NMIRACLE: MULTI-MODAL GENERATIVE MOLECULAR ELUCIDATION FROM IR AND NMR SPECTRA

Federico Ottomano[1], **Yingzhen Li**[2], **Alex M. Ganose**[1]
[1]Department of Chemistry, Imperial College London
[2]Department of Computing, Imperial College London
f.ottomano@imperial.ac.uk

ABSTRACT

Molecular structure elucidation from spectroscopic data is a long-standing challenge in Chemistry, traditionally requiring expert interpretation. We introduce NMIRacle, a two-stage generative framework that builds upon recent paradigms in AI-driven spectroscopy with minimal assumptions. In the first stage, NMIRacle trains a generator to reconstruct molecular structures from count-aware fragment representations, capturing both fragment identities and their occurrences. In the second stage, a spectral encoder maps input spectra (IR, ^{1}H-NMR, ^{13}C-NMR) into a latent embedding used to condition the pre-trained generator, which is fine-tuned for direct spectra-to-molecule generation. This formulation bridges fragment-level chemical modeling with spectral evidence, yielding accurate molecular predictions. Empirical results demonstrate that NMIRacle outperforms existing baselines on molecular elucidation, while maintaining robust performance across increasing levels of molecular complexity. NMIRacle code is publicly available at NMIRacle code is publicly available at https://github.com/fedeotto/nmiracle.

1 INTRODUCTION

Determining the molecular structure of an unknown compound through spectroscopy is a fundamental problem in Chemistry, central to drug discovery, metabolomics, and materials design. This task is challenging due to the combinatorial explosion of possible atomic arrangements: even for molecules with fewer than 36 heavy atoms, the size of drug-like chemical space could exceed $\sim 10^{33}$ (Polishchuk et al., 2013). Techniques including *infrared* (IR) spectroscopy, *nuclear magnetic resonance* (NMR) spectroscopy and *mass spectrometry* (MS) provide complementary yet indirect evidence of the molecular structure, and interpreting them requires integrating heterogeneous and often noisy signals. Traditionally, structure elucidation relies on expert-driven spectral interpretation or database matching. These strategies are limited by subjectivity, the need for extensive chemical expertise, and the inability to identify molecules absent from reference libraries. Recent advances in deep learning have opened new directions for automated elucidation, including (i) cross-modal retrieval systems that learn shared embeddings of spectra and molecular structures (Yang et al., 2021; Jin et al., 2025; Mirza & Jablonka, 2024), and (ii) *de novo* generative frameworks that direcly predict molecular graphs or sequences from spectroscopic evidence (Bohde et al., 2025; Litsa et al., 2023; Guo et al., 2024; Yang et al., 2026). While retrieval-based methods leverage existing databases to identify the closest-matching structures, *de novo* generative approaches do not depend on pre-existing molecular libraries, making them inherently more flexible and capable of proposing novel compounds. However, this fully generative formulation poses substantial challenges: the model must integrate multiple spectra modalities with distinct noise characteristics and resolution biases, and learn a high-dimensional, multimodal mapping from continuous spectra to discrete molecular representations. A more comprehensive discussion of related work in Appendix A.1. Despite the availability of new datasets and benchmarks (Bushuiev et al., 2024; Guo et al., 2024), current spectra-to-molecule generative methods typically exhibit one or more limitations: (i) reliance on a single spectral modality, which neglects complementary patterns (Litsa et al., 2023; Bohde et al., 2025; Bushuiev et al., 2024); (ii) dependence on extensive pre-processing (e.g., peak extraction, multiplet assignment) to convert spectra into symbolic or text-based inputs (Alberts et al., 2023; Yao et al., 2023; Jin et al., 2025); (iii) assumptions of strong prior information, such as chemical formula or molecular scaffold

(Alberts et al., 2024; Wang et al., 2025b), which are rarely available under realistic experimental conditions; iv) limited benchmarking settings, restricted to molecules composed of only a few chemical species (typically C, N, O) and fewer than 20 heavy (non-hydrogen) atoms (Hu et al., 2024).
In this work, we tackle the most challenging formulation of molecular structure elucidation: direct generation of molecular structures from raw, multi-spectra input. We build upon previous established paradigms in data-driven molecular elucidation from spectroscopy with minimal assumptions (Hu et al., 2024). We introduce **NMIRacle**, a generative framework that learns from spectroscopic intensity arrays, the same data produced by experimental instruments, requiring only minimal preprocessing to handle modality-specific signal characteristics. This setup is intentionally difficult, as the model must infer structural constraints from noisy, high-dimensional inputs, but it enables greater realism and generalization across multiple acquisition settings. We evaluate NMIRacle on a multimodal spectroscopic dataset comprising molecules with up to 35 heavy atoms and diverse chemical compositions (Alberts et al., 2024). Our framework consistently obtains strong molecular elucidation performance across a broad range of molecular sizes and structural complexities. We summarize the main contributions of this work below:

- We propose NMIRacle, a generative framework for molecular structure elucidation from spectroscopy, operating directly on combinations of raw IR, ^{1}H-NMR, and ^{13}C-NMR spectra.
- We leverage count-aware fragment representations as an alternative to the binary indicators commonly used in existing frameworks. We demonstrate that capturing fragment occurrences provides a more faithful structural representation of molecules that effectively transfers to the downstream spectra-to-molecule task.
- We design a multi-spectral encoder that fuses raw IR, ^{1}H-NMR, and ^{13}C-NMR signals through intra- and inter-spectral attention.
- We demonstrate strong performance on molecular elucidation and robust generalization to complex molecules under minimal input assumptions.

2 METHODS

2.1 PROBLEM FORMULATION

We formulate molecular structure elucidation as a conditional generative modeling task. Given a set of complementary spectroscopic measurements $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_N\}$, where each $\mathbf{s}_i \in \mathbb{R}^{n_i}$ is a raw intensity vector sampled over the measurement domain of modality i, the goal is to generate the corresponding molecular structure $\mathcal{M}$. While the number and type of modalities N are fixed for a given model instance, our framework can be applied to any subset of available spectroscopic data. In practice, we represent each molecule by its SMILES sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$, which provides full information about atom types and connectivity. We assume access to a dataset $\mathcal{D} = \{(\mathcal{S}^{(m)}, \mathbf{y}^{(m)})\}_{m=1}^{M}$ of paired spectra–molecule examples. The learning objective is to estimate model parameters θ that maximize the likelihood of generating the correct SMILES given the corresponding input spectra:

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{(\mathcal{S},\mathbf{y})\sim\mathcal{D}}[\log p_\theta(\mathbf{y} \mid \mathcal{S})]\,. \tag{1}$$

Each spectral modality provides complementary structural evidence. We focus on three common techniques: IR spectroscopy captures vibrational modes of molecular bonds; *proton-NMR* (^{1}H-NMR) spectroscopy measures hydrogen environments and connectivity; *carbon-NMR* spectroscopy (^{13}C-NMR) probes carbon backbone structure (Clayden et al., 2012, Chapter 13).

2.2 SPECTRA PRE-PROCESSING

We convert raw spectral data from different analytical techniques into unified sequence representations suitable for transformer-based processing.

IR and ^{1}H-NMR We apply minimum amount of pre-processing for these spectra modalities, since peak shapes and relative intensities contain valuable structural information. These are normalized to the $[0, 1]$ range to ensure consistent intensity scales across samples. These continuous intensity profiles are used directly as inputs to the model.

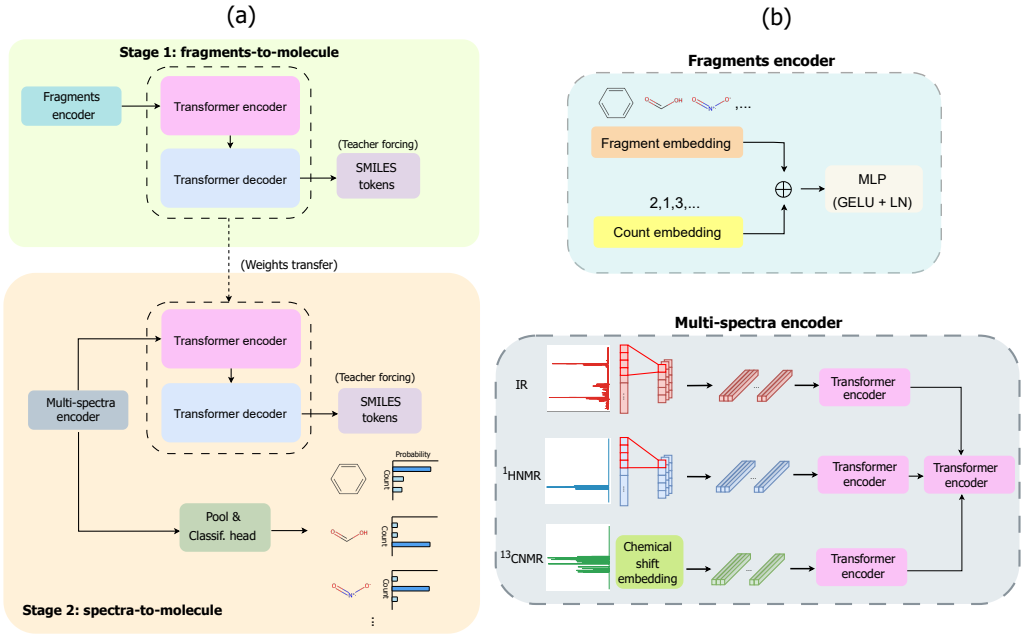

Figure 1: Overview of NMIRacle. (a) The model is trained in two stages: Stage 1 learns a fragment-conditioned molecular generator that reconstructs full molecular structures from count-aware fragment representations, establishing a molecular prior $p_\phi(\mathbf{y} \mid \mathbf{c})$. Stage 2 introduces a multi-spectra encoder that maps combinations of raw IR, ^{1}H-NMR, and ^{13}C-NMR spectra into latent embeddings $\mathbf{z}_\psi(\mathcal{S})$, used to condition the pre-trained generator for direct spectra-to-molecule generation. (b) The architecture integrates (i) a count-aware fragment encoder that embeds molecular fragments and their occurrences, and (ii) a multi-spectra encoder that fuses complementary spectra modalities into a unified latent representation.

^{13}C-NMR For carbon-NMR, peak intensities are not reliable indicators of carbon counts and so, following previous work (Mirza & Jablonka, 2024), we focus on chemical shift positions rather than intensities. We detect peaks in the raw array using SciPy's `find_peaks` function with a threshold of 10% relative to the maximum intensity. The detected peak positions are mapped from array indices to chemical shift values across the 0–220 ppm range. This range is then discretized into 80 equal-width bins of approximately 2.75 ppm each, and we create a binary vector indicating the presence or absence of peaks in each bin.

