Spectro: A multi-modal approach for molecule elucidation using IR and NMR data

Edwin Chacko[†]
Faculty of Applied Science and Engineering
University of Toronto

Rudra Sondhi[†]
Department of Chemistry and Chemical Biology
McMaster University
[†] Contributed equally

Arnav Praveen
Oakwood School, California

Kylie L. Luska
Department of Chemistry
University of Toronto

Rodrigo A. Vargas-Hernández[*]
Department of Chemistry and Chemical Biology
McMaster University
[*]vargashr@mcmaster.ca

Abstract

Molecular structure elucidation is a crucial but fundamentally challenging step in the characterization of materials given the large number of possible structures. Here, we introduce Spectro, an innovative multi-modal approach for molecular elucidation that combines ^{13}C and ^{1}H NMR data with IR. Spectro translates the embedded representations of the spectra into molecular structures using the SELFIES notation. We employed a vision model for the embedded representation of the IR data, which was pretrained to detect relevant functional group peaks in the IR spectra achieving an F1 score of 91%. For NMR data, we utilized LLM2Vec, treating the NMR spectra as text. This integration of multiple spectroscopic techniques allows Spectro to achieve an overall test accuracy of 93% when trained jointly with the vision model for the IR spectra, and 82% when trained with fixed embeddings. Our approach demonstrates the potential of multi-modal learning in tackling complex molecular characterization tasks.

1 Introduction

Molecular structure elucidation from spectroscopic data is among the most complex challenges faced by chemists and materials scientists, as the number of possible structures increases with the number of atoms. This complexity makes the molecular elucidation task formidable, requiring the use of diverse spectroscopic techniques. For instance, mass spectrometry (MS) provides insights into the potential fragments that compose the molecule and its total mass, while infrared (IR) spectroscopy helps identify the molecule's functional groups. Nuclear magnetic resonance (NMR) spectroscopy is particularly valuable as it provides detailed information regarding the connectivity, stereochemistry, and atomic environments, especially for nuclei such as protons (^{1}H) and carbons (^{13}C).

Due to the complexity of molecular elucidation and its extensive application in synthetic chemistry characterization, numerous attempts have been made to automate or assist scientists through algorithms, models, and data-driven processes. Early efforts include GENIUS [1], which employs evolutionary algorithms to generate candidate structures and predict their ^{13}C spectra using a surrogate neural network, demonstrating the potential of combining traditional chemical knowledge with

AI for Accelerated Materials Workshop, NeurIPS 2024.

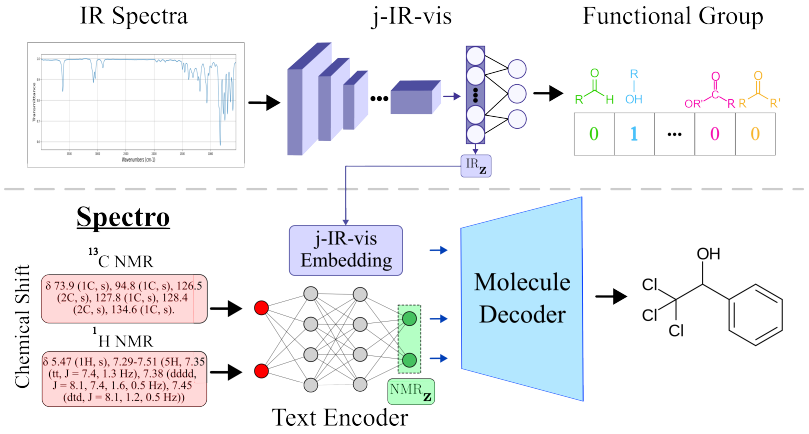

Figure 1: The diagrams of Spectro and j-IR-vis.

statistical algorithms. Other examples include logical algorithms that generate structures consistent with 1D and 2D NMR data, ranking candidates based on the alignment of predicted versus experimental chemical shifts [2–4]. SESAMI [5] interprets mass spectral fragmentation and IR absorption to build partial structures and suggest complete molecules. Some approaches [6] also incorporate techniques like Raman and ion mobility spectrometry to provide additional structural information. Recent machine learning (ML) based approaches have been proposed as well. For example, ML models have been developed to classify IR spectra based on functional groups [7–10] and to predict IR spectra from molecular structures [11, 12]. Given the importance of NMR in molecular elucidation, ML-based models have also been introduced to identify molecular substructures from NMR spectra [6, 13–16] or to match potential molecules with their spectra [17, 14, 18]. Additionally, algorithms have been designed to emulate a spectroscopist's approach to molecular elucidation, identifying small structural fragments from IR and NMR data and then connecting these fragments to build a complete structure [19].

We introduce Spectro, a multi-modal approach for the spectra-to-molecule task that integrates two different spectroscopic techniques. Our method enables a more comprehensive analysis of spectroscopic data while remaining independent of databases during inference. Spectro's workflow can be described as a spectra-to-molecule translation task, composed of three main building blocks: (i) j-IR-vis, a vision model that serves as an image encoder for the IR data, (ii) a text encoder for the 1D ^{1}H and ^{13}C NMR data, and (iii) a molecule decoder that translates the embedded vectors from the IR and NMR encoders into a molecular structure. Fig. 1 illustrates Spectro's complete workflow.

2 Methodology

The dataset used to train Spectro and j-IR-vis was compiled from multiple sources. IR spectra were downloaded from NIST [8, 20] as JDX files and plotted using Matplotlib. The resulting figures were saved in PNG format. For NMR data, we utilized predicted NMR spectral text obtained from NMRium [21]. This predicted NMR data, which simulates the spectral information without requiring experimental measurements, was processed to extract chemical shift information using Selenium WebDriver. For each molecule, the SELFIES and SMILES representations were generated using the SELFIES [22] and RDKit [23] libraries, respectively. All this information was consolidated into a YAML file for ease of access and use in the study. The total dataset comprises $6,833$ molecules, ranging from 1 to 7 functional groups per molecule. We performed an 80-20 split for training and testing both models and all results reported here are based on the test dataset of $1,366$ molecules. For more details regarding the dataset, preprocessing, and tokenization, we refer the reader to the Supplemental Material (SM).

