OMNISPECGEN: WHERE TO FUSE MULTIMODAL SPECTRA FOR MOLECULAR STRUCTURE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

De novo molecular structure prediction from spectral measurements is a fundamental task in chemistry. Although existing models have primarily focused on single modalities or predefined modality sets, the measurements available in practice vary across molecules, calling for flexible use of complementary information from arbitrary subsets of modalities. The heterogeneous formats of these spectra motivate modality-specific processing, but the generation pipeline must transition to joint processing to generate a single molecule. This raises a central architectural question: *To best exploit complementarity, how much processing should remain modality-specific before joint processing begins?* To address this question, we introduce *OmniSpecGen*, a unified framework for comparing four fusion stages within a common diffusion architecture. Our comparison identifies fusion of modality-specific representations before pooling as the most effective strategy, with the resulting model outperforming the current state-of-the-art model for flexible spectrum-to-structure generation across the evaluated multimodal configurations. These findings clarify when to begin joint processing and provide an effective framework for molecular structure elucidation.

1 INTRODUCTION

De novo molecular structure prediction from measured spectra is a fundamental task in chemistry, with applications ranging from identification of reaction products to the structure elucidation of novel natural products (Silverstein et al., 2014). A wide variety of spectrometric techniques are available for this purpose, including infrared (IR), Raman, ultraviolet–visible (UV–Vis), nuclear magnetic resonance (NMR), and mass spectrometry (MS), each probing different aspects of molecular structure through distinct physical and chemical processes. Together, these complementary measurements provide a more comprehensive view of a molecule than any single modality alone, motivating multimodal approaches that jointly interpret multiple types of spectral measurements for reliable molecular structure elucidation (Wu et al., 2025; Priessner et al., 2026).

Recent generative approaches have extended molecular structure elucidation to multimodal spectral inputs, ranging from combinations of a few complementary spectra to broader sets of NMR, IR, Raman, UV–Vis, and MS measurements (Devata et al., 2024; Priessner et al., 2026; Wang et al., 2025; Yu et al., 2026). However, these approaches are generally developed around predefined sets of input modalities. In practice, the available measurements may instead vary from sample to sample—for example, NMR and IR may be available for one molecule, while another may have MS and Raman measurements. Although prior work has examined robustness to missing measurements, the flexible generation setting illustrated in Fig. 1, in which a single model generates molecular structures *de novo* from arbitrary subsets of heterogeneous spectral modalities, remains largely underexplored.

A key difficulty in supporting arbitrary combinations of spectral modalities is that these heterogeneous spectra do not naturally share a common input format. For example, some spectra are represented as continuous curves and others as peak sets, while characteristic peak shapes also vary across modalities. SpectraLLM by Su et al. (2026) sidesteps this difficulty by representing the available spectral measurements in a common language-based format and concatenating them into a single sequence, thereby enabling a single model to handle different combinations of spectral inputs. However, given the substantially different characteristics of individual spectral modalities, *it may be more natural to preserve their modality-specific information using dedicated encoders.* These encoders

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.

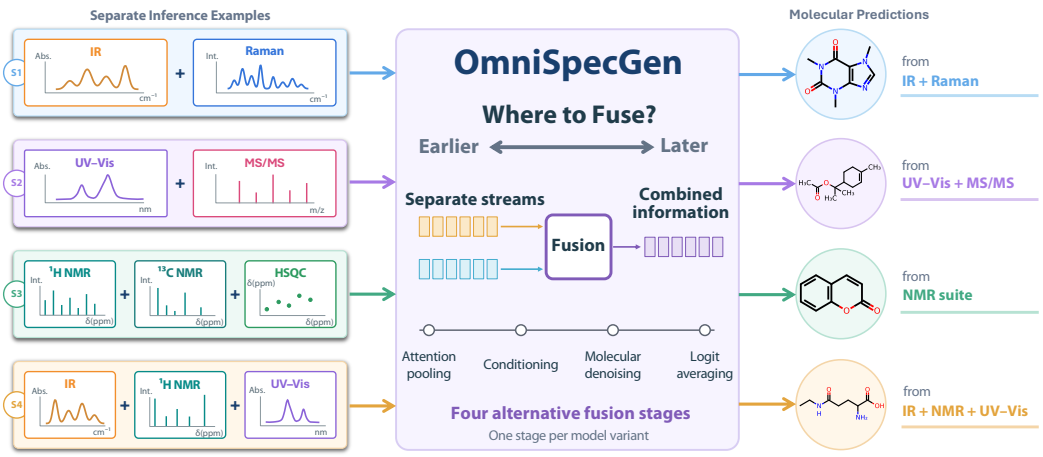

Figure 1: **Flexible multimodal molecular generation and the choice of fusion stage.** Each row illustrates a separate inference example in which OmniSpecGen generates a molecular structure *de novo* from a different subset of spectral measurements. Modality-specific encoders produce separate spectral representations, raising the question of where in the molecular generation pipeline their information should be integrated.

produce a separate spectral representation for each available modality, and these representations can be integrated at different stages to generate a single molecular structure. At one extreme, the model can fuse these representations immediately after encoding for joint processing; at the other, it can make separate predictions conditioned on each modality and combine them only at the output level. This range of possibilities raises a central architectural question: *at which stage should information from different spectral modalities be fused to most effectively exploit their complementarity?*

To address this question, we introduce *OmniSpecGen*, a unified framework for comparing multimodal fusion at different stages within a common generative architecture. Using a GenMol-style discrete diffusion model as the generative backbone (Lee et al., 2025), we identify distinct fusion points along the generation pipeline, spanning spectral representations, conditioning embeddings, molecular denoising, and output predictions. By comparing these fusion points under a unified framework, we determine at which stage multimodal integration can most effectively leverage the complementary information provided by multiple spectra for *de novo* molecular structure elucidation from arbitrary subsets of spectral measurements.

Our contributions are threefold: (i) We introduce OmniSpecGen, a unified framework for comparing four fusion stages within a common diffusion architecture. Each variant supports molecular generation from arbitrary spectral subsets using a single trained model. (ii) We show that molecular structure elucidation performance generally declines as multimodal fusion is delayed to later stages of the generation pipeline. (iii) We demonstrate that the representation-fusion variant achieves higher structural similarity than the state-of-the-art model SpectraLLM across the evaluated multimodal configurations.

2 Preliminaries

We begin by formulating the task of *de novo* molecular structure prediction from arbitrary subsets of spectral modalities and introducing the conditional generative backbone used in this study. Let $\mathcal{M}$ denote the set of available spectral modalities (*e.g.*, IR, NMR, and MS). For each molecule, a subset $\mathcal{S} \subseteq \mathcal{M}$ is observed, yielding a collection of spectral measurements $\mathcal{X} = \{\boldsymbol{x}^m \mid m \in \mathcal{S}\}$. Given $\mathcal{X}$, the goal is to generate the corresponding molecular structure $\boldsymbol{y}$, represented as a molecular string, by modeling $p(\boldsymbol{y} \mid \mathcal{X})$. The subset $\mathcal{S}$ may vary across samples, requiring the model to operate on arbitrary combinations of available spectral modalities. We consider a fully *de novo* setting, without access to a candidate structure database or molecular formula.

2.1 MASKED DISCRETE DIFFUSION

We use a masked discrete diffusion model as the generative backbone (Austin et al., 2021). A molecule is represented as a sequence of SAFE tokens, $\boldsymbol{y} = \boldsymbol{y}_0 = (y_{0,1}, \ldots, y_{0,L})$ (Noutahi et al., 2024). The forward process progressively replaces tokens with a mask token `[MASK]`, producing increasingly corrupted sequences,

$$\boldsymbol{y}_0 \underset{\text{unmask}}{\overset{\text{mask}}{\rightleftarrows}} \boldsymbol{y}_1 \underset{\text{unmask}}{\overset{\text{mask}}{\rightleftarrows}} \boldsymbol{y}_2 \underset{\text{unmask}}{\overset{\text{mask}}{\rightleftarrows}} \cdots \underset{\text{unmask}}{\overset{\text{mask}}{\rightleftarrows}} \boldsymbol{y}_T = [\texttt{MASK}]^L. \tag{1}$$

At each diffusion step t, the denoising network $f_{\boldsymbol{\theta}}$ takes the corrupted sequence $\boldsymbol{y}_t$ and conditioning embeddings $C \in \mathbb{R}^{N_e \times d}$ derived from $\mathcal{X}$ as inputs and predicts token-wise logits: $O_t = f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, C) \in \mathbb{R}^{L \times |\mathcal{V}|}$, where N_e is the number of conditioning vectors, d is their embedding dimension, L is the sequence length, $|\mathcal{V}|$ is the size of the SAFE token vocabulary, and $\boldsymbol{\theta}$ denotes the trainable parameters. Each row of O_t corresponds to a position in the molecular sequence and specifies the predicted logits over possible clean tokens at that position, from which a prediction of the clean sequence $\hat{\boldsymbol{y}}_0$ is obtained.

During generation, the reverse process starts from a fully masked sequence $\boldsymbol{y}_T = [\mathrm{MASK}]^L$. At each step t, the model predicts O_t from the current corrupted sequence $\boldsymbol{y}_t$ and conditioning representation C, and the resulting clean-token predictions are used to progressively unmask the sequence. Repeating this process yields increasingly complete molecular sequences until the final molecular sequence $\boldsymbol{y}_0$ is obtained.

2.2 TRANSFORMER-BASED CONDITIONING

Following GenMol by Lee et al. (2025), we parameterize the denoising network $f_{\boldsymbol{\theta}}$ using a Transformer (Vaswani et al., 2017). The corrupted SAFE sequence $\boldsymbol{y}_t$ is first mapped to a sequence of molecular hidden representations $H^{(0)}_{\text{mol}} \in \mathbb{R}^{L \times d}$. Starting from a standard Transformer encoder, which stacks self-attention and feed-forward sublayers, we replace selected self-attention sublayers with cross-attention to incorporate the conditioning embeddings C. While self-attention updates each molecular representation using information from other molecular tokens, cross-attention instead updates it using the conditioning embeddings, with the molecular representations serving as queries and C as keys and values:

$$\mathrm{CrossAttn}(H_{\text{mol}}, C) = \mathrm{Attn}\left(Q = H_{\text{mol}}, K = C, V = C\right). \tag{2}$$

The resulting molecular representations are finally mapped to logits over the SAFE vocabulary to predict the clean token at each sequence position, yielding $\hat{\boldsymbol{y}}_0$.

3 WHERE TO FUSE: THE OMNISPECGEN FRAMEWORK

We now consider how to construct the conditioning representation C from the available spectral measurements. With a dedicated encoder for each modality, the measurements initially form separate representation streams, raising the question: *how long should these streams remain separate before their information is integrated?* To study this question, we introduce OmniSpecGen, a unified framework that varies the stage of multimodal fusion within the common generative architecture described in Sec. 2.

To identify the possible fusion stages, we begin with the case in which only one spectral modality m is observed. The modality-specific encoder g_m maps the observed spectrum $\boldsymbol{x}^m$ to a sequence of token-level spectral representations,

$$H^m = g_m(\boldsymbol{x}^m) \in \mathbb{R}^{L_m \times d}, \tag{3}$$

where L_m denotes the number of spectral tokens. Each row of H^m is a d-dimensional feature vector that may correspond, for example, to an individual peak or a local region of a continuous spectrum. Since L_m may vary across samples and modalities, we summarize these variable-length representations into a fixed number n_e of embeddings using attention pooling (Lee et al., 2019),

$$E^m = \mathrm{Pool}(H^m) = \mathrm{Attn}\left(Q = Q_{\text{pool}}, K = H^m, V = H^m\right) \in \mathbb{R}^{n_e \times d}, \tag{4}$$

Table 1: **Overview of the four multimodal fusion strategies.** The strategies share a processing hierarchy from spectral representations H to embeddings E, conditioning representations C, and molecular predictions O, while progressively delaying multimodal integration. A superscript m denotes modality-specific processing before fusion.

Fusion strategy	**Embedding** E	**Condition** C	**Molecular generation**
Representation	$\mathrm{Pool}(\mathrm{Concat}_m H^m) \to E$	$T_{\text{cond}}(E) \to C$	$f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, C) \to O$
Latent	$\mathrm{Pool}(H^m) \to E^m$	$T_{\text{cond}}(\mathrm{Concat}_m E^m) \to C$	$f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, C) \to O$
Molecular-interaction	$\mathrm{Pool}(H^m) \to E^m$	$T_{\text{cond}}(E^m) \to C^m$	$f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, \mathrm{Concat}_m C^m) \to O$
Late	$\mathrm{Pool}(H^m) \to E^m$	$T_{\text{cond}}(E^m) \to C^m$	$f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, C^m) \to O^m \to \mathrm{Mean}_m(O^m)$

where $Q_{\text{pool}} \in \mathbb{R}^{n_e \times d}$ is a set of learned queries. The resulting fixed-size spectral embedding $E^m \in \mathbb{R}^{n_e \times d}$ is then processed by a conditioning Transformer T_{cond} to produce $C^m = T_{\text{cond}}(E^m) \in \mathbb{R}^{n_e \times d}$, which serves as the conditioning representation supplied to the molecular denoiser $f_{\boldsymbol{\theta}}$. As introduced in Sec. 2, the denoiser takes the corrupted molecular sequence $\boldsymbol{y}_t$ together with C^m and produces token-wise molecular predictions as $O_t^m = f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, C^m)$.

With a single observed modality, the complete processing pathway is therefore

$$H^m \xrightarrow{\text{Pool}} E^m \xrightarrow{T_{\text{cond}}} C^m \xrightarrow{f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, \cdot)} O_t^m. \tag{5}$$

Fusion strategies. We now extend this construction to multiple observed modalities. The hierarchy in Eq. 5 provides four possible stages at which information from different modalities can be combined: token-level spectral representations H^m, pooled embeddings E^m, conditioning representations C^m, or molecular predictions O_t^m. These choices yield the four fusion strategies summarized in Table 1 and Fig. A1 in Appendix A. We use the superscript m for representations computed separately for each modality and omit it for those incorporating information from multiple modalities. In **representation fusion**, the token-level representations $\{H^m\}_{m \in \mathcal{S}}$ are concatenated and then processed by attention pooling using $N_e = n_e|\mathcal{M}|$ learned query vectors. Attention pooling is thus the first operation to jointly process information across modalities and produces the shared embedding E. In **latent fusion**, each modality's representations are first pooled independently into E^m, after which the modality-specific embeddings are concatenated and jointly processed by T_{cond} to construct the shared conditioning representation C. **Molecular-interaction fusion** preserves modality-specific processing through both pooling and conditioning, yielding $\{C^m\}_{m \in \mathcal{S}}$, which are concatenated only when supplied to the molecular denoiser; cross-modal interaction therefore first occurs within its cross-attention layers. Finally, **late fusion** retains independent modality pathways throughout molecular denoising, producing modality-specific logits $O_t^m = f_{\boldsymbol{\theta}}(\boldsymbol{y}_t, C^m)$ that are combined only at the prediction level as $O_t = \mathrm{Mean}_{m \in \mathcal{S}}(O_t^m)$.

The four strategies delay the fusion from attention pooling to conditioning, molecular denoising, and finally prediction aggregation, shifting the balance between cross-modal interaction and modality-specific processing. Earlier fusion allows complementary cues to inform more stages of processing but requires a larger portion of the network to accommodate arbitrary modality combinations. Fusion before pooling also enables flexible allocation of a shared embedding budget across modalities, although differences in token counts and information density may cause some modalities to dominate joint pooling. These trade-offs make it unclear a priori which fusion stage is most effective.

