Research Square

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

Mechanism-grounded fragmentation trees enable explicit chemical reasoning for electron ionization mass spectrum prediction

Shuangyin Li

shuangyinli@scnu.edu.cn

South China Normal University

Hongyan Sun

South China Normal University

Shuofan Li

State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum

Quan Shi

State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum https://orcid.org/0000-0002-1363-1237

Jianxun Wu

State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum

Chunming Xu

China University of Petroleum (Beijing)

Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10941499/v1

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

Additional Declarations: There is **NO** Competing Interest.

Mechanism-grounded fragmentation trees enable explicit chemical reasoning for electron ionization mass spectrum prediction

Hongyan Sun[1,†], **Shuofan Li**[2,3,†], **Quan Shi**[2,*], **Jianxun Wu**[2], **Chunming Xu**[2], and **Shuangyin Li**[1,*]

[1]School of Computer Science, South China Normal University, Guangzhou 510631, China

[2]State Key Laboratory of Heavy Oil Processing, Petroleum Molecular Engineering Center (PMEC), China University of Petroleum, Beijing 102249, China

[3]College of Science, China University of Petroleum, Beijing 102249, China

[†]These authors contribute equally: Hongyan Sun, Shuofan Li

[*]Correspondence: Shuangyin Li (shuangyinli@scnu.edu.cn) and Quan Shi (sq@cup.edu.cn)

Abstract

Developing predictive models that connect molecular structures with measured spectra while preserving interpretable links to ion formation and fragmentation chemistry remains a central challenge in predictive spectroscopy. Here we present FragTreeMS, a mechanism-grounded framework that decomposes electron ionization (EI) mass spectrum prediction into fragmentation tree generation and repair followed by pathway-conditioned intensity learning. FragTreeMS organizes source-mechanism-product relations into traceable fragmentation trees and uses candidate-guided self-review to identify and recover plausible omissions before intensity prediction. The repaired fragmentation tree provides the pathway from the molecular root to each retained product ion, and these pathways are directly used for intensity prediction. FragTreeMS improves product ion recovery and complete spectrum similarity relative to direct spectrum prediction and established EI-MS baselines. The inferred ion formation processes and associated putative structural fragments are explicitly represented and visualized, advancing mass spectrum prediction toward interpretable chemical reasoning.

Introduction

Developing predictive models that connect molecular structures with experimentally measured spectra while preserving interpretable links to the physicochemical processes underlying spectrum formation remains a central challenge in predictive spectroscopy. This challenge is especially acute in electron ionization mass spectrometry (EI-MS), in which a molecular ion can undergo successive bond cleavages, hydrogen transfers, rearrangements, neutral losses and secondary dissociations [1, 2]. The resulting peaks are therefore not independent outputs, but observable products of connected ion-formation histories. A predicted spectrum may closely match its experimental counterpart while revealing neither which molecular fragment produces a given ion, which chemical process generates it, nor how it arises from preceding fragmentation events. The methodological challenge is therefore to externalize the chemistry normally compressed within a structure-to-spectrum model as an explicit intermediate representation that is inspectable, correctable and directly usable for quantitative spectrum prediction. Fig. 1a illustrates this limitation: conventional structure-to-spectrum prediction reconstructs the final peak pattern while leaving the underlying ion-formation pathways implicit.

Existing computational approaches address different aspects of this challenge. QCEIMS explores fragmentation trajectories using molecular dynamics and quantum chemical calculations, providing physically grounded insight into spectrum formation at substantial computational cost [3]. CFM-EI models probabilistic transitions between molecular fragments and thereby introduces chemically motivated fragmentation states into EI spectrum prediction [4]. More recent neural methods substantially increase prediction efficiency [5]. NEIMS maps molecular fingerprints directly to binned EI spectra [6], whereas RASSP predicts chemically plausible molecular subformulae or atom subsets before reconstructing spectral intensities [7]. These approaches span physical simulation, probabilistic fragmentation and rapid neural prediction [8]. However, no single trainable framework yet provides an explicit, correctable representation of ion formation and subsequently uses it to predict peak intensities. In many efficient predictors, information about local fragments, precursor-product relations, fragmentation mechanisms and fragmentation history is compressed into a global representation or retained only partially. Consequently, failure to generate an ion cannot be readily distinguished from incorrect assignment of its intensity.

Fragmentation-aware representations provide a route beyond this limitation. Fragmentation trees organize dependencies between precursor and product ions and provide structured explanations of tandem mass spectra [9, 10, 11, 12]. Autoregressive generation of fragmentation trees further shows that neural models can construct molecular fragments through successive cleavage events and represent each spectrum as an ordered fragmentation process instead of an unordered vector [13]. An intermediate representation for interpretable EI-MS prediction must, however, satisfy three requirements. First, it must be chemically explicit: every product ion should remain associated with its molecular-fragment or precursor-ion source and a chemically defined ion-formation process [14]. Second, it must be correctable: omission of an intermediate ion can

suppress an entire downstream branch, so incomplete fragmentation trees must be identifiable and repairable before intensity prediction. Third, it must be quantitatively operative: an inferred pathway should not serve only as a post hoc explanation but should directly condition the predicted abundance of its corresponding ion. Together, these properties separate ion recovery from intensity estimation while reconnecting them through a shared mechanism-resolved representation.

Autoregressive language models provide a suitable computational interface for constructing sequences containing varying numbers of chemically related events [15, 16]. Hybrid neural and symbolic systems also show that learned models can operate through explicit chemical structures rather than only through final predictions [17]. Pretrained molecular language models encode transferable chemical information [18], and general language models can be adapted to structured chemical prediction using supervision tailored to the chemical domain [19, 20, 21]. Related developments in mass spectrometry indicate that representations learned through sequence modeling and attention mechanisms can capture informative molecular and spectral structure [22, 23, 24]. However, methods based on language models alone do not guarantee chemically meaningful reasoning. Inherently interpretable or scientifically constrained representations provide a stronger basis for examining how a prediction is produced [25, 26]. Thus, mechanistic grounding requires an explicit output schema, molecule-specific supervision, and constraints that connect generated relations to valid ions and fragmentation structures. Under such supervision, an autoregressive model can externalize each predicted event as a structured source-mechanism-product relation, connecting the successive events into traceable ion-formation pathways and recovering chemically connected relations removed from otherwise complete fragmentation trees.

Thus, we present **FragTreeMS**, a mechanism-grounded framework with a large language model that reformulates direct structure-to-spectrum prediction as **mechanism-grounded fragmentation tree generation and repair**, followed by **pathway-conditioned intensity learning**. A task-adapted language model generates structured source-mechanism-product relations and organizes them into a directed fragmentation tree, in which each product ion retains its molecular-fragment or precursor-ion origin, assigned mechanism and sequential ancestry. During the generation of fragmentation trees, we utilize the structured ion masking to remove selected ions and their associated relations from the complete trees, and the model can learn to recover the missing chemically connected content.

For tree repair, auxiliary ion predictions identify plausible omissions and guide self-review and repair of the initial tree while preserving its existing relations. This mechanism formulates repair as a learned process that completes the fragmentation tree and does not rely on predefined rules to add unsupported peaks. Meanwhile, the repaired tree also provides the candidate ions and pathway structure used for quantitative spectrum prediction. As shown in Fig. 1b, for each ion retained in the repaired tree, FragTreeMS encodes the ordered pathway connecting the molecular root to that ion and uses an ion-specific representation as a

query over its formation history. Thus, the final fragmentation tree provides an explicit interface between ion generation by a language model and intensity learning, as summarized in Fig. 1.

For pathway-conditioned intensity learning, we design the Path Cross-Attention module to preserve information about molecular fragments, precursor states, mechanism identities and event order during intensity learning. Global molecular context is integrated at the candidate level, after which an Inter-Peak Interaction module jointly contextualizes all ions retained in the repaired tree before predicting intensities. Consequently, the model captures both the dependencies within each pathway that govern ion formation and those among peaks that govern relative abundance. Note that in intensity learning, the fragmentation tree is not generated merely to explain a completed spectrum; it constitutes an operative intermediate representation in the computation that produces the spectrum.

To instantiate this formulation, we curate mechanism-resolved supervision using a controlled vocabulary spanning 13 major EI ion-formation and fragmentation processes and evaluate FragTreeMS in a chemical space comprising four hydrocarbon families. FragTreeMS improves product-ion recovery and spectrum similarity relative to direct structure-to-spectrum models, established EI-MS predictors and direct baselines based on language models. Tree repair recovers omitted product ions and incomplete branches, whereas pathway-conditioned intensity learning enables inferred fragmentation histories to contribute directly to intensity assignment. Each predicted peak can consequently be traced to a putative structural fragment, precursor relation, fragmentation mechanism and root-to-ion pathway through an interactive molecule-level visualization. Source code is available at `https://github.com/Swpy39/FragTreeMS`, and the interactive visualization is available at `https://pec.cup.edu.cn/FragTreeMS`. By making ion-formation processes explicit, repairable and predictive, FragTreeMS shifts EI-MS prediction from reproducing spectral patterns toward modeling the chemical processes that generate them. More broadly, it also provides a strategy for grounding generative models in domain knowledge by requiring them to produce, repair and reuse explicit scientific intermediates rather than only final predictions.

Results

FragTreeMS reformulates EI mass spectrum prediction as explicit fragmentation pathway modeling

Conventional structure to spectrum prediction methods typically map a molecular representation directly to peak intensities. Although such models can reproduce aspects of the observed spectral pattern, they do not explicitly identify the structural origin of each product ion, its precursor ion, the associated fragmentation mechanism, or its formation pathway (Fig. 1a). Consequently, when a prediction is incorrect, the error cannot be readily attributed to the omission of a product ion or to inaccurate intensity estimation for an identified ion.

FragTreeMS separates EI mass spectrum prediction into two sequentially coupled stages. Fragmentation tree generation and repair (Stage 1) uses the molecular structure to generate a mechanism resolved fragmentation tree composed of source, mechanism and product relations. Candidate guided self review and repair then recovers product ions omitted during the initial generation step. In pathway-conditioned intensity learning (Stage 2), the repaired tree is converted into formation pathways from the molecular root to individual product ions. These pathways are integrated with global molecular information and interactions among retained ions to predict relative peak intensities (Fig. 1b).

