Supplementary Information for "NMRPeak: a ready-to-use intelligent system for molecular structure elucidation enabled by synergistic cross-modal learning"

Fanjie Xu[1,2,3], Jinyuan Hu[2,4], Jingxiang Zou[4], Junjie Wang[3,5], Boying Huang[6], Zhifeng Gao[3,7], Xiaohong Ji[3,7*], Weinan E[7,8,9], Zhong-Qun Tian[2,4], Fujie Tang[6,1,4*], Jun Cheng[2,1,4*]

[1*]Institute of Artificial Intelligence, Xiamen University, Xiamen, 361005, China.
[2]State Key Laboratory of Physical Chemistry of Solid Surfaces, iChEM, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, 361005, China.
[3]DP Technology, Beijing, 100080, China.
[4]Laboratory of AI for Electrochemistry (AI4EC), Tan Kah Kee Innovation Laboratory (IKKEM), Xiamen, 361005, China.
[5]College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China.
[6]Pen-Tung Sah Institute of Micro-Nano Science and Technology, Discipline of Intelligent Instrument and Equipment, iChEM, Xiamen University, Xiamen, 361005, China.
[7]AI for Science Institute, Beijing, 100080, China.
[8]Center for Machine Learning Research, Peking University, Beijing, 100871, China.
[9]School of Mathematical Sciences, Peking University, Beijing, 100871, China.

*Corresponding author(s). E-mail(s): jixh@dp.tech; tangfujie@xmu.edu.cn; chengjun@xmu.edu.cn;

Contents

Supplementary Notes

Supplementary Note 1. Configuration of the unified token vocabulary.

To facilitate synergistic cross-modal learning, we developed a unified vocabulary for the BART-based [1] architecture, encompassing spectral features, molecular formulas, and structural representations. The total vocabulary for the BART modules consists of **2,954** distinct tokens, categorized as follows:

- **Numerical Tokens (2,380)**: These tokens represent discretized ^{13}C and ^{1}H chemical shifts, as well as ^{1}H coupling constants. The detailed interval partitioning and step sizes, including boundary tokens for out-of-range values, are provided in **Supplementary Table 3**.
- **Categorical Tokens (409)**: This set includes 140 ^{1}H NMR multiplicity patterns, 50 proton integration levels, 118 element types, 99 atom counts, and 2 specific charge state tokens. For a complete list of categorical definitions, refer to **Supplementary Table 4**.
- **SMILES Tokens (158)**: These are generated using the Smirk-based [2] tokenization method.
- **Special Tokens (7)**: These include standard linguistic markers (`[CLS]`, `[PAD]`, `[SEP]`, `[UNK]`, `[MASK]`) and modality-specific boundary tokens (`cnmr_end`, `hnmr_end`) used to delineate spectral segments and maintain sequence order during training and inference.

In contrast, the **Uni-Mol** [3] molecule encoder, which is utilized to understand 3D conformational information, maintains its original specialized vocabulary of **31** tokens.

Supplementary Note 2. Implementation of NMRPeak-P model variants.

To ensure the robustness and accuracy of forward spectrum prediction, we implemented two variants of the prediction module: **NMRPeak-P-Single** and **NMRPeak-P-Multi**. The configurations for their model selection and ensemble strategy are as follows:

- **NMRPeak-P-Single**: This baseline variant utilizes the single best-performing checkpoint identified during the training process. The optimal model is selected by maximizing the `token_acc` metric on the validation set of the curated experimental NMRexp [4] dataset.
- **NMRPeak-P-Multi**: This ensemble variant leverages a multi-model consensus approach to mitigate prediction variance and capture a broader distribution of spectral signatures. Specifically, we utilize ten independent checkpoints saved at regular intervals during the stable convergence phase of training, corresponding to steps 200,000, 250,000, 300,000, 400,000, 500,000, 600,000, 700,000, 800,000, 900,000, and 1,000,000.
- **Ensemble Strategy**: For each input molecular structure, the ten models in the ensemble independently generate candidate spectral token sequences. These candidates are subsequently aggregated and re-ranked using the molecule-to-spectrum ranking function of the NMRPeak-R module.

Supplementary Note 3. Token representation of ^{1}H NMR spectral peaks.

