Leveraging Pre-Trained LMs for Rapid and Accurate Structure Elucidation from 2D NMR Data

Susanna Di Vita
ETH Zurich
sdivita@ethz.ch

Florian Grötschla
ETH Zurich
fgroetschla@ethz.ch

Luca A. Lanzendörfer
ETH Zurich
lanzendoerfer@ethz.ch

Roger Wattenhofer
ETH Zurich
wattenhofer@ethz.ch

Abstract

Molecular structure elucidation from NMR data is a crucial process in chemistry, particularly for applications on small and medium molecules in materials science. Despite advances in computational methods, traditional approaches remain time-consuming and data-intensive, necessitating the exploration of more efficient and automated solutions. We propose a novel application of a pretrained language model (LM) for structure elucidation using 2D NMR data, marking the first instance of such an approach with experimental data. Our method generates SMILES strings representing molecular structures by conditioning on both HSQC peaks and the molecular formula, achieving a 74% accuracy rate. This surpasses the previous state-of-the-art achieved with simulated data. By leveraging a pretrained model, our approach requires significantly less data and compute. To our knowledge, this work is the first to apply LMs to automated structure elucidation on 2D NMR spectra, particularly on experimental data.

1 Introduction

The elucidation of molecular structures from NMR data is a critical task in chemistry, particularly in the context of drug discovery and materials design [1]. Traditionally, this process relies on expert interpretation of NMR spectra or the use of classical computational methods, which can be both time-consuming and data-intensive. In recent years, the advent of deep learning and neural network-based approaches has offered new avenues for automating and accelerating this process [2, 3]. However, existing approaches, including those based on transformers, have focused primarily on 1D NMR data and have largely relied on simulated or synthetic datasets, limiting their practical applicability in real-world scenarios [4].

In this work, we focus on small molecules, which play a crucial role in the design of advanced materials due to their versatile chemical properties and the ability to serve as fundamental building blocks for more complex structures. Small molecules are central to the development of materials with tailored functionalities, such as organic semiconductors, catalysts, and polymers [5, 6]. Their structure elucidation is, therefore, a crucial step in the accelerated design and discovery of new materials.

Transformers, originally designed for natural language processing, excel at capturing complex dependencies within sequences due to their attention mechanisms [7]. This ability to focus on different parts of the input sequence simultaneously makes transformers particularly well-suited for handling the intricate relationships found in 2D NMR data. By treating NMR peaks as a sequence of

Neural Information Processing Systems (NeurIPS) 2024 Workshop on AI for Accelerated Materials Design

information with spatial and intensity values, transformers can leverage their sequence-to-sequence framework to model these complex relationships more effectively than traditional methods.

Inference with pretrained models is considerably more computationally efficient compared to training transformers from scratch, which requires extensive datasets and significant resources for effective sequence generation and generalization [8, 9]. In chemometrics, transfer learning is commonly used to adapt models trained on large, generalized chemical datasets to specific tasks, improving accuracy in the presence of minimal task-specific data [10]. Recent advancements highlight the use of large, general-purpose transformer models, which leverage transfer learning to significantly reducing computational demands while achieving high accuracy [11].

We present a novel approach that leverages a pretrained T5 transformer model for structure elucidation from 2D NMR data. To our knowledge, this marks the first application of transformers to 2D NMR data, and, importantly, the first to do so using experimental data. Our approach conditions the generation of SMILES strings [12], which represent the molecular structure, on the 2D NMR peaks and the molecular formula, achieving an accuracy of 74% and exceeding the current state-of-the-art [4]. By utilizing a pretrained model, we significantly reduce both the amount of data and the compute required for training, achieving substantial reductions in training time, making our approach both efficient and scalable. The reduced data requirements and accelerated training times offered by our approach make it especially well-suited for high-throughput environments, where the ability to process large volumes of experimental data quickly is crucial. Furthermore, our solution is computationally inexpensive, allowing for easy fine-tuning on standard machines commonly available in materials science laboratories.

