CLEF: CLEAVAGE EVENTS GUIDE ELECTRON IONIZATION MASS SPECTRUM PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Mass spectrometry is widely used to identify compounds and analyze molecular structures. Predicting electron ionization (EI) mass spectra from structures can extend reference libraries to compounds without measured spectra. Accurate EI spectrum prediction and reliable retrieval among constitutional isomers are complementary goals. Both depend on modeling how molecular connectivity affects fragment intensities. We introduce CLEF, a structure-guided EI spectrum predictor. It uses atom connectivity, bond types, and candidate cleavage events to score fragment formulas. The events are enumerated from the input structure and matched to formulas by elemental composition. Composition-matched attention updates formula representations with event information. These representations then guide atom pooling and formula scoring, allowing the same formula to receive different weights in constitutional isomers. The model is trained on paired structures and spectra without event annotations. CLEF outperforms six trained baselines in spectrum prediction and isomer retrieval on NIST23 and MoNA, and in MoNA-to-SWGDRUG transfer. With only 25% of NIST23 training data, CLEF achieves spectral accuracy comparable to that of a leading full-data predictor. Code: https://anonymous.4open.science/r/CLEF-F368

1 INTRODUCTION

Gas chromatography coupled with electron ionization mass spectrometry (GC/EI–MS) provides a reproducible and cost-effective means of compound identification. Identification commonly relies on matching measured spectra to reference libraries. Direct matching requires a reference spectrum for the target compound to be present in the library. Expanding coverage requires obtaining compound samples and measuring their spectra. Sample availability and experimental costs limit this process. Some rare compounds and compounds controlled because of addiction risks can be particularly costly to obtain and characterize experimentally. These limitations motivate computational and statistical approaches to spectrum prediction (Wei et al., 2019; Kubo et al., 2023).

Existing computational predictors take three broad approaches. Physical and rule-based methods simulate or enumerate fragmentation chemistry (Grimme, 2013; Koopman & Grimme, 2019; Allen et al., 2016); learned predictors map molecular fingerprints or graphs directly to spectral intensities (Wei et al., 2019; Zhang et al., 2022); and structured-output methods score fragment formulas or atom subsets and render their isotope patterns (Zhu & Jonas, 2023). Most of these predictors are trained and evaluated on large annotated commercial libraries such as NIST, with attention focused on several spectral similarity metrics between the experimental and predicted spectra (National Institute of Standards and Technology, 2023).

Yet existing studies face two problems. First, NIST23 is a commercial library that cannot be redistributed (National Institute of Standards and Technology, 2023); open community collections are few and small, and the two larger ones, MoNA and SWGDRUG, together hold only a small fraction of the NIST library (Fiehn Laboratory, n.d.). The need to purchase such commercial data raises the barrier to independent reproduction and retraining, and how these methods perform under limited open labeled data also remains underexamined. Second, spectrum fidelity does not by itself establish identification reliability: constitutional isomers share a molecular formula and often have similar experimental spectra (Figure 1). A predicted spectrum can achieve a high fit score against its own experimental spectrum, yet an even higher one against the experimental spectrum of a same-formula isomer; without retrieval evaluation among isomers, such potential misidentification is not reflected in fit metrics.

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

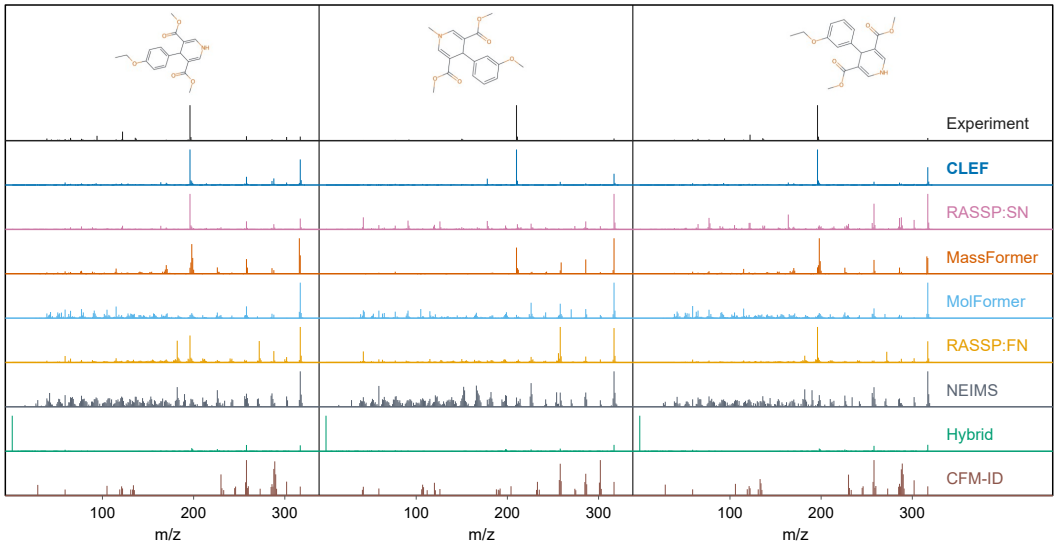

Figure 1: Selected EI spectra for three $C_{17}H_{19}NO_5$ isomers (nominal $M = 317$). Rows show the measurements and the predictions of CLEF and seven baselines (six trained models and the frozen CFM-ID reference). Traces are normalized to their own base peak and shown over m/z 1–330.

To address these gaps, we evaluate EI spectrum predictors along two dimensions. The first is data-efficient generalization: structure–spectrum relationships should extend to unseen molecules and achieve higher spectrum fit metrics with less labeled data. The second is more comprehensive evaluation: while maintaining spectral fidelity, predictions should also preserve the differences that separate same-formula isomers. We accordingly define a retrieval task that is more difficult than standard library search: retrieving the correct structure from a set of same-formula isomers.

We propose CLEF, an EI spectrum predictor that uses composition-matched cleavage events to guide formula scoring. It jointly models atom connectivity and bond types in the molecular graph, cleavage events, and candidate fragment formulas. Candidate events are enumerated deterministically from the input structure and encoded into learned representations. Each formula gathers evidence from composition-matched events to select atom information for its score. Finally, a fixed isotope renderer converts the formula weights into the spectrum. We evaluate the data-efficient generalization of CLEF to unseen molecules and its ability to distinguish constitutional isomers while maintaining spectral fidelity. Our contributions are as follows:

- We develop CLEF, an EI spectrum predictor guided by cleavage events. The model jointly encodes cleavage descriptors and local atom context, and uses composition-matched event evidence to guide atom pooling and fragment-formula scoring without event annotations.
- To assess spectral accuracy and isomer discrimination together, we design same-source and cross-source retrieval protocols for same-formula isomers. Against six trainable baselines, CLEF records the highest means on the four spectrum metrics on MoNA and NIST23 and on the retrieval metrics of both protocols.
- The structure-guided design of CLEF remains effective with limited training data and in cross-library transfer. With 25% of the NIST23 training data, CLEF reaches spectrum-prediction performance close to that of full-data RASSP:SN. Trained on MoNA, CLEF attains the highest means on all four SWGDRUG spectrum metrics among the transferred trainable models.

2 Related Work

We group prior work by regime: GC/EI–MS fragments whole molecules with 70 eV electrons; tandem mass spectrometry (MS/MS) fragments isolated precursor ions by collisions.

GC/EI–MS prediction. Physical simulators reproduce EI fragmentation most directly: QCEIMS and its semiempirical extensions sample fragment distributions by molecular dynamics (Grimme,

2013; Koopman & Grimme, 2019), and QCxMS2 adds reaction-network discovery with quantum-chemical calculations (Gorges & Grimme, 2025); CFM-EI, the EI model of CFM-ID, learns fragmentation probabilities over enumerated fragmentation graphs (Allen et al., 2016). These approaches explicitly model fragmentation pathways but require per-molecule fragmentation computation. RASSP takes a formula-based approach: it scores candidate subformulas or atom subsets and renders their isotope patterns (Zhu & Jonas, 2023). Direct-output predictors fit spectral intensities from structure: NEIMS maps fingerprints to intensities (Wei et al., 2019), graph convolutional models learn atom and bond representations (Zhang et al., 2022; Qu et al., 2024), and Hybrid adds a residual decoder, attention, and chemical masks (Majewski & Łabuda, 2026). MolFormer, a pretrained chemical language model, has been used for limited-data EI prediction (Ross et al., 2022; Muto et al., 2026). We test generalization to unseen structural groups and limited public data. Spectral fit alone does not show whether a predictor preserves differences useful for ranking same-formula candidates. Same-formula isomers can have different spectra even when their candidate fragment formulas are the same. We use possible bond cuts and local atom context to model this difference, and test both spectral fit and isomer retrieval.

Tandem prediction. MassFormer transfers a graph transformer pretrained on DFT-computed HOMO–LUMO gaps to tandem spectra (Young et al., 2024); FIORA predicts fragment contributions from single fragmentation events in local bond neighborhoods (Nowatzky et al., 2025); FraGNNet learns fragment and formula distributions on a fragmentation graph (Young et al., 2025); and SCARF treats the spectrum as a set of subformulae, decoding this set with a prefix tree, atom-type by atom-type, before a second model assigns intensities (Goldman et al., 2023). These designs target collision-induced fragmentation of isolated precursors and do not directly transfer to EI prediction. CLEF uses structure-derived cleavage candidates for EI prediction without simulating reaction trajectories. It combines event and atom representations, using composition-matched cleavage evidence to guide fragment-formula scoring and fixed isotope templates to render spectra.

Prediction for identification. Downstream, identification-oriented systems consume these predicted spectra: pEI extends prediction to a large PubChem collection (Kubo et al., 2023), GCMS-ID adds peak annotation and retention indices (Wakoli et al., 2024), FastEI retrieves a NEIMS-generated library through spectral embeddings (Yang et al., 2023), and CSU-EP aligns experimental and predicted spectra contrastively (Xie et al., 2026). Each pairs a different predictor with its own candidate pools, matching rules, and spectral sources, so identification gains are hard to attribute to the predictor. We hold candidate pools and scoring rules fixed across trainable predictors and evaluate both PubChem candidate search and same-source isomer ranking; MassSpecGym complements this setting for tandem spectra under structural separation (Bushuiev et al., 2024).