The preprocessing of each molecule involved using the SELFIES library to convert SMILES strings into SELFIES. Each SELFIE string was then augmented with "<start>" and "<end>" tokens to mark the sequence boundaries. These augmented SELFIES were tokenized to a maximum length of 45 tokens, ensuring a consistent input size for the model. This tokenization process yielded a vocabulary of 69 unique tokens, covering the full range of molecular substructures in the dataset. The tokenized

representations were used as the target output for the molecule decoder, which was trained to predict these sequences of tokens representing molecular structures.

For the IR spectra, we employ a modified ResNet50 [24], where the size of the input images is 512×512 pixels. We call this model j-IR-vis; 25M parameters total. j-IR-vis embeds each IR spectrum into a $2,048$-dimensional vector ($^{\text{IR}}\mathbf{z}$), which is then processed by a two-layer MLP classification block followed by a sigmoid function. We pretrained j-IR-vis to predict the available 9 functional groups in the dataset using a multi-label classification approach. To address the class imbalance in the dataset (Fig. 4 in the SM) we used the weighted binary cross-entropy loss function, and the weights for each class were calculated based on the inverse of their frequency in the training dataset. See the SM for more details.

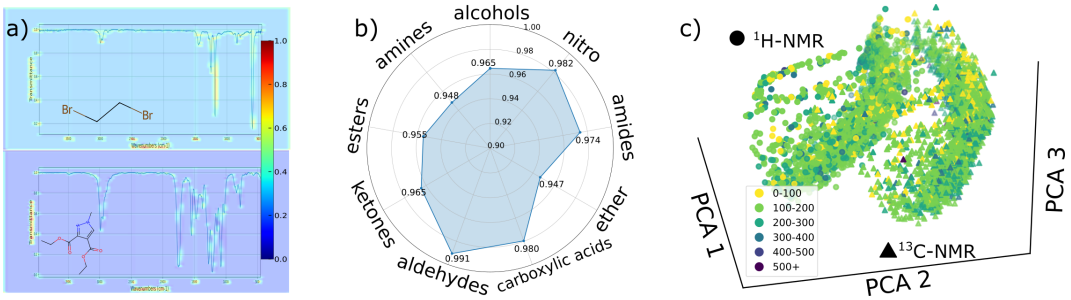

Figure 2: (a) Filters learned by j-IR-vis, the molecular structure is only for reference. (b) Accuracy of j-IR-vis for peak detection. (c) PCA of the NMR data embedded by LLM2Vec, symbols are color-coded according to the molecular weight.

Notably, j-IR-vis achieved a 94% accuracy in identifying the 9 functional groups, and an overall accuracy of 82%, see Fig. 2(b). However, it struggled with low-intensity signals, particularly in detecting functional groups such as amines and ketones. The learned filters by j-IR-vis indicate that it can detect the peak location of key functional groups, and capture the general contour of the IR absorption spectrum. For instance, Fig. 2(a) illustrates some of the relevant filters for the 1,2-Dibromoethane and Diethyl 1-methyl-1H-pyrazole-3,4-dicarboxylate molecules. We also found that the other filters capture the box-plot features of the spectra.

The use of ^{1}H and ^{13}C NMR data for molecular tasks has been primarily dominated by vision models like CNNs or transformers [25–27, 16], as well as graph neural networks [28] for the chemical shifts prediction. In contrast to IR spectra, NMR images are more challenging given the "discrete vertical peak" shape that describes the chemical shift and multiplicity. Here, we leverage the alternative representation of ^{1}H and ^{13}C NMR spectra based on text as it provides detailed information about the chemical environment of the nuclei. For example, for the molecule 2,2,2-Trichloro-1-phenylethanol, the 1D ^{1}H and ^{13}C NMR are,

^{13}C NMR: δ 73.9 (1C, s), 94.8 (1C, s), 126.5 (2C, s), 127.8 (1C, s), 128.4 (2C, s), 134.6 (1C, s). (1)

^{1}H NMR: δ 5.47 (1H, s), 7.29-7.51 (5H, 7.35 (tt, J = 7.4, 1.3 Hz),
7.38 (dddd, J = 8.1, 7.4, 1.6, 0.5 Hz), 7.45 (dtd, J = 8.1, 1.2, 0.5 Hz)), (2)

where in Text 1 for the ^{13}C NMR, each entry shows the chemical shift (δ) in parts per million (ppm), followed by the number of equivalent carbons and the multiplicity. For example, "73.9 (1C, s)" indicates one carbon atom with a chemical shift of 73.9 ppm appearing as a singlet. The notation is similar for ^{1}H NMR but includes more detailed splitting patterns. For instance, "7.35 (tt, J = 7.4, 1.3 Hz)" describes a triplet of triplets centered at 7.35 ppm with coupling constants of 7.4 and 1.3 Hz. The coupling constants represent the interaction between magnetic moments of neighboring nuclei, influencing the splitting of NMR signals. This notation helps chemists interpret the molecular structure based on the NMR spectral data. It's important to note that this NMR data is predicted, and as such, the multiplicity information is theoretical and may appear overly complicated compared to experimental spectra, where some of these couplings might not be resolved or observable.

