Automated Structure Elucidation at Human-Level Accuracy via a Multimodal Multitask Language Model

Marvin Alberts[1,2,3] **Nina Hartramp**[3] **Teodoro Laino**[1,2]
[1]IBM Research [2]NCCR Catalysis [3]University of Zürich
marvin.alberts@ibm.com

Abstract

Structure elucidation is crucial for identifying unknown chemical compounds, yet traditional spectroscopic analysis remains labour-intensive and challenging, particularly when applied to a large number of spectra. Although machine learning models have successfully predicted chemical structures from individual spectroscopic modalities, they typically fail to integrate multiple modalities concurrently, as expert chemists usually do. Here, we introduce a multimodal multitask transformer model capable of accurately predicting molecular structures from integrated spectroscopic data, including Nuclear Magnetic Resonance (NMR) and Infrared (IR) spectroscopy. Trained initially on extensive simulated datasets and subsequently finetuned on experimental spectra, our model achieves Top–1 prediction accuracies up to 96%. We demonstrate the model's capability to leverage synergistic information from different spectroscopic techniques and show that it performs on par with expert human chemists, significantly outperforming traditional computational methods. Our model represents a major advancement toward fully automated chemical analysis, offering substantial improvements in efficiency and accuracy for chemical research and discovery.

1 Introduction

Accurately determining the structure of unknown molecules is essential across many areas of chemistry, including drug discovery, natural product analysis, and materials science. Spectroscopic techniques such as Nuclear Magnetic Resonance (NMR), Infrared (IR), and Mass Spectrometry (MS) are routinely used to analyse molecular structures. Although the acquisition of spectroscopic data has become routine and largely automated, its interpretation remains heavily reliant on expert knowledge, making it time-consuming and frequently a bottleneck in the synthesis workflow.

In recent years, machine learning approaches have been developed to assist with structure elucidation from individual spectroscopic modalities. Several studies have demonstrated that neural networks, particularly those based on transformer architectures, can predict molecular structures from NMR spectra with promising accuracy[1–7]. Similar approaches have been applied to IR data, where models learn to associate vibrational features with functional group patterns[8–12], or to link the information contained in entire spectra to precise molecular structure[13–15]. For mass spectrometry, deep learning methods have been used to infer molecular fingerprints and retrieve candidate structures from large databases[16–20].

Despite these advancements, models that rely on a single modality often encounter limitations. Spectroscopic techniques provide complementary views of molecular structure: NMR reveals local chemical environments, IR identifies characteristic bond vibrations, and MS provides mass and fragmentation data. When used in isolation, each modality can leave ambiguities unresolved, especially in cases involving structurally similar compounds or noisy experimental data.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Design.

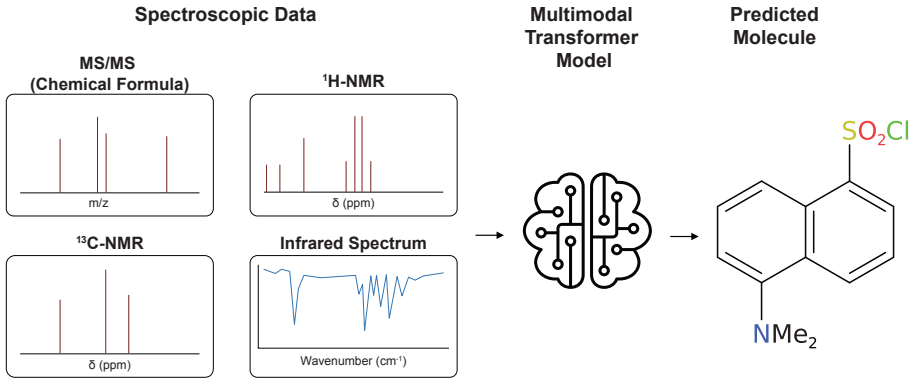


Figure 1: Combining different spectroscopic modalities. Our model is capable of combining three different spectroscopic modalities (IR, ^{1}H-NMR and ^{13}C-NMR spectra) in addition to the chemical formula and based on these inputs predicts a molecular structure.

Drawing inspiration from fields such as natural language processing, computer vision, and audio analysis, the development of multimodal, multitask machine learning models has shown clear benefits when different data sources offer complementary information[21–23]. In language tasks, for example, combining visual cues with text enhances understanding in image captioning or visual question answering[24, 25]. Similarly, in speech recognition, incorporating audio and textual transcriptions improves transcription accuracy in noisy environments[26, 27]. These successes parallel the challenges faced in spectroscopic analysis, where no single modality provides a comprehensive representation of molecular structure. Just as combining text and image strengthens semantic interpretation in language models, integrating a diverse set of analytical data allows for a more accurate and nuanced reconstruction of chemical structures Figure 1. This convergence of evidence across modalities mirrors how chemists approach structure elucidation[28], motivating the design of machine learning models that can fuse spectroscopic inputs in a similarly holistic manner. A few works have already explored the combination of various different spectroscopic modalities, however, most are limited to evaluating model performance only on simulated data and on single inference tasks[29–31].

To effectively exploit the integrated information across different spectroscopic techniques, multitask learning presents a particularly suited approach. Rather than training separate models for each type of spectrum, a multitask architecture enables a single model to learn shared representations across multiple inputs while also optimising for each modality-specific task. This design brings several practical benefits: it improves generalisation, allows the model to operate on incomplete modality combinations, thus reflecting real-world limitations, and promotes more effective use of limited experimental data by incorporating both paired and unpaired spectra. In the context of chemical structure elucidation, multitask learning encourages the model to consistently extract and reconcile information from different spectroscopic sources, improving both accuracy and interpretability.

In this work, we present a multimodal multitask transformer model able to process the molecular formula, ^{1}H-NMR, ^{13}C-NMR and IR spectra from which the molecular structure is predicted as a SMILES string. We demonstrate the effectiveness of our method through a rigorous evaluation on a set of experimental spectra, followed by a comparison of the model's performance with that of human chemists. In addition, we also investigate the synergistic of the different spectroscopic modalities and demonstrate that our model, just like human chemists, is able to glean insights from each modality and combine them to make an informed prediction.

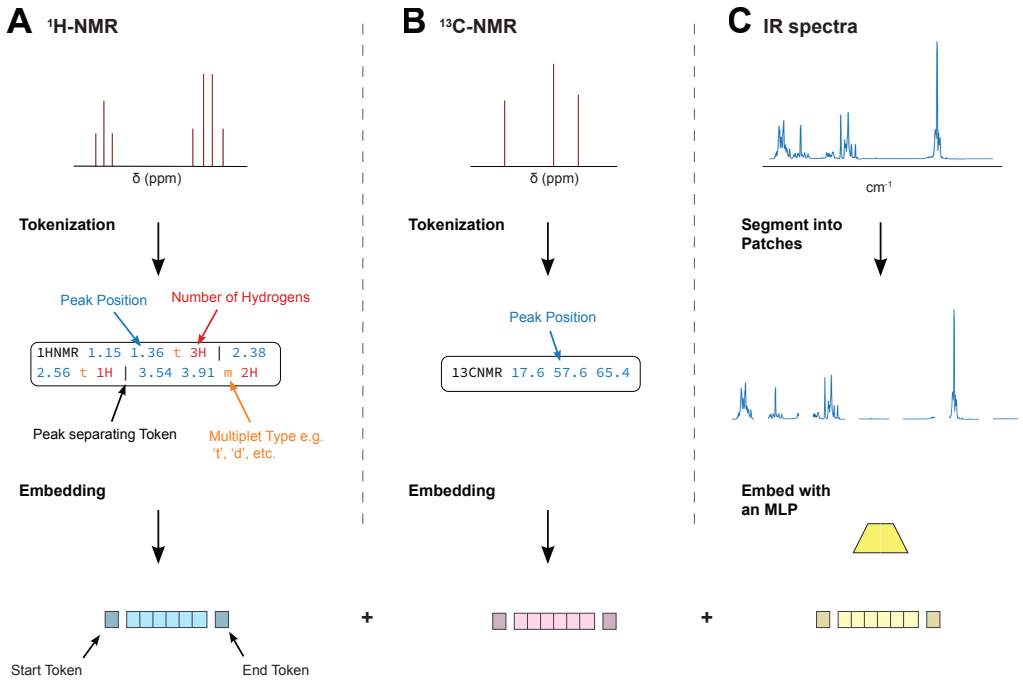


Figure 2: Tokenising NMR and IR spectra. A and B) Tokenisation procedure for ^{1}H-NMR and ^{13}C-NMR spectra respectively. C) IR spectra are segmented into patches, and each patch is projected into the embedding space via an MLP. Bottom: The embeddings of each modality are flanked with a modality start and modality end token and concatenated. The concatenation of all modalities is fed into the model.

2 Results

Fusing Spectroscopic Modalities

When chemists try to identify an unknown compound, they rely on extracting complementary structural information from orthogonal spectroscopic techniques. Here we present an AI model which, just like chemists, is able to leverage the information contained in different spectra to predict the correct molecular structure. A fundamental challenge in this approach lies in determining the optimal strategies to combine the different spectroscopic data types while preserving the information contained in them.

We selected three spectroscopic modalities, ^{1}H-NMR, ^{13}C-NMR and IR spectra, based on the complementary structural information each provides and the availability of experimental data under acceptable licensing terms. Our model predicts molecular structure as SMILES strings based on these spectral inputs combined with the chemical formula[32]. We chose to include the chemical formula as it significantly constrains the possible chemical space the model needs to explore to generate a molecule. In practical applications, the chemical formula can be obtained from mass spectrometry data.

In this work, we built upon the encoder-decoder transformer architecture. This requires that each input modality is converted either into text or a similar discrete representation compatible with a transformer model. For the chemical formula, ^{1}H-NMR, and ^{13}C-NMR spectra, text-based encoding schemes were implemented. The chemical formula is inherently textual, requiring no conversion. For ^{1}H-NMR spectra, a structured annotation capturing peak characteristics (Figure 2 (A)), including the chemical shift (beginning and end of each peak), multiplicity pattern (singlet, doublet, multiplet, etc.)[3], and integration values was used. ^{13}C-NMR spectra were encoded via a simplified annotation recording only the position of each peak (Figure 2 (B)). On the other hand, for IR spectra, a patch-based encoding approach was implemented with each spectrum segmented into frequency bands (e.g.,

Table 1: Comparison between single-task and multitask models. Top: Accuracies of four single-task models on simulated data, zero-shot on experimental data and after finetuning on experimental data. Bottom: Accuracies of one multitask model at either predicting structures from only one spectroscopic modality or all three.

