# Supplementary Information for 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)

# Contents

**Supplementary Methods** **4**

Supplementary Note 1. Record-level curation and canonicalization . . . . . 4
Supplementary Note 2. LLM-assisted fragmentation annotation protocol . . . . . 4
Supplementary Note 3. FITTree serialization and constrained parsing . . . . . 7
Supplementary Note 4. Stage 1 target weighting and auxiliary objectives . . . . . 8
Supplementary Note 5. Recall Repair implementation and diagnostic candidates . . . . . 11
Supplementary Note 6. Pathway construction, encoding and Stage 2 implementation . . . 11
Supplementary Note 7. Evaluation metrics and statistical aggregation . . . . . 12
Supplementary Note 8. Baseline and ablation implementation . . . . . 15

**Supplementary Tables** **17**

Supplementary Table 1. Full-model performance and baseline comparisons across all datasets: Base-peak accuracy and Cosine. . . . . 17
Supplementary Table 2. Full-model performance and baseline comparisons across all datasets: Cos-Pre and Cos-All. . . . . 18
Supplementary Table 3. Full-model performance and baseline comparisons on acyclic alkanes: Base-peak accuracy and Cosine. . . . . 19
Supplementary Table 4. Full-model performance and baseline comparisons on alkenes: Base-peak accuracy and Cosine. . . . . 20
Supplementary Table 5. Full-model performance and baseline comparisons on aromatic hydrocarbons: Base-peak accuracy and Cosine. . . . . 21
Supplementary Table 6. Full-model performance and baseline comparisons on cycloalkanes: Base-peak accuracy and Cosine. . . . . 22
Supplementary Table 7. Full-model performance and baseline comparisons on acyclic alkanes: Cos-Pre and Cos-All. . . . . 23
Supplementary Table 8. Full-model performance and baseline comparisons on alkenes: Cos-Pre and Cos-All. . . . . 24
Supplementary Table 9. Full-model performance and baseline comparisons on aromatic hydrocarbons: Cos-Pre and Cos-All. . . . . 25
Supplementary Table 10. Full-model performance and baseline comparisons on cycloalkanes: Cos-Pre and Cos-All. . . . . 26
Supplementary Table 11. Performance under the Gold condition on acyclic alkanes: Base-peak accuracy and Cosine. . . . . 27
Supplementary Table 12. Performance under the Gold condition on alkenes: Base-peak accuracy and Cosine. . . . . 27
Supplementary Table 13. Performance under the Gold condition on aromatic hydrocarbons: Base-peak accuracy and Cosine. . . . . 28

Supplementary Table 14. Performance under the Gold condition on cycloalkanes: Base-peak accuracy and Cosine. . . . . 28
Supplementary Table 15. Performance under the Gold condition on acyclic alkanes: Cos-Pre and Cos-All. . . . . 29
Supplementary Table 16. Performance under the Gold condition on alkenes: Cos-Pre and Cos-All. . . . . 29
Supplementary Table 17. Performance under the Gold condition on aromatic hydrocarbons: Cos-Pre and Cos-All. . . . . 30
Supplementary Table 18. Performance under the Gold condition on cycloalkanes: Cos-Pre and Cos-All. . . . . 30
Supplementary Table 19. Performance under the Gold condition across all datasets: Base-peak accuracy and Cosine. . . . . 31
Supplementary Table 20. Performance under the Gold condition across all datasets: Cos-Pre and Cos-All. . . . . 31
Supplementary Table 21. Stage 1 fragmentation-tree generation and repair performance across datasets and backbone models. . . . . 32
Supplementary Table 22. Key Stage 1 generation, auxiliary-objective and optimization parameters. . . . . 33
Supplementary Table 23. Key Stage 2 pathway-representation and architecture parameters. 33
Supplementary Table 24. Key Stage 2 optimization, evaluation and reproducibility parameters. . . . . 34

**Supplementary References** **35**

## Supplementary Methods

The main Methods define the resource, model architecture, training objectives and evaluation endpoints. The notes below report only implementation choices that are required for reproduction but are not specified there. Equations and definitions given in the main text are not repeated.

### Supplementary Note 1. Record-level curation and canonicalization

Manual review is performed on each complete annotation record after the language-model-assisted passes described in the main Methods. A retained relation contains a parseable source, a mechanism label from the controlled vocabulary and a product formula consistent with its nominal mass. A sequential relation also contains an internally consistent precursor reference. Exact duplicate relations are removed during canonicalization. The resulting annotations represent chemically plausible assignments at nominal mass resolution and do not claim a unique microscopic gas-phase trajectory when several routes remain compatible with the observed ion.

For downstream processing, raw molecular-fragment strings are preserved in the `smiles_fragment` field and precursor masses are stored in the `precursor_mz` field. Field names, product formulas, nominal masses and mechanism labels are normalized before duplicate detection, so equivalent relations do not remain duplicated because of formatting differences.

### Supplementary Note 2. LLM-assisted fragmentation annotation protocol

**Fragmentation annotation generation with GPT-4.1.** GPT-4.1 generates the initial source-mechanism-product triplets with a separate prompt for each molecular class. For each molecule, the model receives its name, class, SMILES, molecular formula, molecular weight and complete 70 eV EI spectrum. The spectrum is supplied as a peak table containing nominal $m/z$ and relative intensity. For each peak, the model identifies a plausible source, selects a controlled relation label and assigns a product ion formula. It also records whether the peak follows a regular 14 Da series, shows unusually high intensity or is associated with a site that can retain charge. The output follows a fixed JSON schema.

**General annotation constraints for generation.** Each source-mechanism-product triplet links a source to its assigned product ion through a controlled relation. The relation field uses one of the following 13 labels:

- 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;
- Ring cleavage / Ring rearrangement.

When `precursor_mz` is used, the precursor peak must be present and must not be weaker than the proposed product peak. This intensity requirement applies to relations between ions and does not prevent a strong ion from being assigned directly from the molecular structure. If no supported relation is available, the peak is marked as invalid instead of assigning an invented fragment or pathway.

**Chemical constraints for each dataset.** The prompt for each dataset applies the following main rules:

- *Acyclic alkane.* Ordinary C-C bond cleavage that forms an alkyl cation is assigned to Sigma-bond cleavage, and the source fragment retains the relevant carbon skeleton. For a branched molecule, the selected fragment preserves the tertiary or quaternary carbon environment when that site explains charge retention or a strong peak. An ion at even $m/z$ is assigned to Dehydrogenation / Sequential dehydrogenation only when the spectrum contains a suitable precursor with the same number of carbon atoms, more hydrogen atoms and sufficient intensity.
- *Alkene.* The prompt first identifies the position and substitution of each C=C bond and determines whether the double bonds are isolated, conjugated or cumulated. Allylic cleavage is assigned only when cleavage next to C=C produces a connected allylic or conjugated ion; ordinary C-C cleavage without allylic stabilization remains Sigma-bond cleavage. The source fragment preserves the relevant C=C bond and contains the same number of carbon atoms as the product ion, whereas dehydrogenation requires a supported precursor with the same carbon count, more hydrogen atoms and sufficient intensity.
- *Aromatic hydrocarbon.* An intense intact radical cation is retained as the Molecular ion, and an isotope relation is assigned only to a supported $M+1$ or $M+2$ peak. For an aromatic molecule with an alkyl substituent, an $M-15$ ion is assigned to Neutral loss when the structure supports methyl radical loss; Benzylic cleavage is used when cleavage at a benzylic position produces a stabilized ion. Ring cleavage / Ring rearrangement is assigned only when the

aromatic framework is broken, contracted or reorganized; neutral loss and dehydrogenation require a supported precursor and a chemically consistent mass change.

- *Cycloalkane.* Ring cleavage / Ring rearrangement is used for ions that arise from a supported opening or reorganization of the cyclic skeleton, including the main $C_nH_{2n-1}^+$ series for $n$ = 3-7. The source fragment is connected, represents the retained carbon skeleton and normally contains the same number of carbon atoms as the product ion; the complete parent SMILES is not used for a smaller product. Sigma-bond cleavage remains valid for an ion formed from an alkyl side chain, while dehydrogenation requires a supported precursor with the same carbon count, more hydrogen atoms and sufficient intensity.

**Annotation refinement with GPT-4.1.** A second GPT-4.1 pass reviews one generated triplet at a time. It receives the molecular record, complete spectrum, original triplet, fragment checks produced by the code, previously verified triplets and candidate precursor ions or source fragments. These candidates provide supporting evidence and are checked again before use. The model verifies the product formula and nominal mass, the carbon count, the source format, the fragment structure, the selected relation and the measured intensities. It returns `accept` when the original triplet is chemically reasonable and `revise` when a clear error is present. It does not delete a triplet and changes only the incorrect field whenever possible. The observed product ion remains unchanged unless its formula and nominal $m/z$ disagree.

The refinement prompt for each dataset focuses on the following checks:

- *Acyclic alkane.* The judge checks whether an alkyl cation has a connected source fragment and a valid Sigma-bond cleavage assignment. An unsupported `precursor_mz` is replaced with a suitable `smiles_fragment` when one is available, while a fragment containing the relevant branch is retained when it explains charge stabilization. Dehydrogenation is accepted only when carbon number, hydrogen count, precursor evidence and intensity are consistent.