3 METHOD

3.1 PROBLEM FORMULATION

Given a compound's molecular structure, the task is to predict its EI mass spectrum at 70 eV. The input is a molecular graph $G = (V, E)$, where V denotes atoms and E typed bonds. The target is the measured spectrum $\mathbf{y} \in \mathbb{R}^{512}_{\geq 0}$, whose entries record ion intensities at mass-to-charge ratios $m/z = 1, \ldots, 512$. We assume singly charged ions and index these bins by nominal ion mass. Given paired structures and spectra, the goal is to learn a mapping $G \mapsto \hat{\mathbf{y}}$ that predicts the spectrum of an unseen molecule.

3.2 ATOM AND EVENT REPRESENTATIONS

Atom environments. Both cut boundaries and formula scoring require information about atoms and their surroundings. In Figure 2(a), each atom i starts with a feature vector $\mathbf{x}_i \in \mathbb{R}^{45}$ encoding properties such as element identity, valence, aromaticity, and hydrogen count. Batch normalization over valid atoms gives the initial states $\mathbf{H}^{(0)} \in \mathbb{R}^{|V| \times 45}$, which stacks one atom state per row; $\mathbf{H}^{(l)}$ denotes the states after l layers.

A stacked relational graph convolutional network (RGCN) (Schlichtkrull et al., 2018) updates each state from the atom itself and its bonded neighbors. Let $\mathbf{A}_r$ be the binary adjacency matrix for bond type $r \in \{1, 2, 3, 4\}$, covering single, double, triple, and aromatic bonds. We use symmetric normalization $\bar{\mathbf{A}}_r = \mathbf{D}^{-1/2}\mathbf{A}_r\mathbf{D}^{-1/2}$ with the total degree $D_{ii} = \sum_{r,k}(\mathbf{A}_r)_{ik}$ across bond types,

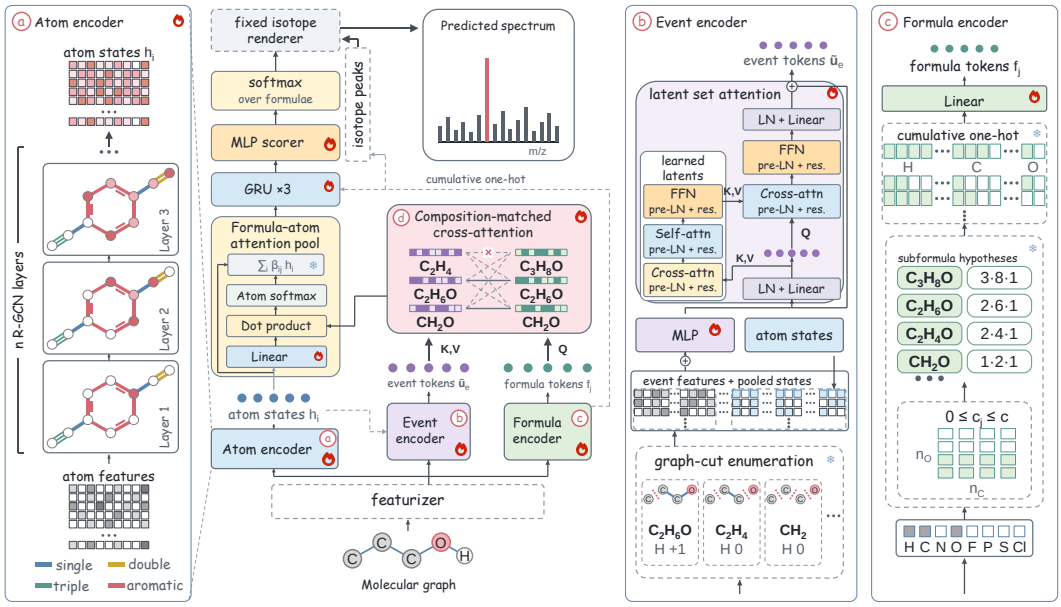

Figure 2: Overview of CLEF. (a) Stacked RGCN layers update atom features into atom states. (b) Graph-cut enumeration supplies cleavage descriptors, combined with pooled atom states and encoded as event tokens. (c) Candidate formulas are encoded as formula tokens. (d) Composition-matched cross-attention updates formula tokens with event evidence; these tokens then guide atom pooling and formula scoring. A fixed isotope renderer converts formula weights into the spectrum. Flames mark learned operations; snowflakes mark fixed operations.

which controls the scale of messages between atoms of different degrees. The update is

$$\mathbf{U}^{(l)} = \mathrm{Dropout}\left[\mathrm{LeakyReLU}\left(\mathbf{H}^{(l)}\mathbf{W}_0^{(l)} + \sum_{r=1}^{4} \bar{\mathbf{A}}_r \mathbf{H}^{(l)} \mathbf{W}_r^{(l)}\right)\right], \tag{1}$$

$$\mathbf{H}^{(l+1)} = \mathrm{LN}\Big(\mathbf{H}^{(l)} + \mathbf{U}^{(l)}\Big), \quad \mathbf{H}^{(1)} = \mathrm{LN}(\mathbf{U}^{(0)}).$$

Layer 0 maps 45 features to width 512 without a residual. Layers 1–15 use the residual update. Here, $\mathbf{W}_0^{(l)}$ and $\mathbf{W}_r^{(l)}$ are learned maps from width d_l to d_{l+1}; $\mathbf{W}_0^{(l)}$ transforms the atom's own state and $\mathbf{W}_r^{(l)}$ its neighbors of bond type r. Separate relation matrices do not constrain different bond types to bond-order-scaled copies of one shared message. After 16 layers, the transpose of row i is the final atom state $\mathbf{h}_i \in \mathbb{R}^{512}$, used by both event encoding and formula scoring.

Graph-cut enumeration. The "graph-cut enumeration" block in Figure 2(b) connects candidate fragments to specific parts of the molecule. We remove one to three heavy-atom bonds and collect the resulting connected components. For each component, we vary its hydrogen count by $\Delta H \in \{-3, -2, -1, 0, 1, 2, 3\}$ and retain compositions within the precursor counts, also adding the molecular-ion candidate with unchanged hydrogen count. Each event e records a component, the number of removed bonds, ΔH, and composition $\mathbf{c}_e^{\mathrm{evt}}$. Different events can share a composition but refer to different components or cut contexts. These are deterministic structural candidates, not simulated reaction trajectories. Their representations and use in formula scoring are learned from spectrum supervision, without manual event annotations. Filtering, deduplication, and capacity limits are specified in Appendix B.

Encoding each event. The "event features + pooled states" block in Figure 2(b) combines fixed cleavage information with learned graph context. Its descriptor $\mathbf{d}_e \in \mathbb{R}^{45}$ records event type, cut count, hydrogen-count change, fragment and complementary compositions, mass, and cut-bond properties. Molecular context is $\bar{\mathbf{h}} = |V|^{-1} \sum_{i \in V} \mathbf{h}_i$, the mean final state of all atoms. Local cut context comes from B_e: at most four unique endpoints of removed bonds crossing the component boundary, selected from both sides in enumeration order. A shared MLP encodes these inputs:

$$\mathbf{u}_e = \mathrm{MLP}_{\mathrm{event}}\left([\mathbf{d}_e; \bar{\mathbf{h}}; \mathrm{mean}_{i \in B_e} \mathbf{h}_i; \max_{i \in B_e} \mathbf{h}_i]\right), \tag{2}$$

where $[\cdot;\cdot]$ denotes concatenation and both boundary summaries are zero when B_e is empty, as for the molecular ion. Events then exchange context through 32 learned latent vectors that attend to the events and are read back in turn, yielding contextual event states $\tilde{\mathbf{u}}_e$, the "event tokens" in the figure. This set-based bottleneck avoids an all-events attention matrix while letting each event carry context from other candidate cleavages before formula matching.

3.3 Event-informed formula scoring

The molecular graph G includes explicit hydrogens and has elemental composition $\mathbf{c} \in \mathbb{N}_0^8$, counting H, C, N, O, F, P, S, and Cl. The candidate set $\mathcal{F}(\mathbf{c})$ contains fragment formulas $\mathbf{c}_j$ satisfying $\mathbf{0} \le \mathbf{c}_j \le \mathbf{c}$; the corresponding data filters are given in Appendix A. The model predicts a nonnegative weight $p_j(G)$ for each candidate, with $\sum_j p_j(G) = 1$. Each candidate's composition determines a fixed isotope template $\mathbf{r}_j \in \mathbb{R}^{512}_{\ge 0}$, where r_{jm} is its intensity at nominal mass m. Their weighted sum gives the predicted spectrum:

$$\hat{y}_m = \sum_{\mathbf{c}_j \in \mathcal{F}(\mathbf{c})} p_j(G) r_{jm}, \qquad m = 1, \dots, 512. \tag{3}$$

Formula tokens and matching events. Figure 2(c) encodes each candidate composition as a cumulative one-hot vector $\boldsymbol{\phi}_j$: a count c activates positions $0, \dots, c$ in its element block. A learned linear layer maps this 276-dimensional encoding to an eight-dimensional formula token $\mathbf{f}_j$.