This formulation enables two experimentally testable assessments: whether the mechanism resolved fragmentation tree recovers the product ions required for subsequent spectrum reconstruction, and whether the formation pathway associated with each ion provides predictive information for intensity estimation beyond ion identity when the product ion set is held constant. The following results address these two aspects in sequence.

Stage 1: Candidate-guided self-review recovers omitted product ions

We first assess whether Stage 1 can generate a product-ion set with high coverage of the annotated ions (Fig. 2a). Across the Qwen3-8B, Llama3.1-8B and Qwen2.5-14B backbones, the initial generation pass recovers a substantial fraction of the annotated product ions, although First Pass Recall varies across molecular classes. Aromatic hydrocarbons show relatively low First Pass Recall, consistent with more complex ion formation patterns and providing a challenging setting for evaluating self-review repair.

Adding candidate ions predicted by the Ion Inventory and Ion Count modules expands the candidate space available for self-review. Recall Repair subsequently incorporates selected candidates with their associated relations into the final fragmentation tree. All three backbones show a positive Repair Gain, with larger improvements observed for molecular classes with lower initial coverage. Strong-ion Recall remains high after repair, indicating that the recovered content includes experimentally intense product ions rather than only

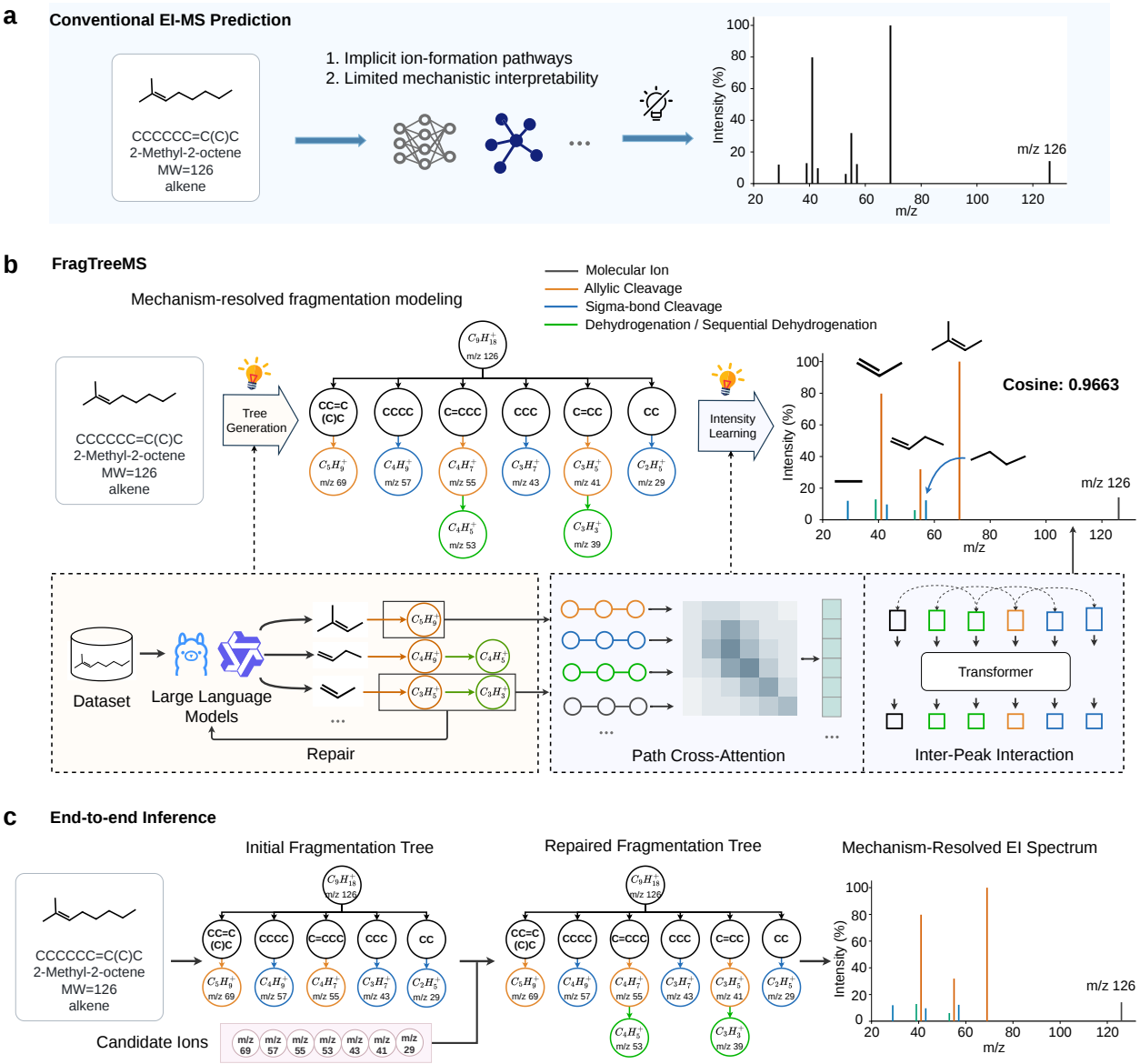


Fig. 1 | FragTreeMS explicitly represents ion formation pathways in EI mass spectrum prediction. a, Conventional structure-to-spectrum prediction maps molecular structure directly to peak intensities while leaving the underlying ion formation processes implicit. **b**, FragTreeMS first generates and repairs a mechanism-resolved fragmentation tree, followed by pathway-conditioned intensity prediction using Path Cross-Attention and Inter-Peak Interaction. **c**, End-to-end inference from initial fragmentation tree generation and candidate-guided repair to reconstruction of a mechanism-resolved EI mass spectrum.

weak or peripheral spectral peaks.

To distinguish the contribution of candidate expansion from that of structured repair, we separately remove ion-aware weighting, similarity regularization and Recall Repair (Fig. 2g). Removing ion-aware weighting or similarity regularization produces only limited changes in Product-ion Recall and Product-ion F2. In contrast, removing Recall Repair causes the largest decrease in both metrics. This result indicates that the performance gain cannot be attributed solely to an expanded candidate set, but is primarily associated

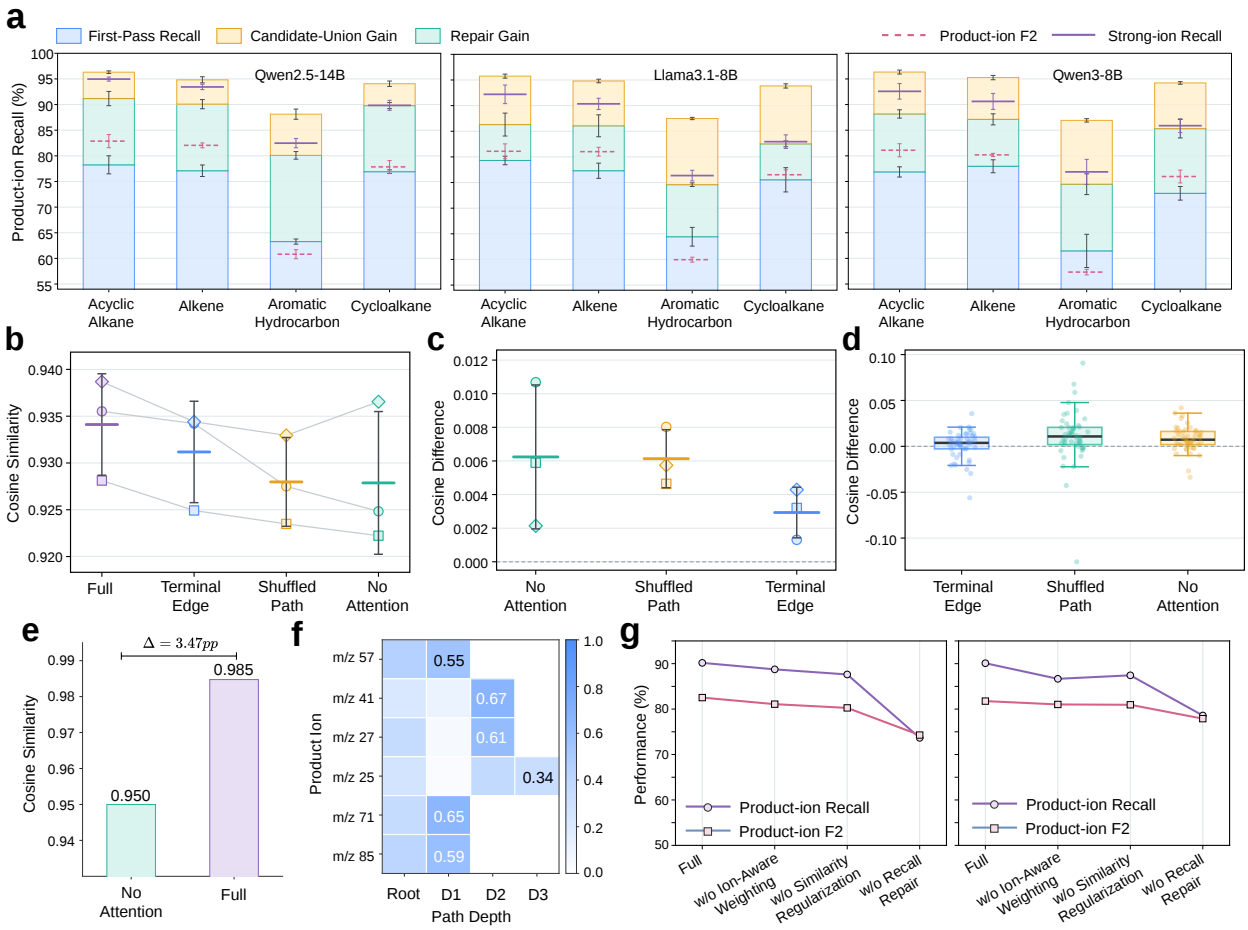


Fig. 2 | Self-review and pathway-conditioned modeling support product ion recovery and spectrum reconstruction. **a**, Product-ion recovery across the three backbone models and four molecular classes, decomposed into First-Pass Recall, Candidate-Union Gain and Repair Gain; Product-ion F2 and Strong-ion Recall are shown as dashed and solid lines, respectively. **b**, Cosine similarity of the Full model and three pathway controls. Terminal-edge retains only the molecular root and terminal ion-formation edge; Shuffled-path assigns complete pathways to different ions within the same molecule; No-attention bypasses Path Cross-Attention and Inter-Peak Interaction. **c**, Mean Cosine differences between the Full model and each control. **d**, Molecule-level paired Cosine differences across 50 molecules. **e**, Full-model and No-attention spectra for a representative molecule. **f**, Path Cross-Attention weights for representative product ions along their formation pathways. **g**, Ablation of ion-aware weighting, similarity regularization and Recall Repair using Product-ion Recall and Product-ion F2. Error bars, where shown, represent the standard deviation across three independently trained runs with different seeds.

with the structured restoration of incomplete fragmentation trees.

