2DNMRGym: An Annotated Experimental Dataset for Atom-Level Molecular Representation Learning in 2D NMR via Surrogate Supervision

Yunrui Li[*]
Brandeis University
yunruili@brandeis.edu

Hao Xu[*]
Harvard Medical School
haxu@bwh.harvard.edu

Pengyu Hong
Brandeis University
hongpeng@brandeis.edu

Abstract

Two-dimensional (2D) Nuclear Magnetic Resonance (NMR) spectroscopy, particularly Heteronuclear Single Quantum Coherence (HSQC) spectroscopy, plays a critical role in elucidating molecular structures, interactions, and electronic properties. However, accurately interpreting 2D NMR data remains labor-intensive and error-prone, requiring highly trained domain experts, especially for complex molecules. Machine Learning (ML) holds significant potential in 2D NMR analysis by learning molecular representations and recognizing complex patterns from data. However, progress has been limited by the lack of large-scale and high-quality annotated datasets. In this work, we introduce **2DNMRGym**, the first annotated experimental dataset designed for ML-based molecular representation learning in 2D NMR. It includes over 22,000 HSQC spectra, along with the corresponding molecular graphs and SMILES strings. Uniquely, 2DNMRGym adopts a surrogate supervision setup: models are trained using algorithm-generated annotations derived from a previously validated method and evaluated on a held-out set of human-annotated gold-standard labels. This enables rigorous assessment of a model's ability to generalize from imperfect supervision to expert-level interpretation. We provide benchmark results using a series of 2D and 3D GNN and GNN transformer models, establishing a strong foundation for future work. 2DNMRGym supports scalable model training and introduces a chemically meaningful benchmark for evaluating atom-level molecular representations in NMR-guided structural tasks. Our data and code is open-source and available at: https://github.com/siriusxiao62/2DNMRGym.

1 Introduction

1.1 Overview

Nuclear Magnetic Resonance (NMR) spectroscopy is a powerful technique that uses the magnetic properties of atomic nuclei to provide detailed insights into the structure and dynamics of chemical compounds [1–3]. It can determine the types, quantities, and spatial arrangements of atoms within molecules and their surrounding chemical environments, from small molecules to material polymers and complex bio-macromolecules. In NMR spectrum analysis, chemists utilize prediction tools to generate chemical shifts from molecular structures, comparing them with experimental values to verify structural assignments. This comparison aids in assessing the accuracy of proposed molecular structures and provides insights into the electronic and spatial environments of atoms within the molecule.

Among NMR techniques, Heteronuclear Single Quantum Coherence (HSQC) spectroscopy [4] stands out as a powerful two-dimensional (2D) Nuclear Magnetic Resonance (NMR) method that has

[*]Equal contribution.

Y. Li et al., 2DNMRGym: An Annotated Experimental Dataset for Atom-Level Molecular Representation Learning in 2D NMR via Surrogate Supervision. *Proceedings of the Fourth Learning on Graphs Conference (LoG 2025)*, PMLR 269, Arizona State University, Phoenix, USA, December 10–12, 2025.

become indispensable for the structural elucidation of complex molecules, especially when traditional one-dimensional (1D) NMR techniques are insufficient [5, 6]. By correlating the chemical shifts of proton nuclei with those of heteronuclei, typically ^{13}C or ^{15}N, via scalar coupling interactions, HSQC enables the precise mapping of interatomic linkages within molecular frameworks. This method is particularly valuable for identifying connectivity patterns between protons and adjacent heteronuclei, thereby providing critical insights into chemical bonding, stereochemistry, and three-dimensional molecular conformation.

Despite recent advancements in the prediction of 1D NMR spectra [7–10] and peak assignment [11], the application of machine learning techniques to 2D NMR, such as HSQC spectra prediction, remains constrained by the scarcity of annotated datasets for training. To the best of our knowledge, no large-scale annotated dataset of experimental HSQC spectra is currently available for training machine learning models. This is primarily due to the significant bottleneck in acquiring, processing, and annotating 2D NMR data. Acquiring HSQC spectra is time-consuming, requires highly sensitive instrumentation, and depends on the availability of pure samples at an appropriate concentration, making the process highly labor-intensive. Typically, a research group can only produce 10-20 high-quality spectra per week. Furthermore, the complexity of molecular structures leads to spectral overlap and signal degeneracy, complicating peak resolution. The presence of multiple chiral centers in molecules can further complicate annotations (see Appendix A for annotation visualizations). This process currently relies heavily on expert interpretation and domain knowledge, often requiring trained chemists with advanced degrees (eg. PhD) and years of experience. Expermental conditions also play a critical role in determining the quality of HSQC spectra. Consequently, the requirement for expensive instruments, labor-intensive sample preparation, and specialized expertise in organic chemistry severely limit the availability of large, annotated datasets.

Since there are no widely adopted, automated solutions exist that provide both accurate peak prediction and atom-level annotation, deep learning methods are well-suited for this task, as it can capture complex molecular patterns and interactions from data. The goal is twofold: (1) to overcome the practical limitations of expert-only interpretation by automating annotation, and (2) to improve the accuracy and scalability of spectral prediction, which can significantly accelerate molecular analysis pipelines. To adapt many state-of-the-art deep learning architecture in this domain, a high-quality, large-scale and annotated HSQC dataset is in great demand. To fill this gap, we introduce the 2DNMRGym dataset (illustrated in Figure 1), including 22,348 experimental HSQC spectra. Among these, 21,869 HSQC spectra with 33,8370 cross peaks were annotated using a recently published algorithm [12] and 479 spectra with 7,310 peaks were manually annotated and cross-validated by three domain experts. Each spectrum includes cross peaks annotated with their corresponding molecular graphs, enabling supervised training and systematic evaluation of models for HSQC peak prediction. What distinguishes 2DNMRGym is its dual-layer annotation strategy: the large-scale algorithm-generated annotations serve as silver-standard supervision for model training, while the expert-labeled subset provides a gold-standard benchmark to evaluate model robustness and generalization. This setup uses surrogate and abundant training labels to enable deep learning methods, and the high quality evaluation dataset to assess the ability of a model to learn meaningful molecular representations at the atom level. As such, the dataset offers a benchmark for existing and future GNN architectures in atom-level representation learning tasks.

In summary, the contribution of this work includes:

- Atom-level annotations and high-quality surrogate/gold-standard labels, which are not available in the source datasets;
- A fine-grained atom-level prediction task, which goes beyond standard graph-level benchmarks and promotes richer molecular representation learning;
- An innovative surrogate learning framework that enables large-scale training and fair model comparison, using silver labels (algorithmic) for broad coverage and golden labels (expert) as a trusted benchmark. This approach scales data while ensuring unbiased evaluation;
- A fully experimental benchmark based on HSQC spectra, which is unique in scope and scale, and to our knowledge, not previously available in the literature.

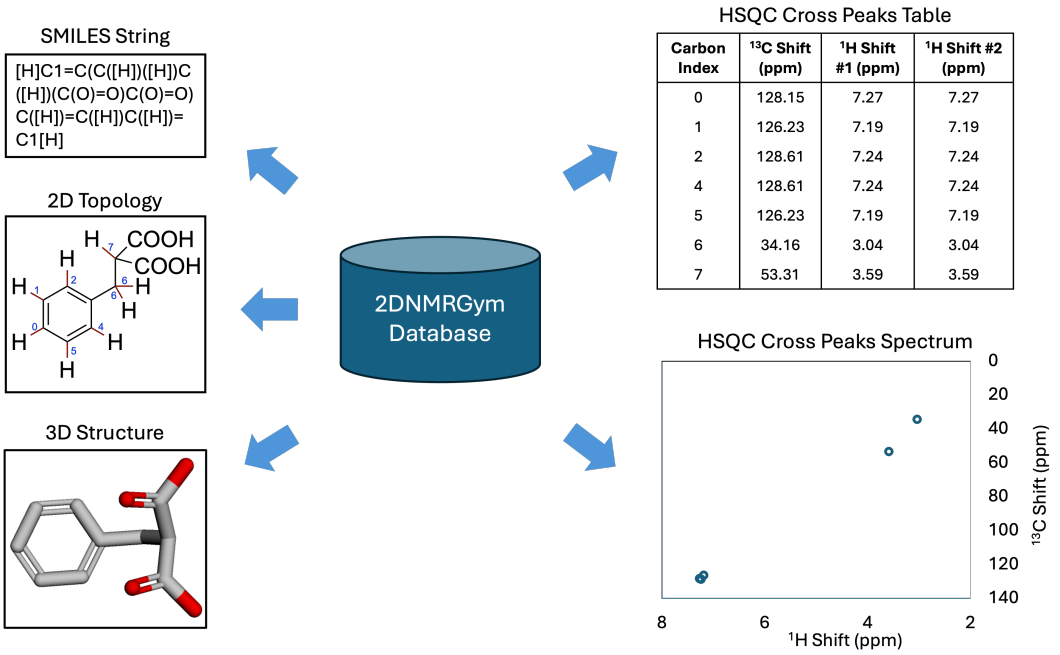

Figure 1: The 2DNMRGym dataset comprises multi-modal components, including the SMILES representation of each molecule and its conversion to a molecular graph. This graph includes both 2D topological structures and Cartesian coordinates for 3D spatial information. The ground truth spectrum is represented as cross peak tables, where the "Carbon Index" maps to the corresponding carbons in the molecular topology graph.