Each token queries events with the same elemental composition: $\mathcal{E}_j = \{e : \mathbf{c}_e^{\mathrm{evt}} = \mathbf{c}_j\}$. Eight attention heads attend from $\mathrm{LN}(\mathbf{f}_j)$ to $\mathrm{LN}(\tilde{\mathbf{u}}_e)$ with width $d_a = 16$ per head:

$$\alpha_{je}^a = \frac{\exp((\mathbf{q}_j^a)^\top \mathbf{k}_e^a / \sqrt{d_a})}{\sum_{e' \in \mathcal{E}_j} \exp((\mathbf{q}_j^a)^\top \mathbf{k}_{e'}^a / \sqrt{d_a})}, \qquad e \in \mathcal{E}_j, \tag{4}$$

$$\mathbf{f}_j' = \mathbf{f}_j + \mathbf{W}_O \left[\Big\Vert_{a=1}^{8} \sum_{e \in \mathcal{E}_j} \alpha_{je}^a \mathbf{v}_e^a \right] + \mathbf{b}_O. \tag{5}$$

Here, $\mathbf{q}_j^a$, $\mathbf{k}_e^a$, and $\mathbf{v}_e^a$ are learned projections of the normalized formula token and event states, $\Vert$ concatenates heads, and the residual retains $\mathbf{f}_j$ while adding event evidence. $\mathbf{W}_O \in \mathbb{R}^{8\times 128}$ and $\mathbf{b}_O \in \mathbb{R}^8$ are the learned output projection and bias. Only matching pairs are evaluated, avoiding a dense formula–event matrix. If $\mathcal{E}_j$ is empty, $\mathbf{f}_j' = \mathbf{f}_j + \mathbf{b}_O$ and the formula can still be scored.

Formula-guided atom pooling and scoring. The updated token $\mathbf{f}_j'$ selects atom information, as shown in the "Formula–atom attention pool" block:

$$\beta_{ij} = \mathrm{softmax}_i \left((\mathbf{W}_A \mathbf{h}_i + \mathbf{b}_A)^\top \mathbf{f}_j' \right), \qquad \mathbf{g}_j = \sum_i \beta_{ij} \mathbf{h}_i, \tag{6}$$

$\mathbf{W}_A \in \mathbb{R}^{8\times 512}$ and $\mathbf{b}_A \in \mathbb{R}^8$ are the learned atom projection and bias. All molecules, including explicit hydrogens, are padded to 48 atom positions. Padded atom states are zero, but their scores remain in the softmax denominator. This scales $\mathbf{g}_j$ relative to normalization over valid atoms only. The scale depends on the molecule and candidate formula, but not on batch composition. Here, β_{ij} weights atom states for formula j and $\mathbf{g}_j$ is the pooled 512-dimensional representation. Starting from $\mathbf{s}_j^{(0)} = \mathbf{g}_j$, three GRU cells with separate parameters, shared across formulas, combine the pooled state with the original composition encoding:

$$\mathbf{s}_j^{(t)} = \mathrm{GRUCell}_t(\boldsymbol{\phi}_j, \mathbf{s}_j^{(t-1)}), \quad t = 1, 2, 3; \qquad p_j = \mathrm{softmax}_j(\mathrm{MLP}_{\mathrm{score}}(\mathbf{s}_j^{(3)})). \tag{7}$$

Thus, events influence atom selection through $\mathbf{f}_j'$, while $\boldsymbol{\phi}_j$ supplies composition at every GRU step; the candidate-wise softmax yields the weights in Equation 3. Numerical masking and architecture settings are in Appendix B.

3.4 Spectrum supervision and training

We supervise the rendered spectrum rather than assigning observed peaks to formulas or events. Define $T(\mathbf{x}) = \sqrt{\mathbf{x} + \epsilon} / \|\sqrt{\mathbf{x} + \epsilon}\|_2$. We add $\epsilon = 10^{-9}$ to each intensity before the square root. For

a batch of B molecules, the loss is

$$\mathcal{L} = \frac{1}{512B} \sum_{b=1}^{B} \left\| T(\hat{\mathbf{y}}^{(b)}) - T(\mathbf{y}^{(b)}) \right\|_2^2. \tag{8}$$

Gradients pass through the fixed renderer to train the scorer, attention modules, and atom and event encoders jointly. Formula weights are intermediate model quantities, not unique fragment assignments from the measured spectrum. We optimize with Adam and select checkpoints on validation spectra. Inference requires only the molecular structure. Appendix B gives the training settings.

4 Experiments

4.1 Experimental settings

Datasets and splits. We study limited-data learning on MoNA, large-scale prediction on NIST23, and transfer from MoNA to SWGDRUG (Fiehn Laboratory, n.d.; National Institute of Standards and Technology, 2023; Ji & Wallace, 2024). The curated datasets contain 3,977, 178,654, and 1,992 molecular structures, respectively. Appendix A gives the structural and spectral filters. We use 80/10/10 splits by structural group. Cyclic molecules are grouped by Bemis–Murcko scaffold (Bemis & Murcko, 1996), and acyclic molecules by anonymous heavy-atom topology, so that test molecules come from structural groups unseen in training. Random and fingerprint-hash splits can both place close structural analogues across training and test sets. This structural overlap exposes the model during training to close analogues of test molecules and may lead to high test scores. Our grouping keeps the defined scaffold and topology groups disjoint across training, validation, and test sets. NIST23 uses one split and MoNA three splits, each with three optimizer seeds. SWGDRUG evaluation uses MoNA-trained checkpoints after removing overlap with the corresponding training and validation sets.

Baselines and metrics. We compare with RASSP:SN, RASSP:FN (Zhu & Jonas, 2023), NEIMS (Wei et al., 2019), Hybrid (Majewski & Łabuda, 2026), MassFormer (Young et al., 2024), MolFormer (Ross et al., 2022), and CFM-ID (its released EI model, CFM-EI; Allen et al., 2016). The EI-native baselines span the three design families reviewed above, so the comparison is not confined to one design. We further adapt MassFormer from tandem spectra and the pretrained MolFormer model to EI. MolFormer tests whether chemical language pretraining helps when EI labels are limited. The original trainable models use the same data splits and validation-based selection; pretrained neural models start from their released weights. CFM-ID is a frozen third-party reference, without retraining or tuning on these splits. Its historical training overlap with NIST23 is not auditable; we therefore do not interpret it as a matched-training comparison. All comparisons require the same evaluation queries, frozen candidate lists, nominal-mass processing, and scoring rules. The accuracy of predicted spectra is measured by four spectrum metrics. DP (Nowatzky et al., 2025) is the cosine similarity of square-root intensities; the transform gives weaker peaks more relative weight, so the metric reflects the overall intensity profile including weak peaks. SDP, the Stein dot product (Zhu & Jonas, 2023), adds the mass and intensity weights of Stein & Scott (1994); errors at large masses receive more weight, so the metric emphasizes high-mass peaks. COS is the cosine similarity of raw intensities and measures unweighted peak-wise agreement. Top_1 is base-peak accuracy, the fraction of queries whose predicted base peak coincides with the observed one.

Beyond spectrum-fitting metrics, we design two isomer-retrieval tasks that test whether predicted spectra can distinguish same-formula isomers. Retrieval ranks candidates by SDP between the predicted candidate spectra and the observed query spectrum. Same-source retrieval uses $K = 5$ candidates: the target and four same-formula isomers from the same test collection. This small candidate set reflects the limited number of same-formula isomers available in the test data. To evaluate discrimination among more candidates, we also designed a cross-source $K = 50$ protocol. Each set contains the test target and 49 structurally similar, same-formula PubChem isomers selected as hard negatives by Morgan fingerprint similarity (Kim et al., 2023). R@1 is the fraction of queries for which the target ranks uniquely first; MRR averages the reciprocal of the target's rank across queries, using average ranks for ties. The two metrics measure top-1 identification accuracy and how highly the target ranks among isomers; higher values are better. Candidate selection, metric formulas, and training settings are detailed in Appendices A, A.1, and B.3, respectively.

Dataset	Method	Spectrum prediction				Same-source (K=5)		Cross-source (K=50)	
		DP ↑	SDP ↑	COS ↑	Top_1 ↑	R@1 ↑	MRR ↑	R@1 ↑	MRR ↑
MoNA	CLEF	**0.773**±0.008	**0.748**±0.019	**0.696**±0.019	**0.446**±0.039	**0.825**±0.052	**0.901**±0.034	**0.348**±0.040	**0.473**±0.033
	RASSP:SN	0.744±0.012	0.718±0.027	0.645±0.024	0.380±0.021	0.807±0.083	0.885±0.049	0.330±0.048	0.445±0.048
	MassFormer	0.731±0.016	0.688±0.015	0.592±0.025	0.292±0.027	0.757±0.052	0.860±0.026	0.253±0.039	0.375±0.035
	RASSP:FN	0.674±0.007	0.499±0.041	0.537±0.016	0.277±0.030	0.569±0.059	0.740±0.036	0.136±0.023	0.243±0.023
	MolFormer	0.691±0.006	0.682±0.014	0.596±0.012	0.308±0.020	0.689±0.052	0.819±0.033	0.215±0.024	0.340±0.019
	NEIMS	0.700±0.010	0.660±0.017	0.619±0.013	0.362±0.022	0.680±0.052	0.819±0.027	0.212±0.024	0.342±0.020
	Hybrid	0.576±0.036	0.687±0.021	0.396±0.068	0.149±0.064	0.640±0.083	0.794±0.048	0.292±0.059	0.422±0.055
	CFM-ID†	0.606±0.013	0.553±0.013	0.600±0.010	0.382±0.015	0.637±0.020	0.787±0.013	0.142±0.013	0.296±0.016
NIST23	CLEF	**0.874**±0.001	**0.877**±0.001	**0.817**±0.003	**0.583**±0.005	**0.740**±0.002	**0.848**±0.001	**0.560**±0.004	**0.679**±0.003
	RASSP:SN	0.839±0.001	0.851±0.002	0.778±0.002	0.529±0.003	0.719±0.002	0.835±0.002	0.518±0.004	0.648±0.003
	MassFormer	0.848±0.001	0.835±0.001	0.737±0.001	0.449±0.002	0.688±0.007	0.816±0.003	0.470±0.001	0.604±0.002
	RASSP:FN	0.838±0.001	0.808±0.007	0.754±0.001	0.512±0.001	0.627±0.007	0.776±0.005	0.400±0.014	0.538±0.013
	MolFormer	0.800±0.001	0.812±0.002	0.710±0.004	0.438±0.008	0.646±0.005	0.790±0.003	0.397±0.003	0.536±0.003
	NEIMS	0.772±0.001	0.780±0.001	0.678±0.001	0.477±0.002	0.623±0.005	0.775±0.002	0.367±0.001	0.508±0.001
	Hybrid	0.631±0.004	0.757±0.001	0.367±0.007	0.108±0.008	0.608±0.004	0.767±0.002	0.326±0.004	0.475±0.001
	CFM-ID†	0.547	0.556	0.538	0.337	0.452	0.660	0.161	0.299

Table 1: Spectrum prediction and isomer retrieval. Trainable models: mean ± sample SD over three NIST23 or nine MoNA split–seed runs; ranking marks are restricted to these trainable models. Bold and underline denote the best and second-best results, respectively. CFM-ID (†) is a frozen EI reference; coverage and protocol details are in Appendix A.2.