2.3 Method Overview

We conceptualize molecular structure generation from spectra as a two-stage conditional generative process, building upon previous work (Hu et al., 2024; Bohde et al., 2025). Figure 1 provides a visual overview of the proposed framework.

Global fragment vocabulary We define a vocabulary of chemical substructures $\mathcal{V} = \{f_1, f_2, \ldots, f_{|\mathcal{V}|}\}$ that serves as a discrete compositional basis for representing fragment compositions of molecules. Specifically, we curate a fragments vocabulary including 991 SMARTS patterns covering a broad range of common organic motifs.

Molecular representation We consider two representations for a molecule $\mathcal{M}$: (i) a fine-grained sequence of SMILES tokens $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ encoding full information about atom types and connectivity; (ii) a coarse fragments vector $\mathbf{c}$ capturing the fragment composition of the underlying molecule. Specifically, $\mathbf{c} = (c_1, c_2, \ldots, c_{|\mathcal{V}|}) \in \mathbb{N}_0^{|\mathcal{V}|}$, where c_j indicates the number of occurrences of fragment f_j in $\mathcal{M}$. This compact representation captures information about both fragment identities and their occurrences.

Two-stage modeling The overall spectra-to-molecule model is trained in two stages: (i) In Stage 1, we pre-train a fragment-conditioned generative model $p_\phi(\mathbf{y} \mid \mathbf{c})$ that learns to reconstruct a molecular SMILES sequence from its corresponding fragment composition $\mathbf{c}$; (ii) In Stage 2, a spectra encoder q_ψ, trained from scratch, maps spectroscopic measurements $\mathcal{S}$ into a continuous embedding $\mathbf{z}_\psi(\mathcal{S})$ that conditions the pre-trained generator from Stage 1. Under this formulation, the latter is fine-tuned to approximate the true conditional distribution

$$p(\mathbf{y} \mid \mathcal{S}) \approx p_\phi(\mathbf{y} \mid \mathbf{z}_\psi(\mathcal{S}))\,. \tag{2}$$

Conceptually, this can be viewed as replacing the marginalization

$$p(\mathbf{y} \mid \mathcal{S}) = \sum_{\mathbf{c}} p(\mathbf{y} \mid \mathbf{c})\, q(\mathbf{c} \mid \mathcal{S})\,, \tag{3}$$

with a deterministic point estimate of the fragment composition induced by the spectral embedding $\mathbf{z}_\psi(\mathcal{S})$. In other words, $\mathbf{z}_\psi(\mathcal{S})$ serves as a continuous surrogate for the (unknown) fragment composition $\mathbf{c}$, enabling the generator to transfer from fragment-conditioned pre-training to spectra-conditioned fine-tuning.

2.3.1 Stage 1: Fragments-to-molecule pre-training

In the first stage, we learn parameters ϕ of a conditional generative model $p_\phi(\mathbf{y} \mid \mathbf{c})$ that reconstructs a molecular SMILES sequence $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ from a corresponding coarse fragments vector $\mathbf{c}$. Previous approaches typically adopt a binary fragment encoding, where each entry $c_j \in \{0, 1\}$ indicates the presence or absence of fragment f_j (Bohde et al., 2025; Hu et al., 2024). In contrast, we employ a count-aware fragment representation, where $c_j \in \mathbb{N}$ denotes the number of occurrences of each fragment in the molecule. This representation provides a more faithful description of molecular composition, enabling the model to capture structural regularities that depend on fragment repetition (e.g., ring patterns, chain extensions). Each fragment type f_j and its associated count c_j are independently embedded:

$$\mathbf{h}_{f_j} = \text{Embed}_f(f_j), \quad \mathbf{h}_{c_j} = \text{Embed}_c(c_j) \in \mathbb{R}^d, \tag{4}$$

where $\text{Embed}_f(\cdot)$ and $\text{Embed}_c(\cdot)$ denote learnable lookup tables for fragment types and occurrences, respectively, while d indicates the hidden dimensionality. The two embeddings are combined through element-wise addition, followed by a non-linear transformation:

$$\mathbf{h}_j = \text{LayerNorm}(\text{MLP}(\mathbf{h}_{f_j} + \mathbf{h}_{c_j})) \in \mathbb{R}^d\,, \tag{5}$$

where MLP denotes a single-hidden-layer perceptron with GELU activation. The resulting set of count-aware fragment embeddings $\{\mathbf{h}_j\}$ serves as input tokens to the transformer encoder, which provides contextualized representations for decoding. Conditioned on this context, the decoder autoregressively predicts SMILES tokens:

$$\mathcal{L}_{\text{Stage1}}(\phi) = \mathbb{E}_{(\mathbf{c},\mathbf{y})}\Big[-\sum_{t=1}^{T} \log p_\phi(y_t \mid y_{<t}, \{\mathbf{h}_j\})\Big], \tag{6}$$

minimizing the standard autoregressive negative log-likelihood.

2.3.2 Stage 2: spectra-to-molecule fine-tuning

In the second stage, we fine-tune the fragment-conditioned generator $p_\phi(\mathbf{y} \mid \mathbf{c})$, previously trained under the count-aware fragment encoding scheme, to map spectroscopic measurements $\mathcal{S}$ directly to molecular SMILES. Rather than conditioning on count-aware fragment encodings $\{\mathbf{h}_j\}$, the model now conditions on latent spectral embeddings produced by a multi-spectral encoder q_ψ (Eq. 2). These embeddings serve as a continuous proxy for the fragment-level representation learned in pre-training, thereby preserving the same generative interface while adapting it to spectral inputs.

Multi-spectra encoder Each input spectrum $\mathbf{s}_i \in \mathbb{R}^{n_i}$ from modality $i \in \{\text{IR}, {}^1\text{H-NMR}, {}^{13}\text{C-NMR}\}$ is processed by a modality-specific encoder $E^{(i)}_{\text{spec}}$. The encoder extracts spectral features and projects them into a shared embedding space of dimension d. For IR and ^{1}H-NMR spectra, we first apply 1D convolutional layers to capture local peak patterns and

compress the signal into feature maps $\mathbf{Z}_i \in \mathbb{R}^{s_i \times c_i}$. A learnable linear projection $P^{(i)} \in \mathbb{R}^{c_i \times d}$ maps these features to token embeddings with hidden dimensionality d. To retain spectral ordering, we add learnable positional encodings $\mathbf{W}_i^{pos} \in \mathbb{R}^{s_i \times d}$:

$$\mathbf{Z}_i^{\text{seq}} = P^{(i)}(\mathbf{Z}_i) + \mathbf{W}_i^{pos}\,. \tag{7}$$

We ablate the impact of learnable positional encodings against sinusoidal positional encodings in Appendix A.3. For ^{13}C-NMR spectra, where inputs are discrete chemical shift indices rather than continuous peaks, we omit positional encodings and instead use a learnable embedding lookup for each non-zero bin index. Each modality sequence $\mathbf{Z}_i^{\text{seq}}$ is then passed to an intra-modal transformer encoder to model local dependencies among peaks within the same spectrum

$$\mathbf{H}_i = \text{TEnc}_{\text{intra}}^{(i)}(\mathbf{Z}_i^{\text{seq}}) \in \mathbb{R}^{s_i \times d}, \tag{8}$$

producing modality-specific embeddings. The encoded modalities are concatenated and fed to a separate, inter-modal transformer encoder:

$$\mathbf{H}_{\text{inter}} = \text{TEnc}_{\text{inter}}([\mathbf{H}_1; \mathbf{H}_2; \mathbf{H}_3]) \in \mathbb{R}^{s \times d}\,, \tag{9}$$

where ; denotes concatenation along the sequence dimension, and $s = \sum_i s_i$ is the resulting sequence length. This enables dedicated learning between distinct modalities (e.g., associating IR absorption bands with ^{1}H chemical shifts linked to the same functional groups). The obtained representation $\mathbf{H}_{\text{inter}}$ replaces the count-aware fragment tokens $\{\mathbf{h}_j\}$ used in Stage 1 (Eq. 6) as contextual input to the pre-trained model p_ϕ, thus conditioning molecular generation directly on spectral features.

Fragment composition head To enhance fragment-level supervision, we adopt a multi-task setup (Hu et al., 2024) optimizing concurrently the model for SMILES reconstruction (Eq. 6) and for predicting fragment compositions. First, the fused representation $\mathbf{H}_{\text{inter}}$ is mean-pooled to a global feature vector:

$$\mathbf{h}_{\text{inter}} = \text{MeanPool}(\mathbf{H}_{\text{inter}}) \in \mathbb{R}^d. \tag{10}$$

Each fragment identity f_j is represented by a one-hot vector $\mathbf{e}_{f_j}$ from the fragment vocabulary. For each fragment, we concatenate $\mathbf{h}_{\text{inter}}$ and $\mathbf{e}_{f_j}$ and predict a categorical distribution over possible counts:

$$p_\psi(c_j \mid \mathcal{S}, f_j) = \text{Softmax}\big(\text{MLP}[\mathbf{h}_{\text{inter}}; \mathbf{e}_{f_j}]\big)\,, \tag{11}$$

where $c_j \in \{c_0, \ldots, c_{\text{max}}\}$, c_{max} represents the maximum observed occurrences of a fragment in a molecule, $\text{MLP}(\cdot)$ denotes a single-hidden-layer perceptron with GELU activation, and ; denotes feature-wise concatenation. This formulation enables the model to learn both fragment presence and occurrence directly from spectral evidence.

Training objective During Stage 2, the pre-trained generator $p_\phi(\mathbf{y} \mid \mathbf{c})$ from Stage 1 is fine-tuned with spectra conditioning, while the spectra encoder q_ψ is trained from scratch. The overall objective combines (i) a sequence-level cross-entropy loss for molecular reconstruction and (ii) a fragment-level cross-entropy loss over discrete fragment occurrences:

$$\begin{aligned}\mathcal{L}_{\text{Stage2}}(\phi, \psi) = \alpha\, \mathbb{E}_{(\mathcal{S},\mathbf{y})}\Big[-\sum_{t=1}^{T} \log p_\phi(y_t \mid y_{<t}, \mathbf{z}_\psi(\mathcal{S}))\Big] \\ + \beta\, \mathbb{E}_{(\mathcal{S},\mathbf{c})}\Big[-\sum_{j=1}^{|\mathcal{V}|} \log p_\psi(c_j \mid \mathcal{S}, f_j)\Big]\,,\end{aligned} \tag{12}$$

where α and β balance the contributions of the two tasks, $p_{\phi,\psi}$ denotes the fine-tuned generator with spectra conditioning, and $p_\psi(c_j \mid \mathcal{S}, f_j)$ parameterizes the fragment composition head (Eq. 11). In practice, we set $\alpha = \beta = 1$. This multi-task setup encourages the latent representation $\mathbf{z}_\psi(\mathcal{S})$ to encode fragment compositions for molecular generation.