- *Alkene.* The judge checks that the source fragment is connected, contains the correct number of carbon atoms and preserves the relevant C=C environment. A valid Allylic cleavage assignment is retained, and Sigma-bond cleavage is changed to Allylic cleavage only when the broken bond and retained ion are genuinely allylic. Dehydrogenation is accepted only when an observed precursor has the same carbon count, more hydrogen atoms and an intensity at least as high as that of the product.

- *Aromatic hydrocarbon.* The judge checks the formula and mass direction before accepting an isotope relation, neutral loss or dehydrogenation step. Supported Molecular ion and Benzylic cleavage assignments are retained, while Ring cleavage / Ring rearrangement requires an actual change to the aromatic framework. An unsupported `precursor_mz` is replaced with a valid aromatic source fragment when possible, and dehydrogenation is accepted only when the precursor evidence and intensity satisfy all constraints.

- *Cycloalkane.* The judge checks that each `smiles_fragment` is connected, occurs in the parent structure and has the carbon skeleton required by the product ion. It retains Ring cleavage / Ring rearrangement for a supported primary cyclic ion and corrects use of the complete parent SMILES for a smaller product. Sigma-bond cleavage is retained for an ion formed from a valid side chain, while dehydrogenation requires a verified precursor with the same carbon count, more hydrogen atoms and sufficient intensity.

**GPT-5-based annotation correction.** After GPT-4.1 refinement, GPT-5 reviews the molecular structure, EI spectrum, initial triplets and revised triplets. It corrects remaining errors in the source, product formula or selected relation and adds a plausible triplet when an observed ion remains unassigned. The final triplets are normalized to the source-mechanism-product format used to construct the fragmentation trees.

## Supplementary Note 3. FITTree serialization and constrained parsing

The curated source-mechanism-product relations are converted into the deterministic FITTree representation used for Stage 1 training. The current implementation retains the legacy field names `fragment_edges` and `ion_edges`, together with the termination token `[GRAPH_END]`. These literal implementation tokens do not alter the tree-based scientific representation used in the manuscript.

Each record contains a count block, a product-ion inventory, a molecular-fragment inventory, fragment-to-ion relations, ion-to-ion relations and a termination marker. Structured fields are tab-delimited and follow this grammar:

```
[COUNTS]    fragments=<int>    ions=<int>
            fragment_edges=<int>    ion_edges=<int>

[IONS]
I0    <positive nominal m/z>    <formula>
      <molecular_ion|fragment_ion|referenced_precursor>
I1    <positive nominal m/z>    <formula>
      <molecular_ion|fragment_ion|referenced_precursor>
...
[/IONS]

[FRAGMENTS]
F0    "<JSON-quoted raw fragment text>"
F1    "<JSON-quoted raw fragment text>"
...
[/FRAGMENTS]

[FRAGMENT_TO_ION]
```

```
F0    I0    <controlled mechanism label>
F1    I2    <controlled mechanism label>
...
[/FRAGMENT_TO_ION]

[ION_TO_ION]
I0    I1    <controlled mechanism label>
I1    I2    <controlled mechanism label>
...
[/ION_TO_ION]

[GRAPH_END]
```

Triplets with a `smiles_fragment` source generate fragment-to-ion relations, whereas triplets with a `precursor_mz` source generate precursor-ion-to-product-ion relations. One ion entry is used for each represented nominal mass. A precursor referenced by an ion-to-ion relation but absent from the explicit product-ion inventory is retained as a `referenced_precursor`. The molecular ion is serialized first when it is present, followed by the remaining ions in descending nominal $m/z$.

The parser reads the region from `[COUNTS]` to the first `[GRAPH_END]` and verifies the expected sections, field counts and local identifiers. Ion entries require a positive nominal mass. Relation entries require valid fragment or ion references, and mechanism labels must belong to the controlled vocabulary. Exact duplicate relations and self-transitions are removed. Non-isotopic precursor-to-product relations are also checked for a compatible nominal-mass direction. The reverse converter returns normalized source-mechanism-product relations in deterministic order.

## Supplementary Note 4. Stage 1 target weighting and auxiliary objectives

Ion-aware target weighting is assigned before subword tokenization. The higher weight applies to the numerical value of `ions=<count>` and to the nominal $m/z$ and molecular-formula fields in the `[IONS]` section. Formatting characters, local identifiers, ion-type fields, fragment text and relation fields retain the ordinary weight. Character spans are mapped to tokenizer offset spans, and each target token receives the mean character weight over its span. Averaging over each character span makes the weighting independent of the subword segmentation of a formula or nominal mass. The two weights are listed in Supplementary Table 22.

The three auxiliary heads use the masked mean of the final hidden states at unpadded positions in the molecular prompt; target FITTree tokens do not contribute to this representation. Ion Inventory applies a sigmoid linear projection, Ion Count applies a Softplus linear projection and Tree Similarity applies an $L_2$-normalized 256-dimensional projection. During training, the valid Ion Inventory range extends to $z^{\max}_{i,\text{train}} = \min(1,024, \max(\text{round}(M_i) + 4, \max(\mathcal{I}_i)))$, where $M_i$ is the molecular weight and $\mathcal{I}_i$ is the set of reference product ions for molecule $i$. Extending the range in this way prevents a curated positive ion from being masked.

For the Ion Inventory objective, the occurrence-based mass weight is defined as:

$$\omega_z = \begin{cases} \mathrm{clip}\left(\sqrt{\frac{f_{\mathrm{med}}}{f_z}}, 1, 3\right), & f_z > 0, \\ 1, & f_z = 0, \end{cases} \tag{S1}$$

where $f_z$ is the occurrence frequency of mass $z$, $f_{\mathrm{med}}$ is the median nonzero frequency and $\mathrm{clip}(x, l, u) = \min(\max(x, l), u)$ restricts $x$ to the interval $[l, u]$. Square-root inverse-frequency scaling emphasizes rare ion masses, whereas clipping limits excessive weights. Molecule-specific validity masking and separate normalization of positive and negative mass positions are introduced for the ion inventory. The positive normalizer is defined as:

$$N_i^+ = \max\left(1, \sum_z v_{iz} a_{iz}\right), \tag{S2}$$

where $i$ indexes molecules, $z$ indexes nominal mass bins, $a_{iz} \in \{0, 1\}$ indicates whether mass $z$ occurs in the reference ion inventory and $v_{iz} \in \{0, 1\}$ indicates whether that mass bin is valid for molecule $i$. Normalization by $N_i^+$ keeps molecules with more observed ions from dominating the positive term and prevents division by zero.

The negative normalizer is defined as:

$$N_i^- = \max\left(1, \sum_z v_{iz}(1 - a_{iz})\right). \tag{S3}$$

Division by $N_i^-$ keeps the negative contribution comparable across molecules with different valid mass ranges and remains well defined when no negative position is available.

The stabilized probability assigned to a negative mass position follows the probability-shifting component of Asymmetric Loss [1] and is defined as:

$$q_{iz}^- = \mathrm{clip}(1 - \pi_{iz} + m_-, \varepsilon, 1), \tag{S4}$$

where $\pi_{iz} \in [0, 1]$ is the predicted probability of ion mass $z$, $m_-$ is the negative margin, $\varepsilon$ is the numerical lower bound and $\mathrm{clip}(x, l, u)$ restricts $x$ to the interval $[l, u]$. Clipping stabilizes the logarithm in the negative term and prevents undefined numerical values.

Combining this probability shift with focal modulation [2], the Ion Inventory loss is defined as:

$$\mathcal{L}_{\mathrm{inventory}} = -\frac{1}{B} \sum_{i=1}^{B} \left[ \frac{\sum_z \alpha_+ \omega_z v_{iz} a_{iz} \log(\max(\pi_{iz}, \varepsilon))}{N_i^+} + \frac{\sum_z v_{iz}(1 - a_{iz}) \pi_{iz}^{\gamma_-} \log(q_{iz}^-)}{N_i^-} \right], \tag{S5}$$

where $B$ is the minibatch size, $\alpha_+$ controls the contribution from positive ions and $\gamma_-$ is the focusing exponent for negative positions.

For Ion Count, the relative under-count error is defined as:

$$u_i = \frac{[n_i - \widehat{n}_i]_+}{\max(n_i, 1)}, \tag{S6}$$

where $n_i$ is the reference number of distinct product ions, $\widehat{n}_i$ is the predicted count and $[x]_+ = \max(x, 0)$ denotes the positive part of $x$. Relative scaling makes $u_i$ comparable across molecules with different ion counts and sets it to zero when the count is not underestimated.

The relative over-count error is defined as:

$$o_i = \frac{[\widehat{n}_i - n_i]_+}{\max(n_i, 1)}. \tag{S7}$$

Conversely, $o_i$ measures the relative excess in the predicted count and becomes zero when the count is not overestimated.

The asymmetric count penalty for molecule $i$ is defined as:

$$\ell_{\text{count},i} = \lambda_{\text{count}} \left( \alpha_{\text{under}} u_i^2 + \alpha_{\text{over}} o_i^2 \right), \tag{S8}$$

where $\lambda_{\text{count}}$ controls the overall count contribution, and $\alpha_{\text{under}}$ and $\alpha_{\text{over}}$ set the respective costs of under-counting and over-counting. Separate coefficients allow missed ions to receive a different penalty from excess predictions.