In summary, our contributions are as follows:

- We are the first to apply a pretrained LM to 2D NMR peaks and molecular formulas for small molecule structure elucidation, achieving 74% accuracy on experimental data.
- We evaluate different LMs and conditioning signals to accelerate the materials discovery pipeline, reducing data and compute requirements, and enabling high-throughput screening of molecular structures.
- We open-source our framework and data augmentation code to facilitate further research in NMR-based AI materials design.[1]

2 Method

In this work, we propose the application of a pretrained T5 transformer model for structure elucidation from 2D NMR peaks and molecular formula. To the best of our knowledge we are the first to propose such an approach for structure elucidation of 2D NMR.

Pre-trained architecture. The T5 model [13], originally developed for general natural language processing (e.g., natural language translation), operates in a sequence-to-sequence framework. We fine-tune a pretrained T5 [14] to map encoded 2D NMR spectral peaks (input) to SMILES strings (output), representing molecular structures. In our experiments, we use T5-small (60M) and T5-large (770M) to test both lightweight and expressive models. The T5 encoder processes the formula alongside the spectral data, generating a high-dimensional latent representation that captures molecular connectivity, even for complex motifs. The decoder part of the T5 model then takes this latent representation and sequentially generates a SMILES string that describes the molecular structure. The decoder operates in a step-by-step manner, predicting the next token (part of the SMILES string) at each step based on the previously generated tokens. To improve accuracy, we utilize beam search [15] to explore multiple candidate sequences, increasing the likelihood of finding the correct structure.

Our approach (see Figure 1) involved fine-tuning a conditional version of the T5 model (T5Conditional), which incorporates additional input to guide the model during training. Rather than treating the molecular formula as a token preceding the peak list, we treat is as secondary input conditioning the model on the formula, ensuring that the generated SMILES string matches the number and type of atoms.

[1] Codebase available at `https://github.com/ETH-DISCO/2DNMR-to-structure`

Tokenization. Inspired by [4], our models encodes 2D NMR peaks as a list of (x, y, intensity) values separated by ¦, where x corresponds to the chemical shift in the proton (^{1}H) dimension and y to the carbon (^{13}C) heteronuclear dimension. Values of x and y rounded to 4 decimal points and intensity normalized between 0 and 1. The intensity reflects the peak height, proportional to the number of contributing nuclei. This structured input, processed as text in the embedding layer, allows the model to interpret both spatial and intensity information as a sequence, similar to a language translation task, facilitating accurate molecular structure reconstruction when conditioned on the molecular formula.

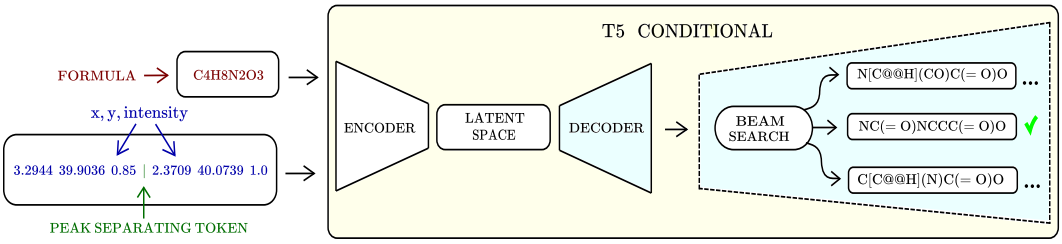

Figure 1: Overview of the tokenization process and T5 model structure for formula and 2D NMR peaks. Beam search explores multiple candidate sequences to generate the most probable SMILES. The molecular formula is used as a conditioning input to ensure atom count consistency.