Given the successful application of large language models (LLM) in chemistry [29], we leverage the use of LLMs as text encoders for the NMR data [30]. In particular, Spectro uses LLM2Vec [31] to build a continuous representation of the ^{1}H and ^{13}C NMR spectra. The input prompt for each NMR consisted of the individual string of the ^{1}H and ^{13}C NMR text (Text 1-2), preceded by the

instruction "Analyze the following NMR spectral data:". This prompt structure helped to provide context for the model, thereby improving the quality of the generated embeddings. Notably, LLM2Vec demonstrates the ability to differentiate between ^{1}H and ^{13}C NMR spectra without any fine-tuning, while also capturing possible correlations with molecular size, see Fig. 2(c). This can be attributed to the bidirectional attention and masked next token prediction architecture in LLM2Vec, which is advantageous for complex data like NMR spectra, where relationships between different parts of the spectrum are essential. For each NMR, the embedding text is a $4,096$-dimensional vector ($^{\text{NMR}}\mathbf{z}$).

The final component of Spectro is a molecule decoder, which utilizes a Recurrent Neural Network (RNN) architecture with four long-short-term memory (LSTM) layers; inspired from Ref. [32]. Spectro's molecule decoder predicts the tokens of the elucidated molecule using the SELFIES representation [22]. For the input, we concatenated the three embedded vectors, $[^{\text{IR}}\mathbf{z}, {}^{\text{NMR}}\mathbf{z}(^{1}\text{H}), {}^{\text{NMR}}\mathbf{z}(^{13}\text{C})]$. Due to the inherent sequential processing capability of RNNs, no additional masking or positional encoding was necessary. The molecule decoder, comprising 42M parameters total, was optimized by minimizing the sparse categorical cross-entropy loss function. We considered two training schemes, jointly training Spectro and j-IR-vis end-to-end, and utilizing fixed embedded representations from pretrained models. Our code is available at `https://github.com/ChemAI-Lab/spectro`.

3 Results and Discussions

j-IR-vis was pretrained to detect the presence of 9 functional groups from IR spectra, specifically, alcohols, amines, esters, ketones, aldehydes, carboxylic acids, ethers, amides, and nitro compounds. Notably, alkenes and alkynes were omitted from j-IR-vis's training due to their characteristically low-intensity signals in IR spectra. j-IR-vis was most successful at predicting aldehydes, amides, carboxylic acids, and nitro groups (Fig. 2(b)), as these functional groups exhibit more than one high-intensity stretching vibration in the 1500-4000 cm^{-1} region of IR spectra. Additional results are presented in Table 1.

j-IR-vis was most challenged at predicting alcohols, amines, ethers, esters, and ketones (Fig. 2(b)). While alcohols and amines possess O$-$H and N$-$H stretching vibrations that produce strong signals between 3100-3500 cm^{-1}, we believe that the model struggled to identify these groups due to the quality of IR spectra obtained from the NIST database. The IR spectra used were primarily based on gas-phase data and the O$-$H and N$-$H stretching signals possessed lower intensities than what would be expected for these functional groups. Ether functional groups exhibit only a C$-$O stretch between 1000-1300 cm^{-1}; however, this signal is contained in the fingerprint region (< 1500 cm^{-1}). The fingerprint region in IR spectra contains many stretching and bending vibrations common to organic molecules (e.g., C$-$C, C$-$O, C$-$N) and as such it may be difficult for j-IR-vis to distinguish the C$-$O signal from other signals in this region. Lastly, ketones and esters show strong signals for the C=O stretch between 1700-1750 cm^{-1}; however, neither of these functional groups exhibit another vibration that allows them to be definitively identified over other carbonyl-based functional groups as seen for aldehydes, amides, and carboxylic acids.

As outlined in the Methodology Section, we tested both the joint training of Spectro with j-IR-vis, and Spectro with fixed IR and NMR embeddings. To quantify the model's accuracy, we employed two metrics, Tanimoto Similarity (TS) and the Number of Wrong Tokens (NWT), defined as, $\text{NWT}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i}^{M} 1 - \delta(y_i = \hat{y}_i)$ where $\mathbf{y}$ and $\hat{\mathbf{y}}$ represent the tokens of the predicted and target molecule respectively, and M is the total number of tokens, $M = 45$. For all molecules, the TS was computed using RDKit with $2,048$ bits.

In Fig. 3 we present the results for the entire test dataset for the TS and NWT metrics. Spectro accurately predicts 88% of molecules (TS = 1) when trained with fixed embeddings from IR, ^{1}H, and ^{13}C NMR data, and the accuracy improves to 93% when jointly trained with j-IR-vis (Fig. 3(a)). A similar accuracy is observed using the NWT metric, with Spectro making no token prediction errors for 91% of the test molecules. Joint training of Spectro provides further insight into molecular structure through fine-tuning of j-IR-vis. Furthermore, Spectro achieved 90.42% and 95.61% on Top-1 and Top-2 sequence-wise accuracy, respectively. Spectro predicts more than four tokens in only 1.4% of the test data (see the insight in Fig. 3(a)). Finally, Fig. 3(b) and Fig. 6(e) in the SM showcase some of Spectro's predictions for the test molecules when using only NMR data, compared to when both IR and NMR embeddings are combined.

In addition to TS and NWT, we assessed Spectro's accuracy using the hydrogen deficiency index (HDI) (Fig. 6(a) in the SM), which measures the degree of unsaturation in molecules, and the

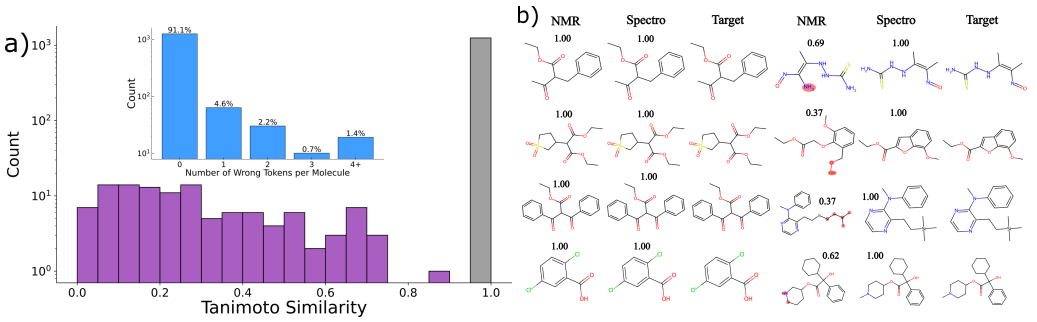

Figure 3: a) Histograms of Tanimoto Similarity and the number of wrong tokens. b) Predicted molecules with Spectro, and the TS.