The Ion Count loss, including the count coefficient, is defined as:

$$\mathcal{L}_{\text{count}} = \frac{1}{B} \sum_{i=1}^{B} \ell_{\text{count},i}. \tag{S9}$$

For Tree Similarity, $\mathcal{I}_i$ and $\mathcal{T}_i$ denote the sets of reference product ions and canonical fragmentation triplets for molecule $i$, respectively, and $J$ denotes Jaccard similarity. The target similarity is $s_{ij} = 0.70J(\mathcal{I}_i, \mathcal{I}_j) + 0.30J(\mathcal{T}_i, \mathcal{T}_j)$. The predicted similarity is $\widehat{s}_{ij} = ((\mathbf{e}_i^{\text{sim}})^{\mathsf{T}} \mathbf{e}_j^{\text{sim}} + 1)/2$, where $i$ and $j$ index molecules and $\mathbf{e}_i^{\text{sim}}$ is the $L_2$-normalized 256-dimensional projection for molecule $i$. The Tree Similarity loss is defined as:

$$\mathcal{L}_{\text{sim}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j) \in \mathcal{P}} \rho_{\delta_{\text{sim}}}(\widehat{s}_{ij} - s_{ij}), \tag{S10}$$

where $\mathcal{P}$ is the set of non-identical molecule pairs in the minibatch and $\rho_{\delta_{\text{sim}}}$ is the Smooth $L_1$ penalty with transition parameter $\delta_{\text{sim}} = 0.05$.

The Smooth $L_1$ penalty used in the Tree Similarity loss is a scaled form of the Huber penalty [3] and is defined as:

$$\rho_\delta(r) = \begin{cases} \dfrac{r^2}{2\delta}, & |r| < \delta, \\ |r| - \dfrac{\delta}{2}, & |r| \geq \delta. \end{cases} \tag{S11}$$

where $r$ is the pairwise similarity residual and $\delta > 0$ is the transition threshold. Quadratic behavior near zero supports smooth optimization, whereas linear growth for larger residuals reduces sensitivity to outlying molecule pairs.

All numerical coefficients and epoch schedules are reported in Supplementary Table 22. Gradients from the three heads update both the parameters of each head and the shared LoRA representation.

## Supplementary Note 5. Recall Repair implementation and diagnostic candidates

When a training record is selected for structured ion masking, each removed product ion is deleted together with all incident relations. A molecular fragment is then pruned if it no longer supports a retained fragment-to-ion relation. The complete tree remains the target for the shared Stage 1 model and all auxiliary heads.

At inference, candidate masses already present in the first-pass tree are removed before self-review. The model performs one review pass only. Its output is normalized to source-mechanism-product relations, united with the first-pass relations and rebuilt into a canonical tree. This relation-level no-delete merge prevents the review pass from removing a supported first-pass relation merely by omitting it from the reviewed sequence.

The candidate-accessibility analysis uses the diagnostic set $\mathcal{U}_i = \mathcal{I}_i^{\mathrm{rep}} \cup \mathcal{C}_i^{\mathrm{aux}}$. This union is used only to determine whether a reference mass is accessible to generation or auxiliary nomination. A candidate-only mass is not treated as a mechanism-resolved relation and does not enter the Stage 2 pathway set or the reported Product-ion Recall.

## Supplementary Note 6. Pathway construction, encoding and Stage 2 implementation

Stage 2 uses one deterministic pathway from the molecular root to each retained product ion. Relations are traversed in canonical order, and the first valid parent discovered for a reachable ion is retained. The pathway is reconstructed by backtracking parent relations to the root. If an ion in an imperfect generated tree cannot be connected to the root, its first valid relation is retained as an orphan element preceded by the root; the ion is therefore not discarded silently. Candidate ions are ordered by descending nominal $m/z$. A molecule retains at most 320 candidates, and a pathway retains at most 128 elements. For a longer pathway, the root and the most recent downstream elements are retained.

Textual fields are normalized in separate namespaces for the molecular root, molecular fragments, mechanisms, ion formulas and products. Each lexical unit is assigned deterministically to a 65,536-entry embedding vocabulary by an MD5 hash; index zero is reserved for padding, and at most 48 lexical units are retained per field. For pathway element $j$ of ion $i$, the numerical feature vector is

$$f_{ij}^{\mathrm{path}} = \left[\frac{z_{ij}^{p}}{1000}, \frac{z_{ij}^{s}}{1000}, \frac{z_{ij}^{s}-z_{ij}^{p}}{1000}, \frac{M_i}{1000}, \frac{\log(1+z_{ij}^{p})}{10}, \frac{\log(1+z_{ij}^{s})}{10}, \frac{z_{ij}^{p}}{M_i^{+}}, \frac{z_{ij}^{s}}{M_i^{+}}, \frac{z_{ij}^{s}-z_{ij}^{p}}{M_i^{+}}\right], \tag{S12}$$

where $z_{ij}^{p}$ and $z_{ij}^{s}$ denote the product and source nominal masses for pathway element $j$, respectively, $M_i$ denotes the molecular mass, and $M_i^{+} = \max(M_i, 1)$ prevents division by zero.

The pathway-element representation is the sum of textual, positional, relation-type, mechanism and projected numerical embeddings. The initial ion query combines product identity with $[z_i/1000, M_i/1000, \log(1+z_i)/10, z_i/M_i^{+}]$. Pathway padding is masked in Path Cross-Attention, and padded candidate positions are masked in Inter-Peak Interaction.

The candidate-level regression weight used in the main-text intensity loss is defined as:

$$w_{mi} = 1 + 2y_{mi}^{1.4} + 1.5\mathbb{I}(y_{mi} \geq 0.30) + 3\mathbb{I}(i = b_m), \tag{S13}$$

where $w_{mi}$ is the regression weight for candidate ion $i$ in molecule $m$, $y_{mi}$ is its experimental base-peak-normalized intensity, $b_m$ is the experimental base-peak index and $\mathbb{I}(\cdot)$ is the indicator function. This weighting retains low-intensity candidates while increasing the contribution of abundant ions and the experimental base peak. Other Stage 2 architecture and optimization settings not specified in the main text are reported in Supplementary Tables 23 and 24.

## Supplementary Note 7. Evaluation metrics and statistical aggregation

**Evaluation conventions.** All methods are evaluated on the same molecule-level test partitions. For molecule $i$, $\mathcal{G}_i$ denotes the set of filtered reference product-ion nominal $m/z$ values, and $\mathcal{P}_i$ denotes the final predicted product-ion set for the condition being evaluated. For FragTreeMS, $\mathcal{P}_i = \mathcal{I}_i^{\text{rep}}$, corresponding to the ions retained in the repaired fragmentation tree. Duplicate annotations at the same nominal mass are collapsed before scoring, and ions are matched by exact nominal $m/z$. The diagnostic candidate-union set is excluded from the final product-ion and spectrum metrics. The metrics that require additional specification are defined below.

**Product-ion F2.** Product-ion F2 combines product-ion precision and recall while assigning four times more weight to recall. The molecule-level score is defined as:

$$F_{2,i} = \begin{cases} \dfrac{5p_i r_i}{4p_i + r_i}, & 4p_i + r_i > 0, \\ 0, & \text{otherwise}, \end{cases} \tag{S14}$$

where $p_i = |\mathcal{P}_i \cap \mathcal{G}_i| / \max(1, |\mathcal{P}_i|)$ is the product-ion precision of molecule $i$, $r_i = |\mathcal{P}_i \cap \mathcal{G}_i| / \max(1, |\mathcal{G}_i|)$ is its product-ion recall, $\mathcal{P}_i$ is the predicted product-ion set and $\mathcal{G}_i$ is the filtered reference product-ion set.

The reported Product-ion F2 is the macro-average across all evaluated molecules:

$$\text{Product-ion F2} = \frac{1}{N} \sum_{i=1}^{N} F_{2,i}, \tag{S15}$$

where $N$ is the number of evaluated molecules.

**Strong-ion Recall.** Strong-ion Recall measures whether the final predicted ion set recovers the

more intense reference ions. Strong reference ions are selected on the original NIST intensity scale before base-peak normalization, and no additional threshold is applied to their predicted intensities:

$$\text{Strong-ion Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\mathcal{P}_i \cap \mathcal{S}_i|}{\max(1, |\mathcal{S}_i|)}, \tag{S16}$$

where $\mathcal{S}_i = \{z \in \mathcal{G}_i : I_{iz}^{\text{NIST}} \geq 100\}$ is the set of strong reference ions for molecule $i$, and $I_{iz}^{\text{NIST}}$ is the unnormalized NIST intensity at nominal mass $z$.

**First-Pass Recall.** First-Pass Recall measures reference-ion recovery after the first deterministic Stage 1 decoding pass and before candidate guidance and self-review:

$$\text{First-Pass Recall} = \frac{1}{N}\sum_{i=1}^{N} R_i\left(\mathcal{I}_i^{(0)}\right), \tag{S17}$$

where $\mathcal{I}_i^{(0)}$ is the product-ion set generated in the first pass and $R_i(\mathcal{A}) = |\mathcal{A} \cap \mathcal{G}_i| / \max(1, |\mathcal{G}_i|)$ is the reference-ion recall attained by an arbitrary nominal-mass set $\mathcal{A}$.