Stage 1 therefore does not merely increase the number of predicted ions. It first establishes the main fragmentation structure, then uses candidate-guided self-review to recover chemically connected relations omitted during the initial generation pass. These results support the repairability of the mechanism-resolved fragmentation tree and its role in providing a more complete product-ion set for subsequent intensity prediction.

Stage 2: Fragmentation pathways provide information beyond product ion identity for intensity prediction

High Product-ion Recall alone does not establish that fragmentation pathways contribute to intensity prediction. The model may instead treat the recovered product ions as an unordered candidate set. We therefore construct three pathway controls that preserve product ion identity while selectively reducing pathway information (Fig. 2b-d).

Terminal-edge condition retains only the molecular root and the terminal ion formation edge. This condition reduces the upstream pathway information available to the model. Shuffled-path condition assigns the complete fragmentation pathway of one ion to another ion from the same molecule. This disrupts the correspondence between each ion and its actual formation history. No-attention condition bypasses Path Cross-Attention and Inter-Peak Interaction. This condition evaluates the joint contribution of pathway encoding and interactions among ions.

Full-model achieves the highest Cosine similarity. Terminal-edge causes a relatively small performance decrease, although the magnitude of this decrease varies across molecules. Shuffled-path preserves the ion set, pathway length and overall model structure. It disrupts only the correspondence between each ion and its own formation history. This control produces a more pronounced performance decrease. The result indicates that the model uses more than the m/z value of each product ion. It also uses the structural correspondence between each ion and its specific formation pathway. No-attention causes a further decrease, indicating that both pathway information and interactions among ions contribute to intensity estimation.

In the paired analysis of all molecules, the Cosine differences between Full-model and both Shuffled-path and No-attention are positive for most molecules (Fig. 2d). The performance loss caused by pathway disruption is therefore distributed across multiple molecules rather than concentrated in a small number of extreme cases. For a representative molecule, removing Path Cross-Attention and Inter-Peak Interaction reduces Cosine similarity from 0.985 to 0.950, corresponding to a decrease of 3.47 percentage points (Fig. 2e). The Path Cross-Attention weights further show that different product ions attend to different positions along their formation pathways, including the molecular root and intermediate fragmentation depths (Fig. 2f).

Together, these findings indicate that FragTreeMS relies on two sequential but distinguishable processes. Stage 1 determines which product ions enter downstream prediction. Stage 2 estimates the relative intensity of each ion using its upstream precursor history, fragmentation mechanisms and event order. The fragmentation tree therefore serves not merely as a post hoc explanation. It functions as an operative intermediate representation for spectral intensity prediction.

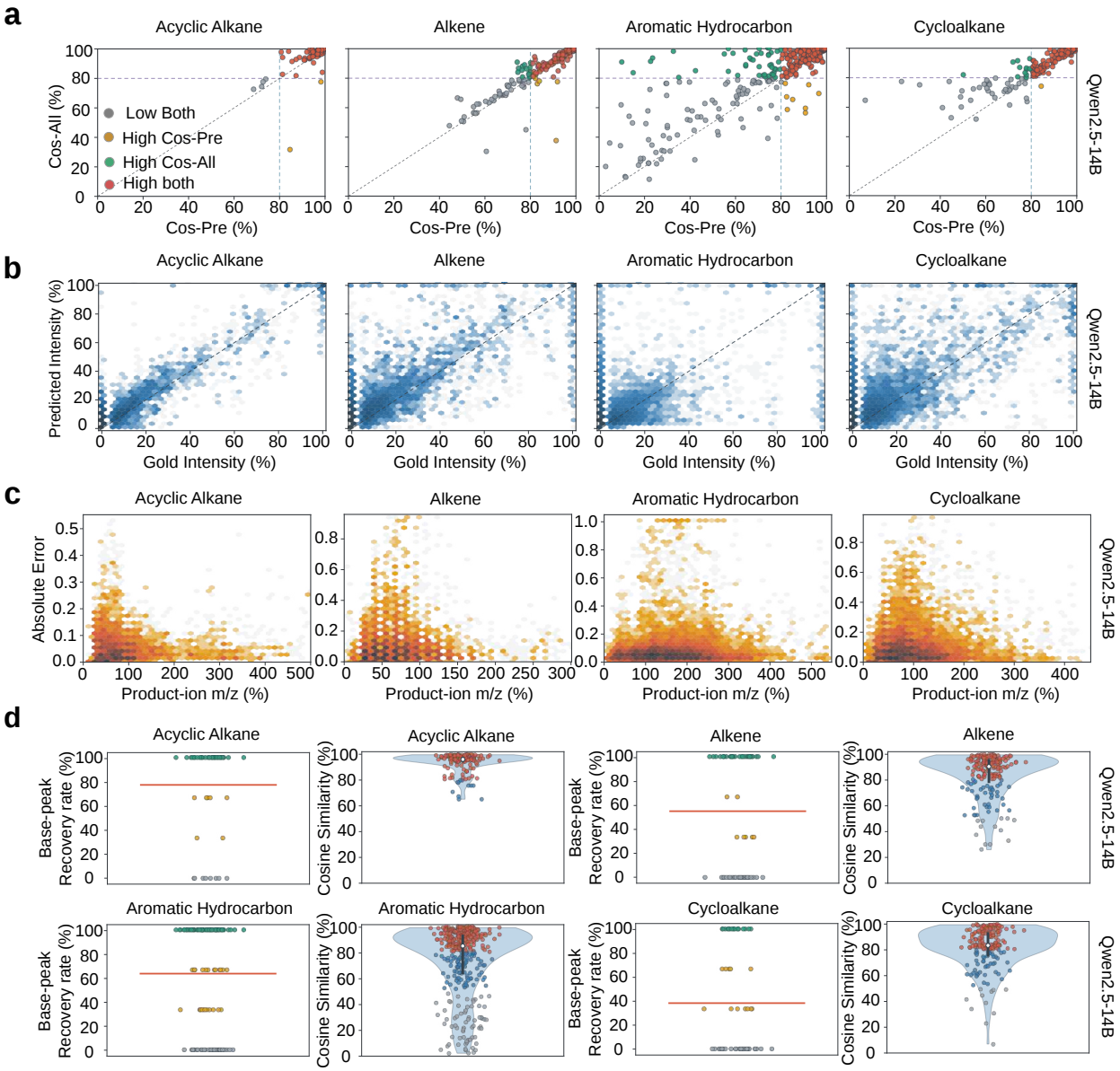


Fig. 3 | Molecule-level spectrum reconstruction and intensity error patterns. All panels show Qwen2.5-14B results across the four hydrocarbon classes. **a**, Molecule-level relationship between Cos-Pre and Cos-All, with points grouped according to low or high values of the two metrics. **b**, Ion-level comparison between predicted and gold intensities. **c**, Absolute intensity error as a function of product-ion m/z. **d**, Molecule-level distributions of Base-peak Accuracy and Cosine similarity for each molecular class.

Molecule level analysis distinguishes ion omission from intensity estimation error

After establishing the contributions of the two stages, we examine the distribution of spectrum reconstruction errors at the molecule and ion levels (Fig. 3). This analysis examines whether the difference between Cos-Pre and Cos-All mainly reflects ion omission, and whether intensity errors are concentrated at particular ion masses or within a small subset of molecules.

Cos-Pre and Cos-All show a positive relationship for most molecules. Molecules with high Cos-Pre but comparatively lower Cos-All generally have accurate intensity estimates for the represented ions but omit some reference ions. Because Cos-All assigns zero intensity to reference ions absent from the prediction, this metric captures the effect of product ion coverage on complete spectrum similarity (Fig. 3a).

Predicted intensities follow the overall trend of the experimental intensities, although the degree of dispersion varies across molecular classes (Fig. 3b). Absolute errors are concentrated mainly in the low to moderate intensity range and show no clear systematic increase with product ion m/z (Fig. 3c). This pattern indicates that intensity estimation errors cannot be attributed simply to increasing ion mass.

The molecular level distributions of Base-peak Accuracy and Cosine similarity further show that model performance is not determined by a small subset of high performing molecules or by a single molecular class. Although the distributions differ across classes, all four classes achieve substantial spectrum reconstruction performance (Fig. 3d). The improvement observed in the aggregate metrics therefore extends across a broad range of molecules.

This molecular and ion level analysis separates two sources of prediction error. Missing product ions directly limit Cos-All, whereas intensity errors for recovered ions reflect the remaining limitations of pathway conditioned intensity learning. The two stage design therefore enables these error sources to be evaluated separately.

FragTreeMS improves overall spectrum reconstruction across prediction settings

After establishing that Stage 1 improves product ion coverage and that Stage 2 uses fragmentation pathway information, we next evaluate whether the complete FragTreeMS framework improves overall spectrum reconstruction (Fig. 4).