3 EXPERIMENTS

3.1 DATASETS

We employ two complementary datasets for our experiments: a molecular pre-training dataset for Stage 1 and a spectra fine-tuning dataset for Stage 2.

Molecular pre-training dataset We build upon an existing molecular dataset employed in previous work (Hu et al., 2024), comprising approximately $\sim$ 3.1M molecules, obtained by combining $\sim$ 3M compounds randomly sampled from the GDB-17 database (Ruddigkeit et al., 2012) with an additional $\sim$ 140k entries sourced from SpectraBase (John Wiley & Sons, Inc.). While this collection provides a large set of molecules, it is chemically-limited, containing only carbon (C), oxygen (O), and nitrogen (N) atoms, and restricted to a maximum of 19 heavy atoms per molecule. Such constraints make it poorly representative of the molecular diversity encountered in experimental settings. To address this limitation, we extend the original pre-training pool with $\sim$ 670k molecules from a recent multimodal spectroscopic dataset introduced by Alberts et al. (2024). This augmentation increases chemical diversity up to 9 distinct elements and extends molecular size up to 35 heavy (non-hydrogen) atoms, thereby exposing the pre-training model to richer compositional and structural variations.

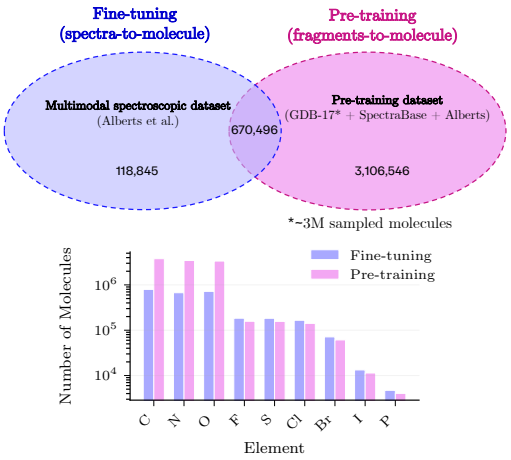

Figure 2: (Top) Venn diagrams illustrate the overlap between the molecular pre-training dataset (derived from GDB-17 and SpectraBase) and the additional molecules incorporated from Alberts et al. (2024) dataset. (Bottom) Element distribution across the utilized datasets, highlighting the broader chemical diversity introduced by the data augmentation.

Spectra fine-tuning dataset We utilize a recently proposed multimodal spectroscopic dataset (Alberts et al., 2024) as the main benchmark for spectra-to-molecule task (Stage 2). It contains over $\sim$ 790k molecules paired with various simulated spectra, including IR, ^{1}HNMR and ^{13}CNMR. We split the dataset into training, validation and test subsets in an 8:1:1 ratio. The training split ($\sim$ 670k SMILES) corresponds to the augmentation performed on the molecular pre-training dataset. Crucially, we ensure no molecules from either pre-training or training data are present in the test set. This guarantees that the final evaluation measures the model's ability to predict entirely unseen molecules, only from spectral evidence.

3.2 BASELINES

We employ different baselines to compare the performance of the proposed approach.

SMILES/SELFIES transformers We implement transformer models that operate on simple concatenations of spectra features. To stay consistent with the proposed methodology, which assumes minimal pre-processing on input spectra, we apply minimal feature extractors: 1D convolutional layers for continuous spectra (IR, ^{1}H-NMR) and a learnable lookup embedding for ^{13}C-NMR peak bins. The resulting representations are concatenated across modalities and provided to an encoder-decoder transformer that generates molecules in either SMILES or SELFIES format. This setup is inspired by the benchmark provided in the work of Alberts et al. (2024), but differs in that we avoid domain-specific pre-processing (e.g., MestreNova (Willcott, 2009) peak extraction) and instead let the neural encoders discover spectral patterns directly from raw data.

NMR2Struct We evaluate NMR2Struct (Hu et al., 2024), a two-stage framework for spectra-to-molecule prediction. In Stage 1, a Transformer-based molecular generator is pre-trained to recon-

struct SMILES sequences from binary fragment indicators. In Stage 2, a spectral encoder is trained to produce embeddings that condition the pre-trained generator, which is fine-tuned for spectra-to-molecule generation. We re-implement NMR2Struct within our experimental setting to ensure a controlled and fair comparison.

Spec2Mol We evaluate Spec2Mol (Litsa et al., 2023), which follows a two-stage, latent-space alignment approach. In Stage 1, a GRU-based SMILES autoencoder is trained to reconstruct molecular SMILES, yielding a continuous latent representation. In Stage 2, a convolutional spectra encoder is trained to minimize the ℓ_2-loss between spectral embeddings and SMILES latent representations produced by the autoencoder encoder. At inference, spectra are embedded using the trained spectra encoder and decoded into SMILES using the autoencoder decoder. We re-implement Spec2Mol within our experimental setting to ensure a controlled and fair comparison.

3.3 Results and discussion

Table 1 reports results for the pre-training stage for models requiring it (NMR2Struct, NMIRacle, and Spec2Mol). Extended results for this stage are provided in Appendix A.2. Table 2 summarizes performance on the spectra-to-molecule generation task across multiple molecular-level metrics, with formal metric definitions given in Appendix A.5. All evaluations follow an enantiomer-aware protocol: predicted and reference molecules are considered equivalent if their canonical SMILES strings match exactly or correspond to enantiomeric (mirror-image) configurations. Details of the evaluation procedure are described in Appendix A.4.

Fragments-to-molecule Table 1 reports fragment-to-molecule reconstruction performance on $10,000$ molecules sampled from the pre-training test set. We compare three distinct pre-training strategies: (i) fragment-based reconstruction with binary fragment indicators (NMR2Struct), (ii) fragment-based reconstruction with count-aware fragment representations (ours), and (iii) SMILES-to-SMILES reconstruction using an autoencoder (Spec2Mol). Fragment-based pre-training poses a fundamentally different and more undetermined reconstruction problem than SMILES autoencoding. Given a set of fragments, the model must infer atom-level connectivity and global molecular topology, none of which are explicitly specified in the input. In terms of fragment-based reconstruction paradigm, incorporating fragment occurrences consistently improves reconstruction performance over binary fragment indicators. Top-1 accuracy increases from 0.63 to 0.70, and Top-10 accuracy from 0.76 to 0.81, while maintaining near-perfect chemical validity. These gains indicate that count-aware fragment representations provide additional, quantitative constraints that help resolve ambiguities in molecular assembly, leading to more faithful reconstructions within the fragment-based paradigm. In contrast, SMILES autoencoding provides a complete description of the molecular graph, making reconstruction substantially easier. Therefore, as expected, Spec2Mol achieves the highest reconstruction accuracy in this stage.

Table 1: Pre-training results on 10,000 test molecules. Results are reported in terms of chemical validity, graph edit distance (MCES), and Top-k accuracies. For continuous metrics, values are reported as mean $\pm$ standard deviation computed across test molecules. * indicates our implementations of baseline approaches.

Pre-training	**Valid** (↑)	**MCES** (↓)	**Top-k Acc.** (↑)		
			1	10	15
NMR2Struct*	**1.00**	$0.92_{\pm 2.32}$	0.63	0.76	0.81
NMIRacle (Ours)	0.97	$\mathbf{0.57}_{\pm 2.32}$	**0.70**	**0.81**	**0.86**
Spec2Mol*	1.00	$0.07_{\pm 0.67}$	0.96	0.98	0.98

Spectra-to-molecule Results for molecular generation from multi-spectra inputs are reported in Table 2. Across all evaluated spectral combinations, NMIRacle consistently achieves the strongest performance. When all three modalities (IR, ^{1}H-NMR, ^{13}C-NMR) are available, NMIRacle attains a Top-1 accuracy of 0.48 and a Top-15 accuracy of 0.66, outperforming NMR2Struct (0.41 / 0.58) and all other baselines. Models that rely on more constrained or less aligned intermediate representations perform substantially worse. For instance, SELFIES-based transformer exhibits weaker performance: while SELFIES guarantees chemical validity, it may also reduce flexibility in conditional generative modeling, leading to worse performance on other structural metrics (Skinnider,

Table 2: Performance comparison for the spectra-to-molecule task across different spectral combinations. Results are reported in terms of chemical validity, structural similarity (Tanimoto), graph edit distance (MCES), string-level distance (Levenshtein), and Top-k accuracies. For continuous metrics, values are reported as mean $\pm$ standard deviation computed across test molecules. * indicates our implementations of existing baseline approaches.