**Repair Gain.** Repair Gain isolates the increase in reference-ion recall produced by candidate-guided self-review and repair:

$$\text{Repair Gain} = \frac{1}{N}\sum_{i=1}^{N}\left[R_i(\mathcal{I}_i^{\text{rep}}) - R_i\left(\mathcal{I}_i^{(0)}\right)\right], \tag{S18}$$

where $\mathcal{I}_i^{\text{rep}}$ is the product-ion set retained after self-review and merging, $\mathcal{I}_i^{(0)}$ is the first-pass set, and $R_i(\mathcal{A}) = |\mathcal{A} \cap \mathcal{G}_i| / \max(1, |\mathcal{G}_i|)$. The merge procedure preserves supported first-pass relations, so the gain reflects reference ions newly retained by the repaired tree.

**Candidate-Union Gain.** Candidate-Union Gain quantifies the additional reference-ion recall available in the complete diagnostic candidate union beyond the ions retained in the repaired tree. It represents candidate coverage rather than final prediction performance:

$$\text{Candidate-Union Gain} = \frac{1}{N}\sum_{i=1}^{N}\left[R_i(\mathcal{U}_i) - R_i(\mathcal{I}_i^{\text{rep}})\right], \tag{S19}$$

where $\mathcal{U}_i = \mathcal{I}_i^{\text{rep}} \cup \mathcal{C}_i^{\text{aux}}$, $\mathcal{C}_i^{\text{aux}}$ is the auxiliary candidate set supplied for review, and $R_i(\mathcal{A}) = |\mathcal{A} \cap \mathcal{G}_i| / \max(1, |\mathcal{G}_i|)$. The candidate union is used only for this diagnostic calculation and is not passed to Stage 2 as the predicted ion inventory.

**Peak-intensity Coverage.** Peak-intensity Coverage measures the fraction of the total filtered reference intensity whose nominal masses are represented in the predicted ion set. It uses reference intensities and therefore evaluates intensity-weighted ion recovery rather than the accuracy of the predicted intensities:

$$\text{Peak-intensity Coverage} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{z\in\mathcal{P}_i\cap\mathcal{G}_i} y_{iz}}{\max\left(\varepsilon, \sum_{z\in\mathcal{G}_i} y_{iz}\right)}, \tag{S20}$$

where $y_{iz} \in [0,1]$ is the base-peak-normalized experimental intensity at nominal mass $z$, and $\varepsilon = 10^{-8}$ prevents division by zero. When first-pass coverage is reported, $\mathcal{P}_i$ is replaced by $\mathcal{I}_i^{(0)}$.

**Cos-Pre.** Cos-Pre evaluates intensity agreement over the predicted ion support. A predicted mass that is absent from the filtered reference spectrum is assigned zero reference intensity, so unsupported predicted ions remain penalized:

$$\text{Cos-pre} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{z\in\mathcal{P}_i} y_{iz}\widehat{y}_{iz}}{\max\left(\varepsilon, \sqrt{\sum_{z\in\mathcal{P}_i} y_{iz}^2}\sqrt{\sum_{z\in\mathcal{P}_i} \widehat{y}_{iz}^2}\right)}, \tag{S21}$$

where $y_{iz}$ and $\widehat{y}_{iz}$ are the base-peak-normalized experimental and predicted intensities at nominal mass $z$, respectively. For $z \notin \mathcal{G}_i$, $y_{iz} = 0$. Thus, Cos-Pre evaluates the predicted spectrum on $\mathcal{P}_i$, penalizes unsupported predicted peaks and does not penalize reference peaks outside the predicted ion set. A molecule with an empty predicted spectrum receives a score of zero.

**Cos-All.** Cos-All evaluates intensity agreement over the complete filtered reference ion support. A reference ion that is absent from the predicted spectrum is assigned zero predicted intensity, so missed reference ions are explicitly penalized:

$$\text{Cos-all} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{z\in\mathcal{G}_i} y_{iz}\widehat{y}_{iz}}{\max\left(\varepsilon, \sqrt{\sum_{z\in\mathcal{G}_i} y_{iz}^2}\sqrt{\sum_{z\in\mathcal{G}_i} \widehat{y}_{iz}^2}\right)}, \tag{S22}$$

where $\widehat{y}_{iz} = 0$ for $z \notin \mathcal{P}_i$. Cos-All therefore penalizes reference ions missed by the predicted ion inventory, whereas predicted masses outside $\mathcal{G}_i$ do not enter this reference-support calculation.

**Base-peak Accuracy.** Base-peak Accuracy is the fraction of molecules for which the nominal mass of the most intense predicted peak matches that of the experimental base peak:

$$\text{Base-peak Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\widehat{b}_i = b_i\right), \tag{S23}$$

where $b_i = \min(\arg\max_{z\in\mathcal{G}_i} y_{iz})$ and $\widehat{b}_i = \min(\arg\max_{z\in\mathcal{P}_i} \widehat{y}_{iz})$ are the experimental and predicted base-peak masses, respectively, $\widehat{y}_{iz}$ is the predicted normalized intensity, and $\mathbb{I}(\cdot)$ is the indicator function. Selecting the smaller nominal $m/z$ resolves intensity ties; an empty predicted spectrum is scored as incorrect.

**Ion Recall@5.** Ion Recall@5 measures how many masses from the complete filtered reference inventory occur among the five highest-intensity predicted peaks:

$$\text{Ion Recall@5} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,\text{Top5}_i(\widehat{y}) \cap \mathcal{G}_i|}{\max(1, |\mathcal{G}_i|)}, \tag{S24}$$

where $\text{Top5}_i(\widehat{y}) \subseteq \mathcal{P}_i$ contains up to five predicted nominal masses ordered by decreasing predicted intensity, with ties resolved by ascending nominal $m/z$. The denominator is the complete number of filtered reference ions rather than five; consequently, the maximum attainable value is below one when $|\mathcal{G}_i| > 5$.

**Statistical aggregation.** Each metric is calculated for every molecule and macro-averaged within each independent training run. Results from seeds 42, 43 and 44 are summarized using the arithmetic mean and sample standard deviation across runs. Error bars denote this across-run standard deviation unless a figure caption specifies a molecule-level distribution. Box plots show the median, interquartile range and whiskers extending to the most extreme observations within 1.5 times the interquartile range. Methods evaluated in a single deterministic run are reported without an artificial variance estimate. Scores are calculated on the $[0, 1]$ scale and multiplied by 100 when displayed as percentages; differences between percentage-valued metrics are reported in percentage points. The matching rule, strong-ion threshold, base-peak tie rule and numerical stabilizer are also listed in Supplementary Table 24.

## Supplementary Note 8. Baseline and ablation implementation

**Shared prompt design.** GPT-5.5 and DeepSeek-V4-Pro use the same system message and user-prompt templates. Every condition supplies the molecule identifier, name, SMILES, molecular formula and molecular weight. No experimental spectrum or reference intensity is included; only the Gold condition receives curated triplets and their product-ion identities. The system message assigns expertise in 70-eV electron ionization mass spectrometry and requests a single JSON object without Markdown or explanatory text. Product ions must use a parseable formula or chemical name with an explicit nominal $m/z$. Predicted intensities lie in $[0, 1]$, and the strongest predicted peak has intensity 1.0. The prompt favors likely major and medium-intensity signals over an exhaustive list of theoretical fragments. A short `reason` is requested for each predicted peak to make the output inspectable, but this field does not enter any evaluation metric.

**Direct (Base).** The Direct condition, denoted Base in the figures, uses one model call. The user prompt asks the model to predict the 70-eV EI spectrum directly from the molecular information and explicitly prohibits triplet generation. The response contains a single `peaks` array. Each entry contains `mz`, `product_ion`, `intensity` and `reason`. The product-ion string includes the same nominal mass as `mz`, and the predicted intensities are normalized to the strongest returned peak. This condition tests spectrum prediction without an explicit fragmentation representation.

**Triplets.** The Predicted Triplets condition, denoted Triplets in the figures, uses two consecutive model calls. The first prompt asks the model to infer source-mechanism-product triplets from the

molecular information. A source begins with either `smiles_fragment:` followed by a fragment SMILES or `precursor_mz:` followed by an integer; the product ion contains an explicit nominal $m/z$. The prompt suggests the controlled mechanism vocabulary defined in Supplementary Note 2. Isotopic peak is omitted from the suggested vocabulary in the reported baseline runs. The response contains only a `triplets` array, and each entry contains `source`, `mechanism` and `product_ion`. The second prompt supplies these generated triplets together with the same molecular information and asks for normalized peak intensities. It identifies the task as peak prediction rather than triplet generation and requests the same `peaks` schema used in the Direct condition. Product ions supported by the supplied triplets are preferred, although the model may add other chemically plausible ions. The triplets therefore provide structural evidence rather than a closed candidate list.

**Gold.** The Gold Triplets condition, denoted Gold in the figures, bypasses triplet generation and supplies the curated, corrected reference triplets directly to the intensity prompt. The molecular fields, intensity instructions, peak limit and output schema are identical to those used in the second call of the Triplets condition. Experimental peak intensities remain hidden. Extra plausible ions are permitted under the same rule. Gold is an oracle diagnostic for intensity prediction with a correct fragmentation representation and is not a deployable inference setting.

**Interpretation of the three conditions.** Direct removes the intermediate fragmentation representation, Triplets introduces a representation generated by the evaluated model, and Gold replaces it with the curated reference representation. Because GPT-5.5 and DeepSeek-V4-Pro receive identical prompt templates, their comparison is not affected by model-specific prompt wording. The comparison between Direct and Triplets assesses whether an explicit triplet representation improves spectrum prediction, whereas the comparison between Triplets and Gold quantifies the performance gap associated with errors in triplet generation.