1.2 Concepts and terminology in chemistry

SMILES. Simplified Molecular Input Line Entry System (SMILES) [13] is a textual representation that employs short ASCII strings to describe chemical molecular structures. This notation system utilizes a series of characters, including alphanumeric symbols and punctuation marks, to represent the atoms, bonds, and connectivity within a molecule.

Chemical shift. Chemical shift is a measure of the resonant frequency of a nucleus relative to a reference standard, expressed in parts per million (ppm), and reflects the electronic environment surrounding the nucleus. In NMR spectroscopy, ^{1}H chemical shifts typically range from 0 to 12 ppm, while ^{13}C chemical shifts span a broader range, from 0 to 220 ppm, due to greater variation in carbon bonding environments. These shifts provide critical information about molecular structure, such as hybridization states, functional groups, and local electron density.

HSQC. HSQC [4] is a 2D NMR spectroscopy technique used to elucidate the structure of molecules by correlating the chemical shifts of hydrogen atoms with those of directly bonded heteronuclei, typically carbon or nitrogen. This technique provides detailed insights into molecular connectivity and is particularly useful for studying complex organic compounds where traditional 1D NMR spectroscopy may not provide sufficient information. HSQC is instrumental in identifying atom-to-atom connections and understanding the molecular architecture of a substance.

Tanimoto similarity. Tanimoto similarity is a widely used metric in cheminformatics for comparing molecular fingerprints, which are typically represented as binary vectors [14]. It quantifies the structural similarity between two molecules based on the presence or absence of shared substructures.

Scaffold. Scaffold refers to the core structural backbone of a molecule, typically consisting of the ring systems and the connecting linkers, with side chains and substituents removed. It represents the central topology that defines a molecule's overall shape and connectivity. In cheminformatics, scaffolds are often used to group molecules by structural similarity and to assess model generalization; for example, Bemis–Murcko scaffolds [15] are commonly used to analyze scaffold diversity and enable tasks like scaffold splitting in molecular datasets.

Hybridization. Hybridization refers to the combination of atomic orbitals (e.g., sp^3, sp^2, sp) to form new orbitals, which dictate the geometry of chemical bonds around an atom. This process affects both the electron distribution and the local chemical environment, factors that are crucial in determining NMR chemical shifts.

Chirality. Chirality is a molecular property where a compound exists as non-superimposable mirror images, usually due to a carbon atom bonded to four different substituents. This stereochemical feature affects the three-dimensional arrangement of atoms, which in turn influences the NMR signals, particularly in chiral environments.

2 Related work

The landscape of NMR databases exhibits a significant disparity in development and structure between 1D and 2D NMR spectra. For instance, the nmrshiftdb2 [16] dataset provides a comprehensive collection of 1D data, serving as an open-access platform for the sharing of chemical shift information. This database is highly structured and extensively utilized across the computational chemistry community, making it a valuable resource for researchers. In contrast, databases that catalog 2D NMR spectra, such as those for HSQC, exhibit less cohesion and a greater degree of specialization, often tailored to specific sub-realms or applications within the field. The Human Metabolome Database (HMDB) [17], for example, is a rich resource that includes detailed HSQC spectra for thousands of metabolites, coupled with extensive metadata on their structures, biochemical properties, and roles in biological systems. This makes HMDB a vital tool for metabolomics research, aiding in the identification and detailed analysis of metabolites across various biological samples. Another dataset, CH-NMR-NP [18], focuses on natural products and provides essential NMR spectral data, including HSQC spectra, for studying complex organic compounds. This dataset supports researchers in chemistry and biology by providing insights into the structure and potential applications of natural products, thus advancing the understanding of their biochemical pathways and therapeutic potentials. These specialized databases are not only repositories of NMR spectra but also rich sources of varied molecular dynamics and functional groups. Each database captures a unique slice of the chemical universe, encompassing a broad spectrum of molecular structures, which are represented as diverse graphs of varying sizes and complexities. This diversity is crucial for the development and evaluation of machine learning techniques, especially in the fields of computational chemistry and bioinformatics. While valuable, these databases were not designed with machine learning tasks in mind and lack the structured annotations necessary for supervised learning. Last year, a multimodal spectroscopic dataset for Chemistry [19] is published, comprising simulated ^{1}H-NMR, ^{13}C-NMR, HSQC-NMR, Infrared, and Mass spectra (positive and negative ion modes) for 790k molecules extracted from chemical reactions in patent data. However, only limited experimental data was collected and provided, among which no experimental HSQC-NMR is included. This highlights the challenge of collecting and annotation experimental HSQC-NMR data at scale.

Furthermore, most existing ML models such as GCN [20], GIN [21], GAT [22], GNN Transformer [23], ComENet [24] and SchNet [25] are trained at the molecule (graph-level) using coarse labels such as molecular properties using datasets like MolecularNet [26], QMugs [27], GEOM [28] etc., rather than capturing the finer atom-level interactions, as required in analyzing NMR spectra. Prior datasets rarely support this granularity, and those that do often rely on simulated data derived from quantum chemistry rather than real experimental spectra.

To address this gap, we introduce 2DNMRGym, a comprehensive, unified repository for experimental 2D NMR data. Unlike previous datasets, 2DNMRGym provides atom-level annotations, linking each cross peak to a specific hydrogen–heteronucleus bond within a molecular graph. The annotation process is labor-intensive and requires expert-level understanding of NMR and organic chemistry. To scale this effort, we adopt a dual-labeling strategy, combining algorithm-generated pseudo labels with a human-annotated subset for evaluation. This enables a unique atom-level representation learning task using surrogate supervision, where models are trained on imperfect algorithmic labels and evaluated against expert-labeled ground truth. In doing so, 2DNMRGym advances beyond traditional molecular fingerprinting and graph-level tasks, offering a new benchmark for fine-grained, chemically grounded prediction that bridges NMR spectroscopy and machine learning. This one-stop resource aims to streamline access and analysis of two-dimensional NMR spectra across various chemical contexts.

3 Constructing the 2DNMRGym dataset

Our 2DNMRGym dataset consists of over 22,000 HSQC spectra, where a small subset of 479 molecules with 7,310 cross peaks were randomly sampled for expert annotation as a held-out test set for evaluation.

Figure 2 summarizes key statistics of the training and test sets, which exhibit similar distributions in terms of total atom count, molecular weight, and Tanimoto similarity, indicating that the test set fairly represents the broader dataset and supports robust model evaluation. On average, molecules contain 58 atoms and have a molecular weight of approximately 400 Daltons. Over 25% of the molecules exceed 75 atoms and 500 Daltons in weight. The Tanimoto similarity plot reveals that most molecule pairs have a similarity score below 0.1, highlighting the structural diversity of the dataset.

