SPECTRE: A Spectral Transformer for Molecule Identification

Wangdong Xu
University of California, San Diego
w6xu@ucsd.edu

Huanru Henry Mao
Jenni AI
henry@jenni.ai

Hyunwoo Kim
Dongguk University
hwkim8906@dongguk.edu

James Jialun Zhao
Carnegie Mellon University
jjzhao2@andrew.cmu.edu

Chen Zhang
University of California, San Diego
beowulf.zc@gmail.com

Byeol Ryu
University of California, San Diego
b2ryu@ucsd.edu

Yiran Xu
University of California, San Diego
y6xu@ucsd.edu

William H. Gerwick *
University of California, San Diego
wgerwick@ucsd.edu

Garrison W. Cottrell *
University of California, San Diego
gcottrell@ucsd.edu

Abstract

Nuclear Magnetic Resonance (NMR) spectroscopy is essential for identifying novel natural products. However, interpreting NMR spectra is time-consuming and requires expertise, leading to the development of computational tools for "structure annotation", which provides an ordered list of similar known molecules to speed up identification.

This work introduces SPECTRE, a state-of-the-art transformer-based model for structure annotation. Key contributions include 1) A novel, entropy-optimized Morgan fingerprint (MF) that can be adjusted for different NMR spectra types. 2) A lightweight, accurate structure annotation method, accepting flexible types of NMR input by Data type dropout (DTD), a new data augmentation technique to handle missing modalities for multi-modal models. As a result, SPECTRE achieves 95.79% accuracy, a 12.18% improvement over the previous SOTA.

Our code is available at here and the dataset is available at here . Unfortunately, we have to remove all the HSQC spectra from the dataset because of intellectual property issue.

1 Introduction

Research on new drugs from penicillin to a variety of anti-cancer compounds often begins with the discovery and analysis of natural products. Nuclear Magnetic Resonance (NMR) spectroscopy is pivotal to natural products' structure analysis. Chemists use both one-dimensional (1D) and two-dimensional (2D) NMR spectroscopy to solve molecular structures. This process typically starts with

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

simple 1D experiments to gain basic information and progresses to more complex 2D experiments to unravel detailed structural features. However, analyzing NMR spectra involves a considerable amount of human reasoning and a high level of expertise, making it a substantial bottleneck in the structure elucidation process. Structure annotation methods have been proved to expedite the identification process by offering clues to molecular structure based on measured similarity to known molecules [9].

Thus, the construction of molecular representations suitable for machine learning models plays an essential role in developing systems that assist chemists in this complex task. Notably, the Morgan Fingerprint (MF) [10] is a bit vector representation of molecules widely used in cheminformatics. Please refer to Appendix A for an explanation of MF's configuration parameters: radius and length.

Building on this foundation, our work introduces SPECTRE, a cutting-edge transformer-based model designed to aid chemists using NMR spectroscopy for molecular identification. We have developed a novel form of Morgan Fingerprints, aiming for the highest entropy representation. SPECTRE is able to predict these advanced fingerprints from NMR data and accurately identify molecules from a pool of over 3,881 candidates. This marks a significant leap forward, improving accuracy by 12.18% and reducing the parameter size by 93.94% over the previous state-of-the-art.

Furthermore, our model addresses one of the critical limitations encountered in prior work: the restriction of input data types. SPECTRE's versatility in accepting a wide array of NMR spectra types, from 2D ^{1}H–^{13}C HSQC NMR to 1D ^{1}H NMR and ^{13}C NMR, and any combination of them, substantially widens its applicability. This flexibility, the first in the field to our knowledge, enables chemists to leverage NMR data more comprehensively for molecule recognition, particularly in low-resource settings. Our proposed data type dropout method randomly removes some of the three types of input NMR types for each training example, leading SPECTRE to learn more robust molecule representations exploiting all three types of NMR spectra and accept molecules with only some particular NMR spectrum type available.

2 Related Work

Molecular identification and representation are crucial in cheminformatics. SMILES [12] offers a human-readable, machine-parsable string format representation and is used in our dataset as molecules' identifiers. Morgan Fingerprints(MFs) [10] map molecular structures to bit vectors and admit variations like the fingerprint proposed in DeepSAT [4](DeepSAT FP) which reduce collisions.