Modality	Multitasking	Simulated		Zero Shot		Finetuned	
		TOP–1 ↑	TOP–5 ↑	TOP–1 ↑	TOP–5 ↑	TOP–1 ↑	TOP–5 ↑
IR	✗	39.72	57.28	1.05	4.21	17.97±5.65	44.37±6.62
^{13}C-NMR	✗	49.35	66.85	**48.42**	**66.32**	**57.75±11.26**	**85.97±9.62**
^{1}H-NMR	✗	60.08	76.74	20.00	31.58	45.83±8.14	72.89±4.21
IR + ^{13}C-NMR + ^{1}H-NMR	✗	**73.40**	**87.83**	17.89	33.68	50.04±11.25	81.26±9.08
IR	✓	33.57	53.21	2.11	11.58	15.97±8.65	47.14±8.30
^{13}C-NMR	✓	41.37	61.80	**53.16**	**82.11**	58.57±9.71	89.61±4.69
^{1}H-NMR	✓	53.44	72.82	31.58	51.58	49.96±5.99	79.22±3.91
IR + ^{13}C-NMR + ^{1}H-NMR	✓	**73.05**	**88.69**	50.53	61.05	**69.10±11.16**	**91.47±5.58**

450-575 cm^{-1}) and projected into the embedding dimension via a multilayer perceptron (MLP), as illustrated in Figure 2 (C) [14]. This approach preserves the spectral patterns present in the data. The model predicts molecules as SMILES strings, inherently a textual representation tokenised according to Schwaller *et al.* [33].

To combine the different modalities, we drew inspiration from advances in multimodal transformer architectures[25] and implemented a token-based modality demarcation strategy. Each embedded modality sequence is flanked by specific modality-start and modality-end tokens that signal boundaries between different input types (Figure 2 (bottom)) before being concatenated and fed into the model. This approach enables the transformer's self-attention mechanism to establish cross-modal relationships while maintaining awareness of each input's specific modality. This strategy also allows flexible handling of incomplete data (e.g. only the chemical formula and the IR spectrum), making it possible for the model to process any combination of the input modalities.

Multitasking outperforms single-task models

As paired experimental spectroscopic data is very scarce, we first pretrained our model on simulated spectra before finetuning on experimental data. We pretrained the models on the Multimodal Spectroscopic Dataset[34], containing paired IR and NMR spectra for 790,000 molecules. During pretraining, the model learns the SMILES string vocabulary, how to assign peaks to specific functional groups, and how to integrate different spectroscopic modalities. However, there still is a significant sim-to-real gap between simulated and experimental spectra. To address this domain shift, we finetuned the models on paired experimental spectroscopic data. In this study, we utilised the dataset published by Van Bramer & Bastin [35], subsequently referred to as the Chemical Education dataset, which contains paired IR and NMR spectra for 171 molecules. All experiments with experimental data were conducted with five-fold cross-validation, with results reported as means and standard deviations across the five folds.

To overcome the limited availability of paired spectral data, we leveraged a multitasking approach. While most structure elucidation models are single-task (e.g., predicting structure from an IR spectrum alone)[3, 13–15], our multitask model can predict molecular structure from any combination of IR, ^{1}H-NMR, and ^{13}C-NMR spectra. The rationale behind this architecture addresses the data scarcity: A multimodal single-task model might reach a local minimum by relying on only one modality while ignoring others. In contrast, our multitask model must learn not only how to perform structure elucidation from all modalities but also from each individual modality. This approach forces the

model to learn from all available data sources, prevents convergence to local minima, and effectively serves as data augmentation by allowing each sample to be represented through various combinations of modalities.

We evaluated the performance of both single-task and multitask models as shown in Table 1. For the single-task approach, we trained four separate models: each using the chemical formula with either IR, ^{1}H-NMR, ^{13}C-NMR, or all three spectral types as input. In contrast, our multitask approach required training only one comprehensive model capable of performing all four tasks. Table 1 reports Top–1 and Top–5 accuracies across three scenarios: performance on simulated data, zero-shot performance on experimental data (without finetuning), and performance after finetuning.

On simulated data, single-task and multimodal multitask models performed comparably. However, the multitask model showed slightly lower performance on individual tasks (e.g., structure prediction from IR spectra alone) relative to their single-task counterparts. For zero-shot performance, both model types achieved their highest accuracy with ^{13}C-NMR data, reflecting the smaller simulation-to-reality gap for these spectra compared to others[34], with performance decreasing in accordance with the diminishing similarity between experimental and simulated spectra.

However, the multitask model demonstrated significantly higher zero-shot performance than its single-task counterpart on the multimodal task, supporting our hypothesis that multitasking enables the model to extract more information from the spectra. The most substantial advantage of multitasking emerged during finetuning. While single-task model performance followed the zero-shot trend (with the ^{13}C-NMR model outperforming even the multimodal model), the multitask approach surpassed the single-task models on the multimodal task. The performance of the multitask model exceeded the single-task multimodal model by approximately 20% in Top–1 accuracy, while performing comparably or slightly better in individual tasks. These results indicate that multitasking facilitates more efficient data use and helps avoid local minima during optimisation, making it ideal for structure elucidation due to the limited amount of experimental data.

Multitasking enables the use of unpaired data

A key advantage of multitask models is their ability to leverage unpaired data. Here, we demonstrate that incorporating unpaired spectral data into the training of our multitask, multimodal architecture leads to a substantial performance gain, even on tasks involving paired data. This offers a potential solution to the scarcity of paired experimental datasets. To evaluate this capability, we augmented our training set with additional IR spectra from the NIST EPA gas–phase library and ^{13}C-NMR spectra from NMRShiftDB2[36, 37]. Due to the limited availability of raw ^{1}H-NMR spectra, we restricted the additional data to the other two modalities.

We first investigated the impact of modest data augmentation by adding IR spectra, ^{13}C-NMR spectra, and paired IR and ^{13}C data for an additional 600 molecules beyond the 171 molecules in our original experimental dataset. To prevent data leakage, we excluded all molecules present in the Chemical Education dataset. As shown in Table 2, performance improved across all modalities with the incorporation of additional data.

Building on these results, we expanded our dataset with an additional 33,088 molecules. These samples included 29,960 additional ^{13}C-NMR spectra, 5,027 IR spectra, and 1,899 samples for which both spectral types were present. As illustrated in Table 2, this saturated performance on the Chemical Education dataset for both the ^{13}C-NMR and multimodal tasks. Notably, performance on the ^{1}H-NMR task also improved despite no direct supplementation with ^{1}H-NMR data, likely attributable to the model's exposure to a more diverse range of molecular structures in the training set.

The IR prediction task remained the most challenging, reaching only 42% Top–1 accuracy even after incorporating 5,027 additional IR spectra. This falls short of a recent study[14], in which the IBM team reported up to 63% Top–1 accuracy on IR spectra. The performance gap is primarily attributable to domain differences between the ATR IR spectra in the Chemical Education dataset and the gas phase IR spectra present in the NIST database. Additionally, there is a larger sim-to-real gap for IR spectra than for NMR spectroscopy, requiring more substantial adaptation during finetuning.

Table 2: Adding unpaired experimental data during finetuning. The performance of the multitasking model finetuned solely on the data present in Chemical Education dataset, with the addition of 600 unpaired samples and 33,000 additional samples.

Modality	Finetuned		+ 600 Samples		+ 33,088 Samples	
	TOP–1 ↑	TOP–5 ↑	TOP–1 ↑	TOP–5 ↑	TOP–1 ↑	TOP–5 ↑
IR	15.97±8.65	47.14±8.30	21.69±6.41	52.86±9.33	42.50±12.12	75.00±6.85
^{13}C-NMR	58.57±9.71	89.61±4.69	63.55±5.84	90.55±5.37	96.25±7.50	97.50±5.00
^{1}H-NMR	49.96±5.99	79.22±3.91	52.77±3.66	78.23±7.27	93.75±12.50	96.75±2.50
IR + ^{13}C-NMR + ^{1}H-NMR	**69.10±11.16**	**91.47±5.58**	**75.41±8.32**	**91.52±1.87**	**96.25±7.50**	98.75±2.50

What are the benefits of multimodality?

While the previous sections demonstrated that a multitasking model is capable of outperforming single-task models, in this section, the synergistic effect between the different modalities is investigated in more detail. For all further experiments, we used the multitasking model trained with an additional 600 unpaired data samples. To gain deeper insight into how the different modalities complement each other, we analysed the model's performance across various subsets of the dataset with specific functional groups.

Figure 3 shows the prediction accuracy across three representative functional groups, esters (A), haloalkanes (B) and ethers (C), using models that predict molecular structure from either individual modalities (IR, ^{13}C- or ^{1}H-NMR spectra) or a combination of all three. Across all functional groups, IR spectroscopy alone reaches the lowest accuracy, while the multimodal approach, integrating all spectroscopic data, consistently achieves the highest accuracy. These results support the notion that different spectroscopic techniques provide complementary structural information.

Interestingly, the relative performance of ^{1}H- and ^{13}C-NMR varies depending on the functional group. For esters, ^{1}H-NMR outperforms ^{13}C NMR, though both are outperformed by the multimodal approach. In contrast, for haloalkanes, ^{13}C NMR demonstrates superior performance compared to ^{1}H NMR. This can be attributed to the characteristic chemical shifts of carbon atoms bonded to halogens, providing a strong identifiable signal in ^{13}C spectra. For ethers, ^{1}H- and ^{13}C-NMR perform approximately equivalently, suggesting that both spectroscopic techniques capture similar levels of structural information.

Figure 3 (D) also showcases two representative examples of molecules generated by the model. Only the first prediction of the model (Top–1) is displayed. In the first example, the model successfully generates the correct molecular structure when using ^{13}C-NMR spectra alone and multimodal data. In the second example, none of the individual prediction tasks yield the exact correct structure. However, it is noteworthy that even when predictions are incorrect, the generated molecules closely resemble the ground truth structures. This pattern of close predictions when model failure occurs highlights an important characteristic of our multitask model: It learns meaningful chemical relationships that allow it to produce chemically reasonable structures even when the exact target molecule is not identified. For instance, incorrect predictions often maintain the same functional groups, similar carbon backbone arrangements, or equivalent stereochemical features as the ground truth molecules. This suggests that the model has developed a robust understanding of spectral-structural relationships rather than merely memorising training examples. An analysis of the similarity between the predicted and ground truth molecules, backing up these claims, is provided in SI section 1.

