Reviewer Response

Summary:

During the NeurIPS 2025 review period, our submission was evaluated by five reviewers, with final scores of 5, 4, 4, 4, 4. The main themes of their feedback are summarized below; point-by-point responses to each reviewer follow.

1. Include complementary metrics such as $R^2$and RMSE, in addition to MAE to more fully assess baseline performance.

2. Add visualizations that reveal error patterns and failure cases.

3. Clarify what makes 2DNMRGym unique and motivate its relevance to the ML community.

4. Add additional citation and use more detailed reference formatting.

In response to these comments, we have revised the manuscript as follows:

1. We have added the R-Squard table as Table 3, and added the RMSE table in the appendix E.

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.

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.

2. We have added more visualization to show model performance on challenges cases in Appendix F.

3. We have added more clarification of the dataset and domain challenges in the "Introduction" section and "Related work" section.

4. We have added more citation and made the citation formatting more comprehensive.

Below are our original point-by-point response and communication with each reviewer.

Reviewer 1 (Rating 4, Confidence 4):

1. The Mean Absolute Error (MAE) is used as the primary metric, but additional metrics (e.g., R^2, RMSE) or visualizations (e.g., scatter plots of predicted vs. actual shifts) could provide a more comprehensive evaluation.

Response:

We have added the RMSE and R-squared results to the appendix, as well as below. We will also add more visualizations in the revised manuscript as suggested.

1. The RMSE table

Model Type	Model	All-test RMSE (C)	All-test RMSE (H)	Few-shot RMSE (C)	Few-shot RMSE (H)	Zero-shot RMSE (C)	Zero-shot RMSE (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)
2D GNN	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)
2D GNN	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)
3D GNN	SchNet	6.1147 (0.0280)	0.4725 (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)
Transformer	GIN-Trans	5.4041 (0.0034)	0.3828 (0.0036)	3.3318 (0.0087)	0.2852 (0.0030)	7.5851 (0.0177)	0.4932 (0.0048)
Transformer	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)

2. The R-Squared table

Model Type	Model	All-test R^2 (C)	All-test R^2 (H)	Few-shot R^2 (C)	Few-shot R^2 (H)	Zero-shot R^2 (C)	Zero-shot R^2 (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)
2D GNN	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)
2D GNN	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 (0.0001)	0.9456 (0.0011)	0.9335 (0.0008)	0.9147 (0.0007)
3D GNN	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)
Transformer	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)
Transformer	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)

2. The paper briefly mentions the dataset's current limitation to HSQC spectra. A deeper discussion on the challenges of extending this work to other NMR techniques (e.g., HMBC, COSY) would be valuable. Beyond HMBC and COSY, are there plans to include other NMR techniques or multi-modal data (e.g., combining NMR with mass spectrometry)?

Response:

Our current dataset focuses on HSQC spectra, which are widely used in NMR due to their relatively fast acquisition and clear ^{1}H–^{13}C correlation signals. Extending the dataset to more

advanced NMR techniques such as HMBC, COSY, or TOCSY presents notable challenges, as these experiments often require significantly longer acquisition times and more complex setups. As a result, it is difficult to obtain large-scale, annotated datasets for these modalities. While small numbers of such spectra are available in databases like HMDB (https://www.hmdb.ca/) and NPMRD (https://np-mrd.org/), data scarcity remains a key barrier. Our team is actively exploring ways to collect and curate these datasets to support the broader research community. As you suggested, we also recognize the potential of incorporating multi-modal data, such as combining NMR with mass spectrometry (MS), especially tandem MS (MS/MS), to enhance molecular structure elucidation. However, a major challenge in this space is the frequent absence of one or more modalities in experimental datasets. For instance, many compounds may have NMR spectra but lack corresponding MS data. This raises an exciting AI challenge: how can we design robust multi-modal models that perform well despite missing modalities? Tackling this will require developing approaches that handle incomplete data effectively, such as modality dropout, data imputation, or weak supervision. We view this as a promising and impactful direction for future research.