Implementation details. For concreteness, Fig. 2 illustrates the architecture of the OmniSpec-Gen variant using representation fusion. The other fusion strategies use the same architectural components and differ only in the stage at which modality-specific streams are first jointly processed, as summarized in Table 1. All hidden representations have dimension $d = 256$. Each modality is summarized using $n_e = 4$ learned pooling queries, while representation fusion uses a shared query set with a matched overall embedding budget. A learned modality embedding is added before representations enter a shared processing stage to preserve their modality identity after fusion. The conditioning network T_{cond} is implemented as a two-layer Transformer encoder with eight attention heads and a feed-forward hidden dimension of 1,024. The molecular denoiser described in Sec. 2 consists of eight Transformer layers arranged as S $\to$ C $\to$ S $\to$ C $\to$ S $\to$ C $\to$ S $\to$ S, where S and C denote self- and cross-attention layers, respectively. Each denoiser layer uses eight attention

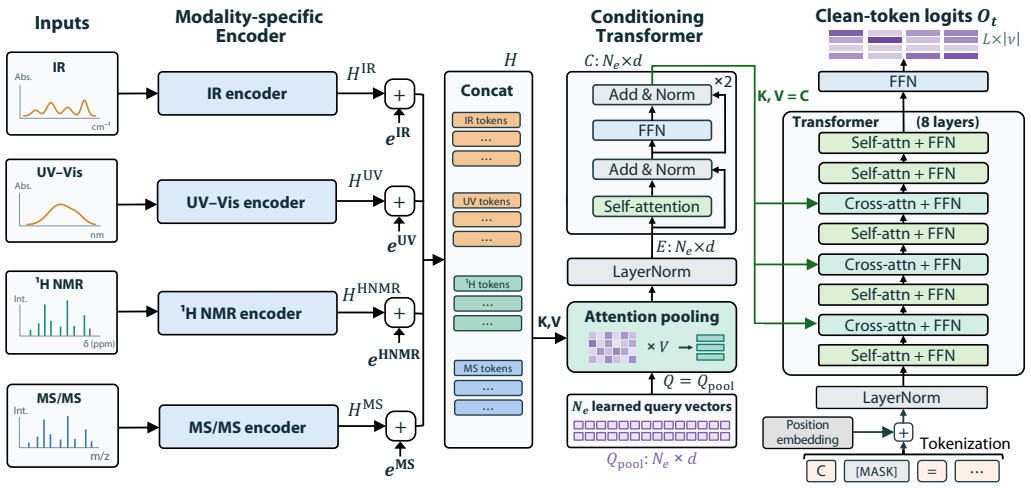

Figure 2: **Overall architecture of the OmniSpecGen variant using representation fusion.** Modality-specific spectral representations, augmented with learned modality embeddings, are concatenated and jointly compressed using N_e learned pooling queries. A two-layer Transformer encoder T_{cond} produces the conditioning representation C, which conditions the molecular Transformer denoiser through cross-attention. The token prediction head maps the final hidden representation at each position to a vector of logits over the vocabulary.

heads and a feed-forward sublayer with a hidden dimension of 1,024. The final molecular representations are mapped to logits over the SAFE vocabulary. The model is trained to reconstruct masked molecular tokens by minimizing a cross-entropy loss. Each variant has approximately 30 million trainable parameters. Details are provided in Appendix A.

4 Experiments

We evaluated OmniSpecGen on flexible multimodal spectrum-to-structure generation, where a single model generates the corresponding molecular structure *de novo* from an arbitrary combination of the available spectral modalities. To enable direct comparison with SpectraLLM (Su et al., 2026), the current state of the art for this task, we followed its experimental setting and used its reported results as the baseline. Training and generation details are provided in Appendix B.

Dataset. Following SpectraLLM's dataset construction protocol, we constructed a unified corpus from four public data sources. **QM9S** provides simulated IR, Raman, and UV–Vis spectra for 129,817 small organic molecules (Zou et al., 2023). **The Multimodal Spectroscopic Dataset** contains simulated IR, mass, and multiple NMR spectra, including ^{1}H NMR, ^{13}C NMR, and HSQC, for 794,403 molecules derived from USPTO reactions (Alberts et al., 2024b). **MassSpecGym** provides 231,104 experimentally acquired high-resolution MS/MS spectra (Bushuiev et al., 2024). **MassBank** contributes 122,746 experimental MS and MS/MS spectra collected under diverse acquisition conditions (Horai et al., 2010). We standardized molecular structures and associated each normalized molecular identity with its available spectra from the four sources. We used the training, validation, and test splits released by SpectraLLM. The resulting corpus contains 927,669 unique molecules, with 740,993, 92,840, and 93,836 molecules in the training, validation, and test sets, respectively.

Evaluation metrics. For evaluation, we generated one candidate molecular structure for each spectral input. Following SpectraLLM, we assessed the generated structures using five complementary metrics. **Validity** is the fraction of generated outputs that can be successfully decoded and sanitized as molecular graphs using RDKit (Landrum et al., 2016). **Tanimoto (ECFP4)** measures the Tanimoto similarity (Bajusz et al., 2015) between the predicted and reference structures using ECFP4 circular fingerprints, which encode local atomic environments within a radius of two bonds (Rogers & Hahn, 2010). **Functional Group** measures the Jaccard similarity between the sets

of 17 common functional-group types detected in the two structures using SMARTS patterns. **Tanimoto (MACCS)** computes the Tanimoto similarity between their 166-bit MACCS structural keys, which represent a predefined collection of molecular substructures (Durant et al., 2002). **Fraggle** measures fragment-level similarity by decomposing the molecules into substructures and identifying their shared structural fragments using RDKit's FraggleSim implementation (Landrum et al., 2016). All four structural similarity metrics range from 0 to 1, with higher values indicating greater similarity and 1 denoting complete agreement under the corresponding metric.

4.1 Fusion strategies and benchmark performance

Tables 2 and 3 compare the performance of the four fusion strategies within OmniSpecGen with the reported results of SpectraLLM (Su et al., 2026) on multimodal input configurations from QM9S and the Multimodal Spectroscopic Dataset, respectively. Each fusion variant uses a single trained model across the different input configurations. Single-modality results on these datasets, along with evaluations on MS spectra from MassBank and MassSpecGym, are provided in Appendix C.

Across these benchmarks, structural similarity generally decreases as fusion is delayed from spectral representations to pooled embeddings, conditioning representations, and finally molecular predictions. For the complete IR+Raman+UV–Vis input on QM9S, ECFP4 similarity decreases from 0.7624 with representation fusion to 0.7282 with latent fusion, 0.7050 with molecular-interaction fusion, and 0.5854 with late fusion. The same ordering holds for all four structural similarity metrics in this configuration. On the Multimodal Spectroscopic Dataset, this ordering also holds across all reported input configurations for ECFP4, functional-group, and MACCS similarities. One possible explanation for this trend is that earlier fusion allows complementary spectral cues to jointly guide the compression and transformation of spectral information. Delaying fusion requires more of this processing to occur independently within each modality, potentially discarding cues that are weak in isolation but informative when interpreted together. Appendix D shows that the advantage of earlier multimodal integration largely persists when the denoising network $f_{\boldsymbol{\theta}}$ incorporates the conditioning embeddings C through in-context concatenation, an established alternative to cross-attention studied in DiT (Peebles & Xie, 2023). Appendix E further provides a broader comparison of conditioning strategies, including adaptive layer normalization.

The fusion strategies also differ in how effectively they use additional spectral modalities. Representation, latent, and molecular-interaction fusion achieve higher structural similarity as spectral modalities are added. For example, with representation fusion on QM9S, ECFP4 similarity increases

Table 2: **Comparison of the four fusion strategies and SpectraLLM on multimodal input configurations from the QM9S dataset.** † denotes values reported in Table 19 of SpectraLLM (Su et al., 2026) for models separately trained on each input configuration, rather than results obtained from our reimplementation. **Bold** and underline indicate the best and second-best results, respectively, for each metric within each input configuration.

Inputs	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
IR + Raman	SpectraLLM†	**99.63%**	0.5320	0.8834	0.7363	0.4806
	Representation	98.80%	**0.7535**	**0.9600**	**0.9081**	**0.8740**
	Latent	98.91%	0.7142	0.9552	0.8932	0.8657
	Mol. interaction	98.71%	0.7026	0.9521	0.8864	0.8591
	Late	99.27%	0.5813	0.9426	0.8373	0.8151
Raman + UV–Vis	SpectraLLM†	**99.45%**	0.4594	0.8431	0.6718	0.4561
	Representation	99.17%	**0.7272**	**0.9566**	**0.8943**	**0.8658**
	Latent	98.84%	0.6599	0.9407	0.8635	0.8443
	Mol. interaction	98.42%	0.6492	0.9286	0.8512	0.8336
	Late	99.37%	0.5177	0.9164	0.7856	0.7893
UV–Vis + IR	SpectraLLM†	**99.63%**	0.2088	0.6791	0.4471	0.3333
	Representation	98.92%	**0.5844**	**0.9439**	**0.8326**	**0.8015**
	Latent	98.92%	0.5104	0.9350	0.7937	0.7898
	Mol. interaction	98.71%	0.4897	0.9270	0.7783	0.7796
	Late	99.46%	0.4478	0.9163	0.7498	0.7584
IR + Raman + UV–Vis	SpectraLLM†	98.53%	0.4053	0.8370	0.6501	0.4479
	Representation	98.89%	**0.7624**	**0.9648**	**0.9134**	**0.8782**
	Latent	99.04%	0.7282	0.9619	0.9026	0.8703
	Mol. interaction	98.88%	0.7050	0.9538	0.8876	0.8594
	Late	**99.53%**	0.5854	0.9461	0.8352	0.8202

Table 3: **Comparison on multimodal input configurations from the Multimodal Spectroscopic Dataset.** NMR denotes the joint use of ^{13}C NMR, ^{1}H NMR, and HSQC spectra. † denotes results reported in Table 4 of SpectraLLM (Su et al., 2026).

Inputs	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
IR + MS/MS	SpectraLLM†	99.25%	0.2300	0.6345	0.4887	0.4566
	Representation	99.02%	**0.2881**	**0.7467**	**0.6097**	0.5189
	Latent	99.09%	0.2746	0.7411	0.5990	**0.5524**
	Mol. interaction	99.11%	0.2640	0.7344	0.5873	0.5432
	Late	**99.75%**	0.2296	0.6707	0.5066	0.5200
NMR	SpectraLLM†	98.92%	0.4151	0.7209	0.6367	0.5862
	Representation	98.68%	**0.4445**	**0.7852**	**0.6970**	0.6258
	Latent	98.60%	0.4268	0.7799	0.6920	**0.6512**
	Mol. interaction	98.66%	0.3973	0.7605	0.6643	0.6313
	Late	**99.85%**	0.2334	0.5939	0.4548	0.5010
NMR + IR	SpectraLLM†	99.60%	0.4121	0.7764	0.6575	0.5809
	Representation	98.49%	**0.5160**	**0.8473**	**0.7675**	0.6763
	Latent	98.46%	0.4854	0.8377	0.7502	**0.6855**
	Mol. interaction	98.43%	0.4525	0.8183	0.7249	0.6617
	Late	**99.81%**	0.2489	0.6272	0.4835	0.5174
NMR + MS/MS	SpectraLLM†	98.37%	0.4518	0.7618	0.6760	0.6063
	Representation	98.49%	**0.4791**	**0.8307**	**0.7422**	0.6521
	Latent	98.44%	0.4576	0.8242	0.7322	**0.6685**
	Mol. interaction	98.43%	0.4220	0.8057	0.7047	0.6428
	Late	**99.84%**	0.2298	0.6053	0.4596	0.5023
NMR + IR + MS/MS	SpectraLLM†	99.79%	0.4875	0.8103	0.7099	0.6222
	Representation	98.37%	**0.5351**	**0.8680**	**0.7854**	0.6881
	Latent	98.30%	0.5081	0.8612	0.7734	**0.6972**
	Mol. interaction	98.20%	0.4703	0.8412	0.7461	0.6689
	Late	**99.86%**	0.2440	0.6323	0.4833	0.5189

from 0.7535 with IR+Raman to 0.7624 with IR+Raman+UV–Vis. On the Multimodal Spectroscopic Dataset, it similarly increases from 0.4445 with NMR to 0.5160 with NMR+IR and 0.5351 with NMR+IR+MS/MS. All four structural similarity metrics improve along both sequences, supporting the model's ability to exploit complementary spectral information. Late fusion, however, does not always benefit from additional modalities. In this strategy, the model performs a separate forward pass conditioned on each spectral modality and averages the resulting logits. The absence of direct interaction between spectral modalities within the denoiser may limit its ability to resolve ambiguities that require jointly interpreting complementary spectral cues.

Beyond the comparison among fusion strategies, the representation-fusion variant outperforms SpectraLLM on all four structural similarity metrics across all multimodal input configurations reported for QM9S and the Multimodal Spectroscopic Dataset. For the complete input configurations, ECFP4 similarity is 0.7624 versus 0.4053 on QM9S and 0.5351 versus 0.4875 on the Multimodal Spectroscopic Dataset. Notably, these results are achieved by a single model of approximately 30M parameters shared across all input configurations, whereas SpectraLLM uses a 32B-parameter backbone. The best-performing fusion stage in our comparison thus yields a model that surpasses the current state of the art, lending practical relevance to the comparison of fusion stages.

We further examine the generalization of representation fusion to unseen modality combinations in Appendix F and provide qualitative comparisons of molecular structures generated by the four fusion strategies and SpectraLLM under different spectral input configurations in Appendix G.

4.2 Fusion timing and pooling capacity

The results in Sec. 4.1 favor representation fusion, but the comparison also involves different pooling budgets at inference. With the default setting of $n_e = 4$, representation fusion uses $n_e|\mathcal{M}| = 28$ shared pooling queries regardless of the observed modality subset, whereas latent fusion uses only $n_e|\mathcal{S}|$ queries. This raises the question of whether the advantage of representation fusion can be explained by its larger active query budget.

To examine this question, we trained both fusion variants with $n_e \in \{1, 2, 4, 8, 16\}$, keeping all other architectural and training settings unchanged. We additionally included three shared pooling queries in total for representation fusion, corresponding to $n_e = 3/7$, and $n_e = 32$ for latent fusion. Figure 3 plots Tanimoto (ECFP4) similarity against the number of active pooling queries

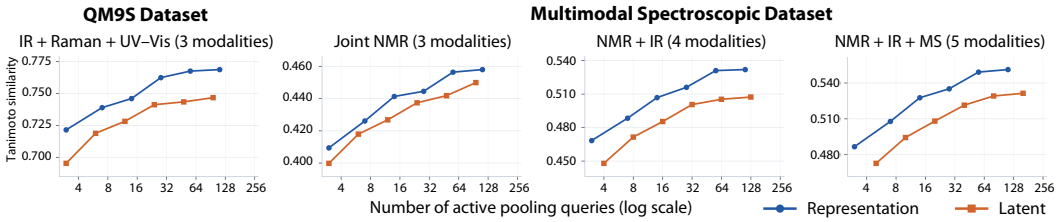

Figure 3: **Effect of pooling capacity on fusion performance.** Tanimoto (ECFP4) similarity is plotted against the number of active pooling queries (log scale) for representation and latent fusion across four input configurations from QM9S and the Multimodal Spectroscopic Dataset. We used $n_e \in \{3/7, 1, 2, 4, 8, 16\}$ for representation fusion and $n_e \in \{1, 2, 4, 8, 16, 32\}$ for latent fusion, keeping all other settings fixed. The active query count is $n_e|\mathcal{M}|$ for representation fusion and $n_e|\mathcal{S}|$ for latent fusion, where $|\mathcal{M}| = 7$ and $\mathcal{S}$ is the observed modality subset. NMR denotes the joint use of ^{1}H NMR, ^{13}C NMR, and HSQC spectra. Lines connecting the data points serve as a guide to the eye.

for four input configurations: IR+Raman+UV–Vis from QM9S, and joint NMR, NMR+IR, and NMR+IR+MS/MS from the Multimodal Spectroscopic Dataset. Here, joint NMR comprises three modalities: ^{1}H NMR, ^{13}C NMR, and HSQC.

