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

Anonymous Author(s)
Affiliation
Address
email

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 (Gunther and Gunther, 1994; Claridge, 2016; Yu *et al.*, 2021). 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.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Among NMR techniques, Heteronuclear Single Quantum Coherence (HSQC) spectroscopy (Bodenhausen and Ruben, 1980) stands out as a powerful two-dimensional (2D) Nuclear Magnetic Resonance (NMR) method that has become indispensable for the structural elucidation of complex molecules, especially when traditional one-dimensional (1D) NMR techniques are insufficient (Bross-Walch *et al.*, 2005; Li and Kang, 2020). 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 (Kwon *et al.*, 2020; Yang *et al.*, 2021; Han *et al.*, 2022; Chen *et al.*, 2024) and peak assignment (Xu *et al.*, 2023), 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. Experimental 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.

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 (Li *et al.*, 2025) 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.

1.2 Concepts and terminology in chemistry

SMILES. Simplified Molecular Input Line Entry System (SMILES) (Weininger *et al.*, 1988) 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 (Bodenhausen and Ruben, 1980) 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.

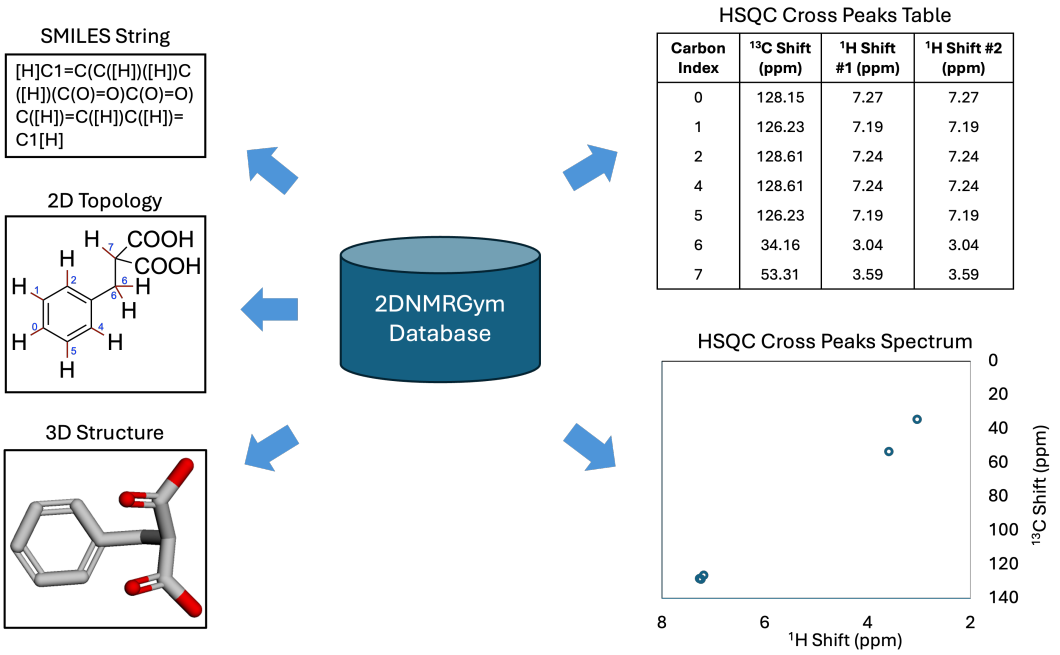

Figure 1: The 2DNMRGym dataset comprises multi-modal components, including the SMILES representation of each molecule and its conversion to a molecular graph. This graph includes both 2D topological structures and Cartesian coordinates for 3D spatial information. The ground truth spectrum is represented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

Tanimoto similarity. Tanimoto similarity is a widely used metric in cheminformatics for comparing molecular fingerprints, which are typically represented as binary vectors (Bajusz *et al.*, 2015). 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 (Bemis and Murcko, 1996) 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 (Steinbeck *et al.*, 2003) 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) (Wishart *et al.*, 2022), 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 (Hayamizu *et al.*, 2015), 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.

