Sim→Exp-MMNMR: A Benchmark for Simulation-to-Experiment Generalization in Multimodal NMR with Chemistry-Aware Metrics

Susanna Di Vita
Department of Computer Science, ETH Zurich
Zurich, Switzerland
sdivita@student.ethz.ch

Abstract

We introduce Sim→Exp-MMNMR, the first benchmark designed to systematically evaluate how well machine learning models and similarity metrics generalize from simulated to experimental nuclear magnetic resonance (NMR) spectra. Unlike prior work, which primarily relies on cosine similarity and simulated data alone, our benchmark features a curated dataset of 1,849 molecules with both simulated and experimental ^{1}H and ^{13}C spectra, standardized under a common solvent and validated for peak alignment. We propose chemistry-aware similarity metrics, including Shift-Marginalized Maximum Mean Discrepancy (SM-MMD), which explicitly account for peak shift uncertainty and calibration noise. Through a suite of four evaluation tasks-including matching, shift robustness, SMILES-to-spectra prediction, and candidate ranking—we show that traditional metrics often misrepresent performance under realistic conditions. Our results demonstrate that SM-MMD offers stronger robustness and structure-awareness, particularly in ^{13}C spectra, suggesting it as a more suitable metric for real-world NMR applications involving domain shift.

1 Introduction

The evaluation of machine learning models for NMR spectroscopy remains surprisingly limited by legacy metrics and synthetic-only datasets. Despite rapid progress in spectrum prediction and inverse structure elucidation, most works still assess models using only simulated spectra and global metrics like cosine similarity [1, 2, 3, 4, 5, 6, 7, 8] , which fail to reflect real-world performance under experimental noise, calibration shifts, and chemical variability.

In this work, we identify and address two key gaps in the current evaluation paradigm. First, standard metrics—chiefly cosine similarity—fail to capture chemically meaningful discrepancies across domains. They are highly sensitive to minor peak misalignments and blind to structural distortions, especially in ^{1}H NMR where peak overlap and solvent effects are prominent. Second, there exists no standardized benchmark for evaluating simulation-to-experiment generalization in NMR under controlled, chemically aligned conditions.

Our contribution. We propose a new set of chemistry-aware metrics, including a handcrafted *Advanced Similarity* score and a distributional kernel-based *Shift-Marginalized Maximum Mean Discrepancy (SM-MMD)*, both designed to overcome the weaknesses of cosine similarity in capturing structure-informed, domain-robust spectral similarity. These metrics are analytically grounded and empirically shown to outperform or match cosine across tasks including retrieval, ranking, match verification, and robustness under synthetic referencing shifts.

To support evaluation, we also release a **benchmark dataset** with *paired simulated and experimental* ^{1}H and ^{13}C spectra for 1,849 molecules, validated for spectral alignment and modality consistency.

All spectra are collected in a unified solvent ($CHCl_3$), with rigorous filtering to remove miscalibrated or ambiguous samples. Our benchmark[1] tasks target key sim-to-real challenges and demonstrate where traditional metrics fail and new ones excel.

2 Dataset

Our benchmark contains 1,849 unique small molecules (see Appendix A for composition) with paired simulated and experimental spectra for ^{1}H and ^{13}C NMR. The final dataset was constructed by aligning simulated spectra from [1] with 123,174 experimental spectra extracted from NMRBank [9] with the highest confidence interval (0.6–1.0), based on common unique SMILES [10]. For each molecule–modality pair, we provide one simulated spectrum and one experimental spectrum, all measured in a single dominant solvent ($CHCl_3$). To ensure correct pairing, we verified peak count consistency within a modality, enforced tight tolerances on the chemical shift ranges (≤ 0.05 ppm for ^{1}H, ≤ 0.5 ppm for ^{13}C), and rejected any pairs with missing or ambiguous peaks. All splits used in the tasks are molecule-disjoint (no SMILES overlap). Each spectrum is released in peak list format as JSON arrays of chemical shift values (δ) in ppm.

3 Benchmark Tasks

Match Verification We evaluate discrimination performance as a binary classification task: matching simulated–experimental spectra versus *hard decoys*. Hard decoys are constructed by selecting *structurally similar compounds*, defined as those with Tanimoto similarity [11] $\tau \in [0.3, 0.8]$ computed using Morgan fingerprints (radius = 2, nBits = 2048). The lower bound excludes trivially unrelated molecules ($\tau < 0.3$), while the upper bound removes near-duplicates that differ only by tautomeric or stereochemical notation ($\tau > 0.8$). This range yields non-identical yet chemically related candidates (analogs, isomers, homologous series), ensuring realistic hard negatives. Positive pairs are $(A_{\text{sim}}, A_{\text{exp}})$; negative pairs are $(A_{\text{sim}}, B_{\text{exp}})$. We report ROC-AUC and PR-AUC with bootstrap confidence intervals ($n = 1000$) and paired significance tests across methods.

Shift-stress Analysis We proceed with assessing the robustness under synthetic referencing offsets $\Delta \in [-0.5, 0.5]$ ppm (1H) and $\Delta \in [-8.0, 8.0]$ ppm (13C) with 21 discrete shift points per spectrum type of our proposed metrics. The system applies artificial chemical shift offsets to experimental spectra and measures how well each similarity metric maintains discrimination between compounds as systematic errors increase. Plot similarity score curves $s(\Delta)$ and report (i) robustness index $R = \frac{1}{\Delta_{max} - \Delta_{min}} \int R(\Delta) d\Delta$ where $R(\Delta) = s(\Delta)/s(0)$ is the retention ratio measuring how much discrimination is preserved, (ii) tolerance points Δ_{95} and Δ_{90}, the largest $|\Delta|$ for which the score remains $\geq 95\%$ and $\geq 90\%$ of baseline respectively, and (iii) local sensitivity $S_0 = \frac{\partial s}{\partial \Delta}|_{\Delta=0}$ with bootstrap confidence intervals ($n = 100$ iterations). This evaluation is critical because real-world NMR spectra often contain systematic referencing errors due to solvent effects, temperature variations, and instrument calibration differences. A robust metric should maintain its ability to distinguish between compounds even when these systematic shifts occur, ensuring reliable compound identification in practical applications where perfect calibration cannot be guaranteed.

Spectral Prediction We trained neural network predictors to generate NMR spectra from SMILES strings using molecular features (Morgan [12] fingerprints with radius=2, nBits=2048, plus 10 molecular descriptors including MolWt, LogP, TPSA, and ring counts). The system processes both 1H and 13C NMR data separately, converting experimental peaks to histogram vectors ($b = 100$ bins, ranges 1H: [0,12] ppm, 13C: [0,220] ppm) with Gaussian smoothing ($\sigma = 1.0$). The neural architecture uses 4 hidden layers [512, 256, 128, 64] with BatchNorm, ReLU activation, Dropout (0.3), and Xavier weight initialization. Training uses AdamW (lr = 0.001, weight decay = 1e-4) with ReduceLROnPlateau scheduling, and optimizes mean squared error (MSE) between the smoothed histogram vectors. Alternative loss functions based on the proposed similarity metrics (SM-MMD and a hybrid MSE+SM-MMD) are evaluated in Appendix D. This evaluation is critical because real-world NMR spectra often contain systematic referencing errors due to solvent effects, temperature variations, and instrument calibration differences.