Across all four input configurations, representation fusion consistently achieves higher ECFP4 similarity than latent fusion at comparable active query counts. This advantage is evident even with only three shared queries, where representation fusion outperforms latent fusion using three to five active queries, depending on the input configuration. Representation fusion also achieves higher similarity with substantially fewer queries: with 28 queries, it outperforms latent fusion at its largest evaluated setting ($n_e = 32$), corresponding to 96, 128, and 160 active queries for IR+Raman+UV–Vis, NMR+IR, and NMR+IR+MS/MS, respectively. For joint NMR, representation fusion with 56 queries likewise outperforms latent fusion with 96 queries.

These comparisons show that the performance gap between representation and latent fusion cannot be explained solely by differences in active query budget. Representation fusion may benefit from both using complementary spectral information during compression and sharing the query budget across modalities without a fixed per-modality allocation. In contrast, latent fusion compresses each modality independently using a fixed number of queries, potentially discarding cues that are weak in isolation but informative when combined with information from other modalities.

5 Related Work

The development of spectrum-based molecular structure elucidation can be viewed as a gradual relaxation of assumptions about the spectral input configuration. Early models were specialized to individual spectroscopic modalities; later systems combined predefined sets of complementary measurements; and recent approaches have begun to support different subsets of spectra within a single model.

Single-modality molecular structure elucidation. Early computational approaches often did not directly construct molecular structures, instead predicting intermediate molecular representations or ranking candidates from a predefined database. In mass spectrometry, methods such as MetFrag and CSI:FingerID use fragmentation evidence or predicted molecular fingerprints to score candidate structures (Wolf et al., 2010; Dührkop et al., 2015), while analogous retrieval-based approaches have also been explored for NMR spectra (Zhang et al., 2017; Yang et al., 2021). The emergence of molecular generative models has shifted the task from selecting the candidate structure that best matches a measurement to generating a molecular structure directly from that measurement. MassGenie and Spec2Mol generate molecular strings from MS/MS spectra without requiring a candidate structure database (Shrivastava et al., 2021; Litsa et al., 2023), whereas other approaches constrain generation using molecular formulas, predicted fingerprints, or related structured priors (Stravs et al., 2022; Bohde et al., 2025). Related developments have extended direct molecular structure

generation to NMR and vibrational spectral inputs, including raw one-dimensional NMR measurements (Hu et al., 2024) and IR spectra (Alberts et al., 2024a). More recent studies have explored diffusion- and flow-based formulations for spectrum-conditioned molecular generation (Yang et al., 2025; Nie & Gao, 2026). Together, these studies demonstrate the feasibility of *de novo* structure generation from individual spectroscopic modalities. However, by operating on a single modality at a time, they cannot exploit the complementary information available across different types of spectra, which can help resolve ambiguities that remain in any individual measurement.

Multimodal molecular structure elucidation. An early line of work in multimodal molecular structure elucidation focuses on predefined sets of complementary modalities, including IR together with ^{13}C NMR (Devata et al., 2024), combinations of IR, Raman, and UV–Vis spectra (Wang et al., 2025), and broader combinations of NMR, IR, Raman, and MS measurements (Wu et al., 2025; Priessner et al., 2026). These approaches demonstrate the benefit of combining complementary spectral information, but are generally designed for fixed input configurations or require separate training or adaptation for particular modality combinations. More recently, MM-Spectrum (Yu et al., 2026) uses sparse mixture-of-experts routing to process NMR, IR, and MS inputs, with robustness evaluated by removing individual modalities. Nevertheless, these approaches primarily target particular multimodal configurations rather than supporting arbitrary subsets of available spectral modalities within a single model.

Beyond these fixed settings, recent studies have begun to accommodate variable spectral inputs within a limited range of spectral modalities, including combinations of IR and Raman spectra (Lu et al., 2026) and multiple types of NMR experiments (Xue et al., 2025). SpectraLLM (Su et al., 2026) extends this flexibility to a broader range of spectral modalities within a single model, addressing the same flexible spectrum-to-structure setting considered in this work. It serializes heterogeneous spectral measurements into a common language-based representation and concatenates the resulting representations into a single input sequence. While this approach provides a unified interface across modalities, language-based serialization may not fully exploit the continuous nature and modality-specific structure of spectral signals. An alternative is to use dedicated encoders for each spectral modality. Doing so makes the extent of modality-specific processing an explicit architectural choice: *how much processing should be performed separately for each modality before joint processing begins?* Our work addresses this question through a comparison of fusion at multiple stages within a common generative framework.

6 Conclusion and limitations

We investigated *de novo* molecular structure prediction from arbitrary subsets of spectral modalities using a single model. To capture the distinct characteristics of heterogeneous spectra, we adopted modality-specific encoders, raising the question of how deep the separate modality streams should go before being merged for joint processing. We examined this question using *OmniSpecGen*, which compares four fusion stages within a common masked discrete diffusion backbone. Our results show that representation fusion, which integrates spectral information before compression through attention pooling, yields the strongest overall performance. Across the evaluated multimodal configurations, the resulting model also surpasses the reported results of SpectraLLM with a substantially smaller model. Together, these findings identify where multimodal information should be integrated when modality-specific encoders are used, and provide an effective framework for molecular structure elucidation from flexible spectral inputs.

Our study has several limitations that motivate further investigation. First, we study fusion timing within a masked discrete diffusion framework, and whether our findings generalize to other generative backbones, such as autoregressive models (Bagal et al., 2022), remains to be tested. Second, the modality-specific encoder architectures are kept fixed throughout our comparison, leaving architectural choices, such as encoding depth, open for further exploration. Third, our multimodal fusion analysis focuses on simulated spectra. Extending this analysis to experimental multimodal measurements would help clarify how acquisition conditions and measurement noise affect the relative benefits of different fusion stages. Finally, the current framework treats the available spectral subset as given. An interesting extension is to quantify uncertainty in molecular structure predictions given the observed spectra and estimate the expected uncertainty reduction from each additional modality, thereby guiding which type of spectrum to acquire next.

AI USE STATEMENT

We used generative AI tools to assist with translation during the preparation of this manuscript. All other tasks for which disclosure is required were either performed without generative AI assistance or were not applicable to this work. We also used these tools for literature review, language editing, and refinement of the manuscript's structure. We reviewed the entire manuscript to ensure that it accurately reflects our intended scientific claims, and we independently verified all references. We take full responsibility for the final content of this work, including all text and scientific claims.

ETHICS STATEMENT

This work aims to support chemical analysis through molecular structure prediction from spectral measurements. Our experiments use existing public datasets of molecular structures and spectra and do not involve human participants or personal data. Potential risks include incorrect molecular identification and misuse of structure-prediction methods in applications involving hazardous compounds. Generated structures should therefore be treated as candidate hypotheses and independently validated before use in safety-critical applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the source code in the supplementary materials and will make it publicly available on GitHub upon acceptance. Sections 2 and 3 describe the generative framework and the four fusion strategies, while Appendix A details the model architecture and modality-specific spectral encoders. Section 4 describes the dataset sources, molecular standardization, training, validation, and test splits, and evaluation metrics. Appendix B specifies the training objective, optimization settings, modality-subset sampling procedure, molecular sequence processing, and generation algorithm and its parameters, together with the implementation framework and computational resources used.

REFERENCES

Marvin Alberts, Teodoro Laino, and Alain C. Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. *Commun. Chem.*, 7:268, 2024a. URL https://doi.org/10.1038/s42004-024-01341-w.

Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024b. URL https://doi.org/10.52202/079017-3996.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *Proceedings of the Twenty-ninth ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS 2024)*, 2024. URL https://doi.org/10.1145/3620665.3640366.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*, 2021. URL https://openreview.net/forum?id=h7-XixPCAL.

Viraj Bagal, Rishal Aggarwal, P. K. Vinod, and U. Deva Priyakumar. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.*, 62(9):2064–2076, 2022. URL https://doi.org/10.1021/acs.jcim.1c00600.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.*, 7:20, 2015. URL https://doi.org/10.1186/s13321-015-0069-3.

Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra. In *Proceedings of the Forty-second International Conference on Machine Learning (ICML 2025)*, 2025. URL https://proceedings.mlr.press/v267/bohde25a.html.

Roman Bushuiev, Anton Bushuiev, Niek F. de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, Marcus Ludwig, Nils A. Haupt, Apurva Kalia, Corinna Brungs, Robin Schmid, Russell Greiner, Bo Wang, David S. Wishart, Li-Ping Liu, Juho Rousu, Wout Bittremieux, Hannes Rost, Tytus D. Mak, Soha Hassoun, Florian Huber, Justin J. J. van der Hooft, Michael A. Stravs, Sebastian Böcker, Josef Sivic, and Tomáš Pluskal. MassSpecGym: A benchmark for the discovery and identification of molecules. In *Advances in Neural Information Processing Systems (NeurIPS 2024)*, 2024. URL https://doi.org/10.52202/079017-3491.

Roman Bushuiev, Anton Bushuiev, Raman Samusevich, Corinna Brungs, Josef Sivic, and Tomáš Pluskal. Self-supervised learning of molecular representations from millions of tandem mass spectra using DreaMS. *Nat. Biotechnol.*, 44:630–640, 2026. URL https://doi.org/10.1038/s41587-025-02663-3.

Sriram Devata, Bhuvanesh Sridharan, Sarvesh Mehta, Yashaswi Pathak, Siddhartha Laghuvarapu, Girish Varma, and U. Deva Priyakumar. DeepSPInN – deep reinforcement learning for molecular structure prediction from infrared and ^{13}C NMR spectra. *Digit. Discov.*, 3(4):818–829, 2024. URL https://doi.org/10.1039/d4dd00008k.

Kai Dührkop, Huibin Shen, Marvin Meusel, Juho Rousu, and Sebastian Böcker. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U.S.A.*, 112(41):12580–12585, 2015. URL https://doi.org/10.1073/pnas.1509788112.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.*, 42(6):1273–1280, 2002. URL https://doi.org/10.1021/ci010132r.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *The Twenty-ninth IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 2016. URL https://doi.org/10.1109/CVPR.2016.90.

Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, Kenichi Tanaka, Satoshi Tanaka, Ken Aoshima, Yoshiya Oda, Yuji Kakazu, Miyako Kusano, Takayuki Tohge, Fumio Matsuda, Yuji Sawada, Masami Yokota Hirai, Hiroki Nakanishi, Kazutaka Ikeda, Naoshige Akimoto, Takashi Maoka, Hiroki Takahashi, Takeshi Ara, Nozomu Sakurai, Hideyuki Suzuki, Daisuke Shibata, Steffen Neumann, Takashi Iida, Ken Tanaka, Kimito Funatsu, Fumito Matsuura, Tomoyoshi Soga, Ryo Taguchi, Kazuki Saito, and Takaaki Nishioka. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.*, 45(7):703–714, 2010. URL https://doi.org/10.1002/jms.1777.

Frank Hu, Michael S. Chen, Grant M. Rotskoff, Matthew W. Kanan, and Thomas E. Markland. Accurate and Efficient Structure Elucidation from Routine One-Dimensional NMR Spectra Using Multitask Machine Learning. *ACS Cent. Sci.*, 10(11):2162–2170, 2024. URL https://doi.org/10.1021/acscentsci.4c01132.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. In *The Thirty-fourth Conference on Uncertainty in Artificial Intelligence (UAI 2018)*, 2018. URL http://auai.org/uai2018/proceedings/papers/313.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *The Third International Conference on Learning Representations (ICLR 2015)*, 2015. URL https://arxiv.org/abs/1412.6980.

Greg Landrum, Brian Kelley, Paolo Tosco, Sriniker, NadineSchneider, Riccardo Vianello, Adalke, AlexanderSavelyev, Samo Turk, Gedeck, JP, Strets123, JLVarjo, Axel Pahl, Patrick Fuller, DoliathGavid, Maciej Wójcikowski, Matt Swain, Gianluca Sforna, Michał Nowotka, Pzc, Jan Domański, David Hall, Baoilleach, Jan Holst Jensen, Brian Cole, Wolf-Guido Bolick, Nhfechner, DavidACosgrove, and Christos Kannas. RDKit: Open-Source Cheminformatics Software, 2016. URL https://www.rdkit.org/.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. In *Proceedings of the Thirty-sixth International Conference on Machine Learning (ICML 2019)*, 2019. URL https://proceedings.mlr.press/v97/lee19d/lee19d.pdf.

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Gopal Paliwal, Weili Nie, and Arash Vahdat. GenMol: A Drug Discovery Generalist with Discrete Diffusion. In *Proceedings of the Forty-second International Conference on Machine Learning (ICML 2025)*, 2025. URL https://openreview.net/forum?id=KM7pXWG1xj.

Eleni E. Litsa, Vijil Chenthamarakshan, Payel Das, and Lydia E. Kavraki. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun. Chem.*, 6:132, 2023. URL https://doi.org/10.1038/s42004-023-00932-3.

Xinyu Lu, Hao Ma, Hui Li, Jia Li, Yi Rong, Yuqiang Li, Tong Zhu, Guokun Liu, and Bin Ren. Vib2Mol: from vibrational spectra to molecular structures—a unified deep learning framework. *Chem. Sci.*, 2026. URL https://doi.org/10.1039/d6sc04559f.

Jianan Nie and Peng Gao. FlowMS: Flow Matching for De Novo Structure Elucidation from Mass Spectra. In *ICLR 2026 Workshop on Foundation Models for Science: Real-World Impact and Science-First Design*, 2026. URL https://openreview.net/forum?id=7H0fSu6gz1.

Emmanuel Noutahi, Cristian Gabellini, Michael Craig, Jonathan S. C. Lim, and Prudencio Tossou. Gotta be SAFE: a new framework for molecular design. *Digit. Discov.*, 3(4):796–804, 2024. URL https://doi.org/10.1039/d4dd00019f.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *The Nineteenth IEEE/CVF International Conference on Computer Vision (ICCV 2023)*, 2023. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Martin Priessner, Richard J. Lewis, Isak Lemurell, Magnus J. Johansson, Jonathan M. Goodman, Jon Paul Janet, and Anna Tomberg. Advancing Structure Elucidation with a Flexible Multi-Spectral AI Model. *Angew. Chem. Int. Ed.*, 65(2):e17611, 2026. URL https://doi.org/10.1002/anie.202517611.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5): 742–754, 2010. URL https://doi.org/10.1021/ci100050t.

Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B. Kell. MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra. *Biomolecules*, 11(12):1793, 2021. URL https://doi.org/10.3390/biom11121793.

Robert M. Silverstein, Francis X. Webster, David J. Kiemle, and David L. Bryce. *Spectrometric Identification of Organic Compounds*. John Wiley & Sons, 8th edition, 2014.