The consistently superior performance of the multimodal approach demonstrates that integrating complementary spectroscopic techniques enables the model to overcome the limitations of individual modalities. This synergistic effect is particularly valuable for complex structural determination tasks where ambiguity in one spectral domain can often be resolved through information available in another.

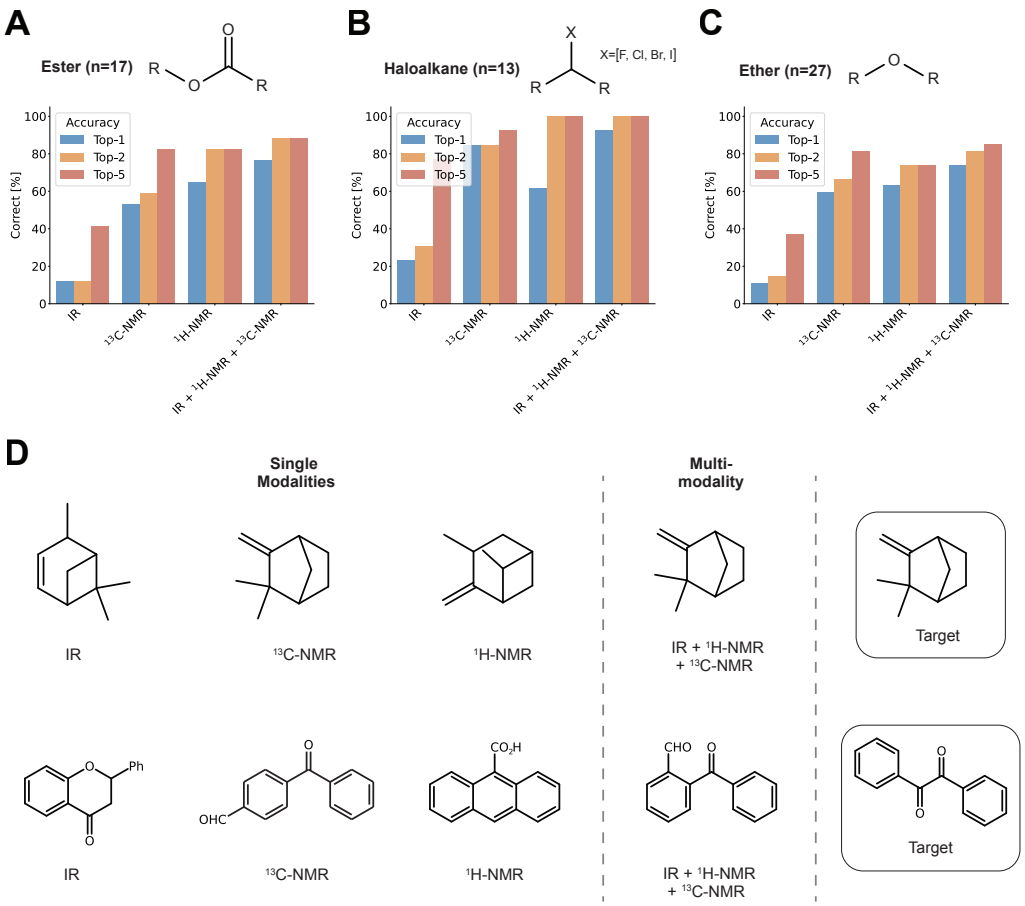


Figure 3: Combining different spectroscopic modalities: Top: Performance of the model on molecules containing either esters (A), ethers (B) or haloalkanes (C). Providing the model with multiple spectroscopic modalities leads to an increase in performance. (D): Two example predictions from the model, with predictions from single modalities as well as the combination. On top, an example where the correct molecule was predicted and at the bottom, a failure case in which the model did not predict the correct molecule.

Our Model achieves expert-level accuracies

As a final demonstration of our model's capabilities, we conducted a comparative evaluation against human expertise. We measured IR, ^{1}H- and ^{13}C-NMR spectra for 16 organic molecules, carefully selecting compounds absent from the Chemical Education dataset to prevent data leakage. Three PhD-level chemists participated in this comparison, each receiving the same input as the model: the molecular formula and raw spectroscopic data. Both human experts and the model were allowed up to five structural predictions per sample. As shown in Table 3, our model achieves performance comparable to that of expert chemists, with both reaching approximately 65% Top–1 and 80% Top–5 accuracy. While the sample size is relatively small, the results are statistically consistent and support meaningful comparisons. Importantly, the advantage of our computational approach extends beyond accuracy: It's efficiency. Where human experts typically require 10 minutes or more to analyse the spectra of a single compound, our model generates predictions in under one second. This highlights the potential for a significant increase in throughput for structure elucidation tasks while maintaining expert-level accuracy.

Table 3: Human vs. AI performance. Performance of the three models presented in this work compared against three Phd-level chemists.

	Top-1	Top-2	Top-5	Top-10
Finetuning	46.25±9.35	56.25±6.85	75.00±3.95	77.50±5.00
Finetuning + 600 Samples	56.25±5.59	65.00±7.50	77.50±5.00	80.00±2.50
Finetuning + 30k Samples	65.00±3.06	72.50±3.06	76.25±2.50	81.25±3.95
Human	62.50±5.10	77.08±2.95	79.17±5.89	N/A.

3 Conclusion

In this study, we have developed a multimodal multitask transformer model. Our findings demonstrate that a multitasking approach significantly outperforms single-task models by leveraging complementary information across different spectroscopic modalities, making more efficient use of limited experimental data. The model's architecture enables the exploitation of unpaired experimental spectroscopic data, a significant benefit as paired datasets are notoriously sparse, resulting in Top–1 accuracies of up to 96% on the Chemical Education dataset.

To validate our approach under real-world conditions, we conducted a comparative evaluation using 16 novel experimental compounds measured specifically for this study. When benchmarked against three PhD-level chemists, our model achieved comparable accuracy (65% Top–1 and 80% Top–5), while delivering results in under one second per compound, a significant improvement over the 10+ minutes required by human experts. Even when predictions were not exact matches, the model generated chemically reasonable structures that shared significant similarities with ground truth molecules.

These findings promise a shift towards AI-assisted structure elucidation in chemistry. Rather than replacing human expertise, we envision a collaborative workflow where AI models provide rapid initial predictions from spectroscopic data, allowing chemists to focus their expertise on verification, refinement, and interpretation of the results. This human-AI partnership promises to considerably accelerate the structure elucidation process, reducing a major bottleneck in chemical research, drug discovery, and materials development.

References

1. Jonas, E. *Deep imitation learning for molecular inverse problems* in *Advances in Neural Information Processing Systems* **32** (2019).
2. Sridharan, B., Mehta, S., Pathak, Y. & Priyakumar, U. D. Deep Reinforcement Learning for Molecular Inverse Problem of Nuclear Magnetic Resonance Spectra to Molecular Structure. *The Journal of Physical Chemistry Letters* **13,** 4924–4933 (2022).
3. Alberts, M., Zipoli, F. & Vaucher, A. *Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models* ChemRxiv: 10.26434/chemrxiv-2023-8wxcz. 2023.
4. Schilter, O., Alberts, M., Zipoli, F., Vaucher, A. C., Schwaller, P. & Laino, T. *Unveiling the Secrets of* ^{1}*H-NMR Spectroscopy: A Novel Approach Utilizing Attention Mechanisms* in *NeurIPS 2023, AI4Science Workshop* (2023).
5. Hu, F., Chen, M. S., Rotskoff, G. M., Kanan, M. W. & Markland, T. E. Accurate and Efficient Structure Elucidation from Routine One-Dimensional NMR Spectra Using Multitask Machine Learning. *ACS Central Science* **10,** 2162–2170 (2024).
6. Devata, S., Sridharan, B., Mehta, S., Pathak, Y., Laghuvarapu, S., Varma, G. & Priyakumar, U. D. DeepSPInN – deep reinforcement learning for molecular structure prediction from infrared and 13C NMR spectra. *Digital Discovery* **3,** 818–829 (2024).
7. Alberts, M., Hartrampf, N. & Laino, T. *From Spectra to Structure: AI-Powered* ^{31}*P-NMR Interpretation* ChemRxiv: 10.26434/chemrxiv-2025-5bd0b. 2025.

8. Fine, J. A., Rajasekar, A. A., Jethava, K. P. & Chopra, G. Spectral deep learning for prediction and prospective validation of functional groups. *Chemical Science* **11,** 4618–4630 (2020).
9. Enders, A. A., North, N. M., Fensore, C. M., Velez-Alvarez, J. & Allen, H. C. Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Analytical Chemistry* (2021).
10. Jung, G., Jung, S. G. & Cole, J. M. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chemical Science* **14,** 3600–3609 (2023).
11. Chandan Kanakala, G., Sridharan, B. & Deva Priyakumar, U. Spectra to structure: contrastive learning framework for library ranking and generating molecular structures for infrared spectra. *Digital Discovery* **3,** 2417–2423 (2024).
12. Lee, G., Shim, H., Cho, J. & Choi, S.-I. Machine-Learning Approach to Identify Organic Functional Groups from FT-IR and NMR Spectral Data. *ACS Omega* **10,** 12717–12723 (2025).
13. Alberts, M., Laino, T. & Vaucher, A. C. Leveraging infrared spectroscopy for automated structure elucidation. *Communications Chemistry* **7,** 1–11 (2024).
14. Alberts, M., Zipoli, F. & Laino, T. *Setting New Benchmarks in AI-driven Infrared Structure Elucidation* ChemRxiv: 10.26434/chemrxiv-2025-9p2dw. 2025.
15. Wu, W., Leonardis, A., Jiao, J., Jiang, J. & Chen, L. Transformer-Based Models for Predicting Molecular Structures from Infrared Spectra Using Patch-Based Self-Attention. *The Journal of Physical Chemistry A* **129,** 2077–2085 (2025).
16. Dührkop, K., Nothias, L.-F., Fleischauer, M., Reher, R., Ludwig, M., Hoffmann, M. A., Petras, D., Gerwick, W. H., Rousu, J., Dorrestein, P. C. & Böcker, S. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* **39,** 462–471. ISSN: 1546-1696 (2021).
17. Huber, F., van der Burg, S., van der Hooft, J. J. J. & Ridder, L. MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *Journal of Cheminformatics* **13,** 84 (2021).
18. Hoffmann, M. A., Nothias, L.-F., Ludwig, M., Fleischauer, M., Gentry, E. C., Witting, M., Dorrestein, P. C., Dührkop, K. & Böcker, S. High-confidence structural annotation of metabolites absent from spectral libraries. *Nature Biotechnology* **40,** 411–421 (2022).
19. Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nature Methods* **19,** 865–870. (2025) (2022).
20. Litsa, E. E., Chenthamarakshan, V., Das, P. & Kavraki, L. E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry* **6,** 1–12 (2023).
21. Li, J., Li, D., Xiong, C. & Hoi, S. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation* arXiv:2201.12086. 2022.
22. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z. & Gao, J. *Vision-Language Pre-training: Basics, Recent Advances, and Future Trends* arXiv:2210.09263. 2022.
23. Liu, H., Li, C., Wu, Q. & Lee, Y. J. *Visual Instruction Tuning* arXiv:2304.08485. 2023.
24. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. *Learning Transferable Visual Models From Natural Language Supervision* arXiv:2103.00020. 2021.
25. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A. & Simonyan, K. *Flamingo: a Visual Language Model for Few-Shot Learning* arXiv:2204.14198. 2022.
26. Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M. & Zeghidour, N. *AudioLM: a Language Modeling Approach to Audio Generation* arXiv:2209.03143. 2023.
27. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z. & Zhang, C. *SALMONN: Towards Generic Hearing Abilities for Large Language Models* arXiv:2310.13289. 2024.
28. Breitmaier, E. *Structure Elucidation by NMR in Organic Chemistry: A Practical Guide* (John Wiley & Sons, Ltd, 2002).
29. Mirza, A. & Jablonka, K. M. *Elucidating structures from spectra using multimodal embeddings and discrete optimization* ChemRxiv: 10.26434/chemrxiv-2024-f3b18-v2. 2024.