Furthermore, most existing ML models such as GCN (Kipf and Welling, 2016), GIN (Xu *et al.*, 2018), GAT (Velickovic *et al.*, 2017), GNN Transformer (Wu *et al.*, 2021), ComENet (Wang *et al.*, 2022) and SchNet (Schütt *et al.*, 2018) are trained at the molecule (graph-level) using coarse labels such as molecular properties using datasets like MolecularNet (Wu *et al.*, 2018), QMugs (Isert *et al.*, 2022), GEOM (Axelrod and Gomez-Bombarelli, 2022) 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.

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.

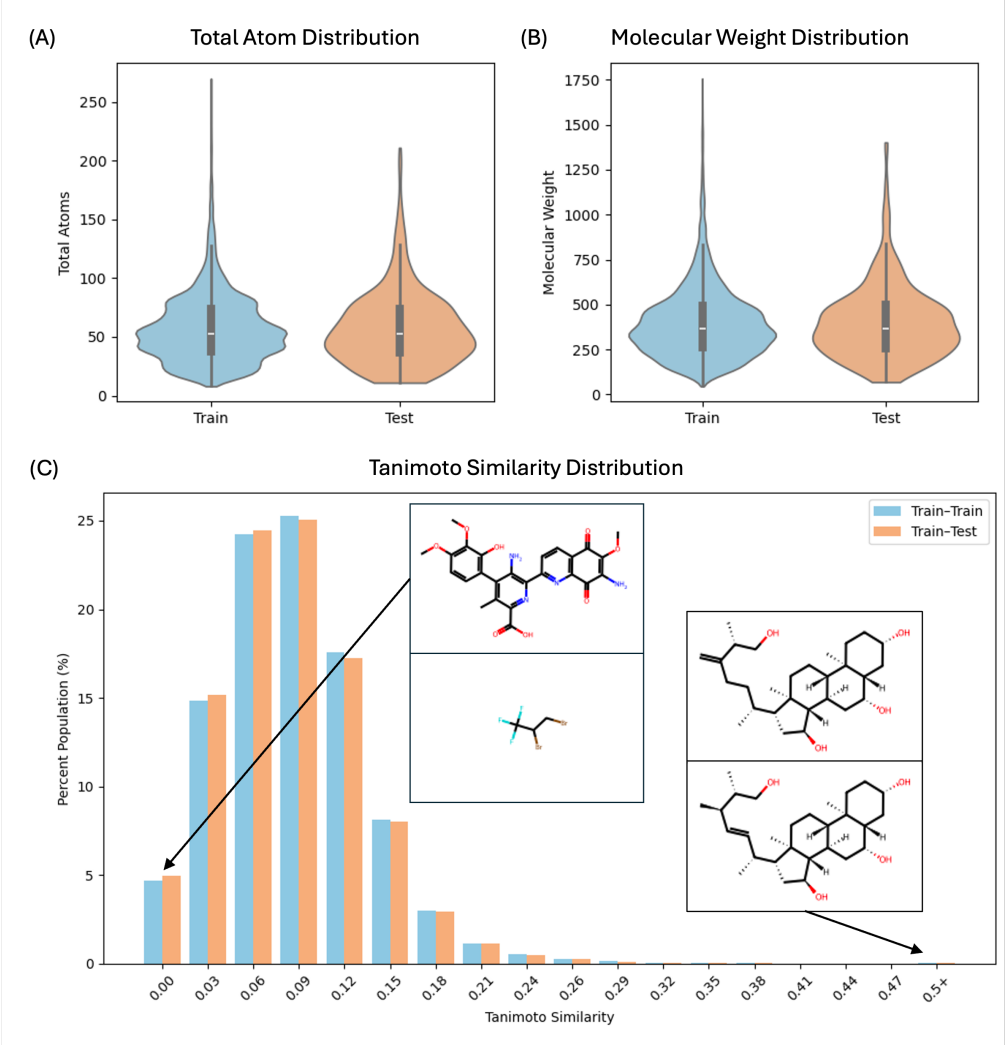

Figure 2: Data statistics by number of atoms, molecular weight, and tanimoto similarity.