Peter St. John, Dejun Lin, Polina Binder, Malcolm Greaves, Vega Shah, John St. John, Adrian Lange, Patrick Hsu, Rajesh Illango, Arvind Ramanathan, Anima Anandkumar, David H. Brookes, Akosua Busia, Abhishaike Mahajan, Stephen Malina, Neha Prasad, Sam Sinai, Lindsay Edwards, Thomas Gaudelet, Cristian Regep, Martin Steinegger, Burkhard Rost, Alexander Brace, Kyle Hippe, Luca Naef, Keisuke Kamata, George Armstrong, Kevin Boyd, Zhonglin Cao, Han-Yi Chou, Simon Chu, Allan dos Santos Costa, Sajad Darabi, Eric Dawson, Kieran Didi, Cong Fu, Mario Geiger, Michelle Gill, Darren J. Hsu, Gagan Kaushik, Maria Korshunova, Steven Kothen-Hill, Youhan Lee, Meng Liu, Micha Livne, Zachary McClure, Jonathan Mitchell, Alireza Moradzadeh, Ohad Mosafi, Youssef Nashed, Saee Paliwal, Yuxing Peng, Sara Rabhi, Farhad Ramezanghorbani, Danny Reidenbach, Camir Ricketts, Brian C. Roland, Kushal Shah, Tyler Shimko, Hassan Sirelkhatim, Savitha Srinivasan, Abraham C. Stern, Dorota Toczydlowska, Srimukh Prasad Veccham, Niccolò Alberto Elia Venanzi, Anton Vorontsov, Jared Wilber, Isabel Wilkinson, Wei Jing Wong, Eva Xue, Cory Ye, Xin Yu, Yang Zhang, Guoqing Zhou, Becca Zandstein, Alejandro Chacon, Prashant Sohani, Maximilian Stadler, Christian Hundt, Feiwen Zhu, Christian Dallago, Bruno Trentini, Emine Kucukbenli, Saee Paliwal, Timur Rvachov, Eddie Calleja, Johnny Israeli, Harry Clifford, Risto Haukioja, Nicholas Haemel, Kyle Tretina, Neha Tadimeti, and Anthony B. Costa. BioNeMo Framework: a modular, high-performance library for AI model development in drug discovery. arXiv preprint arXiv:2411.10548, 2024. URL https://arxiv.org/abs/2411.10548.

Michael A. Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. MSNovelist: de novo structure generation from mass spectra. *Nat. Methods*, 19:865–870, 2022. URL https://doi.org/10.1038/s41592-022-01486-3.

Yunyue Su, Jiahui Chen, Zao Jiang, Zhenyi Zhong, Liang Wang, Qiang Liu, and Zhaoxiang Zhang. SpectraLLM: Uncovering the Ability of LLMs for Molecular Structure Elucidation from Multi-Spectral Data. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026. URL https://openreview.net/forum?id=J5XUzUW8o3.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Liang Wang, Yu Rong, Tingyang Xu, Zhenyi Zhong, Zhiyuan Liu, Pengju Wang, Deli Zhao, Qiang Liu, Shu Wu, Liang Wang, and Yang Zhang. DiffSpectra: Molecular Structure Elucidation from Spectra using Diffusion Models. arXiv preprint arXiv:2507.06853, 2025. URL https://arxiv.org/abs/2507.06853.

Sebastian Wolf, Stephan Schmidt, Matthias Müller-Hannemann, and Steffen Neumann. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinform.*, 11:148, 2010. URL https://doi.org/10.1186/1471-2105-11-148.

Sheng Wu, Shengjiao Ji, Mingzhi Yuan, Jun Jiang, Yi Luo, and Wei Hu. Multimodal Machine Learning for Molecular Identification Based on IR, Raman, and NMR Spectra. *J. Phys. Chem. Lett.*, 16 (44):11495–11501, 2025. URL https://doi.org/10.1021/acs.jpclett.5c02592.

Xi Xue, Hanyu Sun, Jingying Sun, Luc Patiny, Xiangying Liu, Kai Chen, Jingjie Yan, Liangning Li, Xue Liu, Shu Xu, Dongming Zhang, Yafeng Deng, Yingda Zang, Yaling Gong, Jie Ma, and Xiaojian Wang. NMRMind: A Transformer-Based Model Enabling the Elucidation from Multidimensional NMR to Structures. *Anal. Chem.*, 97(41):22603–22614, 2025. URL https://doi.org/10.1021/acs.analchem.5c03783.

Qingsong Yang, Binglan Wu, Xuwei Liu, Bo Chen, Wei Li, Gen Long, Xin Chen, and Mingjun Xiao. DiffNMR: Diffusion Models for Nuclear Magnetic Resonance Spectra Elucidation. arXiv preprint arXiv:2507.08854, 2025. URL https://arxiv.org/abs/2507.08854.

Zhuo Yang, Jianfei Song, Minjian Yang, Lin Yao, Jiahua Zhang, Hui Shi, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Cross-Modal Retrieval between ^{13}C NMR Spectra and Structures for Compound Identification Using Deep Contrastive Learning. *Anal. Chem.*, 93(50):16947–16955, 2021. URL https://doi.org/10.1021/acs.analchem.1c04307.

Hai-Tao Yu, Nan Min, Zheng Fang, Hongyu Zhan, Yusen Tan, Yuhan Wang, and Jun Xia. MM-Spectrum: Multimodal Multi-spectral Molecular Structural Elucidation with a Stable MoE Framework. In *Proceedings of the Forty-third International Conference on Machine Learning (ICML 2026)*, 2026. URL https://openreview.net/forum?id=YFva63G3zA.

Chen Zhang, Yerlan Idelbayev, Nicholas Roberts, Yiwen Tao, Yashwanth Nannapaneni, Brendan M. Duggan, Jie Min, Eugene C. Lin, Erik C. Gerwick, Garrison W. Cottrell, and William H. Gerwick. Small Molecule Accurate Recognition Technology (SMART) to Enhance Natural Products Research. *Sci. Rep.*, 7:14243, 2017. URL https://doi.org/10.1038/s41598-017-13923-x.

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nat. Comput. Sci.*, 3: 957–964, 2023. URL https://doi.org/10.1038/s43588-023-00550-y.

CONTENTS

1 Introduction **1**

2 Preliminaries **2**

2.1 Masked discrete diffusion 3

2.2 Transformer-based conditioning 3

3 Where to fuse: The OmniSpecGen framework **3**

4 Experiments **5**

4.1 Fusion strategies and benchmark performance 6

4.2 Fusion timing and pooling capacity 7

5 Related work **8**

6 Conclusion and limitations **9**

A Architecture details **16**

A.1 Overall model architecture 16

A.2 Modality-specific spectral encoders 17

B Training and generation details **20**

C Single-modality evaluation **21**

D Results with an alternative conditioning formulation **21**

E Comparison of conditioning strategies **24**

F Generalization to unseen modality combinations **25**

G Qualitative comparisons **25**

A ARCHITECTURE DETAILS

A.1 OVERALL MODEL ARCHITECTURE

All four fusion strategies are instantiated using the same basic architectural components: modality-specific spectral encoders g_m, attention pooling, a conditioning Transformer T_{cond}, and a molecular Transformer denoiser $f_{\boldsymbol{\theta}}$. All hidden representations have dimension $d = 256$, and the same architectural specifications are used across the fusion strategies. The strategies differ primarily in whether each component operates jointly on multiple modalities or independently on modality-specific representations, as illustrated in Fig. A1.

Each modality-specific encoder g_m transforms its input spectrum into a sequence of spectral representations $H^m \in \mathbb{R}^{L_m \times d}$. A learned modality embedding $\boldsymbol{e}^m \in \mathbb{R}^d$ is broadcast across token positions and added to the representations to encode modality identity. In representation fusion, this embedding is added to the token-level representations H^m before concatenation and joint attention pooling, thereby preserving explicit information about the source modality of each token. In the modality-wise pooling strategies, it is added to the pooled embeddings E^m.

By default, we set the per-modality embedding budget to $n_e = 4$. Representation fusion concatenates the token-level spectral representations and applies a single attention-pooling module with $N_e = n_e|\mathcal{M}|$ learned queries. These queries constitute one shared query set and are not assigned to individual modalities. The other strategies apply attention pooling independently to each modality using n_e learned queries. Thus, when all modalities are present, the two constructions produce the same total number of spectral embeddings.

The resulting embeddings are processed by T_{cond}, implemented as a two-layer Transformer encoder. In representation fusion, T_{cond} operates on the output of joint attention pooling. In latent fusion, it operates on the concatenation of the independently pooled embeddings. In molecular-interaction and late fusion, it is applied independently to each modality-specific embedding E^m. Molecular-interaction fusion then concatenates the resulting conditioning representations $\{C^m\}_{m\in\mathcal{S}}$ before molecular denoising, whereas late fusion retains the modality-specific pathways through the molecular denoiser and averages their output logits.

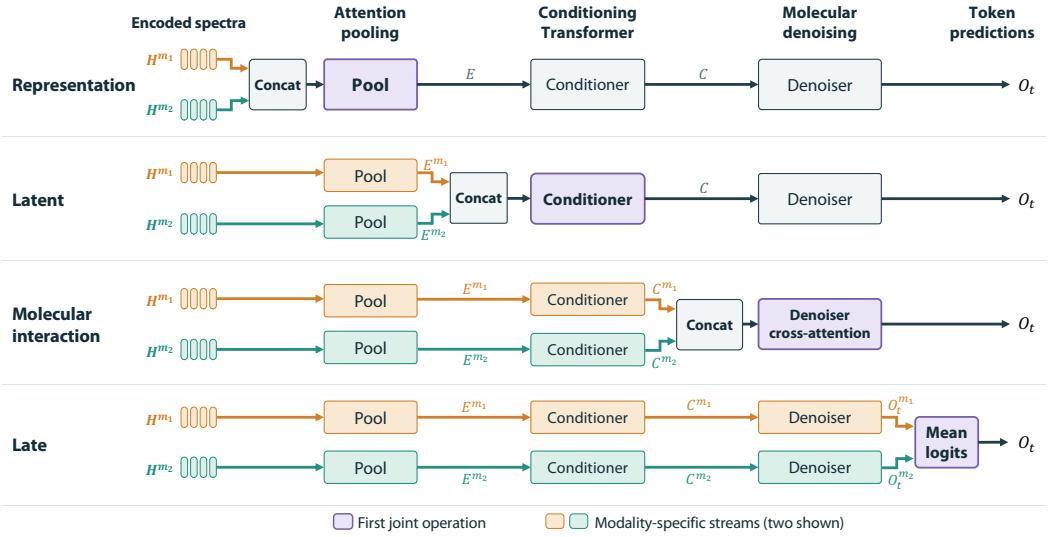

Figure A1: **The four fusion strategies in OmniSpecGen.** Highlighted blocks indicate the first joint operation across modalities: attention pooling in representation fusion, the conditioning Transformer in latent fusion, denoiser cross-attention in molecular-interaction fusion, and logit averaging in late fusion. Two modality streams are shown for clarity. Late fusion applies the same denoiser to the same corrupted sequence $\boldsymbol{y}_t$ under separate modality conditions and averages the output logits at each denoising step.

Table A1: **Architectural specifications shared across the four fusion strategies.** The strategies use the same component designs and hidden dimensions but differ in whether each component is applied jointly or modality-wise.

Inputs: Spectral representations $\{H^m\}_{m\in S}$ and corrupted SAFE sequence $\boldsymbol{y}_t$

Modules	Output	Details
		Spectral fusion and conditioning branch
Modality embedding	Unchanged	One learned 256-dimensional embedding per spectral modality, broadcast over the corresponding representations. *Representation fusion:* Added before joint pooling. *Other strategies:* Added after modality-wise pooling.
Attention pooling	$(N_e, 256)$ or $(n_e, 256) \times \lvert\mathcal{S}\rvert$	*Representation fusion:* Jointly pools the concatenated spectral representations using a single set of $N_e = n_e\lvert\mathcal{M}\rvert$ learned queries. *Other strategies:* Independently pool each available modality using $n_e = 4$ learned queries. Heads: 8; per-head dimension: 32; dropout: 0. Output LayerNorm; no residual connection or FFN sublayer.
Conditioning Transformer	Unchanged	A two-layer Transformer encoder. *Representation and latent fusion:* Applied after fusion. *Other strategies:* Applied modality-wise before fusion. Heads: 8; per-head dimension: 32. FFN: $256 \to 1024 \to 256$ (Linear–ReLU–Linear). Post LayerNorm with $\epsilon = 10^{-5}$; dropout: 0.
		Molecular denoising branch
Molecular input embedding	$(L, 256)$	Sum of learned SAFE token embeddings and learned absolute positional embeddings. Maximum sequence length: 200. LayerNorm; dropout: 0; no explicit timestep embedding.
Transformer denoiser	$(L, 256)$	Blocks: 8 in the order $\mathrm{S}\to\mathrm{C}\to\mathrm{S}\to\mathrm{C}\to\mathrm{S}\to\mathrm{C}\to\mathrm{S}\to\mathrm{S}$, where S and C denote self-attention and cross-attention, respectively. Heads: 8; per-head dimension: 32. FFN: $256 \to 1024 \to 256$ (Linear–ReLU–Linear). Post LayerNorm with $\epsilon = 10^{-5}$; dropout: 0.
Token prediction head	$(L, \lvert\mathcal{V}\rvert)$	$256 \to 256 \to \lvert\mathcal{V}\rvert$ (Linear–GELU–LayerNorm–Linear). The output projection shares its weight matrix with the input SAFE token embedding; a learned vocabulary bias is added.
Logit aggregation	$(L, \lvert\mathcal{V}\rvert)$	Used only in late fusion. The modality-wise logits are averaged over the available modalities: $O_t = \lvert\mathcal{S}\rvert^{-1}\sum_{m\in\mathcal{S}} O_t^m$.

Molecular structures are converted to SAFE strings with bracketed attachment-point notation and tokenized using the pretrained SAFE tokenizer released with SAFE (Noutahi et al., 2024), based on the `datamol-io/safe-gpt` vocabulary. We augment its vocabulary with the bracket delimiters and bracketed attachment-point tokens used by our representation. The corrupted SAFE sequence $\boldsymbol{y}_t$ is mapped to molecular hidden representations by summing learned token and absolute positional embeddings, followed by layer normalization. No explicit timestep embedding is used. The molecular denoiser consists of eight Transformer layers arranged in the order $\mathrm{S}\to\mathrm{C}\to\mathrm{S}\to\mathrm{C}\to\mathrm{S}\to\mathrm{C}\to\mathrm{S}\to\mathrm{S}$, where S and C denote self-attention and cross-attention layers, respectively. In each cross-attention layer, the molecular representations serve as queries, while the relevant conditioning representation serves as keys and values. Finally, the token prediction head maps the molecular hidden representations to logits over the SAFE vocabulary. Table A1 summarizes the architectural specifications and their application in the four fusion strategies.

A.2 Modality-specific spectral encoders

We use two types of modality-specific encoders according to the input representation: one for continuous spectra represented as intensity sequences and the other for variable-length peak sets. Both

produce the spectral representations H^m defined in the main text, which are passed to the subsequent attention-pooling stage.

