MIST predictor on the prescribed training fold, showing that the quality of the conditioning signal remains a major limitation under the evaluated setting.

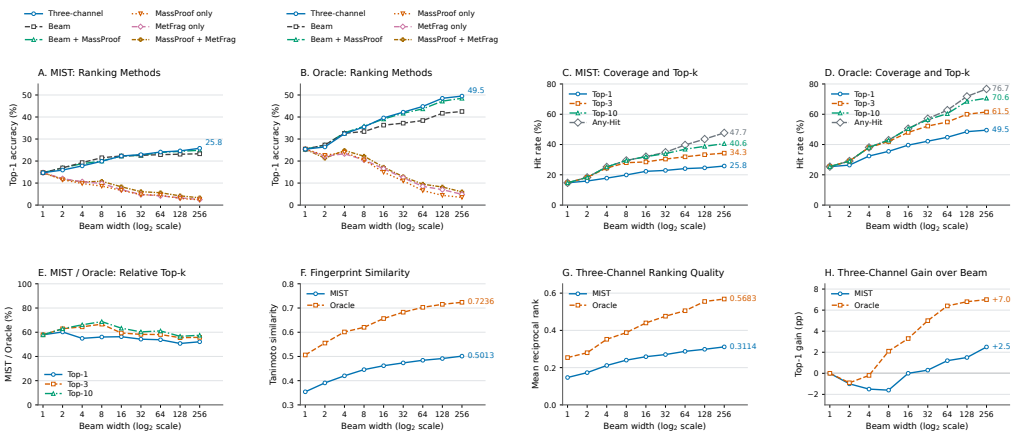


Figure 4: Wider beams expand candidate coverage and increase joint-ranking gains while physical-only ranking deteriorates. A–B compare six Top-1 scoring arms under MIST and Oracle; C–D show pool Any-Hit (the target InChIKey-14 occurs at any rank) together with three-channel Top-1/3/10; reranking a fixed pool leaves Any-Hit unchanged. E plots $100 \times \text{Top-}k_{\text{MIST}} / \text{Top-}k_{\text{Oracle}}$ for the separate evaluation series, F compares reported Tanimoto similarity, G compares three-channel mean reciprocal rank (MRR), and H shows three-channel Top-1 minus Beam Top-1 in percentage points (pp). Pool-normalized weights are Beam:MassProof $= 1 : 0.5$, Beam:MassProof:MetFrag $= 1 : 0.5 : 0.5$, and MassProof:MetFrag $= 1 : 1$; all beam axes use a base-2 logarithmic scale.

Table 4: De novo structure identification on MassSpecGym under the formula-conditioned (bonus) protocol (Bushuiev et al., 2024). Baseline values are as published (Bohde et al., 2025; 2026; Che et al., 2026; Han et al., 2025; Sun et al., 2026; Zhao et al., 2026), cross-checked against the audited MassSpecGym leaderboard (Liu et al., 2026). Notice the last two rows are based on oracle fingerprints rather than deployed systems.

System	**Decoding and conditioning**	**Top-1 (%)**	**Top-10 (%)**
MS-BART (Han et al., 2025)	SELFIES decoding, MIST-predicted fingerprint	1.07	1.11
DiffMS (Bohde et al., 2025)	Graph diffusion, spectrum	2.30	4.25
MBGen (Sun et al., 2026)	Edge diffusion, spectrum and formula	7.58	12.54
MIST + MolForge, corrected (Bohde et al., 2026; Liu et al., 2026; Neo et al., 2025)	SMILES decoding, MIST-predicted fingerprint and formula	10.73	14.48
FRIGID (Bohde et al., 2026)	SAFE-masked diffusion, MIST-predicted fingerprint and formula	18.29	22.00
DualLGD (Che et al., 2026)	Line-graph diffusion, spectrum and formula	23.89	33.08
MS-GPT (Zhao et al., 2026)	SAFE decoding, MIST posterior and formula	23.91	28.65
SpectraForge (this work)	Mass-ASM decoding, MIST-predicted fingerprint and formula	25.8	40.6
MIST + MolForge, ground-truth fingerprint (Neo et al., 2025)	SMILES decoding, ground-truth fingerprint and formula	46.00	59.28
SpectraForge, Oracle conditioning (this work)	Mass-ASM decoding, ground-truth fingerprint and formula	49.5	70.6

4.5 MassProof on CASMI Candidate Pools

We evaluate MassProof on fixed candidate pools from CASMI 2016 (127 positive-ion queries; 184,036 ChemSpider entries). Candidates are grouped by InChIKey-14; ties receive average rank,

Table 5: CASMI 2016 results. Top-k columns report query counts; strict Top-1 also includes percentages.