2.1 Dataset

Our data set consists of 943 unique spectra taken from the Human Metabolomics Database [16] and NP-MRD [17]. Due to data availability and to maintain homogeneity, the spectra are all HSQC. All spectra correspond to experimental measurements previously defined as "Excellent" (76%) or "Very good" (24%) quality. This dataset focuses exclusively on HSQC spectra for small-sized and medium-sized molecules, all of which have SMILES strings with lengths less than 70 characters. The spectra exhibit well-resolved peaks with clear cross-peaks corresponding to heteronuclear scalar couplings between 1H and 13C nuclei with chemical shifts spread over a broad range ([-0.5097, 11.3513] for 1H and [-4.9942, 195.9225] for 13C), reflecting diverse functional groups and environments. The selected SMILES contain a limited variety of atoms, primarily including carbon (C), hydrogen (H), oxygen (O), nitrogen (N), and halogens (F, Cl, Br, I). They include up to 4 simple ring systems (e.g., benzene rings), 12 basic functional groups (e.g., -OH, -NH2, -COOH), including sequences of up to 14 single (-), double (=), and triple (#) bonds and 2 chirality centers (@@) [18].

3 Results

Given the novel application of transformers in the field of automated structure elucidation from 2D NMR data, and the lack of prior work in this specific area, we use an established baseline for 1D NMR [4], applying the same training parameters. This baseline serves as a comparison for evaluating the effectiveness of the pretrained T5 model versus a transformer trained from scratch. We applied the baseline to 2D NMR as well as to (1) only 1D proton spectra (1H), (2) only 1D carbon spectra (13C), and (3) joint 1D proton/carbon spectra (1H+13C). These control experiments allow us to assess whether the model is truly learning from the additional 2D information. We also tested the baseline on 2D NMR with the ChemBERTa tokenizer, which is trained on millions of SMILES strings, to determine whether its domain-specific handling of SMILES data would improve the model's ability to process molecular structures. All models were run for 200k steps on an RTX 3070 GPU. In addition to accuracy, our evaluation metrics include Top-5 and Top-10 measures, which assess whether the correct SMILES string appears within the top 5 or top 10 predictions generated by the model. We also measure Validity, which tracks the percentage of syntactically valid SMILES strings generated (checked with RDKit [19]), ensuring the model outputs chemically plausible structures. The percentage of correctly-generated isomers is also measured as explained in the Appendix section.

As observable from Table 1, the baseline model failed to leverage the complex relationships between the 2D NMR peaks and the molecular formula within the constraints of the given training steps. This difficulty likely stemmed from the model's inability to sufficiently capture the complex patterns due to

Table 1: Overview of model performance and training time. Base+ChemBERTa refers to the baseline approach with the ChemBERTa tokenizer. Seq. refers to models treating the molecular formula as a token, and Cond. refers to conditional models incorporating the molecular formula as additional input during training. We report Accuracy, Top-5, Top-10, Validity, Isomers, and training time (TT).

Model	Acc. (%)	Top-5 (%)	Top-10 (%)	Val. (%)	Iso. (%)	TT (hours)
Baseline 1H	30.0	43.0	47.0	96.0	25.0	8.0
Baseline 13C	35.0	47.0	52.0	97.0	30.0	8.2
Baseline 1H+13C	42.0	50.0	57.0	98.0	36.0	10.0
Baseline 2D	46.5	53.2	60.1	98.7	41.0	11.0
Base+ChemBERTa	47.7	54.1	61.3	99.0	43.5	11.0
Cross-val Formula	37.8	44.0	50.3	76.5	31.2	6.8
Cross-val Peaks	30.2	36.1	41.0	73.8	25.9	5.9
Seq. T5-small	68.8	77.5	82.1	99.8	60.2	**3.5**
Seq. T5-large	68.3	76.8	81.4	99.6	58.9	4.3
Cond. T5-small	**74.0**	82.8	87.9	**100**	70.5	5.5
Cond. T5-large	73.8	**83.2**	**88.3**	**100**	**70.8**	6.7