Continuous spectra. Each continuous spectrum is provided to the encoder as an intensity vector obtained by dense, uniform sampling along the spectral axis. We process this vector using a one-dimensional residual CNN (He et al., 2016) to extract local intensity patterns while downsampling the sequence. We construct a uniformly spaced grid over $[-1, 1]$ with the same number of points as the downsampled sequence. Each grid coordinate is mapped to a d-dimensional position embedding using an MLP, and the resulting embeddings are added to the corresponding CNN features. A token-wise MLP transforms these position-aware features into H^m. Each output embedding therefore represents the intensity pattern within a local region of the input spectrum together with its location along the spectral axis.

Peak sets. For spectra represented as variable-length sets of peaks, each peak is described by its spectral coordinates and associated attributes, such as intensity. We first embed each peak into a vector using MLPs and then apply a Transformer encoder following the Set Transformer design (Lee et al., 2019). The encoder consists of stacked blocks comprising multi-head self-attention and a token-wise feed-forward network, with residual connections and layer normalization. Self-attention models relationships among peaks within the same spectrum, so each output embedding in H^m corresponds to one input peak while also incorporating information from other peaks in the spectrum.

Implementation details. We implement modality-specific encoders for seven spectral modalities: IR, Raman, UV–Vis, ^{13}C NMR, ^{1}H NMR, HSQC, and MS. IR, Raman, and UV–Vis use the continuous-spectrum encoder shown in Fig. A2. The NMR and MS modalities use the peak-set encoders shown in Figs. A3 and A4, respectively. All encoders produce spectral representations of dimension $d = 256$.

For *IR, Raman, and UV–Vis spectra*, each CNN consists of two residual blocks with kernel sizes 9 and 5, respectively, and 256 output channels in both blocks. Each block contains two convolutions: the first uses stride 2 to approximately halve the sequence length, and the second uses stride 1 with padding to preserve the reduced length. A 1×1 convolution with stride 2 on the shortcut matches the sequence length and channel dimension for residual addition. ReLU is applied after the first convolution and after the residual addition. We assign evenly spaced coordinates in $[-1, 1]$ to the downsampled sequence and embed them using a position MLP with layer widths $1 \to 256 \to 256$. We add these position embeddings to the CNN features and apply a token-wise output MLP with layer widths $256 \to 256 \to 256$ to obtain H^m. Both MLPs use ReLU between their linear layers. Batch normalization and dropout are disabled throughout these encoders.

For *NMR spectra*, each ^{13}C NMR peak is described by its chemical shift and intensity, each ^{1}H NMR peak by its chemical shift and hydrogen count, and each HSQC peak by its ^{13}C and ^{1}H chemical shifts and hydrogen count. Each scalar component is embedded using a separate MLP with layer widths $1 \to 256 \to 256$ and ReLU between the two linear layers. For each peak, the resulting component embeddings are summed to form a single peak token. These tokens are then processed by a four-layer Transformer encoder with a hidden dimension of 256, eight attention heads, a feed-forward hidden dimension of 1,024, and zero dropout to produce H^m.

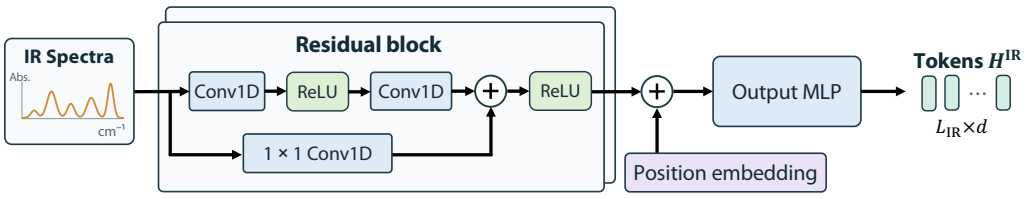

Figure A2: **Continuous-spectrum encoder.** Two residual convolutional blocks extract local spectral features and downsample the intensity sequence. Position embeddings are added to the resulting features, and a token-wise MLP produces the spectral representations H^m. IR is shown as an example; IR, Raman, and UV–Vis use the same architecture with separate parameters for each modality.

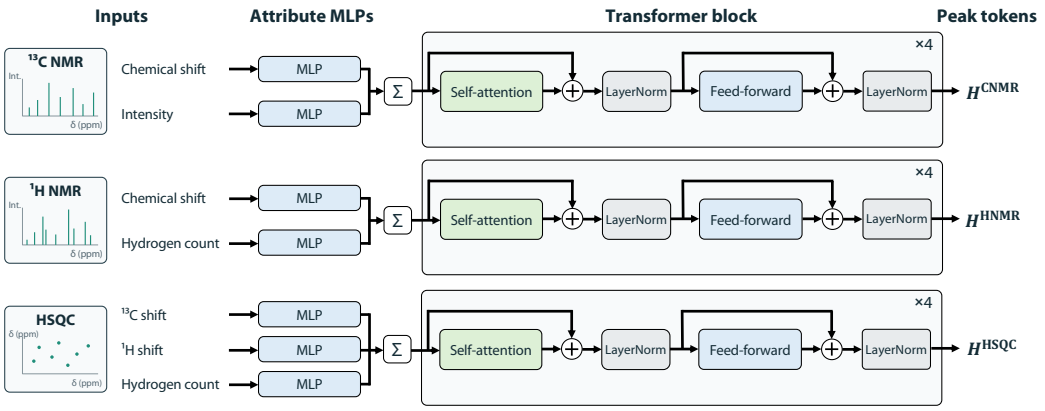

Figure A3: **NMR peak-set encoders.** MLPs embed the scalar attributes of each peak, and their outputs are summed to form a peak token. Four Transformer blocks apply self-attention and feed-forward transformations, each followed by residual addition and layer normalization, to produce H^m. The ^{13}C NMR, ^{1}H NMR, and HSQC encoders have separate parameters.

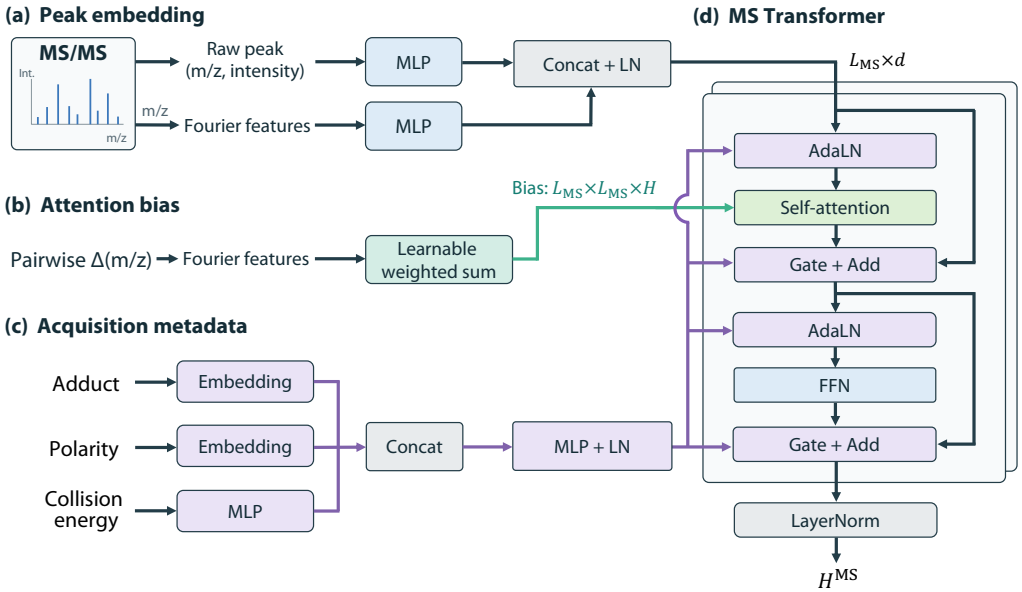

Figure A4: **MS peak-set encoder.** **(a)** Embeddings of Fourier-encoded m/z values and raw $(m/z, \text{intensity})$ pairs are concatenated and layer-normalized to form peak tokens. **(b)** Fourier features of pairwise m/z differences provide head-specific biases added to attention logits before softmax. **(c)** Adduct, polarity, and collision energy are embedded into a conditioning vector. **(d)** This vector modulates normalization and residual gates through AdaLN-Zero in two Transformer blocks. A final layer normalization produces H^{MS}.

For *MS spectra*, each peak is described by its m/z value and intensity. We construct peak tokens using a two-branch embedding scheme inspired by DreaMS (Bushuiev et al., 2026). The first branch converts the m/z value into sine and cosine Fourier features at multiple frequencies and processes them with an MLP. The second branch jointly processes the raw m/z value and intensity using a separate MLP. The two outputs are concatenated and layer-normalized to form a 256-dimensional peak token. These tokens are processed by a two-layer Transformer encoder with a hidden dimension of 256, eight attention heads, a feed-forward hidden dimension of 1,024, and zero dropout. To explicitly incorporate relationships between peak positions, each self-attention layer adds a learned

bias to the attention logits before softmax. In our implementation, this bias is computed separately for each attention head as a weighted sum of sine and cosine features of pairwise m/z differences.

To account for the dependence of mass spectra on measurement conditions, we explicitly provide the adduct, ionization polarity, and collision energy as additional inputs. The adduct and polarity are represented by learned embeddings, while the collision energy is embedded using an MLP with layer widths $1 \to 256 \to 256$ and ReLU between the linear layers. We represent missing collision energies with a learned embedding. These three embeddings are concatenated and processed by an MLP with layer widths $768 \to 256 \to 256$, SiLU between the linear layers, and a final layer normalization to obtain a condition vector. This vector conditions each MS Transformer block through AdaLN-Zero (Peebles & Xie, 2023). Specifically, it determines feature-wise scale and shift parameters applied to the normalized peak tokens before the self-attention and feed-forward sublayers, together with gates that scale the outputs of these sublayers before residual addition. The linear projections producing the scale, shift, and gate parameters are initialized to zero, so each block initially acts as an identity mapping. A final layer normalization produces the spectral representations H^{MS}.

B Training and generation details

Training details. We trained all models for 600 epochs with a batch size of 256 using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon_{\text{Adam}} = 10^{-9}$, and no weight decay (Kingma & Ba, 2015). The learning rate was linearly warmed up from zero to 3×10^{-4} over the first six epochs and subsequently decayed using a cosine schedule. We clipped the global ℓ_2 norm of the gradients to 1.0 before each optimizer update. During the final 100 epochs, we held the learning rate fixed and applied stochastic weight averaging (Izmailov et al., 2018), updating the parameter average after each optimizer step and using the averaged parameters for evaluation. To support arbitrary spectral subsets, we expose the model to diverse modality combinations during training. For each training example, we first choose the number of input modalities uniformly from one to the number available. We then uniformly select that many available modalities as conditioning inputs.

Each molecular structure is encoded as a SAFE token sequence consisting of a beginning-of-sequence token `[BOS]`, the tokens representing the molecule, and an end-of-sequence token `[EOS]`. Sequences are truncated to a maximum total length of 200 tokens, including `[BOS]` and `[EOS]`, and shorter sequences are padded with additional `[EOS]` tokens to match the longest sequence in each mini-batch. These `[EOS]` padding tokens are included in attention and contribute to the training loss when masked.

The model learns to reconstruct the clean SAFE sequence from its masked version, using the sampled spectral subset as conditioning. For each training example, we independently sample a diffusion time $t \sim \mathcal{U}(0, 1)$ and mask tokens with marginal probability $1 - \alpha(t)$, where $\alpha(t) = 1 - (1 - \epsilon_{\text{noise}})t$ and $\epsilon_{\text{noise}} = 10^{-3}$. We optimize a time-weighted cross-entropy loss over the masked positions:

$$\mathcal{L}(\boldsymbol{\theta}) = -\mathbb{E}\left[\frac{w(t)}{L}\sum_{i=1}^{L}\mathbf{1}[y_{t,i} = \texttt{[MASK]}]\log p_{\boldsymbol{\theta}}\left(y_{0,i} \mid \boldsymbol{y}_t, C\right)\right]. \tag{A1}$$

Here, $p_{\boldsymbol{\theta}}$ denotes the token-wise predictive distribution obtained from the denoiser logits, and L is the padded sequence length. The time-dependent weight is $w(t) = 1/t$. The expectation is taken over training examples, conditioning subsets, diffusion times, and masking patterns. In practice, losses are normalized by the padded sequence length and then averaged over the mini-batch. We implemented OmniSpecGen in PyTorch (Ansel et al., 2024) and used the MDLM implementation in the BioNeMo Framework (St. John et al., 2024) for diffusion training. Training took approximately 32 hours on a single NVIDIA GeForce RTX 5090 GPU with 32 GB of VRAM.

Generation details. Given the observed spectra, we generate a molecular structure by progressively unmasking a SAFE sequence using confidence-based sampling following GenMol by Lee et al. (2025). The initial sequence contains a fixed beginning-of-sequence token `[BOS]` followed by mask tokens. At each iteration k, the denoising network produces token logits conditioned on the current sequence and the observed spectra. We obtain the token distributions $p_{k,i}$ $(i = 1, \ldots, L)$

by applying the softmax function to these logits over the vocabulary excluding the mask token. For each masked position i, we sample a candidate token $\tilde{y}_{k,i} \sim p_{k,i}$ and assign it a confidence score based on its predicted probability and an additive Gumbel perturbation:

$$s_{k,i} = \log p_{k,i}(\tilde{y}_{k,i}) + r\left(1 - \frac{k}{K-1}\right) g_{k,i}, \qquad g_{k,i} \sim \mathrm{Gumbel}(0,1), \tag{A2}$$

where K is the total number of iterations, $k \in \{0, \ldots, K-1\}$, and r controls the initial perturbation strength. We fill the highest-scoring masked position with its sampled token and retain all previously filled positions. The remaining masked positions are predicted again in the next iteration, while the perturbation strength decreases linearly to zero.

For evaluation, we set the generation sequence length to $L = 100$ tokens, including the initial `[BOS]` token, and used $r = 0.5$. With one token unmasked per iteration, generation proceeds for $K = L - 1 = 99$ iterations. After the final iteration, we remove the initial `[BOS]` token and retain the sequence preceding the first end-of-sequence `[EOS]` token. The resulting token sequence is decoded into a SAFE string and then converted to SMILES for evaluation.

C Single-modality evaluation

Tables A2, A3, and A4 report single-modality results on QM9S, the Multimodal Spectroscopic Dataset, MassBank, and MassSpecGym. We used the same trained models as in the multimodal experiments reported in the main text, but provided only one spectral modality at inference. The representation-fusion variant outperforms SpectraLLM on all four structural similarity metrics for every modality in QM9S and for IR and all three NMR modalities in the Multimodal Spectroscopic Dataset. These results support the effectiveness of the proposed framework across a broad range of single-modality inputs.

The four variants do not exhibit a consistent performance ordering across single-modality inputs, in contrast to the results for multimodal inputs reported in Sec. 4.1. Representation fusion achieves the highest ECFP4 similarity among the variants for IR, Raman, and each NMR modality, whereas late fusion performs best on all four structural similarity metrics for UV–Vis. With only one input modality, no cross-modal fusion occurs at inference, so the ordering observed with multimodal inputs does not necessarily carry over to this setting. However, the variants differ in how modalities interact during multimodal training, which can influence the representations learned for each modality and contribute to their differences in single-modality performance.

On MS-only inputs, our models achieve broadly comparable performance to SpectraLLM, with relative strengths varying across datasets and metrics. All four variants improve functional-group similarity on MS/MS inputs from the Multimodal Spectroscopic Dataset, while molecular-interaction fusion outperforms SpectraLLM on three of the four similarity metrics on MassBank. On MassSpecGym, representation fusion achieves similar functional-group similarity, although SpectraLLM obtains higher ECFP4, MACCS, and Fraggle similarities. The lack of consistent improvements on MS-only inputs may partly reflect the composition of the training corpus: MassBank and MassSpecGym provide only MS measurements, leaving many MS training samples without complementary spectral modalities and thus limiting opportunities to benefit directly from cross-modal fusion.