Across the Qwen3-8B, Llama3.1-8B and Qwen2.5-14B backbone models, FragTreeMS achieves high Base-peak Accuracy and Cosine similarity (Fig. 4a). Qwen2.5-14B achieves a Cosine similarity of 83.7%, compared with 81.3% for NEIMS. Llama3.1-8B achieves a Base-peak Accuracy of 58.2%, compared with 55.2% for NEIMS. The three backbone models show similar overall performance, indicating that the improvement does not depend on a particular language model.

NEIMS is retrained using the same fixed data split as FragTreeMS and therefore provides a reference under matched training conditions. CFM-ID is evaluated using publicly released parameters and serves as an external reference under different training conditions. This distinction provides a more appropriate interpretation of the comparisons between FragTreeMS and the reference methods.

We next compare the complete FragTreeMS framework with the Base and Triplets language model settings (Fig. 4a,b). The Base setting directly predicts the EI mass spectrum from molecular information. The Triplets setting first generates source, mechanism and product triplets and then predicts intensities for the product

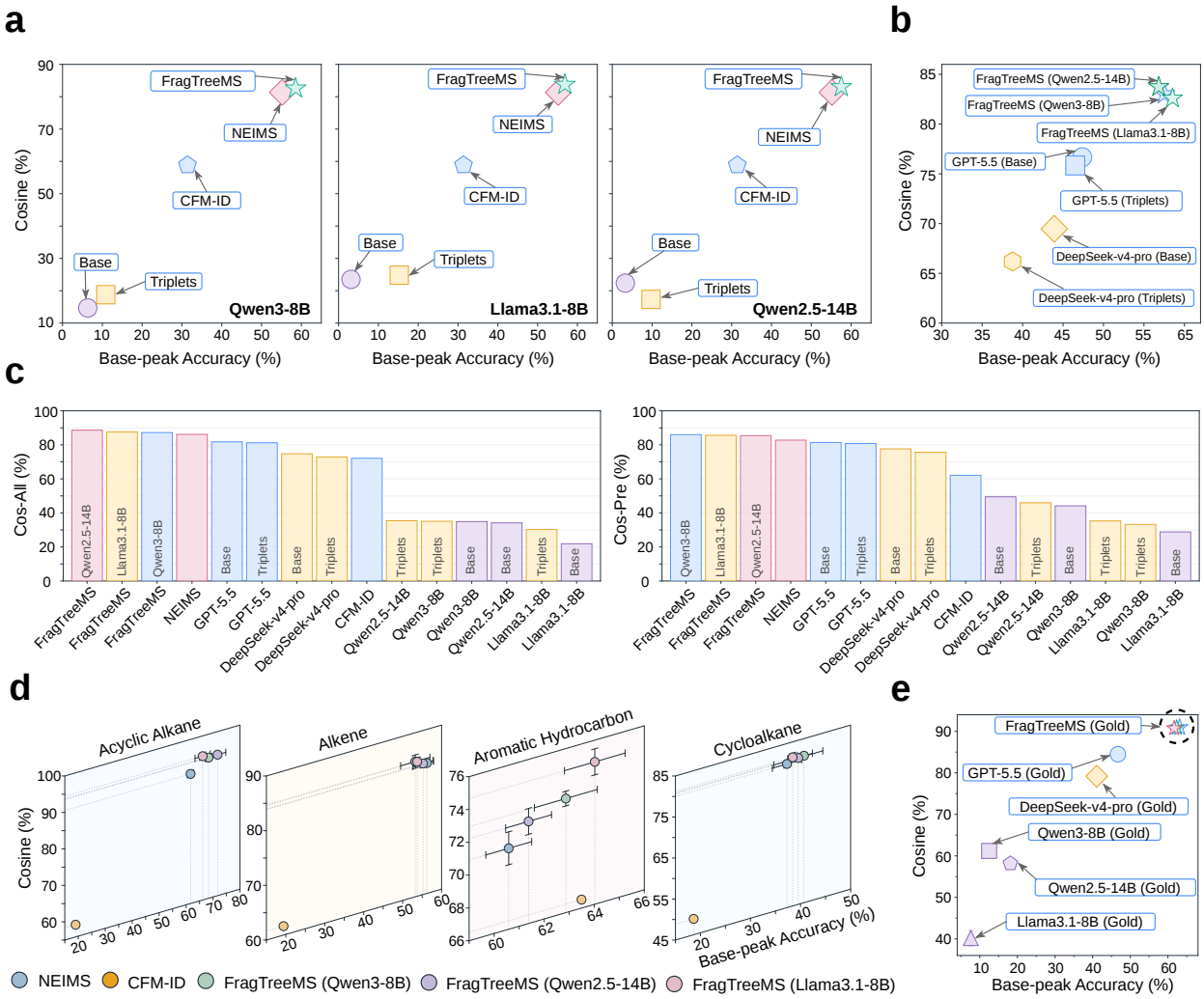


Fig. 4 | End-to-end spectrum reconstruction across prediction settings and reference methods. **a**, Overall Base-peak Accuracy and Cosine similarity for FragTreeMS and the corresponding Base and Triplets controls using Qwen3-8B, Llama3.1-8B and Qwen2.5-14B backbones; NEIMS and CFM-ID are included as reference methods. **b**, Overall Base-peak Accuracy and Cosine similarity for the three FragTreeMS configurations and the Base and Triplets controls based on GPT-5.5 and DeepSeek-v4-pro. **c**, Ranked comparison of Cos-All and Cos-Pre across prediction settings, with methods ordered independently according to the corresponding similarity score. **d**, Base-peak Accuracy and Cosine similarity for FragTreeMS, NEIMS and CFM-ID across the four molecular classes. **e**, e, Overall comparison under the Gold condition using annotated triplets comprising source, mechanism and product information as input to Stage 2 to isolate intensity prediction from errors in Stage 1 ion recovery.

ions contained in the generated triplets. Both settings perform below the complete FragTreeMS framework. The best FragTreeMS result achieves a Cosine similarity approximately 7.1% higher than GPT-5.5 under the Base setting and approximately 14.3% higher than DeepSeek-V4-Pro under the Base setting.

These comparisons indicate that the improvement does not result solely from separating triplet generation from intensity prediction. The complete performance gain requires candidate guided tree repair, ion specific formation pathways and pathway conditioned intensity learning.

The Cos-Pre and Cos-All rankings further distinguish intensity estimation from product ion coverage

(Fig. 4c). The three FragTreeMS backbone models occupy the highest positions for both metrics. Cos-Pre measures intensity agreement over the product ions represented by the model. Cos-All additionally penalizes reference ions that are not recovered. FragTreeMS therefore improves both intensity assignment for recovered product ions and reconstruction of the complete reference spectrum.

FragTreeMS achieves higher Cosine similarity than NEIMS across all four molecular classes, with improvements ranging from approximately 0.9 to 3.9% (Fig. 4d). Aromatic hydrocarbons show more substantial ion omission during the initial Stage 1 generation pass, yet their final Cosine similarity still improves by approximately 3.7%. The consistent class specific improvement indicates that the overall gain is not driven by a single molecular class.

Finally, the Gold condition provides the annotated source, mechanism and product triplets directly to Stage 2 (Fig. 4e). This condition is not a complete end to end evaluation. It is a diagnostic setting that isolates the intensity prediction capability of Stage 2 from errors in Stage 1 ion recovery. The stable results across backbone models and molecular classes under the Gold condition support the independent contribution of the pathway conditioned intensity module.

Overall, the performance advantage of the complete FragTreeMS framework reflects the combined contributions of candidate guided ion recovery, ion specific fragmentation pathways and pathway conditioned intensity learning. It cannot be attributed solely to candidate expansion or backbone selection.

FragTreeMS remains robust across test molecule similarity to the training set and molecular complexity

Finally, we examine the behavior of FragTreeMS across differences between test molecules and the training set, together with variation in molecular complexity (Fig. 5).

For each test molecule, we calculate its ECFP4 Tanimoto similarity to the nearest training molecule. We then compare this structural similarity with fragment triplet Jaccard similarity and ion triplet Jaccard similarity (Fig. 5a,b). Higher structural similarity between a test molecule and its nearest training molecule does not necessarily correspond to greater fragment triplet or ion triplet overlap. This result indicates that molecular structural similarity is not a sufficient proxy for mechanistic similarity. The test set therefore contains molecules with diverse relationships to the training set at both the structural and fragmentation levels.

We next stratify molecules according to the BertzCT complexity descriptor. For each molecular class, the thresholds separating the low, medium and high complexity groups are calculated from the 33rd and 67th percentiles of the corresponding training subset only. Test molecules are then assigned to these training derived intervals. This procedure prevents information from the test set from being used to define the complexity groups (Fig. 5c-e).