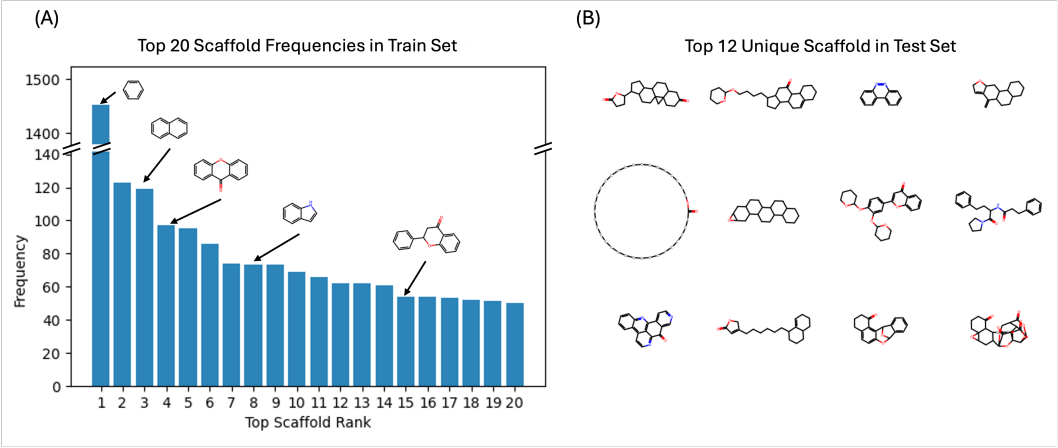

Figure 3: Scaffold analysis for training and test dataset.

3.1 Collection of HSQC spectra and SMILES

We meticulously curated 22,157 experimental spectra, along with NMR conditions and molecular CAS Registry numbers, which were extracted from the Human Metabolome Database(HMDB) (Wishart *et al.*, 2022) (CC-NC-4.0 licence), and CH-NMR-NP (Hayamizu *et al.*, 2015) for each molecule were extracted from PubChem (Kim *et al.*, 2023) (CC-BY-4.0 license) using their CAS Registry numbers. The corresponding SMILES for each molecule were extracted from PubChem (Kim *et al.*, 2023) using their CAS numbers.

3.2 Generation of molecular graphs

Molecular graphs with stable 3D structures are derived from SMILES strings using the RDKit (Landrum, 2013) 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.

Furthermore, using the RDKit (Landrum, 2013) 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 (Li *et al.*, 2025) 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.

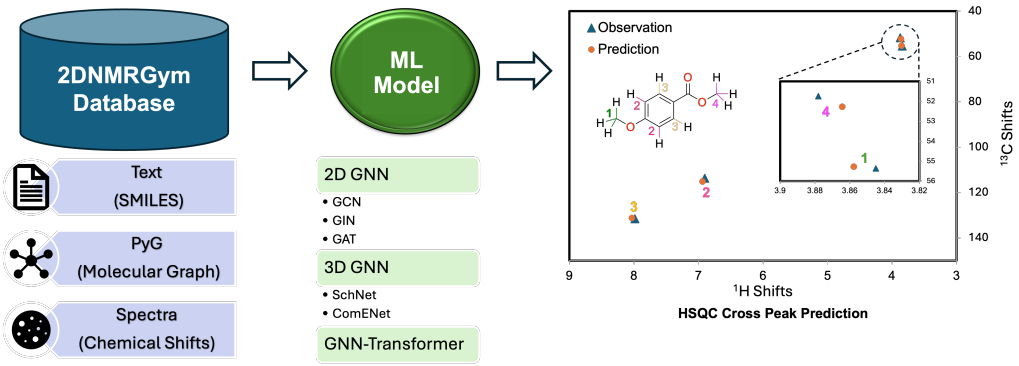

Figure 4: A demonstration workflow using 2DNMRGym dataset to train GNN models. The learnt graph representation from these benchmark models can be evaluated in the downstream HSQC cross peak prediction task.