30. Tan, X. *A Transformer Based Generative Chemical Language AI Model for Structural Elucidation of Organic Compounds* arXiv:2410.14719. 2024.
31. Priessner, M., Lewis, R., Janet, J. P., Lemurell, I., Johansson, M., Goodman, J. & Tomberg, A. *Enhancing Molecular Structure Elucidation: MultiModalTransformer for both simulated and experimental spectra* ChemRxiv: 10.26434/chemrxiv-2024-zmmnw. 2024.
32. Weininger, D. SMILES, a chemical language and information system. *Journal of Chemical Information and Computer Sciences* **28,** 31–36 (1988).
33. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. & Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **5,** 1572–1583 (2019).
34. Alberts, M., Schilter, O., Zipoli, F., Hartrampf, N. & Laino, T. *Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry* in *Advances in Neural Information Processing Systems* **37** (2024), 125780–125808.
35. Van Bramer, S. E. & Bastin, L. D. Spectroscopy Data for Undergraduate Teaching. *Journal of Chemical Education* **100,** 3897–3902 (2023).
36. Stein, S. E. *NIST Standard Reference Database 35: NIST/EPA Gas-Phase Infrared Database - JCAMP Format* Accessed: 2025-03-28. 2008. `https://www.nist.gov/srd/nist-standard-reference-database-35`.
37. Kuhn, S., Kolshorn, H., Steinbeck, C. & Schlörer, N. Twenty years of nmrshiftdb2: A case study of an open database for analytical chemistry. *Magnetic Resonance in Chemistry* **62,** 74–83 (2024).
38. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25,** 1157–1174 (2004).
39. *MNova* (Accessed March 21, 2025). `https://mestrelab.com/software/mnova/`.
40. *RDKit* (Accessed April 14, 2025). `https://www.rdkit.org/`.
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
42. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L. & Liu, T.-Y. *On Layer Normalization in the Transformer Architecture* arXiv:2002.04745. 2020.
43. Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. N. *Convolutional Sequence to Sequence Learning* arXiv:1705.03122. 2017.
44. Shazeer, N. *GLU Variants Improve Transformer* arXiv:2002.05202. 2020.

A Methods

A.1 Data

When training our model, we first pretrain it on simulated data before finetuning and evaluating it on experimental spectra.

Simulated Data: The simulated data used in this study is sourced from an earlier paper introducing a multimodal spectroscopic dataset[34]. We use all molecules present in the dataset, only filtering out those also present in one of the experimental datasets. The IR spectra in this dataset were simulated with molecular dynamics using the GAFF forcefield, whereas the ^{1}H- and ^{13}C-NMR spectra were generated using MestreNova[38, 39]. This dataset is filtered, excluding molecules with a heavy atom count (all atoms except for hydrogen) outside the range of 5 to 35 and molecules containing elements other than carbon, hydrogen, oxygen, nitrogen, sulphur, phosphorus, silicon, boron, and the halogens.

Chemical Education Dataset: To finetune our model, we primarily use the chemical education data set published by Van Bramer & Bastin [35]. This dataset is designed for teaching undergraduate students how to analyse and interpret the spectra of molecules. In total, Van Bramer & Bastin [35] measured spectra for 254 molecules, however, for only 205 of them the IR, ^{1}H- and ^{13}C-NMR spectra were present. ^{1}H- and ^{13}C-NMR were annotated automatically leveraging MestreNova. Specifically, MestraNova's "AutoMultipletAnalysis" function was applied to each spectrum, yielding the beginning, end, integration and type of each peak. The integrations were rounded to the closest positive non-zero integer, i.e. all numbers below one were rounded to one. No human correction of the annotations was carried out. IR spectra were extracted manually, interpolated to a range of 450-4000cm^{-1} with a resolution of 2cm^{-1} and normalised such that the maximum absorption corresponds to 1.0. In addition, we filtered out any molecules containing atoms other than the ones contained in the simulated dataset, but due to the small size of the dataset, no filtering based on the heavy atom count was carried out. In total, this yielded 171 molecules with annotated IR, ^{1}H- and ^{13}C-NMR spectra.

Experimental ^{13}C-NMR and IR spectra: During finetuning we add additional experimental ^{13}C-NMR and IR spectra to the training dataset. The ^{13}C-NMR spectra were extracted from NMRShiftDB2, whereas IR spectra were sourced from the NIST EPA Gas-Phase database[36, 37]. All entries in both datasets were filtered to remove molecules either outside of the heavy atom range or containing elements not contained in the simulated dataset, as well as molecules present in the Chemical Education dataset, eliminating the possibility of data leakage.

Measured Experimental Data: All ^{1}H- and ^{13}C measured as part of this study were annotated with the same procedure as for the Chemical Education dataset, with IR spectra also processed in the same manner.

A.2 Tokenisation and Preprocessing

The chemical formula, ^{1}H- and ^{13}C-NMR spectra and molecules represented as SMILES were provided to the model as text with the tokenisation procedures outlined below. IR spectra were embedded via patches. We prepend a modality start token and append a modality end token before the start and end of each modality.

Chemical Formula: The chemical formula was tokenised with the following regular expression: `([A-Z]{1}[a-z]?[0-9]*)`

^{1}H-NMR spectra: All ^{1}H-NMR spectra were first converted into a text representation following [3]. For each peak in the spectrum, this representation provides information on the beginning and end in ppm, type (e.g. singlet, multiplet) and integration. All numeric values were discretised by rounding to two decimal points. "|" is used as a separator between two peaks. As an example, the spectrum `1.24 1.39 t 3H | 1.89 2.14 q 2H |` would contain two peaks, one triplet integrating to three hydrogens and one quartet integrating to two hydrogens.

^{13}C-NMR spectra: ^{13}C spectra were tokenised in a similar fashion to ^{1}H-NMR spectra converting them to a text representation. However, for these spectra, only the position of each peak

was considered and all peak positions in ppm were rounded to one decimal point. The spectrum `12.1 27.8 127.5` contains three peaks.

IR spectra: In contrast to the other modalities, we did not encode IR spectra as text and divided the spectrum into patches and projected each patch into the embedding space via a multilayer perceptron (MLP). For all experiments we used a patch size of 75.

Molecules: All molecules were canonicalised using RDKit[40] and tokenised using the same regular expression as employed by Schwaller *et al.* [33]

A.3 Model

The model employed in this work follows the encoder-decoder transformer architecture. Building upon the original implementation by Vaswani *et al.* [41] we leverage post layer normalisation[42], learned positional embeddings[43] and gated linear units[44]. The following hyperparameters were used to construct the model:

```
Layers: 6
Heads: 8
Embedding Dimension: 512
Feedforward Dimension: 2048
```

A.4 Training

Train-test splitting: For pretraining the data an 70/20/10 train, test and validation split was used. All finetuning experiments were carried out with five-fold cross-validation using the same seed to ensure reproducibility, also with a 70/20/10 train, test and validation split.

Training settings: Training of the models was carried out on two Nvidia A100 GPUs with an average pretraining time of $\sim$ 20h. When evaluating models, the best validation checkpoint was used. For each training run, the below listed training parameters were used. No distinction was made between pretraining and finetuning experiments:

```
Epochs: 60
Optimiser: AdamW
Learning Rate: 0.001
Dropout: 0.1
Warmup steps: 8000
Adam beta_1: 0.9
Adam beta_2: 0.999
Batch size: 128
```

A.5 Experimental Data Collection

A.5.1 Chemicals

2,6-Lutidine, 9-Fluoroenylmethylcarbazate, N-Boc-propargylamine, 2-Nitrofuran and Propylene Carbonate were purchased from SigmaAldrich; 4-Chloro-5-Iodo-7H-pyrrolo[2,3]dyrimidine, Boc-Gly-Oh and N-Chlorosuccinimidine were purchased from Tokyo Chemical Industry; Dansyl Chloride and Methyl Glycolate were purchased from Carl Roth GmbH; Mono-tert-butyl-succinate and N-Methylmorpholine were purchased from Apollo Scientific; 2-Ethoxycinnamic acid and 3-Chloroperbenzoic acid were purchased from Fluorochem Ltd; 2,4,6-Trimethylbenzyl alcohol was purchased from BLDPharm and Phenanthraquinone was purchased from Acros Organics.

A.5.2 Analytical Procedures

IR Spectra: IR data were acquired using a PerkinElmer Spectrum Two FTIR spectrometer with a diamond anvil ATR attachment (450–4000 $cm^{\hat{}}1$, 8 scans, 2 $cm^{\hat{}}1$ resolution).

^{1}H-NMR Spectra and ^{13}C-NMR Spectra: Approximately 50mg of each compound was dissolved in deuterated chloroform, and the solutions were transferred into 5mm NMR tubes for analysis. ^{1}H-

and ^{13}C-NMR spectra were recorded on a Bruker AV2-402 (400 MHz) in deuterated chloroform at room temperature. Chemical shifts are expressed in parts per million (ppm) and are calibrated using residual protic solvent as internal reference.