molecular weight (MW) (Fig. 6(b) in the SM). Spectro shows a linear correlation between the predicted and true values of HDI and MW, indicating its effectiveness in capturing the structure and size of the molecule. Finally, our results further demonstrate that Spectro is not biased toward specific molecular sizes (Fig. 6(d) in the SM), as the mean of the TS for wrongly predicted molecules (TS<1) is around 0.25–0.3 across all molecular weights.

4 Summary

Drawing inspiration from chemists, we propose a multi-modal approach for molecule elucidation, leveraging powerful encoding models tailored to individual spectroscopy techniques. Spectro combines IR and NMR spectral information translating it to a SELFIES representation using a decoder. When trained jointly with j-IR-vis, Spectro achieves a test accuracy of 93%. The multi-modal approach allows each encoder to learn relevant information from its individual task and be easily pretrained. For IR spectra, we pretrained j-IR-vis with data that only requires knowledge of functional groups, not the entire molecular structure. For NMR, we leverage the text notation of NMR spectra, demonstrating that existing models like LLM2Vec can encode relevant molecular information. This approach could be further extended with more tailored LLM-based encoders for these representations.

In the future, we aim to extend Spectro to other spectroscopy techniques, particularly 2D NMR experiments (COSY, HSQC, HMBC) which provide critical information about atomic connectivity through cross-peak correlations. For these 2D NMR data, j-IR-vis's vision-based approach could be adapted to predict cross-peak locations and intensities, enabling the model to capture additional structural features beyond chemical shifts. The modular nature of Spectro's pipeline allows for straightforward integration of new spectroscopic data sources - the embeddings from these additional techniques could be concatenated with the existing IR and 1D NMR embeddings to enhance molecular structure prediction. Additionally, we plan to combine it with multi-task learning approaches for scenarios where specific data are missing, paving the way for more efficient and accurate structure elucidation in various chemical applications.

5 Broader Impact

Spectro represents a significant advancement in chemistry by tackling the complex challenge of molecular structure elucidation without relying on prior structural information. This innovative framework, independent of extensive preprocessing pipelines, is highly adaptable and can be extended to various spectroscopy techniques, thereby broadening its applications in chemical analysis. In self-driving laboratories (SDLs), Spectro could automate the characterization of molecules and provide deeper insights into the structures of synthesized compounds, enhancing the computationally guided platform. Furthermore, in educational settings, Spectro can support undergraduate students by aiding their understanding of molecular structure elucidation, a crucial but complex skill. Approaches like Spectro could enrich both research and educational environments. A potential limitation of Spectro is its reliance on automated pipelines without the existence of a validation scheme, which may mischaracterize chemicals in SDLs and hinder the development of essential structure elucidation skills in future chemists. However, this limitation can be mitigated with the increasing data being generated by SDLs.

6 Acknowledgments

This research was partly enabled by support from the Digital Research Alliance of Canada and NSERC Discovery Grant No. RGPIN-2024-06594.

References

[1] J. Meiler and M. Will. Genius: a genetic algorithm for automated structure elucidation from 13c nmr spectra. *Journal of the American Chemical Society*, 124 9:1868–70, 2002. doi: 10.1021/JA0109388.

[2] Alexei V. Buevich and Mikhail E. Elyashberg. Synergistic combination of case algorithms and dft chemical shift predictions: A powerful approach for structure elucidation, verification, and revision. *Journal of Natural Products*, 79(12):3105–3116, 2016. doi: 10.1021/acs.jnatprod.6b00799. URL https://doi.org/10.1021/acs.jnatprod.6b00799. PMID: 28006916.

[3] Darcy C. Burns, Eugene P. Mazzola and William F. Reynolds. The role of computer-assisted structure elucidation (case) programs in the structure elucidation of complex natural products. *Nat. Prod. Rep.*, 36:919–933, 2019. doi: 10.1039/C9NP00007K. URL http://dx.doi.org/10.1039/C9NP00007K.

[4] Mikhail Elyashberg and Antony Williams. Acd/structure elucidator: 20 years in the history of development. *Molecules*, 26(21), 2021. ISSN 1420-3049. doi: 10.3390/molecules26216623. URL https://www.mdpi.com/1420-3049/26/21/6623.

[5] Morton E. Munk. Computer-based structure determination: Then and now. *Journal of Chemical Information and Computer Sciences*, 38(6):997–1009, 1998. doi: 10.1021/ci980083r. URL https://doi.org/10.1021/ci980083r.

[6] Matevž Pesek, Andraž Juvan, Jure Jakoš, Janez Košmrlj, Matija Marolt and Martin Gazvoda. Database independent automated structure elucidation of organic molecules based on ir, ^{1}h nmr, ^{13}c nmr, and ms data. *Journal of Chemical Information and Modeling*, 61(2):756–763, 2021. doi: 10.1021/acs.jcim.0c01332. URL https://doi.org/10.1021/acs.jcim.0c01332. PMID: 33378192.

[7] D. Ricard, C. Cachet, D. Cabrol-Bass and T. P. Forrest. Neural network approach to structural feature recognition from infrared spectra. *Journal of Chemical Information and Computer Sciences*, 33(2):202–210, 1993. doi: 10.1021/ci00012a004. URL https://doi.org/10.1021/ci00012a004.