## Supplementary Tables

**Supplementary Table 1 |** Full-model performance and baseline comparisons across all datasets: Base-peak accuracy and Cosine. For each seed, metrics are averaged across the four datasets to assess overall performance. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.473672 \pm 0.022425$ | $0.766679 \pm 0.004604$ |
| GPT | GPT-5.5 | Triplets | $0.465017 \pm 0.021807$ | $0.758379 \pm 0.008583$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.439137 \pm 0.011554$ | $0.694744 \pm 0.007960$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.387552 \pm 0.005669$ | $0.661877 \pm 0.005582$ |
| Local | Qwen3-8B | Direct | $0.032728 \pm 0.012365$ | $0.223422 \pm 0.004079$ |
| Local | Qwen3-8B | Triplets | $0.097345 \pm 0.007182$ | $0.173181 \pm 0.024315$ |
| Local | Llama3.1-8B | Direct | $0.063396 \pm 0.044287$ | $0.146684 \pm 0.059439$ |
| Local | Llama3.1-8B | Triplets | $0.108659 \pm 0.002752$ | $0.187644 \pm 0.000904$ |
| Local | Qwen2.5-14B | Direct | $0.031304 \pm 0.000000$ | $0.235105 \pm 0.000334$ |
| Local | Qwen2.5-14B | Triplets | $0.152492 \pm 0.001146$ | $0.248185 \pm 0.000051$ |
| Baseline | NEIMS | – | $0.551889 \pm 0.012243$ | $0.813358 \pm 0.002221$ |
| Baseline | CFM-ID | – | $0.313848 \pm 0.000000$ | $0.588455 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.575272 \pm 0.009958$ | $0.830246 \pm 0.004002$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.582036 \pm 0.012525$ | $0.826379 \pm 0.001219$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.568318 \pm 0.009949$ | $0.837374 \pm 0.004962$ |

**Supplementary Table 2 |** Full-model performance and baseline comparisons across all datasets: Cos-Pre and Cos-All. For each seed, metrics are averaged across the four datasets to assess overall performance. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.813865 \pm 0.003015$ | $0.817509 \pm 0.006797$ |
| GPT | GPT-5.5 | Triplets | $0.807943 \pm 0.007230$ | $0.812128 \pm 0.007785$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.776415 \pm 0.005547$ | $0.747541 \pm 0.004114$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.756361 \pm 0.006271$ | $0.728794 \pm 0.004311$ |
| Local | Qwen3-8B | Direct | $0.441617 \pm 0.016790$ | $0.349234 \pm 0.007841$ |
| Local | Qwen3-8B | Triplets | $0.332652 \pm 0.039276$ | $0.351104 \pm 0.024636$ |
| Local | Llama3.1-8B | Direct | $0.288344 \pm 0.069167$ | $0.218917 \pm 0.059964$ |
| Local | Llama3.1-8B | Triplets | $0.352875 \pm 0.012889$ | $0.302506 \pm 0.004606$ |
| Local | Qwen2.5-14B | Direct | $0.494663 \pm 0.000782$ | $0.341458 \pm 0.000426$ |
| Local | Qwen2.5-14B | Triplets | $0.459654 \pm 0.001450$ | $0.354257 \pm 0.000687$ |
| Baseline | NEIMS | – | $0.827888 \pm 0.000619$ | $0.861711 \pm 0.002850$ |
| Baseline | CFM-ID | – | $0.621136 \pm 0.000000$ | $0.721040 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.860016 \pm 0.001147$ | $0.872072 \pm 0.002572$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.856218 \pm 0.001024$ | $0.875836 \pm 0.001239$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.854368 \pm 0.003852$ | $0.886422 \pm 0.004195$ |

**Supplementary Table 3 |** Full-model performance and baseline comparisons on acyclic alkanes: Base-peak accuracy and Cosine. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.653333 \pm 0.050332$ | $0.877066 \pm 0.002031$ |
| GPT | GPT-5.5 | Triplets | $0.586667 \pm 0.030551$ | $0.881832 \pm 0.004462$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.566667 \pm 0.050332$ | $0.805479 \pm 0.008450$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.500000 \pm 0.034641$ | $0.807021 \pm 0.012732$ |
| Local | Qwen3-8B | Direct | $0.026667 \pm 0.046188$ | $0.415192 \pm 0.003829$ |
| Local | Qwen3-8B | Triplets | $0.000000 \pm 0.000000$ | $0.257756 \pm 0.093169$ |
| Local | Llama3.1-8B | Direct | $0.093333 \pm 0.161658$ | $0.201308 \pm 0.197001$ |
| Local | Llama3.1-8B | Triplets | $0.260000 \pm 0.017321$ | $0.334283 \pm 0.027731$ |
| Local | Qwen2.5-14B | Direct | $0.000000 \pm 0.000000$ | $0.502651 \pm 0.000000$ |
| Local | Qwen2.5-14B | Triplets | $0.500000 \pm 0.000000$ | $0.599656 \pm 0.000000$ |
| Baseline | NEIMS | – | $0.666667 \pm 0.011547$ | $0.905453 \pm 0.004098$ |
| Baseline | CFM-ID | – | $0.240000 \pm 0.000000$ | $0.583563 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.733333 \pm 0.011547$ | $0.935911 \pm 0.008254$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.766667 \pm 0.030551$ | $0.936754 \pm 0.003339$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.711565 \pm 0.028306$ | $0.944392 \pm 0.009138$ |

**Supplementary Table 4 |** Full-model performance and baseline comparisons on alkenes: Base-peak accuracy and Cosine. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.368159 \pm 0.034469$ | $0.760805 \pm 0.000566$ |
| GPT | GPT-5.5 | Triplets | $0.402985 \pm 0.039489$ | $0.759233 \pm 0.002022$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.358209 \pm 0.000000$ | $0.682056 \pm 0.002779$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.363184 \pm 0.017234$ | $0.662679 \pm 0.011477$ |
| Local | Qwen3-8B | Direct | $0.029851 \pm 0.000000$ | $0.234291 \pm 0.003817$ |
| Local | Qwen3-8B | Triplets | $0.019900 \pm 0.008617$ | $0.133116 \pm 0.007760$ |
| Local | Llama3.1-8B | Direct | $0.064677 \pm 0.008617$ | $0.172239 \pm 0.021574$ |
| Local | Llama3.1-8B | Triplets | $0.049751 \pm 0.017234$ | $0.140690 \pm 0.004487$ |
| Local | Qwen2.5-14B | Direct | $0.044776 \pm 0.000000$ | $0.217778 \pm 0.000000$ |
| Local | Qwen2.5-14B | Triplets | $0.059701 \pm 0.000000$ | $0.226599 \pm 0.000000$ |
| Baseline | NEIMS | – | $0.562189 \pm 0.017234$ | $0.838366 \pm 0.001162$ |
| Baseline | CFM-ID | – | $0.194030 \pm 0.000000$ | $0.616345 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.532338 \pm 0.037561$ | $0.847261 \pm 0.012049$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.552239 \pm 0.025852$ | $0.839617 \pm 0.002956$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.537313 \pm 0.000000$ | $0.845927 \pm 0.007391$ |

**Supplementary Table 5 |** Full-model performance and baseline comparisons on aromatic hydrocarbons: Base-peak accuracy and Cosine. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.703704 \pm 0.009164$ | $0.754877 \pm 0.012988$ |
| GPT | GPT-5.5 | Triplets | $0.661376 \pm 0.012123$ | $0.734967 \pm 0.016204$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.566138 \pm 0.012123$ | $0.673157 \pm 0.009090$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.449735 \pm 0.025512$ | $0.611009 \pm 0.007435$ |
| Local | Qwen3-8B | Direct | $0.057445 \pm 0.003273$ | $0.137006 \pm 0.019784$ |
| Local | Qwen3-8B | Triplets | $0.329932 \pm 0.047131$ | $0.228260 \pm 0.014406$ |
| Local | Llama3.1-8B | Direct | $0.027778 \pm 0.006873$ | $0.079085 \pm 0.019434$ |
| Local | Llama3.1-8B | Triplets | $0.062736 \pm 0.001309$ | $0.116508 \pm 0.013139$ |
| Local | Qwen2.5-14B | Direct | $0.063492 \pm 0.000000$ | $0.103497 \pm 0.001336$ |
| Local | Qwen2.5-14B | Triplets | $0.050265 \pm 0.004582$ | $0.079425 \pm 0.000206$ |
| Baseline | NEIMS | – | $0.605820 \pm 0.009164$ | $0.709330 \pm 0.010022$ |
| Baseline | CFM-ID | – | $0.634921 \pm 0.000000$ | $0.665174 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.628635 \pm 0.012563$ | $0.729529 \pm 0.004474$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.613757 \pm 0.009164$ | $0.722205 \pm 0.007924$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.640212 \pm 0.012123$ | $0.746765 \pm 0.007939$ |