B Tanimoto Similarity

To determine how similar the predictions of the model are to the ground truth, we calculate the Tanimoto similarity between the ground truth and the top five predictions of the model. For these experiments, we excluded all samples for which the model predicts the correct molecule within the Top–5 predictions. The multitasking model, finetuned with 600 additional spectra, was utilised with the Tanimoto similarity distribution displayed in Figure 4. The median Tanimoto similarity for when the model is supplied the IR, ^{1}H-NMR or ^{13}C-NMR are 0.467, 0.464 and 0.522, respectively. On the other when all three modalities are supplied, the median Tanimoto similarity rises to 0.547

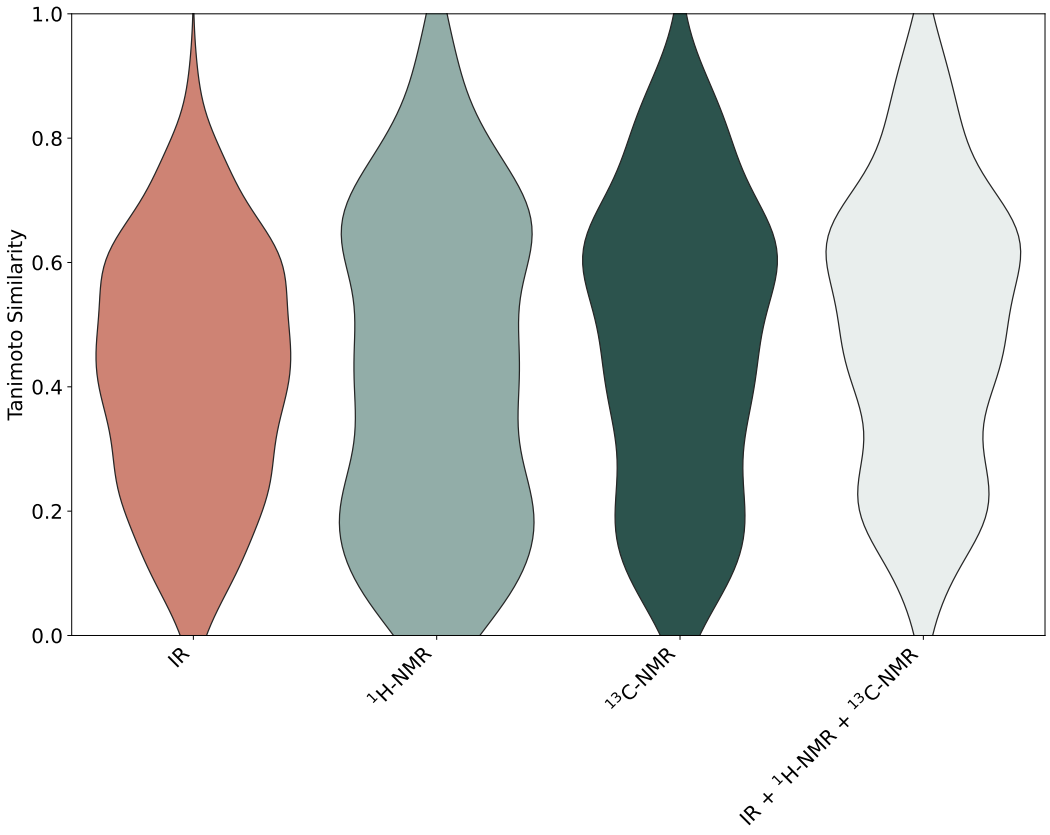
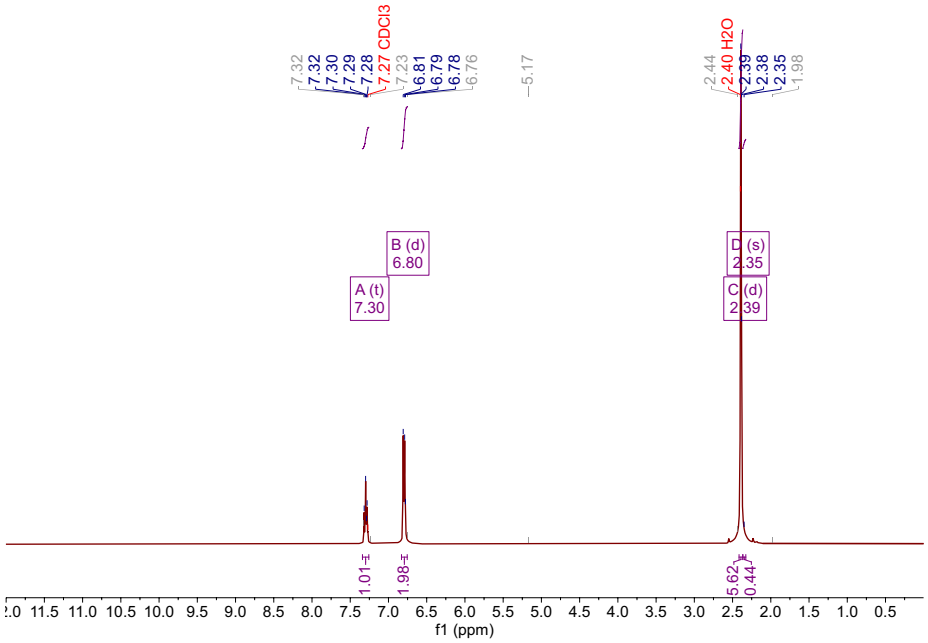


Figure 4: Tanimoto Similarity for the top five predictions for different modalities.