the relatively small size of the dataset, which may not have provided enough variability and examples for the model to generalize effectively. The differences in performance stem from the information each NMR type provides. Proton (1H) spectra offer limited structural details, while carbon (13C) spectra's wider ppm range provide more framework-specific insights, slightly boosting accuracy. Combined 1H+13C data provides complementary information from both environments but yields lower accuracy compared to 2D NMR data, which contains detailed cross-peak correlations that map atomic connectivity, indicating that the model effectively learns from the additional structural information present in the 2D spectra. The T5-large model shows slightly lower accuracy but better Top-5 and Top-10 performance because it has a larger capacity to explore and rank multiple candidate sequences, allowing it to identify and generate more relevant SMILES strings within the top predictions. Cross-attention is a powerful mechanism that allows transformers to focus on different parts of input sequences simultaneously, potentially enhancing the model's ability to capture relationships between data sources such as peaks and molecular formulas [20, 21]. In our study, we explored two cross-attention strategies: (1) cross-attention solely among 2D NMR peaks, and (2) cross-attention between the peaks and the molecular formula. As observable in Table 1, both approaches performed significantly worse than the baseline. The presence of experimental noise, combined with the high variability in the length of peak lists, caused significant challenges in relationship extraction. The model incorrectly correlated the number of peaks with the abundance of certain atoms in the formula, leading to poor generalization.

4 Conclusion

In this work, we demonstrated that fine-tuning a pretrained sequence-to-sequence model such as T5 can effectively tackle the challenge of molecular structure elucidation from 2D NMR spectra. Our approach outperformed previous state-of-the-art methods, while also being the first to operate on real experimental data. By leveraging transfer learning, we achieved high performance with relatively small datasets, reducing the need for large-scale data collection and extensive computational resources. This makes our method particularly suitable for resource-constrained environments such as standard chemometric laboratories. Unlike prior approaches that trained custom transformers from scratch, we show that fine-tuning on curated datasets of small molecules not only reduces training time but also improves accuracy. With 74% accuracy in generating correct SMILES representations, our method demonstrates the potential of pre-trained transformers to manage the complexities of 2D NMR data, offering a practical solution for small molecule structure elucidation.

Future research could explore hybrid models that combine transformers with convolutional neural networks to better capture spatial relationships in 2D NMR data. Furthermore, it could be interesting to develop long-context sequence transformers tailored to larger molecules, leveraging extended datasets to improve performance on more complex structures.

References

[1] Mikhail Elyashberg. Identification and structure elucidation by nmr spectroscopy. *TrAC Trends in Analytical Chemistry*, 69:88–97, 2015.

[2] Sarah Lindley, Yiyang Lu, and Diwakar Shukla. The experimentalist's guide to machine learning for small molecule design. *ACS Applied Bio Materials*, 7, 08 2023.

[3] Yunrui Li, Hao Xu, and Pengyu Hong. Ai-enabled prediction of nmr spectroscopy: Deducing 2-d nmr of carbohydrate, 2024.

[4] M Alberts, F Zipoli, and AC Vaucher. Learning the language of nmr: Structure elucidation from nmr spectra using transformer models. *ChemRxiv*, 2023. This content is a preprint and has not been peer-reviewed.

[5] A Mubarik, F Shafiq, HR Wang, et al. Theoretical design and evaluation of efficient small donor molecules for organic solar cells. *Journal of Molecular Modeling*, 29:373, 2023.

[6] Q Zhu and S Hattori. Organic crystal structure prediction and its application to materials design. *Journal of Materials Research*, 38:19–36, 2023.

[7] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[8] Zhuohan Li, Joseph E Gonzalez, Kurt Keutzer, and Dan Klein. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale. *arXiv preprint arXiv:2207.00032*, 2022.

[9] Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. Can transformer models generalize via in-context learning beyond pretraining data? In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2024.

[10] Herim Han and Sunghwan Choi. Transfer learning from simulation to experimental data: Nmr chemical shift predictions. *The Journal of Physical Chemistry Letters*, 12(14):3662–3668, 2021.