[8] Marvin Alberts, Teodoro Laino and Alain C. Vaucher. Leveraging Infrared Spectroscopy for Automated Structure Elucidation, May 2023. URL https://chemrxiv.org/engage/chemrxiv/article-details/645df5cbf2112b41e96da616.

[9] Laura Hannemose Rieger, Max Wilson, Tejs Vegge and Eibar Flores. Understanding the patterns that neural networks learn from chemical spectra. *Digital Discovery*, 2(6):1957–1968, Dec 2023. doi: https://doi.org/10.1039/D3DD00203A. URL https://pubs.rsc.org/en/content/articlelanding/2023/dd/d3dd00203a#cit46.

[10] Sriram Devata, Bhuvanesh Sridharan, Sarvesh Mehta, Yashaswi Pathak, Siddhartha Laghuvarapu, Girish Varma and U. Deva Priyakumar. DeepSPInN – deep reinforcement learning for molecular structure prediction from infrared and 13C NMR spectra. *Digital Discovery*, 3(4):818–829, April 2024. ISSN 2635-098X. doi: 10.1039/D4DD00008K. URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d4dd00008k. Publisher: RSC.

[11] Charles J McGill, Michael Forsuelo, Yanfei Guan and William H Green. Predicting infrared spectra with message passing neural networks. *Journal of Chemical Information and Modeling*, 61(6):2594–2609, May 2021. doi: https://doi.org/10.1021/acs.jcim.1c00055. URL https://dspace.mit.edu/bitstream/handle/1721.1/131020/Chemprop_IR_JCIM.pdf;jsessionid=4C22A3E51A7484A89092CDA15729E882?sequence=1.

[12] Hao Ren, Hao Li, Qian Zhang, Li Liang, W. Guo, Fang Huang, Yi Luo and Jun Jiang. A machine learning vibrational spectroscopy protocol for spectrum prediction and spectrum-based structure recognition. *Fundamental Research*, 2021. doi: 10.1016/J.FMRE.2021.05.005.

[13] Gogulan Karunanithy and D. Flemming Hansen. Fid-net: A versatile deep neural network architecture for nmr spectral reconstruction and virtual decoupling. *Journal of Biomolecular NMR*, 75(4):179–191, May 2021. ISSN 1573-5001. doi: 10.1007/s10858-021-00366-w. URL https://doi.org/10.1007/s10858-021-00366-w.

[14] Zhaorui Huang, Michael S. Chen, Cristian P. Woroch, Thomas E. Markland and Matthew W. Kanan. A framework for automated structure elucidation from routine NMR spectra. *Chem. Sci.*, 12(46):15329–15338, December 2021. ISSN 2041-6539. doi: 10.1039/D1SC04105C. URL https://pubs.rsc.org/en/content/articlelanding/2021/sc/d1sc04105c. Publisher: The Royal Society of Chemistry.

[15] Chongcan Li, Yong Cong and Weihua Deng. Identifying molecular functional groups of organic compounds by deep learning of nmr data. *Magnetic Resonance in Chemistry*, 60(11):1061–1069, 2022. doi: https://doi.org/10.1002/mrc.5292. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mrc.5292.

[16] Thomas Specht, Justus Arweiler, Johannes Stüber, Kerstin Münnemann, Hans Hasse and Fabian Jirasek. Automated nuclear magnetic resonance fingerprinting of mixtures. *Magnetic Resonance in Chemistry*, 62(4):286–297, 2024. doi: https://doi.org/10.1002/mrc.5381. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/mrc.5381.

[17] Alexander Howarth, Kristaps Ermanis and Jonathan M. Goodman. Dp4-ai automated nmr data analysis: straight from spectrometer to structure. *Chem. Sci.*, 11:4351–4359, 2020. doi: 10.1039/D0SC00442A. URL http://dx.doi.org/10.1039/D0SC00442A.

[18] Oliver Schilter, Marvin Alberts, Federico Zipoli, Alain Vaucher, Philippe Schwaller and Teodoro Laino. Unveiling the secrets of ^{1}h-NMR spectroscopy: A novel approach utilizing attention mechanisms. In *NeurIPS 2023 AI for Science Workshop*, 2023. URL https://openreview.net/forum?id=TScjG5zoB0.

[19] Matevž Pesek, Andraž Juvan, Jure Jakoš, Janez Košmrlj, Matija Marolt and Martin Gazvoda. Database independent automated structure elucidation of organic molecules based on ir, 1h nmr, 13c nmr, and ms data. *Journal of Chemical Information and Modeling*, 61(2):756–763, Dec 2020. doi: https://doi.org/10.1021/acs.jcim.0c01332. URL https://pubs.acs.org/doi/10.1021/acs.jcim.0c01332.

[20] NIST Office of Data and Informatics. Nist chemistry webbook. URL https://webbook.nist.gov/chemistry/.

[21] Luc Patiny, Damien Jeannerat, Michael Wenk, Hamed Musallam, Julien Wist, Johannes Liermann, Nils Schloerer, Alejandro Bolanos, Daniel Kostro and Michaël Zasso. Visualize, analyze and process nmr spectra online. URL https://www.nmrium.org/.

[22] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100*Machine Learning: Science and Technology*, 1(4):045024, oct 2020. doi: 10.1088/2632-2153/aba947. URL https://dx.doi.org/10.1088/2632-2153/aba947.

[23] Greg Landrum. Rdkit: Open-source cheminformatics software. 2016. URL https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4.

[24] Kaiming He, X. Zhang, Shaoqing Ren and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

[25] Denis Andzheevich Sapegin and Joseph C. Bear. Structure seer – a machine learning model for chemical structure elucidation from node labelling of a molecular graph. *Digital Discovery*, 3:186–200, 2024. doi: 10.1039/D3DD00178D. URL http://dx.doi.org/10.1039/D3DD00178D.