NMR-based structure annotation has rapidly advanced with the advent of deep learning [1, 4]. Alberts, Zipoli, and Vaucher predict SMILES from 1D NMR spectra and then compute MF for molecular retrieval. DeepSAT uses CNNs to process 2D ^{1}H–^{13}C HSQC spectra for classification, molecular weight prediction, and their variation of MFs. SPECTRE, however, is the first to accept flexible combinations of NMR inputs, including 1D ^{1}H, 1D ^{13}C, and 2D ^{1}H–^{13}C HSQC spectra.

Addressing missing modalities in multi-modal machine learning has been tackled by methods such as ModDrop [8] and ModDrop++ [6], which are CNNs where each channel represent a modality and missing modalities are handled by zeroing out channels. Cheerla and Gevaert use a multimodal autoencoder and drop entire feature vectors for missing modalities, rescaling the weights for input to the next stage. In contrast, our Data Type Dropout (DTD) uses delimiter tokens to combine all modalities as one input sequence, with adjacent <start><end> delimiters indicating missing modalities. ESM3 [3] uses BERT-style training to learn a model that can complete patterns of structure, sequence, and function, allowing the generation of new molecules with desired properties.

3 Dataset

In our work, we collected 2D ^{1}H–^{13}C HSQC NMR spectra from the JEOL database [7] and simulated some additional 2D ^{1}H–^{13}C HSQC NMR spectra using the ACD Labs' Spectrus Processor, whose license prevents us from releasing the simulated HSQC spectra to public. We in total have 2D HSQC spectra of 137,267 molecules, together with their chemical names, SMILES strings, and molecular weight. Meanwhile, we collected 1D ^{1}H NMR and 1D ^{13}C NMR spectra of 155,815 natural products from NP-MRD, the largest natural product NMR database. Duplicated molecules with stereodescriptors, which indicate a right-handed and left-handed stereocenter, are filtered out in our dataset. Combining the two sources, we created SPECTRE-DB (SDB), a dataset of 39,563 molecules with all three NMR spectra types, to compare different NMR combinations and choose their entropy-based MF configurations. The rest of the molecules, which have partial NMR types missing, are used to train models targeted to single-NMR-combination and DTD-based models.

4 Methods

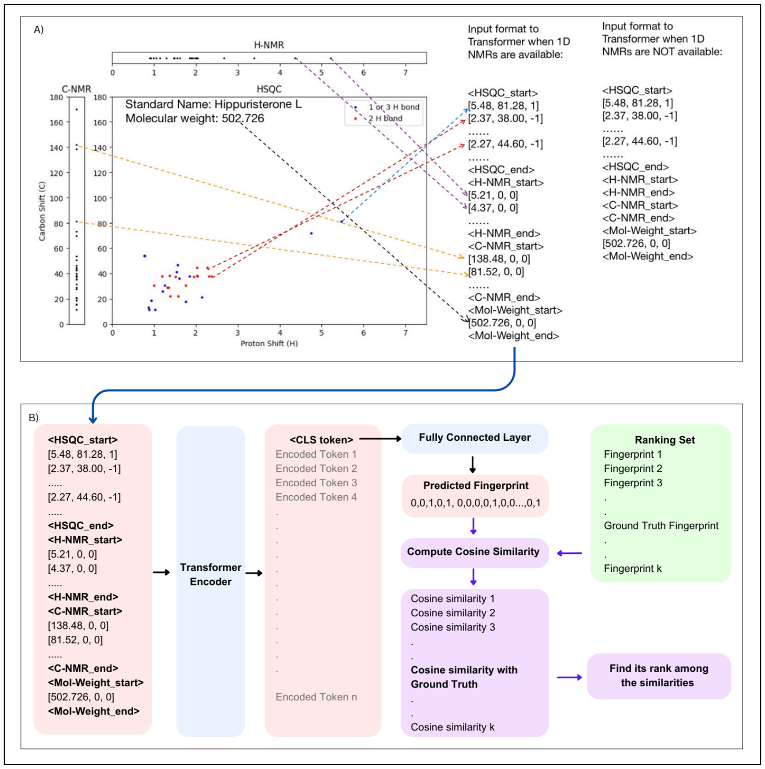

Figure 1: SPECTRE Overview. A) demonstrates how NMR spectra and molecular weight are encoded and provides an additional example when 1D NMRs are not available. Angle brackets indicate classification tokens and input source delimiter tokens. At the 3rd column of HSQC spectra, a negative sign means the carbon is bound with 1 or 3 hydrogen(s) and a positive sign means 2 hydrogens. B) shows that the input sequence is encoded by the transformer encoder and only the <CLS> token is used to predict MF. Molecules will be retrieved based on cosine similarity. The ground truths of NMR input are used only for measuring performance during validation and testing.