To accurately capture the spectral profile of ^{1}H NMR signals, which often appear as multiplet ranges rather than single sharp resonances, we utilize a dual-token representation strategy for each peak. For experimental peaks reported with defined boundaries, the tokenizer converts the maximum chemical shift and the minimum chemical shift into two corresponding numerical tokens. These tokens are appended sequentially to the peak's token set, preserving the inherent width of the multiplet signal. In scenarios where specific range boundaries are unavailable, the system fallbacks to the peak's centroid chemical shift. To maintain a consistent sequence length and structural representation across all ^{1}H signals, the centroid token is duplicated and appended twice. This ensures that the downstream transformer architecture processes a uniform input dimensionality regardless of the raw data format.

Supplementary Note 4. Model size configuration and computational cost.

To investigate the impact of model scale on performance, we additionally evaluated larger model configurations for the NMRPeak system modules, including NMRPeak-P, NMRPeak-R, and NMRPeak-G. The detailed hyperparameter settings of the base configurations used in this work are summarized in Supplementary Table 5.

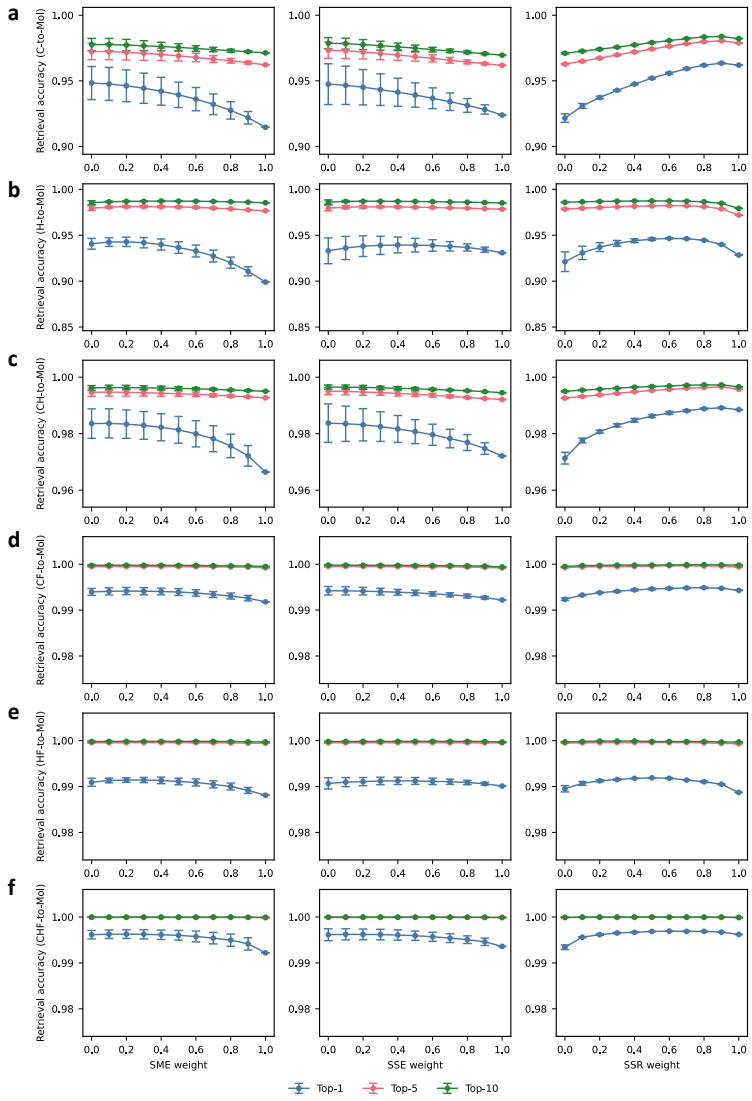
Empirical results indicate that large-scale configurations consistently yield improved performance compared to their base counterparts, primarily due to increased model capacity and representational power. However, these performance gains come at the cost of substantially higher computational overhead, including longer training time, increased GPU memory consumption, and reduced training efficiency.

Considering the trade-off between performance and computational cost, we adopt the base configurations for all experiments reported in this work. This choice ensures competitive performance while maintaining reasonable training time and resource requirements.