[1] Benchmark dataset, metric implementations, and full experiment code available at `https://github.com/SusannaDiV/SimExp-MMNMR`.

Candidate ranking We further evaluate metric performance in candidate selection scenarios using fixed candidate sets of $n = 30$ compounds selected sequentially from the dataset. The system takes an experimental spectrum and ranks candidate molecules by their spectral similarity, measuring how consistently different metrics order the same set of candidates. Performance is measured using rank correlations (Spearman ρ) between different metrics, Top-K analysis showing candidate overlap across metrics, and overall ranking stability. This evaluation is important because it reveals whether different metrics produce similar candidate rankings, which is critical for practical applications where chemists need consistent results regardless of which metric they choose.

4 Metrics

Cosine Similarity As a baseline, we include the standard cosine similarity between two normalized peak intensity vectors $u, v \in \mathbb{R}^d$:

$$\mathrm{CosSim}(u, v) = \frac{\langle u, v \rangle}{\|u\|_2 \, \|v\|_2} = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2} \, \sqrt{\sum_{i=1}^{d} v_i^2}}.$$

Cosine similarity takes values in $[-1, 1]$ and provides a simple measure of shape agreement between spectra, independent of absolute intensity scaling.

Advanced NMR Similarity Our proposed *Advanced NMR Similarity* metric combines peak matching via Hungarian [13] algorithm, density correlation, chemical shift significance, and peak count similarity with weights $[0.4, 0.3, 0.2, 0.1]$ respectively. The weights $[0.4, 0.3, 0.2, 0.1]$ are not learned or tuned; they are heuristically chosen to reflect chemical intuition. Exact peak matching (S_1, $w_1 = 0.4$) provides the strongest evidence that two spectra correspond to the same molecular structure, followed by global peak distribution similarity (S_2, $w_2 = 0.3$), and chemical-shift significance (S_3, $w_3 = 0.2$). Peak count consistency (S_4, $w_4 = 0.1$) receives the lowest weight because it captures only coarse structural information. The weights encode this priority ordering and are chosen for interpretability rather than numerical optimization. The complete definition and proof of correctness can be found in Appendix B.

Shift-Marginalized Maximum Mean Discrepancy (SM-MMD). We evaluate a kernel-based metric because NMR spectra are naturally *distributions of peaks*, not aligned vectors. Metrics such as cosine similarity operate on absolute peak positions and are therefore highly sensitive to global referencing shifts or small peak misalignments. In contrast, SM-MMD compares spectra in a reproducing kernel Hilbert space (RKHS), where distance reflects distributional similarity. Our formulation analytically marginalizes over all global shifts within $[-S, S]$, making the metric inherently shift-invariant. We use a multi-scale Gaussian kernel with bandwidths $\sigma \in \{0.05, 0.1, 0.15\}$ for ^{1}H and $\sigma \in \{1.0, 2.0, 5.0\}$ for ^{13}C (weights $[0.5, 0.3, 0.2]$), capturing both narrow and broad peak features. Importantly, the kernel formulation is fully differentiable, enabling SM-MMD to be used directly as a training loss for generative modeling (Appendix D). The complete definition and proof of correctness can be found in Appendix B-C.

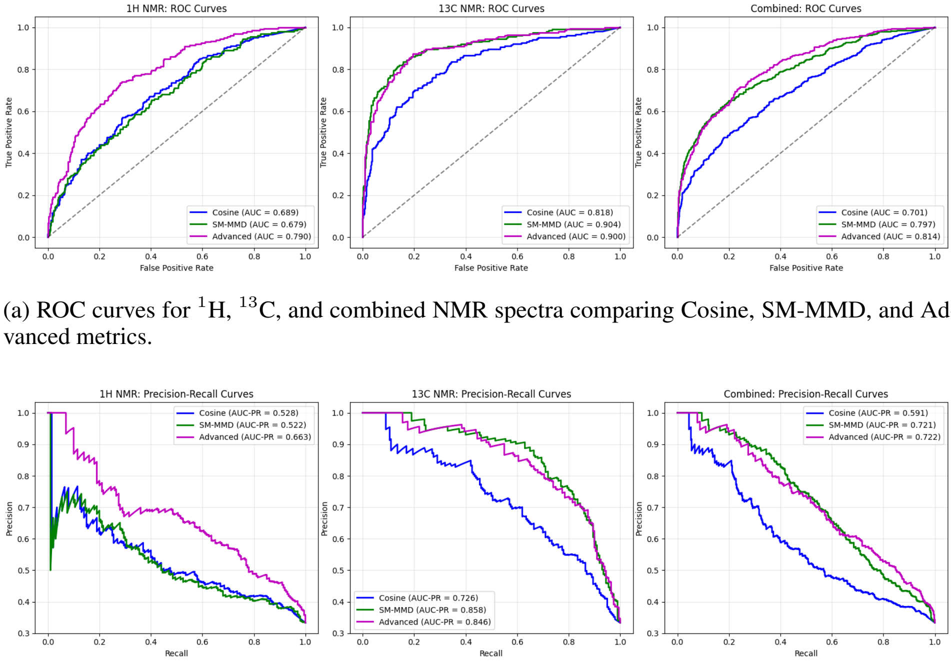
5 Results

While cosine similarity provides a simple baseline for comparing spectra, it treats each spectrum as a fixed high-dimensional vector. This approach implicitly assumes perfect peak alignment and equal importance of all chemical shifts. As a result, even small global referencing shifts or minor peak-picking errors can lead to disproportionately low similarity scores, despite the underlying spectra being chemically identical. Moreover, cosine similarity is insensitive to local distributional structure: two spectra with similar peak densities but slight misalignments may appear very dissimilar.

Our proposed shift-marginalized MMD (SM-MMD) overcomes these limitations by directly comparing the *peak distributions* rather than fixed vectors. By marginalizing over all referencing shifts within a window $[-S, S]$ (with S being 0.15 ppm for 1H, 2.0 for 13C) and incorporating peak intensities as weights, SM-MMD remains stable under global shifts and robust to small local perturbations. The kernel formulation further enables distributional comparison in a reproducing kernel Hilbert space, providing a mathematically rigorous, shift-tolerant, and intensity-aware similarity measure for NMR spectra. All the detailed results to our experiments can be found in Appendix D.

This phenomenon is evident on synthetic shift-stress analysis, where we deliberately applied global referencing offsets ($\Delta \in [-0.5, 0.5]$ ppm for ^{1}H and $\Delta \in [-8.0, 8.0]$ ppm for ^{13}C) to experimental spectra and measured normalized retention ratios $s(\Delta)/s(0)$. Cosine similarity and the Advanced NMR metric both peak at $\Delta = 0$, then decay rapidly with increasing shift, because they rely on absolute peak positions: cosine operates on histogram bins, and the Advanced metric enforces explicit peak–peak matching within fixed tolerance windows. Any global shift moves peaks to different bins or outside tolerance bounds, causing similarity to drop. In contrast, SM-MMD maintains nearly constant similarity across the entire shift range, because the shift-marginalized kernel analytically integrates over all referencing offsets in $[-S, S]$ and therefore compares spectra based on their peak distributions rather than their absolute positions.

