Neurips Review and Meta Review

Meta Review:

This work presents a new dataset of 2D NMR spectroscopy data, focusing on HSQC (Heteronuclear Single Quantum Coherence) spectra. Such data is interesting and can support significant advances in machine learning methods for chemistry applications. Indeed, it seems clear from the paper, reviews, and discussion, that this is indeed the first large-scale annotated dataset of its type. Moreover, on the topic of annotations, the authors propose here a combination of "gold standard" hand-labeled samples and "silver standard" pseud-labels obtained algorithmically through a surrogate learning approach, providing a practical solution for supporting large-scale training and reliable evaluation. Reviewers have raised some concerns initially regarding novelty, significance, interpretability, or potential for new insights, but as far as I can see, these were all sufficiently addressed in rebuttal, and can reasonably be addressed in the final revision of the manuscript. Following the rebuttal and discussion period, all reviewers were satisfied with the clarifications provided by the authors, and have either raised or maintained their ratings on the positive side, ranging from accept to borderline accept (noting in their final justification that they indeed support accepting the work - even where technically only marking borderline accept). Therefore, I confidently recommend accepting this work.

Reviewer 1:

Summary:

The paper introduces 2DNMRGym, a novel dataset for machine learning applications in 2D NMR spectroscopy, specifically Heteronuclear Single Quantum Coherence spectra. The dataset comprises over 22,000 experimental HSQC spectra, with annotations derived from both algorithmic predictions and expert validation. The authors benchmark several GNN models for atom-level chemical shift prediction, demonstrating the utility of the dataset for molecular representation learning.

Strengths Contributions:

1. The paper addresses a significant gap in the field by providing the first large-scale, annotated experimental dataset for 2D NMR (HSQC) spectra. This fills a critical need for ML applications in NMR analysis.
2. The combination of algorithm-generated (silver-standard) and expert-annotated (gold-standard) labels is innovative. It allows for scalable training while maintaining rigorous evaluation standards.
3. The paper evaluates a diverse set of GNN models (2D, 3D, and transformer-based) on the dataset, providing a solid foundation for future research.
4. The dataset, code, and detailed experimental settings are openly shared, facilitating reproducibility and further research.

Limitations Weaknesses:

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

Ethical Considerations: No, there are no or only very minor ethics concerns
Ethical Comments:

If these concerns can be addressed, I am willing to improve my rating.

Dataset Code Accessibility: Yes
Dataset Code Comments:

Data and code is open-source and available at: https://github.com/siriusxiao62/2DNMRGym.

Additional Feedback:

If these concerns can be addressed, I am willing to improve my rating.

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:

Borderline accept. I thank the authors for their detailed rebuttal. I am generally satisfied with the clarifications provided.

Reviewer 2:

Summary:

This work proposes a new chemistry-related dataset, 2DNMRGym, based on tthe 2D NMR data that can be used for atom-level molecular representation learning. To address the significant workload associated with manual annotation, the authors employ a surrogate supervision strategy: the majority of the labels are generated algorithmically, resulting in pseudo-labels, while a smaller subset is annotated by human experts to serve as gold-standard references. Several machine learning models are then evaluated on this dataset to benchmark atom-level representation learning performance.

Strengths Contributions:

- The paper is well-written and easy to follow, even for readers without a strong background in chemistry (such as myself).
- The proposed dataset appears to be the first large-scale collection based on 2D NMR spectroscopy data, which could be a valuable contribution to the broader scientific community.
- The surrogate supervision strategy—combining algorithmically generated pseudo-labels with a smaller set of expert-annotated gold-standard labels—offers a practical solution to the high cost of manual annotation.
- A range of machine learning models are employed to provide a comprehensive benchmark of atom-level representation learning performance on the proposed dataset.

Limitations Weaknesses:

Despite the merits of this work, I have several concerns regarding its soundness:

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.

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.

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.

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.

Please let me know if I missed anything

Ethical Considerations: No, there are no or only very minor ethics concerns
Dataset Code Accessibility: Yes
Dataset Code Comments:

The dataset is presented and located in github with clear instructions on its usage

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes

Final Justification:

Most of my concerns raised in the review previously have been mitigated, including the clarification on novelty, the motivation to using DL-based approaches, and the fair comparison between different methods.