4.2 Results

On the small public MoNA dataset, CLEF had the highest mean scores among the six trainable baselines (Table 1). This held for all four spectrum metrics and for R@1 and MRR in both retrieval protocols. The same pattern held on the large commercial NIST23 library. Together with the NIST23 ablations, these results support the role of cleavage events and their local atom context in learning fragmentation-related representations for spectrum prediction and isomer retrieval. CLEF thus effectively uses molecular structure information, achieving the best mean performance among the compared trainable models in both limited-data and large-scale settings.

To test whether the small-data advantage extended beyond MoNA, we constructed nested training subsets of NIST23. Validation data, test data and retrieval candidates remained fixed. CLEF had higher means than RASSP:SN on all four spectrum metrics and all four retrieval metrics at every tested fraction, from 1% to 100% (Figure 3). With 1% of the data, it reached 0.7319 DP, compared with 0.6770 for RASSP:SN. With 25%, it reached 0.8491 DP, above the 0.8388 of full-data RASSP:SN.

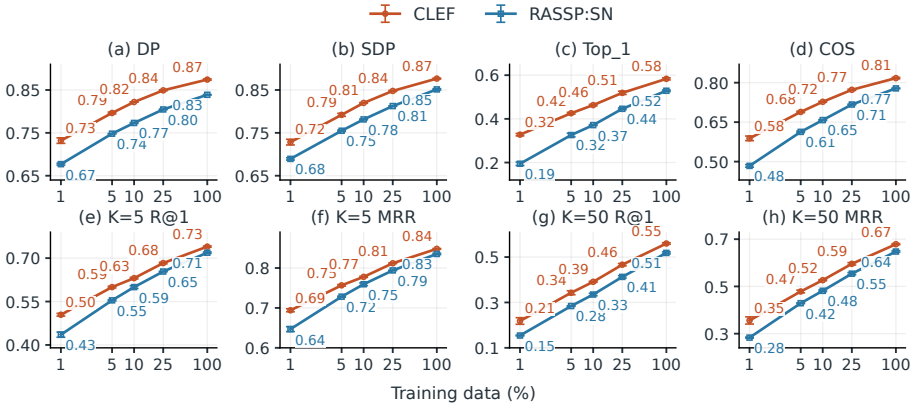

Figure 3: NIST23 learning curves: spectrum metrics (top); retrieval ($K = 5$ same-source, $K = 50$ cross-source; bottom). Labels: truncated means (2 d.p.); error bars: sample SD (3 seeds).

We then asked whether learning from the small MoNA dataset could extend to a new library. Using only MoNA-trained checkpoints, CLEF had the highest means on all four SWGDRUG spectrum metrics among the transferred trainable models (Table 2). Its DP was 0.724, compared with 0.648 for RASSP:SN. The compatible SWGDRUG pool had a higher median exact mass than MoNA: 231.16 versus 172.03 Da. The transfer results show that CLEF can predict spectra in a new library despite a shift in molecular mass distribution.

Dataset	Method	Spectrum prediction			
		DP ↑	SDP ↑	COS ↑	Top_1 ↑
SWGDRUG	CLEF	**0.724** ± 0.007	**0.651** ± 0.013	**0.605** ± 0.018	**0.413** ± 0.027
	RASSP:SN	<u>0.648</u> ± 0.005	<u>0.569</u> ± 0.009	<u>0.439</u> ± 0.015	<u>0.212</u> ± 0.018
	MassFormer	0.561 ± 0.006	0.498 ± 0.006	0.331 ± 0.007	0.150 ± 0.011
	RASSP:FN	0.546 ± 0.018	0.432 ± 0.034	0.297 ± 0.013	0.128 ± 0.009
	MolFormer	0.478 ± 0.010	0.506 ± 0.004	0.308 ± 0.008	0.127 ± 0.010
	NEIMS	0.568 ± 0.006	0.517 ± 0.006	0.354 ± 0.006	0.160 ± 0.005
	Hybrid	0.385 ± 0.018	0.493 ± 0.009	0.155 ± 0.021	0.040 ± 0.022
	CFM-ID†	0.552 ± 0.000	0.500 ± 0.001	0.534 ± 0.000	0.338 ± 0.001

Table 2: SWGDRUG spectrum evaluation. Trainable models transfer from MoNA, with mean ± sample SD over nine split–seed runs. Bold and underline denote the best and second-best results, respectively. CFM-ID (†) is a frozen external reference, excluded from ranking. Details are in Appendix A.2.

4.3 Ablation and event attribution

Ablation analysis. We trained and evaluated model variants on NIST23 to examine each branch's contribution. The graph-only model predicts spectra directly from atom properties and connectivity. The formula branch restricts candidate fragment formulas to the precursor's element counts. The predicted spectrum is a weighted sum of their isotope templates. Formula-guided atom pooling combines elemental composition with molecular structure for scoring. The event branch adds cleavage evidence with the same elemental composition as each candidate formula. This evidence updates formula representations and guides atom selection and the weights used to render the spectrum. Consistent with this design, adding formula scoring and then event information improved spectrum prediction and isomer retrieval (Table 3a).

We further tested whether the enumerated events provided useful fragmentation-related information. The formula–graph null replaces event inputs with features built from candidate formulas and the global graph state. These features pass through the event encoder and fusion modules without actual event descriptors or cut-endpoint information. This tests whether the replacement inputs can preserve predictive performance. The shuffled control tests whether event descriptors need to match their local atom context. Within each molecule, it exchanges descriptors among events with the same type, cut count, and hydrogen-count change. The exchanged descriptors are concatenated with cut-endpoint atom states from other events. This applies during both training and evaluation. Both controls had lower spectrum and retrieval scores than CLEF (Table 3b). These results show that cleavage events and their local atom context help the model learn fragmentation-related representations for spectrum prediction and isomer retrieval.

Variant	Spectrum prediction				Same-source $K = 5$		Cross-source $K = 50$	
	DP	SDP	COS	Top_1	R@1	MRR	R@1	MRR
(a) Conditioning branches								
Graph	0.7691	0.5051	0.6303	0.3463	0.4814	0.6806	0.1832	0.3134
Graph + formula	0.8091	0.7437	0.7001	0.4401	0.5492	0.7240	0.2845	0.4277
CLEF	**0.8729**	**0.8758**	**0.8147**	**0.5782**	**0.7386**	**0.8477**	**0.5567**	**0.6762**
(b) Event-descriptor alignment								
CLEF	**0.8729**	**0.8758**	**0.8147**	**0.5782**	**0.7386**	**0.8477**	**0.5567**	**0.6762**
Shuffled descriptors	0.8577	0.8525	0.7887	0.5464	0.6945	0.8202	0.4824	0.6166
Formula–graph null	0.8072	0.7471	0.6965	0.4313	0.5483	0.7228	0.2920	0.4319

Table 3: Bold marks the best value per column within each block.

Event-erasure analysis. We next asked whether the model relies on events that receive high attention. If these events carry useful evidence, removing them should harm predictions more than removing the same number of other events. The erasure takes place at the composition-matched cross-attention shown in Figure 2(d). With the trained model frozen, we mask selected events at this step so that they no longer contribute to the formula update. Atom states, event tokens and all learned weights stay unchanged, and the remaining attention weights are renormalized. We compared

removal of the highest-attention, random, and lowest-attention events in equal numbers. Across all three training seeds, removing high-attention events caused the largest loss in spectrum fit (Figure 4a). The effect therefore depended on which events were removed, not just how many.

We then asked whether these events also help distinguish same-formula isomers. We applied event erasure to both predictions in each isomer pair. The paired margin measures how much better predictions match their own measured spectra than the other isomer's spectrum. A larger margin makes the correct pairing easier to distinguish from the crossed pairing. Removing high-attention events reduced this margin more than random or low-attention removal (Figure 4b). This links the attended events to structural information used for isomer discrimination, as well as spectrum fit.

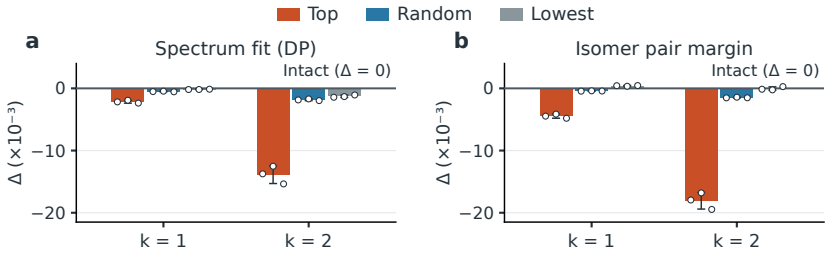

Figure 4: NIST23 frozen-model event erasure: (a) DP; (b) same-formula isomer-pair margin. The axis is erased minus intact (10^{-3}); negative bars indicate decreases. Orange/blue/grey mark highest-attention/random/lowest-attention removal; k is events removed per selected formula. Bars/points/error bars show three-seed means/individual seeds/sample SD.

Observation variation. Community libraries vary in provenance and quality (Fiehn Laboratory, n.d.; Wallace et al., 2017); tuning, counting noise, and spectrum extraction alter relative ion abundances and weak peaks (Kelly & Bell, 2018; Stein, 1999). On curated commercial NIST23, we simulate these changes by perturbing observed spectra only, with matched draws across models. Under these simulated acquisition and processing variations, CLEF retains the highest DP and SDP among seven predictors in all eight conditions (Figure 5), maintaining the strongest spectral agreement. Together with the limited-data learning and cross-library transfer results, this provides complementary evidence for generalization to community libraries with limited data and variable quality.