The robustness index R (area under the retention curve) shows SM-MMD ($R = 0.9993$) and SM-JS ($R = 0.9974$) maintain near-perfect score retention across the full shift range as observable in Figure 3, while cosine and the Advanced NMR metric degrade rapidly under even small shifts. Across both nuclei, SM-MMD achieves robustness indices of 0.999 for ^{1}H and ^{13}C, far exceeding Cosine (0.633 for ^{1}H, 0.410 for ^{13}C) and the Advanced metric (0.221 for ^{1}H, 0.336 for ^{13}C), confirming its near shift-invariance even under severe referencing perturbations (Figure 3 (c)). Local sensitivity analysis confirms this trend: SM-JS achieves $S_0 = 0.0040\,[0.0002, 0.0118]$ versus $S_0 = 0.1930\,[0.0131, 0.4975]$ for the Advanced metric, indicating far lower sensitivity to infinitesimal referencing errors. All shift-marginalized methods retain $> 95\%$ of their baseline scores across the entire range, producing flat, high robustness curves where similarity remains nearly constant even under severe misalignment. In contrast, cosine shows narrow, peaked curves with steep drops, confirming its lack of tolerance to realistic referencing errors.



(a) ROC curves for ^{1}H, ^{13}C, and combined NMR spectra comparing Cosine, SM-MMD, and Advanced metrics.

(b) Precision–Recall curves for ^{1}H, ^{13}C, and combined NMR spectra comparing Cosine, SM-MMD, and Advanced metrics.

Figure 1: (A) ROC and (B) Precision–Recall curves across ^{1}H, ^{13}C, and combined NMR spectra for Cosine, SM-MMD, and Advanced metrics.

The same trend appears in the candidate ranking task of Figure 4. Across 30 candidate molecules, cosine similarity produced the lowest scores ($0.0000-0.5340$), the Advanced NMR metric yielded moderate values ($0.1848-0.4657$), and SM-MMD returned the highest ($0.6045-0.9127$) due to its robustness to peak shifts. Ranking correlations showed strong agreement between cosine and

Advanced NMR ($\rho = 0.92$) but only moderate alignment with SM-MMD ($\rho \approx 0.67$), reflecting its focus on distributional rather than peak-wise similarity. Top-k analysis confirmed partial overlap: no candidate ranked first across all metrics, but six appeared in every top-10 list.

In the match verification task (Figure 1), the Advanced NMR metric achieves the highest classification accuracy, reaching AUC = 0.900 (ROC) and AUC-PR = 0.846 for ^{13}C NMR, while SM-MMD performs competitively with AUC = 0.797 (ROC) and AUC-PR = 0.721 in the combined setting. Cosine similarity lags behind on both metrics, indicating that handcrafted features or shift-marginalization improve discrimination between matching and decoy spectra beyond simple vector comparisons. However, SM-MMD exhibits nucleus-specific behavior: while it significantly improves ^{13}C discrimination (AUC-ROC = 0.904 vs. cosine's 0.818), its ^{1}H performance (AUC-ROC = 0.679) remains comparable to cosine (0.689). This disparity arises not from noise or instrument resolution, but from peak congestion and spectral scale. In ^{1}H NMR, different molecules produce dense clusters of peaks within a narrow 0–12 ppm range, with inter-molecule separations of only 0.1–0.3 ppm—comparable to SM-MMD's kernel bandwidth ($\sigma = 0.15$ ppm). At this scale, the shift-marginalized kernel smooths nearby peaks, causing distinct molecules to appear artificially similar. In contrast, ^{13}C NMR peaks are well separated (10–30 ppm apart across a 220 ppm range), so the kernel bandwidth ($\sigma = 2$ ppm) remains selective enough to distinguish different molecules. Thus, SM-MMD excels when peak spacing is larger than the kernel resolution (^{13}C), but becomes less discriminative in highly congested spectral domains (^{1}H).

Finally, in the SMILES-to-spectra prediction task with 400 predicted spectra (200 ^{1}H, 200 ^{13}C), SM-MMD again achieved the highest similarity (0.9066 ± 0.0751 overall), followed by cosine (0.8358 ± 0.1119) and the Advanced NMR metric (0.6124 ± 0.1084), as observable in Figure 5. Interestingly, cosine performs surprisingly well here because the prediction model outputs spectra in the same reference frame as the training data, so global shifts and large misalignments — the main failure modes of cosine — rarely occur. Nevertheless, SM-MMD still achieves higher scores, especially for ^{1}H NMR (0.9573 vs. 0.8451), because it remains robust to local peak-picking errors and amplitude noise that cosine does not handle. The Advanced NMR metric yields lower absolute scores (0.6693 for ^{1}H, 0.5554 for ^{13}C) because its strict peak-matching and density components penalize every deviation, providing interpretability at the cost of conservative similarity values.

6 Conclusion

We present Sim→Exp-MMNMR, a 1H-13C benchmark to quantify simulation-to-experiment generalization in NMR spectroscopy. Our key contributions include (1) a new high-quality dataset of aligned simulated and experimental ^{1}H and ^{13}C spectra, and (2) domain-aware similarity metrics such as SM-MMD that outperform standard metrics like cosine under realistic shift, spectral prediction, and candidate ranking scenarios. Across multiple tasks, we find that traditional cosine similarity underperforms—particularly in scenarios with domain shift or structural ambiguity, whereas SM-MMD provides more robust and chemically meaningful discrimination. Future extensions include incorporating HSQC spectra into the benchmark to fully support 2D multimodal evaluation, extending the framework to solvent-dependent domain shift and cross-instrument variability, and applying our metrics in structure elucidation pipelines.

References

[1] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *arXiv preprint*, 2024.

[2] Qingsong Yang, Binglan Wu, Xuwei Liu, Kai Chen, Wei Li, Gen Long, Xin Chen, and Mingjun Xiao. Diffnmr: Diffusion models for nuclear magnetic resonance spectra elucidation. *arXiv preprint arXiv:2507.08854*, 2025.

[3] Tanvir Sajed, Zinat Sayeeda, Brian L. Lee, Mark Berjanskii, Fei Wang, Vasuk Gautam, and David S. Wishart. Accurate prediction of ^{1}h nmr chemical shifts of small molecules using machine learning. *Metabolites*, 14(5):290, 2024.

[4] H.W. Kim, C. Zhang, R. Reher, et al. Deepsat: Learning molecular structures from nuclear magnetic resonance data. *Journal of Cheminformatics*, 15(71), 2023.

[5] Federico Zipoli, Marvin Alberts, and Teodoro Laino. Ir–nmr multimodal computational spectra dataset for 177k patent-extracted organic molecules. *Scientific Data*, 12:1375, 2025.

[6] Xi Xue, Hanyu Sun, Jingying Sun, Luc Patiny, Xiangying Liu, Kai Chen, Jingjie Yan, Liangning Li, Xue Liu, Shu Xu, Dongming Zhang, Yafeng Deng, Yingda Zang, Yaling Gong, Jie Ma, and Xiaojian Wang. Nmrmind: A transformer-based model enabling the elucidation from multidimensional nmr to structures. *Analytical Chemistry*, 97(41):22603–22614, 2025.

[7] Anonymous. Spectrallm: Uncovering the ability of llms for molecule structure elucidation from multi-spectra. In *International Conference on Learning Representations (ICLR), under review*, 2026.