3. Include an analysis of cases where models perform poorly. For example, are errors concentrated in specific molecular scaffolds or regions of the chemical shift range? For the 23 molecules where the algorithmic annotations were partially correct, what were the common sources of error? Were they related to specific structural features or spectral artifacts?

Response:

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. We appreciate the reviewer's suggestion and will highlight this finding, together with some visualizations of the annotation errors, in the revised manuscript.

4. Add more visual examples of HSQC spectra with annotations, especially for cases with spectral overlap or degeneracy, to illustrate the annotation challenges.

Response:

We will include additional visual examples of HSQC spectra with annotations in the revised manuscript. Specifically, we will highlight cases with spectral overlap and chemical shift degeneracy to better illustrate the inherent challenges in annotation. These examples will help clarify the complexity of atom-level assignments and demonstrate the value of our benchmark.

Reviewer 2 (Score 4, Confidence 3):

1. *Limited Novelty in Dataset Source and Preparation*

One concern lies in the contribution related to dataset construction. If I understand correctly, the 2DNMRGym dataset is derived by combining two existing sources—HMDB and CH-NMR-NP, as described around lines 115–118. As such, it's difficult to fully appreciate the novelty or effort in data collection itself. The main contributions then appear to be the post-processing steps: generating pseudo-labels and gold-standard labels, and computing additional metadata (e.g., via RDKit). While I do not doubt the effort involved, I believe the authors should more clearly highlight what makes this dataset distinct from previous work—particularly emphasizing the surrogate labeling strategy. This clarification is especially important for readers to better appreciate the contributions of this work like myself, who lack a strong chemistry background, and for the broader machine learning audience at this venue. Making the unique value proposition of the dataset more explicit would strengthen the work considerably.

Response:

While our dataset builds on HMDB and CH-NMR-NP, these sources provide only raw, unannotated HSQC spectra, which are not directly usable for machine learning. Curating a fully experimental HSQC dataset requires substantial domain expertise (typically at the Ph.D. level) to interpret ^{1}H–^{13}C correlations, resolve overlapping signals, and perform extensive annotation and validation. This process is complex, labor-intensive, and central to our contribution. Importantly, by framing HSQC analysis as a fine-grained atom-level prediction task grounded in real experimental data, our benchmark creates opportunities for evaluating and developing models that go beyond traditional graph-level property prediction. This can help drive advances in molecular representation learning, transfer learning, and semi-supervised modeling. We believe this resource will be valuable for the ML community interested in learning from rich, structured scientific data.

Our work introduces:

(1) Atom-level annotations and high-quality surrogate/gold-standard labels, which are not available in the source datasets;

(2) A fine-grained atom-level prediction task, which goes beyond standard graph-level benchmarks and promotes richer molecular representation learning;

(3) 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.

For comparison, a recent NeurIPS dataset paper ("A Multimodal Spectroscopic Dataset for Chemistry") relies primarily on simulated HSQC data with minimal experimental validation and does not include real HSQC measurements.[1] This highlights the challenge of obtaining real

HSQC data at scale, and we believe our dataset fills an important gap. We will clarify these contributions more explicitly for the broader ML audience in the revised manuscript.

[1] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry. In Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track, 2024.

2. Lack of Motivation for Deep Learning in This Context

I think the topic of this work that uses ML or DL-based approaches for molecule prediction and the corresponding atom-level representation learning is a bit lack of motivations. For example, a natural question that arises in my mind why is deep learning necessary here? Is the task inherently difficult for traditional methods, or is the goal to achieve better performance? While I assume the latter, this is never clearly stated or justified in the paper. A brief discussion of the practical limitations of non-ML or classical methods, and how ML models improve upon them, would help contextualize the contribution and clarify the problem setting.