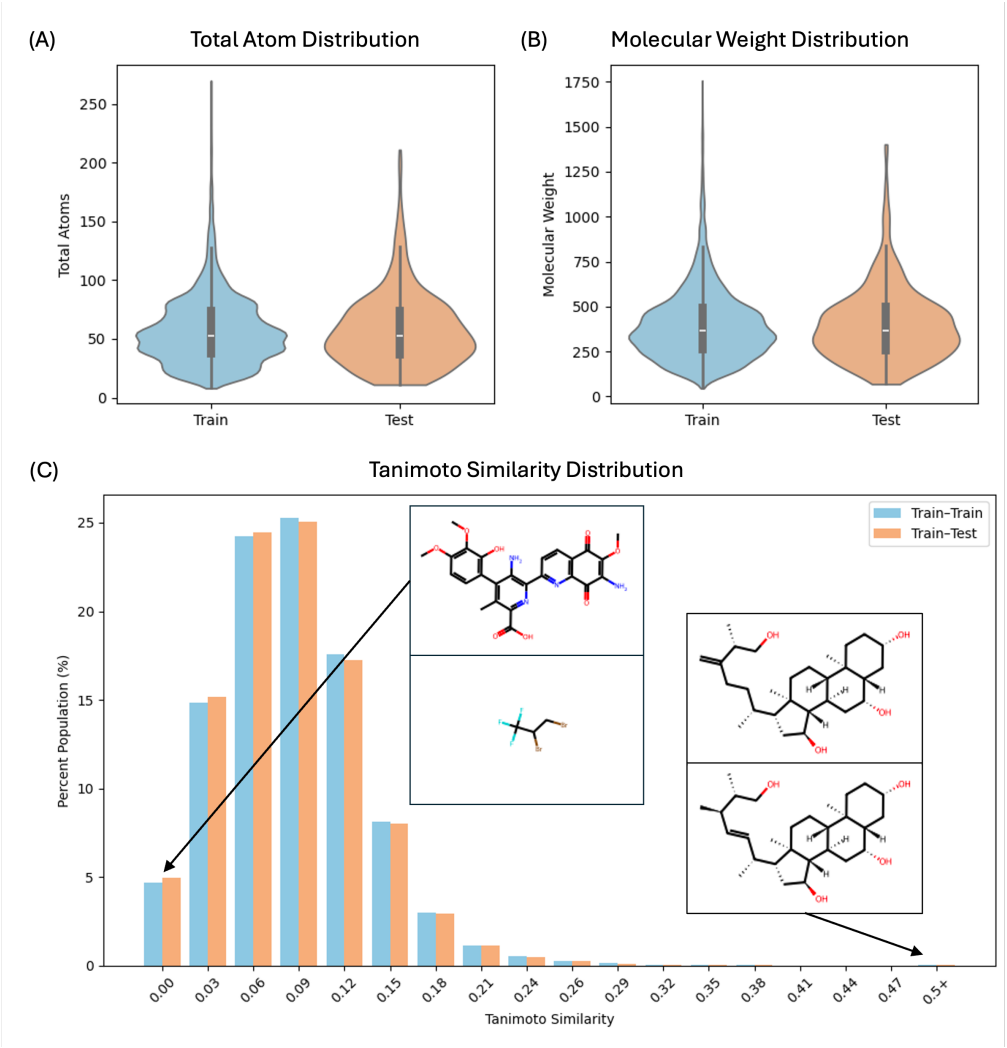

Figure 2: Data statistics by number of atoms, molecular weight, and tanimoto similarity.

To enable few-shot and zero-shot learning, we performed scaffold analysis for both the training and testing dataset. The test dataset contains 397 unique scaffolds, 148 of which are novel scaffolds that can be used for zero-shot learning. For scaffolds that appeared less than 10 times in the training set, they are used for few-shot learning. Figure 3 summarizes the distribution and top scaffolds in the data.

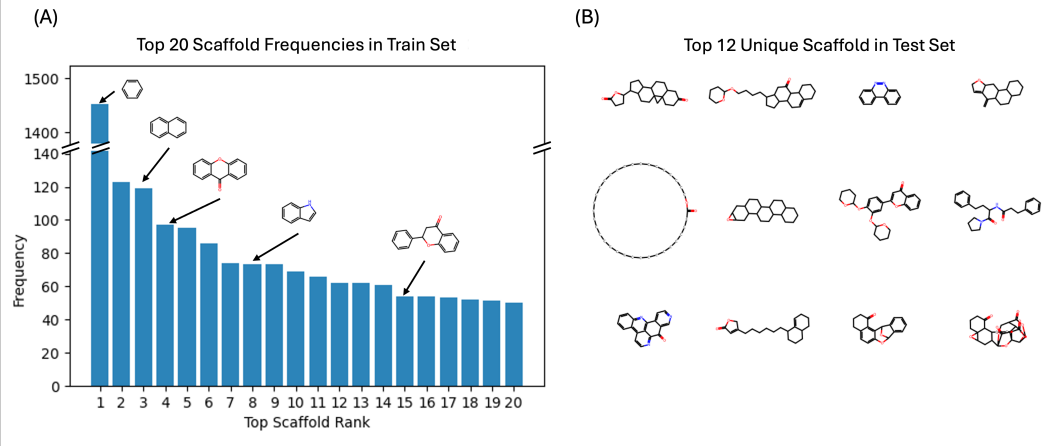

Figure 3: Scaffold analysis for training and test dataset.

3.1 Collection of HSQC spectra and SMILES

We meticulously curated 22,348 experimental HSQC spectra, together with their associated NMR conditions and CAS Registry Numbers. The spectra were obtained from the Human Metabolome Database (HMDB) [17] (CC-NC-4.0 license) and CH-NMR-NP [18]. For each molecule, the corresponding CAS Registry Numbers and SMILES representations were retrieved from PubChem [29].

3.2 Generation of molecular graphs

Molecular graphs with stable 3D structures are derived from SMILES strings using the RDKit [30] package, and formatted in Python Geometric format for computational processing. In the process of converting SMILES representations into molecular graphs, challenges arose with disjoint graphs, primarily due to the presence of floating ions. To ensure data quality and model accuracy, these anomalies are systematically identified and excluded from the dataset. Additionally, certain SMILES strings fail to yield energy-stable 3D structures despite multiple optimization attempts. These instances suggest structural inconsistencies or complexities that RDKit cannot resolve adequately. Such unstable entries are also eliminated to maintain the structural integrity and reliability of our dataset. This meticulous preprocessing ensures that our dataset only includes high-quality, consistent molecular graphs that are suitable for subsequent analysis and modeling. For 3D-based models (e.g., SchNet[25], ComENet[24]), molecular geometries were generated from SMILES when the lowest-energy conformer was retained. While this ensures consistency across molecules, it does not account for conformational diversity in solution, which can influence NMR shifts. Incorporating conformer ensembles or Boltzmann-weighted sampling could further improve 3D model robustness.

Furthermore, using the RDKit [30] package, we enrich the molecular graphs with node and edge features to infuse domain-specific insights into our Chemistry-Informed ML development. Three features are provided for each node: atomic type, chirality, and hybridization. Also, two features are considered for each edge: bond type and bond direction. Bond types include Single, Double, Triple, and Aromatic, each reflecting a distinct configuration of electron sharing between atoms. Bond direction includes None, EndUpRight, and EndDownRight, primarily representing stereochemistry in double bonds. ML practitioners have the option to incorporate these hand-crafted, domain-specific features in the model training process, which not only helps in understanding how traditional chemical knowledge translates into computational predictions but also explores how machine learning techniques can uncover patterns and relationships that might elude conventional domain expertise. This dual approach allows our models to benefit from established chemical theory while potentially discovering novel insights into molecular behavior that could redefine our understanding of NMR shifts and molecular interactions. Such findings could provide valuable contributions to the field, suggesting new areas of research or improvements to existing chemical theories.

3.3 Annotation process

Silver-standard labels. We use a framework proposed in [12] to generate pseudo lables for 21,869 molecules. This model was first trained on extensive 1D NMR data, which establishes a robust foundation for understanding basic molecular interactions and chemical shift patterns. Afterwards, the model was fine-tuned on a diverse set of 2D NMR data, enhancing its ability to generalize across different molecular structures and solvent environments. With an accurate prediction of 2D NMR cross peaks, the model uses a matching algorithm to assign the predicted cross peaks to the most plausible observed peaks in the HSQC spectra, thus creating a direct linkage between each observed peak and its corresponding C–H bonds within the molecular graph. To test its annotation capability, we compared the annotation generated by this model to the expert annotations on our test dataset. Table 1 displays the result. Out of the 479 test molecules, the algorithm accurately annotates all peaks for 456 of the molecules (95.21%). For the remaining 23 molecules, the model was able to annotate 81.56% of the peaks accurately.