4.1 Baseline models

To benchmark atom-level cross-peak prediction, we evaluate several representative GNN architectures. For 2D GNNs, we include GCN (Kipf and Welling, 2016), which performs neighborhood aggregation with normalized message passing; GIN (Xu *et al.*, 2018), designed for maximal expressive power in distinguishing graph structures; and GAT (Velickovic *et al.*, 2017), which introduces attention mechanisms to weight neighbor contributions adaptively. We also incorporate GNN-Transformer (Wu *et al.*, 2021), 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 (Schütt *et al.*, 2018), which leverages continuous-filter convolutions to model spatial interactions, and ComENet (Wang *et al.*, 2022), 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 is 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.

4.3 Benchmark results

All experiments were run using one V100 GPU. The performance of the baseline models is summarized in Table 2. 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.

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.

For all GNN models, adding the transformer component in model architecture generally boosts performance and reduces variances. 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 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.

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. 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.

6 Acknowledgment

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

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

T. D. Claridge, *High-resolution NMR techniques in organic chemistry*, Elsevier, 2016, vol. 27.

H.-Y. Yu, S. Myoung and S. Ahn, *Magnetochemistry*, 2021, **7**, 121.

G. Bodenhausen and D. J. Ruben, *Chemical Physics Letters*, 1980, **69**, 185–189.

N. Bross-Walch, T. Kühn, D. Moskau and O. Zerbe, *Chemistry & biodiversity*, 2005, **2**, 147–177.

Q. Li and C. Kang, *Molecules*, 2020, **25**, 2974.

Y. Kwon, D. Lee, Y.-S. Choi, M. Kang and S. Kang, *Journal of chemical information and modeling*, 2020, **60**, 2024–2030.

Z. Yang, M. Chakraborty and A. D. White, *Chemical science*, 2021, **12**, 10802–10809.

J. Han, H. Kang, S. Kang, Y. Kwon, D. Lee and Y.-S. Choi, *Physical Chemistry Chemical Physics*, 2022, **24**, 26870–26878.

H. Chen, T. Liang, K. Tan, A. Wu and X. Lu, *Journal of Cheminformatics*, 2024, **16**, 132.

H. Xu, Z. Zhou and P. Hong, *arXiv preprint arXiv:2311.13817*, 2023.

Y. Li, H. Xu, A. Kumar, D.-S. Wang, C. Heiss, P. Azadi and P. Hong, *Communications chemistry*, 2025, **8**, 51.

D. Weininger, A. Weininger and J. Weininger, *J Chem Inf Comput Sci*, 1988, **28**, 31–36.

D. Bajusz, A. Rácz and K. Héberger, *Journal of cheminformatics*, 2015, **7**, 1–13.

G. W. Bemis and M. A. Murcko, *Journal of medicinal chemistry*, 1996, **39**, 2887–2893.

C. Steinbeck, S. Krause and S. Kuhn, *Journal of chemical information and computer sciences*, 2003, **43**, 1733–1739.

D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee *et al.*, *Nucleic acids research*, 2022, **50**, D622–D631.

K. Hayamizu, K. Asakura and T. Kurimoto, 57th Experimental Nuclear Magnetic Resonance Conference, Pittsburgh, PA, 2015.

T. N. Kipf and M. Welling, *arXiv preprint arXiv:1609.02907*, 2016.

K. Xu, W. Hu, J. Leskovec and S. Jegelka, *arXiv preprint arXiv:1810.00826*, 2018.

P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, *stat*, 2017, **1050**, 10–48550.

Z. Wu, P. Jain, M. Wright, A. Mirhoseini, J. E. Gonzalez and I. Stoica, *Advances in neural information processing systems*, 2021, **34**, 13266–13279.