Response:

The task addressed in this work is inherently challenging: HSQC spectra contain complex information about atom–atom couplings, especially in larger or more flexible molecules. Acquiring experimental data is already resource-intensive, but annotating that data (assigning each cross-peak to a specific atom pair) is even more time-consuming. This process currently relies heavily on expert interpretation and domain knowledge, often requiring trained chemists with advanced degrees (eg. PhD) and years of experience.

Unlike 1D NMR, which is more amenable to rule-based analysis, 2D spectra such as HSQC encode higher-order structural relationships that are difficult to model algorithmically. The interpretation of 2D NMR spectra such as HSQC has traditionally required expert reasoning and manual inspection to resolve signal overlap, deduce connectivities, and assign atoms, particularly in structurally complex compounds.[2] Classical (non-ML) computational methods for HSQC analysis are limited, and no widely adopted, automated solutions exist that provide both accurate peak prediction and atom-level annotation. Therefore, 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.

We will include this discussion in the updated manuscript to better contextualize the problem and highlight the potential impact of ML-based approaches in this space.

[2] Bross-Walch, Nadja, et al. "Strategies and tools for structure determination of natural products using modern methods of NMR spectroscopy." Chemistry & biodiversity 2.2 (2005): 147-177.

3. Missing Comparison to the Pseudo-Labeling Method

Since the pseudo-labels used in the dataset are generated via an existing method (which also appears to be ML-based), I find it puzzling that the benchmark results in Table 2 do not include a direct comparison to this original labeling method. Such a comparison would be crucial for understanding how well the GNN baselines perform relative to the model used to generate the training labels. Without this, it is difficult to interpret the MAE values in Table 2. After all, this setup closely resembles a teacher-student or data distillation paradigm, where we wouldn't expect student models to outperform the teacher—especially if the pseudo-labels are noisy. Without an "apples-to-apples" comparison, the benchmarking results feel incomplete and their implications unclear.

Response:

The silver-standard labels are used only for training to enable scalability, while all evaluations are performed on gold-standard, expert-annotated labels. This setup ensures that model performance is assessed independently of any potential bias introduced during pseudo-labeling. To ensure "apples-to-apples" comparison, we have included the performance of the original labeling method to Table 2. The comparison of this method and the best benchmark results are shown below. As expected, the benchmark results fall short compared to the original method, suggesting many possible ML directions for further improvements.

Model	13C MAE (ppm)	1H MAE (ppm)
TransPeakNet (Original)	2.025 (0.129)	0.167 (0.006)
GIN-Trans (Benchmark)	2.348 (0.031)	0.198 (0.000)

4. Inconsistent Evaluation Metrics Between Tables

A more minor but still confusing issue is the use of different evaluation metrics across Table 1 and Table 2. Table 1 reports accuracy (suggesting a classification setup), while Table 2 reports MAE (suggesting regression). This discrepancy makes it hard to understand the nature of the underlying prediction tasks. Line 204 mentions that a matching algorithm is used to assign C–H bonds within the molecular graph, which may partially explain the metric switch—but for readers unfamiliar with chemistry (like myself), the rationale for these choices is unclear. A brief explanation of why different tasks and metrics are used—and what each is measuring—would significantly improve readability and interpretability.

Response:

The use of accuracy in Table 1 is to illustrate the validation of pseudo-label quality. Specifically, it assesses how well the automatically generated labels match known assignments in a subset of data. In contrast, the main task of this work is atom-level peak prediction, which is formulated as a regression problem and evaluated using mean absolute error (MAE), as shown in Table 2. We will further clarify this distinction in the manuscript.

Reviewer 3 (Score 5, Confidence 3):

1. The benchmark task is limited to cross-peak shift prediction. While this is a meaningful task, the work could be strengthened by demonstrating additional downstream applications such as peak assignment, structure elucidation, or molecule retrieval.