[8] Yeongsu Kwon, Donghoon Lee, Yoonsu S. Choi, et al. Molecular search by nmr spectrum based on evaluation of matching between spectrum and molecule. *Scientific Reports*, 11:20998, 2021.

[9] Qinggong Wang, Wei Zhang, Mingan Chen, Xutong Li, Zhaoping Xiong, Jiacheng Xiong, Zunyun Fu, and Mingyue Zheng. Nmrextractor: leveraging large language models to construct an experimental nmr database from open-source scientific publications. *Chemical Science*, 16:11548–11558, 2025.

[10] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28:31–36, 1988.

[11] T. Tanimoto. An elementary mathematical theory of classification and prediction. Technical Report Technical Report, IBM Corporation, New York, 1958.

[12] H. L. Morgan. The generation of a unique machine description for chemical structures — a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965.

[13] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

[14] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[15] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel Two-Sample test. In *Journal of Machine Learning Research*, volume 13, pages 723–773, 2012.

[16] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert R. G. Lanckriet, and Bernhard Schölkopf. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

A Dataset Composition

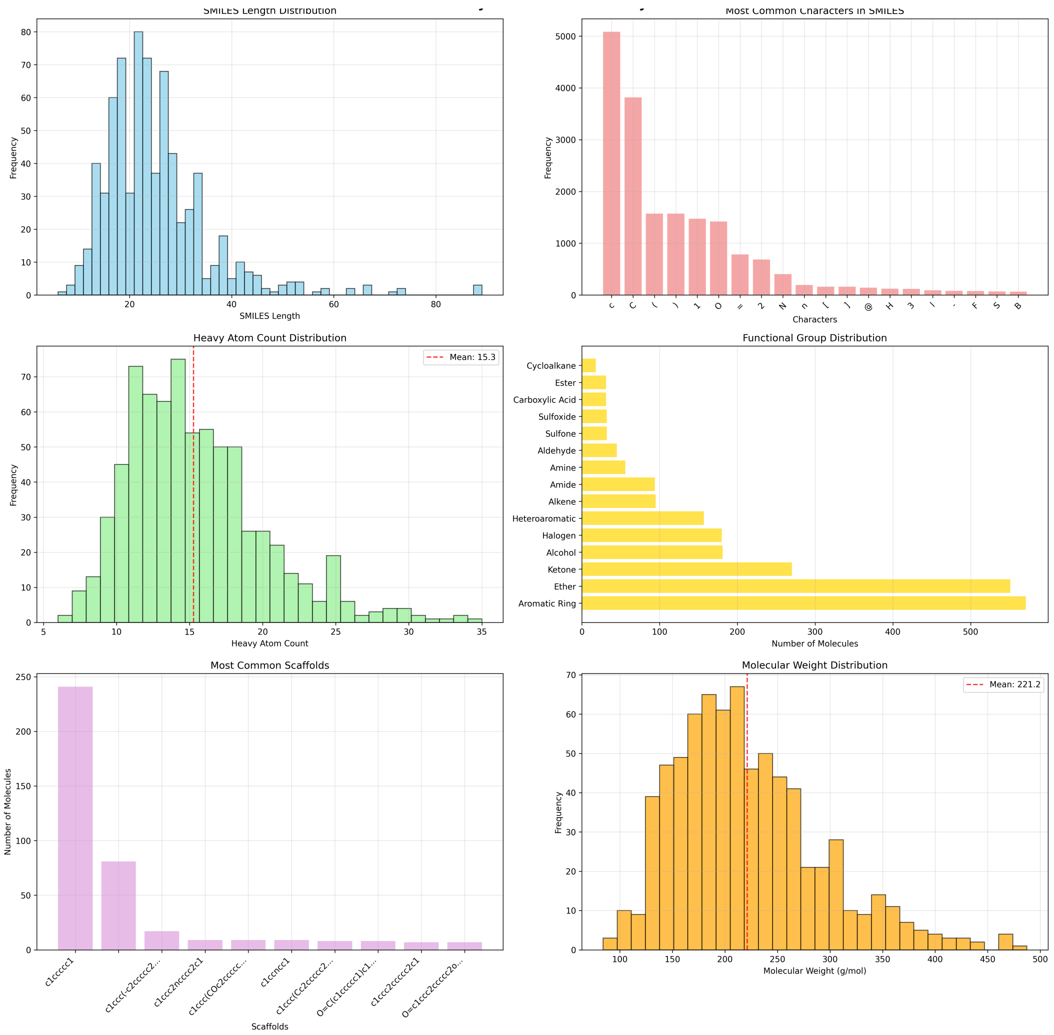


Figure 2: Summary plots for the Sim-EXP-MMNMR dataset: SMILES length and character frequencies, heavy atom counts, functional group tallies, common scaffolds, and molecular weight distribution.

The dataset's SMILES lengths are concentrated in the low-to-mid 20s, indicating moderate structural complexity, and the most common characters are C/c and =, consistent with aliphatic and aromatic carbons plus frequent double bonds. Heavy atom counts follow a unimodal distribution with a mean near **15.3** heavy atoms per molecule. In terms of functional groups, aromatic rings are the most prevalent, with ethers and ketones also frequent, followed by alcohols and halogens, suggesting a bias toward common drug-like motifs. A few aromatic scaffolds dominate the head of the distribution, with a long tail of rarer frameworks. Finally, the molecular weight histogram centers in the 200–250 g/mol band, with a mean around **221.2** g/mol, squarely within typical small-molecule ranges.

B Metrics Definition

Advanced NMR Similarity Metric We propose an *Advanced NMR Similarity Metric* that decomposes spectrum similarity into four chemically-interpretable components. For peak sets $A = \{(p_i, a_i)\}_{i=1}^{m}$ and $B = \{(q_j, b_j)\}_{j=1}^{n}$ with normalized intensities $\sum_i a_i = \sum_j b_j = 1$, the similarity metric is defined as:

$$\text{Sim}(A, B) = \sum_{k=1}^{4} w_k \cdot S_k(A, B) \tag{1}$$

where $w = [0.4, 0.3, 0.2, 0.1]$ are the component weights. The *peak matching component* $S_1(A, B)$ employs optimal assignment theory:

$$S_1(A, B) = 1 - \frac{\min_{\sigma \in \Pi_{m,n}} \sum_{i=1}^{m} C_{i,\sigma(i)}}{m} \tag{2}$$

where $\Pi_{m,n}$ is the set of partial permutations, $m = \min(|A|, |B|)$, and the cost matrix $C_{ij} = \min(|\Delta_{ij}|/(3\tau), 1)$ with $\Delta_{ij} = |p_i - q_j|$ and tolerance τ (0.3 ppm for ^{1}H, 2.0 ppm for ^{13}C). The optimal assignment is computed using the Hungarian algorithm. The *density similarity component* $S_2(A, B)$ measures distribution correlation:

$$S_2(A, B) = \max(0, \rho(\text{hist}_A, \text{hist}_B)) \tag{3}$$

where hist_A and hist_B are normalized histograms with 50 bins over ranges $[0, 12]$ ppm (^{1}H) or $[0, 220]$ ppm (^{13}C), and ρ is the Pearson correlation coefficient. The *chemical significance component* $S_3(A, B)$ captures shift importance:

$$S_3(A, B) = 1 - \left| \frac{1}{|A|} \sum_{i=1}^{|A|} \left(\frac{p_i}{R_{\text{NMR}}} \right)^2 - \frac{1}{|B|} \sum_{j=1}^{|B|} \left(\frac{q_j}{R_{\text{NMR}}} \right)^2 \right| \tag{4}$$

where R_{NMR} is the appropriate range (12.0 or 220.0 ppm). The *peak count component* $S_4(A, B)$ measures structural consistency:

$$S_4(A, B) = 1 - \frac{||A| - |B||}{\max(|A|, |B|, 1)} \tag{5}$$

Note. The peak-count component $S_4(A, B)$ does not satisfy the triangle inequality when interpreted as a distance (i.e., $d_4 = 1 - S_4$), because its normalization uses the pair-dependent denominator $\max(|A|, |B|, 1)$. This adaptive scaling changes the distance scale from comparison to comparison and can yield cases where $d_4(A, C) > d_4(A, B) + d_4(B, C)$. This is not an issue in our setting: S_4 is only a bounded similarity term (weight $w_4 = 0.1$) inside a composite score and is never used as a standalone metric for kernel methods, geometric embedding, or nearest-neighbor search.

This decomposition provides interpretable similarity assessment while maintaining mathematical rigor through optimal assignment theory and statistical correlation measures. The bounded output $\text{Sim}(A, B) \in [0, 1]$ ensures intuitive interpretation where higher values indicate greater spectral similarity.

B.1 Shift-Marginalized Maximum Mean Discrepancy (SM-MMD)

We propose *Shift-Marginalized MMD (SM-MMD)* for comparing NMR spectra in peak-list form. Each spectrum is represented as a set of peaks with normalized intensities:

$$A = \{(p_i, a_i)\}_{i=1}^{m}, \quad B = \{(q_j, b_j)\}_{j=1}^{n}, \quad \sum_i a_i = \sum_j b_j = 1.$$

Here p_i, q_j denote chemical shifts (in ppm), while a_i, b_j are normalized peak areas (or heights).

Shift-marginalized kernel Let the base Gaussian kernel [14] be

$$k_\sigma(t) = \exp\left(-\frac{t^2}{2\sigma^2} \right).$$

To ensure robustness against global referencing shifts, we average this kernel over all shifts $s \in [-S, S]$:

$$k_{\sigma,S}(p,q) = \frac{1}{2S}\int_{-S}^{S} k_\sigma(p-q-s)\,\mathrm{d}s$$
$$= \frac{\sigma\sqrt{\pi/2}}{2S}\Big[\mathrm{erf}\Big(\frac{p-q+S}{\sqrt{2}\sigma}\Big) - \mathrm{erf}\Big(\frac{p-q-S}{\sqrt{2}\sigma}\Big)\Big]. \tag{6}$$

This kernel is *shift-marginalized*: small global shifts Δ do not significantly affect the similarity.

Shift-Marginalized MMD We build on the maximum mean discrepancy (MMD) framework [15], which measures discrepancies between probability distributions via their mean embeddings in an RKHS. The squared SM-MMD between A and B is defined as

$$\mathrm{MMD}^2_{\sigma,S}(A,B) = \sum_{i,i'} a_i a_{i'}\, k_{\sigma,S}(p_i,p_{i'}) + \sum_{j,j'} b_j b_{j'}\, k_{\sigma,S}(q_j,q_{j'}) - 2\sum_{i,j} a_i b_j\, k_{\sigma,S}(p_i,q_j). \tag{7}$$

The similarity score is then reported as

$$\mathrm{Sim}(A,B) = 1 - \mathrm{MMD}^2_{\sigma,S}(A,B),$$

Multi-scale extension For additional robustness, we use a multi-scale mixture of kernels with normalized weights $\sum_\ell w_\ell = 1$:

$$\mathrm{MMD}^2(A,B) = \sum_\ell w_\ell\, \mathrm{MMD}^2_{\sigma_\ell,S}(A,B),$$

with weights $w_\ell \geq 0$.

C Proofs

Shift-Marginalized Kernel Properties Let

$$k_{\sigma,S}(x,y) = \frac{1}{2S}\int_{-S}^{S} \exp\Big(-\frac{(x-y-s)^2}{2\sigma^2}\Big)\,\mathrm{d}s.$$

Then $k_{\sigma,S}(x,y)$ satisfies the following:

1. **Symmetry:** $k_{\sigma,S}(x,y) = k_{\sigma,S}(y,x)$ for all x, y.
2. **Positive definiteness:** For any $x_1,\dots,x_n \in \mathbb{R}$ and $c_1,\dots,c_n \in \mathbb{R}$,

$$\sum_{i,j} c_i c_j k_{\sigma,S}(x_i,x_j) \geq 0.$$

3. **Boundedness:** $0 \leq k_{\sigma,S}(x,y) \leq 1$ for all x, y (with appropriate normalization).

Proof. **1. Symmetry.** For any x, y,

$$k_{\sigma,S}(x,y) = \frac{1}{2S}\int_{-S}^{S} e^{-(x-y-s)^2/(2\sigma^2)}\,\mathrm{d}s.$$

Let $u = x - y - s$ so $\mathrm{d}s = -\mathrm{d}u$. Reversing the integration limits yields the same value since $e^{-u^2/(2\sigma^2)} = e^{-(-u)^2/(2\sigma^2)}$. Replacing $x - y$ by $y - x$ leaves the integral unchanged:

$$k_{\sigma,S}(y,x) = \frac{1}{2S}\int_{-S}^{S} e^{-(y-x-s)^2/(2\sigma^2)}\,\mathrm{d}s = \frac{1}{2S}\int_{-S}^{S} e^{-(x-y+s)^2/(2\sigma^2)}\,\mathrm{d}s = k_{\sigma,S}(x,y).$$

2. Positive definiteness. The Gaussian kernel $k_\sigma(t) = e^{-t^2/(2\sigma^2)}$ is positive definite on $\mathbb{R}$. Since

$$k_{\sigma,S}(x,y) = \frac{1}{2S}\int_{-S}^{S} k_\sigma(x-y-s)\,\mathrm{d}s$$

is a nonnegative integral (convex combination) of positive-definite kernels, $k_{\sigma,S}$ remains positive definite by standard closure properties.

3. Boundedness. We have $0 < k_\sigma(t) \le 1$, with equality $k_\sigma(0) = 1$ and $k_\sigma(t) \to 0$ as $|t| \to \infty$. Since $k_{\sigma,S}(x, y)$ is the average of $k_\sigma(x - y - s)$ over $s \in [-S, S]$,

$$0 \le k_{\sigma,S}(x, y) \le \frac{1}{2S} \int_{-S}^{S} 1 \, \mathrm{d}s = 1.$$

□

Multi-scale case Positive definiteness ensures that this kernel can be used inside any kernel-based method (e.g., SVMs, MMD) and that MMD remains a squared distance in an RKHS. The following result shows that the multi-scale kernel inherits all key properties from its single-scale components.

Let

$$k_{\mathrm{MS}}(x, y) = \sum_{\ell=1}^{L} w_\ell \, k_{\sigma_\ell,S}(x, y), \qquad w_\ell \ge 0, \ \sum_\ell w_\ell = 1.$$

Then k_{MS} is symmetric, positive definite, and satisfies $0 \le k_{\mathrm{MS}}(x, y) \le 1$.