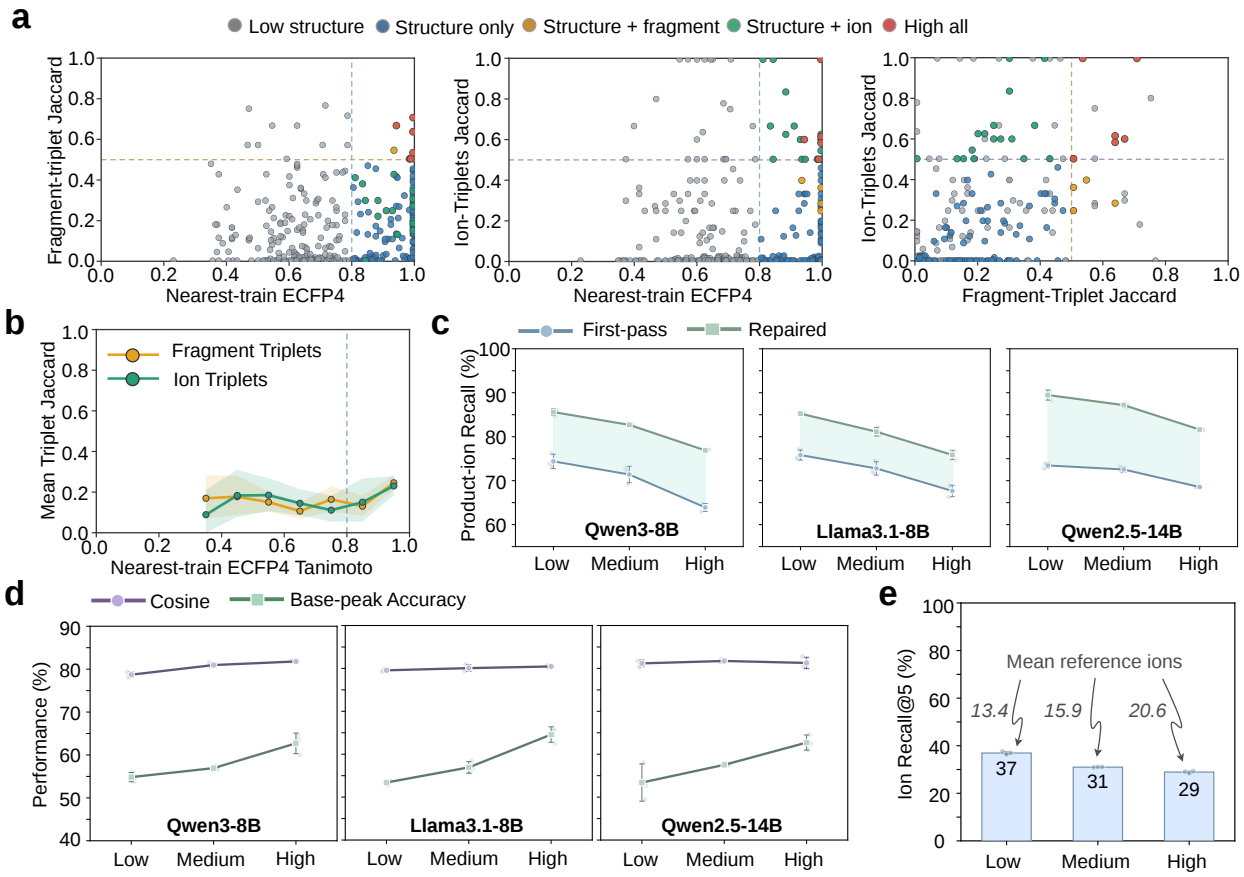


Fig. 5 | Chemical space coverage, robustness to molecular complexity. **a**, Similarity of each test molecule to its nearest training molecule, measured by ECFP4 Tanimoto similarity, fragment triplet Jaccard similarity and ion triplet Jaccard similarity. Colours indicate the four similarity categories. **b**, Mean fragment triplet Jaccard similarity and ion triplet Jaccard similarity for test molecules grouped according to their ECFP4 Tanimoto similarity to the nearest training molecule. **c**, Product-ion Recall before Recall Repair (First-pass), after Recall Repair (Repaired) across low, medium, high BertzCT complexity groups for three backbone models. **d**, Cosine similarity and Base-peak Accuracy across the same complexity groups. **e**, Ion Recall@5 across complexity groups. Values above the bars indicate the mean filtered reference ion count. Low, medium, high complexity groups use class specific training set BertzCT tertiles. Error bars, where shown, represent the standard deviation across different seeds.

As BertzCT complexity increases, Product-ion Recall during the First Pass decreases. This trend is consistent with more challenging ion formation patterns in structurally complex molecules. Recall Repair improves Product-ion Recall across all three complexity groups and all three backbone models. The improvement remains evident in the high complexity group (Fig. 5c). These results indicate that candidate guided self-review can compensate for part of the increased difficulty encountered by Stage 1.

The final spectrum level metrics remain more stable than the initial ion recovery performance. Cosine similarity shows no clear decrease with increasing complexity, and Base-peak Accuracy does not exhibit systematic deterioration (Fig. 5d). Ion Recall@5 shows a more pronounced decrease. This decrease coincides with an increase in the mean number of filtered reference ions from 13.4 in the low complexity group to 15.9 in the medium complexity group and 20.6 in the high complexity group. Because Ion Recall@5 can recover

no more than five reference masses, the metric becomes increasingly stringent as the complete reference ion set expands. Its decrease should therefore not be interpreted simply as a proportional loss in overall spectrum reconstruction performance (Fig. 5e).

Together, these results show that the test set covers diverse structural and fragmentation relationships relative to the training set. They also show that FragTreeMS maintains stable final spectrum reconstruction as molecular complexity increases. The framework absorbs part of the increased difficulty at the fragmentation tree stage through candidate guided self-review, then converts the repaired structured pathways into relatively stable intensity predictions in Stage 2.

Discussion

The central significance of FragTreeMS lies in reformulating EI mass spectrum prediction as structured inference over ion formation events rather than direct mapping from molecular structure to peak intensities. The framework connects ion recovery, fragmentation mechanisms and intensity estimation through an explicit fragmentation tree (Fig. 1). Its scientific significance is reflected in three related aspects.

First, FragTreeMS separates product ion recovery from intensity estimation. In direct structure to spectrum prediction, missing ions and inaccurate intensity assignments are combined in the final spectral error. FragTreeMS first determines the product ions and their structural relations, then estimates their intensities according to the corresponding formation pathways (Fig. 2a,g). This separation distinguishes errors caused by unrecovered ions from errors caused by inaccurate intensity estimation. Cos-Pre therefore reflects the agreement for represented ions, whereas Cos-All also accounts for reference ions that are not recovered. Molecule level and ion level analyses support this distinction by showing that differences between Cos-Pre and Cos-All mainly reflect incomplete product ion coverage, whereas intensity errors are not systematically associated with higher product ion mass (Fig. 3a-c). The Gold condition further indicates that Stage 2 retains predictive capacity when annotated source, mechanism and product triplets are provided, thereby isolating intensity prediction from errors in Stage 1 ion recovery (Fig. 4e). This separation improves error attribution and provides independent targets for optimizing ion recovery and intensity prediction.

Second, FragTreeMS represents EI ion formation as an explicit, inspectable and repairable chemical intermediate. Hierarchical relations between precursor and product ions have long provided structured interpretations of mass spectra [9, 10, 12]. Recent autoregressive fragmentation models further show that neural networks can construct structured intermediate states through sequential fragmentation events [13]. FragTreeMS extends these ideas by linking each product ion to its structural origin, precursor ion, fragmentation mechanism, event order and complete path from the molecular root (Fig. 2b,c). The fragmentation tree is generated before spectrum reconstruction and can be completed through Recall Repair with candidate guidance from Ion Inventory and Ion Count (Fig. 2a,g). Incompleteness in the initial generation is consequently treated as a recoverable state. The resulting representation can be inspected at both ion and pathway levels, which makes the predicted chemical process accessible for analysis and verification.

Third, FragTreeMS makes complete fragmentation pathways quantitatively useful for intensity prediction. Stage 2 encodes the upstream formation history of each ion and integrates this information with global molecular features and interactions among ions. The pathway controls show that retaining the product ion set and pathway length is insufficient when the correspondence between an ion and its own formation history is disrupted. Shuffled path therefore produces a clear performance decrease, whereas bypassing Path Cross Attention and Inter Peak Interaction causes a further decrease (Fig. 2b-d). The molecule level paired analysis shows that these effects are distributed across multiple molecules rather than being

driven by a small number of difficult cases (Fig. 3c,d). Attention analysis further indicates that different product ions use different positions along their formation pathways (Fig. 2e,f). The fragmentation tree therefore functions as a predictive intermediate rather than an explanation added after prediction. It connects local fragmentation events with global spectrum reconstruction and provides precursor history, mechanism information, event order and upstream structural context for relative intensity estimation.

Taken together, these three contributions clarify why FragTreeMS differs from methods that merely append chemical information to a spectrum predictor. NEIMS primarily learns a mapping from molecular representations to spectra [6]. RASSP introduces structural constraints through molecular subformulae or atom subsets [7]. QCEIMS and CFM ID incorporate physically calculated or probabilistic fragmentation processes [3, 4]. FragTreeMS combines generative modeling with a unified mechanism vocabulary and organizes local source, mechanism and product relations into complete pathways that can be repaired and reused for quantitative prediction. Comparisons with the Base and Triplets settings show that separating triplet generation from intensity prediction alone is insufficient (Fig. 4a,b). The full benefit depends on candidate guided tree repair, ion specific formation pathways and intensity prediction conditioned on these pathways.

The consistency of FragTreeMS across backbone models and molecular classes indicates that its performance is primarily associated with the task decomposition and structured intermediate representation rather than with a particular language model (Fig. 4a,d). The chemical space analysis further shows that structural similarity between a test molecule and its nearest training molecule does not fully determine fragment triplet or ion triplet overlap (Fig. 5a,b). Structural similarity is therefore an incomplete proxy for mechanistic similarity. Within the evaluated chemical space, increasing molecular complexity makes initial product ion recovery more difficult, whereas Recall Repair mitigates part of this difficulty and final spectrum level metrics remain comparatively stable (Fig. 5c-e). These findings support the use of explicit and repairable intermediates for maintaining predictive stability across varied molecular settings.

The present study focuses on a representative vertical-domain setting for EI-MS, with molecular classes covering characteristic carbon skeletons commonly encountered in natural organic mixtures and a 13-category mechanism vocabulary spanning major EI ion-formation and fragmentation processes. Because FragTreeMS is built around a domain-specific mechanism system, a structured intermediate representation, repairable generation and downstream quantitative reuse, the same framework can be adapted to other specialized scientific settings by constructing corresponding mechanism-resolved resources for domain-specific expert models.

More broadly, FragTreeMS illustrates a scientific AI paradigm that differs from purely end-to-end mapping. Many scientific observations are not direct outputs of molecules, materials or biological systems, but arise from structured sequences of physical, chemical or biological processes. When these processes are compressed only

into latent representations, even accurate predictions provide limited information about whether the model relies on scientifically meaningful processes. FragTreeMS shows that such processes can instead be represented as explicit scientific intermediates that can be generated, inspected, repaired and reused in prediction. In EI-MS, this intermediate is a mechanism-resolved fragmentation tree; more generally, the same principle can extend to predictive problems in which observations arise from structured sequences of mechanistic events. FragTreeMS therefore provides not only an approach for improving mass spectrum prediction, but also a general strategy for coupling interpretable scientific reasoning with quantitative machine learning.