L. Wang, Y. Liu, Y. Lin, H. Liu and S. Ji, *Advances in Neural Information Processing Systems*, 2022, **35**, 650–664.

K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *The Journal of Chemical Physics*, 2018, **148**,.

Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, *Chemical science*, 2018, **9**, 513–530.

C. Isert, K. Atz, J. Jiménez-Luna and G. Schneider, *Scientific Data*, 2022, **9**, 273.

S. Axelrod and R. Gomez-Bombarelli, *Scientific Data*, 2022, **9**, 185.

S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu *et al.*, *Nucleic acids research*, 2023, **51**, D1373–D1380.

G. Landrum, *Release*, 2013, **1**, 4.

W. Bremser, *Analytica Chimica Acta*, 1978, **103**, 355–365.

K. W. Wiitala, T. R. Hoye and C. J. Cramer, *Journal of Chemical Theory and Computation*, 2006, **2**, 1085–1092.

N. Mills, *ChemDraw Ultra 10.0 CambridgeSoft*, 2006.

M. R. Willcott, *MestRe nova*, 2009.

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.

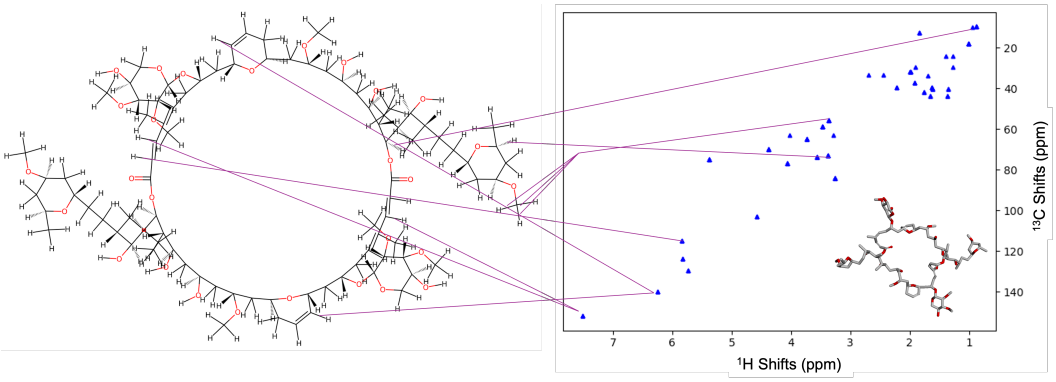

Figure 5: An annotation example. To avoid overcrowded, only a few "C-H bond – peak" associations are shown. For a large molecule with complex structure like this, aligning the chemical bonds with the cross peaks is extremely difficult due to signal overlap and degeneracy. The bottom-right of the HSQC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

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.

HOSE codes HOSE (Bremser, 1978) 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) (Wiitala *et al.*, 2006) 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* (Mills, 2006) and *Mestrenova* (Willcott, 2009). They can serve as baselines for Machine Learning based methods.

C Benchmark GNN models

C.0.1 2D GNN models

GCN Graph Convolutional Networks (GCNs) (Kipf and Welling, 2016) 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) (Xu *et al.*, 2018) 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) (Velickovic *et al.*, 2017) 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 (Wu *et al.*, 2021) 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.0.2 3D GNN models

ComENet ComENet (Wang *et al.*, 2022) 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 (Schütt *et al.*, 2018). 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 3: 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 4: 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	–	–

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims are accurate and precise in both the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our dataset is discussed in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does nt include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All model hyperparameters, random seeds and dataset are disclosed for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All code and data is released.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All training and testing details are provided in Section 4.3 and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The error measures are discussed in Section 4.2 and results provided in Section 4.3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computing resources are discussed in Section 4.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All previous work is properly cited.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no perceivable negative social impacts of our work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our data and model do not have perceivable high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All benchmark models and relative dataset is properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All related dataset an code is submitted.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.