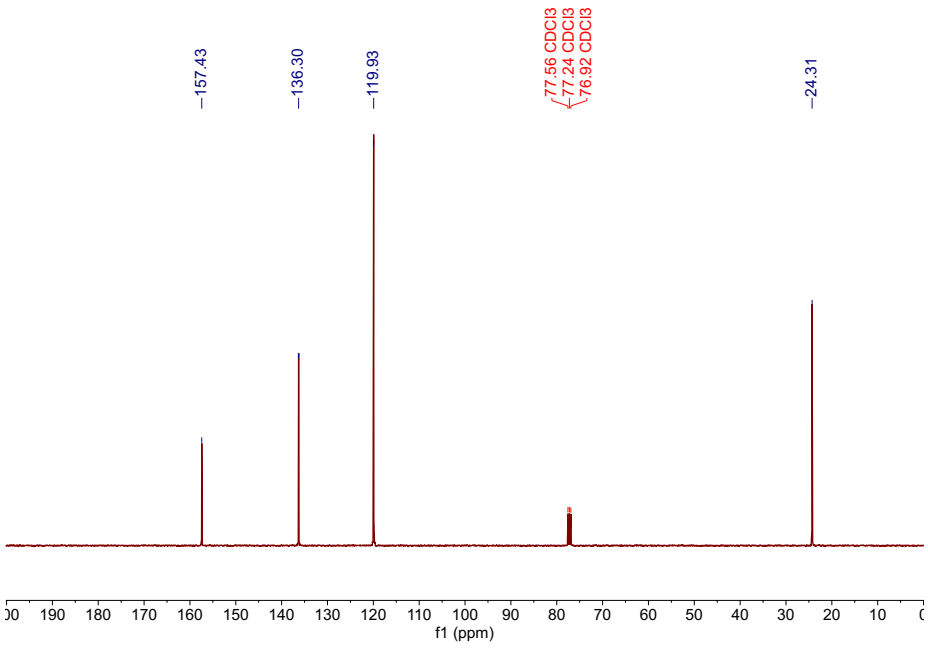
C Experimental Spectra

2,6-Lutidine

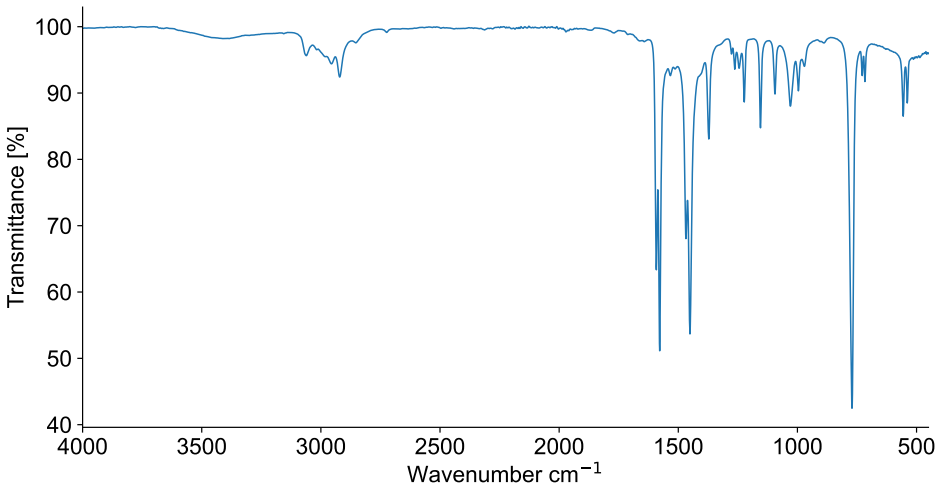
^{1}H-NMR



^{13}C-NMR

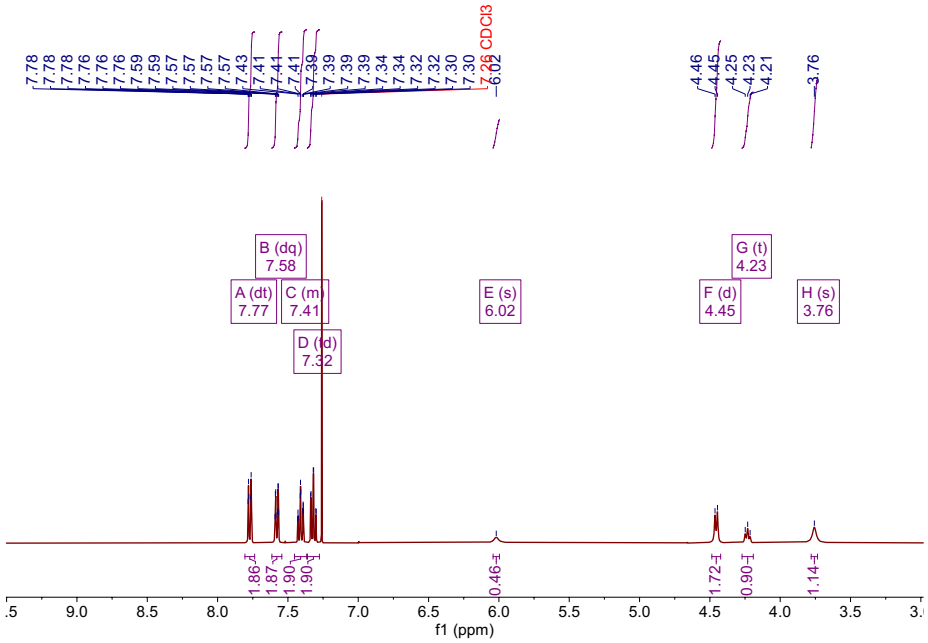


IR

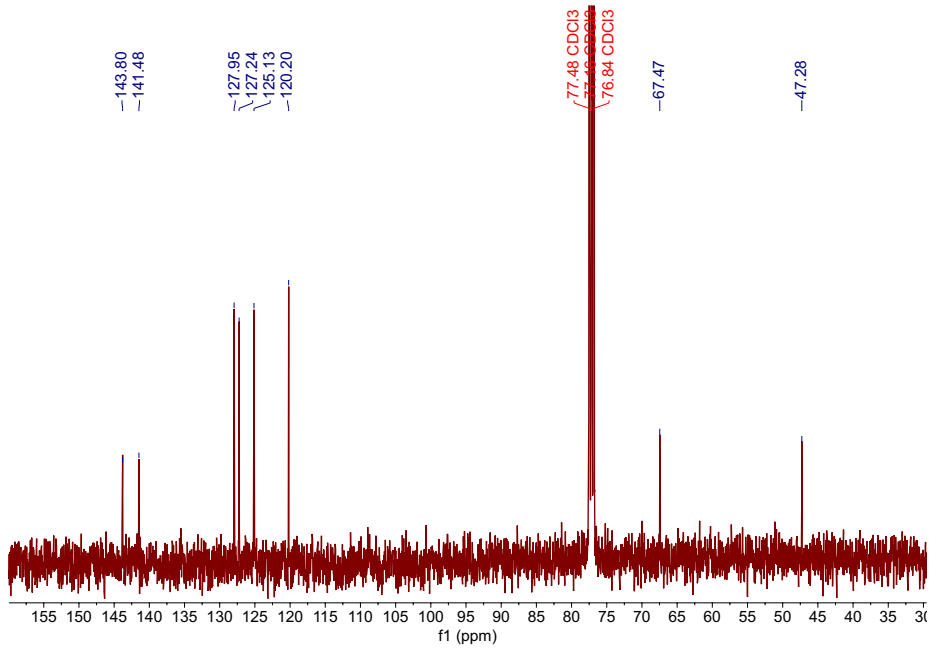


9-Fluoroenylmethylcarbazate

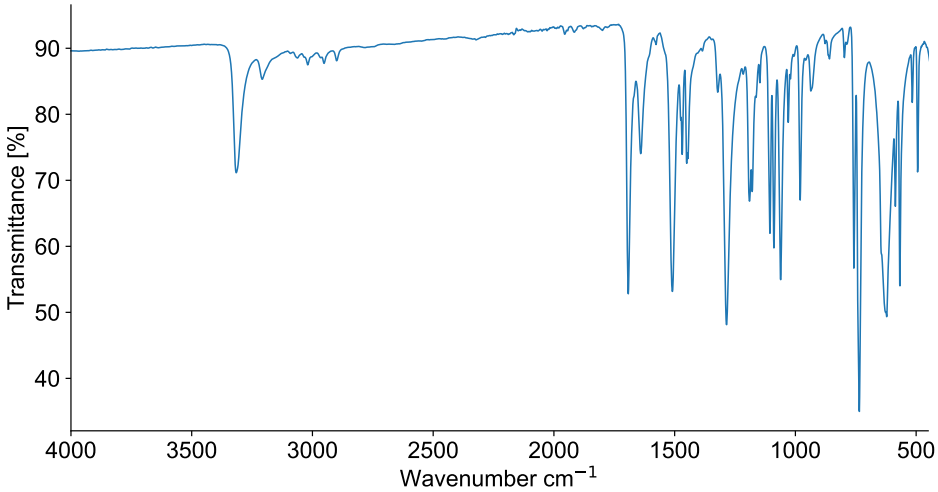
^{1}H-NMR



^{13}C-NMR

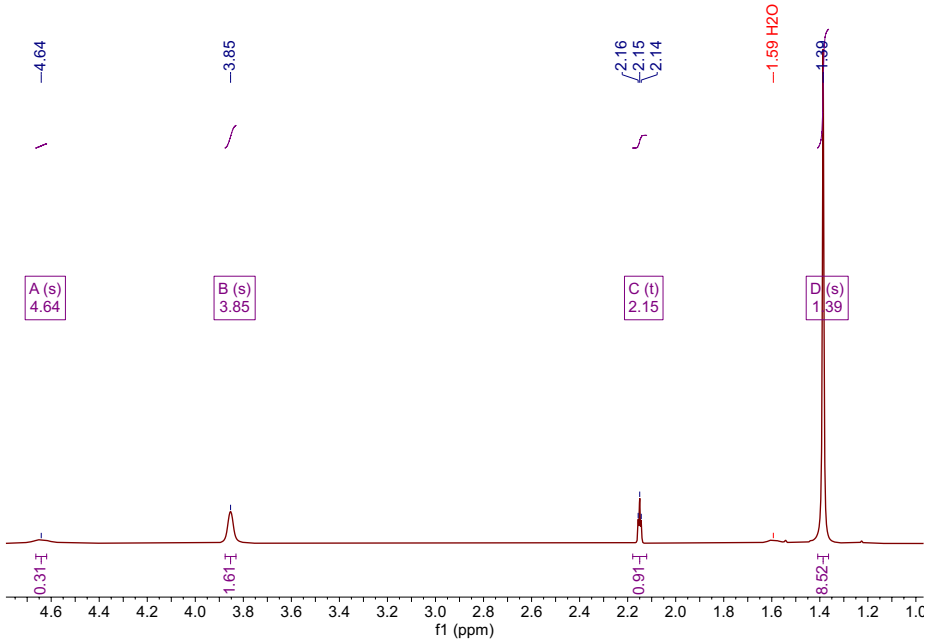

143.80
141.48
127.95
127.24
125.13
120.20
77.48 CDCl3
76.84 CDCl3
67.47
47.28
f1 (ppm)