[26] Lin Yao, Minjian Yang, Jianfei Song, Zhuo Yang, Hanyu Sun, Hui Shi, Xue Liu, Xiangyang Ji, Yafeng Deng and Xiaojian Wang. Conditional molecular generation net enables automated structure elucidation based on 13c nmr spectra and prior knowledge. *Analytical Chemistry*, 95(12):5393–5401, 2023. doi: 10.1021/acs.analchem.2c05817. URL https://doi.org/10.1021/acs.analchem.2c05817. PMID: 36926883.

[27] Frank Hu, Michael S. Chen, Grant M. Rotskoff, Matthew W. Kanan and Thomas E. Markland. Accurate and efficient structure elucidation from routine one-dimensional nmr spectra using multitask machine learning, 2024. URL https://arxiv.org/abs/2408.08284.

[28] Eric Jonas and Stefan Kuhn. Rapid prediction of nmr spectral properties with quantified uncertainty. *Journal of Cheminformatics*, 11(1), Aug 2019. doi: https://doi.org/10.1186/s13321-019-0374-3. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-019-0374-3.

[29] Mayk Caldas Ramos, Christopher J. Collison and Andrew D. White. A review of large language models and autonomous agents in chemistry, 2024. URL https://arxiv.org/abs/2407.01603.

[30] Marvin Alberts, Federico Zipoli and Alain C. Vaucher. Learning the language of nmr: Structure elucidation from nmr spectra using transformer models. *ChemRxiv*, 2023. doi: 10.26434/chemrxiv-2023-8wxcz.

[31] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders, 2024. URL https://arxiv.org/abs/2404.05961.

[32] Kohulan Rajan, Achim Zielesny and Christoph Steinbeck. DECIMER: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1):65, October 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00469-w. URL https://doi.org/10.1186/s13321-020-00469-w.

A Supplemental Material

A.1 Dataset information

The dataset used in this study includes a diverse range of molecular structures, featuring 20 distinct functional groups and 20 distinct heavy atoms, including C, N, O, F, P, S, Cl, Br, Si, among others. The molecules have a mean molecular weight of 195.9 amu (σ = 90.4 amu), ranging from 25.0 to 1776.9 amu, and contain an average of 13.4 (σ = 6.4) heavy atoms. The dataset had a total of 6,833 molecules which were oversampled 3 times when training all models. As shown in Fig.4(a), the most prevalent functional groups were alcohols, ethers, and aromatics, while enols, hydrazones, and imines were the least common. Due to their low occurrence, these latter groups were omitted from the j-IR-vis model to prevent class imbalance issues. Fig.4(b) illustrates that the majority of molecules in the dataset contained an average of two functional groups. This distribution of functional groups and molecular complexity provided a robust foundation for training and evaluating the Spectro model, ensuring its ability to handle a wide variety of molecular structures commonly encountered in chemical analysis.

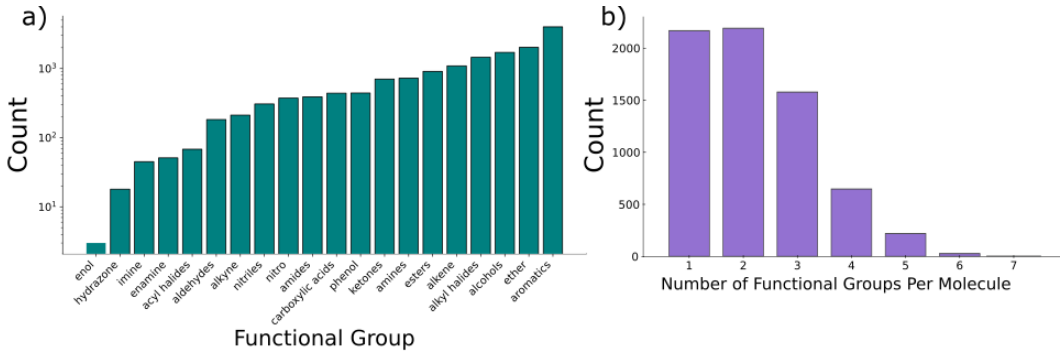

Figure 4: a) Number of molecules for each functional group. b) Number of molecules as a function of the number of functional groups.

A.2 SELFIES Preprocessing and Tokenization

The choice of SELFIES representation over SMILES was motivated by its inherent chemical validity guarantees. SELFIES' self-referencing nature ensures that even if the model makes prediction errors, the resulting molecule will still be chemically valid. In contrast, SMILES' less constrained syntax can lead to invalid predictions when used with machine learning models.

The tokenization process for SELFIES involved several steps. First, each SELFIE string was split at bracket delimiters ('[]'), which naturally separate the molecular fragments in the SELFIES notation. These fragments were collected across the entire dataset to create a comprehensive dictionary, with each unique fragment assigned a numerical token. This process yielded our final vocabulary of 69 tokens. The maximum length of 45 tokens was determined by analyzing the token length distribution across the dataset, ensuring coverage of all molecular structures while maintaining computational efficiency.

A.3 j-IR-vis model

The j-IR-vis model, based on ResNet50 [24], was trained with all layers unfrozen to allow for optimal prediction of functional groups. An MLP was attached to the ResNet50 backbone, incorporating a dropout layer with a rate of 0.5 to prevent overfitting. The dataset was split into 80% training and 20% testing sets. To optimize the training process, we employed a learning rate scheduler (`tf.keras.callbacks.ReduceLROnPlateau`) with a reduction factor of 0.2, patience of 4 epochs, and a minimum learning rate of 1e-6. The ReduceLROnPlateau scheduler dynamically adjusts the learning rate by monitoring the validation loss and reducing the learning rate when this metric has stopped improving. This adaptive approach helps in fine-tuning the model's performance, especially in later stages of training by responding to plateaus in the validation performance. L2

regularization was applied with a coefficient of 0.001 to further mitigate overfitting. The initial learning rate was set to 1e-4, with a weight decay of 0.1. To improve model generalization and handling of uncertain labels, we implemented label smoothing with a factor of 0.1, parameterized as $\text{smoothed\_labels} = \text{labels} \cdot (1 - \text{smoothing}) + \frac{\text{smoothing}}{9}$.