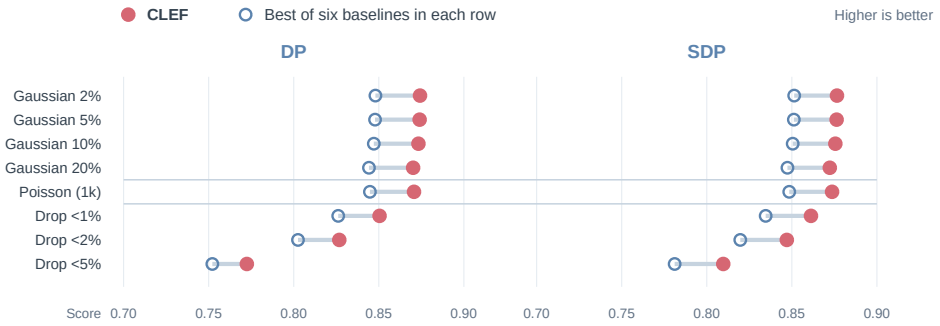

Figure 5: NIST23 DP/SDP under observation perturbations. Filled markers show CLEF and open markers show the strongest of six baselines in each condition. Points are three-seed means; full scores are in Appendix C.2.

5 Conclusion

CLEF learns from cleavage events and local atom context to predict EI spectra without fragment-level annotations. It achieves the best mean spectrum and isomer-retrieval scores among the compared trainable models across data scales. It is also data-efficient and transfers to a new spectral library. These findings show that structure-derived cleavage information can improve spectral accuracy while preserving differences useful for distinguishing same-formula isomers.

AI USE STATEMENT

AI tools assisted with literature search, language polishing, figure scripting, and manuscript consistency checks.

REPRODUCIBILITY STATEMENT

We will open-source the code, data construction scripts, configuration files, and related materials. The NIST23 spectra are distributed under a commercial license, and we do not redistribute them; the released split files and processing code reconstruct the training and evaluation sets from a licensed copy. MoNA and SWGDRUG are openly available, and the PubChem decoys derive from open structures. Appendix A describes data preparation, splits, candidate pools, and metrics. Appendix B gives model and training details, followed by the analysis protocols and additional results. We have released the code of CLEF in an anonymous repository: https://anonymous.4open.science/r/CLEF-F368.

REFERENCES

Felicity Allen, Allison Pon, Russ Greiner, and David Wishart. Computational prediction of electron ionization mass spectra to assist in GC/MS compound identification. *Analytical Chemistry*, 88(15): 7689–7697, 2016. doi: 10.1021/acs.analchem.6b01622.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi: 10.1021/jm9602928.

Roman Bushuiev et al. MassSpecGym: A benchmark for the discovery and identification of molecules. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3491. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c6c31413d5c53b7d1c343c1498734b0f-Abstract-Datasets_and_Benchmarks_Track.html.

Fiehn Laboratory. MassBank of North America (MoNA). Mass spectral repository, n.d. URL https://mona.fiehnlab.ucdavis.edu/. Accessed 20 June 2026.

Samuel Goldman, John Bradshaw, Jiayi Xin, and Connor W. Coley. Prefix-tree decoding for predicting mass spectra from molecules. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Johannes Gorges and Stefan Grimme. QCxMS2: A program for the calculation of electron ionization mass spectra via automated reaction network discovery. *Physical Chemistry Chemical Physics*, 27: 6899–6911, 2025. doi: 10.1039/D5CP00316D.

Stefan Grimme. Towards first principles calculation of electron impact mass spectra of molecules. *Angewandte Chemie International Edition*, 52(24):6306–6312, 2013. doi: 10.1002/anie.201300158.

Weihua Ji and William E. Wallace. Comprehensive data evaluation methods used in developing the SWGDRUG mass spectral reference library for seized drug identification. *Analytical Chemistry*, 96(42):17004–17012, 2024. doi: 10.1021/acs.analchem.4c04425.

Kristin Kelly and Suzanne Bell. Evaluation of the reproducibility and repeatability of GCMS retention indices and mass spectra of novel psychoactive substances. *Forensic Chemistry*, 7:10–18, 2018. doi: 10.1016/j.forc.2017.11.002.

Sunghwan Kim et al. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

Jeroen Koopman and Stefan Grimme. Calculation of electron ionization mass spectra with semiempirical GFNn-xTB methods. *ACS Omega*, 4(12):15120–15133, 2019. doi: 10.1021/acsomega.9b02011.

Ayumi Kubo, Azusa Kubota, Haruki Ishioka, Takuhiro Hizume, Masaaki Ubukata, Kenji Nagatomo, Takaya Satoh, Mitsuyoshi Yoshida, and Fuminori Uematsu. Construction of a mass spectrum library containing predicted electron ionization mass spectra prepared using a machine learning model and the development of an efficient search method. *Mass Spectrometry*, 12(1):A0120, 2023. doi: 10.5702/massspectrometry.A0120.

Bartosz Majewski and Marta Łabuda. Hybrid deep learning model for EI-MS spectra prediction. *International Journal of Molecular Sciences*, 27(3):1588, 2026. doi: 10.3390/ijms27031588.

Satoki Muto, Akiko Kumada, and Masahiro Sato. Prediction of mass spectra using large chemical language models and verification of adaptability in data-scarce domains. *Applied Physics Letters*, 129(5):051903, 2026. doi: 10.1063/5.0326955.

National Institute of Standards and Technology. NIST23: Updates to the NIST tandem and electron ionization spectral libraries. NIST Standard Reference Database 1A, 2023. URL https://www.nist.gov/programs-projects/nist23-updates-nist-tandem-and-electron-ionization-spectral-libraries. Licensed database; accessed 9 July 2026.

Yannek Nowatzky, Francesco Friedrich Russo, Jan Lisec, Alexander Kister, Knut Reinert, Thilo Muth, and Philipp Benner. FIORA: Local neighborhood-based prediction of compound mass spectra from single fragmentation events. *Nature Communications*, 16:2298, 2025. doi: 10.1038/s41467-025-57422-4.

Chen Qu, Barry I. Schneider, Anthony J. Kearsley, Walid Keyrouz, and Thomas C. Allison. Applying graph neural network models to molecular property prediction using high-quality experimental data. *Artificial Intelligence Chemistry*, 2:100050, 2024. doi: 10.1016/j.aichem.2024.100050.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web*, pp. 593–607, 2018. doi: 10.1007/978-3-319-93417-4_38.

Stephen E. Stein. An integrated method for spectrum extraction and compound identification from gas chromatography/mass spectrometry data. *Journal of the American Society for Mass Spectrometry*, 10(8):770–781, 1999. doi: 10.1016/S1044-0305(99)00047-1.

Stephen E. Stein and Donald R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society for Mass Spectrometry*, 5 (9):859–866, 1994. doi: 10.1016/1044-0305(94)87009-8.

Julia Wakoli, Afia Anjum, Tanvir Sajed, Eponine Oler, Fei Wang, Vasuk Gautam, Marcia LeVatte, and David S. Wishart. GCMS-ID: A webserver for identifying compounds from gas chromatography mass spectrometry experiments. *Nucleic Acids Research*, 52(W1):W381–W389, 2024. doi: 10.1093/nar/gkae425.

William E. Wallace, Weihua Ji, Dmitrii V. Tchekhovskoi, Karen W. Phinney, and Stephen E. Stein. Mass spectral library quality assurance by inter-library comparison. *Journal of the American Society for Mass Spectrometry*, 28(4):733–738, 2017. doi: 10.1007/s13361-016-1589-4.

Jennifer N. Wei, David Belanger, Ryan P. Adams, and D. Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Central Science*, 5(4):700–708, 2019. doi: 10.1021/acscentsci.9b00085.

Ting Xie, Yaoyu Zhuang, Jiongyu Wang, Hailiang Zhang, Chuanxiu Yu, Zhimin Zhang, and Hongmei Lu. CSU-EP: Contrastive learning for unifying experimental and predicted EI-MS spectra. *Analytical Chemistry*, 98(30):22119–22131, 2026. doi: 10.1021/acs.analchem.6c01233. URL https://pubs.acs.org/doi/10.1021/acs.analchem.6c01233.

Qiong Yang, Hongchao Ji, Zhenbo Xu, Yiming Li, Pingshan Wang, Jinyu Sun, Xiaqiong Fan, Hailiang Zhang, Hongmei Lu, and Zhimin Zhang. Ultra-fast and accurate electron ionization mass spectrum matching for compound identification with million-scale in-silico library. *Nature Communications*, 14:3722, 2023. doi: 10.1038/s41467-023-39279-7. URL https://www.nature.com/articles/s41467-023-39279-7.

Adamo Young, Hannes Röst, and Bo Wang. Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, 6:404–416, 2024. doi: 10.1038/s42256-024-00816-8.

Adamo Young, Fei Wang, David S. Wishart, Bo Wang, Russell Greiner, and Hannes Röst. FraGNNet: A deep probabilistic model for tandem mass spectrum prediction. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2404.02360.

Baojie Zhang, Jun Zhang, Yi Xia, Peng Chen, and Bing Wang. Prediction of electron ionization mass spectra based on graph convolutional networks. *International Journal of Mass Spectrometry*, 475: 116817, 2022. doi: 10.1016/j.ijms.2022.116817.

Richard Licheng Zhu and Eric Jonas. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical Chemistry*, 95(5):2653–2663, 2023. doi: 10.1021/acs.analchem.2c02093.

A DATA, BASELINES, AND EVALUATION DETAILS

The experiments use NIST23 (National Institute of Standards and Technology, 2023), MoNA (Fiehn Laboratory, n.d.), and SWGDRUG v3.14 (Ji & Wallace, 2024). Acquisition-energy metadata are handled according to their availability in each source. The MoNA preparation retains EI–MS records, removes records explicitly reporting an electron energy other than 70 eV, and distinguishes records reporting 70 eV from those with unreported energy; the latter are retained. The energy parser accepts reported values within 0.51 eV of 70 eV. SWGDRUG uses the EI spectra released in version 3.14. Structural compatibility requires a single covalent component, elements from $\{\mathrm{H}, \mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}, \mathrm{P}, \mathrm{S}, \mathrm{Cl}\}$, exact molecular mass at most 400 Da, at most 48 atoms including explicit hydrogens, and at most 8,192 subformulas. Hydrogens are standardized before counting atoms. Connectivity-level consensus is constructed before partitioning, producing 178,654 NIST23 structures and 3,977 MoNA targets from 5,265 compatible source records.