Spectra	Model	Valid (↑)	Tanimoto (↑)			MCES (↓)	Lev. (↓)	Top-k Acc. (↑)			
			Morgan	MACCS	RDKit			1	5	10	15
^{1}H + ^{13}C-NMR	SMILES Transformer	1.00	$0.72_{\pm 0.25}$	$0.90_{\pm 0.13}$	$0.77_{\pm 0.24}$	$4.21_{\pm 4.40}$	$7.01_{\pm 8.33}$	0.25	0.31	0.36	0.38
	SELFIES Transformer	1.00	$0.61_{\pm 0.26}$	$0.85_{\pm 0.14}$	$0.67_{\pm 0.24}$	$5.70_{\pm 4.57}$	$10.27_{\pm 9.65}$	0.15	0.20	0.23	0.24
	NMR2Struct*	1.00	$0.79_{\pm 0.24}$	$0.93_{\pm 0.11}$	$0.82_{\pm 0.23}$	$3.17_{\pm 4.15}$	$5.26_{\pm 7.58}$	0.35	0.41	0.47	0.50
	Spec2Mol*	1.00	$0.32_{\pm 0.13}$	$0.69_{\pm 0.12}$	$0.45_{\pm 0.14}$	$10.78_{\pm 3.78}$	$17.76_{\pm 9.61}$	0.00	0.00	0.01	0.01
	NMIRacle (Ours)	1.00	$\mathbf{0.82}_{\pm 0.23}$	$\mathbf{0.94}_{\pm 0.11}$	$\mathbf{0.85}_{\pm 0.22}$	$\mathbf{2.72}_{\pm 3.96}$	$\mathbf{4.49}_{\pm 7.13}$	**0.39**	**0.45**	**0.52**	**0.56**
IR + ^{1}H-NMR	SMILES Transformer	1.00	$0.77_{\pm 0.24}$	$0.93_{\pm 0.10}$	$0.81_{\pm 0.23}$	$3.45_{\pm 4.11}$	$5.79_{\pm 7.77}$	0.30	0.36	0.42	0.45
	SELFIES Transformer	1.00	$0.64_{\pm 0.26}$	$0.88_{\pm 0.13}$	$0.70_{\pm 0.13}$	$5.16_{\pm 4.46}$	$9.52_{\pm 9.57}$	0.18	0.22	0.27	0.28
	NMR2Struct*	1.00	$0.82_{\pm 0.23}$	$0.95_{\pm 0.09}$	$0.85_{\pm 0.21}$	$2.71_{\pm 3.86}$	$4.09_{\pm 6.65}$	0.38	0.44	0.51	0.54
	Spec2Mol*	1.00	$0.29_{\pm 0.12}$	$0.67_{\pm 0.12}$	$0.42_{\pm 0.13}$	$10.70_{\pm 3.78}$	$17.77_{\pm 8.91}$	0.00	0.00	0.00	0.00
	NMIRacle (Ours)	1.00	$\mathbf{0.86}_{\pm 0.21}$	$\mathbf{0.96}_{\pm 0.08}$	$\mathbf{0.89}_{\pm 0.19}$	$\mathbf{2.06}_{\pm 3.49}$	$\mathbf{3.52}_{\pm 6.50}$	**0.45**	**0.50**	**0.59**	**0.63**
IR + ^{1}H-NMR + ^{13}C-NMR	SMILES Transformer	1.00	$0.76_{\pm 0.22}$	$0.93_{\pm 0.10}$	$0.80_{\pm 0.22}$	$3.59_{\pm 4.09}$	$6.22_{\pm 8.04}$	0.28	0.34	0.40	0.42
	SELFIES Transformer	1.00	$0.64_{\pm 0.26}$	$0.88_{\pm 0.12}$	$0.71_{\pm 0.24}$	$5.07_{\pm 4.39}$	$9.48_{\pm 9.50}$	0.17	0.22	0.26	0.28
	NMR2Struct*	1.00	$0.84_{\pm 0.22}$	$0.96_{\pm 0.09}$	$0.87_{\pm 0.20}$	$2.39_{\pm 3.67}$	$4.18_{\pm 6.97}$	0.41	0.47	0.55	0.58
	Spec2Mol*	1.00	$0.35_{\pm 0.15}$	$0.72_{\pm 0.12}$	$0.47_{\pm 0.15}$	$9.85_{\pm 4.28}$	$17.45_{\pm 9.94}$	0.00	0.01	0.01	0.01
	NMIRacle (Ours)	1.00	$\mathbf{0.88}_{\pm 0.20}$	$\mathbf{0.97}_{\pm 0.07}$	$\mathbf{0.90}_{\pm 0.18}$	$\mathbf{1.82}_{\pm 3.29}$	$\mathbf{3.21}_{\pm 6.23}$	**0.48**	**0.53**	**0.61**	**0.66**
IR + ^{13}C-NMR	SMILES Transformer	1.00	$0.55_{\pm 0.25}$	$0.85_{\pm 0.13}$	$0.64_{\pm 0.22}$	$6.37_{\pm 4.31}$	$11.94_{\pm 9.83}$	0.09	0.12	0.14	0.15
	SELFIES Transformer	1.00	$0.47_{\pm 0.23}$	$0.81_{\pm 0.13}$	$0.58_{\pm 0.21}$	$7.22_{\pm 4.22}$	$13.83_{\pm 10.10}$	0.05	0.08	0.09	0.10
	NMR2Struct*	1.00	$0.59_{\pm 0.26}$	$0.87_{\pm 0.13}$	$0.68_{\pm 0.23}$	$5.71_{\pm 4.39}$	$10.84_{\pm 9.84}$	0.12	0.16	0.20	0.21
	Spec2Mol*	1.00	$0.35_{\pm 0.15}$	$0.73_{\pm 0.12}$	$0.47_{\pm 0.15}$	$9.42_{\pm 4.19}$	$17.05_{\pm 10.14}$	0.00	0.01	0.01	0.01
	NMIRacle (Ours)	1.00	$\mathbf{0.63}_{\pm 0.26}$	$\mathbf{0.88}_{\pm 0.12}$	$\mathbf{0.71}_{\pm 0.23}$	$\mathbf{5.22}_{\pm 4.38}$	$\mathbf{9.84}_{\pm 9.62}$	**0.14**	**0.19**	**0.23**	**0.24**

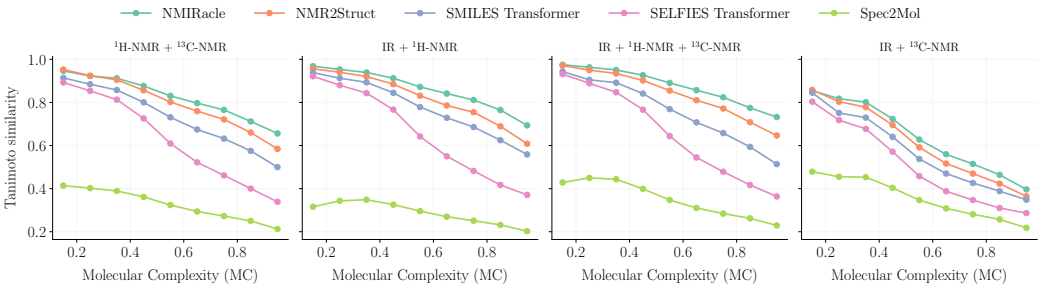

Figure 3: Model performance across molecular complexity bins for different spectral combinations. NMIRacle maintains higher Tanimoto similarity even for structurally-rich molecules.

2024), and making it harder for the model to resolve structural ambiguities from IR and NMR spectra. Spec2Mol also underperforms in the spectra-to-molecule task, despite its strong pre-training results. This highlights a limitation of SMILES autoencoding as a pre-training strategy for multi-spectra conditional generation: in our setting, aligning spectral representations with a continuous SMILES latent space proves significantly more difficult than conditioning generation on an intermediate, chemically-grounded fragment representation. Similar trends have been observed in recent spectra-to-molecule benchmarks (Bohde et al., 2025). Additional qualitative success cases are reported in Appendix A.11, illustrating molecules that are correctly recovered by NMIRacle while all competing methods fail.

Scaling to more complex molecules To assess models' robustness with respect to molecular size and structural diversity, we introduce an empirical *molecular complexity* (MC) index:

$$MC := \frac{1}{3}\left(\frac{N_h}{N_{h_{\max}}} + \frac{N_u}{N_{u_{\max}}} + \frac{N_r}{N_{r_{\max}}}\right), \tag{13}$$

where N_h, N_u, and N_r denote the number of heavy atoms, unique elements, and rings, respectively. $N_{(\cdot)_{\max}}$ corresponds to the 99-th percentile of the corresponding distribution. Molecules are partitioned into ten complexity bins, and Tanimoto similarity (based on Morgan fingerprints) is reported

per bin in Figure 3 across different spectral combinations. NMIRacle consistently outperforms baseline models across all complexity bins, maintaining a stable performance lead even as the overall Tanimoto similarity declines for structurally-rich molecules.

3.4 Summary of additional studies

We provide in Appendix A additional analyses that further characterize the proposed framework. In particular:

- In Appendix A.2 we report extended results for the pre-training stage, including additional molecular similarity metrics.
- In Appendix A.3 we present ablation studies isolating the contribution of learnable positional encodings and explicit inter-modal attention, showing that both components yield consistent performance gains.
- In Appendix A.4 we detail the enantiomer-aware evaluation protocol adopted in the main results. We further examine alternative evaluation criteria that relax stereochemical constraints, showing that a non-negligible fraction of errors arises from stereochemical ambiguity rather than incorrect connectivity.
- In Appendix A.6 we provide a detailed analysis of model failure cases, identifying fragment misprediction as the dominant source of error across different spectral settings.
- In Appendix A.7 we analyze fragment occurrence prediction under the multi-task objective (Eq. 12), showing that accurately predicted fragments rapidly cover most of the fragment occurrence space, highlighting practical utility for common structural motifs.

4 Conclusion

Motivated by recent advances in spectra-to-molecule machine learning, we present NMIRacle, a two-stage generative framework for molecular structure elucidation from combinations of raw IR, ^{1}H-NMR, and ^{13}C-NMR spectra. Our framework builds upon previous efforts towards spectra-to-molecule modeling with minimal assumptions. Our approach combines a count-aware fragment prior with a hierarchical, multi-spectra encoder, enabling informative spectral conditioning. Across multiple evaluation settings, NMIRacle achieves consistently strong molecular elucidation performance and exhibits robust generalization to structurally-complex molecules. Overall, NMIRacle provides a flexible foundation for realistic, data-driven molecular elucidation from spectral evidence. Additional discussion related to limitations and future directions is provided in A.8.

5 Acknowledgments

The authors acknowledge the AI for Chemistry: AIchemy hub for funding (EPSRC grant EP/Y028775/1 and EP/Y028759/1).

References

Marvin Alberts, Federico Zipoli, and Alain Vaucher. Learning the language of NMR: structure elucidation from NMR spectra using transformer models. In *AI for Accelerated Materials Design-NeurIPS 2023 Workshop*, 2023.

Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808, 2024.

Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. DiffMS: Diffusion generation of molecules conditioned on mass spectra. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=EvILcv2v8L.

Roman Bushuiev, Anton Bushuiev, Niek de Jonge, Adamo Young, Fleming Kretschmer, Raman Samusevich, Janne Heirman, Fei Wang, Luke Zhang, Kai Dührkop, et al. Massspecgym: A benchmark for the discovery and identification of molecules. *Advances in Neural Information Processing Systems*, 37:110010–110027, 2024.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning*, pp. 6140–6157. PMLR, 2023.

Jonathan Clayden, Nick Greeves, Stuart Warren, and Peter Wothers. *Organic Chemistry*. Oxford University Press, 2 edition, 2012.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. Domain-agnostic molecular generation with chemical feedback. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9rPyHyjfwP.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. Can LLMs solve molecule puzzles? a multi-modal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024.