IR

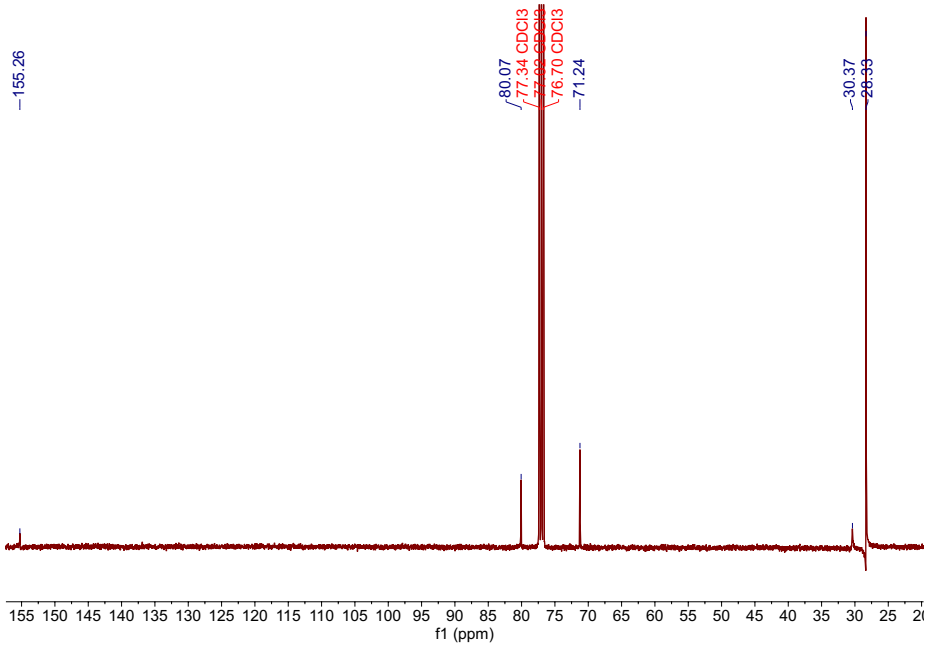

Transmittance [%]
Wavenumber cm^{-1}

N-Boc-propargylamine

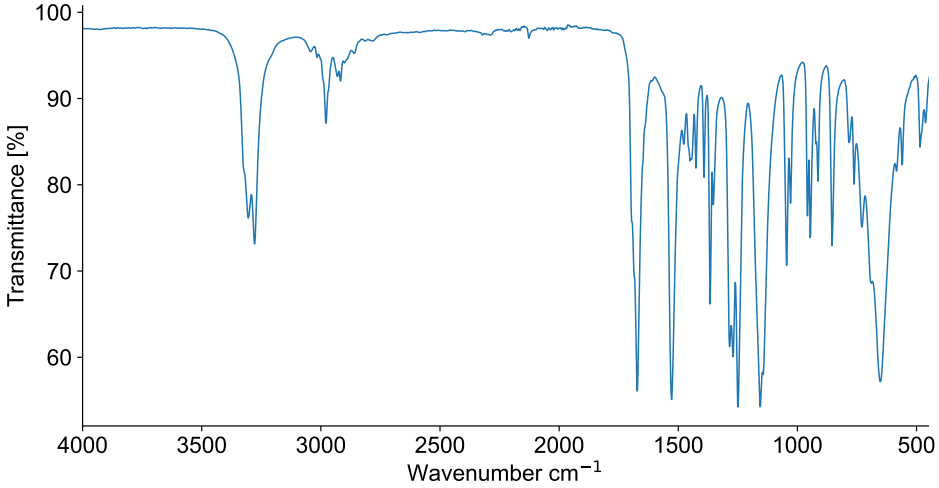
^{1}H-NMR



^{13}C-NMR

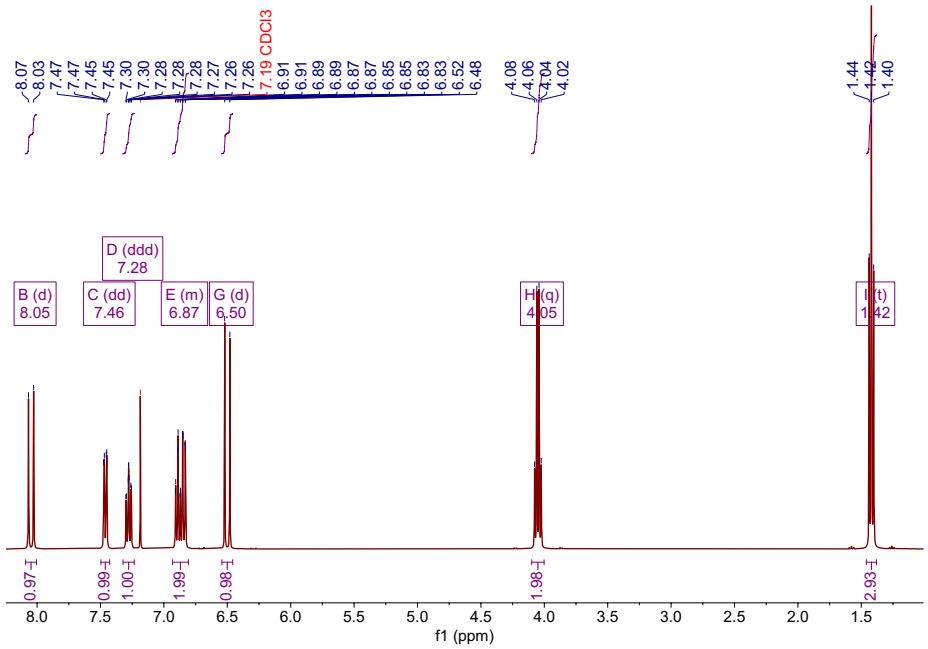


IR

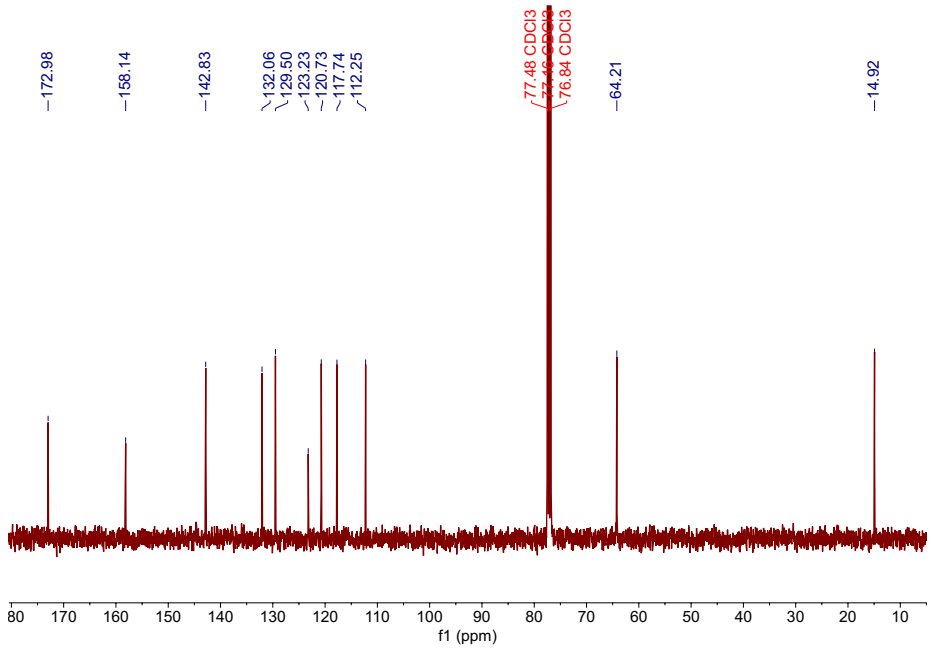


2-Ethoxycinnamic acid

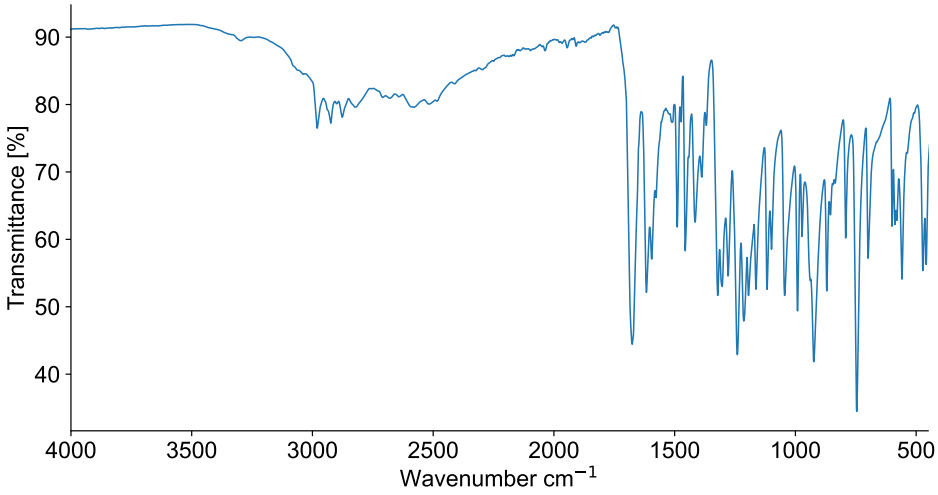
^{1}H-NMR



^{13}C-NMR

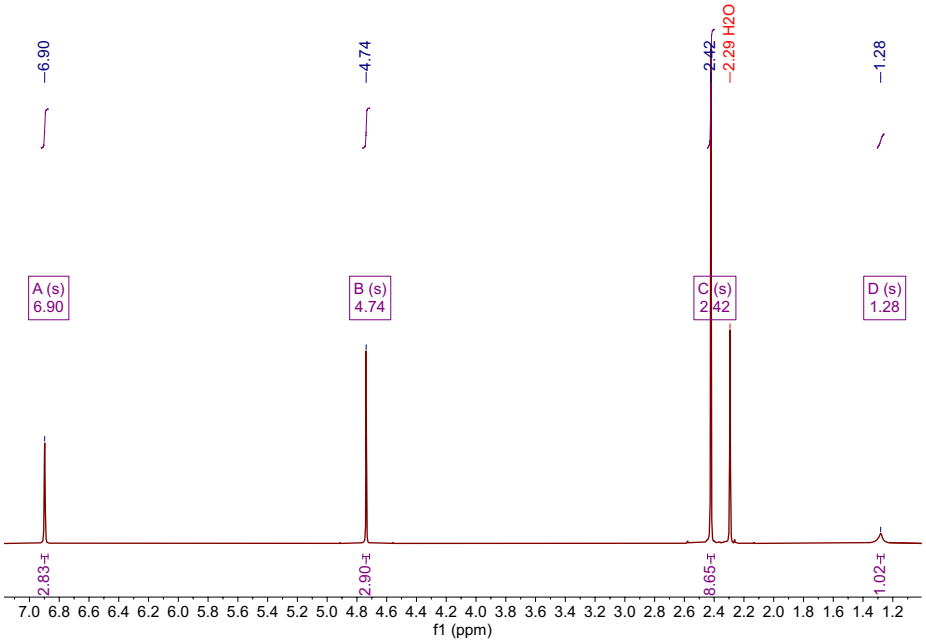

172.98
158.14
142.83
132.06
129.50
123.23
120.73
117.74
112.25
77.48 CDCl3
77.16 CDCl3
76.84 CDCl3
64.21
14.92
f1 (ppm)