Method	Strict Top-1	Top-3	Top-10	MRR
MetFrag 2.3 (Ruttkies et al., 2016)	15 (11.81%)	33	61	0.2393
MS-FINDER 1.62	32 (25.20%)	49	58	0.3381
MAGMa	14 (11.02%)	–	–	–
CFM-ID (Machine Learning)	27 (21.26%)	52	78	0.3462
CSI:FingerID (Machine Learning) (Dührkop et al., 2015)	70 (55.12%)	90	100	0.6488
CSI:IOKR (Machine Learning)	53 (41.73%)	78	90	0.5332
MassProof	33 (25.98%)	45	72	0.3472

and strict Top-1 requires a unique winner. The fixed score combines cascade-fragment coverage, precursor-mass gating, and kinetic bridge evidence, with SCScore tie-breaking. Table 5 compares CASMI 2016 results with non-learning and learning-based competition entries (Dührkop et al., 2015; Schymanski et al., 2017): MassProof identifies 33 queries at strict Top-1 and reaches 0.3472 MRR.

5 Conclusion

Scientific Compilation turns scientific rules into executable generation semantics. Mass-ASM instantiates this paradigm by binding representation, state transitions, conservation laws, and acceptance to one execution contract. The abstraction shifts from validity-aware representations and decoding constraints to executable scientific programs whose semantics support both generation and interpretation. The resulting division of labor is precise: execution enforces legality, a learned prior guides budget-limited search, and spectral evidence refines selection. The random-weight control demonstrates why legal actions alone do not provide effective navigation.

The beam-width study shows that wider search increases the value of physical correction when the model prior is retained, while the Oracle–MIST gap makes the conditioning interface a priority for improvement. Better spectrum-to-fingerprint prediction and models that internalize physical evidence offer complementary routes forward without changing the execution contract. Mass-ASM is currently the sole instantiation of Scientific Compilation. It motivates a broader research program: scientific compilation defines the feasible space and reshapes which algorithms can explore it efficiently. Representation, conditioning, and search therefore become joint design choices.

AI use statement

In this work, we used generative AI tools for research ideation and execution: proposing and refining hypotheses, implementing and debugging experimental code, and assisting in interpreting experimental results. We have not used generative AI tools for designing the experimental methodology, formulating mathematical propositions, providing elements of proofs, generating synthetic datasets, cleaning or rearranging datasets, supporting qualitative or thematic data analysis, or translation; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: retrieving and discovering related work, drafting sections of the manuscript, and polishing the writing for readability. We have reviewed all AI-assisted work: AI-generated text was checked and revised by the authors, and reported numbers were verified against machine-audited experimental records. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for more details.

Ethics statement

This work studies computational methods for molecular structure elucidation from tandem mass spectrometry and does not involve human subjects, personal data, or biological experimentation. All data sources are used under their respective licenses: the commercial NIST 20 library is not

redistributed, and all other datasets follow the MassSpecGym protocol. Improved structure elucidation supports metabolomics, forensics, and drug discovery; as with any chemical-structure prediction tool, outputs are hypotheses that require standard analytical validation, and we encourage responsible use consistent with applicable safety and regulatory practices.

Reproducibility statement

Our pipeline is designed to be reproducible end to end. The Mass-ASM interpreter derives every mask, acceptance decision, and failure mode from explicit deterministic rules, so the audit results in Section 4.2 are exact machine-verified counts rather than estimates. The random-weight control fixes the architecture, data subsample, seed, and decoding parameters (Section 4.3), and the beam-width sweeps, conditioning variants, and score-fusion weights are fully specified in Sections 3 and 4.4. The Supplementary Information specifies the vocabulary, state fields, hydrogen accounting, initialization and verification conditions, and shared mass accounting (Sections D1–D6), enabling independent reimplementation of the interpreter and of the evaluation protocol.

References

M. Bohde, M. Manjrekar, R. Wang, S. Ji, and C. W. Coley. DiffMS: diffusion generation of molecules conditioned on mass spectra. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *PMLR*, pp. 4737–4756, 2025.