Response:

We appreciate the suggestion and agree that these downstream tasks are important directions. The current dataset is indeed applicable to tasks such as peak assignment, structure elucidation, and molecule retrieval by appropriately framing spectral data as input and structural outputs or matches as targets. In fact, the original method used to create the pseudo-labels used an automated annotation approach, which is a promising approach to address the annotation bottleneck in NMR analysis. However, such annotation fundamentally depends on accurate spectra prediction, which is precisely the challenge our dataset aims to address.

While these downstream applications are compelling, they remain long-standing open problems in the chemistry and cheminformatics communities, and there is currently a lack of robust AI models capable of solving them reliably. Conducting benchmarks using simple models may not yield meaningful insights at this stage. That said, we will explicitly mention the potential applicability of our dataset to these tasks and encourage future work in these directions.

2. While overall annotation accuracy is reported there is limited discussion of failure cases. Understanding these failure modes would provide useful guidance for users of the dataset.

Response:

We agree that analyzing failure cases is important for understanding the limitations of both the dataset and models trained on it. In the revised manuscript, we will include a discussion of cases where the annotations were partially accurate, potentially due to mismatches in crowded spectral regions, symmetric atoms, and molecules with rare functional groups. By highlighting these failure cases, we aim to provide users with clearer expectations of where current models struggle and to guide future improvements in model design and dataset usage.

Reviewer 4 (Score 4, Confidence 4):

1. **Limited novelty beyond dataset construction**: The core methodological contributions are minimal, with the paper mainly describing dataset curation rather than introducing new models or insights.

Response:

This is a submission to the Datasets and Benchmarks Track. The main contribution is a large-scale, fully experimental HSQC dataset—rare, hard to curate, and highly valuable. In addition, we provide comprehensive baseline models to support future method development.

Respectfully, your comment on the "limited novelty" of our work does not capture its contributions. Our innovation lies in the careful curation of the dataset and the adoption of a surrogate learning approach to enable large-scale machine learning. Additionally, we employ the gold standard as a universal benchmark for evaluating all models, ensuring unbiased performance comparisons.

2. **Shallow benchmarks**: The baseline experiments rely on standard GNN architectures with routine hyperparameter tuning, and do not deeply analyze failure modes or structural patterns specific to 2D NMR.

Response:

We respectfully disagree with this assessment. Our manuscript includes comprehensive benchmarking across major model families, including 2D GNNs, 3D GNNs, and transformer-based architectures, with detailed analysis provided in Lines 266–280. It is unclear what is specifically meant by "failure modes or structural patterns specific to 2D NMR." If more specific suggestions had been provided instead of general comments, we would have been glad to address them directly.

3. **Lack of chemical interpretability**: There is insufficient discussion about whether the learned representations capture chemically meaningful patterns, beyond reporting numerical MAE metrics.

Response:

This comment suggests some misunderstanding of the task and its evaluation standards, possibly due to limited background in chemistry. Our atom-level chemical shift prediction task uses MAE, which is the standard and chemically meaningful metric: small deviations directly correlate with accurate structural and electronic interpretations by chemists. Strong model performance inherently reflects the learning of chemically relevant patterns.

Furthermore, the ability to generalize across diverse molecular contexts such as zero-shot and few-shot learning further supports the chemical relevance of the learned representations.

4. **Potential annotation biases**: The "silver-standard" pseudo labels depend on a prior model trained on related data, potentially introducing systematic biases. The paper does not rigorously analyze this.

Response:
It is unclear what specific bias is being referred to. The silver-standard labels are used only for training to enable scalability, while all evaluations are performed on gold-standard, expert-annotated labels. This setup ensures that model performance is assessed independently of any potential bias introduced during pseudo-labeling.

5. **Related work coverage is insufficient**: The paper fails to discuss *Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry* (NeurIPS 2024), which introduced a significantly larger dataset of 79,000 HSQC spectra and more diverse benchmarks. Ignoring this prior work weakens the novelty claims and gives an incomplete perspective.