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 Central Science*, 10(11):2162–2170, 11 2024. doi: 10.1021/acscentsci.4c01132. URL https://doi.org/10.1021/acscentsci.4c01132.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2323–2332. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/jin18a.html.

Wengong Jin, Dr.Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4839–4848. PMLR, 2020. URL https://proceedings.mlr.press/v119/jin20a.html.

Yongqi Jin, Jun-Jie Wang, Fanjie Xu, Xiaohong Ji, Zhifeng Gao, Linfeng Zhang, Guolin Ke, Rong Zhu, et al. NMR-Solver: Automated structure elucidation via large-scale spectral matching and physics-guided fragment optimization. *arXiv preprint arXiv:2509.00640*, 2025.

John Wiley & Sons, Inc. SpectraBase. https://spectrabase.com.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. Mol-LLaMA: Towards general understanding of molecules in large molecular language model. In *NeurIPS 2025 AI for Science Workshop*, 2025. URL https://openreview.net/forum?id=TTeQ3bOwsL.

Fleming Kretschmer, Jan Seipp, Marcus Ludwig, Gunnar W. Klau, and Sebastian Böcker. Small molecule machine learning: All models are wrong, some may not even be useful. *bioRxiv*, 2023. doi: 10.1101/2023.03.27.534311.

Gregory Landrum. RDKit: Open-source cheminformatics. URL https://www.rdkit.org.

Joongwon Lee, Seonghwan Kim, and Wou Youn Kim. FragFM: Efficient fragment-based molecular generation via discrete flow matching. *arXiv preprint arXiv:2502.15805*, 2025.

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. *Communications Chemistry*, 6(1):132, 2023. doi: 10.1038/s42004-023-00932-3. URL https://doi.org/10.1038/s42004-023-00932-3.

Gang Liu, Jiaxin Xu, Tengfei Luo, and Meng Jiang. Graph diffusion transformers for multi-conditional molecular generation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 8065–8092. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0260.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073, March 2024b. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.108073. URL http://dx.doi.org/10.1016/j.compbiomed.2024.108073.

Adrian Mirza and Kevin Maik Jablonka. Elucidating structures from spectra using multimodal embeddings and discrete optimization. *ChemRxiv*, 2024. doi: 10.26434/chemrxiv-2024-f3b18-v2.

Qizhi Pei, Rui Yan, Kaiyuan Gao, Jinhua Zhu, and Lijun Wu. 3D-MolT5: Leveraging discrete structural information for molecule-text modeling. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eGqQyTAbXC.

P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design*, 27(8): 675–679, 2013. doi: 10.1007/s10822-013-9672-4. URL https://doi.org/10.1007/s10822-013-9672-4.

Lars Ruddigkeit, Ruud van Deursen, Lorenz C. Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11):2864–2875, 11 2012. doi: 10.1021/ci300415d. URL https://doi.org/10.1021/ci300415d.

Michael A. Skinnider. Invalid smiles are beneficial rather than detrimental to chemical language models. *Nature Machine Intelligence*, 6(4):437–448, 2024. doi: 10.1038/s42256-024-00821-x. URL https://doi.org/10.1038/s42256-024-00821-x.

Gary Tom, Edwin Yu, Naruki Yoshikawa, Kjell Jorner, and Alán Aspuru-Guzik. Stereochemistry-aware string-based molecular generation. *PNAS Nexus*, 4(11):pgaf329, 2025. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgaf329. URL https://doi.org/10.1093/pnasnexus/pgaf329.

Liang Wang, Yu Rong, Tingyang Xu, Zhenyi Zhong, Zhiyuan Liu, Pengju Wang, Deli Zhao, Qiang Liu, Shu Wu, and Yang Zhang. DiffSpectra: Molecular structure elucidation from spectra using diffusion models. *arXiv preprint arXiv:2507.06853*, 2025a.

Yinkai Wang, Xiaohui Chen, Liping Liu, and Soha Hassoun. MADGEN: Mass-spec attends to de novo molecular generation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=78tc3EiUrN.

Mark Robert Willcott. MestRe nova. *Journal of the American Chemical Society*, 131(36):13180–13180, 2009. doi: 10.1021/ja906709t. URL https://doi.org/10.1021/ja906709t.

Yutong Xie, Chence Shi, Hao Zhou, Yuwei Yang, Weinan Zhang, Yong Yu, and Lei Li. MARS: Markov molecular sampling for multi-objective drug discovery. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=kHSu4ebxFXY.

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. *Materials Futures*, 5(1):015601, 2026.

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. *Analytical Chemistry*, 93(50):16947–16955, 12 2021. doi: 10.1021/acs.analchem.1c04307. URL https://doi.org/10.1021/acs.analchem.1c04307.

Lin Yao, Minjian Yang, Jianfei Song, Zhuo Yang, Hanyu Sun, Hui Shi, Xue Liu, Xiangyang Ji, Yafeng Deng, and Xiaojian Wang. Conditional molecular generation net enables automated structure elucidation based on ^{13}C NMR spectra and prior knowledge. *Analytical Chemistry*, 95(12):5393–5401, 03 2023. doi: 10.1021/acs.analchem.2c05817. URL https://doi.org/10.1021/acs.analchem.2c05817.

Huaisheng Zhu, Teng Xiao, and Vasant G Honavar. 3M-Diffusion: Latent multi-modal diffusion for language-guided molecular structure generation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=DomBynQsqt.

A APPENDIX

A.1 RELATED WORK

Data-driven molecular elucidation from spectroscopy Molecular structure elucidation has recently emerged as a benchmark for multimodal AI, with several works addressing the task under diverse settings. Guo et al. (2024) introduced MolPuzzle, a zero-shot benchmark framing structure elucidation as a multi-step reasoning task integrating spectral analysis, property inference, and functional groups assembly. MassSpecGym (Bushuiev et al., 2024) provides standardized metrics and curated datasets for molecular *de novo* generation and retrieval from mass spectra. Alberts et al. (2024) present a large-scale multimodal dataset of $\sim$790k molecules paired with simulated spectroscopic data, providing a unified benchmark for multimodal structure elucidation. Mirza & Jablonka (2024) align spectral and molecular embeddings via contrastive learning for cross-modal retrieval, further employing a genetic algorithm to introduce novelty among retrieved molecular candidates. DiffNMR (Yang et al., 2026) employs a conditional discrete diffusion model to perform de novo molecular structure elucidation from NMR spectra, iteratively refining the molecular graph structure and using a two-stage pre-training for enhanced spectra-molecule alignment. Spec2Mol (Litsa et al., 2023) reconstructs molecular SMILES via a gated recurrent unit (GRU)–based autoencoder, then aligns a convolution-based spectral encoder to the molecular latent space, enabling direct reconstruction from mass spectra. DiffMS (Bohde et al., 2025) introduces a discrete diffusion framework for molecular graph generation, in which a graph transformer denoises adjacency matrices conditioned on molecular fingerprints predicted from mass spectra. NMR2Struct (Hu et al., 2024) employs an autoregressive multi-task setup in which a fragment-based generative model is reused for joint SMILES generation and fingerprint prediction conditioned on input spectra. DiffSpectra (Wang et al., 2025a) introduces a diffusion-based framework for elucidating 3D molecules from Ultraviolet-visible (UV-Vis), IR, and Raman spectra, employing an SE(3)-equivariant architecture to jointly infer the 2D topology and 3D geometry of the molecule. Despite these advances, most AI-driven molecular elucidation methods remain single-modality, reliant on pre-processed inputs unavailable from experimental data, and evaluated on small molecules, falling short of realistic, multi-spectra elucidation scenarios.

Conditional generative models for molecules Conditional molecular generation represents a central paradigm in AI-driven Chemistry, enabling the targeted design under textual, structural or multi-modal constraints. Text-driven inverse design via large language models (Edwards et al., 2022; Fang et al., 2024; Christofidellis et al., 2023; Pei et al., 2025), graph-based conditional diffusion models (Liu et al., 2024a), and multi-modal molecular pipelines integrating images, text, and graphs (Zhu et al., 2024; Kim et al., 2025; Liu et al., 2024b) have significantly broadened the range of conditioning modalities. However, these methods rarely incorporate experimental observables. Spectroscopic signals provide physically grounded, high-dimensional evidence of molecular structure, but exhibit instrument-dependent noise and distributions that challenge generic multimodal architectures (Guo et al., 2024).

Fragment-based molecular generative modeling Motif-level modeling represents a flexible inductive bias for molecular generation, operating over chemically-meaningful molecular fragments rather than individual atoms. Pioneered by methods like JT-VAE (Jin et al., 2018) and HierVAE (Jin et al., 2020), which introduced hierarchical generation tree-structured scaffold representations, the field has advanced to methods such as MARS (Xie et al., 2021), which uses GNN-guided Markov Chain Monte Carlo for iterative fragment editing toward multi-objective property optimization, and FragFM (Lee et al., 2025), which employs a coarse-to-fine autoencoder combining fragment-level graph generation with atom-level reconstruction. Despite their promise, fragment-based generative approaches have rarely been explored in conditional settings, particularly when the conditioning signal consists of experimental observables such as spectroscopic measurements.

A.2 PRE-TRAINING RESULTS

Table 3: Full performance comparison for the pre-training stage across different models. Results are reported in terms of structural similarity (Tanimoto), graph edit distance (MCES), string-level distance (Levenshtein), and Top-k accuracies. For continuous metrics, values are reported as mean $\pm$ standard deviation computed across test molecules. * indicates our implementations of baseline approaches.

MODEL	VALID (↑)	TANIMOTO (↑)			MCES (↓)	LEV. (↓)	TOP-k ACC. (↑)			
		MORGAN	MACCS	RDKIT			1	5	10	15
NMR2STRUCT*	**1.00**	$0.93_{\pm 0.15}$	$\mathbf{0.99}_{\pm 0.03}$	$0.95_{\pm 0.12}$	$0.92_{\pm 2.32}$	$1.61_{\pm 4.34}$	0.63	0.68	0.76	0.81
NMIRACLE (OURS)	0.97	$\mathbf{0.96}_{\pm 0.12}$	$\mathbf{0.99}_{\pm 0.04}$	$\mathbf{0.97}_{\pm 0.10}$	$\mathbf{0.57}_{\pm 2.32}$	$\mathbf{1.07}_{\pm 4.41}$	**0.70**	**0.73**	**0.81**	**0.86**
SPEC2MOL*	1.00	$0.99_{\pm 0.05}$	$1.00_{\pm 0.01}$	$1.00_{\pm 0.03}$	$0.07_{\pm 0.67}$	$0.12_{\pm 1.31}$	0.96	0.97	0.98	0.98

A.3 ABLATION STUDIES

We conduct ablation experiments to gain further insights on the choice of architectural components. Specifically, we examine: (i) the effect of learnable positional encodings for spectra compared to fixed sinusoidal encodings from prior work (Hu et al., 2024), and (ii) the role of the inter-modal transformer encoder for inter-spectral integration versus a simple concatenation of independently-processed spectra. As shown in Figure 4, results are reported as relative performance with respect to the full NMIRacle configuration, which serves as the reference (blue bars). For increasing metrics, relative performance is computed as $\frac{\text{Current value}}{\text{Reference value}}$, and as $\frac{\text{Reference value}}{\text{Current value}}$ for decreasing metrics. Both components provide consistent improvements, highlighting the benefits of adaptive spectral representation and explicit cross-spectra attention.