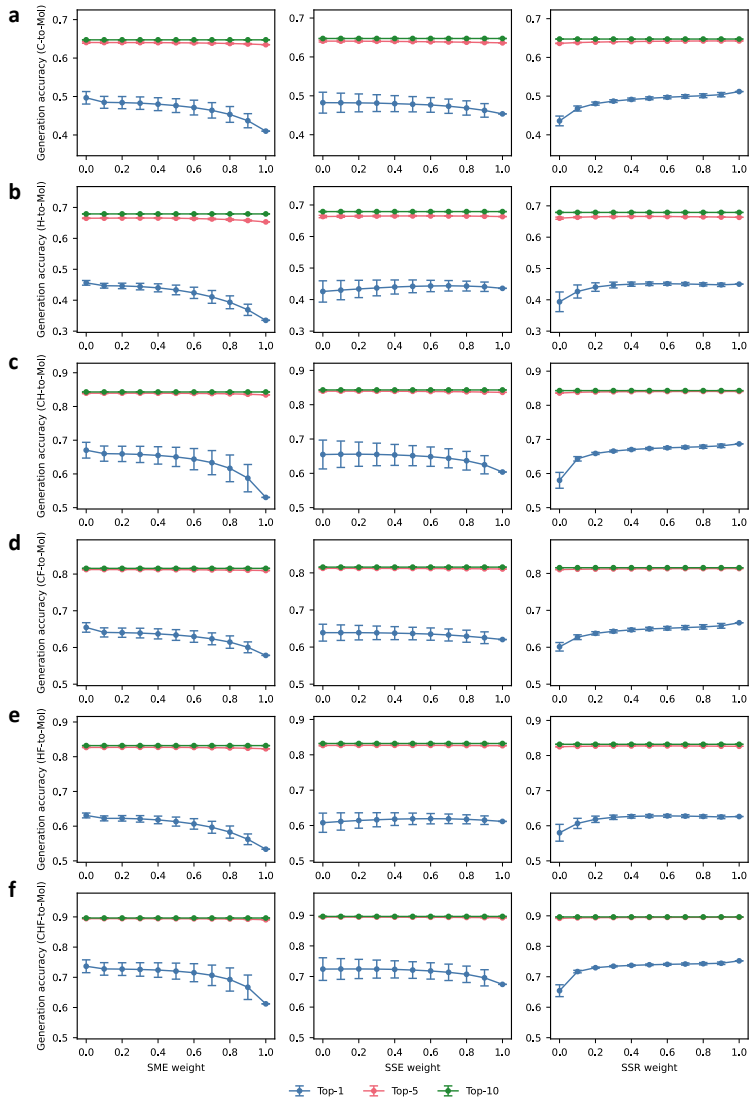
Supplementary Note 5. Effect of beam size on generation performance.

To assess the influence of decoding strategy on molecular generation, we evaluated NMRPeak-G with different beam sizes. The detailed stereochemistry-aware generation accuracies for beam size = 10 and beam size = 100 are summarized in Supplementary Table 10 and Supplementary Table 15, respectively. We observe that increasing the beam size generally improves Top-1 accuracy, as a larger beam allows the decoder to explore a broader candidate space and increases the likelihood that the correct structure appears among the highest-ranked candidates. In particular, a larger beam size enhances the probability that correct molecules are promoted to earlier candidate positions during decoding. However, the overall performance differences between beam size = 10 and beam size = 100 remain relatively small. This suggests that the generation model already captures strong structural priors, and that moderate beam sizes are sufficient to achieve competitive performance. Considering the trade-off between decoding efficiency and marginal accuracy gains, beam size = 10 provides a practical balance between computational cost and generation quality.

Supplementary Figures



Supplementary Figure. 1. Contribution of individual similarity metrics of NMRPeak-R. a–f, Impact of weight variations for each component on retrieval accuracy across diverse configurations, with data points and error bars representing the mean and variance, respectively. For all panels, SME, SSE, and SSR denote spectrum-to-molecule embedding, spectrum-to-spectrum embedding, and spectrum-to-spectrum rule-based similarity, respectively. C and H correspond to ^{13}C and ^{1}H NMR peaks, and F denotes molecular formula constraints.



Supplementary Figure. 2. Contribution of individual similarity metrics of NMRPeak-G. a–f, Impact of weight variations for each component on generation accuracy across diverse configurations, with data points and error bars representing the mean and variance, respectively. For all panels, SME, SSE, and SSR denote spectrum-to-molecule embedding, spectrum-to-spectrum embedding, and spectrum-to-spectrum rule-based similarity, respectively. C and H correspond to ^{13}C and ^{1}H NMR peaks, and F denotes molecular formula constraints.

Supplementary Tables

Supplementary Table 1. Data statistics of the MST-NMR [5] dataset across training, validation, and test splits.