Table 1: Pseudo-label Accuracy

Fully-Correct Molecule (%)	Peak Accuracy (%) for Partial-Correct Molecule
95.21%	81.56%

Golden-standard labels. The test dataset, comprising 479 molecules, underwent a rigorous multi-step annotation and validation process involving three domain experts to ensure the accuracy and reliability of labels used for model evaluation. The experts all have more than 10 years of experience in Organic Chemistry and NMR analysis, from Harvard University, Boston College and University of Georgia. Initially, all molecules were annotated by Expert A. Afterwards, the dataset was split into two subsets, each independently annotated and cross-checked by Expert B and Expert C. In cases of disagreement between the initial and secondary annotations, the molecule was flagged and reviewed by the third expert to resolve inconsistencies. The final consensus annotation agreed upon by at least two experts was recorded as the ground truth.

4 2DNMRGym benchmark

To guide Machine Learning (ML) practitioners using 2DNMRGym, we provide benchmarks for cross peak prediction, an atom level representation learning task, described in Section 1 and Figure 4. Models are evaluated on the held-out test set annotated by domain experts to ensure high-quality assessment. In addition to overall performance, we report results under few-shot and zero-shot evaluation settings to assess generalization. Specifically, a test molecule is considered few-shot if its scaffold appears fewer than 10 times in the training set, and zero-shot if its scaffold is not observed at all during training.

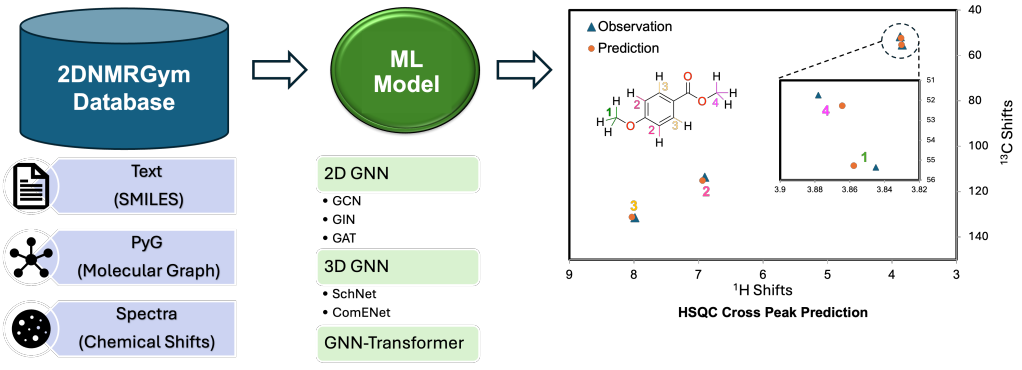

Figure 4: A demonstration workflow using 2DNMRGym dataset to train GNN models. The learnt graph representation from these benchmark models can be evaluated in the downstream HSQC cross peak prediction task.

4.1 Baseline models

To benchmark atom-level cross-peak prediction, we evaluate several representative GNN architectures. For 2D GNNs, we include GCN [20], which performs neighborhood aggregation with normalized message passing; GIN [21], designed for maximal expressive power in distinguishing graph structures; and GAT [22], which introduces attention mechanisms to weight neighbor contributions adaptively. We also incorporate GNN-Transformer [23], a hybrid model combining GNNs with global self-attention and structural encodings to capture both local and long-range dependencies, which has shown strong performance on chemical and biological benchmarks. For 3D molecular graphs, we consider SchNet [25], which leverages continuous-filter convolutions to model spatial interactions, and ComENet [24], which ensures full utilization of 3D geometric information within a 1-hop neighborhood. Together, these models provide a diverse baseline for evaluating atom-level representation learning on our 2DNMRGym dataset. The model details are included in Appendix C.

4.2 Training and evaluation

Train/validation split. In our experiments, the data is randomly split into 80% for training, 20% for model selection, and the expert-annotated test dataset is used for model evaluation. For each model, we repeat the experiments using random seeds of 0, 42 and 66 and report the mean and standard deviatiion of Mean Absolute Error (MAE).

Pre-processing. The value ranges of the ^{13}C- and ^{1}H-shifts are quite different, 0 - 200 ppm for ^{13}C versus 0-12 ppm for ^{1}H. To reduce bias and achieve better training, we normalized them to make their value range comparable by dividing ^{13}C-shifts by 200 and ^{1}H-shifts by 10.

Error measurement. As 2D NMR captures atomic interactions in two dimensions, specifically ^{13}C-shift and ^{1}H-shift, the model is trained using the Mean Absolute Error (MAE) of ^{13}C-shifts and ^{1}H-shifts, assigning them equal weights. The evaluation of the model's performance for both shifts are conducted using the MAE values calculated from the original values of the ^{13}C- and ^{1}H-shifts without normalization. This approach ensures that the model's predictions are assessed directly against the experimental chemical shift values, without any scaling or normalization, providing an unbiased assessment of its predictive capabilities for the two types of atomic interactions captured in 2D NMR spectra. MAE was selected as both the training loss and evaluation metric to align with prior NMR shift prediction literature and ensure interpretability in ppm units.

4.3 Benchmark results

All experiments were run using one V100 GPU. The performance of the baseline models is summarized in Table 2 and Table 3. Additional error measurement can be found in Appendix E. For each model, we adjusted its hyperparameters, including the hidden dimensions for GNN node representations, the hidden dimensions for edge representations (where applicable), the number of GNN layers, and the hidden channels of MLP layers for ^{13}C-shifts and ^{1}H-shifts predictions. Additionally for ComENet, we tune the number of layers inside the interaction module for node and edges during message passing. For SchNet, we also tune the number of filters in its filter-generating network. All models in this experiment are trained for 100 epochs with batch size set to 32.

For all GNN models, adding the transformer component in model architecture generally improve performance and reduces variances, while not to a large margin. This suggests that while HSQC correlations are primarily local, through-space interactions and solvent effects can modulate the observed shifts, indicating a nuanced interplay between local and global molecular features. Among GNN architectures, GIN models perform the best in our task due to their strong discriminative power, which is essential for capturing subtle structural variations that influence NMR shifts. Unlike GCN and GAT, GIN uses injective aggregation functions that better preserve node uniqueness within molecular graphs. Compared to GAT models, GIN is also architecturally simpler and tends to be more robust, especially when the dataset contains noise or biases introduced by silver standard labeling. This robustness makes GIN more reliable in learning meaningful representations from limited or noisy training data.

HSQC spectra primarily reflect short-range correlations governed by the 2D molecular structure, such as connectivity, atom types, hybridization, and chirality. These features, which are directly encoded in our graph representations, are sufficient to capture the stereoelectronic environments that determine

Model Type	Model	All-test MAE		Few-shot MAE		Zero-shot MAE	
		^{13}C	^{1}H	^{13}C	^{1}H	^{13}C	^{1}H
2D GNN	GCN	3.035 (0.039)	0.229 (0.002)	3.014 (0.011)	0.227 (0.001)	3.103 (0.038)	0.242 (0.002)
	GIN	2.370 (0.007)	0.203 (0.003)	**2.274 (0.022)**	0.192 (0.002)	**2.587 (0.005)**	0.230 (0.003)
	GAT	2.574 (0.045)	0.206 (0.004)	2.524 (0.042)	0.201 (0.003)	2.811 (0.066)	0.226 (0.003)
3D GNN	ComENet	3.143 (0.018)	0.238 (0.003)	3.178 (0.015)	0.233 (0.002)	3.348 (0.042)	0.262 (0.003)
	SchNet	3.156 (0.022)	0.240 (0.001)	3.183 (0.014)	0.239 (0.001)	3.369 (0.031)	0.261 (0.001)
Transformer	GCN-Trans	2.911 (0.044)	0.221 (0.003)	2.869 (0.036)	0.215 (0.004)	3.017 (0.055)	0.241 (0.004)
	GIN-Trans	**2.348 (0.031)**	**0.198 (0.000)**	2.281 (0.016)	**0.188 (0.001)**	2.620 (0.039)	**0.228 (0.003)**
	GAT-Trans	2.543 (0.097)	0.206 (0.005)	2.493 (0.104)	0.200 (0.006)	2.740 (0.079)	0.228 (0.005)

Table 2: Comparison of MAE in ppm for ^{13}C and ^{1}H chemical shift predictions across different GNN models. The best model parameters are documented in Appendix D.