We use the transformer encoder architectures for the prediction of molecules' Morgan Fingerprints from their NMR spectra and molecular weights. Figure 1(a) illustrates the pre-processing of multi-sourced inputs across different scenarios, showcasing the flexibility of SPECTRE. For missing NMR types, <start> and <end> tokens are adjacently placed to indicate the absence of that modality.

Transformers have two advantages over CNNs. First, they handle inputs of varying lengths more effectively and can process multi-sourced data simultaneously because of their attention mechanism, making them ideal for incorporating molecular weight and handling missing NMR data using delimiters. Second, while CNNs process the entire 2D ^{1}H–^{13}C HSQC NMR spectra as images where most values are zero and therefore cause extreme inefficiency, transformers only use peak coordinates in the form of $(proton\_shift, carbon\_shift)$, reducing redundancy. This improvement is evident in model size and performance: SPECTRE has 10.3M parameters, converging in 40 epochs (80 seconds/epoch on an Nvidia RTX 3090), while our implementation of DeepSAT, a CNN-based model, has 173M parameters, needing 70 epochs (100 seconds/epoch) for convergence. This highlights the transformer's efficiency and suitability for complex chemical informatics tasks.

When training SPECTRE, we applied the Noam scheduler proposed in [11], batch size of 64, and Adam optimizer [5]. Our early-stop metric is the rank-1 score, which is explained in section 4.

4.1 Model's Prediction Target: Entropy-Based Morgan Fingerprint

A key limitation of traditional Morgan Fingerprints (usually with radius 2) is their redundancy, with many bits frequently zero. This reduces their ability to capture distinct molecular features. To address this, combining MFs of different radii has been proposed as a solution.

Kim, et al. [4] proposed "a 6144-bit Morgan Fingerprint created by concatenating three 2048-bit fingerprints for radii 0, 1, and 2". This method captures a wider range of chemical structures, but redundancy remains an issue, as many bits are still predominantly zero, indicating the need for further refinement.

We propose a novel entropy-based fingerprinting technique to reduce redundancy. Starting with traditional MFs for radii 0 to a specific value n (in our case, 15), we concatenate them to form a super-fingerprint of $(n + 1) * 6144$ bits for every molecule, capturing a broad range of structural features. For each bit in this vector, we compute its entropy across the training set, then sort them, keeping the 6,144 bits with the highest entropy. We form a more compact and informative version of fingerprint by picking the bits by the recorded indices. We refer to these as the R0 to R$_n$ entropy-based MF.

4.2 Model Architecture

Our model architecture is depicted as Figure 1(b). It consists of a transformer encoder, a classifier, and a retrieval system using cosine similarity for ranking the results.

The encoder has 8 transformer encoder layers with 8 attention heads. The embedding size of 384 is divided into 180, 180, and 24 for positional encoding of the 3 components of each token: <C-shift>, <H-shift>, and <sign>. The first embedding, the <cls> token, is passed to a fully connected layer and a sigmoid layer to predict a 6144-bit MF. Binary cross-entropy loss is used to update model weights.

During validation and testing, we de-duplicated the validation and test sets based on canonical SMILES strings to form candidate pools, called "ranking sets," consisting of 3,942 and 3,881 molecules, respectively. The model predicts a 6144-bit MF and cosine similarities between the predicted MF and each MF of the molecules from the ranking set are computed. A "Rank-k" score represents the probability that the predicted MF ranks higher than the k-th most similar in the set. The rank-1 score therefore indicates the accuracy of correctly identifying the input molecule.