Split	All samples	^{13}C	^{1}H	^{13}C & ^{1}H
Train	679,186	679,186	679,186	679,186
Valid	35,747	35,747	35,747	35,747
Test	79,437	79,437	79,437	79,437
Full dataset	794,370	794,370	794,370	794,370

Supplementary Table 2. Data statistics of the NMRexp [4] dataset across training, validation, and test splits.

Split	All samples	^{13}C	^{1}H	^{13}C & ^{1}H
Train	920,796	786,240	805,224	670,668
Valid	48,463	41,383	42,366	35,286
Test	107,696	91,830	94,163	78,297
Full dataset	1,076,955	919,453	941,753	784,251

Supplementary Table 3. Discretization strategy for numerical tokens in the NMRPeak tokenizer. Different regions and step sizes are defined for ^{13}C chemical shift, ^{1}H chemical shift, and ^{1}H coupling constants.

Type	Region	Step size	Tokens
^{13}C chemical shift (ppm)			
	$(-\infty, -6)$	-	1
	$(-6, 0)$	0.5	12
	$(0, 200)$	0.1	2000
	$(200, 210)$	0.5	20
	$(210, 220)$	1.0	10
	$(220, 250)$	30.0	1
	$(250, \infty)$	-	1
	Total ^{13}C chemical shift tokens	-	2045
^{1}H chemical shift (ppm)			
	$(-\infty, -1)$	-	1
	$(-1, 0)$	0.1	10
	$(0, 12)$	0.05	240
	$(12, 13)$	0.1	10
	$(13, 14)$	0.2	5
	$(14, 16)$	0.5	4
	$(16, \infty)$	-	1
	Total ^{1}H chemical shift tokens	-	271
^{1}H coupling constant (Hz)			
	$(0, 20)$	0.5	40
	$(20, 30)$	1.0	10
	$(30, 50)$	2.0	10
	$(50, 60)$	5.0	2
	$(60, 300)$	240.0	1
	$(300, \infty)$	-	1
	Total ^{1}H coupling constant tokens	-	64
Total numerical tokens			**2380**

Supplementary Table 4. Discretization strategy for categorical tokens in the NMRPeak tokenizer. These tokens represent discrete chemical attributes and symbolic information, now including molecular charge states.

Category	Definition / Range	Tokens
^{1}H NMR Peak Attributes		
Multiplicity	140 predefined splitting patterns (s, d, t, m, etc.)	140
Integration	Integer integration values (proton counts) in range [1, 50]	50
Molecular Formula Attributes		
Element Type	Atomic numbers for element identification in range [1, 118]	118
Atom Count	Number of atoms per element in range [2, 100]	99
Charge State	Positive and negative	2
Total Categorical Tokens	(Excluding Special/Boundary Tokens)	**409**

Supplementary Table 5. Hyperparameter configurations for the NMRPeak system modules (NMRPeak-P, NMRPeak-R, and NMRPeak-G).

Hyperparameter	NMRPeak-P	NMRPeak-R	NMRPeak-G
Architecture	Uni-Mol [3] encoder[a] + Bart [1] decoder[b]	Uni-Mol [3] encoder[a] + Bart [1] encoder[b]	Bart [1] encoder[b] + Bart [1] decoder[b]
Number of GPUs	8	1	8
Learning rate	3×10^{-4}	3×10^{-4}	3×10^{-4}
Batch size	16	512	16
Contrastive temperature	-	0.1	-
Warmup steps	60,000	15,000	60,000
Total steps	1,000,000	250,000	1,000,000
Dropout	0	0	0
Update frequency	1	1	1

[a] Load pretrained weights into the base architecture and unfreeze for training;
[b] Base architecture, trained from scratch.

Supplementary Table 6. Comparison of peak-aware similarity scores for NMRPeak-P-Single and NMRPeak-P-Multi on the NMRexp [4] test set.

Nucleus Type	NMRPeak-P-Single	NMRPeak-P-Multi
^{13}C	0.9384	**0.9391**
^{1}H	0.9314	**0.9321**

Supplementary Table 7. Top-k molecular generation accuracy using ground-truth spectra (NMRexp), NMRPeak-P-Single, and NMRPeak-P-Multi as inputs to the frozen NMRPeak-G model.