D Results with an alternative conditioning formulation

To assess whether our findings on the stage of multimodal fusion depend on the conditioning formulation, we repeated the comparison of the four fusion strategies using in-context concatenation instead of cross-attention. In this alternative conditioning strategy, the spectral conditioning representations are concatenated with the molecular hidden representations H_{mol} and jointly processed by self-attention. We replace the three cross-attention layers in the original denoiser with self-attention layers, so that all eight Transformer layers jointly process the concatenated sequence. Like cross-attention, in-context concatenation accepts conditioning embeddings as a sequence of tokens, allowing us to compare the four fusion stages by varying how these embeddings are constructed and supplied to the denoiser. Although adaptive layer normalization (adaLN) is another well-established

Table A2: **Comparison of the four fusion strategies and SpectraLLM on the QM9S dataset with single-modality inputs.** † denotes values reported in Table 19 of SpectraLLM (Su et al., 2026) for models separately trained on each input modality, rather than results obtained from our reimplementation. **Bold** and underline indicate the best and second-best results, respectively, for each metric within each input configuration.

Inputs	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
IR	SpectraLLM†	**100.00%**	0.1474	0.5309	0.3436	0.2686
	Representation	98.66%	**0.5610**	**0.9331**	**0.8170**	**0.7852**
	Latent	98.90%	0.4813	0.9229	0.7731	0.7710
	Mol. interaction	98.74%	0.4683	0.9167	0.7601	0.7654
	Late	99.07%	0.4836	0.9243	0.7755	0.7752
Raman	SpectraLLM†	98.90%	0.3251	0.7986	0.5748	0.3958
	Representation	98.90%	**0.7169**	**0.9499**	**0.8862**	**0.8547**
	Latent	98.91%	0.6418	0.9293	0.8505	0.8357
	Mol. interaction	98.41%	0.6351	0.9212	0.8412	0.8269
	Late	**99.01%**	0.5557	0.9245	0.8128	0.8029
UV–Vis	SpectraLLM†	**100.00%**	0.0777	0.3600	0.2081	0.2137
	Representation	99.15%	0.1358	0.5401	0.3268	0.4497
	Latent	99.13%	0.1353	0.5371	0.3299	0.5049
	Mol. interaction	99.21%	0.1349	0.5358	0.3270	0.4955
	Late	99.49%	**0.1445**	**0.5608**	**0.3493**	**0.5235**

Table A3: **Comparison of the four fusion strategies and SpectraLLM on the Multimodal Spectroscopic Dataset with single-modality inputs.** † denotes values reported in Table 4 of SpectraLLM (Su et al., 2026).

Inputs	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
IR	SpectraLLM†	**99.63%**	0.1720	0.6023	0.4031	0.3906
	Representation	99.36%	**0.2510**	**0.6712**	**0.5285**	0.4727
	Latent	99.45%	0.2394	0.6653	0.5143	**0.5225**
	Mol. interaction	99.37%	0.2315	0.6569	0.5048	0.5129
	Late	99.57%	0.2380	0.6595	0.5128	0.5222
MS/MS	SpectraLLM†	99.64%	**0.1844**	0.4929	0.4254	0.4282
	Representation	99.69%	0.1687	0.6038	0.4301	0.3978
	Latent	99.64%	0.1660	0.6027	0.4245	0.4410
	Mol. interaction	99.72%	0.1683	0.6090	0.4309	0.4407
	Late	**99.80%**	0.1744	**0.6152**	**0.4371**	**0.4525**
^{13}C NMR	SpectraLLM†	99.64%	0.1016	0.4249	0.2952	0.3607
	Representation	98.83%	**0.3783**	**0.7488**	**0.6423**	0.5743
	Latent	98.76%	0.3541	0.7395	0.6302	**0.6034**
	Mol. interaction	98.68%	0.3390	0.7266	0.6123	0.5855
	Late	**99.90%**	0.0821	0.3846	0.1578	0.2365
^{1}H NMR	SpectraLLM†	99.10%	0.0720	0.3329	0.2203	0.2572
	Representation	99.53%	**0.2363**	**0.5796**	0.4770	0.4469
	Latent	99.57%	0.2331	0.5777	**0.4774**	**0.5040**
	Mol. interaction	99.42%	0.2263	0.5703	0.4687	0.4959
	Late	**99.69%**	0.2220	0.5707	0.4634	0.4949
HSQC NMR	SpectraLLM†	**99.64%**	0.2058	0.5495	0.4392	0.4274
	Representation	99.21%	**0.3348**	**0.6751**	**0.5829**	0.5326
	Latent	99.20%	0.3218	0.6686	0.5734	**0.5759**
	Mol. interaction	99.08%	0.3095	0.6579	0.5604	0.5636
	Late	99.49%	0.2640	0.6319	0.5160	0.5326

Table A4: **Comparison on the MassBank and MassSpecGym datasets.** † denotes values reported in Table 3 of SpectraLLM (Su et al., 2026).

Dataset	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
MassBank	SpectraLLM†	98.44%	0.1286	0.4539	**0.3787**	0.3150
	Representation	98.08%	0.1337	0.4540	0.3304	0.2861
	Latent	98.48%	0.1368	0.4715	0.3347	0.3440
	Mol. interaction	97.93%	**0.1408**	**0.4724**	0.3390	**0.3510**
	Late	**99.30%**	0.1249	0.4385	0.3008	0.3180
MassSpecGym	SpectraLLM†	**99.74%**	**0.1533**	0.5003	**0.4723**	0.3610
	Representation	99.08%	0.1303	0.4984	0.4139	0.3027
	Latent	99.00%	0.1278	**0.5048**	0.4133	0.3570
	Mol. interaction	99.15%	0.1291	0.4980	0.4157	**0.3631**
	Late	99.34%	0.1183	0.4736	0.3766	0.3376

conditioning strategy studied in diffusion Transformers (Peebles & Xie, 2023), its vector-based formulation requires aggregating the spectral tokens into a single conditioning vector before denoising, introducing an additional aggregation step that complicates the comparison of fusion stages. We therefore compare adaLN with the other conditioning strategies separately in Appendix E. In that comparison, cross-attention achieves better overall performance than adaLN, supporting its use as the primary conditioning strategy for our comparison of fusion stages.

Tables A5 and A6 show that the advantage of representation fusion largely persists under in-context concatenation. Representation fusion outperforms latent fusion across all reported multimodal input configurations on the QM9S dataset and achieves the highest ECFP4, functional-group, and MACCS similarities among the four fusion variants on the Multimodal Spectroscopic Dataset. Under in-context concatenation, however, molecular-interaction fusion outperforms latent fusion in some input configurations, while latent fusion retains a substantial advantage in others. This may reflect differences in how the two conditioning strategies enable cross-modal interaction within the denoiser. In molecular-interaction fusion, cross-attention uses modality-specific conditioning representations as keys and values, integrating their information through updates to the molecular tokens. In-context concatenation additionally allows conditioning tokens from different modalities to interact directly through self-attention. Although fusion still begins within the denoiser, this direct inter-

Table A5: **Comparison of the four fusion strategies on QM9S with multimodal inputs, using in-context concatenation conditioning. Bold** and underline indicate the best and second-best results, respectively, for each metric within each input configuration.

Inputs	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
IR + Raman	Representation	99.26%	**0.7517**	**0.9649**	**0.9122**	**0.8752**
	Latent	99.23%	0.7140	0.9601	0.8986	0.8609
	Mol. interaction	98.99%	0.6688	0.9404	0.8655	0.8480
	Late	**99.59%**	0.4764	0.9246	0.7798	0.7831
Raman + UV–Vis	Representation	99.05%	**0.7248**	**0.9566**	**0.8967**	**0.8647**
	Latent	99.10%	0.6730	0.9457	0.8725	0.8414
	Mol. interaction	98.80%	0.6861	0.9458	0.8752	0.8537
	Late	**99.70%**	0.4215	0.8963	0.7261	0.7554
UV–Vis + IR	Representation	99.38%	**0.5674**	**0.9474**	**0.8274**	**0.7989**
	Latent	99.31%	0.4977	0.9390	0.7895	0.7693
	Mol. interaction	99.35%	0.1793	0.7111	0.4334	0.5761
	Late	**99.63%**	0.3686	0.8936	0.6900	0.7299
IR + Raman + UV–Vis	Representation	99.30%	**0.7586**	**0.9675**	**0.9177**	**0.8808**
	Latent	99.23%	0.7283	0.9642	0.9048	0.8654
	Mol. interaction	98.97%	0.6828	0.9472	0.8746	0.8535
	Late	**99.79%**	0.4543	0.9199	0.7600	0.7723

Table A6: **Comparison of the four fusion strategies on the Multimodal Spectroscopic Dataset with multimodal inputs, using in-context concatenation conditioning.** NMR denotes the joint use of ^{13}C NMR, ^{1}H NMR, and HSQC spectra.

Inputs	Method	Validity ↑	Tanimoto ECFP4 ↑	Functional Group ↑	Tanimoto MACCS ↑	Fraggle ↑
IR + MS/MS	Representation	99.32%	**0.2805**	**0.7440**	**0.6079**	**0.5165**
	Latent	99.34%	0.2742	0.7394	0.6009	0.5110
	Mol. interaction	99.72%	0.1721	0.6175	0.4482	0.4454
	Late	**99.91%**	0.1922	0.6244	0.4489	0.4887
NMR	Representation	98.80%	**0.4257**	**0.7791**	**0.6900**	0.6142
	Latent	98.89%	0.4094	0.7751	0.6833	0.6062
	Mol. interaction	98.76%	0.4124	0.7760	0.6842	**0.6457**
	Late	**99.88%**	0.0880	0.3966	0.1744	0.2536
NMR + IR	Representation	98.74%	**0.4966**	**0.8471**	**0.7655**	**0.6671**
	Latent	98.82%	0.4681	0.8391	0.7477	0.6482
	Mol. interaction	98.76%	0.4217	0.7929	0.7004	0.6510
	Late	**99.93%**	0.1319	0.4925	0.2530	0.3528
NMR + MS/MS	Representation	98.74%	**0.4613**	**0.8295**	**0.7400**	0.6439
	Latent	98.62%	0.4395	0.8216	0.7292	0.6293
	Mol. interaction	98.58%	0.4395	0.8194	0.7279	**0.6612**
	Late	**99.90%**	0.0968	0.4331	0.2003	0.3059
NMR + IR + MS/MS	Representation	98.58%	**0.5157**	**0.8693**	**0.7853**	**0.6802**
	Latent	98.64%	0.4915	0.8651	0.7737	0.6651
	Mol. interaction	98.61%	0.4477	0.8310	0.7388	0.6658
	Late	**99.93%**	0.1294	0.4983	0.2634	0.3704

action provides an additional opportunity to combine complementary spectral cues during molecular generation.

In summary, representation fusion achieves the strongest overall structural similarity performance under in-context concatenation, while late fusion generally performs worst. The relative ordering of latent and molecular-interaction fusion changes for some input configurations compared with cross-attention, indicating that the conditioning mechanism can affect the relative performance of these intermediate fusion stages. Nevertheless, the advantage of integrating modality-specific spectral information before pooling persists under both cross-attention and in-context concatenation.

E Comparison of conditioning strategies

Following the conditioning design space explored in DiT (Peebles & Xie, 2023), we consider three representative conditioning mechanisms for diffusion Transformers: cross-attention (used in the main experiments), in-context conditioning (used in Appendix D), and adaptive layer normalization (adaLN). These mechanisms incorporate conditional information through cross-attention over conditioning representations, joint self-attention over molecular and conditioning tokens, and feature-wise modulation of molecular hidden representations, respectively. To isolate the effect of the conditioning mechanism, we fix the multimodal fusion strategy to representation fusion.

Both cross-attention and in-context conditioning operate directly on the fused conditioning representation as a sequence of tokens. By contrast, for adaLN, we concatenate all conditioning embeddings into a single vector and linearly project it to the model's hidden dimension. In each Transformer block ℓ, the projected conditioning vector $\boldsymbol{c}$ is mapped to separate feature-wise scale, shift, and residual gating parameters for the self-attention and MLP branches. For each branch $b \in \{\mathrm{attn}, \mathrm{mlp}\}$, the normalized molecular hidden representation is modulated as

$$\mathrm{adaLN}_{\ell,b}(\boldsymbol{h}_i, \boldsymbol{c}) = (1 + \gamma_{\ell,b}(\boldsymbol{c})) \odot \mathrm{LN}(\boldsymbol{h}_i) + \beta_{\ell,b}(\boldsymbol{c}), \tag{A3}$$

where $\gamma_{\ell,b}(\boldsymbol{c})$ and $\beta_{\ell,b}(\boldsymbol{c})$ denote the feature-wise scale and shift parameters, respectively. The output of each branch is additionally scaled by a feature-wise gate before being added to the residual stream. The same modulation and gating parameters are applied across all molecular token positions, and $\odot$ denotes element-wise multiplication.

Table A7 compares the conditioning strategies for multimodal inputs using two metrics: Tanimoto (ECFP4) similarity and Fraggle similarity. Cross-attention performs best overall, achieving the highest ECFP4 similarity for every input configuration and the highest Fraggle score for seven of the nine configurations. In-context concatenation performs similarly to cross-attention overall and achieves slightly higher Fraggle similarity for IR+Raman and IR+Raman+UV–Vis on QM9S. In contrast, adaLN yields the lowest ECFP4 and Fraggle scores for every configuration. This consistent gap suggests that compressing the concatenated conditioning embeddings into a single vector before denoising may create an information bottleneck for fine-grained spectral cues. These results support our choice of cross-attention as the default conditioning mechanism in the main architecture, while showing that in-context concatenation remains a competitive alternative.

Table A7: **Comparison of three conditioning strategies using representation fusion on multimodal input configurations from QM9S and the Multimodal Spectroscopic Dataset.** NMR denotes the joint use of ^{13}C NMR, ^{1}H NMR, and HSQC spectra. **Bold** and underline indicate the best and second-best results, respectively, for each metric within each input configuration.

Dataset	Inputs	Tanimoto ECFP4 ↑			Fraggle ↑		
		Cross-attention	In-context	adaLN	Cross-attention	In-context	adaLN
QM9S	IR + Raman	**0.7535**	0.7517	0.7188	0.8740	**0.8752**	0.8529
	Raman + UV–Vis	**0.7272**	0.7248	0.6907	**0.8658**	0.8647	0.8405
	UV–Vis + IR	**0.5844**	0.5674	0.5452	**0.8015**	0.7989	0.7793
	IR + Raman + UV–Vis	**0.7624**	0.7586	0.7270	0.8782	**0.8808**	0.8562
Multimodal Spectroscopic	IR + MS/MS	**0.2881**	0.2805	0.2675	**0.5189**	0.5165	0.4972
	NMR	**0.4445**	0.4257	0.3947	**0.6258**	0.6142	0.5885
	NMR + IR	**0.5160**	0.4966	0.4473	**0.6763**	0.6671	0.6244
	NMR + MS/MS	**0.4791**	0.4613	0.4210	**0.6521**	0.6439	0.6093
	NMR + IR + MS/MS	**0.5351**	0.5157	0.4651	**0.6881**	0.6802	0.6386

F GENERALIZATION TO UNSEEN MODALITY COMBINATIONS

We investigated whether representation fusion can integrate information from modality combinations not encountered during training. To construct such unseen combinations, we restricted each training example to at most k modalities, with $k \in \{2, 3, 4\}$, while retaining the same set of modality types. Combinations containing more than k modalities were therefore excluded from training. We evaluated the resulting models on the Multimodal Spectroscopic Dataset using three NMR modalities jointly, NMR combined with either IR or MS/MS, and all five modalities. Table A8 reports Tanimoto (ECFP4) similarity alongside the unrestricted-training reference; † identifies combinations held out by the training restriction.