4.3 Data Type Dropout(DTD) Technique

We deployed a new data-augmentation method named "Data Type Dropout" specifically for models to adapt for uncertain input modality availability. In particular, when the input sequence is constructed, each modality(2D ^{1}H–^{13}C HSQC NMR, 1D ^{1}H NMR, and 1D ^{13}C NMR spectra) are randomly omitted from the input sequences. Therefore, DTD effectively expands the dataset because each molecule is trained by a different input NMR combination at each different epoch. This approach ensures that the model becomes adept at interpreting all possible NMR input combinations, leading to capability to accept any combination of input modalities and better generalizability.

5 Experiments

In this section, we present our experiment results. For each model, we have 3 experiment trials and report their average performances. In addition, molecular weight, which is easily accessible in most chemistry labs, is always used as a part of model input when training SPECTRE models. Our evaluation is mainly based on rank-1 score. Appendix B displays more detailed experiment data and includes rank-5 score, mean rank, cosine similarity to ground truth, and F1-score of MF prediction, providing a comprehensive analysis of model performance.

5.1 Uncovering the Most Efficient Morgan Fingerprint for each Type of Input

This section explores experimentation with different entropy-based MFs as target outputs for each NMR input type, aiming to identify the most suitable MF for each and compare the identification capability of each type of NMR. Table 1 presents the results of comparing the rank-1 score of each

NMR input combination type using different MFs, including the R0-to-Rn MF (with n from 1 to 5) and DeepSAT FP. Experiments are conducted on the SPECTRE-DB dataset where all three types of NMR and molecular weight are accessible to ensure fair comparison. For more detailed comparison results, please refer to Appendix B.

We found that using multiple types of NMR data consistently improves ranking scores, whereas the previous SOTA method, DeepSAT, only uses 2D ^{1}H–^{13}C HSQC as input. Further, our entropy-based MFs outperform DeepSAT FP. Contrary to the belief that 2D ^{1}H–^{13}C HSQC spectra offer more structural information, results from the last three rows of Table 1 indicate that 1D ^{13}C-NMR is the most informative. This suggests that transformers may better utilize information from carbon atoms without hydrogen, which 2D spectra lack, compared to human analysis.

Rank-1 Score↑	R0-1 MF	R0-2 MF	R0-3 MF	R0-4 MF	R0-5 MF	DeepSAT FP
All 3 NMRs	90.82%	93.50%	93.92%	**94.12%**	93.71%	92.58%
C NMR and H NMR	89.26%	92.56%	92.80%	**92.86%**	92.24%	91.96%
HSQC and C NMR	89.56%	92.52%	92.81%	**92.95%**	92.36%	91.42%
HSQC and H NMR	85.05%	89.56%	89.99%	**90.04%**	89.58%	88.40%
Only C NMR	86.21%	89.52%	**89.66%**	89.58%	89.13%	88.29%
Only H NMR	77.17%	82.27%	82.06%	**82.44%**	82.04%	81.25%
Only HSQC	75.76%	81.37%	**81.63%**	81.61%	80.89%	79.11%

Table 1: Rank-1 score of each NMR input combination type using different MFs. Each value is measured on separate models trained with its corresponding NMR inputs and target MF.

5.2 Achieving Higher Performance by Data Type Dropout

The main advance of our model over previous state-of-the-art approaches is primarily due to the incorporation of multiple types of input NMRs. However, oftentimes not all three types of NMR spectra are available.

One solution is to train separate models for each NMR input combination using the most efficient MF identified in subsection 5.1, though model performance is limited by dataset size. We compare this approach with our Data Type Dropout (DTD) method (subsection 4.3), which augments the dataset by training the model with incomplete data. Using DTD, we train SPECTRE with the R0-R4 entropy-based MF, the most effective model experimentally. Each NMR type's dropout rate is adjusted for even distribution of each NMR type combination during training. We compare the effectiveness of these two strategies in Table 2 and find that DTD improves performance when multiple NMR types are used but reduces effectiveness when only a single NMR type is available. This suggests that DTD is an effective strategy for enhancing a model's ability to learn relations and interactions across multiple modalities.