Model Type	Model	All-test R^2		Few-shot R^2		Zero-shot R^2	
		^{13}C	^{1}H	^{13}C	^{1}H	^{13}C	^{1}H
2D GNN	GCN	0.9784 (0.0002)	0.9680 (0.0002)	0.9889 (0.0001)	0.9781 (0.0001)	0.9591 (0.0001)	0.9453 (0.0001)
	GIN	**0.9822 (0.0002)**	**0.9713 (0.0002)**	**0.9926 (0.0001)**	0.9827 (0.0003)	**0.9626 (0.0005)**	0.9472 (0.0001)
	GAT	0.9811 (0.0004)	0.9709 (0.0003)	0.9916 (0.0003)	0.9813 (0.0005)	0.9615 (0.0005)	0.9479 (0.0001)
3D GNN	ComENet	0.9589 (0.0004)	0.9411 (0.0009)	0.9681 (0.0004)	0.9456 (0.0011)	0.9335 (0.0008)	0.9147 (0.0007)
	SchNet	0.9602 (0.0003)	0.9349 (0.0004)	0.9697 (0.0004)	0.9328 (0.0009)	0.9364 (0.0005)	0.9132 (0.0004)
Transformer	GCN-Trans	0.9794 (0.0004)	0.9679 (0.0006)	0.9902 (0.0003)	0.9792 (0.0006)	0.9602 (0.0006)	0.9440 (0.0009)
	GIN-Trans	0.9823 (0.0000)	0.9708 (0.0005)	0.9929 (0.0000)	**0.9825 (0.0004)**	0.9626 (0.0002)	**0.9473 (0.0010)**
	GAT-Trans	0.9812 (0.0007)	0.9704 (0.0006)	0.9919 (0.0004)	0.9818 (0.0008)	0.9620 (0.0006)	0.9469 (0.0008)

Table 3: Comparison of R^2 for ^{13}C and ^{1}H chemical shift predictions across different GNN and Transformer models. The best results in each column are highlighted in bold.

chemical shifts. In contrast, 3D models like ComENet or SchNet rely on atomic coordinates that may not be optimal, as a molecule can adopt many possible conformers in solution. When only a single RDKit-embedded conformer is used, 3D models risk learning from spurious geometrical patterns or overfitting to noise in the 3D structure, leading to degraded performance compared to 2D models.

To investigate the sources of model error, we grouped the 23 molecules with partially correct annotations based on their molecular scaffolds. While the errors did not cluster around a specific scaffold type or chemical shift range, we observed that structurally complex molecules, such as those containing flexible ring systems and multiple chiral centers, tended to exhibit higher annotation errors.

Additionally, some atoms are pseudo-chemically symmetric, existing in nearly identical chemical and electronic environments, which may further complicate accurate peak assignment. These results and visualizations are shown in Appendix F.

5 Discussion and conclusion

Our curated 2DNMRGym dataset is the first experimental, centralized, annotated, and high-quality dataset for learning atom-level molecular representation in the 2D NMR space. Significant effort was invested in the database's construction, with the cross-validation from three domain experts. Our dataset includes multimodal inputs such as text and graphs, and covers a wide range of molecules of varying sizes and scaffolds, providing valuable insights for evaluating representation learning models. To establish benchmark results, we tested a variety of 2D and 3D GNN models to predict HSQC cross peaks from molecular topologies/structures, paving the way for more advanced machine learning models for predicting HSQC cross peaks. The benchmarking results indicate that GIN stands out among the 2D and 3D GNN models that we have tried. This highlights the potential for developing 3D GNN models to capture spatial information such as chirality centers and hybridization, for atom-level tasks, which is potentially a major advance in NMR spectroscopy. Also, since each molecule is represented by a single low-energy conformer generated by RDKit in the 3D representation, it may not fully capture the conformational ensemble relevant to NMR shifts, where multiple states coexist in solution. Future work incorporating conformational averaging or Boltzmann weighting could provide a more physically faithful representation. There is plenty of room for improvements in prediction precision, aiming for an ideal MAE of less than 2 ppm for ^{13}C and less than 0.1 ppm for ^{1}H.

Currently, the database contains only HSQC experimental data, which was generated to interrogate C–H interactions. Nevertheless, we expect the models trained on this HSQC data can be easily adapted or fine-tuned for other types of 2D NMR data. Looking ahead, the 2DNMRGym dataset is poised for further expansion to include a broader range of NMR techniques, such as HMBC and COSY, which probe different aspects of atomic interactions within molecules. Such expansions will enable the development of more advanced ML techniques for analyzing a wider array of NMR spectra, facilitating a more integrated approach to molecular characterization. Additionally, we note that although an accurate spectra prediction given the molecular SMILE representation aids in structural elucidation and discrimination between isomers, it is not a direct structural prediction from spectra. Framing spectral data as input and structural output as targets remains a long-standing challenge in the chemistry community, and the 2DNMRGym dataset can also serve as a valuable asset for the development of effective deep learning model architectures and pipelines.

Author Contributions

Y.Li contributed in conceptualization, software, analysis, validation, investigation, visualization, methodology and writing. X. Hao contributed in conceptualization, analysis, investigation, data curation and writing. P. Hong contributed in supervision, resources, funding acquisition, project administration and writing review.

Acknowledgements

This work was supported by GlycoMIP, a National Science Foundation (NSF) Materials Innovation Platform funded through Cooperative Agreement DMR-1933525, as well as NSF OAC 1920147. We also want to thank all the expert annotators: Dr. Hao Xu from Harvard Medical School, Dr. Duo-Sheng Wang from Boston College, and Dr. Ambrish Kumar from University of Georgia, Athens.

References

[1] Harald Gunther and Harald Gunther. *NMR spectroscopy: basic principles, concepts, and applications in chemistry*. John Wiley & Sons Chichester, UK, 1994. 1

[2] Timothy DW Claridge. *High-resolution NMR techniques in organic chemistry*, volume 27. Elsevier, 2016.

[3] Hyo-Yeon Yu, Sangki Myoung, and Sangdoo Ahn. Recent applications of benchtop nuclear magnetic resonance spectroscopy. *Magnetochemistry*, 7(9):121, 2021. 1

[4] Geoffrey Bodenhausen and David J Ruben. Natural abundance nitrogen-15 nmr by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, 69(1):185–189, 1980. 1, 3

[5] Nadja Bross-Walch, Till Kühn, Detlef Moskau, and Oliver Zerbe. Strategies and tools for structure determination of natural products using modern methods of nmr spectroscopy. *Chemistry & biodiversity*, 2(2):147–177, 2005. 2

[6] Qingxin Li and CongBao Kang. A practical perspective on the roles of solution nmr spectroscopy in drug discovery. *Molecules*, 25(13):2974, 2020. 2

[7] Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. Neural message passing for nmr chemical shift prediction. *Journal of chemical information and modeling*, 60(4):2024–2030, 2020. 2

[8] Ziyue Yang, Maghesree Chakraborty, and Andrew D White. Predicting chemical shifts with graph neural networks. *Chemical science*, 12(32):10802–10809, 2021.

[9] Jongmin Han, Hyungu Kang, Seokho Kang, Youngchun Kwon, Dongseon Lee, and Youn-Suk Choi. Scalable graph neural network for nmr chemical shift prediction. *Physical Chemistry Chemical Physics*, 24(43):26870–26878, 2022.