Proof. Each $k_{\sigma_\ell,S}$ is symmetric and positive definite by the theorem above. A nonnegative linear combination preserves these properties:

$$k_{\mathrm{MS}}(x, y) = \sum_\ell w_\ell k_{\sigma_\ell,S}(x, y) = \sum_\ell w_\ell k_{\sigma_\ell,S}(y, x) = k_{\mathrm{MS}}(y, x),$$

and for any $c_1, \dots, c_n$,

$$\sum_{i,j} c_i c_j k_{\mathrm{MS}}(x_i, x_j) = \sum_\ell w_\ell \sum_{i,j} c_i c_j k_{\sigma_\ell,S}(x_i, x_j) \ge 0.$$

Finally, since each $k_{\sigma_\ell,S}(x, y) \in [0, 1]$ and $\sum_\ell w_\ell = 1$,

$$0 \le k_{\mathrm{MS}}(x, y) \le \sum_\ell w_\ell \cdot 1 = 1.$$

□

Characteristic property. The kernel $k_{\sigma,S}$ is translation-invariant and can be written as the convolution of a Gaussian with a uniform window:

$$k_{\sigma,S}(x, y) = (k_\sigma * u_S)(x - y),$$

where u_S denotes the uniform density on $[-S, S]$. Its spectral density is given by the product of the Gaussian spectral density and a sinc factor arising from the uniform window. This spectral density is nonnegative and strictly positive on open intervals (its zeros occur only at a discrete set). Therefore, the spectral measure has support with nonempty interior.

By the standard characterization of characteristic kernels (e.g., [16]), any translation-invariant positive-definite kernel whose spectral measure has nonempty interior support is *characteristic* on $\mathbb{R}$. Hence $k_{\sigma,S}$ is characteristic, and any finite nonnegative mixture of such kernels preserves this property.

Consequently, the multi-scale SM-MMD induced by k_{MS} defines a *metric* on the space of probability measures over peak positions (with intensities as weights).

Non-negativity of SM-MMD Non-negativity ensures that MMD is a valid distance measure (actually a pseudometric, becoming a true metric if the kernel is characteristic). The next result uses the positive-definiteness established above to show that the MMD constructed from this kernel inherits non-negativity because it is a squared Hilbert space distance.

[Non-negativity of SM-MMD] For any peak sets $A = \{(p_i, a_i)\}_i$ and $B = \{(q_j, b_j)\}_j$ with $a_i, b_j \ge 0$ and $\sum_i a_i = \sum_j b_j = 1$, the shift-marginalized maximum mean discrepancy satisfies

$$\mathrm{MMD}^2_{\sigma,S}(A, B) \ \ge \ 0.$$

Proof (RKHS view). Because $k_{\sigma,S}$ is positive definite (as shown earlier), it induces a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with feature map $\varphi(\cdot)$. Define the mean embeddings of the discrete measures represented by the peak lists:

$$\mu_A = \sum_i a_i\, \varphi(p_i), \qquad \mu_B = \sum_j b_j\, \varphi(q_j).$$

By standard MMD theory,

$$\mathrm{MMD}^2_{\sigma,S}(A,B) = \|\mu_A - \mu_B\|^2_{\mathcal{H}} = \langle \mu_A, \mu_A \rangle + \langle \mu_B, \mu_B \rangle - 2\langle \mu_A, \mu_B \rangle.$$

Expanding the inner products in $\mathcal{H}$ using the kernel $k_{\sigma,S}$ gives exactly the formula

$$\sum_{i,i'} a_i a_{i'} k_{\sigma,S}(p_i, p_{i'}) + \sum_{j,j'} b_j b_{j'} k_{\sigma,S}(q_j, q_{j'}) - 2 \sum_{i,j} a_i b_j k_{\sigma,S}(p_i, q_j).$$

Because this is the squared norm $\|\mu_A - \mu_B\|^2_{\mathcal{H}}$, it is always non-negative:

$$\mathrm{MMD}^2_{\sigma,S}(A,B) \geq 0.$$

□

As each spectrum is represented as a finite weighted set of peaks $\{(p_i, a_i)\}$, where p_i is the chemical shift (in ppm) and a_i is the normalized peak area, $\sum_i a_i = 1$, this defines a discrete probability measure on $\mathbb{R}$ (or $\mathbb{R}^2$ for 2D NMR), and the SM-MMD measures the squared RKHS distance between the mean embeddings of these two measures. Because squared distances in any Hilbert space are always non-negative, the SM-MMD inherits this property automatically, confirming it a valid dissimilarity measure for NMR peak distributions.

Analytical Marginalization Correctness Let $k_\sigma(t) = \exp\big(-t^2/(2\sigma^2)\big)$ with $\sigma > 0$ be the standard Gaussian kernel on $\mathbb{R}$, and let $\delta \sim \mathrm{Uniform}[-S, S]$ with $S > 0$. Define the *shift-marginalized kernel* as the expectation of k_σ under an additive random shift δ:

$$\tilde{k}(x,y) \;:=\; \mathbb{E}_\delta\big[k_\sigma((x-y)+\delta)\big].$$

Then, for all $x, y \in \mathbb{R}$,

$$\tilde{k}(x,y) = \frac{1}{2S}\int_{-S}^{S} e^{-\frac{(x-y-s)^2}{2\sigma^2}}\, ds = \frac{\sigma\sqrt{\pi/2}}{2S}\left[\mathrm{erf}\Big(\tfrac{x-y+S}{\sqrt{2}\sigma}\Big) - \mathrm{erf}\Big(\tfrac{x-y-S}{\sqrt{2}\sigma}\Big)\right] = k_{\sigma,S}(x,y).$$

Proof. The expectation over a uniform δ on $[-S, S]$ simply corresponds to averaging over all possible shift values in that interval:

$$\tilde{k}(x,y) = \mathbb{E}_\delta[k_\sigma((x-y)+\delta)] = \frac{1}{2S}\int_{-S}^{S} k_\sigma((x-y)+s)\, ds.$$

Because the Gaussian kernel $k_\sigma(t) = e^{-t^2/(2\sigma^2)}$ is an *even function* ($k_\sigma(t) = k_\sigma(-t)$) and the averaging interval $[-S, S]$ is *symmetric around zero*, the integral above is unchanged if we replace $s \mapsto -s$. Therefore,

$$\int_{-S}^{S} k_\sigma((x-y)+s)\, ds = \int_{-S}^{S} k_\sigma((x-y)-s)\, ds.$$

This gives

$$\tilde{k}(x,y) = \frac{1}{2S}\int_{-S}^{S} e^{-\frac{(x-y-s)^2}{2\sigma^2}}\, ds.$$

Now we evaluate this Gaussian integral in closed form. Set $u = \frac{x-y-s}{\sqrt{2}\sigma}$ so that $s = x - y - \sqrt{2}\sigma u$ and $ds = -\sqrt{2}\sigma\, du$. The limits transform as follows:

- when $s = -S$, $u = \frac{x-y+S}{\sqrt{2}\sigma}$,

- when $s = S$, $u = \frac{x-y-S}{\sqrt{2}\sigma}$.