Spectral Type	Top-k	NMRexp [4]	NMRPeak-P-Single	NMRPeak-P-Multi
C	Top-1	0.4227	0.4535	**0.4620**
	Top-5	0.6116	0.6514	**0.6620**
	Top-10	0.6475	0.6874	**0.6974**
CF	Top-1	0.5893	0.6197	**0.6256**
	Top-5	0.7898	0.8216	**0.8293**
	Top-10	0.8155	0.8454	**0.8526**
H	Top-1	0.4373	0.4909	**0.4983**
	Top-5	0.6361	0.7034	**0.7139**
	Top-10	0.6790	0.7439	**0.7541**
HF	Top-1	0.6109	0.6600	**0.6662**
	Top-5	0.8052	0.8501	**0.8570**
	Top-10	0.8320	0.8737	**0.8792**
CH	Top-1	0.6384	0.6801	**0.6888**
	Top-5	0.8177	0.8593	**0.8691**
	Top-10	0.8430	0.8820	**0.8914**
CHF	Top-1	0.7129	0.7473	**0.7542**
	Top-5	0.8779	0.9067	**0.9112**
	Top-10	0.8963	0.9222	**0.9261**

C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.

Supplementary Table 8. Molecule-to-spectrum (M2S) retrieval accuracy of NMRPeak-R on the NMRexp [4] test set.

Spectral Type	Variant	Top-1	Top-5	Top-10
C	NMRPeak-R-M2S	0.9116	0.9617	0.9703
H	NMRPeak-R-M2S	0.8981	0.9764	0.9854

C and H correspond to ^{13}C and ^{1}H NMR spectra, respectively.

Supplementary Table 9. Spectrum-to-molecule (S2M) retrieval accuracy of different scoring variants in NMRPeak-R.

Spectral Type	Variant	Top-1	Top-5	Top-10
C	NMRPeak-R-S2M-SME	0.9146	0.9622	0.9713
	NMRPeak-R-S2M-SSE	0.9239	0.9618	0.9696
	NMRPeak-R-S2M-SSR	0.9620	0.9789	0.9822
	NMRPeak-R-S2M-Combine (0.1,0.0,0.9)	**0.9640**	**0.9809**	**0.9841**
CF	NMRPeak-R-S2M-SME	0.9918	0.9992	0.9995
	NMRPeak-R-S2M-SSE	0.9922	0.9992	0.9994
	NMRPeak-R-S2M-SSR	0.9943	0.9994	0.9998
	NMRPeak-R-S2M-Combine (0.2,0.0,0.8)	**0.9950**	**0.9996**	**0.9999**
H	NMRPeak-R-S2M-SME	0.8990	0.9766	0.9854
	NMRPeak-R-S2M-SSE	0.9309	0.9784	0.9852
	NMRPeak-R-S2M-SSR	0.9284	0.9720	0.9792
	NMRPeak-R-S2M-Combine (0.2,0.2,0.6)	**0.9475**	**0.9824**	**0.9875**
HF	NMRPeak-R-S2M-SME	0.9881	0.9994	0.9997
	NMRPeak-R-S2M-SSE	0.9901	0.9995	0.9997
	NMRPeak-R-S2M-SSR	0.9887	0.9993	0.9997
	NMRPeak-R-S2M-Combine (0.4,0.1,0.5)	**0.9920**	**0.9996**	**0.9998**
CH	NMRPeak-R-S2M-SME	0.9664	0.9927	0.9950
	NMRPeak-R-S2M-SSE	0.9721	0.9921	0.9945
	NMRPeak-R-S2M-SSR	0.9885	0.9957	0.9967
	NMRPeak-R-S2M-Combine (0.1,0.0,0.9; $\lambda = 0.5$)	**0.9894**	**0.9966**	**0.9973**
CHF	NMRPeak-R-S2M-SME	0.9922	0.9999	0.9999
	NMRPeak-R-S2M-SSE	0.9936	0.9999	0.9999
	NMRPeak-R-S2M-SSR	0.9968	0.9999	1.0000
	NMRPeak-R-S2M-Combine (0.0,0.1,0.9; $\lambda = 0.6$)	**0.9970**	**1.0000**	**1.0000**

C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.
Combine denotes the best-performing linear combination of SME, SSE, and SSR with coefficients (α, β, γ). For CH and CHF inputs, λ denotes the weighting factor between ^{13}C and ^{1}H modalities in SSR. For single-modality inputs, $\lambda = 0$.