	Model w/ DTD		Model w/o DTD	
Model Input	Rank-1↑	sampled rate	Rank-1↑	training set size
All 3 NMRs	95.79%	5.0%	94.12%	31,740
C NMR and H NMR	94.72%	17.5%	92.91%	152,981
HSQC and C NMR	94.86%	5.0%	92.83%	34,567
HSQC and H NMR	91.95%	5.0%	89.77%	31,747
Only C NMR	91.59%	17.5%	92.71%	155,808
Only H NMR	82.51%	17.5%	88.90%	152,988
Only HSQC	84.77%	32.5%	88.79%	109,694

Table 2: Performance comparison of models trained on all available NMR data. For single-NMR-combination models, each row represents separate models trained with the best-performing MF from subsection 5.1 and molecules filtered by the availability of corresponding NMR input combination. For models trained using the DTD technique, all NMR spectra are utilized, and dropout rates are adjusted to maximally balance the occurrence of each NMR combination, with the sampling rates indicated above. Each row represents model weights saved at different checkpoints, based on the rank-1 score of the corresponding NMR combination.

6 Conclusion

We introduced SPECTRE, a transformer-based approach for structure elucidation of natural products via NMR spectroscopy, leveraging a novel high-entropy Morgan Fingerprint. Our contributions include developing a robust molecular representation and a lightweight model that accurately predicts these fingerprints from flexible NMR spectra availability by the Data Type Dropout technique. SPECTRE surpasses the previous state-of-the-art, DeepSAT, in accuracy with far fewer parameters. This approach expedites structure identification, potentially accelerating pharmaceutical development.

7 Limitations and Future Work

A key limitation is the exclusion of solvent effects, as our dataset lacks solvent metadata, which could improve model accuracy. Incorporating high-quality Mass Spectrometer data is another potential enhancement. Interestingly, 1D ^{13}C-NMR achieves higher accuracy than 2D ^{1}H–^{13}C HSQC, contrary to chemists' expectations. Future work will explore why transformer models interpret NMR data differently than humans. We also aim to leverage deep learning for predicting SMILES strings from NMR data, advancing natural product identification and therapeutic development.

References

[1] M. Alberts, F. Zipoli, and A. C. Vaucher. “Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models”. In: *ChemRxiv* (2023). This content is a preprint and has not been peer-reviewed. DOI: 10.26434/chemrxiv-2023-8wxcz. URL: https://doi.org/10.26434/chemrxiv-2023-8wxcz.

[2] Anusha Cheerla and Olivier Gevaert. “Deep learning with multimodal representation for pancancer prognosis prediction”. In: *Bioinformatics* 35.14 (2019), pp. i446–i454. DOI: 10.1093/bioinformatics/btz342. URL: https://doi.org/10.1093/bioinformatics/btz342.

[3] Thomas Hayes et al. “Simulating 500 million years of evolution with a language model”. In: *bioRxiv* (2024). DOI: 10.1101/2024.07.01.600583. URL: https://doi.org/10.1101/2024.07.01.600583.

[4] H.W. Kim, C. Zhang, R. Reher, et al. “DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data”. In: *J Cheminform* 15 (2023), p. 71. DOI: 10.1186/s13321-023-00738-4. URL: https://doi.org/10.1186/s13321-023-00738-4.

[5] Diederik P Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).

[6] Heng Liu et al. “ModDrop++: A Dynamic Filter Network with Intra-subject Co-training for Multiple Sclerosis Lesion Segmentation with Missing Modalities”. In: *ArXiv* (2022). URL: https://arxiv.org/abs/2203.04959.

[7] JEOL Ltd. *Natural Product NMR-DB ’CH-NMR-NP’*. 2024. URL: https://ch-nmr-np.jeol.co.jp/en/nmrdb/.

[8] Natalia Neverova et al. “ModDrop: Adaptive multi-modal gesture recognition”. In: *ArXiv* (2014). URL: https://arxiv.org/abs/1501.00102.

[9] R. Reher et al. “A Convolutional Neural Network-Based Approach for the Rapid Annotation of Molecularly Diverse Natural Products”. In: *Journal of the American Chemical Society* 142.9 (2020), pp. 4114–4120. DOI: 10.1021/jacs.9b13786. URL: https://doi.org/10.1021/jacs.9b13786.

[10] D. Rogers and M. Hahn. “Extended-Connectivity Fingerprints”. In: *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754. DOI: 10.1021/ci100050t. URL: https://doi.org/10.1021/ci100050t.

[11] A. Vaswani et al. “Attention Is All You Need”. In: (2017). arXiv:1706.03762. URL: https://arxiv.org/abs/1706.03762.