Substituting into the integral:

$$\frac{1}{2S} \int_{-S}^{S} e^{-\frac{(x-y-s)^2}{2\sigma^2}} \, ds = \frac{\sigma\sqrt{2}}{2S} \int_{(x-y-S)/(\sqrt{2}\sigma)}^{(x-y+S)/(\sqrt{2}\sigma)} e^{-u^2} \, du.$$

The antiderivative of e^{-u^2} is the error function $\mathrm{erf}(u)$, so we obtain

$$\tilde{k}(x, y) = \frac{\sigma\sqrt{\pi/2}}{2S} \left[\mathrm{erf}\!\left(\tfrac{x-y+S}{\sqrt{2}\sigma}\right) - \mathrm{erf}\!\left(\tfrac{x-y-S}{\sqrt{2}\sigma}\right) \right].$$

Finally, by definition this equals $k_{\sigma,S}(x, y)$, the *shift-marginalized kernel* used in SM-MMD. Hence the expectation $\mathbb{E}[k_\sigma((x - y) + \delta)]$ over $\delta \sim \mathrm{Uniform}[-S, S]$ indeed produces the closed-form expression in Eq. (6).

$$\tilde{k}(x, y) = k_{\sigma,S}(x, y)$$

This result formally guarantees that the *analytically marginalized kernel* used in SM-MMD exactly equals the expected kernel value over random global shifts, so the kernel similarity is provably robust to any referencing offset up to magnitude S. □

D Differentiable Loss Functions for Spectral Prediction

Advanced NMR Similarity (non-differentiable) The Advanced NMR metric in Appendix A cannot be used for gradient-based learning because two components are non-differentiable: (i) S_1 requires solving a *discrete peak assignment* via the Hungarian algorithm, and (ii) S_4 depends on *peak counting* through a hard threshold. Both introduce argmin/threshold operations with zero gradients, preventing backpropagation.

Differentiable SM-MMD loss For spectral prediction experiments, we adapt SM-MMD to make it fully differentiable and suitable for end-to-end optimization. SM-MMD as an evaluation metric (Appendix B–C) operates on *sparse peak lists*, which is natural for comparing experimental spectra, whereas neural networks produce fixed-size *dense histograms*. Converting histograms back into peak lists would require non-differentiable operations (thresholding, peak picking), so during training we apply SM-MMD directly to the histogram outputs: (i) each histogram bin is treated as a pseudo-peak with its height as weight, (ii) all pairwise kernel interactions are computed in a single batched GPU operation, and (iii) MMD^2 is returned as the loss to be minimized. Importantly, the differentiable loss and the evaluation metric use the *same* analytically shift-marginalized kernel $k_{\sigma,S}(x, y)$ and identical MMD formulation; only the input representation differs (sparse peaks vs. dense histograms). For training efficiency, we use a *single* bandwidth σ, whereas evaluation uses a *multi-scale mixture* (three bandwidths with normalized weights). Thus, SM-MMD functions both as a rigorous evaluation metric and as an end-to-end differentiable training objective for learning NMR spectra from SMILES.

Hybrid loss (MSE + SM-MMD) We also evaluate a Hybrid loss that combines pixel-level accuracy (MSE) with chemically meaningful distributional similarity (SM-MMD), defined as $\mathcal{L}_{\text{hybrid}} = \alpha \cdot \mathrm{MSE} + (1 - \alpha) \cdot (10 \cdot \mathrm{MMD}^2)$, where $\alpha \in [0, 1]$ controls the trade-off. MSE enforces bin-wise agreement and stabilizes optimization, while SM-MMD matches the global spectral distribution and remains invariant to referencing shifts. The constant factor 10 balances the magnitude of the two terms so neither dominates during training.

Results As observable ifn Figure 6, models trained with SM-MMD loss consistently outperform MSE-trained models on chemically meaningful metrics. For ^{1}H NMR, SM-MMD-trained models achieve median SM-MMD similarity of 0.95 (IQR: 0.94–0.96) and Advanced NMR similarity of 0.72 (IQR: 0.68–0.77), compared to 0.93 and 0.70 for MSE-trained models. For ^{13}C NMR, the advantage is even more pronounced: SM-MMD-trained models reach median SM-MMD of 0.94 and Advanced NMR of 0.74 (IQR: 0.68–0.81), versus 0.89 and 0.59 for MSE. Hybrid loss ($\alpha = 0.5$)

gives intermediate performance. While MSE-trained models achieve lower MSE by definition, their poor SM-MMD and Advanced scores indicate that they fail to learn the underlying peak *distribution*. Finally, the finer histogram resolution used for ^{1}H NMR (100 bins, 0.06 ppm/bin) is critical: it allows SM-MMD to match sharp, closely spaced peaks that would otherwise be smeared out.

E Results

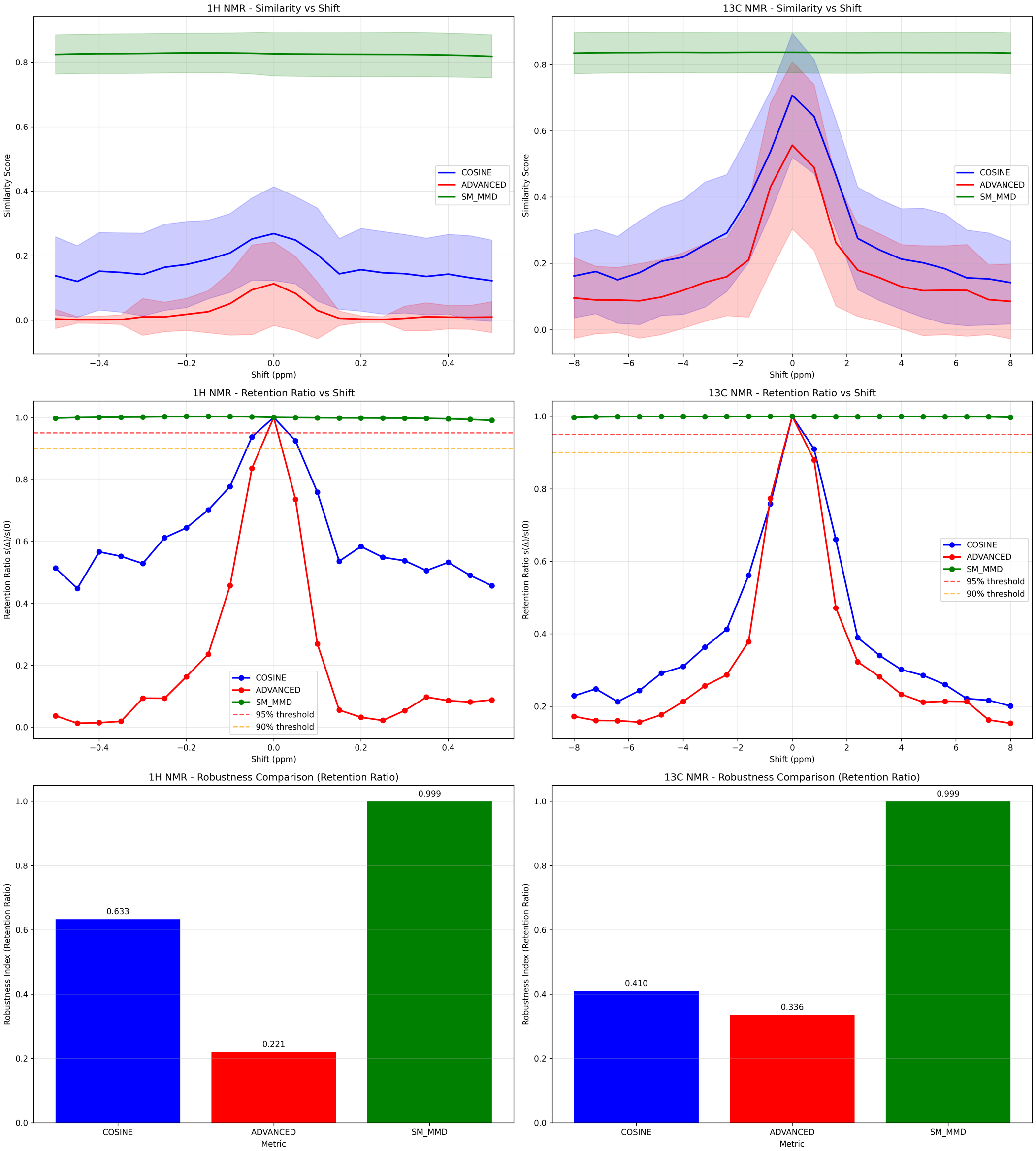
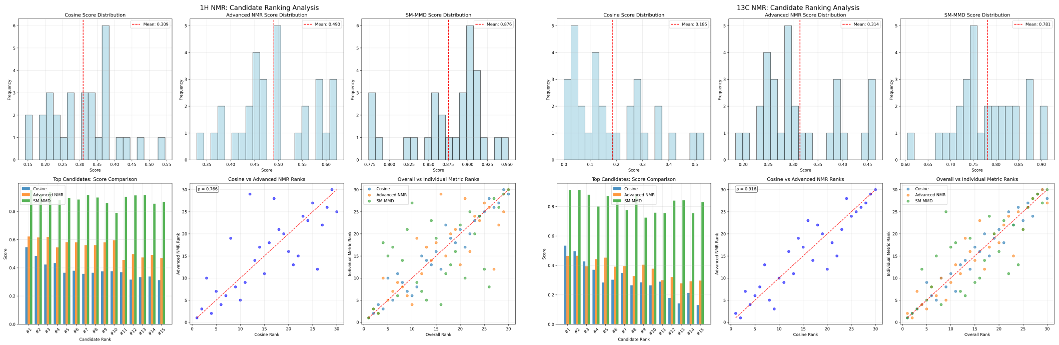