**Supplementary Table 6 |** Full-model performance and baseline comparisons on cycloalkanes: Base-peak accuracy and Cosine. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.169492 \pm 0.000000$ | $0.673970 \pm 0.005540$ |
| GPT | GPT-5.5 | Triplets | $0.209040 \pm 0.019571$ | $0.657484 \pm 0.018394$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.265537 \pm 0.009786$ | $0.618284 \pm 0.015788$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.237288 \pm 0.033898$ | $0.566799 \pm 0.005833$ |
| Local | Qwen3-8B | Direct | $0.016949 \pm 0.000000$ | $0.107198 \pm 0.004178$ |
| Local | Qwen3-8B | Triplets | $0.039548 \pm 0.009786$ | $0.073590 \pm 0.002555$ |
| Local | Llama3.1-8B | Direct | $0.067797 \pm 0.000000$ | $0.134104 \pm 0.000252$ |
| Local | Llama3.1-8B | Triplets | $0.062147 \pm 0.009786$ | $0.159093 \pm 0.013720$ |
| Local | Qwen2.5-14B | Direct | $0.016949 \pm 0.000000$ | $0.116494 \pm 0.000000$ |
| Local | Qwen2.5-14B | Triplets | $0.000000 \pm 0.000000$ | $0.087060 \pm 0.000000$ |
| Baseline | NEIMS | – | $0.372881 \pm 0.029357$ | $0.800285 \pm 0.008179$ |
| Baseline | CFM-ID | – | $0.186441 \pm 0.000000$ | $0.488737 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.406780 \pm 0.016949$ | $0.808281 \pm 0.006271$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.395480 \pm 0.048928$ | $0.806941 \pm 0.002157$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.384181 \pm 0.009786$ | $0.812412 \pm 0.010162$ |

**Supplementary Table 7 |** Full-model performance and baseline comparisons on acyclic alkanes: Cos-Pre and Cos-All. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.900019 \pm 0.005660$ | $0.904769 \pm 0.004067$ |
| GPT | GPT-5.5 | Triplets | $0.898011 \pm 0.004630$ | $0.910420 \pm 0.002490$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.857485 \pm 0.009998$ | $0.836424 \pm 0.005319$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.872981 \pm 0.009362$ | $0.837186 \pm 0.008228$ |
| Local | Qwen3-8B | Direct | $0.576999 \pm 0.099822$ | $0.569467 \pm 0.046960$ |
| Local | Qwen3-8B | Triplets | $0.406098 \pm 0.174560$ | $0.494728 \pm 0.106325$ |
| Local | Llama3.1-8B | Direct | $0.245045 \pm 0.222805$ | $0.284288 \pm 0.178406$ |
| Local | Llama3.1-8B | Triplets | $0.461095 \pm 0.000854$ | $0.507916 \pm 0.061815$ |
| Local | Qwen2.5-14B | Direct | $0.618964 \pm 0.000000$ | $0.710238 \pm 0.000000$ |
| Local | Qwen2.5-14B | Triplets | $0.809634 \pm 0.000000$ | $0.670017 \pm 0.000000$ |
| Baseline | NEIMS | – | $0.908635 \pm 0.004037$ | $0.928007 \pm 0.003679$ |
| Baseline | CFM-ID | – | $0.594277 \pm 0.000000$ | $0.730301 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.956745 \pm 0.003375$ | $0.951586 \pm 0.008948$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.953551 \pm 0.003280$ | $0.954572 \pm 0.001681$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.955055 \pm 0.003640$ | $0.959452 \pm 0.007731$ |

**Supplementary Table 8 |** Full-model performance and baseline comparisons on alkenes: Cos-Pre and Cos-All. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.807569 \pm 0.001829$ | $0.797967 \pm 0.002575$ |
| GPT | GPT-5.5 | Triplets | $0.809937 \pm 0.003510$ | $0.798937 \pm 0.000670$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.768269 \pm 0.001479$ | $0.728014 \pm 0.001282$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.765351 \pm 0.009900$ | $0.717120 \pm 0.006880$ |
| Local | Qwen3-8B | Direct | $0.489641 \pm 0.009738$ | $0.364202 \pm 0.011917$ |
| Local | Qwen3-8B | Triplets | $0.333149 \pm 0.001329$ | $0.277518 \pm 0.019893$ |
| Local | Llama3.1-8B | Direct | $0.279166 \pm 0.026149$ | $0.286707 \pm 0.029272$ |
| Local | Llama3.1-8B | Triplets | $0.280360 \pm 0.001526$ | $0.227689 \pm 0.017402$ |
| Local | Qwen2.5-14B | Direct | $0.532967 \pm 0.000000$ | $0.335687 \pm 0.000000$ |
| Local | Qwen2.5-14B | Triplets | $0.437472 \pm 0.000000$ | $0.413227 \pm 0.000000$ |
| Baseline | NEIMS | – | $0.843673 \pm 0.001640$ | $0.875260 \pm 0.001664$ |
| Baseline | CFM-ID | – | $0.636876 \pm 0.000000$ | $0.705785 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.866492 \pm 0.008746$ | $0.872890 \pm 0.009462$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.859523 \pm 0.006373$ | $0.862195 \pm 0.005750$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.860304 \pm 0.001586$ | $0.872613 \pm 0.008596$ |

**Supplementary Table 9 |** Full-model performance and baseline comparisons on aromatic hydrocarbons: Cos-Pre and Cos-All. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.793457 \pm 0.002124$ | $0.832590 \pm 0.015173$ |
| GPT | GPT-5.5 | Triplets | $0.787511 \pm 0.007799$ | $0.817280 \pm 0.012710$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.745499 \pm 0.009245$ | $0.754184 \pm 0.002188$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.678067 \pm 0.009897$ | $0.727925 \pm 0.007497$ |
| Local | Qwen3-8B | Direct | $0.328831 \pm 0.047519$ | $0.286872 \pm 0.019631$ |
| Local | Qwen3-8B | Triplets | $0.395031 \pm 0.012312$ | $0.420562 \pm 0.033313$ |
| Local | Llama3.1-8B | Direct | $0.253969 \pm 0.030113$ | $0.113487 \pm 0.028109$ |
| Local | Llama3.1-8B | Triplets | $0.268795 \pm 0.032174$ | $0.230270 \pm 0.013266$ |
| Local | Qwen2.5-14B | Direct | $0.348728 \pm 0.003126$ | $0.148362 \pm 0.001703$ |
| Local | Qwen2.5-14B | Triplets | $0.311416 \pm 0.005802$ | $0.140915 \pm 0.002749$ |
| Baseline | NEIMS | – | $0.744570 \pm 0.010938$ | $0.793676 \pm 0.006708$ |
| Baseline | CFM-ID | – | $0.707704 \pm 0.000000$ | $0.794714 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.785627 \pm 0.005821$ | $0.808642 \pm 0.004375$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.775949 \pm 0.002935$ | $0.830335 \pm 0.000002$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.778389 \pm 0.004405$ | $0.842908 \pm 0.009425$ |

**Supplementary Table 10 |** Full-model performance and baseline comparisons on cycloalkanes: Cos-Pre and Cos-All. FragTreeMS uses the complete training pipeline. LLM baselines use the Direct and Triplets settings. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Direct | $0.754417 \pm 0.007342$ | $0.734709 \pm 0.006666$ |
| GPT | GPT-5.5 | Triplets | $0.736312 \pm 0.021893$ | $0.721876 \pm 0.016935$ |
| DeepSeek | DeepSeek-V4-Pro | Direct | $0.734409 \pm 0.009234$ | $0.671543 \pm 0.015782$ |
| DeepSeek | DeepSeek-V4-Pro | Triplets | $0.709045 \pm 0.015754$ | $0.632944 \pm 0.003152$ |
| Local | Qwen3-8B | Direct | $0.370999 \pm 0.005120$ | $0.176398 \pm 0.007883$ |
| Local | Qwen3-8B | Triplets | $0.196329 \pm 0.006473$ | $0.211608 \pm 0.021203$ |
| Local | Llama3.1-8B | Direct | $0.375196 \pm 0.002397$ | $0.191185 \pm 0.004068$ |
| Local | Llama3.1-8B | Triplets | $0.401249 \pm 0.018712$ | $0.244152 \pm 0.012722$ |
| Local | Qwen2.5-14B | Direct | $0.477993 \pm 0.000000$ | $0.171544 \pm 0.000000$ |
| Local | Qwen2.5-14B | Triplets | $0.280096 \pm 0.000000$ | $0.192867 \pm 0.000000$ |
| Baseline | NEIMS | – | $0.814672 \pm 0.006767$ | $0.849900 \pm 0.008116$ |
| Baseline | CFM-ID | – | $0.545687 \pm 0.000000$ | $0.653362 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Full model | $0.831200 \pm 0.003447$ | $0.855170 \pm 0.013863$ |
| FragTreeMS | Llama3.1-8B | Full model | $0.835849 \pm 0.002481$ | $0.856242 \pm 0.002156$ |
| FragTreeMS | Qwen2.5-14B | Full model | $0.823722 \pm 0.009004$ | $0.870716 \pm 0.006977$ |

**Supplementary Table 11 |** Performance under the Gold condition on acyclic alkanes: Base-peak accuracy and Cosine. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.586667 \pm 0.030551$ | $0.907552 \pm 0.005809$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.500000 \pm 0.060000$ | $0.872992 \pm 0.007013$ |
| Local | Qwen3-8B | Gold | $0.000000 \pm 0.000000$ | $0.583401 \pm 0.013157$ |
| Local | Llama3.1-8B | Gold | $0.013333 \pm 0.023094$ | $0.472441 \pm 0.072576$ |
| Local | Qwen2.5-14B | Gold | $0.250000 \pm 0.000000$ | $0.643020 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Gold | $0.740000 \pm 0.020000$ | $0.975773 \pm 0.001273$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.773333 \pm 0.023094$ | $0.974245 \pm 0.001907$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.720000 \pm 0.000000$ | $0.973532 \pm 0.001840$ |