Methods

Construction of the SMILES-EI-Triplet resource

Reference electron ionization (EI) spectra are obtained from the NIST 20 Mass Spectral Library. The curated SMILES-EI-Triplet resource contains 3,713 molecules and 52,997 source-mechanism-product triplets spanning four hydrocarbon classes: acyclic alkanes, alkenes, cycloalkanes and aromatic hydrocarbons. A vocabulary comprising 13 categories of ion formation and fragmentation processes is used consistently throughout annotation and modeling: Molecular ion, Isotopic peak, Alpha-cleavage, Sigma-bond cleavage, Benzylic cleavage, Allylic cleavage, McLafferty rearrangement, Neutral loss, Retro-Diels–Alder fragmentation, Hydrogen transfer, Radical-ion rearrangement, Dehydrogenation / Sequential dehydrogenation and Ring cleavage / Ring rearrangement.

Each ion formation event is represented by the structured triplet shown in Eq. 1:

$$t_{ik} = (s_{ik}, r_{ik}, p_{ik}) \,, \tag{1}$$

where s_{ik}, r_{ik} and p_{ik} denote the source, assigned mechanism and product ion, respectively. Sources are encoded either as a molecular substructure (`smiles_fragment`) or as a precursor ion (`precursor_mz`), representing direct ion formation from a molecular fragment and subsequent ion formation from a precursor ion relations, respectively. We use *relation* to denote the underlying chemical relation and *triplet* to denote its structured annotation representation.

Before annotation, peaks with intensities below 50 on the original NIST scale are removed. Consecutive peaks separated by no more than two nominal mass units are grouped, and peaks below 10% of the group-specific maximum are discarded. Retained spectra are subsequently normalized to the experimental base peak. Additional implementation details for spectrum preprocessing and record canonicalization are provided in Supplementary Note 1.

Fragmentation annotations are generated through a workflow combining language model assistance with manual review [19]. We use GPT-4.1 to generate candidate triplets from the molecular name, SMILES, molecular formula, molecular mass and filtered spectrum. Incomplete or internally inconsistent relations are then reviewed and fixed. In case of model bias, GPT-5 is subsequently used for further checking and completion under the 13-category mechanism vocabulary and constraints on molecular structure, elemental composition, compatibility between precursor and product ions, and mass consistency. Only records that pass manual review are retained in the curated resource. The language-model-assisted annotation protocol and class-specific chemical constraints are described in Supplementary Note 2.

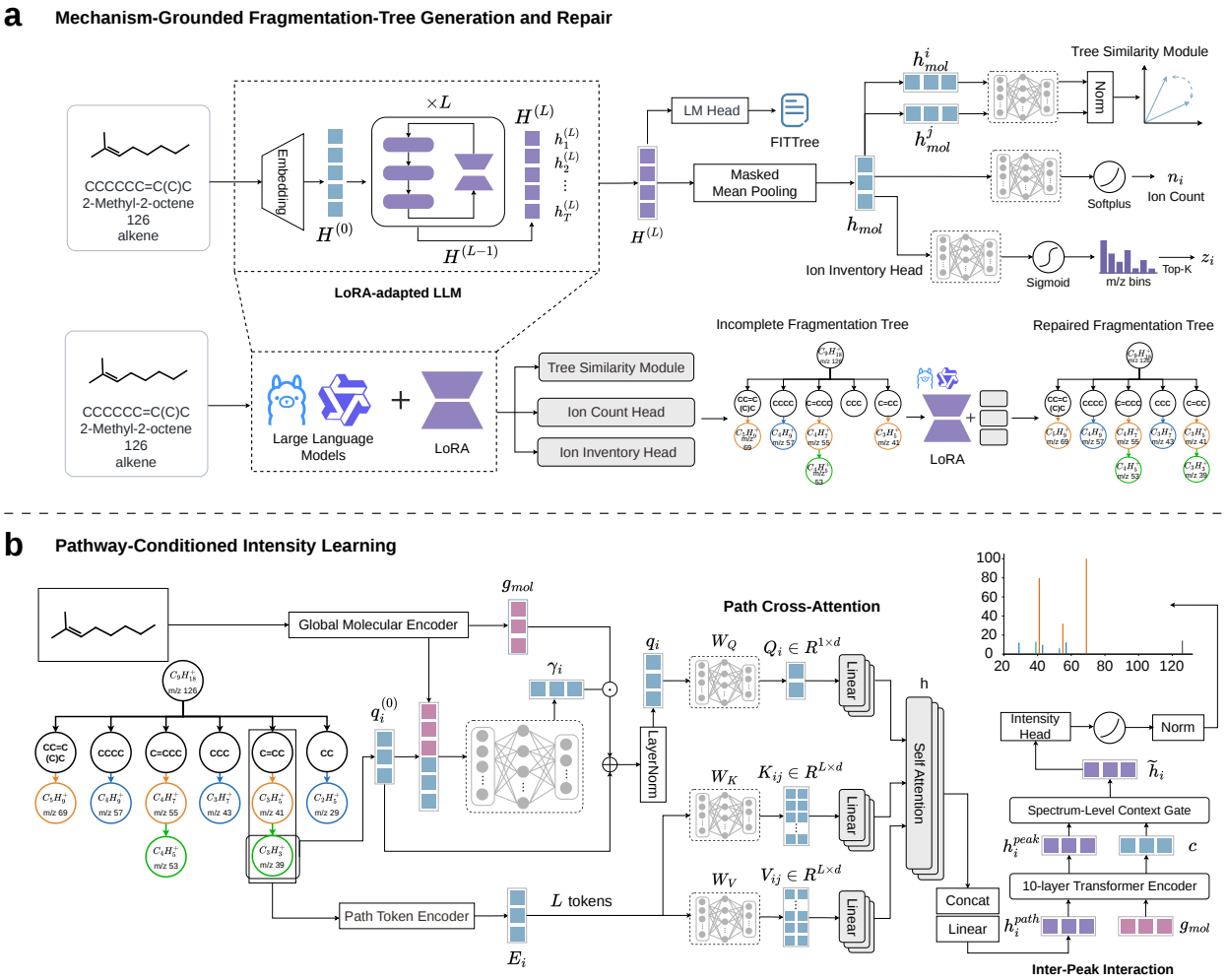


Fig. 6 | Architecture of FragTreeMS. a, A language model adapted with LoRA generates mechanism-resolved fragmentation trees. Structured ion masking is used to construct incomplete training trees, while supervision from Ion Inventory, Ion Count and Tree Similarity jointly shapes the shared representation used for Recall Repair. During inference, predictions from Ion Inventory and Ion Count are used to identify potentially missing candidate ions and guide review and repair of the initial tree. **b**, Pathways from the molecular root to individual ions are first integrated with global molecular context through Path Cross-Attention. Inter-Peak Interaction then models dependencies among the retained ions, and the resulting representations are used to predict normalized ion intensities and reconstruct the EI mass spectrum.

FragTreeMS architecture

The FragTreeMS consists of two sequentially coupled components (Fig. 6): mechanism-grounded fragmentation tree generation and repair (Stage 1), and pathway-conditioned intensity learning (Stage 2). Stage 1 maps the molecular structure to a mechanism-resolved fragmentation tree and uses structured ion masking to learn how to recover tree content that is deliberately removed. Stage 2 converts the repaired fragmentation tree into pathways from the molecular root to individual ions and predicts relative ion intensities by integrating pathway information, global molecular information and context among peaks.

Fragmentation trees provide a graph-based representation of precursor-product dependencies in mass

spectra [9, 10]. For a molecule i, the fragmentation tree G_i can be represented by:

$$G_i = \left(V_i^F, V_i^I, E_i^{F\to I}, E_i^{I\to I}\right), \tag{2}$$

where V_i^F and V_i^I denote molecular fragment nodes and ion nodes, respectively. $E_i^{F\to I}$ denotes the direct relation from a molecular fragment to an ion, whereas $E_i^{I\to I}$ denotes the direct relation from a precursor ion node to a product ion node. After canonicalization, each product ion is assigned a single parent. The fragmentation tree is then serialized using a deterministic structured text format, FITTree, for language model training. The FITTree grammar, parser validation and deterministic reverse conversion are provided in Supplementary Note 3.

Stage 1: mechanism-grounded fragmentation-tree generation and repair

For Stage 1, the input contains the molecular name, SMILES, molecular formula, molecular mass and compound class, with x_i denoting the input for molecule i. A language model adapted with LoRA generates the serialized fragmentation tree autoregressively. Tokens representing product ions and molecular formulas are assigned higher weights than other target tokens. The generation objective function $\mathcal{L}_{\text{tree}}$, normalized for each molecule, is defined as:

$$\mathcal{L}_{\text{tree}} = -\frac{1}{B}\sum_{i=1}^{B}\frac{\sum_{t=1}^{T_i} w_{it}\log P_\theta\left(y_{it} \mid x_i, y_{i,<t}\right)}{\sum_{t=1}^{T_i} w_{it}}, \tag{3}$$

where T_i denotes the number of target tokens for molecule i, and w_{it} is the weight assigned to target token t of molecule i. y_{it} denotes the target token at position t, and $y_{i,<t}$ denotes all target tokens preceding position t. The normalization within each molecule prevents longer tree serializations from receiving disproportionate weight in the minibatch optimization objective.

A masked mean pooling operation over the prompt hidden states produces a shared molecular representation that is used for three auxiliary objectives: Ion Inventory, Ion Count and Tree Similarity. Ion Inventory performs sparse multilabel prediction over nominal $m/z = 1, \ldots, 1,024$ using an asymmetric focal objective [27]; Ion Count estimates the number of distinct product ions and applies a stronger penalty to underestimation than to overestimation; Tree Similarity regularizes pairwise relations between molecular representations using the Jaccard similarities of product ions and triplets [28]. The complete objective function for Stage 1 is defined as:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{tree}} + \lambda_{\text{ion}}(e)\left(\mathcal{L}_{\text{inventory}} + \mathcal{L}_{\text{count}}\right) + \lambda_{\text{sim}}(e)\mathcal{L}_{\text{sim}}, \tag{4}$$

here e denotes the training epoch and $\lambda_{\text{ion}}(e)$ and $\lambda_{\text{sim}}(e)$ are epoch-dependent coefficients. All three auxiliary objectives update their head-specific parameters and the shared LoRA representation. Target weighting, construction of the auxiliary heads and complete auxiliary loss definitions are provided in Supplementary Note 4.