Cyclic molecules are grouped by their Bemis–Murcko scaffold (Bemis & Murcko, 1996); acyclic molecules are grouped by anonymous heavy-atom topology. A group is assigned wholly to training, validation, or test. NIST23 contains 39,453 such groups and uses one split with 142,906/17,869/17,879 training/validation/test structures. The three MoNA splits contain 3,189/394/394, 3,178/399/400, and 3,189/395/393 structures, respectively. Connectivity and scaffold intersections between phases are zero. Each split is trained with optimizer seeds 1234, 2345, and 3456; checkpoint selection uses validation data. Data-efficiency experiments use nested subsets of NIST23 training connectivities and leave validation and test fixed. The cross-source $K = 50$ learning curves use the repaired per-seed export at 1–25% and the existing full-data reference endpoints at 100%.

SWGDRUG provides 1,992 compatible structures and is used to evaluate MoNA-trained checkpoints. For each MoNA split, we remove SWGDRUG compounds sharing training or validation connectivity, exact spectra, or near-identical spectra with square-root-intensity cosine at least 0.995. The resulting spectrum cohorts contain 1,767/1,768/1,774 compounds. These same cohorts are used for every method's external spectrum comparison.

Retrieval uses an observed spectrum to rank predicted spectra of candidate structures. For same-source $K = 5$, eligible neutral-formula groups contain at least five distinct connectivities in the same test collection. Each member serves as a query; its candidate set contains the positive and the four closest remaining structures by Morgan radius-2, 2,048-bit Tanimoto similarity (Rogers & Hahn, 2010). Membership is selected before a deterministic permutation of candidate order. MoNA S1/S2/S3 contribute 6/6/8 formula groups and 47/40/48 queries; NIST23 contributes 782 groups and 6,870 queries. Each eligible group contributes one query per member, so query count and formula-group count describe different aspects of the evaluation.

Cross-source $K = 50$ candidates comprise one query-source positive and 49 same-formula, distinct-connectivity PubChem decoys (Kim et al., 2023), selected by Morgan2 similarity. Only queries with sufficient eligible decoys enter this protocol: MoNA contributes 285/289/286 queries and NIST23 contributes 10,156. Candidate membership is shared across methods; the same-source and cross-source scores refer to their respective query populations.

A.1 EVALUATION METRICS

Spectrum prediction. Let $\mathcal{M} \subset \mathbb{R}_{>0}$ be a finite set of aligned m/z coordinates at the chosen spectral resolution. Write $\mathbf{y} = (y_\mu)_{\mu\in\mathcal{M}}$ and $\hat{\mathbf{y}} = (\hat{y}_\mu)_{\mu\in\mathcal{M}}$ for the nonnegative observed and predicted intensities, assigning zero to absent peaks. For nonzero vectors, define

$$C(\mathbf{a}, \mathbf{b}) = \frac{\sum_{\mu\in\mathcal{M}} a_\mu b_\mu}{\sqrt{\sum_{\mu\in\mathcal{M}} a_\mu^2}\sqrt{\sum_{\mu\in\mathcal{M}} b_\mu^2}}, \qquad [S(\mathbf{y})]_\mu = \mu^3 y_\mu^{0.6}. \tag{9}$$

For nonzero predictions, the four spectrum metrics are

$$\begin{aligned} \mathrm{DP} &= \frac{\sum_{\mu\in\mathcal{M}} \sqrt{\hat{y}_\mu y_\mu}}{\sqrt{\sum_{\mu\in\mathcal{M}} \hat{y}_\mu}\sqrt{\sum_{\mu\in\mathcal{M}} y_\mu}}, \qquad \mathrm{SDP} = C(S(\hat{\mathbf{y}}), S(\mathbf{y})), \\ \mathrm{COS} &= C(\hat{\mathbf{y}}, \mathbf{y}), \qquad\qquad \mathrm{Top_1} = \mathbf{1}[a(\hat{\mathbf{y}}) = a(\mathbf{y})]. \end{aligned} \tag{10}$$

where $a(\mathbf{y}) = \min \arg\max_{\mu \in \mathcal{M}} y_\mu$ selects the smallest m/z coordinate when multiple peaks share the maximum intensity. Equivalently, DP is $C(\sqrt{\hat{\mathbf{y}}}, \sqrt{\mathbf{y}})$: its square-root transform gives weaker peaks more relative weight. SDP adds the mass and intensity weights of Stein & Scott (1994). In this study, the spectra are binned at unit nominal-mass resolution, so $\mathcal{M} = \{1, \ldots, 512\}$ and array index zero represents $m/z = 1$. The NIST23/MoNA spectrum scripts add 10^{-10} to the cosine denominator; the external-spectrum and retrieval scripts use the nonzero denominator directly. No epsilon is added to the intensities before these evaluation transforms. A finite all-zero prediction receives zero on all four spectrum metrics and remains in the evaluation denominator. Observed spectra must have nonzero norm.

Candidate retrieval. For query q and candidate c, the ranking score is $s_{qc} = C(S(\mathbf{y}_q), S(\hat{\mathbf{y}}_c))$, clipped to $[0, 1]$ for numerical round-off. The ranking evaluator rejects zero-norm query or candidate spectra. Let s_q^+ denote the positive candidate's score. A candidate is tied with it when $|s_{qc} - s_q^+| \leq 10^{-12} + 10^{-10}|s_q^+|$. If g_q candidates score strictly higher after excluding ties, and t_q candidates are tied including the positive, its midrank is

$$r_q = 1 + g_q + \frac{t_q - 1}{2}, \qquad \mathrm{R@1} = \frac{1}{Q}\sum_{q=1}^{Q} \mathbf{1}[r_q \leq 1], \qquad \mathrm{MRR} = \frac{1}{Q}\sum_{q=1}^{Q} \frac{1}{r_q}. \tag{11}$$

Thus, a tie for first place is not a top-one hit. The same definitions apply to the $K = 5$ and $K = 50$ candidate sets described above.

Aggregation. Within an evaluation cell, spectrum metrics average over molecules and retrieval metrics average over eligible queries. Let z_c be a cell mean. Reported means and sample standard deviations are

$$\bar{z} = \frac{1}{C}\sum_{c=1}^{C} z_c, \qquad s_z = \sqrt{\frac{1}{C-1}\sum_{c=1}^{C}(z_c - \bar{z})^2}. \tag{12}$$

For the main comparisons, cells receive equal weight: $C = 3$ optimizer seeds for NIST23, and $C = 9$ split–seed combinations for MoNA and SWGDRUG. The learning curves use $C = 3$ at each training fraction. Their error bars show this sample standard deviation.

A.2 Baseline adaptation

All comparisons reuse the same evaluation queries, candidate lists, nominal m/z binning, and physical SDP ranking rule.

MassFormer. MassFormer was released for tandem MS as a graph transformer initialized from chemical-property pretraining (Young et al., 2024). We adapted it to the EI task from these released weights: data splits, the three optimizer seeds, effective batch size, update and exposure caps, and validation-based checkpoint selection match CLEF, while the model keeps its native loss and learning-rate schedule (polynomial decay with a 2×10^{-4} peak).

MolFormer. MolFormer is a large pretrained chemical language model (Ross et al., 2022). It has also been used for EI spectrum prediction (Muto et al., 2026). We adapted its released weights to EI under our matched protocol; the learning rate is held constant at 1×10^{-4} after a validation-stability adjustment, and part of the pretrained representation is frozen.

Hybrid. Hybrid is a recent GNN–ResNet predictor for EI spectra (Majewski & Łabuda, 2026). No official implementation was available, so we trained a paper-faithful re-implementation under the same data, splits, seeds, and validation-based selection as the other trainable baselines: a matched task adaptation, not a reproduction of official code.

CFM-ID. The CFM-ID entry uses the official generic electron-ionization model bundled with the CFM-ID v3.3 release (`ei_ms_model`; CFM-EI, Allen et al., 2016), loaded as a frozen third-party reference: no retraining, no parameter tuning, and one prediction per structure. Its historical training overlap with NIST23 is not auditable, so it is not interpreted as a matched-training comparison. On NIST23, 8 spectrum, 10 same-source K=5, and 23 cross-source K=50 queries remain on hold (17,871/17,879; 6,860/6,870; 10,133/10,156 scored); the three MoNA splits hold 4/2/3 cross-source

queries, and SWGDRUG is fully covered. CFM-ID metrics are means over scored queries (†), not full-roster or paired results; the model is not MoNA-trained, so its SDs reflect split and coverage variation only, and rounded SDs of 0.000 are nonzero.

B MODEL DETAILS

B.1 CANDIDATE CONSTRUCTION AND FEATURES

The formula enumerator takes the Cartesian product of the possible counts of each supported element. Molecules whose complete candidate set exceeds 8,192 are excluded. The cumulative one-hot blocks have widths $(50, 46, 30, 30, 30, 30, 30, 30)$ for H, C, N, O, F, P, S, and Cl. A count c activates positions $0, \ldots, c$ of its block. The isotope renderer stores up to 12 peaks for each formula. The zero-count formula has a zero-intensity template. Thus, formula weights sum to one, while the rendered spectrum is normalized as part of the loss and evaluation.

Event construction enumerates connected components after one to three heavy-atom bond cuts. It tests at most 250,000 cut combinations and adds hydrogen-count changes from -3 to 3. Candidates with negative counts, counts above the precursor, invalid mass, or improper components are removed. The maximum component fraction is 0.98. The molecular ion with unchanged hydrogen counts is added separately. Candidates are deduplicated by the selected atom component, number of cuts, and hydrogen-count change. Deterministic balanced selection retains at most 8,192 events. Multiple events can still map to one formula.