**Supplementary Table 12 |** Performance under the Gold condition on alkenes: Base-peak accuracy and Cosine. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.373134 \pm 0.059701$ | $0.841573 \pm 0.005723$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.398010 \pm 0.034469$ | $0.799187 \pm 0.023033$ |
| Local | Qwen3-8B | Gold | $0.019900 \pm 0.008617$ | $0.608221 \pm 0.001916$ |
| Local | Llama3.1-8B | Gold | $0.039801 \pm 0.008617$ | $0.337617 \pm 0.006043$ |
| Local | Qwen2.5-14B | Gold | $0.044776 \pm 0.000000$ | $0.549018 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Gold | $0.611940 \pm 0.000000$ | $0.900984 \pm 0.005902$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.582090 \pm 0.051703$ | $0.896190 \pm 0.000389$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.582090 \pm 0.029851$ | $0.892653 \pm 0.002237$ |

**Supplementary Table 13 |** Performance under the Gold condition on aromatic hydrocarbons: Base-peak accuracy and Cosine. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.687831 \pm 0.025512$ | $0.835571 \pm 0.010394$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.507937 \pm 0.027493$ | $0.734125 \pm 0.004700$ |
| Local | Qwen3-8B | Gold | $0.411754 \pm 0.042876$ | $0.705069 \pm 0.014641$ |
| Local | Llama3.1-8B | Gold | $0.217687 \pm 0.005891$ | $0.435684 \pm 0.005837$ |
| Local | Qwen2.5-14B | Gold | $0.362434 \pm 0.009164$ | $0.651241 \pm 0.001525$ |
| FragTreeMS | Qwen3-8B | Gold | $0.738095 \pm 0.020998$ | $0.882536 \pm 0.003739$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.730159 \pm 0.027493$ | $0.879109 \pm 0.009385$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.746032 \pm 0.007937$ | $0.886839 \pm 0.010437$ |

**Supplementary Table 14 |** Performance under the Gold condition on cycloalkanes: Base-peak accuracy and Cosine. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.220339 \pm 0.016949$ | $0.794910 \pm 0.005126$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.237288 \pm 0.000000$ | $0.761332 \pm 0.016971$ |
| Local | Qwen3-8B | Gold | $0.067797 \pm 0.000000$ | $0.551080 \pm 0.006767$ |
| Local | Llama3.1-8B | Gold | $0.033898 \pm 0.000000$ | $0.361272 \pm 0.000050$ |
| Local | Qwen2.5-14B | Gold | $0.067797 \pm 0.000000$ | $0.483344 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Gold | $0.435028 \pm 0.025890$ | $0.878528 \pm 0.007217$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.406780 \pm 0.000000$ | $0.885930 \pm 0.004348$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.418079 \pm 0.009786$ | $0.876228 \pm 0.008558$ |

**Supplementary Table 15 |** Performance under the Gold condition on acyclic alkanes: Cos-Pre and Cos-All. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.907611 \pm 0.005779$ | $0.917947 \pm 0.002656$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.873565 \pm 0.006270$ | $0.881885 \pm 0.007070$ |
| Local | Qwen3-8B | Gold | $0.584583 \pm 0.013495$ | $0.592121 \pm 0.020709$ |
| Local | Llama3.1-8B | Gold | $0.484519 \pm 0.074182$ | $0.556920 \pm 0.106377$ |
| Local | Qwen2.5-14B | Gold | $0.712375 \pm 0.000000$ | $0.706381 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Gold | $0.975773 \pm 0.001273$ | $0.975773 \pm 0.001273$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.974245 \pm 0.001907$ | $0.974245 \pm 0.001907$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.973532 \pm 0.001840$ | $0.973532 \pm 0.001840$ |

**Supplementary Table 16 |** Performance under the Gold condition on alkenes: Cos-Pre and Cos-All. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.842558 \pm 0.005933$ | $0.848489 \pm 0.005690$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.800391 \pm 0.022409$ | $0.801525 \pm 0.023457$ |
| Local | Qwen3-8B | Gold | $0.608309 \pm 0.001916$ | $0.608221 \pm 0.001916$ |
| Local | Llama3.1-8B | Gold | $0.397542 \pm 0.008573$ | $0.361955 \pm 0.006051$ |
| Local | Qwen2.5-14B | Gold | $0.584675 \pm 0.000000$ | $0.552947 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Gold | $0.900984 \pm 0.005902$ | $0.900984 \pm 0.005902$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.896190 \pm 0.000389$ | $0.896190 \pm 0.000389$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.892653 \pm 0.002237$ | $0.892653 \pm 0.002237$ |

**Supplementary Table 17 |** Performance under the Gold condition on aromatic hydrocarbons: Cos-Pre and Cos-All. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.838570 \pm 0.011743$ | $0.856746 \pm 0.008913$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.736613 \pm 0.004875$ | $0.758293 \pm 0.005575$ |
| Local | Qwen3-8B | Gold | $0.708598 \pm 0.011512$ | $0.711712 \pm 0.015729$ |
| Local | Llama3.1-8B | Gold | $0.473186 \pm 0.012888$ | $0.447362 \pm 0.008359$ |
| Local | Qwen2.5-14B | Gold | $0.679767 \pm 0.002377$ | $0.674060 \pm 0.002442$ |
| FragTreeMS | Qwen3-8B | Gold | $0.882536 \pm 0.003739$ | $0.882536 \pm 0.003739$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.879109 \pm 0.009385$ | $0.879109 \pm 0.009385$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.886839 \pm 0.010437$ | $0.886839 \pm 0.010437$ |

**Supplementary Table 18 |** Performance under the Gold condition on cycloalkanes: Cos-Pre and Cos-All. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.800953 \pm 0.004429$ | $0.805455 \pm 0.005912$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.767509 \pm 0.017391$ | $0.768866 \pm 0.014151$ |
| Local | Qwen3-8B | Gold | $0.553690 \pm 0.005706$ | $0.552082 \pm 0.008504$ |
| Local | Llama3.1-8B | Gold | $0.428370 \pm 0.003947$ | $0.367821 \pm 0.000561$ |
| Local | Qwen2.5-14B | Gold | $0.602305 \pm 0.000000$ | $0.483592 \pm 0.000000$ |
| FragTreeMS | Qwen3-8B | Gold | $0.878528 \pm 0.007217$ | $0.878528 \pm 0.007217$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.885930 \pm 0.004348$ | $0.885930 \pm 0.004348$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.876228 \pm 0.008558$ | $0.876228 \pm 0.008558$ |

**Supplementary Table 19 |** Performance under the Gold condition across all datasets: Base-peak accuracy and Cosine. For each seed, metrics are averaged across the four datasets to assess overall performance. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Base-peak accuracy | Cosine |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.466993 \pm 0.017155$ | $0.844901 \pm 0.000869$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.410809 \pm 0.010426$ | $0.791909 \pm 0.007738$ |
| Local | Qwen3-8B | Gold | $0.124863 \pm 0.008565$ | $0.611943 \pm 0.004779$ |
| Local | Llama3.1-8B | Gold | $0.076180 \pm 0.005092$ | $0.401754 \pm 0.018183$ |
| Local | Qwen2.5-14B | Gold | $0.181252 \pm 0.002291$ | $0.581656 \pm 0.000381$ |
| FragTreeMS | Qwen3-8B | Gold | $0.631266 \pm 0.005342$ | $0.909455 \pm 0.001807$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.623090 \pm 0.016382$ | $0.908869 \pm 0.002640$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.616550 \pm 0.006000$ | $0.907313 \pm 0.004902$ |

**Supplementary Table 20 |** Performance under the Gold condition across all datasets: Cos-Pre and Cos-All. For each seed, metrics are averaged across the four datasets to assess overall performance. Values are mean ± s.d. across seeds 42, 43 and 44.

| System | Model | Setting | Cos-Pre | Cos-All |
|---|---|---|---|---|
| GPT | GPT-5.5 | Gold | $0.847423 \pm 0.000921$ | $0.857159 \pm 0.001208$ |
| DeepSeek | DeepSeek-V4-Pro | Gold | $0.794520 \pm 0.007437$ | $0.802642 \pm 0.008135$ |
| Local | Qwen3-8B | Gold | $0.613795 \pm 0.004346$ | $0.616034 \pm 0.006505$ |
| Local | Llama3.1-8B | Gold | $0.445904 \pm 0.018454$ | $0.433515 \pm 0.025877$ |
| Local | Qwen2.5-14B | Gold | $0.644781 \pm 0.000594$ | $0.604245 \pm 0.000610$ |
| FragTreeMS | Qwen3-8B | Gold | $0.909455 \pm 0.001807$ | $0.909455 \pm 0.001807$ |
| FragTreeMS | Llama3.1-8B | Gold | $0.908869 \pm 0.002640$ | $0.908869 \pm 0.002640$ |
| FragTreeMS | Qwen2.5-14B | Gold | $0.907313 \pm 0.004902$ | $0.907313 \pm 0.004902$ |