The IR spectroscopy images underwent several preprocessing steps to enhance their quality and consistency for model training. Initially, all images were rescaled to a uniform size of 512x512 pixels, ensuring consistent input dimensions for the model. The rescaled images were then converted to grayscale, reducing the complexity of the input data while retaining the essential spectral information. To maintain compatibility with standard image processing architectures, the single-channel grayscale images were expanded to three channels by repeating the intensity values across all axes. Further augmentation and normalization were performed using the Python library "albumentations". This process involved applying random brightness and contrast adjustments, rotation, Gaussian noise, and Gaussian blur to increase the diversity of the training data and improve model robustness. Finally, the images were normalized to have a mean of 0.8 and a standard deviation of 0.2. This normalization approach worked best for the dataset, highlighting the relevant peaks in the IR image.

For j-IR-vis, we employed the Weighted Binary Cross-Entropy (WBCE) loss function,

$$\text{WBCE}_\ell(y, \hat{y}) = (1 - y) \cdot \hat{y} + (1 + (q - 1) \cdot y) \cdot \left(\log(1 + e^{-|\hat{y}|}) + \max(-\hat{y}, 0) \right), \tag{3}$$

where y is the true preak label, $\hat{y}$ is the predicted logit, and q is the positive class weight defined as, $\text{q} = 0.5 \cdot \frac{1+N_-}{N_+}$. The WBCE loss function effectively managed class imbalance in our dataset, allowing us to optimize model performance for predicting different functional groups.

As presented in Fig. 2(a), j-IR-vis analyzes the entire IR spectrum, rather than focusing on specific peaks, integrating multiple features simultaneously. This holistic view allows the model to detect complex patterns and correlations across different spectral regions that might be missed when examining individual peaks in isolation. By considering the overall shape and subtle variations in the spectrum, j-IR-vis can capture the combined effects of multiple functional groups and their interactions.

This approach enables the model to extract rich contextual information about the molecule's structure, identifying relationships between spectral features that aren't obvious in traditional peak-based analysis. The ability to process and integrate information from the entire spectrum at once gives j-IR-vis a unique advantage in deducing functional groups, as it can leverage a more comprehensive set of spectral characteristics than conventional methods.

Functional Group	**Precision**	**Recall**	**F1 score**	**Support**
Aldehydes	0.92	0.97	0.94	60
Nitro	0.95	0.85	0.90	134
Amides	0.87	0.79	0.83	112
Carboxylic Acids	0.87	0.94	0.90	126
Ketones	0.83	0.88	0.86	205
Amines	0.79	0.90	0.84	210
Esters	0.84	0.96	0.90	258
Alcohols	0.99	0.93	0.95	511
Ether	0.95	0.92	0.94	594

Table 1: j-IR-vis additional results for each functional group.

In Table 1 we report the precision, recall, and F1 scores for the identification of the 9 different functional groups. The model exhibits high discriminative power, with F1 scores ranging from 0.83 to 0.95. Notably, alcohols achieve the highest F1 score (0.95) with near-perfect precision (0.99), indicating minimal false positives. Ethers and aldehydes follow closely (F1: 0.94), despite the significant disparity in their support sizes (594 vs. 60). This suggests effective learning even with class imbalance. Amides show the lowest F1-score (0.83), primarily due to a lower recall (0.79), indicating a higher false negative rate. The precision-recall trade-off is evident in groups like nitro (precision: 0.95, recall: 0.85) and amines (precision: 0.79, recall: 0.90), suggesting potential thresholding adjustments could optimize performance. The consistently high metrics across varied support sizes demonstrate the model's robustness to class distribution variations in spectral data

classification. We also displayed the confusion matrix of j-IR-vis in Fig. 5, further corroborating the high accuracy for aldehydes, carboxylic acids, and esters. From Fig. 5 we can observe that j-IR-vis has some challenges predicting ethers, as they are often misclassified as amines or esters, which aligns with the lower precision observed for amines. The confusion between amides and ketones is also apparent, explaining the lower F1 score for amides.

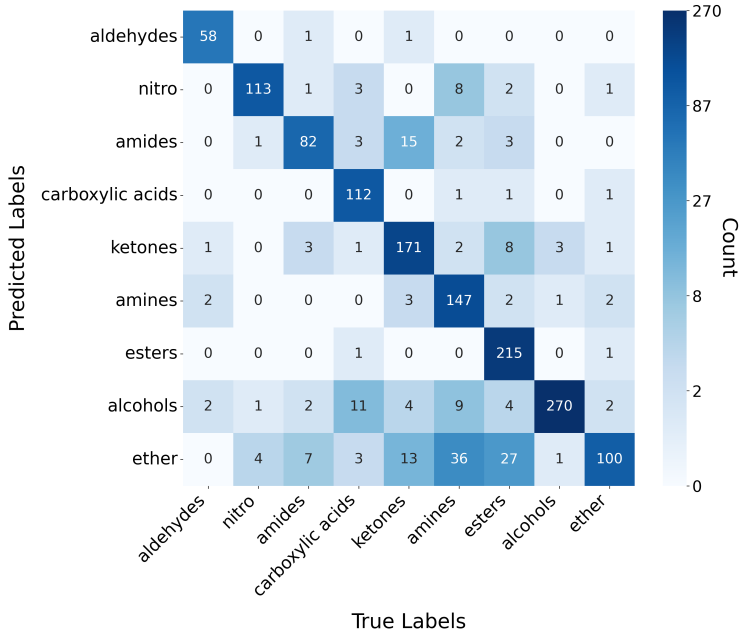

Figure 5: Confusion matrix showing the performance of j-IR-vis in predicting functional groups from IR spectra.