Table 4 gives the descriptor used in Equation 2. Composition counts are divided by the precursor's explicit-H atom count. Masses are divided by 511, and the cut count is divided by three. Cut-bond properties are proportions within the cut set. The descriptor counts cut endpoints with multiplicity. For neural boundary pooling, we instead keep at most four unique endpoints of cuts crossing the selected component, in enumeration order, and use their mean and maximum atom states. The molecular-ion event has no cut endpoints and uses zero boundary summaries.

Feature group	Dimensions
Event type and cut count	$3 + 1$
Hydrogen-count-change one-hot	7
Fragment and complementary element counts	$8 + 8$
Exact mass, rounded mass, component atom fraction	3
Cut-bond type proportions	4
Ring, aromatic, and conjugated cut proportions	3
Cut-endpoint element proportions	8

Table 4: The 45-dimensional event descriptor.

B.2 ATTENTION AND SCORING DETAILS

The graph encoder has 16 layers and uses 45-dimensional atom features, masked input batch normalization, and four independent relation transforms. Relation adjacency matrices use symmetric normalization, with degrees computed from the union of the bond relations. The first layer maps input features to width 512. Each layer applies LeakyReLU and dropout; later layers then add their inputs as residuals. Layer normalization follows the residual addition, and an output atom mask sets padded states to zero. The first layer is normalized without residual addition.

The event MLP has two Linear–LayerNorm–ReLU–Dropout layers of width 256. The latent set module has 32 learned vectors of width 256 and one block with eight heads. Each attention and feed-forward sublayer uses pre-layer normalization and a residual update. Feed-forward layers expand the width by four. LayerNorm–Linear projections map events into and out of the module, with an outer residual from the original event embedding.

Formula–event attention projects the eight-dimensional formula token and 256-dimensional event state into a total attention width of 128. Element counts are mapped to exact-match indices before attention. Each event is associated with its matching formula index; grouped softmax and scatter-add perform Equations 4–5. Attention dropout is 0.1 during training. When a formula has no matching events, its event aggregate is zero; the output projection bias in Equation 5 is retained. The projection weight and bias both start at zero.

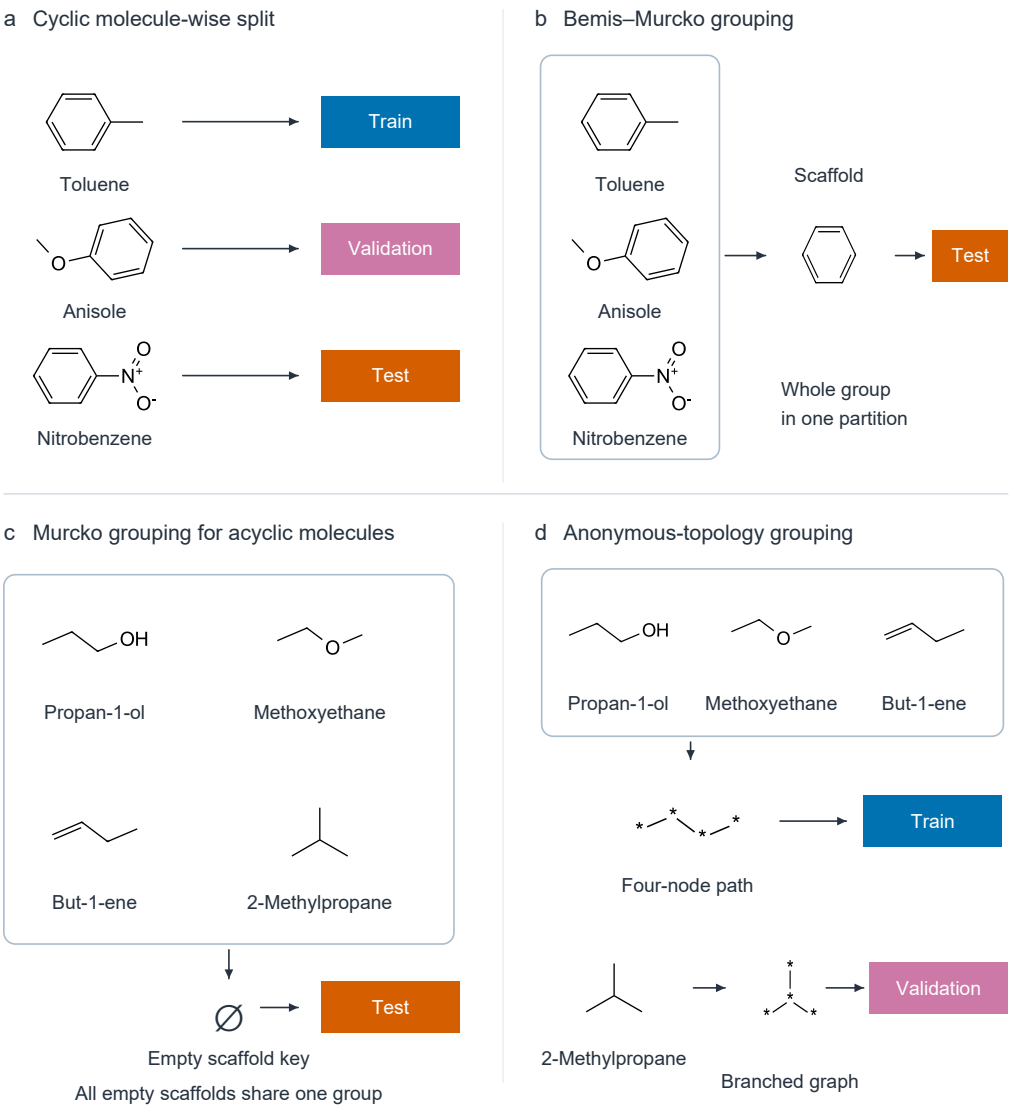

Figure 6: Structural grouping for data partitioning. (a) Molecule-wise assignment can separate a common cyclic scaffold. (b) Bemis–Murcko grouping keeps it together. (c) All acyclic examples have an empty Murcko scaffold and share one group. (d) Anonymous heavy-atom graphs separate path and branched topologies. Each group stays within one partition.

Formula-conditioned atom attention projects atom keys to width eight and uses an unscaled dot product. Its softmax runs over the padded atom axis; atom values are zero-masked before pooling. The three GRU cells have separate parameters, input width 276, and hidden width 512. The scorer maps 512 to 128 dimensions with ReLU, applies nine further 128-dimensional Linear–ReLU blocks, then layer normalization and a scalar output layer. Padded formula scores are set to -100 before the candidate softmax; this is a finite numerical mask rather than exact exclusion from its denominator. The final prediction uses formula weights and isotope templates; events contribute through the formula representations.

B.3 Training settings

Structural splitting. Bemis–Murcko scaffolds retain rings and the linkers between them after removing side chains (Bemis & Murcko, 1996). We use the canonical scaffold with its atom and bond types as the group key. Acyclic molecules all have an empty Murcko scaffold. Applying that key directly would place all acyclic molecules in one group, as panel (c) shows. We instead group

their full heavy-atom graphs after removing element labels and bond orders. This anonymous graph preserves the number of atoms and their connectivity. For example, the three unbranched molecules in Figure 6 share a four-node path, while isobutane gives a four-node star. Whole groups are assigned to the training, validation, and test sets toward an 80/10/10 ratio. No group is split between sets.

Optimization. CLEF uses Adam with learning rate 8×10^{-5}, microbatches of six, and 20 gradient-accumulation steps, giving an effective batch of 120. NIST23 samples 60,000 examples per epoch; MoNA samples 6,000 per epoch. Dropout is 0.1 in the graph, event, latent, and cross-attention modules. Training uses full precision. Validation is evaluated every five epochs. Early stopping tracks validation SDP. NIST23 stops after four consecutive checks (20 epochs) without improving the best score by more than 0.001; MoNA stops after five consecutive checks (25 epochs) without a strictly higher score. A qualifying improvement resets the counter. Checkpoint selection uses validation SDP. The loss in Equation 8 has no additional mass weighting. Its square-root transform uses $\epsilon = 10^{-9}$. All candidate features and isotope tables are computed from structures. MassFormer initializes from chemical-property pretraining; MolFormer initializes from chemical language pretraining. Their EI adaptations retain their own learning-rate schedules. The trainable comparison models select checkpoints using validation spectra; the released CFM-ID EI model remains frozen.

C Additional model analyses

C.1 Event-erasure analysis

The probe measures how removing attended events changes formula weights. We use three frozen NIST23 checkpoints trained with seeds 1234, 2345, and 3456. For molecule x, let S_x be its highest-weight formulas with at least four matching events, keeping at most eight formulas. For each selected formula, event ranks use the mean attention across eight heads, breaking ties by local event identifier. We remove the same top-k events from every head, for $k = 1, 2$, and renormalize the remaining matching weights. Matched random removal draws k events uniformly ten times. We also remove the lowest-attention events as a further comparison.

All interventions take place after latent encoding. Atom and latent states remain fixed, and the learned output-projection bias is retained. If p_j is the original formula weight and $p_j^{(-A)}$ its weight after removing event set A for that formula, define

$$D_j(A) = p_j - p_j^{(-A)}, \qquad G = \frac{1}{|\mathcal{X}'|} \sum_{x \in \mathcal{X}'} \frac{1}{|S_x|} \sum_{j \in S_x} \left[D_j(A_{\text{top}}) - \frac{1}{10} \sum_{d=1}^{10} D_j(A_{\text{random},d}) \right]. \quad (13)$$

Negative drops are retained. Of the 17,879 test molecules, 17,874 have eligible formulas and enter $\mathcal{X}'$. The remaining five are no-ops in the spectrum-level evaluation. The 95% intervals in Table 5 use 10,000 paired InChIKey-cluster bootstrap samples and measure test-molecule sampling uncertainty.