[12] D. Weininger. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”. In: *Journal of Chemical Information and Computer Sciences* 28.1 (1988), pp. 31–36. DOI: 10.1021/ci00057a005. URL: https://doi.org/10.1021/ci00057a005.

[13] Shifa Zhong and Xiaohong Guan. “Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants’ Activities and Properties”. In: *ACS Publications* (2023). DOI: 10.1021/acs.est.3c02198.s002. URL: https://doi.org/10.1021/acs.est.3c02198.s002.

A Appendix1: Morgan Fingerprint Explanation in Detail

Morgan Fingerprint(MF) encodes molecular structures into bit vectors and hence the cosine between two MFs is a way to measure the similarity of structure. MFs have seen widespread application in cheminformatics. Its algorithm breaks down a molecule into fragments based on the radius of molecular bonds, compares the fragments against a pre-compiled fragments library, and converts the results into a set of bit vectors by using a hash function, allowing for possible collisions. The algorithm has two parameters: the length of the vector (hash table size) and the radius: These fingerprints start at radius 0, which maps all atoms into the vector, and expand to radius 1, mapping the structure of each atom's immediate one-bond neighborhood into the bit vector, and so on until it finishes the specified radius.

Count-based MFs include a count of each structure rather than just a binary entry [13]. However, one cannot reproduce the molecule from its MF. In other words, MF can be calculated based on SMILES string but one cannot derive the SMILES string from MF.

B Appendix2: Most efficient entropy-based Fingerprint for each input combination type

Here we present the results of training SPECTRE with different target outputs. All molecules in the training, validation, and test set have all three types of NMR: 2D ^{1}H–^{13}C HSQC, 1D ^{13}C NMR, 1D ^{1}H NMR.

Notably, when the radii go beyond 4, the performance drops along with the radius increasing. This is because as the radius increases, the presence of each active has a more ambiguous meaning since it can result from a molecule subgroup with any size of radius between 0 to the selected radius. Therefore, inspired by DeepSAT FP, we also created another version of an entropy-based fingerprint where each active bit corresponds to a molecule subgroup with a specific radius. However, this method doesn't show any experimental success and we hypothesize that this method leads to the problem of sparsity again and has suboptimal performance.

While we aim to maximize MF entropy, the fingerprints still consist of mostly zeros, with entropy increasing as more radii are included. Therefore, comparing F1 scores across models using the same MF is more meaningful than comparing across different MFs. For instance, the R0-R1 MF has the lowest entropy, which means it contains the most zeros as it hashes only substructures with radii 0 and 1. Therefore, it has high F1 scores in despite of low rank-1/rank-5 scores.

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	90.82%	97.57%	5.81	0.9353	0.9374
R0 to R2 MF	93.50%	98.48%	4.42	0.8823	0.8886
R0 to R3 MF	93.92%	98.66%	2.78	0.8506	0.8589
R0 to R4 MF	**94.12%**	98.76%	3.65	0.8377	0.8473
R0 to R5 MF	93.71%	98.78%	2.91	0.8371	0.8453
R0 to R6 MF	93.26%	98.63%	2.98	0.8321	0.8416
R0 to R7 MF	93.47%	**98.82%**	3.51	0.8420	0.8501
R0 to R8 MF	93.10%	98.65%	**2.19**	0.8462	0.8541
R0 to R9 MF	92.93%	98.55%	2.55	0.8466	0.8546
R0 to R10 MF	92.64%	98.51%	3.03	0.8469	0.8557
R0 to R11 MF	92.41%	98.53%	3.14	0.8489	0.8582
R0 to R12 MF	92.43%	98.56%	2.88	0.8561	0.8647
R0 to R13 MF	92.17%	98.15%	3.08	0.8583	0.8667
R0 to R14 MF	91.88%	98.34%	2.68	0.8626	0.8707
R0 to R15 MF	91.82%	98.50%	2.77	0.8663	0.8744
DeepSAT FP	92.58%	98.29%	3.41	0.8297	0.8389