[10] Haochen Chen, Tao Liang, Kai Tan, Anan Wu, and Xin Lu. Gt-nmr: a novel graph transformer-based approach for accurate prediction of nmr chemical shifts. *Journal of Cheminformatics*, 16 (1):132, 2024. 2

[11] Hao Xu, Zhengyang Zhou, and Pengyu Hong. Enhancing peak assignment in 13c nmr spectroscopy: A novel approach using multimodal alignment. *arXiv preprint arXiv:2311.13817*, 2023. 2

[12] Yunrui Li, Hao Xu, Ambrish Kumar, Duo-Sheng Wang, Christian Heiss, Parastoo Azadi, and Pengyu Hong. Transpeaknet for solvent-aware 2d nmr prediction via multi-task pre-training and unsupervised learning. *Communications chemistry*, 8(1):51, 2025. 2, 7

[13] D Weininger, A Weininger, and JL Weininger. Smiles (simplified molecular input line entry system). *J Chem Inf Comput Sci*, 28:31–36, 1988. 3

[14] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13, 2015. 3

[15] Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996. 3

[16] Christoph Steinbeck, Stefan Krause, and Stefan Kuhn. Nmrshiftdb constructing a free chemical information system with open-source components. *Journal of chemical information and computer sciences*, 43(6):1733–1739, 2003. 4

[17] David S Wishart, AnChi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, Zinat Sayeeda, Siyang Tian, Brian L Lee, et al. Hmdb 5.0: the human metabolome database for 2022. *Nucleic acids research*, 50(D1):D622–D631, 2022. 4, 6

[18] KY Hayamizu, K Asakura, and T Kurimoto. An open access nmr database for organic natural products "ch-nmr-np". In *57th Experimental Nuclear Magnetic Resonance Conference, Pittsburgh, PA*, 2015. 4, 6

[19] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808, 2024. 4

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4, 8, 13

[21] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018. 4, 8, 13

[22] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017. 4, 8, 14

[23] Zhanghao Wu, Paras Jain, Matthew Wright, Azalia Mirhoseini, Joseph E Gonzalez, and Ion Stoica. Representing long-range context for graph neural networks with global attention. *Advances in neural information processing systems*, 34:13266–13279, 2021. 4, 8, 14

[24] Limei Wang, Yi Liu, Yuchao Lin, Haoran Liu, and Shuiwang Ji. Comenet: Towards complete and efficient message passing for 3d molecular graphs. *Advances in Neural Information Processing Systems*, 35:650–664, 2022. 4, 6, 8, 14

[25] Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), 2018. 4, 6, 8, 14

[26] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018. 4

[27] Clemens Isert, Kenneth Atz, José Jiménez-Luna, and Gisbert Schneider. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data*, 9(1):273, 2022. 4

[28] Simon Axelrod and Rafael Gomez-Bombarelli. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. 4

[29] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023. 6

[30] Greg Landrum. Rdkit documentation. *Release*, 1(1-79):4, 2013. 6

[31] W Bremser. Hose—a novel substructure code. *Analytica Chimica Acta*, 103(4):355–365, 1978. 13

[32] Keith W Wiitala, Thomas R Hoye, and Christopher J Cramer. Hybrid density functional methods empirically optimized for the computation of 13c and 1h chemical shifts in chloroform solution. *Journal of Chemical Theory and Computation*, 2(4):1085–1092, 2006. 13

[33] Nancy Mills. Chemdraw ultra 10.0 cambridgesoft, 2006. 13

[34] Mark Robert Willcott. Mestre nova, 2009. 13

A Annotation challenges

2D NMR annotation, which involves associating the chemical shifts of each atom pair with the observed signals from experiments, is a highly challenging task. Using the HSQC spectrum as an example, the signals observed in the 2D spectrum correspond to the chemical shifts of hydrogen atoms directly bonded to heteronuclei, typically ^{13}C or ^{15}N. Annotating these signals requires accurately mapping the observed cross-peaks to specific hydrogen-heteronucleus pairs within the molecule. However, this process is complicated by several factors, including spectral overlap, signal degeneracy, and sensitivity to experimental conditions.

Spectral overlap occurs when multiple signals appear at similar chemical shift values, making it difficult to distinguish and assign them correctly. This issue is exacerbated in larger molecules with numerous hydrogen-heteronucleus pairs, leading to increased signal density and potential overlap. Additionally, signal degeneracy, where multiple atom pairs share the same chemical shift, further complicates the annotation process. Figure 5 shows an example of a large molecule in our dataset. Moreover, the observed chemical shifts are highly sensitive to the experimental conditions, such as temperature, solvent, pH, and sample concentration. Even slight variations in these conditions can cause detectable shifts in the signals, making it challenging to reliably match the experimental data with reference values or theoretical predictions.

B Additional Concepts and terminology in chemistry

Solvent. A solvent, typically a liquid, is used to dissolve other substances (solutes), resulting in the formation of a solution. In the context of HSQC spectroscopy, solvent selection is paramount due to its profound influence on the chemical environment of the sample, thereby affecting the observed chemical shifts in NMR spectra. These shifts serve as pivotal indicators for accurately interpreting molecular structures as solvents can alter interactions such as hydrogen bonding, change molecular conformations, and affect the dynamics within a molecule. Thus, selecting an appropriate solvent and understanding its influence is essential for achieving precise and meaningful HSQC spectral analysis.

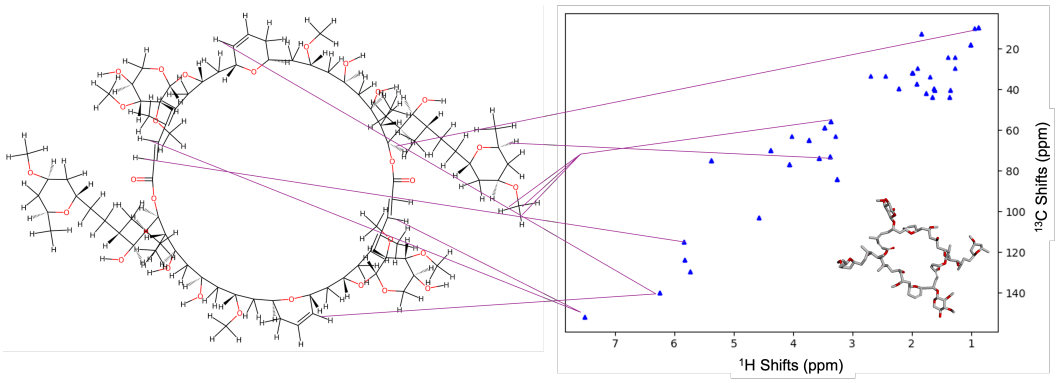

Figure 5: An annotation example. To avoid overcrowded, only a few "C-H bond – peak" associations are shown. For a large molecule with complex structure like this, aligning the chemical bonds with the cross peaks is extremely difficult due to signal overlap and degeneracy. The bottom-right of the HSQC spectrum shows a 3D abstract skeleton of the molecule.

HOSE codes. HOSE [31] codes are a method used in NMR spectroscopy for predicting chemical shifts. These codes function by encoding the structural environment of a nucleus in concentric spheres, capturing the types and positions of neighboring atoms up to several bonds away. Each sphere represents a distinct "shell" of neighbors, and the method relies on a database of known chemical shifts to predict the shift for a given atom based on its specific environment. This approach is empirical, utilizing accumulated historical data to make predictions.

DFT. Density Functional Theory (DFT) [32] is a quantum mechanical method used to investigate the electronic properties of molecules and solids. In the context of NMR, DFT can be used to calculate chemical shifts by simulating the electronic environment around nuclei. This involves solving the Schrödinger equation for electrons in a molecule under the influence of a magnetic field, allowing for the prediction of NMR properties based on fundamental physical principles. DFT is known for its accuracy and ability to handle complex molecules, though it is computationally intensive compared to empirical methods like HOSE codes.

Traditional tools in chemistry. Two software tools are commonly used for processing, visualizing, simulating, and analyzing NMR spectral data, *ChemDraw* [33] and *Mestrenova* [34]. They can serve as baselines for Machine Learning based methods.

C Benchmark GNN models

C.1 2D GNN models

GCN. Graph Convolutional Networks (GCNs) [20] is designed to efficiently learn node representations by leveraging the graph's structural information. The update rule for a GCN layer is formulated as follows:

$$h_v^{(k+1)} = \sigma\left(W^{(k)} \sum_{u \in \mathcal{N}(v) \cup \{v\}} \frac{1}{\sqrt{\deg(v)\deg(u)}} h_u^{(k)}\right), \tag{1}$$

where $h_v^{(k)}$ represents the feature vector of node v at layer k, $\mathcal{N}(v)$ denotes the set of neighbors of node v, $W^{(k)}$ is the weight matrix at the k-th layer, and σ is a non-linear activation function (e.g., ReLU), and $\deg(v)$ and $\deg(u)$ are the degrees of nodes v and u, respectively. This approach, by normalizing based on node degrees, mitigates the problem of scale differences in node degrees, thus ensuring stable training and effective feature learning.