Reviewer 3:

Summary:

This paper introduces 2DNMRGym, a large-scale, experimentally-annotated 2D NMR (HSQC) dataset. This dataset fills a crucial gap in machine learning for NMR spectroscopy, where high-quality, annotated 2D spectra are extremely scarce. The dual-labeling strategy—using algorithm-generated silver-standard annotations for training and expert-labeled gold-standard annotations for evaluation—is well-motivated. It balances scalability with evaluation reliability and provides a realistic setup for weak supervision in scientific domains.

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. While overall annotation accuracy is reported there is limited discussion of failure cases, such as chemical environments where the pseudo-labeling struggles. Understanding these failure modes would provide useful guidance for users of the dataset.

Strengths Contributions:

This paper introduces 2DNMRGym, a large-scale, experimentally-annotated 2D NMR (HSQC) dataset. This dataset fills a crucial gap in machine learning for NMR spectroscopy, where high-quality, annotated 2D spectra are extremely scarce. The dual-labeling strategy—using algorithm-generated silver-standard annotations for training and expert-labeled gold-standard annotations for evaluation—is well-motivated. It balances scalability with evaluation reliability and provides a realistic setup for weak supervision in scientific domains.

Limitations Weaknesses:

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.

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.

Ethical Considerations: No, there are no or only very minor ethics concerns
Dataset Code Accessibility: Yes
Rating: 5: Accept: Technically solid paper, with high impact on at least one sub-area of AI or moderate-to-high impact on more than one area of AI, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes

Reviewer 4:

Summary:

This paper introduces 2DNMRGym, the annotated experimental dataset for two-dimensional nuclear magnetic resonance (2D NMR) spectroscopy, with a focus on HSQC spectra. The authors curate over 22,000 experimental spectra, annotated using a dual-layer strategy: algorithm-generated pseudo-labels (silver-standard) for scalable training, and a smaller subset of human-verified gold-standard labels for evaluation. They benchmark a set of 2D and 3D GNN models on the cross-peak prediction task. The dataset and code are open-sourced.

Strengths Contributions:

1. **Relevance**: 2D NMR spectroscopy is a domain with significant challenges for manual interpretation, and this dataset addresses a clear need.
2. **Large-scale experimental data**: The dataset of over 22,000 experimental HSQC spectra is valuable.
3. **Open-source**: Making both data and code publicly available improves reproducibility and benefits the community.

Limitations Weaknesses:

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

Ethical Considerations: No, there are no or only very minor ethics concerns
Dataset Code Accessibility: Yes
Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:

I decided to raise my score to Borderline Accept, since the authors added more insightful analysis in the rebuttal and addressed my other concerns. The proposed dataset is meaningful for advancing development in the related field.

Reviewer 5:

Summary:

This work introduced 2DNMRGym, the first large-scale, annotated experimental dataset for machine learning on 2D NMR data, specifically HSQC spectra. The authors provided 22,000+ spectra paired with molecular graphs and SMILES strings to support molecular representation learning. A surrogate supervision setup was also proposed, using algorithm-generated labels for training and expert annotations for more robust evaluation. The authors also benchmarked various 2D and 3D GNN and GNN-transformer models for atom-level prediction task. The dataset and code are also open-sourced.

Strengths Contributions:

The main strength of the paper is:

- This work proposed the first large scale annotated dataset of 22,348 experimental HSQC spectra. Data acquisition is time-consuming and requires sensitive instrumentation.
- The authors proposed a dual-layer annotation strategy, algorithm-generated annotations are silver-standard supervision, while the expert-labeled subset provides a gold-standard set.
- The paper is easy to follow and the code and dataset is open-sourced.

Limitations Weaknesses:

limitations and weakness of the work is as follows:

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

[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.

Ethical Considerations: No, there are no or only very minor ethics concerns
Ethical Comments:

No ethical concern.

Dataset Code Accessibility: Yes

Dataset Code Comments:

Yes, the code and dataset is provided on the project github.

Rating: 4: Borderline accept: Technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct Acknowledgement: Yes
Responsible Reviewing Acknowledgement: Yes
Final Justification:

I recommend for the acceptance of this work. The authors have addressed my concerns to a sufficient degree and the collected HSQC spectra data would be beneficial for the community.