The models can benefit from combining modalities even when the resulting combination was excluded from training. With $k = 3$, adding IR to NMR produces an unseen four-modality combination and increases similarity from 0.4177 to 0.4639; using all five modalities further improves similarity to 0.4756. With $k = 4$, adding MS/MS to NMR and IR likewise forms an unseen combination and improves similarity from 0.4929 to 0.5073. These results indicate that the learned fusion mechanism can integrate additional spectral information without having been trained on the exact combination used at inference.

Nevertheless, all restricted models fall below the unrestricted reference across the reported input conditions. This gap is not confined to held-out combinations: for $k = 4$, NMR and IR form a combination permitted during training, yet similarity remains lower than that of the unrestricted reference. One possible explanation is that joint training on richer modality combinations helps the model learn correspondences among complementary spectral cues. These correspondences may also support predictions from smaller subsets, so excluding richer combinations could reduce performance even on combinations permitted during training. Representation fusion therefore exhibits partial generalization to the held-out combinations, while matching the performance of unrestricted training remains challenging.

Table A8: **Generalization of representation fusion to unseen modality combinations on the Multimodal Spectroscopic Dataset, measured by Tanimoto (ECFP4) similarity.** Training inputs are restricted to the indicated maximum number of simultaneous modalities. † marks combinations excluded from training by this restriction. "No limit" denotes unrestricted training. **Bold** indicates the best results for each input configuration.

Max. training modalities	NMR (3 modalities)	NMR + IR (4 modalities)	NMR + MS/MS (4 modalities)	All modalities (5 modalities)
2	$0.3089^{\dagger}$	$0.3600^{\dagger}$	$0.3311^{\dagger}$	$0.3768^{\dagger}$
3	0.4177	$0.4639^{\dagger}$	$0.4394^{\dagger}$	$0.4756^{\dagger}$
4	0.4309	0.4929	0.4594	$0.5073^{\dagger}$
No limit	**0.4445**	**0.5160**	**0.4791**	**0.5351**

G QUALITATIVE COMPARISONS

We complement the quantitative evaluation with molecular structure visualizations. We randomly sampled six molecules from each of the QM9S and Multimodal Spectroscopic Dataset test sets and visualize the predictions generated by the four OmniSpecGen fusion variants and SpectraLLM (Figs. A5–A14). For SpectraLLM, we generated predictions using its publicly available checkpoint. The same reference molecules and spectral input configurations were used across methods. Within each dataset, rows correspond to molecules 1–6 from top to bottom, with the reference structures in the leftmost column and predictions under seven input configurations in the remaining columns. These configurations cover all nonempty combinations of IR, Raman, and UV–Vis for QM9S, and of IR, NMR, and MS/MS for the Multimodal Spectroscopic Dataset. Here, NMR denotes the joint use of three modalities: ^{1}H NMR, ^{13}C NMR, and HSQC. "All spectra" denotes the complete set of spectral inputs for the corresponding dataset. Each prediction cell displays one generated structure, its molecular formula, and its Tanimoto (ECFP4) similarity to the reference. Green outlines and check marks indicate exact matches to the reference structure.

Representation fusion can recover reference structures by combining spectral inputs that yield incorrect predictions when used separately. On the QM9S dataset, this occurs for molecule 1 with IR+UV–Vis, molecule 3 with IR+Raman, and molecule 6 with either IR+UV–Vis or Raman+UV–Vis: in each case, neither constituent modality produces the correct structure on its own (see Fig. A5). Similar examples appear in the Multimodal Spectroscopic Dataset, where molecules 1 and 2 are correctly reconstructed only when all spectra are provided, while molecule 6 is recovered from IR+NMR despite incorrect predictions from both IR and the NMR suite alone (see Fig. A10). These examples illustrate how representation fusion can exploit complementary spectral information to recover structures that remain incorrectly reconstructed under reduced input configurations.

The fusion strategies differ in how successfully they use multimodal inputs. For the QM9S examples shown in Figs. A5–A8, all four strategies correctly reconstruct multiple reference molecules under multimodal input configurations. On the Multimodal Spectroscopic Dataset shown in Figs. A10–A13, none of the strategies correctly reconstructs any of the six displayed molecules from IR or MS/MS alone, whereas multimodal inputs yield successful reconstructions for all strategies except late fusion. Note that the NMR suite itself constitutes a multimodal input comprising three modalities. In particular, representation fusion correctly reconstructs all six molecules under at least one of the two configurations, IR+NMR and all spectra, compared with three molecules for latent fusion, one for molecular-interaction fusion, and none for late fusion. Thus, in these Multimodal Spectroscopic Dataset examples, which involve more modalities and more complex molecular structures, earlier fusion enables accurate reconstruction across a broader set of molecules.

Although structural similarity to the ground truth generally improves with more spectral modalities (Tables 2 and 3), adding modalities can still cause errors in molecules that were correctly predicted from fewer modalities, even with representation fusion. On the QM9S dataset, molecule 1 is correctly reconstructed from Raman but not from Raman+UV–Vis, while molecule 5 is recovered from Raman or Raman+UV–Vis but not from all spectra (see Fig. A5). On the Multimodal Spectroscopic Dataset, adding MS/MS to IR+NMR changes correct predictions into incorrect ones for molecules 5 and 6; adding MS/MS to the NMR suite similarly affects molecules 3 and 5 (see Fig. A10). The overall molecular frameworks are largely preserved, but details such as chain length, substituent positions, and local connectivity become inaccurate. These changes may reflect interference during multimodal integration from inputs that yield inaccurate predictions on their own, affecting structural details correctly recovered from the other spectra.

Despite these limitations, the comparison with SpectraLLM also reveals a substantial difference in absolute reconstruction accuracy. As shown in Figs. A9 and A14, SpectraLLM produces only two exact reconstructions among the 42 displayed predictions on the QM9S dataset and none among the 42 predictions on the Multimodal Spectroscopic Dataset. In these displayed examples, our model achieves higher reconstruction accuracy with approximately 30M parameters, whereas SpectraLLM uses a 32B-parameter backbone.

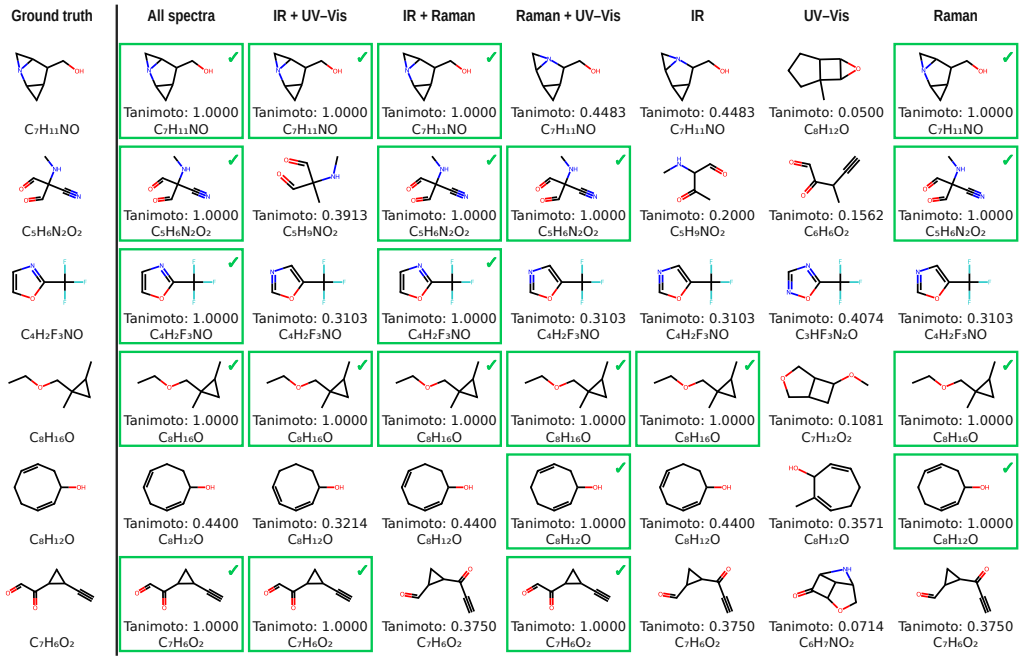

Figure A5: **Molecular structures generated using representation fusion on QM9S.** Six reference molecules (left column) and OmniSpecGen predictions under all seven nonempty subsets of IR, Raman, and UV–Vis. "All spectra" denotes IR+Raman+UV–Vis. Below each predicted structure, we report its Tanimoto (ECFP4) similarity to the corresponding reference and its molecular formula. Green outlines and check marks indicate exact matches to the reference structures. Rows are numbered 1–6 from top to bottom in the accompanying discussion. The same targets and column order are used in the following four figures.

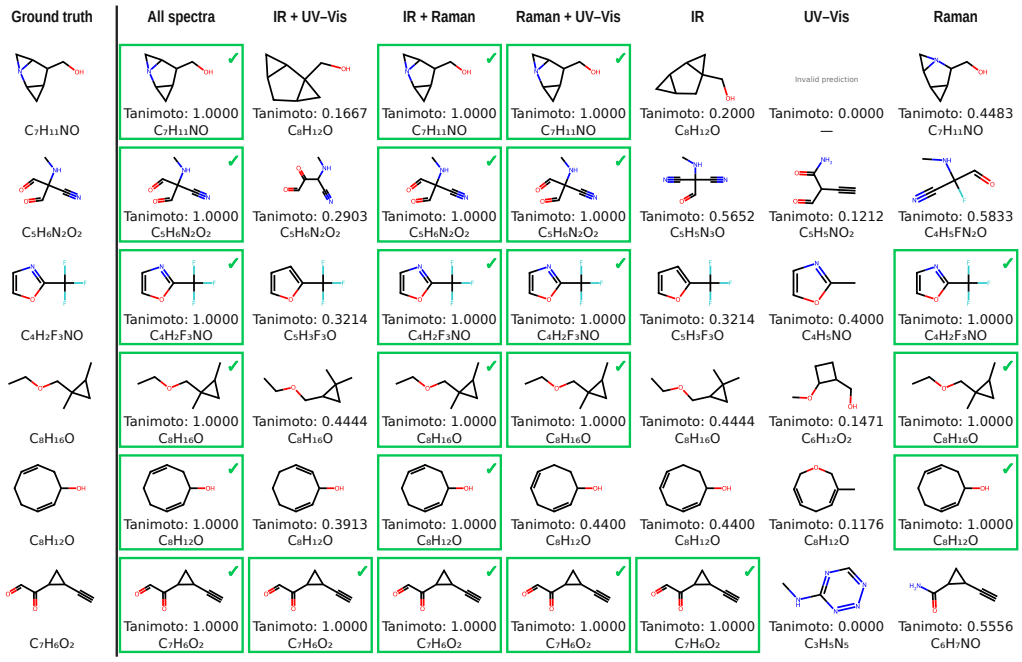

Figure A6: **Molecular structures generated using latent fusion on QM9S.** OmniSpecGen predictions for the same six reference molecules and seven spectral input subsets as in Figure A5. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures. The UV–Vis prediction for molecule 1 is marked as invalid and is retained in the visualization.

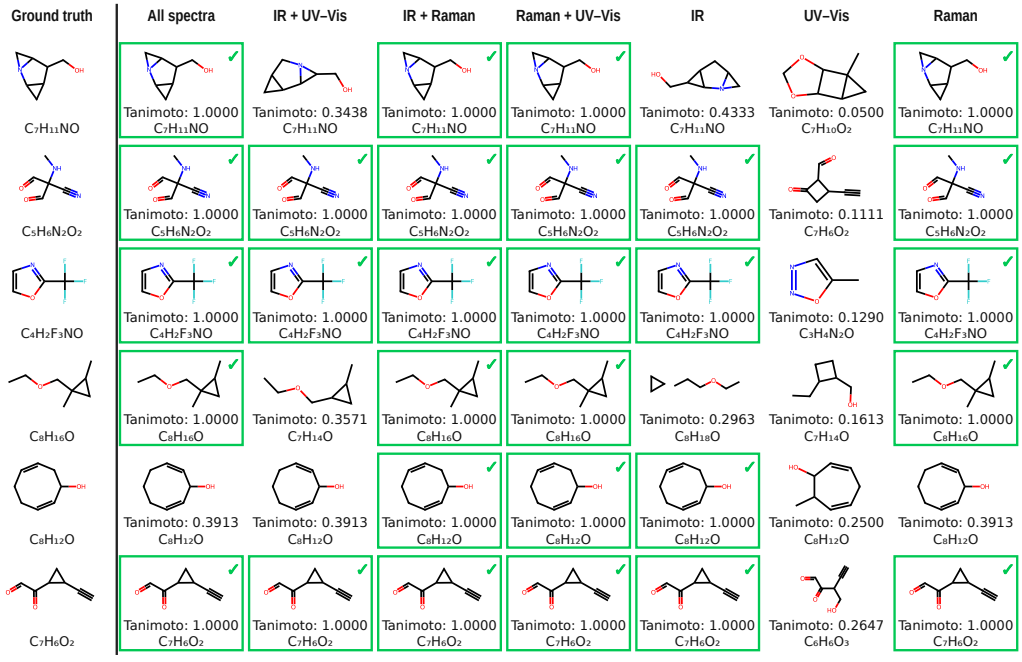

Figure A7: **Molecular structures generated using molecular-interaction fusion on QM9S.** OmniSpecGen predictions for the same six reference molecules and seven spectral input subsets as in Figure A5. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures.

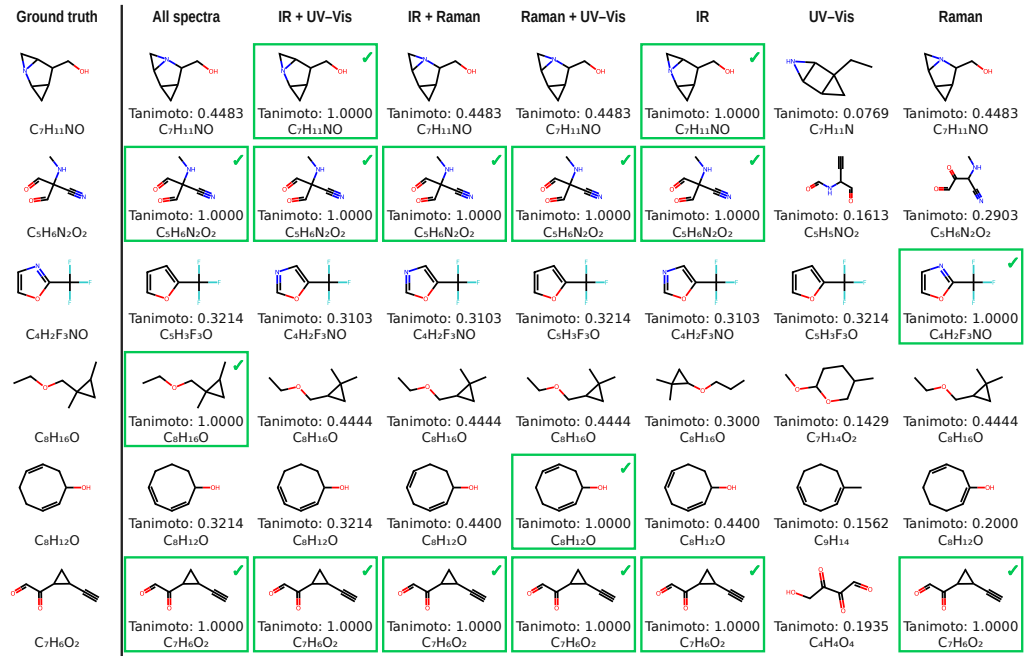

Figure A8: **Molecular structures generated using late fusion on QM9S.** OmniSpecGen predictions for the same six reference molecules and seven spectral input subsets as in Figure A5. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures.