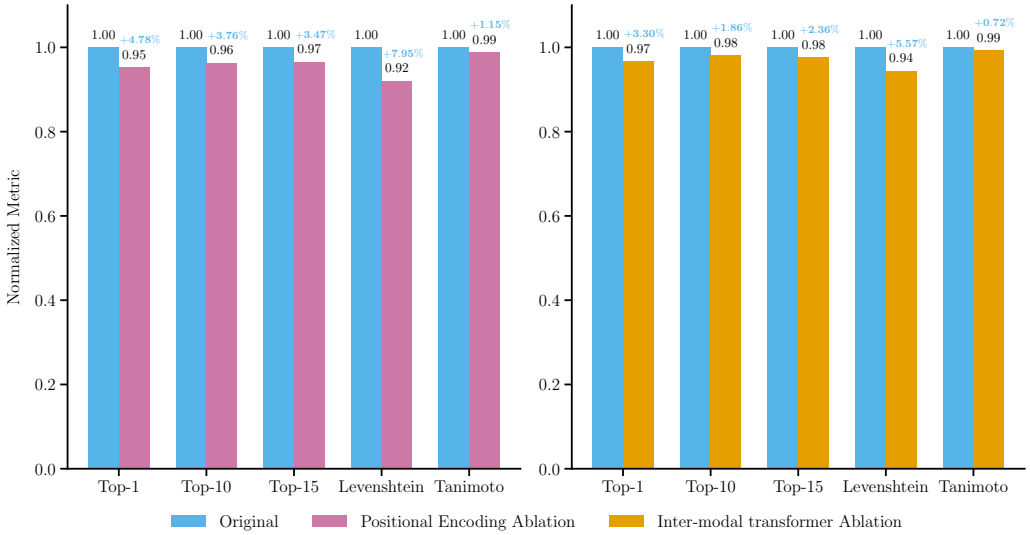

Figure 4: Ablation results comparing the impact of learnable positional encodings (left) and inter-modal transformer encoder (*right*). Bars report relative performance with respect to the full NMIRacle configuration (blue).

A.4 EVALUATION CRITERIA

In our main results we utilize an enantiomer-aware evaluation protocol. We adopt this scheme because standard IR and NMR spectra are inherently agnostic to absolute stereochemistry, thus making it chemically infeasible to demand full stereochemical resolution. As illustrated in Figure 5, this protocol considers a prediction correct if it is either an exact match or the enantiomer

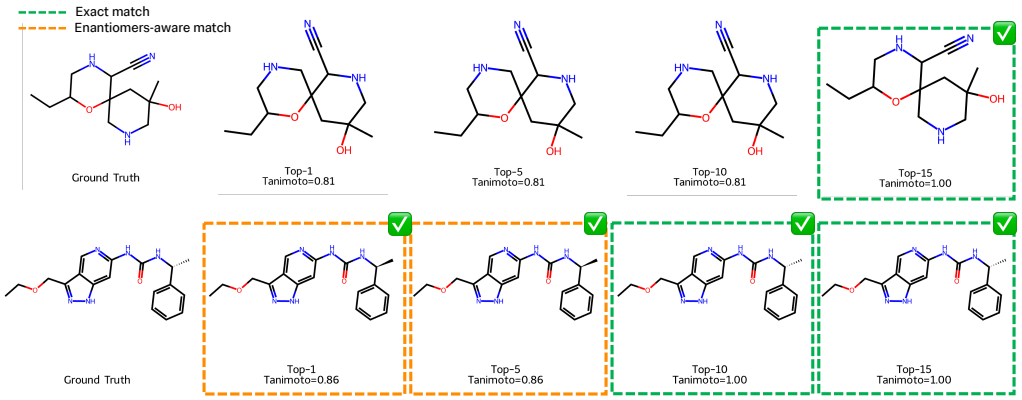

Figure 5: Illustration of the enantiomer-aware evaluation scheme. (a) If the generated and reference molecules share identical canonical SMILES, the prediction is counted as an (exact) match (green). (b) If the generated molecule represents the enantiomer of the reference (i.e., a mirror-image configuration), it is likewise treated as an (equivalent) match (orange). This criterion reflects the fact that IR and NMR spectra generally cannot resolve absolute stereochemistry.

(mirror-image configuration) of the ground truth. To analyze the distribution of mismatches across different levels of structural fidelity, we compare the model's performance under various evaluation criteria, as detailed in Figure 6. Three distinct protocols are considered: (i) *exact match* requires a perfect string-level correspondence between generated and ground-truth SMILES, including atom ordering and stereochemical specification; (ii) *enantiomer-aware* evaluation, adopted as the primary protocol in the main text, relaxes this constraint by treating enantiomeric molecules (i.e., mirror images with identical connectivity) as equivalent, reflecting the limited chirality sensitivity of IR and NMR spectra; (iii) *constitutional* criterion represents the most 'permissive' case, obtained by recomputing ground truth and generated SMILES with `RDKit` (Landrum), using `Chem.MolToSmiles(isomeric=False)` function. This operation canonicalizes molecules solely by their bonding topology, disregarding stereochemical information, and thus evaluates whether the predicted structure matches the correct constitutional framework. We observe consistent improvements under the constitutional metric, highlighting that a fraction of mismatches arise from stereochemical ambiguities rather than errors in molecular connectivity. For instance, top-15 accuracy increases from 0.66 to 0.69 for the IR + ^{1}H-NMR + ^{13}C-NMR combination, and from 0.56 to 0.59 for ^{1}H-NMR + ^{13}C-NMR. These results suggest potential practical value when the goal is structure elucidation up to constitutional isomerism, rather than full stereochemical resolution.

A.5 Metrics

We evaluate model performance through metrics that capture both generation quality and molecular similarity with respect to the corresponding ground truth. For each input spectra $\mathcal{S}$, the model produces a ranked set of k molecular candidates $\hat{\mathcal{Y}}_k = \{\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_k\}$ sampled from $p_\theta(\mathbf{y} \mid \mathcal{S})$ and ranked by their average, per-token log-likelihood under the model.

Validity Fraction of generated molecules that satisfy basic chemical constraints (e.g., valid atom valences). Validity is computed using RDKit's sanitization routines:

$$\text{Validity} = \mathbb{E}_{\mathcal{S}\sim\mathcal{D}}\left[\mathbb{1}\left\{\exists\,\hat{\mathbf{y}} \in \hat{\mathcal{Y}}_k : \hat{\mathbf{y}} \text{ is chemically valid}\right\}\right].$$

A score of 1 indicates that at least one valid molecule is generated for every spectra input.

Top-k Accuracy Measures whether the ground truth SMILES $\mathbf{y}$ appears among the top-k generated candidates for a given set of input spectra $\mathcal{S}$. This captures the model's ability to exactly recover the target structure when allowed multiple guesses:

$$\text{Top-}k\text{ Acc} = \mathbb{E}_{(\mathcal{S},\mathbf{y})\sim\mathcal{D}}\left[\mathbb{1}\Big(\exists\,\hat{\mathbf{y}} \in \hat{\mathcal{Y}}_k : \hat{\mathbf{y}} = \mathbf{y}\Big)\right].$$

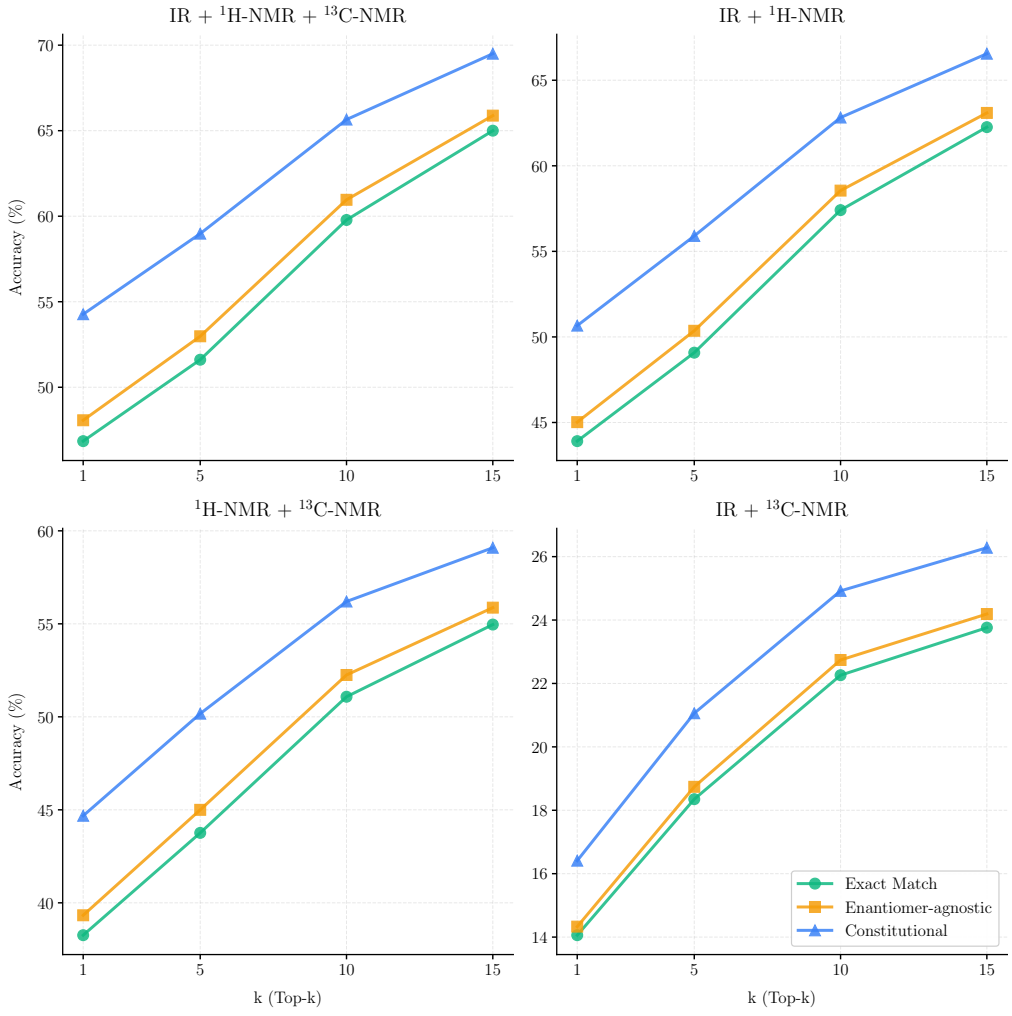

Figure 6: Comparison of molecular generation performance under different evaluation criteria, reported as top-k accuracy for various spectral combinations. Results highlight how relaxing stereochemical constraints (from exact, to enantiomer-aware, to constitutional) affects the Top-k for molecular generation.