Supplementary Table 10. Generation accuracy of NMRPeak-G with beam size = 100 on the NMRexp [4] test set. Stereochemistry-aware accuracy (Isomer = True) is compared against topology-only accuracy (Isomer = False) from top-1 to top-100 candidates.

Modality	Setting	Top-1	Top-5	Top-10	Top-20	Top-50	Top-100
C	Isomer=True	0.4240	0.6166	0.6621	0.6981	0.7370	0.7618
	Isomer=False	0.4849	0.6368	0.6791	0.7132	0.7505	0.7741
CF	Isomer=True	0.5899	0.7951	0.8330	0.8607	0.8863	0.8971
	Isomer=False	0.6667	0.8178	0.8513	0.8763	0.9003	0.9103
H	Isomer=True	0.4379	0.6402	0.6908	0.7280	0.7688	0.7939
	Isomer=False	0.4960	0.6639	0.7097	0.7448	0.7832	0.8074
HF	Isomer=True	0.6111	0.8093	0.8465	0.8719	0.8945	0.9045
	Isomer=False	0.6866	0.8351	0.8671	0.8899	0.9103	0.9194
CH	Isomer=True	0.6389	0.8199	0.8504	0.8729	0.8939	0.9043
	Isomer=False	0.7152	0.8444	0.8694	0.8889	0.9075	0.9170
CHF	Isomer=True	0.7129	0.8798	0.9039	0.9200	0.9339	0.9395
	Isomer=False	0.7955	0.9057	0.9246	0.9375	0.9491	0.9537

C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.

Supplementary Table 11. Generation accuracy of NMRPeak-G with beam size = 10, trained on MST-NMR [5] dataset and evaluated on the MST-NMR test set.

Modality	Variant	Top-1	Top-5	Top-10
C	MST [5] (baseline)	0.3637	0.5226	0.5594
	NMRPeak-G-Raw	**0.5123**	**0.6817**	**0.7114**
CF	MST [5] (baseline)	0.4869	0.6656	0.7002
	NMRPeak-G-Raw	**0.6348**	**0.8060**	**0.8276**
H	MST [5] (baseline)	0.5078	0.6696	0.7020
	NMRPeak-G-Raw	**0.5879**	**0.7580**	**0.7882**
HF	MST [5] (baseline)	0.6322	0.7900	0.8163
	NMRPeak-G-Raw	**0.7265**	**0.8748**	**0.8905**
CH	MST [5] (baseline)	0.6474	0.8021	0.8246
	NMRPeak-G-Raw	**0.7515**	**0.8898**	**0.9068**
CHF	MST [5] (baseline)	0.6977	0.8445	0.8652
	NMRPeak-G-Raw	**0.7965**	**0.9209**	**0.9325**

MST [5] (baseline) refers to results obtained by training the model from scratch using the official source code released by the authors.
Only stereochemistry-aware accuracy (Isomer = True) is reported. Topology-only results are omitted for clarity but follow similar trends.
C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.
Raw denotes generation results without re-ranking.

Supplementary Table 12. Generation accuracy of NMRPeak-G with beam size = 10, trained on MST-NMR [5] dataset and evaluated on the NMRexp test set.

Modality	Variant	Top-1	Top-5	Top-10
C	MST [5] (baseline)	0.0168	0.0328	0.0402
	NMRPeak-G-Raw	**0.0283**	**0.0527**	**0.0624**
CF	MST [5] (baseline)	0.0641	0.1194	0.1379
	NMRPeak-G-Raw	**0.1002**	**0.1768**	**0.1977**
H	MST [5] (baseline)	0.0036	0.0080	0.0109
	NMRPeak-G-Raw	**0.0065**	**0.0141**	**0.0183**
HF	MST [5] (baseline)	0.0343	0.0665	0.0784
	NMRPeak-G-Raw	**0.0533**	**0.1008**	**0.1155**
CH	MST [5] (baseline)	0.0307	0.0572	0.0678
	NMRPeak-G-Raw	**0.0515**	**0.0931**	**0.1098**
CHF	MST [5] (baseline)	0.0757	0.1336	0.1527
	NMRPeak-G-Raw	**0.1227**	**0.2097**	**0.2306**

MST [5] (baseline) refers to results obtained by training the model from scratch using the official source code released by the authors.
Only stereochemistry-aware accuracy (Isomer = True) is reported. Topology-only results are omitted for clarity but follow similar trends.
C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.
Raw denotes generation results without re-ranking.