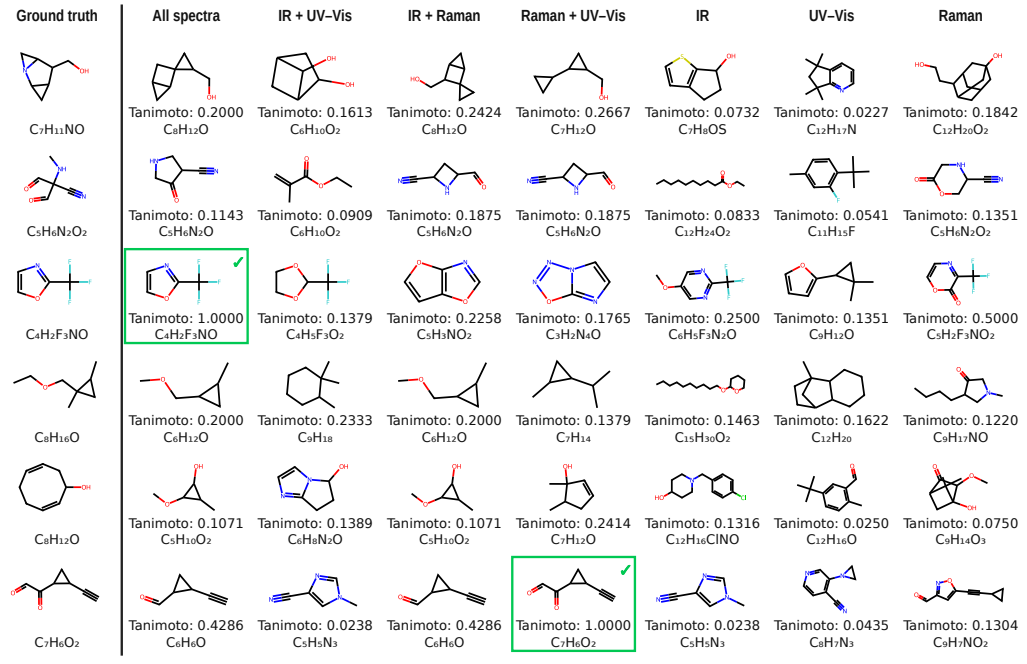

Figure A9: **Molecular structures generated by SpectraLLM on QM9S.** Predictions for the same six reference molecules and seven spectral input subsets as in Figure A5. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures.

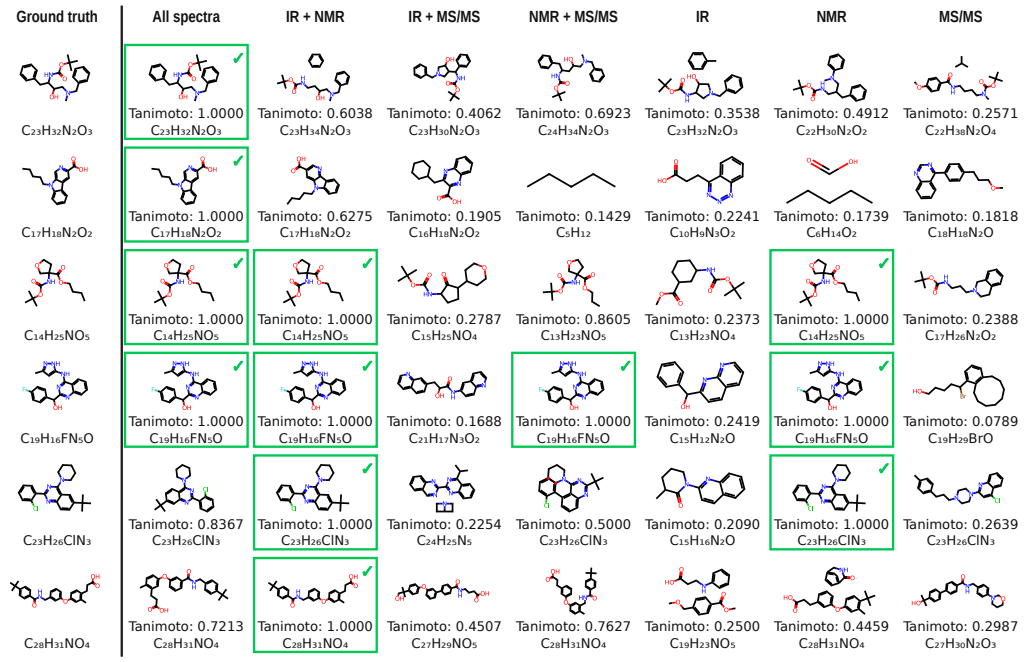

Figure A10: **Molecular structures generated using representation fusion on the Multimodal Spectroscopic Dataset.** Predictions for six randomly selected test molecules under seven spectral input configurations. NMR denotes the joint use of ^{1}H NMR, ^{13}C NMR, and HSQC; "All spectra" denotes IR+NMR+MS/MS. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures.

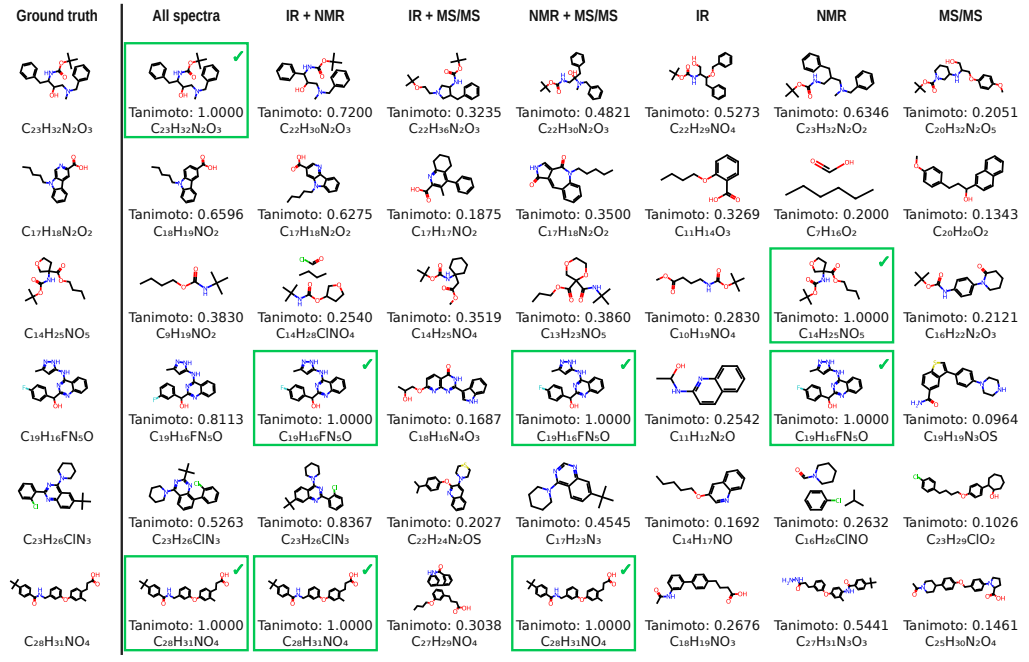

Figure A11: **Molecular structures generated using latent fusion on the Multimodal Spectroscopic Dataset.** Predictions for the same six molecules and spectral input configurations as in Figure A10. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures.

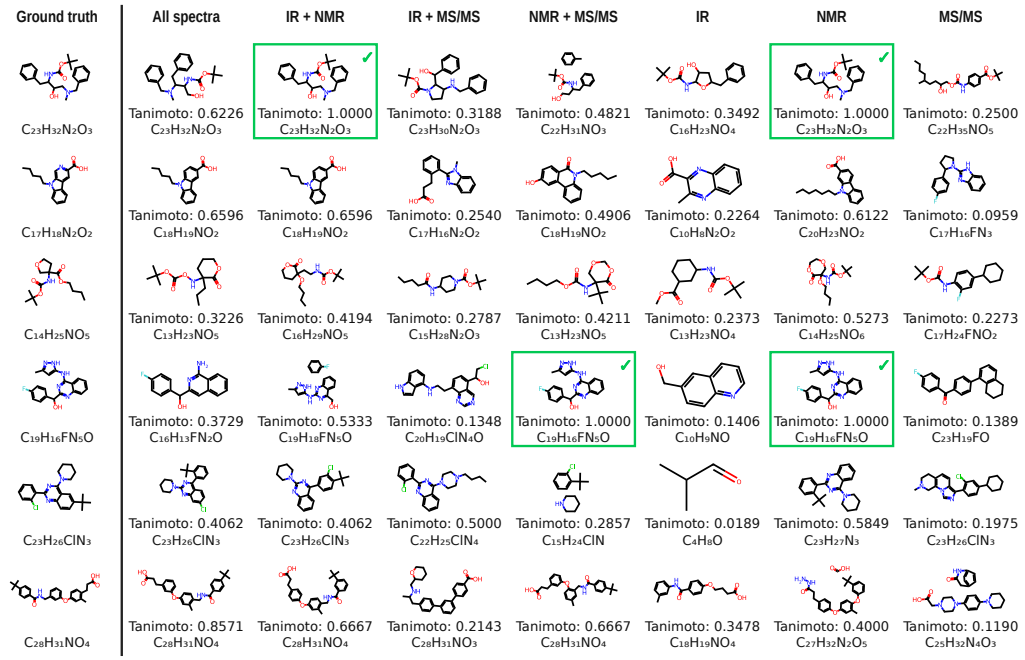

Figure A12: **Molecular structures generated using molecular-interaction fusion on the Multimodal Spectroscopic Dataset.** Predictions for the same six molecules and spectral input configurations as in Figure A10. Numbers report Tanimoto similarity to the reference; green outlines and check marks indicate exact matches to the reference structures.

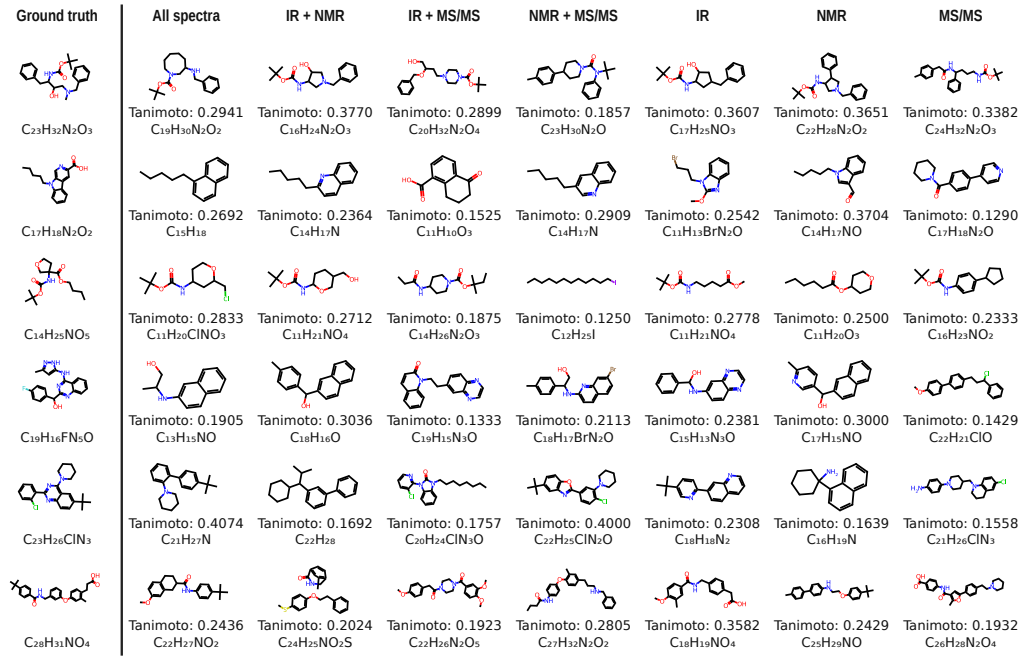

Figure A13: **Molecular structures generated using late fusion on the Multimodal Spectroscopic Dataset.** Predictions for the same six molecules and spectral input configurations as in Figure A10. Numbers report Tanimoto similarity to the reference; none of the displayed predictions exactly matches its reference structure.

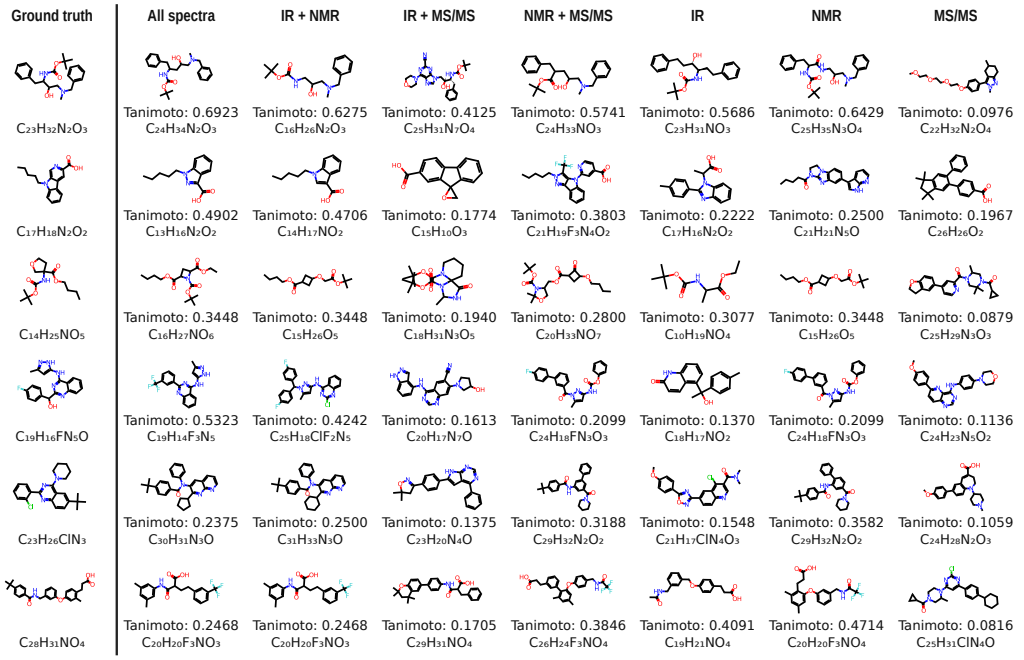

Figure A14: **Molecular structures generated by SpectraLLM on the Multimodal Spectroscopic Dataset.** Predictions for the same six molecules and spectral input configurations as in Figure A10. Numbers report Tanimoto similarity to the reference; none of the displayed predictions exactly matches its reference structure.