Figure 3: Shift–stress evaluation for ^{1}H and ^{13}C NMR: similarity vs. applied shift (top), retention ratio vs. shift (middle), and robustness index (retention ratio at 0.1 ppm for ^{1}H / 2 ppm for ^{13}C) comparison (bottom).



^{1}H NMR: Candidate Ranking Analysis

^{13}C NMR: Candidate Ranking Analysis

Figure 4: Candidate retrieval results for ^{1}H and ^{13}C NMR. Top row: score distributions (Cosine, Advanced NMR, SM-MMD). Bottom row: top-candidate score comparison and rank concordance plots.

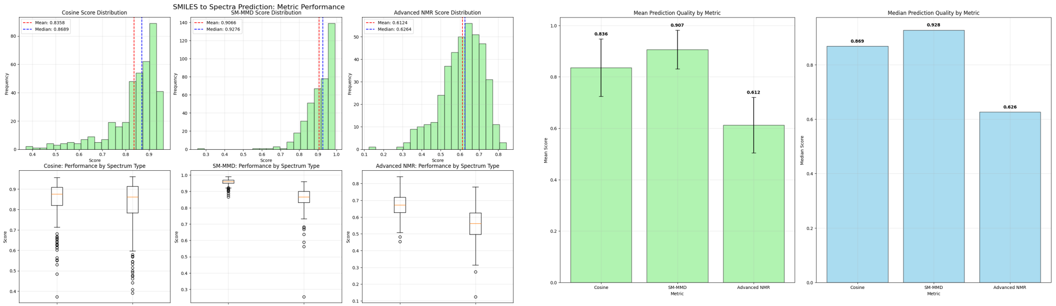


Figure 5: SMILES→Spectra prediction performance. Left: score distributions and per-modality boxplots for Cosine, SM-MMD, and Advanced NMR metrics. Right: bar charts summarizing mean (with SD) and median scores.

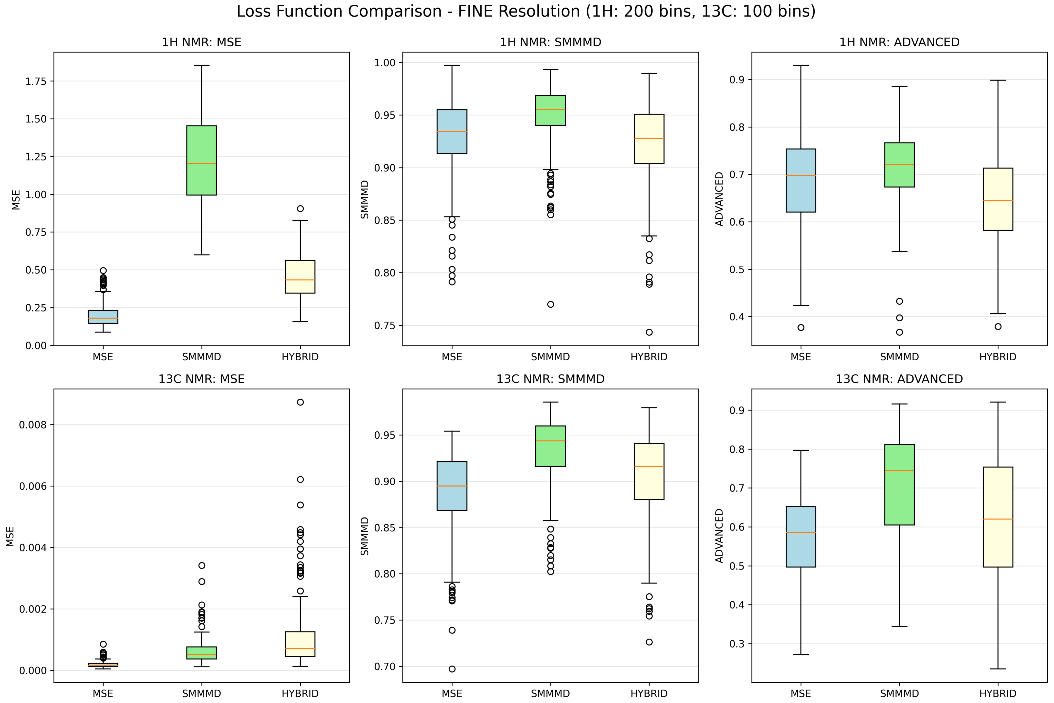


Figure 6: Training loss comparison for spectral prediction. Box plots showing model performance when trained with MSE, SM-MMD, or Hybrid loss for ^{1}H (top) and ^{13}C (bottom) NMR spectra. Columns correspond to different evaluation metrics: MSE (lower is better), SM-MMD, and Advanced NMR similarity (higher is better). Histograms use 200 bins for ^{1}H and 100 bins for ^{13}C, reflecting the finer peak structure of ^{1}H