Supplementary Table 13. Generation accuracy of NMRPeak-G with beam size = 10, trained on NMRexp [4] dataset and evaluated on the NMRexp test set

Modality	Variant	Top-1	Top-5	Top-10
C	MST [5] (baseline)	0.3042	0.4659	0.5029
	NMRPeak-G-Raw	0.4227	0.6116	0.6474
	NMRPeak-G-Rerank-Base	0.4227	0.6128	0.6474
	NMRPeak-G-Rerank-SME	0.4100	0.6348	0.6474
	NMRPeak-G-Rerank-SSE	0.4536	0.6362	0.6474
	NMRPeak-G-Rerank-SSR	0.5117	0.6425	0.6474
	NMRPeak-G-Rerank-Combine (0.0,0.0,1.0)	**0.5117**	**0.6425**	**0.6474**
CF	MST [5] (baseline)	0.4645	0.6631	0.7001
	NMRPeak-G-Raw	0.5894	0.7898	0.8155
	NMRPeak-G-Rerank-Base	0.6044	0.8015	0.8155
	NMRPeak-G-Rerank-SME	0.5784	0.8096	0.8155
	NMRPeak-G-Rerank-SSE	0.6203	0.8105	0.8155
	NMRPeak-G-Rerank-SSR	0.6664	0.8131	0.8155
	NMRPeak-G-Rerank-Combine (0.0,0.0,1.0)	**0.6664**	**0.8131**	**0.8155**
H	MST [5] (baseline)	0.3299	0.5000	0.5406
	NMRPeak-G-Raw	0.4373	0.6361	0.6790
	NMRPeak-G-Rerank-Base	0.4373	0.6375	0.6790
	NMRPeak-G-Rerank-SME	0.3349	0.6532	0.6790
	NMRPeak-G-Rerank-SSE	0.4356	0.6632	0.6790
	NMRPeak-G-Rerank-SSR	0.4500	0.6634	0.6790
	NMRPeak-G-Rerank-Combine (0.0,0.7,0.3)	**0.4626**	**0.6658**	**0.6790**
HF	MST [5] (baseline)	0.5130	0.7078	0.7422
	NMRPeak-G-Raw	0.6109	0.8052	0.8320
	NMRPeak-G-Rerank-Base	0.6264	0.8190	0.8320
	NMRPeak-G-Rerank-SME	0.5337	0.8229	0.8320
	NMRPeak-G-Rerank-SSE	0.6116	0.8258	0.8320
	NMRPeak-G-Rerank-SSR	0.6263	0.8263	0.8320
	NMRPeak-G-Rerank-Combine (0.0,0.5,0.5)	**0.6365**	**0.8271**	**0.8320**
CH	MST [5] (baseline)	0.5222	0.7034	0.7351
	NMRPeak-G-Raw	0.6384	0.8177	0.8430
	NMRPeak-G-Rerank-Base	0.6385	0.8202	0.8430
	NMRPeak-G-Rerank-SME	0.5304	0.8343	0.8430
	NMRPeak-G-Rerank-SSE	0.6038	0.8360	0.8430
	NMRPeak-G-Rerank-SSR	0.6868	0.8404	0.8430
	NMRPeak-G-Rerank-Combine (0.0,0.0,1.0; $\lambda = 0.7$)	**0.6868**	**0.8404**	**0.8430**
CHF	MST [5] (baseline)	0.6093	0.7903	0.8164
	NMRPeak-G-Raw	0.7129	0.8779	0.8963
	NMRPeak-G-Rerank-Base	0.7272	0.8890	0.8963
	NMRPeak-G-Rerank-SME	0.6121	0.8906	0.8963
	NMRPeak-G-Rerank-SSE	0.6743	0.8919	0.8963
	NMRPeak-G-Rerank-SSR	0.7522	0.8949	0.8963
	NMRPeak-G-Rerank-Combine (0.0,0.0,1.0; $\lambda = 1.0$)	**0.7522**	**0.8949**	**0.8963**

MST [5] (baseline) refers to results obtained by training the model from scratch using the official source code released by the authors.
Only stereochemistry-aware accuracy (Isomer = True) is reported. Topology-only results are omitted for clarity but follow similar trends.
C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.
Raw denotes generation results without re-ranking. Base denotes the default re-ranking configuration, where generated candidates are filtered by removing invalid or duplicate SMILES using RDKit [6] validation, and molecular formula constraints (when available) are enforced to further prune the candidate space before re-ranking.
Combine denotes the best-performing linear combination of SME, SSE, and SSR with coefficients (α, β, γ). For CH and CHF inputs, λ denotes the weighting factor between ^{13}C and ^{1}H modalities in SSR. For single-modality inputs, $\lambda = 0$.