Table 3: Performance of various MFs when all three NMRs are available

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	89.26%	96.48%	8.38	0.9305	0.9337
R0 to R2 MF	92.56%	97.65%	5.37	0.8784	0.8860
R0 to R3 MF	92.80%	97.98%	4.35	0.8467	0.8570
R0 to R4 MF	**92.86%**	**98.00%**	**3.75**	0.8368	0.8473
R0 to R5 MF	92.24%	97.77%	4.76	0.8307	0.8420
R0 to R6 MF	92.09%	97.80%	4.63	0.8334	0.8443
R0 to R7 MF	92.14%	97.94%	4.84	0.8418	0.8516
R0 to R8 MF	91.79%	97.87%	5.04	0.8436	0.8535
R0 to R9 MF	91.77%	97.69%	4.32	0.8444	0.8547
R0 to R10 MF	91.56%	97.82%	4.49	0.8453	0.8558
R0 to R11 MF	91.37%	97.59%	5.25	0.8477	0.8589
R0 to R12 MF	90.83%	97.41%	4.95	0.8513	0.8618
R0 to R13 MF	90.89%	97.44%	4.91	0.8567	0.8671
R0 to R14 MF	90.80%	97.62%	4.70	0.8612	0.8717
R0 to R15 MF	90.26%	97.48%	5.89	0.8644	0.8748
DeepSAT FP	91.96%	97.58%	4.55	0.8291	0.8403

Table 4: Performance of various MFs when only 1D ^{13}C-NMR and ^{1}H-NMR are available

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	89.56%	96.96%	5.84	0.9267	0.9293
R0 to R2 MF	92.52%	97.83%	5.15	0.8760	0.8830
R0 to R3 MF	92.81%	98.30%	3.43	0.8432	0.8525
R0 to R4 MF	**92.95%**	98.52%	3.16	0.8319	0.8415
R0 to R5 MF	92.36%	98.28%	3.14	0.8294	0.8392
R0 to R6 MF	92.11%	98.13%	3.09	0.8313	0.8403
R0 to R7 MF	92.14%	98.40%	3.15	0.8340	0.8430
R0 to R8 MF	91.87%	98.08%	**2.74**	0.8389	0.8480
R0 to R9 MF	91.79%	98.21%	3.15	0.8408	0.8496
R0 to R10 MF	91.43%	98.02%	3.29	0.8395	0.8490
R0 to R11 MF	91.12%	97.97%	3.54	0.8422	0.8520
R0 to R12 MF	91.04%	97.98%	3.34	0.8475	0.8568
R0 to R13 MF	90.89%	97.97%	**2.74**	0.8516	0.8610
R0 to R14 MF	91.06%	97.89%	3.41	0.8564	0.8654
R0 to R15 MF	90.41%	97.79%	4.01	0.8608	0.8698
DeepSAT FP	91.42%	97.81%	3.96	0.8166	0.8265

Table 5: Performance of various MFs when 2D ^{1}H–^{13}C HSQC and 1D ^{13}C-NMR are available

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	85.05%	95.11%	9.86	0.9108	0.9136
R0 to R2 MF	89.56%	96.69%	6.85	0.8576	0.8647
R0 to R3 MF	89.99%	97.20%	6.25	0.8247	0.8339
R0 to R4 MF	**90.04%**	**97.22%**	4.58	0.8194	0.8281
R0 to R5 MF	89.58%	97.11%	5.12	0.8137	0.8234
R0 to R6 MF	88.87%	97.08%	4.85	0.8175	0.8263
R0 to R7 MF	88.61%	96.88%	4.68	0.8199	0.8288
R0 to R8 MF	88.27%	96.97%	5.37	0.8227	0.8316
R0 to R9 MF	88.27%	97.08%	3.96	0.8237	0.8324
R0 to R10 MF	88.32%	96.63%	5.76	0.8220	0.8320
R0 to R11 MF	87.35%	96.52%	4.82	0.8222	0.8325
R0 to R12 MF	87.13%	96.61%	5.67	0.8288	0.8389
R0 to R13 MF	87.33%	96.45%	5.71	0.8331	0.8434
R0 to R14 MF	87.27%	96.51%	4.69	0.8389	0.8485
R0 to R15 MF	86.73%	96.40%	5.81	0.8415	0.8512
DeepSAT FP	88.40%	96.41%	6.19	0.8022	0.8119

Table 6: Performance of various MFs 2D ^{1}H–^{13}C HSQC and 1D ^{13}H-NMR are available