Maximum Common Edge Subgraph (MCES) We employ the graph edit distance between a predicted molecule and the corresponding ground truth, following the implementation by Kretschmer et al. (2023). Given that we denote this distance by d_{mces}, then the corresponding metric is calculated as:

$$\text{MCES} = \mathbb{E}_{(\mathcal{S},\mathbf{y})\sim\mathcal{D}}\left[\min_{\hat{\mathbf{y}}\in\hat{\mathcal{Y}}_k} d_{\text{mces}}(\hat{\mathbf{y}},\mathbf{y})\right].$$

Levenshtein distance Quantifies the minimum number of single-character edits (insertions, deletions, substitutions) required to transform a string s into a string t. Given that we denote this distance as $d_{\text{Lev}}(s,t)$, then the corresponding metric is calculated as:

$$\text{LevDist} = \mathbb{E}_{(\mathcal{S},\mathbf{y})\sim\mathcal{D}}\left[\min_{\hat{\mathbf{y}}\in\hat{\mathcal{Y}}_k} d_{\text{Lev}}(\hat{\mathbf{y}},\mathbf{y})\right].$$

Fingerprint-based similarity To capture substructural similarity, we compute similarity scores using different molecular fingerprints. Let $f_{\text{fp}}(\mathbf{y})$ be a fingerprint vector of type fp $\in$

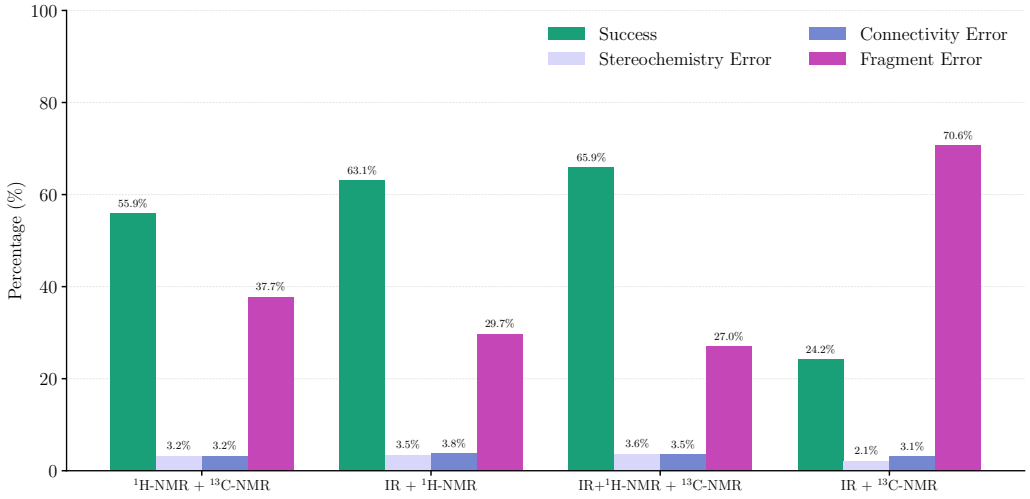

Figure 7: Analysis of model failures across error categories for different spectral combinations, alongside the corresponding success rates. Errors are dominated by incorrect fragment predictions, highlighting the need for improved fragment-level representations and tighter spectra–structure alignment.

$\{\text{Morgan}, \text{MACCS}, \text{RDKit}\}$. For each type, the similarity is:

$$\text{Sim}_{\text{fp}} = \mathbb{E}_{(\mathcal{S},\mathbf{y})\sim\mathcal{D}} \left[\max_{\hat{\mathbf{y}}\in\hat{\mathcal{Y}}_k} \text{Tanimoto}\left(f_{\text{fp}}(\hat{\mathbf{y}}), f_{\text{fp}}(\mathbf{y})\right) \right],$$

where the Tanimoto coefficient is:

$$\text{Tanimoto}(a,b) = \frac{a \cdot b}{\|a\|_1 + \|b\|_1 - a \cdot b}.$$

We utilize Morgan, MACCS, and RDKit fingerprints. All fingerprints are represented as binary vectors, where each bit indicates the presence (1) or absence (0) of a given feature. In this setting, the ℓ_1-norm $\|\cdot\|_1$ corresponds to the number of active bits in the fingerprint.

A.6 Failure modes

To better characterize the limitations of the proposed framework, we conduct a systematic analysis of failure cases, summarized in Figure 7 and grouped into three main categories: (i) *stereochemistry error* where the model correctly generates the constitutional isomer but fails to reproduce the correct stereochemical configuration of the ground truth molecule, as measured under the enantiomer-aware evaluation protocol; (ii) *connectivity error* in which all relevant functional groups are correctly predicted according to the fragment vocabulary $\mathcal{V}$, but the underlying atomic connectivity is incorrect; (iii) *fragment errors*, where the model fails to predict the correct functional groups or produces spurious fragments not present in the ground truth molecule. We observe that the dominant source of failure arises from the misprediction of fragment compositions, suggesting that the model struggles to infer the correct functional group from ambiguous or overlapping spectral evidence. This limitation may stem from the intrinsic ambiguity of spectra-to-molecule mapping or from limited granularity in the fragment vocabulary. While expanding the vocabulary could increase representational expressiveness, it may also amplify combinatorial complexity and learning difficulty. Future work will explore adaptive or hierarchical fragment vocabularies and uncertainty-aware modeling to mitigate these challenges. Moreover, we observe that recent explorations in stereochemistry-aware molecular generation (Tom et al., 2025) could help reduce the stereochemistry-related errors observed in our analysis.

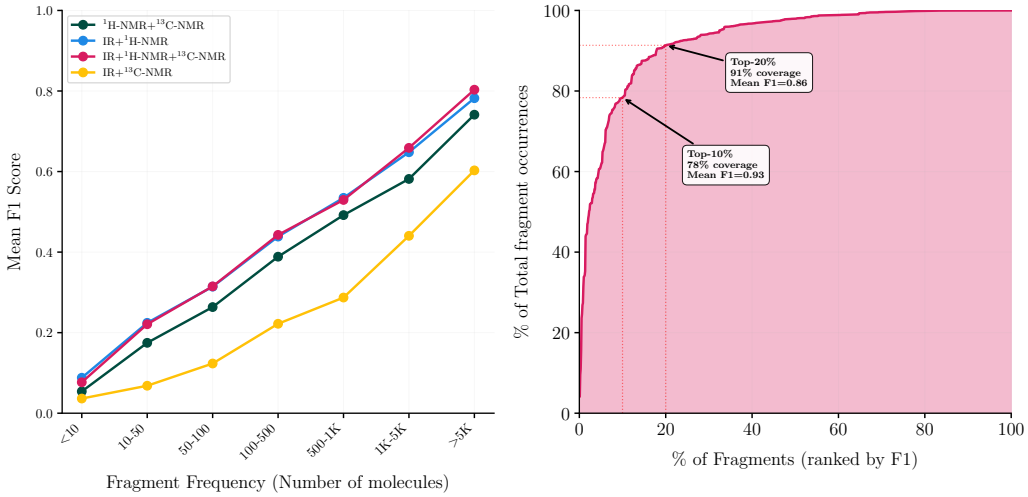

Figure 8: (*Left*) Mean F1-score across fragments grouped by molecular frequency. (*Right*) Percentage of all fragment instances in the dataset covered by the top-$X\%$ of fragments ranked by mean F1-score (IR + ^{1}H-NMR + ^{13}C-NMR), showing that high-scoring fragments account for most observed cases.

Table 4: Fragment count prediction across spectral combinations. Micro- and macro-averaged classification metrics.

	MICRO-AVERAGED				MACRO-AVERAGED			
SPECTRA	ACCURACY	PRECISION	RECALL	F1	ACCURACY	PRECISION	RECALL	F1
^{1}H + ^{13}C-NMR	0.98	0.91	0.78	0.84	0.98	0.60	0.32	0.38
IR + ^{1}H-NMR	0.98	0.90	0.83	0.86	0.98	0.60	0.37	0.43
IR + ^{1}H + ^{13}C-NMR	0.98	0.92	0.82	0.87	0.98	0.63	0.37	0.43
IR + ^{13}C-NMR	0.97	0.90	0.70	0.79	0.97	0.54	0.20	0.26

A.7 FRAGMENT OCCURRENCES PREDICTION

We evaluate fragment predictions using macro- and micro-averaged metrics (Table 4). Macro metrics weight all fragments equally, while micro metrics aggregate over all predictions and therefore reflect fragment frequency. We report two complementary criteria: count accuracy, which measures how well the model estimates fragment occurrences, and presence-based metrics (precision, recall, F1), which assess detection irrespective of count. All models achieve high count accuracy (≈ 0.97–0.98), though this metric is dominated by absent fragments ($c_j = 0$) and is less informative under strong class imbalance. Presence-based metrics provide a more discriminative assessment. The full multi-spectral model (IR + ^{1}H-NMR + ^{13}C-NMR) achieves the strongest performance, with a micro-averaged F1 of 0.87 (precision 0.92, recall 0.82), indicating that complementary spectral modalities substantially improve fragment identification. In contrast, macro-averaged F1 scores are notably lower (e.g., 0.43 for the full model), reflecting the difficulty of predicting rare fragments. Figure 8 further illustrates these trends. Mean fragment-level F1 increases with fragment prevalence, indicating that more common fragments are predicted more accurately. Complementarily, a cumulative coverage analysis that ranks fragments by their per-fragment F1 shows that the top 10% of fragments account for approximately 78% of all fragment occurrences. This demonstrates that the fragments predicted most reliably by the model also correspond to the most prevalent motifs in the dataset. Consequently, strong performance is concentrated on chemically common fragments, while lower macro-averaged scores primarily reflect the intrinsic difficulty of accurately predicting rare fragments under severe class imbalance. Addressing this limitation will likely require targeted data curation or rebalancing strategies to improve generalization across the full fragment vocabulary.