All six checkpoint–k gaps are positive. In the spectrum-level probe, we erase events for every selected formula together and recompute the rendered spectrum. For random removal, the ten scores are averaged within each molecule before aggregation. Table 6 aggregates all three training seeds. Removing the top two events gives 0.8604 DP and 0.5508 Top_1, compared with 0.8724 and 0.5786 for random removal. Across seeds, its additional drop relative to random removal is 0.01204 ± 0.00129 DP and 0.02787 ± 0.00130 Top_1 (mean $\pm$ sample SD). Top-k removal gives the lowest DP, SDP, COS, and Top_1 among the three rules at every checkpoint and both values of k. Table 7 reports each checkpoint. The analysis links attention to support for formula hypotheses in the model. The probe uses the structure-derived candidates defined in Section 3.2.

Seed	k	Top drop	Random drop	G	95% interval
1234	1	0.00592	0.00058	0.00534	[0.00526, 0.00542]
1234	2	0.01449	0.00160	0.01289	[0.01274, 0.01304]
2345	1	0.00528	0.00051	0.00478	[0.00469, 0.00486]
2345	2	0.01357	0.00143	0.01214	[0.01198, 0.01230]
3456	1	0.00621	0.00055	0.00566	[0.00557, 0.00575]
3456	2	0.01540	0.00153	0.01387	[0.01370, 0.01404]

Table 5: Signed hypothesis-weight gap G under event erasure. Intervals are 95% molecule-cluster bootstrap intervals.

Erased events	k	DP	SDP	COS	Top_1
None	0	0.8743 ± 0.0015	0.8766 ± 0.0011	0.8174 ± 0.0026	0.5832 ± 0.0048
Top	1	0.8721 ± 0.0016	0.8755 ± 0.0011	0.8113 ± 0.0028	0.5750 ± 0.0050
Random	1	0.8738 ± 0.0015	0.8763 ± 0.0011	0.8161 ± 0.0027	0.5817 ± 0.0043
Lowest	1	0.8741 ± 0.0015	0.8764 ± 0.0011	0.8172 ± 0.0027	0.5832 ± 0.0044
Top	2	0.8604 ± 0.0025	0.8660 ± 0.0017	0.7886 ± 0.0041	0.5508 ± 0.0052
Random	2	0.8724 ± 0.0015	0.8754 ± 0.0011	0.8132 ± 0.0028	0.5786 ± 0.0043
Lowest	2	0.8730 ± 0.0015	0.8753 ± 0.0011	0.8146 ± 0.0029	0.5790 ± 0.0041

Table 6: Spectrum scores after event erasure on NIST23. Mean $\pm$ sample SD across three training seeds. Random removal averages ten draws within each molecule.

Seed	Erased events	k	DP	SDP	COS	Top_1
1234	None	0	0.8729	0.8758	0.8147	0.5782
1234	Top	1	0.8707	0.8748	0.8086	0.5697
1234	Random	1	0.8724	0.8755	0.8133	0.5770
1234	Lowest	1	0.8727	0.8756	0.8143	0.5784
1234	Top	2	0.8592	0.8656	0.7863	0.5458
1234	Random	2	0.8710	0.8746	0.8104	0.5740
1234	Lowest	2	0.8715	0.8744	0.8114	0.5748
2345	None	0	0.8758	0.8779	0.8200	0.5878
2345	Top	1	0.8739	0.8768	0.8143	0.5796
2345	Random	1	0.8753	0.8776	0.8188	0.5855
2345	Lowest	1	0.8756	0.8777	0.8197	0.5869
2345	Top	2	0.8633	0.8679	0.7933	0.5561
2345	Random	2	0.8741	0.8767	0.8160	0.5825
2345	Lowest	2	0.8745	0.8765	0.8169	0.5829
3456	None	0	0.8741	0.8761	0.8177	0.5837
3456	Top	1	0.8717	0.8749	0.8109	0.5756
3456	Random	1	0.8735	0.8758	0.8163	0.5826
3456	Lowest	1	0.8740	0.8760	0.8176	0.5843
3456	Top	2	0.8587	0.8646	0.7861	0.5504
3456	Random	2	0.8721	0.8749	0.8132	0.5794
3456	Lowest	2	0.8730	0.8751	0.8155	0.5793

Table 7: Event-erasure spectrum scores for each training checkpoint. Every score averages 17,879 test molecules.

Paired-margin probe. The paired-margin probe extends the intervention to same-formula isomer pairs, the comparison that decides candidate rankings. For observed spectra y_a, y_b and predictions $\widehat{y}_a, \widehat{y}_b$, the margin

$$M_{ab} = \tfrac{1}{2}[\mathrm{SDP}(y_a, \widehat{y}_a) + \mathrm{SDP}(y_b, \widehat{y}_b) - \mathrm{SDP}(y_a, \widehat{y}_b) - \mathrm{SDP}(y_b, \widehat{y}_a)] \tag{14}$$

is positive when each prediction matches its own measurement better than the counterpart's. Pairs follow the same same-formula, distinct-connectivity matching rules as retrieval; the 43,839 pairs cover 3,230 formula groups, and we report all pairs, their nearest-neighbor union, and mutual-nearest pairs. Atom and latent states stay frozen, and events are removed from both candidate predictions under the same three rules ($k = 1, 2$; random removal averages ten draws). The intervention acts on CLEF's event attention and has no baseline counterpart. Figure 7 shows the margin changes for the three rules.

The full margin is 0.2353 ± 0.0012 (mean $\pm$ sample SD across the three checkpoints). Removing the top-attended event ($k = 1$) lowers it by 4.47×10^{-3}, roughly ten times the 0.42×10^{-3} drop under random removal, while lowest-attention removal leaves it essentially unchanged (-0.41×10^{-3}); at $k = 2$ the top-event drop grows to 18.07×10^{-3} against 1.49×10^{-3} for random removal, and the ordering top $>$ random $>$ lowest holds in all three pair panels (nearest-neighbor union: 4.02 and 16.58×10^{-3}; mutual-nearest: 3.59 and 15.19×10^{-3}). The events the model attends to most are therefore associated with the differences between same-formula isomers, and removing them weakens the comparison that ranks same-formula candidates. Error bars in Figure 7 are the sample SD across the three checkpoints.

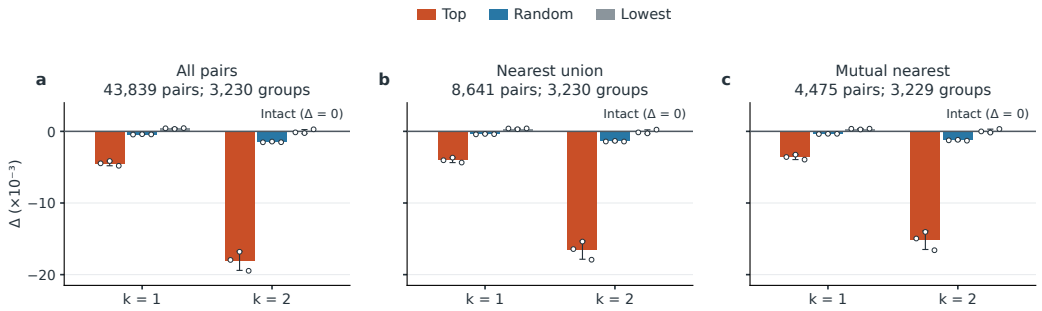

Figure 7: Same-formula isomer-pair margin under frozen-model event erasure: (a) all pairs, (b) nearest-neighbor union, and (c) mutual-nearest pairs; panel headings give pair and molecular-formula-group counts. The vertical axis is erased minus intact (10^{-3}): zero marks the intact model, and farther-downward bars indicate larger margin losses. Orange/blue/grey mark highest-attention/random/lowest-attention removal; k is events removed per selected formula. Bars/points/error bars show means/individual values/sample SD across three checkpoints; random removal averages ten draws per checkpoint.

C.2 Observation-perturbation scores

Tables 8 and 9 list all seven predictors behind Figure 5. Only measured spectra are perturbed; predictions stay fixed, and all models use matched perturbation instances. Gaussian, Poisson, and threshold rows represent intensity noise, sampling at 1,000 base-peak counts, and removing peaks below the stated base-peak fraction, respectively. Scores are three-seed means; higher is better. SN/FN = RASSP:SN/FN; MF/LF = MassFormer/MolFormer.

Perturbation	CLEF	SN	FN	NEIMS	Hybrid	MF	LF
Gaussian 2%	0.8742	0.8388	0.8379	0.7718	0.6307	0.8480	0.7998
Gaussian 5%	0.8740	0.8386	0.8377	0.7717	0.6305	0.8478	0.7996
Gaussian 10%	0.8733	0.8379	0.8370	0.7710	0.6301	0.8471	0.7990
Gaussian 20%	0.8702	0.8350	0.8341	0.7686	0.6279	0.8442	0.7963
Poisson (1k)	0.8707	0.8357	0.8341	0.7663	0.6269	0.8448	0.7964
Threshold 1%	0.8504	0.8196	0.8106	0.7194	0.6003	0.8261	0.7749
Threshold 2%	0.8268	0.7999	0.7848	0.6744	0.5760	0.8025	0.7493
Threshold 5%	0.7724	0.7522	0.7267	0.5853	0.5263	0.7465	0.6906

Table 8: NIST23 DP under the eight observation perturbations in Figure 5.

Perturbation	CLEF	SN	FN	NEIMS	Hybrid	MF	LF
Gaussian 2%	0.8765	0.8514	0.8075	0.7800	0.7574	0.8346	0.8119
Gaussian 5%	0.8763	0.8512	0.8073	0.7799	0.7572	0.8344	0.8117
Gaussian 10%	0.8756	0.8505	0.8067	0.7792	0.7568	0.8338	0.8111
Gaussian 20%	0.8723	0.8474	0.8036	0.7768	0.7536	0.8306	0.8080
Poisson (1k)	0.8736	0.8485	0.8045	0.7762	0.7547	0.8319	0.8091
Threshold 1%	0.8612	0.8346	0.7888	0.7483	0.7401	0.8199	0.7958
Threshold 2%	0.8470	0.8197	0.7731	0.7210	0.7267	0.8065	0.7798
Threshold 5%	0.8097	0.7811	0.7339	0.6600	0.6935	0.7715	0.7386

Table 9: NIST23 SDP under the eight observation perturbations in Figure 5.