Response:

We are aware of this work and have added the citation in the revised manuscript. We did not include it initially because it is based entirely on simulated HSQC spectra and contains no experimental HSQC data. To evaluate similarity with real spectra, the authors collected experimental data for only 251 molecules—excluding HSQC. This highlights the substantial effort required to obtain and annotate experimental 2D NMR data, particularly HSQC. Our dataset, constructed entirely from fully experimental HSQC spectra, directly fills this gap.

This distinction is clearly stated in both our manuscript and the NeurIPS 2024 paper itself, and appears to have been overlooked in the review.

6. **Reference formatting**: The bibliography is incomplete and lacks proper paper titles, making it hard to verify the cited works and reducing the clarity of the literature discussion.

Response:

According to the NeurIPS 2025 format guidelines, "any choice of citation style is acceptable as long as you are consistent." That said, we will include paper titles in the revised manuscript.

Additional Communications with Reviewer 4:

Reviewer:

Thank you to the authors for your detailed rebuttal.

Regarding shallow benchmarks and chemical interpretability: I apologize for not expressing my comments clearly in my previous review. What I meant was that the analysis of the benchmark experimental results in the paper could be made more in-depth, rather than questioning the completeness of the evaluated model architectures. Specifically, MAE is a relatively coarse, molecule-level metric. If the paper could further point out for which types of atoms these models make more accurate predictions, for which types they are less accurate, and discuss the underlying reasons, such deeper analysis could provide more valuable insights for future research. I agree with the authors that submissions to the benchmark track are not evaluated solely on technical novelty, but whether they can provide valuable insights to the field is important.

Considering that my other concerns have been partially addressed, I will accordingly raise my score.

Response:

Thank you for the clarification and for taking the time to revisit our submission.

Regarding your comment on "which types of atoms these models make more accurate predictions," we hope to get some clarification. Since our model only predicts the chemical shifts of carbon and hydrogen atoms, are you referring to the types of local molecular environments (e.g., scaffolds, chiral centers, functional groups, etc.)? If so, we have conducted additional analysis during the rebuttal stage. While the errors did not cluster around specific scaffolds or fall into distinct chemical shift ranges, we observed that structurally complex molecules (for example, those with flexible ring systems or multiple chiral centers) tended to have higher annotation errors. Moreover, atoms that are pseudo-chemically symmetric and exist in nearly identical environments (e.g., single-bonded methylene groups in similar positions) also presented challenges and led to higher prediction errors. We will include this discussion in the main text and add visualizations of these examples in the appendix of our modified manuscript.

If there are any remaining concerns that you feel have only been partially addressed, we would be grateful if you could point them out. We are happy to provide further clarification or additional details as needed. The discussion period is still ongoing, and we will do our best to address any remaining questions or feedback.

We sincerely appreciate your thoughtful feedback and your decision to raise the score. Thank you again.

Reviewer 5 (Score 4, Confidence 4):

1. Lack of recent graph transformer baselines such as GraphGPS[1] or GPS++[2].

[1] Rampášek L, Galkin M, Dwivedi VP, Luu AT, Wolf G, Beaini D. Recipe for a general, powerful, scalable graph transformer. Advances in Neural Information Processing Systems. 2022 Dec 6;35:14501-15. [2] Masters D, Dean J, Klaeser K, Li Z, Maddrell-Mander S, Sanders A, Helal H, Beker D, Fitzgibbon AW, Huang S, Rampášek L. GPS++: Reviving the Art of Message Passing for Molecular Property Prediction. Transactions on Machine Learning Research.