[11] Chengwei Zhang, Yushuang Zhai, Ziyang Gong, Hongliang Duan, Yuan-Bin She, Yun-Fang Yang, and An Su. Transfer learning across different chemical domains: virtual screening of organic materials with deep learning models pretrained on small molecule and chemical reaction data. *Journal of Cheminformatics*, 16(1):89, 2024.

[12] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[13] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[14] Hugging Face. T5 - transformers documentation. `https://huggingface.co/docs/transformers/model_doc/t5`, 2023. Accessed: August 2024.

[15] Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*, 2017.

[16] DS Wishart, AC Guo, E Oler, et al. Hmdb 5.0: the human metabolome database for 2022. *Nucleic Acids Research*, 50(D1):D622–D631, Jan 2022.

[17] D David, S Sayeeda, Z Budinski, et al. Np-mrd: the natural products magnetic resonance database. *Nucleic Acids Research*, 50(D1):D665–D677, Jan 2022.

[18] Edward D. Zanders. *Background to Chemistry of Small and Large Molecules*, pages 31–56. Springer International Publishing, Cham, 2020.

[19] RDKit: Open source cheminformatics. `https://www.rdkit.org`, 2024.

[20] Luis H.M. Torres, Bernardete Ribeiro, and Joel P. Arrais. Multi-scale cross-attention transformer via graph embeddings for few-shot molecular property prediction. *Applied Soft Computing*, 153:111268, 2024.

[21] Xinyu Wang, Le Sun, Chuhan Lu, and Baozhu Li. A novel transformer network with a cnn-enhanced cross-attention mechanism for hyperspectral image classification. *Remote Sensing*, 16(7), 2024.

[22] Eric Jonas. Deep imitation learning for molecular inverse problems. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[23] J. A. Smith and M. T. Brown. *Principles of Nuclear Magnetic Resonance Spectroscopy*. Academic Press, 2015.

[24] M. Badertscher, K. Bischofberger, and E. Pretsch. Isomer generators in structure elucidation. *TrAC Trends in Analytical Chemistry*, 16(5):234–241, 1997.

[25] A Escobedo-Morales, L Tepech-Carrillo, A Bautista-Hernández, JH Camacho-García, D Cortes-Arriagada, and E Chigo-Anota. Effect of chemical order in the structural stability and physicochemical properties of b12n12 fullerenes. *Scientific Reports*, 9(1):16521, Nov 2019.

[26] G Tahıl, F Delorme, D Le Berre, É Monflier, A Sayede, and S Tilloy. Stereoisomers are not machine learning's best friends. *Journal of Chemical Information and Modeling*, 64(14):5451–5469, Jul 2024. Epub 2024 Jul 1.

A Appendix

Data Augmentation. The selected spectra have been augmented to obtain a dataset of 190k records. We introduced a shift of up to ±0.5 ppm in the x-dimension and up to ±5 ppm in the y-dimension. These specific values were chosen based on the typical range of chemical shift variations observed in experimental NMR spectra, as defined in [22]. The rationale behind these shifts lies in the natural chemical environment's impact on NMR signals, where small perturbations in chemical structure or environmental conditions can lead to minor variations in chemical shifts [23].

Isomers. In structure elucidation, a known problem is given by generating the correct structure for different isomers [24]. In materials science, the performance characteristics of a material, such as conductivity or stability, can be heavily influenced by the specific isomeric form of its constituent molecules [25]. In this study, we ensured that our dataset included a diverse set of isomers to fully capture the complexity and challenge of the structure elucidation task. By including isomers, we ensure the model does not place excessive focus on the chemical formula, while also testing its ability to capture subtle differences in atomic connectivity and spatial arrangement [26]. In our model molecules with identical atom count and molecular mass can only be distinguished by structure by relying heavily on the analysis of the NMR peaks. In order to train our model to generalize well starting from a rather limited dataset, we ensured 12% of our dataset to be isomers (97 records).