Supplementary Table 14. Generation accuracy of NMRPeak-G with beam size = 10, trained on NMRexp [4] dataset and evaluated on the NMRexp test set simulated by NMRPeak-P-Single.

Modality	Variant	Top-1	Top-5	Top-10
C	MST [5] (baseline)	0.3232	0.4967	0.5349
	NMRPeak-G-Raw	**0.4535**	**0.6514**	**0.6874**
CF	MST [5] (baseline)	0.4879	0.6945	0.7316
	NMRPeak-G-Raw	**0.6197**	**0.8216**	**0.8454**
H	MST [5] (baseline)	0.3543	0.5415	0.5846
	NMRPeak-G-Raw	**0.4909**	**0.7034**	**0.7439**
HF	MST [5] (baseline)	0.5489	0.7502	0.7832
	NMRPeak-G-Raw	**0.6600**	**0.8501**	**0.8737**
CH	MST [5] (baseline)	0.5599	0.7458	0.7771
	NMRPeak-G-Raw	**0.6801**	**0.8593**	**0.8820**
CHF	MST [5] (baseline)	0.6419	0.8226	0.8474
	NMRPeak-G-Raw	**0.7473**	**0.9067**	**0.9222**

MST [5] (baseline) refers to results obtained by training the model from scratch using the official source code released by the authors.
Only stereochemistry-aware accuracy (Isomer = True) is reported. Topology-only results are omitted for clarity but follow similar trends.
C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.
Raw denotes generation results without re-ranking.

Supplementary Table 15. Generation accuracy of NMRPeak-G with beam size = 10, trained on NMRexp [4] dataset and evaluated on the NMRexp test set simulated by NMRPeak-P-Multi.

Modality	Variant	Top-1	Top-5	Top-10
C	MST [5] (baseline)	0.3283	0.4857	0.5422
	NMRPeak-G-Raw	**0.4620**	**0.6620**	**0.6974**
CF	MST [5] (baseline)	0.4925	0.6992	0.7366
	NMRPeak-G-Raw	**0.6256**	**0.8293**	**0.8526**
H	MST [5] (baseline)	0.3627	0.5521	0.5954
	NMRPeak-G-Raw	**0.4983**	**0.7139**	**0.7541**
HF	MST [5] (baseline)	0.5574	0.7593	0.7915
	NMRPeak-G-Raw	**0.6662**	**0.8570**	**0.8792**
CH	MST [5] (baseline)	0.5681	0.7581	0.7892
	NMRPeak-G-Raw	**0.6888**	**0.8691**	**0.8914**
CHF	MST [5] (baseline)	0.6491	0.8307	0.8553
	NMRPeak-G-Raw	**0.7542**	**0.9112**	**0.9261**

MST [5] (baseline) refers to results obtained by training the model from scratch using the official source code released by the authors.
Only stereochemistry-aware accuracy (Isomer = True) is reported. Topology-only results are omitted for clarity but follow similar trends.
C and H correspond to ^{13}C and ^{1}H NMR peaks, respectively. F denotes molecular formula constraints.
Raw denotes generation results without re-ranking.

References

[1] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020)

[2] Wadell, A., Bhutani, A., Viswanathan, V.: Smirk: An atomically complete tokenizer for molecular foundation models. arXiv preprint arXiv:2409.15370 (2024)

[3] Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., Ke, G.: Uni-mol: A universal 3d molecular representation learning framework. In: The Eleventh International Conference on Learning Representations (2023)

[4] Wang, J.-J., Jin, Y., Zhi, C.-Y., Liu, Y.-J., Huang, X.-H., Xu, F., Ji, X., Fang, X., Tao, H., E, W., *et al.*: Nmrexp: A database of 3.3 million experimental nmr spectra. Scientific Data **12**(1), 1954 (2025)

[5] Alberts, M., Schilter, O., Zipoli, F., Hartrampf, N., Laino, T.: Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. Advances in Neural Information Processing Systems **37**, 125780–125808 (2024)

[6] Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., Gedeck, P., Jones, G., Kawashima, E., Nealschneider, D., et al.: rdkit/rdkit: 2025_03_1 (q1 2025) release. Zenodo (2025)