GIN. Graph Isomorphism Networks (GIN) [21] are introduced to enhance the ability of GNNs to capture the structural nuances of graphs more effectively. Traditional GNN models often struggle

to distinguish non-isomorphic graphs due to their limited expressiveness, akin to the Weisfeiler-Lehman (WL) graph isomorphism test. GINs are designed to address this issue by achieving maximal expressiveness in distinguishing graph structures. The general update rule for a GIN model is defined as follows:

$$h_v^{(k+1)} = \text{MLP}^{(k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(k)} + \sum_{u\in\mathcal{N}(v)} h_u^{(k)}\right), \tag{2}$$

where $h_v^{(k)}$ is the feature vector of node v at layer k, $\mathcal{N}(v)$ denotes the set of neighbors of node v, $\text{MLP}^{(k)}$ represents a multi-layer perceptron used at the k-th layer, $\epsilon^{(k)}$ is a learnable parameter or a fixed scalar that can be tuned to adjust the model's sensitivity to the central node's features.

GAT. Graph Attention Networks (GATs) [22] incorporates the mechanism of attention into the GNN by dynamically assigning importance to nodes within a local neighborhood. The core update rule for a GAT model is expressed as follows:

$$h_v^{(k+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)\cup\{v\}} \alpha_{vu}^{(k)} W^{(k)} h_u^{(k)}\right), \tag{3}$$

where $h_v^{(k)}$ is the representation of node v at layer k, $\mathcal{N}(v)$ denotes the neighbors of node v, $W^{(k)}$ is a weight matrix for the k-th layer, $\alpha_{vu}^{(k)}$ represents the attention coefficient between nodes v and u, and σ is a nonlinear activation function. The attention coefficients $\alpha_{vu}^{(k)}$ are computed through a learnable function of the features of nodes v and u, allowing the model to focus more on relevant features during aggregation.

GNN transformer. The GNNTrans [23] model introduces a hybrid architecture that combines the expressive power of Graph Neural Networks (GNNs) with the global attention mechanism of Transformers to better capture both local and long-range dependencies in graph-structured data. By integrating structural encodings and a novel graph token, the model effectively handles graph-level tasks, achieving state-of-the-art performance on multiple benchmarks. This approach bridges the gap between sequential attention models and relational inductive biases in graphs. The model also achieves promising results on biological and chemical benchmarks, making it a suitable benchmark for our dataset.

C.2 3D GNN models

ComENet. ComENet [24] offers an efficient message passing network designed specifically for 3D GNNs. It incorporates a new message passing scheme that ensures complete utilization of 3D information by operating within a 1-hop neighborhood, achieving both global and local completeness.

SchNet. SchNet is another 3D GNN architecture designed for modeling atomic-scale interactions within molecules and materials [25]. It employs a unique continuous-filter convolutional approach to capture the complex interatomic forces and represents interatomic distances through a radial basis function expansion using a flexible number of Gaussian functions.

D Model parameters

The optimal hyperparameters for each model in Table 2 are summarized below. For each model type, extensive parameter tuning was conducted. The number of GNN layers tested included 3, 4, 5, 6, with hidden dimensions of 256, 374, 512. Prediction head configurations evaluated included [256, 128], [128, 64], [256], [128]. Solvent embedding dimensions were selected from 16, 32. For the Transformer module, the hidden dimensions considered were 128, 256, the number of attention heads 2, 3, 4, feedforward dimensions 256, 512, and the number of Transformer layers 3, 4, 5.

Table 4: Model configurations for transformer GNN models

Batch size	GNN type	GNN layer	Hid dim	Pred head (C)	Pred head (H)	Solvent emb (C)	Solvent emb (H)	Trans hid dim	Num of heads	Trans ff dim	Trans layer
32	gin	5	512	[128, 64]	[128, 64]	16	16	128	4	512	3
32	gcn	5	512	[128, 64]	[128, 64]	16	16	128	4	256	5
32	gat	5	512	[128, 64]	[128, 64]	16	16	128	2	512	5

Table 5: Model configurations for GNN-only models

Batch size	GNN type	GNN layer	Hidden dim	Pred head (C)	Pred head (H)	Solvent emb (C)	Solvent emb (H)	Filters	Gaussians
32	gat	5	512	[128, 64]	[128, 64]	32	16	–	–
32	gat	5	512	[128, 64]	[128, 64]	32	32	–	–
32	gcn	5	512	[128, 64]	[128, 64]	32	32	–	–
32	gin	5	512	[128, 64]	[128, 64]	32	16	–	–
32	gin	5	512	[128, 64]	[128, 64]	32	32	–	–
32	schnet	3	512	[128, 64]	[128, 64]	16	16	128	50
32	comenet	6	512	[128, 64]	[128, 64]	16	16	–	–

E Model Comparison (RMSE)

Besides the MAE and R-squared tables (Table 2 and Table 3 in the main text), we also compared the model performance in RMSE.

Model Type	Model	All-test RMSE		Few-shot RMSE		Zero-shot RMSE	
		^{13}C	^{1}H	^{13}C	^{1}H	^{13}C	^{1}H
2D GNN	GCN	5.9709 (0.0217)	0.4009 (0.0011)	4.1757 (0.0144)	0.3195 (0.0004)	7.9259 (0.0126)	0.5028 (0.0005)
	GIN	**5.4207 (0.0335)**	0.3798 (0.0012)	**3.3935 (0.0243)**	**0.2837 (0.0021)**	**7.5856 (0.0490)**	0.4941 (0.0006)
	GAT	5.5895 (0.0601)	**0.3821 (0.0022)**	3.6252 (0.0604)	0.2947 (0.0037)	7.6914 (0.0538)	**0.4905 (0.0004)**
3D GNN	ComENet	6.2520 (0.0398)	0.4448 (0.0042)	5.0769 (0.0137)	0.4032 (0.0049)	8.1323 (0.0597)	0.5292 (0.0026)
	SchNet	6.1147 (0.0280)	0.4275 (0.0017)	4.8947 (0.0406)	0.4593 (0.0036)	7.9078 (0.0365)	0.5348 (0.0014)
Transformer	GCN-Trans	5.8389 (0.0617)	0.4016 (0.0035)	3.9218 (0.0616)	0.3110 (0.0045)	7.8195 (0.0603)	0.5085 (0.0039)
	GIN-Trans	5.4041 (0.0343)	0.3828 (0.0036)	3.3318 (0.0087)	0.2852 (0.0030)	7.5851 (0.0177)	0.4932 (0.0048)
	GAT-Trans	5.5714 (0.0972)	0.3857 (0.0038)	3.5712 (0.0876)	0.2912 (0.0061)	7.6412 (0.0593)	0.4953 (0.0038)

Table 6: Comparison of RMSE in ppm for ^{13}C and ^{1}H chemical shift predictions across different GNN and Transformer models. Best results in each column are highlighted in bold.

F Additional model error analysis

We observed that structurally complex molecules, such as those containing flexible ring systems and multiple chiral centers, tended to exhibit higher annotation errors. Additionally, some atoms are pseudo-chemically symmetric, existing in nearly identical chemical and electronic environments, which may further complicate accurate peak assignment.