The initial fragmentation tree is obtained by inference using the trained model described above. Missing local nodes in the initially generated tree may interrupt entire downstream fragmentation pathways. We therefore need to improve the model to recognize incomplete trees and recover the missing content, thereby improving the completeness of the resulting fragmentation tree. We refer to this process as Recall Repair.

Recall repair is learned through structured ion masking rather than through a separately trained repair model, following the denoising principle of reconstructing a complete target from a corrupted input [29]. For the training of recall repair, we convert each training sample into repair example with probability p_{repair} . For each selected record, a fraction $r_i \sim \mathcal{U}(r_{\min}, r_{\max})$ is sampled, and the corresponding proportion of product ions together with their associated relations is removed while ensuring that at least one product ion remains. The molecular information, incomplete fragmentation tree and masses of the intentionally removed product ions together form the repair prompt, while the original complete fragmentation tree remains the target output. Direct generation and repair share the same adapted language model and optimization objective in Stage 1.

At inference, the model first generates an initial fragmentation tree $G_i^{(0)}$ through deterministic decoding. Ion Inventory and Ion Count then provide candidate ion guidance for self-review, whereas Tree Similarity is used only during training. Candidate masses are restricted to the mass range defined for each molecule from $z_{\min} = 1$ to $z_i^{\max} = \min(1, 024, \max(1, \text{round}(M_i) + 4))$, where M_i denotes the molecular mass. The auxiliary candidate set $\mathcal{C}_i^{\text{aux}}$ for repair is defined as:

$$\mathcal{C}_i^{\text{aux}} = \{z \in [z_{\min}, z_i^{\max}] \,|\, \pi_{iz} \geq \tau_{\text{inv}}\} \cup \text{TopK}_{z \in [z_{\min}, z_i^{\max}]} (\boldsymbol{\pi}_i; k_i) \,, \tag{5}$$

where π_{iz} denotes the Ion Inventory score predicted for mass z of molecule i, and $\boldsymbol{\pi}_i$ denotes the corresponding vector of Ion Inventory scores over the valid m/z range. τ_{inv} is the threshold used to retain candidate masses from Ion Inventory, and k_i denotes the number of additional high-scoring masses selected for molecule i by the TopK operation. Thus, masses with Ion Inventory scores exceeding τ_{inv} are combined with the k_i highest-scoring masses within the valid m/z range to form the auxiliary candidate set $\mathcal{C}_i^{\text{aux}}$. k_i is controlled by the predicted Ion Count according to $k_i = \min\left(K_{\max}, \lceil \eta_{\text{count}} \max(n_{\min}, \widehat{n}_i) \rceil + \Delta_k, z_i^{\max}\right)$, where $\widehat{n}_i$ denotes the Ion Count predicted for molecule i , $n_{\min}$ specifies the minimum ion count used when determining the candidate budget, and η_{count} is a scaling factor applied to the predicted count. Δ_k provides an additional candidate margin to reduce the risk of excluding potentially missing ions, while $K_{\max}$ defines the maximum number of candidates allowed. $z_i^{\max}$ further constrains the candidate number according to the valid mass

range of molecule i.The ceiling operation ensures that the resulting candidate budget is an integer. Together, these terms determine the number ki of high-scoring masses additionally selected from the Ion Inventory predictions.

Masses already present in the initial fragmentation tree $G_i^{(0)}$ are removed from the candidate list. If potentially missing candidate ions remain, the missing candidate ions are reorganized together with the initial fragmentation tree and provided to the adapted language model, which reviews the incomplete tree again and attempts to recover the missing content. This feedback-guided second pass is related to language-model self-refinement [30], but uses molecule-specific candidate evidence and preserves supported first-pass relations. The reviewed output is then merged with the initial fragmentation tree without deleting supported relations generated in the first pass, yielding G_i^{rep}. Only product ions retained in the final repaired tree are passed to Stage 2. Structured ion masking, candidate filtering, relation-preserving merging and diagnostic candidate analysis are detailed in Supplementary Note 5.

Stage 2: pathway-conditioned intensity learning

Each product ion retained in the final fragmentation tree is associated with a unique pathway from the molecular root to that ion. Direct relations from a molecular structure to an ion originate from the molecular root, whereas sequential fragmentation proceeds through relations involving precursor ions. Each element along the pathway encodes textual information, pathway position, relation type, mechanism category and information derived from mass, including the source mass, product mass, mass difference and corresponding variants normalized by molecular mass. Deterministic pathway construction, handling of disconnected ions and complete feature encoding are provided in Supplementary Note 6.

Each product ion is represented by a query vector specific to that ion, which combines ion identity and mass features. The global molecular structure is encoded independently using a Morgan fingerprint with chirality information computed by RDKit [31]. A gated residual fusion then combines this molecular representation with the ion query before pathway attention is applied. Let E_i denote the encoded pathway from the molecular root to the ion, and let q_i denote the fused ion query. Path Cross-Attention uses the ion representation as the query and the pathway elements as keys and values. For attention head h, the computation is defined as:

$$\mathrm{head}_i^{(h)} = \mathrm{softmax}\left(\frac{(q_i W_Q^{(h)})(E_i W_K^{(h)})^{\mathsf{T}}}{\sqrt{d_h}} + M_i^{\mathrm{path}}\right)(E_i W_V^{(h)}), \tag{6}$$

where $W_Q^{(h)}$, $W_K^{(h)}$ and $W_V^{(h)}$ are the query, key and value projections for head h, M_i^{path} masks padded pathway positions, d is the model dimension and $d_h = d/H$ for H attention heads. The head outputs are combined with the ion query through residual feed-forward processing to produce the pathway-conditioned

ion representation.

The pathway-conditioned representations of all retained ions are subsequently processed by Inter-Peak Interaction, as shown in Fig. 6. A projected global molecular token is prepended to the ion sequence and jointly contextualized by a bidirectional Transformer encoder [32, 33]. The resulting spectrum-level context is returned to each ion through a gated residual connection. We then use a Softplus intensity head to predict non-negative ion abundances, which are normalized by the strongest valid predicted ion within each molecule.

Intensity regression at the candidate ion level uses a weighted Smooth L_1 objective [34]. The regression objective is defined as:

$$\mathcal{L}_{\text{int}} = \frac{1}{B}\sum_{m=1}^{B} \frac{\sum_{i\in\mathcal{C}_m^{\text{rep}}} w_{mi}\,\text{SmoothL1}_{\delta_{\text{int}}}\left(\widehat{I}_{mi} - y_{mi}\right)}{\sum_{i\in\mathcal{C}_m^{\text{rep}}} w_{mi}}, \tag{7}$$

where B is the minibatch size, $\mathcal{C}_m^{\text{rep}}$ is the repaired product-ion set for molecule m, $\widehat{I}_{mi}$ and y_{mi} are the predicted and base-peak-normalized reference intensities, w_{mi} is the ion-specific regression weight and δ_{int} is the transition parameter of the Smooth L_1 penalty.

Candidate-level regression weighting and the remaining Stage 2 implementation settings are provided in Supplementary Note 6. To account for experimental ions that are absent from the repaired fragmentation tree, the predicted intensity of these missing reference masses is set to zero. The Cosine similarity of the complete spectrum is defined by:

$$\text{Cosine}_m = \frac{\sum_{z\in\mathcal{C}_m^{\text{rep}}} \widehat{I}_{mz} y_{mz}}{\sqrt{\sum_{z\in\mathcal{C}_m^{\text{rep}}} \widehat{I}_{mz}^2}\sqrt{\sum_{z\in\mathcal{G}_m} y_{mz}^2} + \varepsilon}, \tag{8}$$

where $\mathcal{C}_m^{\text{rep}}$ denotes the final repaired set of product ions, and $\mathcal{G}_m$ denotes the complete filtered reference spectrum. The final objective of Stage 2 combines intensity regression at the candidate ion level with cosine similarity loss at the spectrum level, as defined as:

$$\mathcal{L}_{\text{stage2}} = \lambda_{\text{int}}\mathcal{L}_{\text{int}} + \lambda_{\text{spec}}\mathcal{L}_{\text{spec}}, \tag{9}$$

where λ_{int} and λ_{spec} denote the weights assigned to the intensity regression objective and the spectrum similarity objective, respectively. In our experiments, λ_{spec} is set larger than λ_{int}, indicating that spectrum similarity is given greater importance during optimization.

Training and reproducibility

Molecules are split at the molecule level into training and test sets at a 9:1 ratio, and used for all backbone models, baselines and FragTreeMS.

Stage 1 uses Qwen3-8B, Llama3.1-8B-Instruct and Qwen2.5-14B-Instruct as alternative backbone models. The pretrained parameters are frozen and adapted using LoRA [35]. Recall Repair training uses $p_{\text{repair}} = 0.25$, $r_i \sim \mathcal{U}(0.10, 0.30)$, $\tau_{\text{inv}} = 0.20$ and $K_{\max} = 64$. Stage 1 is trained for 30 epochs using AdamW [36]. The learning rates are 3×10^{-5} for the LoRA parameters and 1×10^{-4} for the auxiliary heads, followed by cosine learning rate decay. Additional Stage 1 generation, auxiliary-objective and optimization settings are reported in Supplementary Table 22.