**Supplementary Table 21 |** Stage 1 fragmentation-tree generation and repair performance across datasets and backbone models. Results correspond to the complete Stage 1 pipeline. Product-ion Recall measures the recovery of reference product ions, Product-ion F2 assigns greater weight to recall, Strong-ion Recall measures the recovery of high intensity ions, First-Pass Recall measures ion recovery before repair, and Repair Gain is the increase in recall obtained through repair. Values are mean ± sample s.d. across seeds 42, 43 and 44. For All datasets, metrics are first averaged across the four datasets within each seed and then summarized across seeds.

| Model | Product-ion Recall | Product-ion F2 | Strong-ion Recall | First-Pass Recall | Repair Gain |
|---|---|---|---|---|---|
| **Acyclic alkanes** | | | | | |
| Qwen3-8B | $0.881683 \pm 0.008116$ | $0.810880 \pm 0.012776$ | $0.925786 \pm 0.015134$ | $0.768622 \pm 0.009972$ | $0.113061 \pm 0.016402$ |
| Llama3.1-8B | $0.863054 \pm 0.022347$ | $0.810927 \pm 0.014470$ | $0.922013 \pm 0.018130$ | $0.793014 \pm 0.008292$ | $0.070040 \pm 0.014090$ |
| Qwen2.5-14B | $0.911694 \pm 0.013917$ | $0.828682 \pm 0.012752$ | $0.949686 \pm 0.003928$ | $0.782444 \pm 0.017594$ | $0.129250 \pm 0.023109$ |
| **Alkenes** | | | | | |
| Qwen3-8B | $0.871156 \pm 0.010808$ | $0.801842 \pm 0.003087$ | $0.906141 \pm 0.015589$ | $0.779655 \pm 0.012633$ | $0.091501 \pm 0.021472$ |
| Llama3.1-8B | $0.860658 \pm 0.021377$ | $0.810224 \pm 0.008549$ | $0.903438 \pm 0.010952$ | $0.773027 \pm 0.014819$ | $0.087631 \pm 0.036114$ |
| Qwen2.5-14B | $0.900745 \pm 0.008978$ | $0.820455 \pm 0.004773$ | $0.934338 \pm 0.005225$ | $0.770913 \pm 0.011230$ | $0.129832 \pm 0.002767$ |
| **Aromatic hydrocarbons** | | | | | |
| Qwen3-8B | $0.745935 \pm 0.019489$ | $0.560120 \pm 0.028120$ | $0.774914 \pm 0.017261$ | $0.599310 \pm 0.032028$ | $0.146626 \pm 0.012728$ |
| Llama3.1-8B | $0.745847 \pm 0.003577$ | $0.599583 \pm 0.004819$ | $0.763574 \pm 0.010326$ | $0.644263 \pm 0.018262$ | $0.101584 \pm 0.016831$ |
| Qwen2.5-14B | $0.800899 \pm 0.007388$ | $0.608151 \pm 0.008779$ | $0.824742 \pm 0.008987$ | $0.632763 \pm 0.004990$ | $0.168137 \pm 0.002977$ |
| **Cycloalkanes** | | | | | |
| Qwen3-8B | $0.852992 \pm 0.018072$ | $0.759647 \pm 0.012489$ | $0.858788 \pm 0.013494$ | $0.726903 \pm 0.013262$ | $0.126089 \pm 0.027864$ |
| Llama3.1-8B | $0.830326 \pm 0.004005$ | $0.766014 \pm 0.008380$ | $0.820303 \pm 0.019547$ | $0.764280 \pm 0.013184$ | $0.066046 \pm 0.016678$ |
| Qwen2.5-14B | $0.897973 \pm 0.005904$ | $0.778427 \pm 0.012347$ | $0.898485 \pm 0.009805$ | $0.769057 \pm 0.003574$ | $0.128916 \pm 0.006590$ |
| **All datasets** | | | | | |
| Qwen3-8B | $0.837942 \pm 0.001015$ | $0.733122 \pm 0.013616$ | $0.866407 \pm 0.002266$ | $0.718623 \pm 0.010992$ | $0.119319 \pm 0.011344$ |
| Llama3.1-8B | $0.824971 \pm 0.006172$ | $0.746687 \pm 0.002832$ | $0.852332 \pm 0.002551$ | $0.743646 \pm 0.009320$ | $0.081325 \pm 0.014133$ |
| Qwen2.5-14B | $0.877828 \pm 0.002914$ | $0.758929 \pm 0.004436$ | $0.901813 \pm 0.003524$ | $0.738794 \pm 0.003319$ | $0.139034 \pm 0.004006$ |

**Supplementary Table 22 |** Key Stage 1 generation, auxiliary-objective and optimization parameters.

| Component | Parameter | Value |
|---|---|---|
| Stage 1 LoRA | Rank $r$ | 96 |
| Stage 1 LoRA | Alpha $\alpha$ | 192 |
| Stage 1 LoRA | Dropout probability $p$ | 0.05 |
| Stage 1 LoRA | Target modules | q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj |
| Stage 1 batching | Per-device batch size | 2 |
| Stage 1 batching | Gradient accumulation steps | 4 |
| Stage 1 token limits | Maximum training sequence length | 8192 |
| Stage 1 token limits | Maximum input tokens | 6144 |
| Stage 1 token limits | Maximum generated tokens | 8192 |
| Stage 1 token limits | Minimum generated tokens | 256 |
| Stage 1 generation | Temperature | 0.75 |
| Stage 1 generation | Top-$p$ | 0.95 |
| Stage 1 generation | Top-$k$ | 64 |
| Tree-generation loss | Ordinary target weight | 1.0 |
| Tree-generation loss | Ion-related target weight | 1.5 |
| Ion Inventory | Positive frequency-weight range | 1.0–3.0 |
| Ion Inventory | Positive focal coefficient $\alpha_+$ | 2.0 |
| Ion Inventory | Negative focal coefficient $\gamma_-$ | 4.0 |
| Ion Inventory | Negative margin $m_-$ | 0.05 |
| Ion Inventory | Numerical stabilizer $\varepsilon$ | $10^{-8}$ |
| Ion Count | Under-count penalty $\alpha_{\text{under}}$ | 2.0 |
| Ion Count | Over-count penalty $\alpha_{\text{over}}$ | 0.5 |
| Ion Count | Count-loss coefficient $\lambda_{\text{count}}$ | 0.20 |
| Stage 1 objective | Ion-loss coefficient schedule | 0.5–1.0 |
| Stage 1 objective | Similarity-loss coefficient schedule | 0–0.015–0.03 |
| Stage 1 optimization | AdamW coefficient $\beta_1$ | 0.9 |
| Stage 1 optimization | AdamW coefficient $\beta_2$ | 0.95 |
| Stage 1 optimization | Weight decay | 0.01 |
| Stage 1 optimization | Warm-up ratio | 0.05 |
| Stage 1 optimization | Gradient clipping threshold | 1.0 |
| Stage 1 optimization | Numerical precision | bfloat16 |

**Supplementary Table 23 |** Key Stage 2 pathway-representation and architecture parameters.

| Component | Parameter | Value |
|---|---|---|
| Path representation | Hashed vocabulary size | 65,536 |
| Path representation | Maximum lexical units per field | 48 |
| Path representation | Maximum pathway length | 128 |
| Path representation | Maximum product ions per molecule | 320 |
| Stage 2 architecture | Feed-forward expansion factor | 4 |
| Stage 2 architecture | Dropout probability | 0.08 |

Continued on next page

| Component | Parameter | Value |
|---|---|---|
| Stage 2 architecture | Intensity-head linear layers | 3 |

**Supplementary Table 24 |** Key Stage 2 optimization, evaluation and reproducibility parameters.

| Component | Parameter | Value |
|---|---|---|
| Stage 2 objective | Smooth $L_1$ transition coefficient | 0.04 |
| Stage 2 objective | Curated-tree replay probability | 0.08 |
| Stage 2 batching | Training batch size | 8 |
| Stage 2 batching | Gradient accumulation steps | 1 |
| Stage 2 optimization | AdamW coefficient $\beta_1$ | 0.9 |
| Stage 2 optimization | AdamW coefficient $\beta_2$ | 0.95 |
| Stage 2 optimization | Weight decay | 0.01 |
| Stage 2 optimization | Gradient clipping threshold | 1.0 |
| Stage 2 optimization | Warm-up ratio | 0.06 |
| Stage 2 optimization | Minimum learning-rate ratio | 0.05 |
| Data partition | Fixed partition seed | 42 |
| Repeated training | Independent training seeds | 42, 43, 44 |
| Metric aggregation | Within-run aggregation | Molecule-level macro-average |
| Metric aggregation | Across-run aggregation | Mean and standard deviation |
| Ion matching | Mass alignment | Exact nominal $m/z$ |
| Strong-ion evaluation | Original NIST intensity threshold | 100 |
| Base-peak evaluation | Tie handling | Ascending nominal $m/z$ |
| Spectrum metrics | Numerical stabilizer $\varepsilon$ | $10^{-8}$ |

## Supplementary References


1. Ridnik, T. *et al.* Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 82–91 (2021).
2. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* 2980–2988 (2017).
3. Huber, P. J. in *Breakthroughs in statistics: Methodology and distribution* 492–518 (Springer, 1992).