A.8 OUTLOOK AND FUTURE DIRECTIONS

This section discusses the primary limitations encountered within the proposed framework and situates them within broader challenges in spectra-to-molecule learning, thereby outlining key directions for future research.

Scale of molecular pre-training The fragment-to-molecule pre-training currently leverages a corpus of $\sim$ 3.7M SMILES, which offers a solid foundation but remains modest compared to the scale of contemporary chemical databases such as PubChem ($\sim$100M molecules). Given the observed scaling trends in related molecular generative frameworks (Bohde et al., 2025), extending pre-training to larger, more chemically diverse datasets could further enrich the learned molecular prior p_ϕ, enhancing both fragment composition modeling and downstream spectral elucidation. Future work will investigate large-scale, fragment-conditioned pre-training to assess potential performance gains and improved generalization to rare or complex structures.

2D vs 3D molecular representations In this work, we focus on the generation of 2D molecular structures, i.e., atomic connectivity and bond types, rather than explicit 3D conformations. While recent approaches aim to jointly infer 2D and 3D structures from spectra (Wang et al., 2025a), 2D representations remain a central abstraction for molecular structure elucidation. In particular, 2D graphs capture the core chemical composition and topology of molecules, are directly comparable to existing benchmark datasets, and allow for efficient and scalable generative modeling. Moreover, many spectroscopic modalities provide strong constraints on functional groups and connectivity, even when 3D conformations are ambiguous or undetermined. Incorporating explicit 3D geometry and conformational modeling is an important direction for future work, but lies outside the scope of the present study.

Choice of the prior p_ϕ In this work, the molecular prior p_ϕ is learned by reconstructing molecular structures from a coarse, fragment-based representation that encodes both fragment identities and their occurrences. This extends previous binary fragment formulations (Bohde et al., 2025; Hu et al., 2024) by introducing a count-aware model that better reflects the underlying molecular topology. However, the question of what constitutes an *optimal* prior for downstream spectra-to-molecule task remains open. An interesting research direction is to study how the design of p_ϕ (e.g. by incorporating additional relational structure such as fragment connectivity, local bonding patterns, or hierarchical composition) affects the learnability and transferability of this mapping. In other words, future work should explore priors that are not only chemically-faithful but also *spectroscopically-aligned*, facilitating transferability to the spectra-to-molecule stage.

Simulated vs experimental spectra All results in this work are based on simulated spectra generated from computational pipelines that approximate experimental conditions. While this provides consistent supervision, real-world spectra are subject to noise, baseline distortions, solvent effects, and instrument-specific artifacts that may introduce significant distribution shifts. Adapting the model to such data will require domain adaptation strategies or fine-tuning on curated experimental datasets to ensure robustness under practical laboratory conditions.

Approximation of fragment inference In Eq. 2, the mapping from spectra $\mathcal{S}$ to fragment composition $\mathbf{c}$ is treated deterministically via $\mathbf{z}_\psi(\mathcal{S})$. This neglects inherent ambiguity in the inverse mapping from spectra to substructures (multiple molecular configurations may correspond to highly similar spectral signatures). Future work could relax this assumption by introducing stochastic or variational inference, thereby capturing uncertainty over fragment compositions and improving robustness to ambiguous input spectra.

A.9 IMPLEMENTATION DETAILS

Pre-training All models requiring a pre-training stage (NMIRacle, NMR2Struct, and Spec2Mol) are trained using a batch size of 1024 and a weight decay of 1×10^{-5}. For NMIRacle and NMR2Struct, we use a learning rate of 1×10^{-5} and $\beta = (0.9, 0.98)$. For Spec2Mol, we follow the original implementation settings with a learning rate of 1×10^{-4}. While NMIRacle and

NMR2Struct are pre-trained on fragments-to-molecule generation, Spec2Mol utilizes a SMILES reconstruction (autoencoding) task.

Spectra-to-molecule fine-tuning For the downstream task, models are trained for a maximum of 300 epochs using a batch size of 64 and the same learning rates as in the pre-training stage (with a learning rate of 1×10^{-5} used for the Transformer baselines not leveraging a pre-training stage). We utilize early stopping with a patience of 10 epochs based on the validation loss specific to each architecture:

- **NMIRacle (Ours) and NMR2Struct:** The multi-task objective combining SMILES reconstruction with fragment-level supervision. Crucially, we monitor the *fragment count* loss for NMIRacle, compared to the *binary presence* classification for NMR2Struct.
- **Spec2Mol:** The weighted combination of ℓ_2 latent alignment loss and SMILES reconstruction loss.
- **SMILES/SELFIES Transformers:** The standard sequence cross-entropy loss.

For generation, all models utilize top-k sampling with $k = 5$ and temperature $T = 1.0$.

A.10 ARCHITECTURE DETAILS

In Table 5 we present a summary of the main architectural components of NMIRacle.

Table 5: NMiracle Model Architectural Details

Component	Parameter	Value
Global	Model Dimension (d)	128
Fragment Encoder	Vocabulary Size	991
	Embedding Dimension	128
	Activation	GELU
Transformer Model (p_ϕ)	Encoder Layers	6
	Decoder Layers	6
	Attention Heads	8
	FFN Dimension	1024
	Dropout	0.1
	Activation	ReLU
Multispectra Encoder (q_ψ)	Kernel Size 1	5
	Pool Size 1	12
	Out Channels 1	64
	Kernel Size 2	9
	Pool Size 2	20
	Out Channels 2	128
	Transformer encoder layers	2
	Attention Heads	4
	Activation	ReLU
	^{13}C-NMR binary bins	80
Fragment composition head	Hidden dimension	256
	Activation	GELU

A.11 SUCCESS CASES

Figure 9 illustrates representative success cases under full multi-spectral conditioning (IR, ^{1}H-NMR, ^{13}C-NMR), where NMIRacle correctly recovers the target molecular structure, while all competing methods fail under the same enantiomer-aware evaluation protocol.

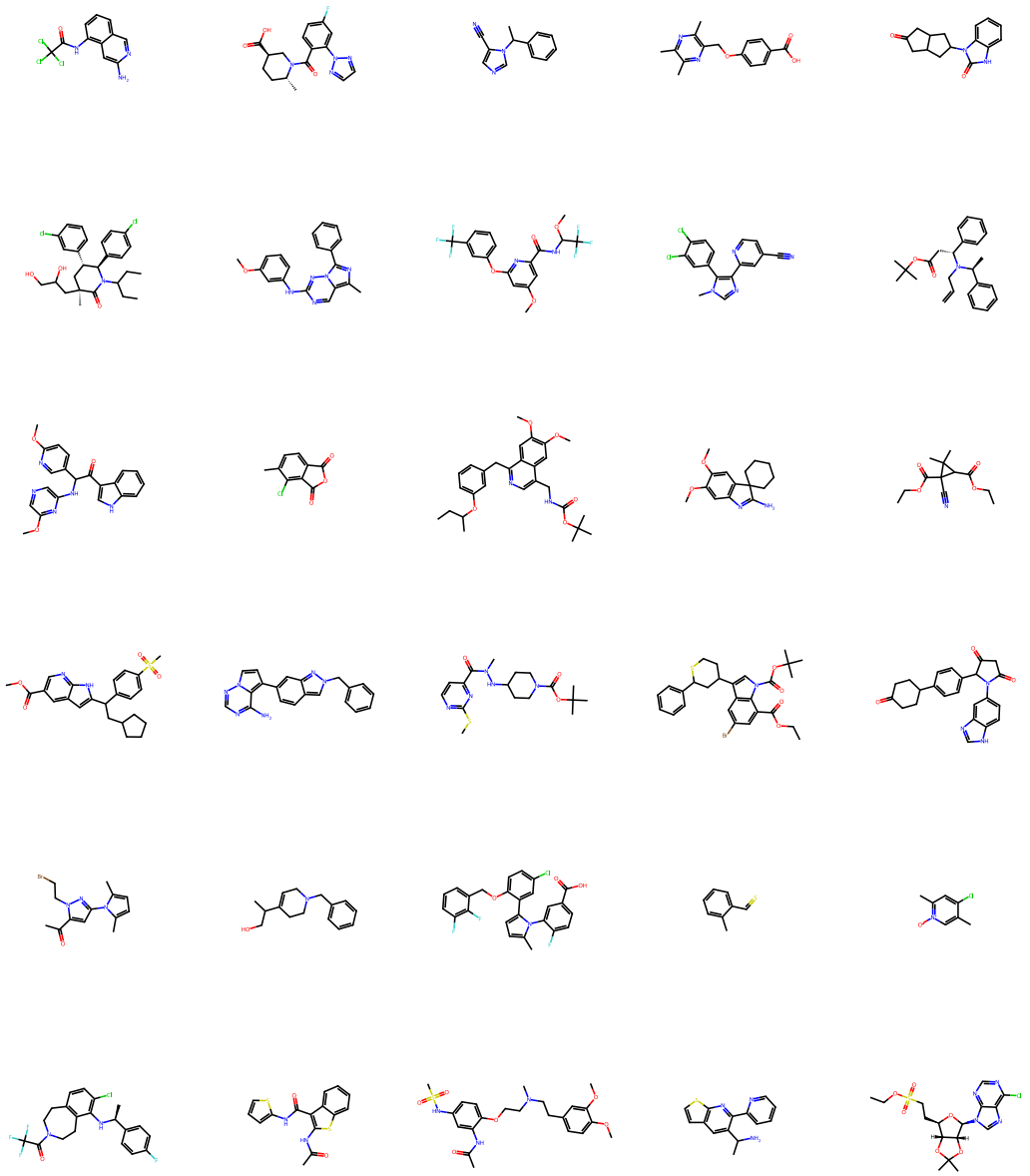

Figure 9: Representative success cases in which NMIRacle correctly predicts the ground-truth molecule, while all baseline models fail. For each example, the prediction from NMIRacle matches the ground truth structure under the enantiomer-aware evaluation protocol, whereas competing methods do not recover the correct molecular structure among their top-k candidates.