Stage 2 uses a Morgan fingerprint with 4,096 bits and radius 2, hidden dimensionality $d = 1,152$, $H = 18$ attention heads and $L_{\text{inter}} = 10$ Transformer layers. We set $\lambda_{\text{int}} = 0.05$ and $\lambda_{\text{spec}} = 3.00$ in Eq. 9. Training uses AdamW with an initial learning rate of 5×10^{-5}, cosine decay, bfloat16 mixed precision and an exponential moving average with decay 0.999 [37] for 350 epochs. The implementation uses PyTorch [38] and Hugging Face Transformers [39]. Additional pathway-representation and architecture settings are reported in Supplementary Table 23. Optimization, evaluation and reproducibility settings are reported in Supplementary Table 24.

Evaluation and statistical analysis

Fragmentation tree generation is evaluated primarily using Product-ion Recall. Let $\mathcal{P}_i$ and $\mathcal{G}_i$ denote the predicted and reference sets of product ion nominal m/z values for molecule i, respectively. Product-ion Recall is defined as:

$$\text{Product-ion Recall} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{P}_i \cap \mathcal{G}_i|}{\max(1, |\mathcal{G}_i|)}, \tag{10}$$

where N is the number of evaluated molecules. The score is therefore a molecule-level macro-average under exact nominal-mass matching.

Product-ion F2 and Strong-ion Recall are additionally used as complementary measures of product-ion recovery. Product-ion F2 combines product-ion precision and recall while assigning greater weight to recall, whereas Strong-ion Recall evaluates the recovery of reference product ions designated as strong according to the predefined intensity criterion.

Spectrum prediction is evaluated using Base-peak Accuracy, Cos-Pre, Cos-All, Cosine similarity over the complete spectrum and Ion Recall@5. Base-peak Accuracy is the fraction of molecules for which the m/z of the highest intensity predicted peak matches the experimental base peak. Cos-Pre measures intensity agreement over peaks represented by the model, whereas Cos-All measures agreement over the complete filtered reference peak set. Cosine similarity over the complete spectrum is calculated according to Eq. 8 and therefore penalizes both missing reference peaks and unsupported predicted peaks. Ion Recall@5 measures the

fraction of filtered reference ion masses recovered among the five predicted peaks with the highest intensities. Complete metric definitions and statistical aggregation procedures are provided in Supplementary Note 7.

Baselines

FragTreeMS is compared with the Direct (Base) and Triplets language model baselines, as well as NEIMS and CFM-ID. The language model comparison includes GPT-5.5, DeepSeek-v4-pro, Qwen3-8B, Llama3.1-8B and Qwen2.5-14B. In the Direct setting, each model predicts the EI spectrum directly from molecular information. In the Triplets setting, each model first generates source-mechanism-product triplets from the molecular information and then predicts intensities for the product ions contained in its generated triplets. Conditions labeled "Base" in Fig. 4 correspond to the Direct setting. NEIMS [6] is retrained on the same fixed data split used for FragTreeMS. CFM-ID [4] is evaluated using its publicly released parameters without retraining and is therefore treated as an external reference rather than a comparison under matched training conditions. Exact prompt templates and implementation details for the Direct, Triplets and Gold conditions are provided in Supplementary Note 8.

Data availability

Project data that the authors are permitted to redistribute, including the curated annotations of fragmentation mechanisms and the random partitioning scheme applied at the molecule level, are available in the FragTreeMS repository at `https://github.com/Swpy39/FragTreeMS`. The underlying reference EI spectra originate from the NIST 20 Mass Spectral Library and remain subject to the applicable NIST licensing terms; these raw licensed records are therefore not redistributed by the authors.

Code availability

Source code for data processing, fragmentation-tree generation and repair, pathway-conditioned intensity learning and evaluation is available at `https://github.com/Swpy39/FragTreeMS`. An interactive visualization of molecule-level fragmentation trees and predicted spectra is available at `https://pec.cup.edu.cn/FragTreeMS`.

References

[1] Ásgeirsson, V., Bauer, C. A. & Grimme, S. Quantum chemical calculation of electron ionization mass spectra for general organic and inorganic molecules. *Chemical Science* **8**, 4879–4895 (2017). URL https://doi.org/10.1039/c7sc00601b. https://pubs.rsc.org/sc/article-pdf/8/7/4879/5763447/c7sc00601b.pdf.

[2] Wang, S., Kind, T., Bremer, P. L., Tantillo, D. J. & Fiehn, O. Quantum chemical prediction of electron ionization mass spectra of trimethylsilylated metabolites. *Analytical chemistry* **94**, 1559–1566 (2022).

[3] Grimme, S. Towards first principles calculation of electron impact mass spectra of molecules. *Angewandte Chemie International Edition* **52**, 6306–6312 (2013). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201300158. https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.201300158.

[4] Allen, F., Pon, A., Greiner, R. & Wishart, D. Computational prediction of electron ionization mass spectra to assist in gc/ms compound identification. *Analytical chemistry* **88**, 7689–7697 (2016).

[5] Zhang, B., Zhang, J., Xia, Y., Chen, P. & Wang, B. Prediction of electron ionization mass spectra based on graph convolutional networks. *International Journal of Mass Spectrometry* **475**, 116817 (2022). URL https://www.sciencedirect.com/science/article/pii/S1387380622000227.

[6] Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science* **5**, 700 (2019).

[7] Zhu, R. L. & Jonas, E. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical Chemistry* **95**, 2653–2663 (2023). URL https://doi.org/10.1021/acs.analchem.2c02093. https://pubs.acs.org/ancham/article-pdf/95/5/2653/8708975/ac2c02093.pdf.

[8] Devata, S., Cleaves, H. J., Dimandja, J., Heist, C. A. & Meringer, M. Comparative evaluation of electron ionization mass spectral prediction methods. *Journal of the American Society for Mass Spectrometry* **34**, 1584–1592 (2023).

[9] Rasche, F., Svatos, A., Maddula, R. K., Bottcher, C. & Bocker, S. Computing fragmentation trees from tandem mass spectrometry data. *Analytical Chemistry* **83**, 1243–1251 (2010). URL https://doi.org/10.1021/ac101825k. https://pubs.acs.org/ancham/article-pdf/83/4/1243/10042380/ac101825k.pdf.

[10] Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *Journal of cheminformatics* **8**, 5 (2016).

[11] Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using csi:fingerid. *Proceedings of the National Academy of Sciences* **112**, 12580–12585 (2015). URL https://www.pnas.org/doi/abs/10.1073/pnas.1509788112. https://www.pnas.org/doi/pdf/10.1073/pnas.1509788112.

[12] Dührkop, K. *et al.* Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature methods* **16**, 299–302 (2019).

[13] Goldman, S., Li, J. & Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks. *Analytical Chemistry* **96**, 3419–3428 (2024). URL https://doi.org/10.1021/acs.analchem.3c04654. https://pubs.acs.org/ancham/article-pdf/96/8/3419/18539624/ac3c04654.pdf.

[14] Ruttkies, C., Schymanski, E., Wolf, S., Hollender, J. & Neumann, S. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. j cheminform 8: 3 (2016).

[15] Schwaller, P. *et al.* Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **5**, 1572 (2019).

[16] Edwards, C. *et al.* Translation between molecules and natural language. In Goldberg, Y., Kozareva, Z. & Zhang, Y. (eds.) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 375–413 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022). URL https://aclanthology.org/2022.emnlp-main.26/.

[17] Segler, M. H., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature* **555**, 604–610 (2018).

[18] Ross, J. *et al.* Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **4**, 1256–1264 (2022).

[19] Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence* **6**, 161–169 (2024).

[20] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023).

[21] M. Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nature machine intelligence* **6**, 525–535 (2024).

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

[23] Young, A., Röst, H. & Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence* **6**, 404–416 (2024).

[24] Xu, Y., Ma, Y., Xu, W., Yang, Z. & Ting, K. M. A large language model for deriving spectral embeddings for accurate compound identification in mass spectrometry. *Communications Chemistry* **8**, 326 (2025).

[25] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**, 206–215 (2019).

[26] Karniadakis, G. E. *et al.* Physics-informed machine learning. *Nature Reviews Physics* **3**, 422–440 (2021).

[27] Ridnik, T. *et al.* Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 82–91 (IEEE, 2021).

[28] Jaccard, P. Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles* **37**, 547–579 (1901).

[29] Lewis, M. *et al.* 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*, 7871–7880 (2020).

[30] Madaan, A. *et al.* Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems* **36**, 46534–46594 (2023).

[31] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling* **50**, 742–754 (2010).

[32] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

[33] Lee, J. *et al.* Set transformer: A framework for attention-based permutation-invariant neural networks. In Chaudhuri, K. & Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 3744–3753 (PMLR, 2019). URL `https://proceedings.mlr.press/v97/lee19d.html`.

[34] Girshick, R. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448 (2015).

[35] Hu, E. J. *et al.* LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations* (2022). URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[36] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019). URL https://openreview.net/forum?id=Bkg6RiCqY7.

[37] Polyak, B. T. & Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization* **30**, 838–855 (1992).

[38] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019).

[39] Wolf, T. *et al.* Transformers: State-of-the-art natural language processing. In Liu, Q. & Schlangen, D. (eds.) *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45 (Association for Computational Linguistics, Online, 2020). URL https://aclanthology.org/2020.emnlp-demos.6/.

Author contributions

Hongyan Sun and Shuofan Li conceived the study and developed the methodology. Hongyan Sun developed the software, performed the computational experiments and formal analyses, prepared the visualizations, and wrote the original draft. Shuofan Li contributed to dataset construction, data analysis, investigation, and methodological validation. Quan Shi contributed to the investigation and methodological validation, acquired funding, provided resources, and supervised the study. Jianxun Wu contributed to dataset construction and data analysis. Chunming Xu acquired funding, provided resources, and supervised the study. Shuangyin Li contributed to methodology development, supervised the study, and revised the original draft. All authors reviewed and edited the manuscript and approved the final version.

Funding

The authors would like to acknowledge the financial support from the National Science and Technology Major Project, 2025ZD1607600. Shuofan Li acknowledges the financial support from the National Natural Science Foundation of China (No. 22608612) and the Science Foundation of China University of Petroleum, Beijing (No. 2462026XKBH009).

Competing interests

The authors declare no competing interests.

Supplementary Files

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

- FragTreeMSSupplyment.pdf