M. Bohde, H. Liu, M. Manjrekar, M. Lederbauer, S. Ji, R. Wang, and C. W. Coley. FRIGID: scaling diffusion-based molecular generation from mass spectra at training and inference time. *arXiv preprint arXiv:2604.16648*, 2026.

R. Bushuiev, A. Bushuiev, N. F. de Jonge, et al. MassSpecGym: a benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

X. Che, X. Du, and D. Xu. Unlocking high-fidelity molecular generation from mass spectra via dual-stream line graph diffusion. *arXiv preprint arXiv:2605.07048*, 2026.

H. Dai, Y. Tian, B. Dai, S. Skiena, and L. Song. Syntax-directed variational autoencoder for structured data. In *International Conference on Learning Representations*, 2018.

K. Dührkop, H. Shen, M. Meusel, J. Rousu, and S. Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences*, 112(41):12580–12585, 2015. doi: 10.1073/pnas.1509788112.

K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu, and S. Böcker. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods*, 16(4):299–302, 2019. doi: 10.1038/s41592-019-0344-8.

S. Goldman, J. Wohlwend, M. Stražar, G. Haroush, R. J. Xavier, and C. W. Coley. Annotating metabolite mass spectra with domain-inspired chemical formula transformers. *Nature Machine Intelligence*, 5:965–979, 2023. doi: 10.1038/s42256-023-00708-3.

V. Gupta, H. Qiang, H.-H. Chung, E. Herbst, and M. A. Skinnider. Comprehensive curation and harmonization of small-molecule MS/MS libraries in Spectraverse. *Analytical Chemistry*, 98(5): 3934–3943, 2026. doi: 10.1021/acs.analchem.5c06256.

Y. Han, P. Wang, K. Yu, X. Chen, and L. Chen. MS-BART: unified modeling of mass spectra and molecules for structure elucidation. *Advances in Neural Information Processing Systems*, 38, 2025.

Y. Hong, Y. Ye, and H. Tang. Machine learning in small-molecule mass spectrometry. *Annual Review of Analytical Chemistry*, 18(1):193–215, 2025. doi: 10.1146/annurev-anchem-071224-082157.

M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik. Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1:045024, 2020. doi: 10.1088/2632-2153/aba947.

M. J. Kusner, B. Paige, and J. M. Hernández-Lobato. Grammar variational autoencoder. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *PMLR*, pp. 1945–1954, 2017.

H. Liu, R. Bushuiev, I. Lightheart, et al. MassSpecGym in the wild: uncovering and correcting evaluation pitfalls in AI-driven molecule discovery. *arXiv preprint arXiv:2606.19624*, 2026.

N. K. N. Neo, J. Lim, Y. Z. P. Ngoui, X. T. S. Koh, and B. Shen. One small step with fingerprints, one giant leap for de novo molecule generation from mass spectra. *arXiv preprint arXiv:2508.04180*, 2025.

C. Ruttkies, E. L. Schymanski, S. Wolf, J. Hollender, and S. Neumann. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *Journal of Cheminformatics*, 8:3, 2016. doi: 10.1186/s13321-016-0115-9.

T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, L. Zettlemoyer, N. Cancedda, and T. Scialom. Toolformer: language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*, 2023.

T. Scholak, N. Schucher, and D. Bahdanau. PICARD: parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9895–9901, 2021. doi: 10.18653/v1/2021.emnlp-main.779.

E. L. Schymanski, C. Ruttkies, M. Krauss, et al. Critical assessment of small molecule identification 2016: automated methods. *Journal of Cheminformatics*, 9:22, 2017. doi: 10.1186/s13321-017-0207-1.

M. A. Stravs, K. Dührkop, S. Böcker, and N. Zamboni. MSNovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022. doi: 10.1038/s41592-022-01486-3.

X. Sun, W. Wei, J. Rao, J. Xie, and Y. Yang. De novo molecular generation from mass spectra via many-body enhanced diffusion. *arXiv preprint arXiv:2602.01643*, 2026.

W. E. Wallace and A. S. Moorthy. NIST mass spectrometry data center standard reference libraries and software tools: application to seized drug analysis. *Journal of Forensic Sciences*, 68(5):1484–1493, 2023. doi: 10.1111/1556-4029.15284.

X. Zhao, Y. Liu, Z. Li, W. Zheng, F. Zhu, X. Yang, Y. Jin, and Y. Xu. MS-GPT: rethinking MS/MS de novo structure elucidation as spectrum-induced posterior querying of a molecule-language model. *arXiv preprint arXiv:2607.23607*, 2026.