Response:

These two papers provide a universal framework for transformer based GNNs, as well as promising results on Chemistry datasets. We have added these two papers in our references. During the rebuttal period, we made active efforts to integrate the GraphGPS pipeline into our framework, as the GPS++ is a specific configuration from the general recipe proposed by the GraphGPS framework. However, several technical challenges (detailed below) limited our ability to fully adapt and tune the model for optimal performance, within the given timeframe. Nonetheless, we were able to test the GraphGPS + RWSE setup (same as the ZINC benchmark, with modifications for node-level regression task), with which we obtained the following preliminary results:

Evaluation Data	C MAE (ppm)	H MAE (ppm)
All-test MAE	4.4681	0.2465
Few-shot MAE	4.3720	0.2350
Zero-shot MAE	4.7240	0.2741

We acknowledge that with more extensive parameter tuning and task-specific adaptations, GraphGPS performance could likely improve. As such, these results should not be interpreted as a fair comparison to our other benchmarks. We plan to include a better-optimized configuration in the camera-ready version by running more experiments.

Several technical mismatches limited full integration.

1. Task misalignment: GraphGPS is primarily designed for graph-level or node-level classification tasks, whereas our objective involves fine-grained, atom-pair-level regression between carbon and hydrogen atoms, which is beyond the default task formulations supported by the GraphGPS pipeline.
2. Non-uniform output structure: Our benchmark models generate outputs only for specific node pairs (e.g., C–H pairs), and incorporates pairwise relationships, requiring a custom loss and evaluation routine that the uniform-output paradigm of GraphGPS does not readily accommodate.
3. Solvent-aware modeling: Our benchmark architectures dynamically inject solvent-class embeddings into node features during prediction. This conditioning is critical for

chemical shift accuracy but requires per-node customization. This is currently not implemented in the GraphGPS pipeline, which could also impact model performance.

We view the variety of GraphGPS configurations as a valuable addition to the GNN+Transformer architecture we have already benchmarked. Due to the implementation challenges outlined above, we aim to produce a more competitive configuration in our camera-ready version.

2. the paper only includes atom-level task while graph level and link level also exist for molecular property prediction. More diverse task types might also be considered.

Response:

We appreciate the suggestion and agree that more diverse task types are important directions. Our benchmark focuses on atom-level prediction, which is well aligned with HSQC spectra, as they provide atom-resolved $^{1}H–^{13}C$ correlation signals. The dataset is also applicable to graph-level tasks, such as structure elucidation and molecule retrieval, by framing spectral data as input and structural outputs as targets. However, these tasks remain long-standing challenges in the chemistry community, and current AI models are not yet robust enough to solve them reliably. Benchmarking with simple models may not yield meaningful insights at this stage. That said, we will explicitly mention the potential of our dataset to support such tasks and encourage future work in this direction. Regarding link-level prediction, we note that this is not meaningful in the context of HSQC, as these spectra do not directly encode bond-level information between heteroatoms, such information typically requires HMQC or HMBC experiments.

3. What is the performance benefit of using GNN architectures over the algorithm used for silver-standard labels? It already achieves 95.21% on the test set, do we expect the ML models to perform better or more efficientt? How long does it take to generate pseudo labels for 21,869 molecules?

Response:

Thank you for the question. The reported 95.21% accuracy reflects annotation accuracy, which is used to validate the quality of the silver-standard labels. However, the primary goal for the 2DNMRGym is to improve chemical shift prediction, which is treated as a regression task and evaluated using mean absolute error (MAE). Using the 2DNMRGym, ML communities now have a large-scale dataset with atom-level labeling, which enables many directions in graph representation learning, structural elucidation and molecule retrieval tasks. The possible directions include:

1. Improved prediction accuracy by leveraging richer representations of molecular structure and dynamics.

2. Develop architectures for semi-supervised learning, where models can be trained on partially labeled data and evaluated using gold-standard labels.
3. Greater generalizability across molecular scaffolds and experimental conditions. With the right model architectures and training strategies, it is possible to achieve significantly lower MAEs, thereby enabling more accurate downstream applications such as peak assignment and structure elucidation.

Regarding efficiency, generating pseudo-labels for 21,869 molecules is computationally fast. On average, it takes approximately 0.2 seconds per small molecule (<500 Da) and 0.3 seconds per large molecule (>500 Da). While we used a single V100 GPU, inference remains fast even on CPU.