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	86.21%	94.15%	15.86	0.9184	0.9229
R0 to R2 MF	89.52%	95.78%	11.26	0.8650	0.8751
R0 to R3 MF	**89.66%**	**96.29%**	9.84	0.8312	0.8443
R0 to R4 MF	89.58%	96.27%	9.79	0.8203	0.8336
R0 to R5 MF	89.13%	96.05%	10.10	0.8192	0.8324
R0 to R6 MF	89.02%	96.17%	9.40	0.8209	0.8338
R0 to R7 MF	89.03%	96.07%	9.48	0.8235	0.8367
R0 to R8 MF	87.96%	96.16%	9.69	0.8296	0.8416
R0 to R9 MF	87.90%	95.90%	8.36	0.8305	0.8431
R0 to R10 MF	88.07%	95.87%	9.45	0.8274	0.8418
R0 to R11 MF	87.45%	95.62%	10.87	0.8307	0.8450
R0 to R12 MF	87.47%	95.74%	9.72	0.8367	0.8502
R0 to R13 MF	87.57%	95.84%	**8.38**	0.8393	0.8527
R0 to R14 MF	87.36%	95.67%	9.20	0.8437	0.8568
R0 to R15 MF	87.10%	95.47%	9.13	0.8514	0.8649
DeepSAT FP	88.29%	95.41%	10.92	0.8114	0.8252

Table 7: Performance of various MFs when only 1D ^{13}C-NMR is available

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	77.17%	89.97%	33.08	0.8888	0.8940
R0 to R2 MF	82.27%	92.84%	25.78	0.8305	0.8412
R0 to R3 MF	82.06%	93.25%	18.84	0.7989	0.8124
R0 to R4 MF	**82.44%**	**93.47%**	17.44	0.7917	0.8057
R0 to R5 MF	82.04%	93.30%	16.40	0.7907	0.8039
R0 to R6 MF	81.64%	92.97%	18.17	0.7910	0.8045
R0 to R7 MF	81.50%	93.22%	16.91	0.7981	0.8108
R0 to R8 MF	80.92%	92.83%	17.93	0.8009	0.8139
R0 to R9 MF	80.59%	92.61%	18.99	0.7991	0.8127
R0 to R10 MF	80.67%	92.66%	18.14	0.7977	0.8125
R0 to R11 MF	80.29%	92.33%	**16.22**	0.8014	0.8161
R0 to R12 MF	79.89%	92.21%	19.01	0.8053	0.8203
R0 to R13 MF	79.28%	92.01%	20.75	0.8095	0.8241
R0 to R14 MF	79.26%	91.91%	18.95	0.8140	0.8283
R0 to R15 MF	79.24%	91.83%	17.80	0.8186	0.8331
DeepSAT FP	81.25%	92.49%	20.08	0.7766	0.7909

Table 8: Performance of various MFs when only 1D ^{1}H-NMR is available

Model Input	Rank-1↑	Rank-5↑	Mean Rank↓	Cosine Sim↑	F1-score↑
R0 to R1 MF	75.76%	90.10%	19.72	0.8781	0.8813
R0 to R2 MF	81.37%	93.11%	12.28	0.8168	0.8236
R0 to R3 MF	**81.63%**	**93.91%**	11.62	0.7914	0.8010
R0 to R4 MF	81.61%	93.81%	10.13	0.7852	0.7946
R0 to R5 MF	80.89%	93.62%	9.90	0.7813	0.7912
R0 to R6 MF	79.84%	93.43%	10.22	0.7805	0.7905
R0 to R7 MF	79.59%	93.04%	9.78	0.7844	0.7943
R0 to R8 MF	79.07%	92.73%	10.69	0.7888	0.7986
R0 to R9 MF	77.93%	92.84%	**9.17**	0.7949	0.8037
R0 to R10 MF	78.10%	92.62%	11.05	0.7991	0.8081
R0 to R11 MF	77.96%	92.51%	10.58	0.8026	0.8119
R0 to R12 MF	77.82%	92.29%	10.07	0.7930	0.8038
R0 to R13 MF	77.31%	92.12%	11.53	0.7949	0.8062
R0 to R14 MF	76.71%	91.54%	10.35	0.7985	0.8096
R0 to R15 MF	76.07%	91.42%	11.00	0.8032	0.8144
DeepSAT FP	79.11%	92.53%	13.36	0.7569	0.7660

Table 9: Performance of various MFs when only 2D ^{1}H–^{13}C HSQC is available