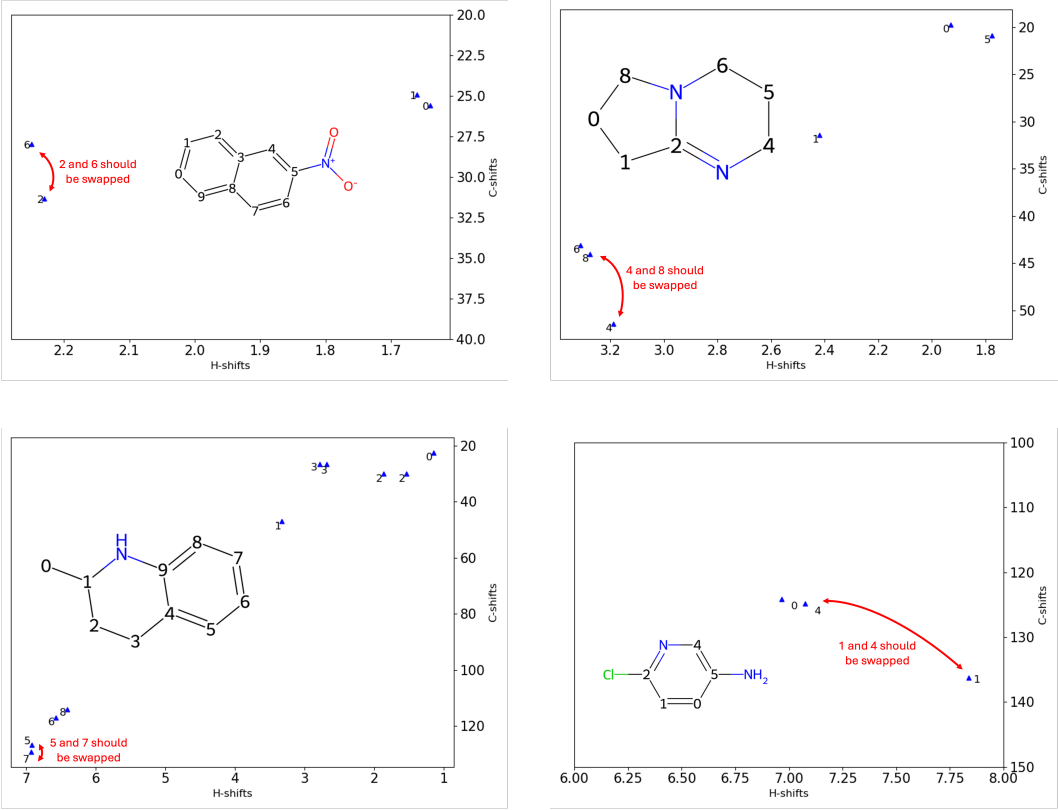

Figure 6: In molecules with pseudo-chemically symmetric atoms, the local environments are effectively indistinguishable at typical spectral resolution, leading to near-degenerate chemical shifts; this underdetermines one-to-one atom–peak assignment and leads to annotation error.

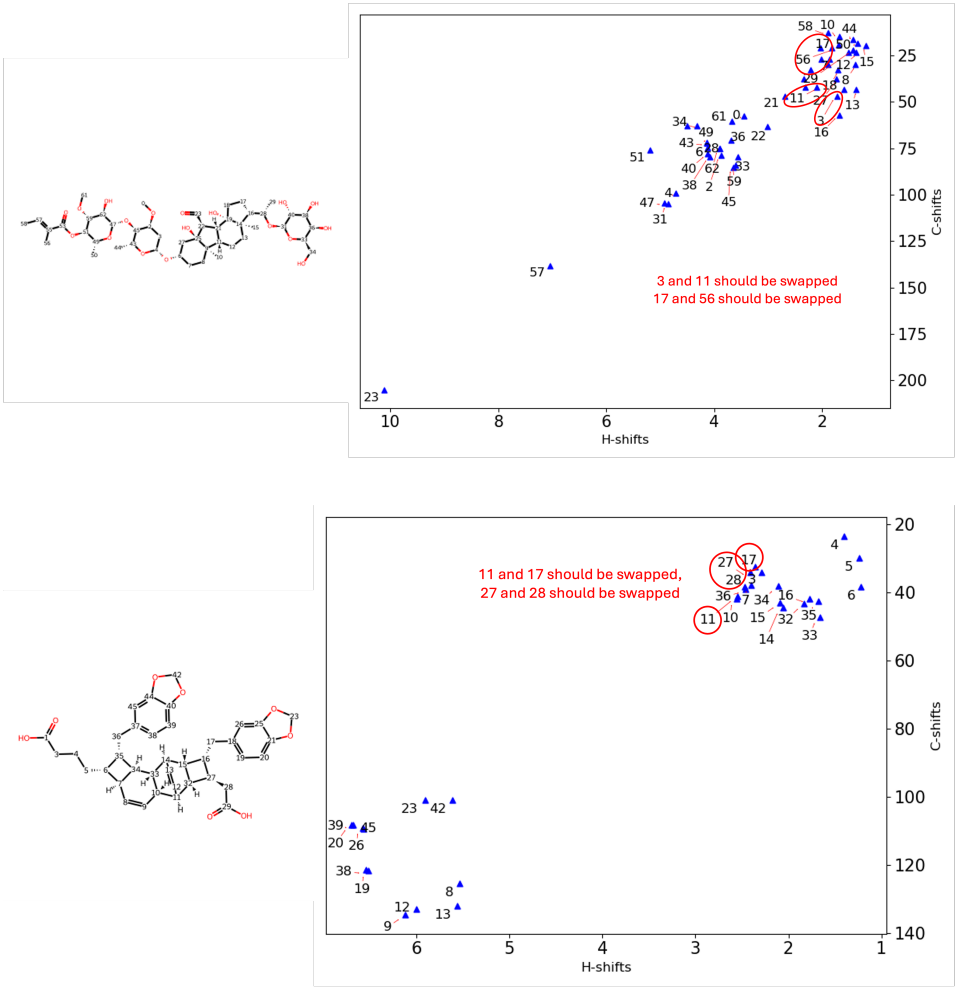

Figure 7: Multiple-ring architectures concentrate many near-isotopic chemical environments into narrow chemical-shift windows. Ring-current anisotropy, repeated CH motifs, and conformational averaging in flexible rings yield near-degenerate shifts and broadened/overlapping HSQC cross-peaks. Together these effects crowd the spectrum and underdetermine one-to-one assignments, increasing annotation error. (More figure in the next page)

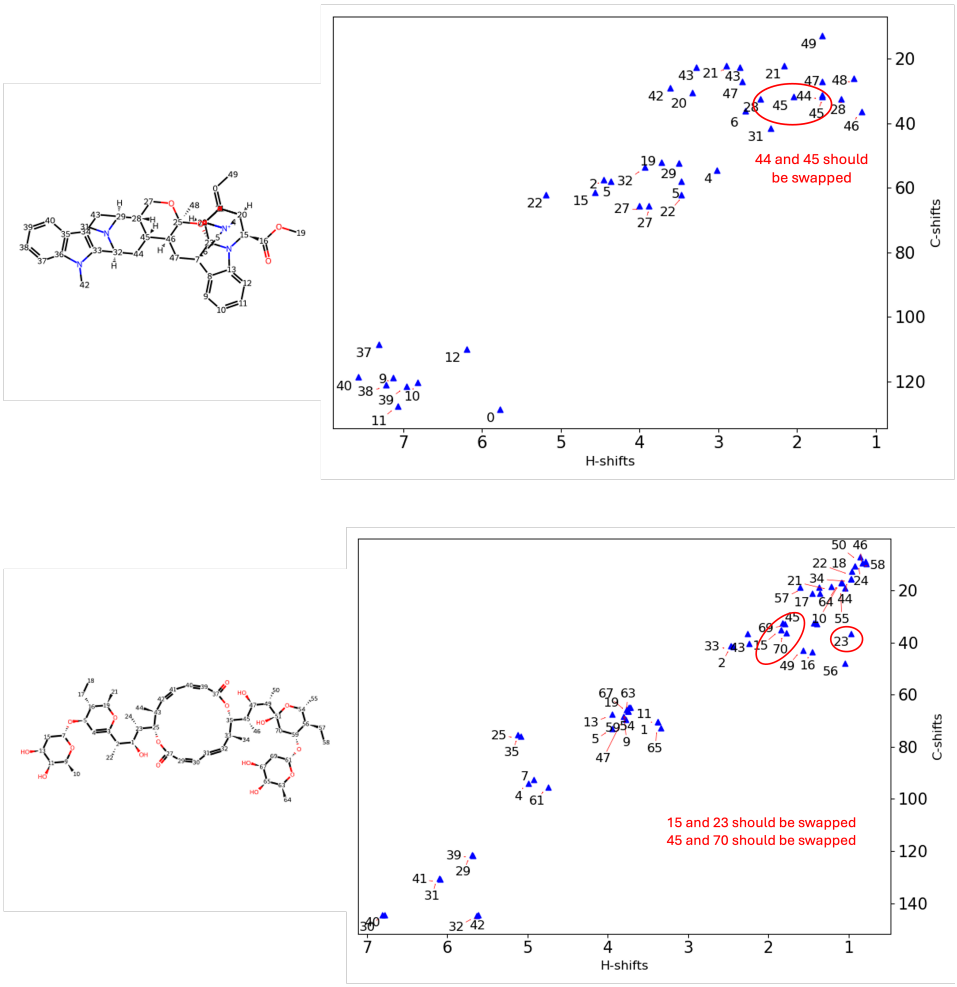

Figure 7: Multiple-ring architectures (continued).