j-IR-vis was most successful at predicting aldehydes, amides, carboxylic acids, and nitro groups (Fig. 2(b)). In general, these functional groups exhibit more than one high-intensity stretching vibration in the diagnostic region (1500-4000 cm^{-1}) of IR spectra. Aldehydes, amides, and carboxylic acid all show a strong signal for the C=O stretching vibration between 1640-1740 cm^{-1} and a second accompanying signal that allows that to be distinguished from other carbonyl-based groups: (1) aldehydes possess a C-H stretch (two signals between 2700-2900 cm^{-1}); (2) primary and secondary amides a N−H stretch (one or two signals between 3100-3500 cm^{-1}); and (3) carboxylic acids a O−H stretch (one signal between 2400-3400 cm^{-1}). Nitro groups show a N−O stretching vibrations that produce two signals between 1475-1550 and 1290-1360 cm^{-1}.

A.4 Molecule Decoder

For the task of SELFIE token prediction, Spectro employs a sparse categorical cross-entropy loss function, implemented using `tf.keras.losses.sparse_categorical_crossentropy`. This loss function is well-suited for SELFIE token prediction as it efficiently handles multi-class classification problems, aligning with the nature of token prediction for molecules represented as text. The sparse categorical cross-entropy loss is defined as,

$$\text{SCCE}(y, \hat{y}) = -\sum_{i=1}^{N} \hat{y}_i \log(y_i), \tag{4}$$

where $\hat{y}$ and y represent the true and predicted tokes, respectively, in the probability domain. N is the number of samples.

Spectro's molecular decoder is designed for embeddings-to-molecule tasks, where the embeddings are derived from j-IR-vis for the IR spectra, and LLM2Vec for the 1D NMR spectra, where the molecule is represented using the SELFIES-token space. The block that reads the embeddings and predicts the sequence of tokens is an RNN with four LSTM layers. The total number of parameters of the decoder is 62M and for the joint model with j-IR-vis are 67M.

A.5 Additional Results

Here, we present additional results for molecular structure prediction from IR and NMR spectra. Additional predicted molecules by Spectro are displayed in Fig. 6(e). To further validate the proposed model for the molecule elucidation task, we also studied the model's accuracy for different-sized molecules quantified through the hydrogen deficiency index (HDI) and the molecular weight (MW). The HDI quantifies the number of multiple bonds and rings in a molecule, providing some additional information about the molecules' structure. The predicted vs the true HDI for all the test molecules is displayed in Fig. 6(a), where a linear trend can be observed. The HDI prediction achieved an R^2 value of 0.96 (Fig. 6(a)), indicating a strong correlation between the predicted and true values. These results further demonstrate the robustness of the model.

Additionally, we found that Spectro correctly predicts the MW. In Fig. 6(b), we demonstrate a linear trend between predicted and true MW. Furthermore, in Fig. 6(c) and Fig. 6(d), we show that for wrongly predicted molecules, where the Tanimoto Similarity (TS) is lower than one (TS<1), the model is not biased to different size molecules neither a specific functional group.

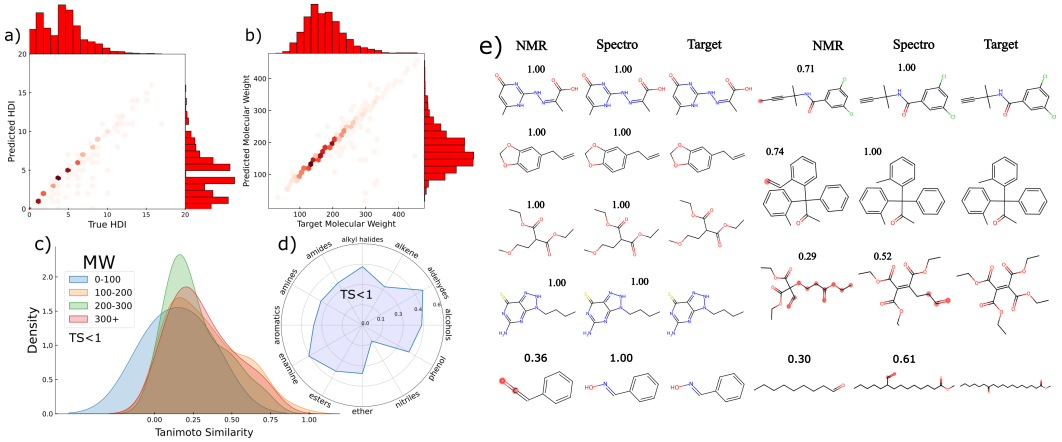

Figure 6: a) Predicted vs True HDI. b) Predicted vs True molecular weight. c) Molecular weight of the predicted molecules with TS lower than 1. d) Tanimoto similarity of the predicted molecules with TS < 1 for different functional groups. e) Predicted molecules and the TS.

A.6 Hardware information

All experiments were conducted using Nvidia 3080 and Nvidia 2070 GPUs with 10 and 8 GBs each, on a system equipped with 46GB RAM, 256GB storage, and an Intel Core i5-8400 CPU. j-IR-vis and all Spectro variants were trained using a single Nvidia RTX 3080 for 8 and 20 hours respectively. Due to memory constraints, extracting the NMR embedding vectors required 6 GPUs (1,3080 and 5,2070s). Validation on all models was done using the mentioned GPUs. The total compute time for the presented models was roughly 300 GPU hours. Data preprocessing was performed on the aforementioned CPU. The complete project, including preliminary experiments and hyperparameter tuning not reported in the main paper, consumed an estimated total of 500 GPU hours. All computations were performed on a custom personal remote server.