IR

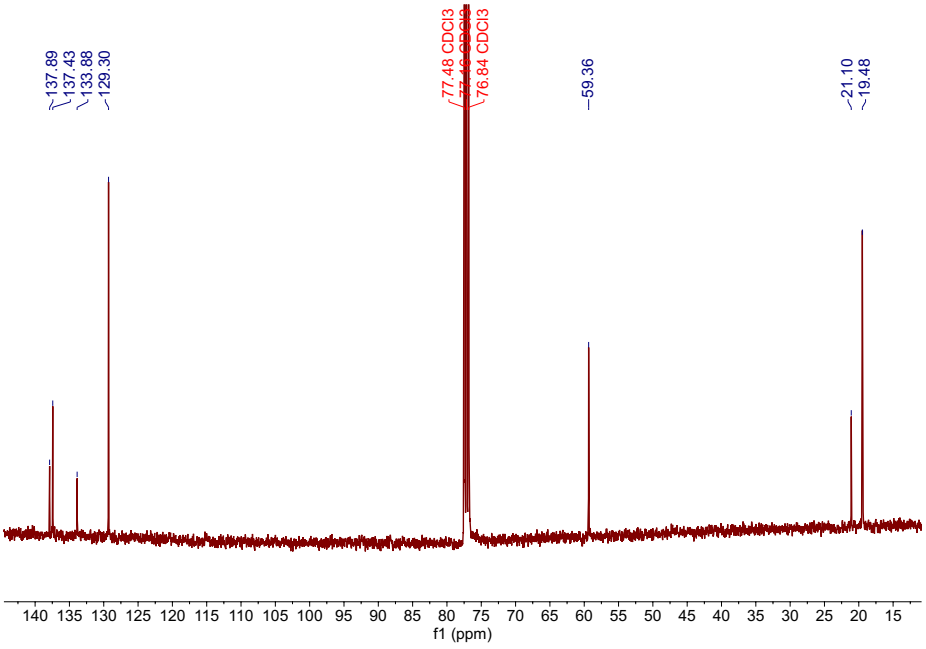

Transmittance [%]
Wavenumber cm^{-1}

2,4,6-Trimethylbenzyl alcohol

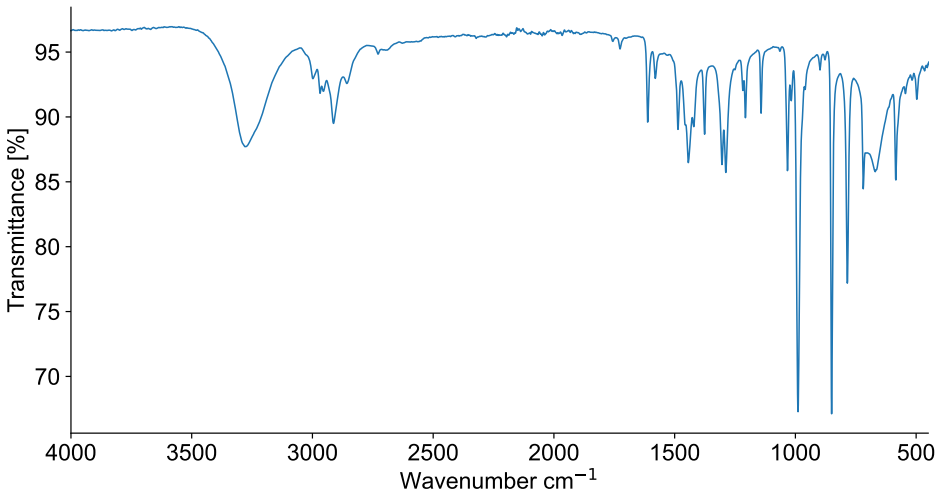
^{1}H-NMR



^{13}C-NMR

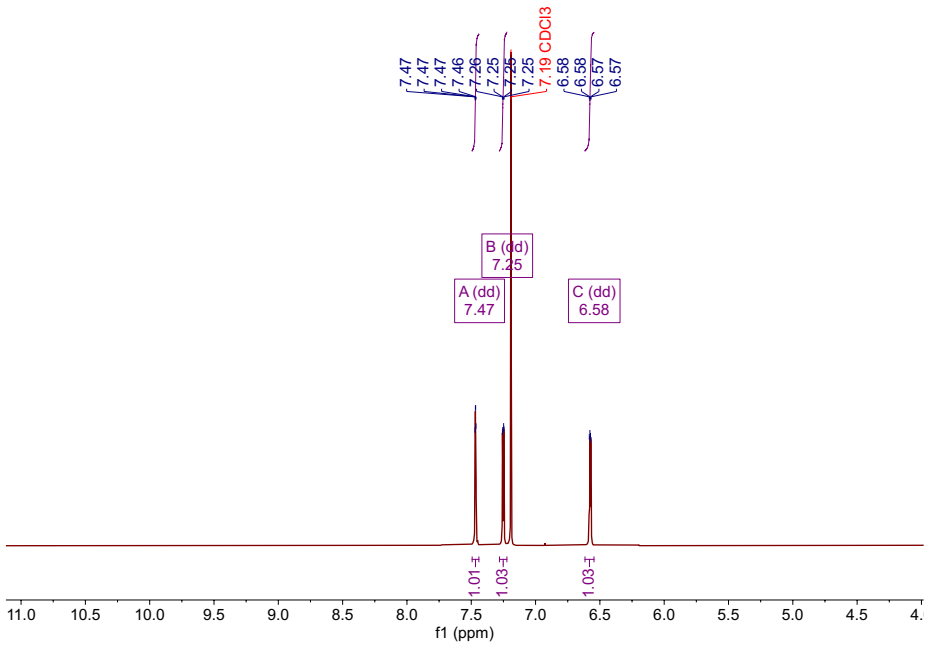


IR

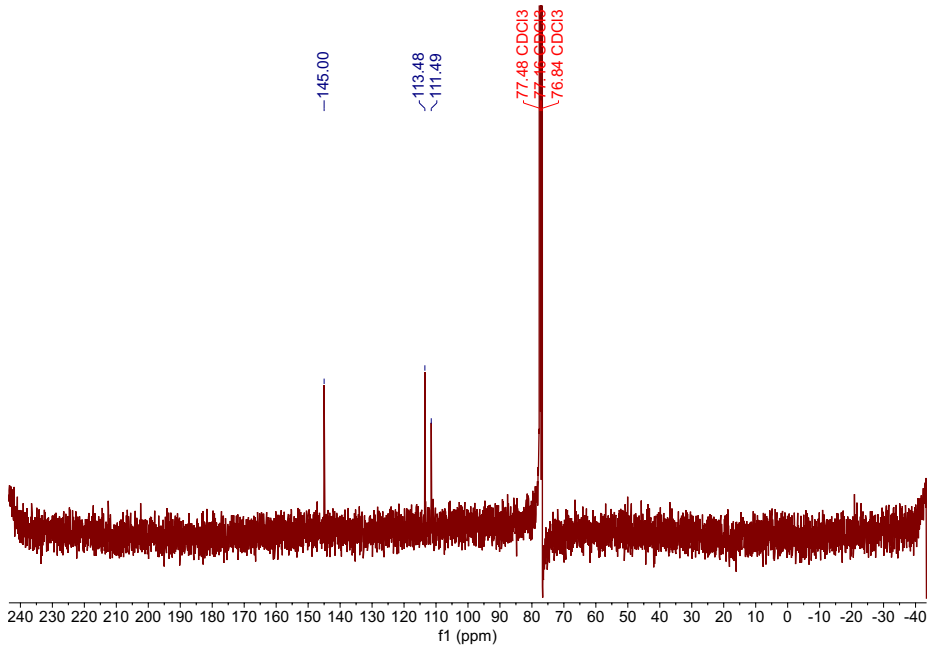


2-Nitrofuran

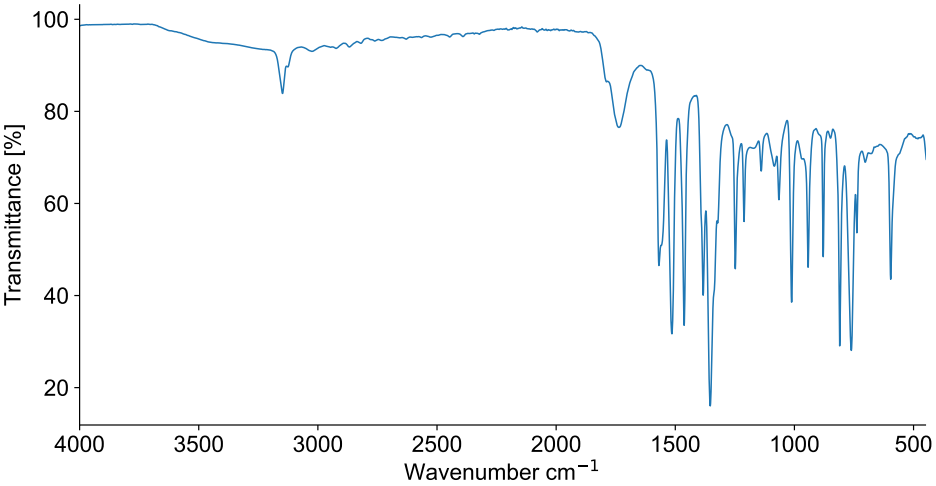
^{1}H-NMR



^{13}C-NMR

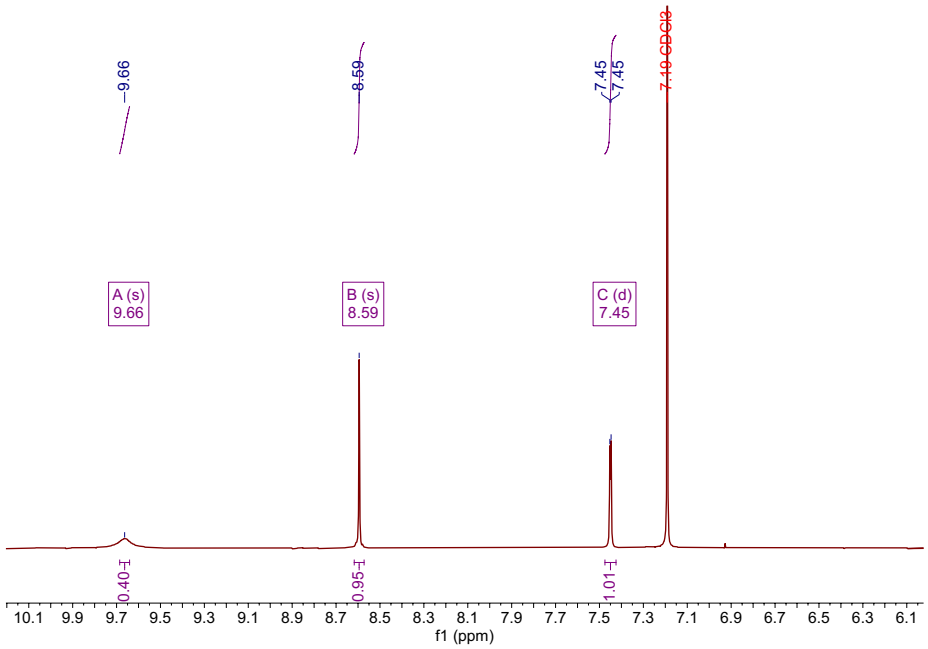

145.00
113.48
111.49
77.48 CDCl3
77.16 CDCl3
76.84 CDCl3
240 230 220 210 200 190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10 0 -10 -20 -30 -40
f1 (ppm)

IR

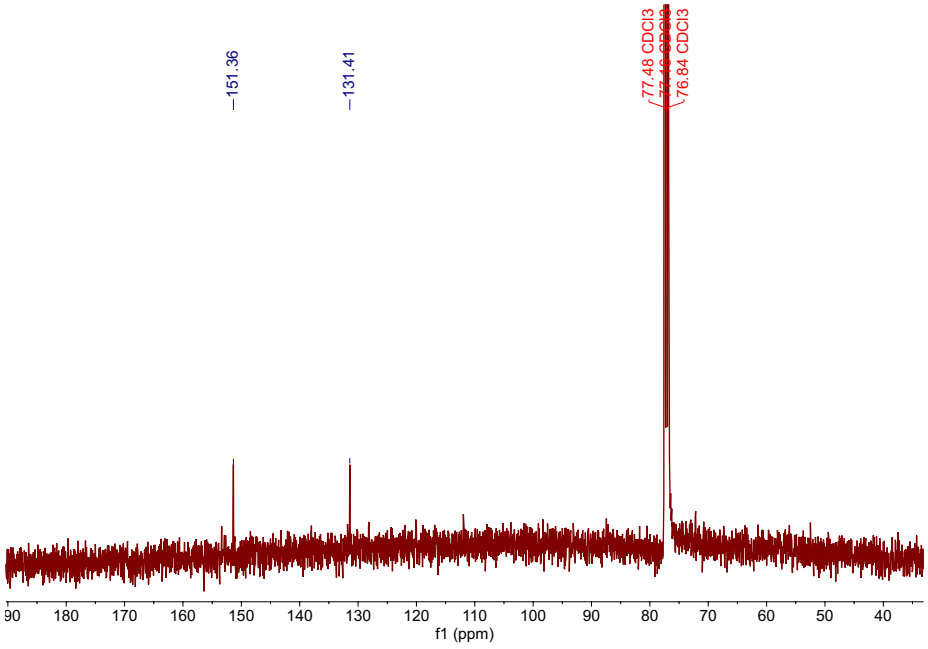

100
80
60
40
20
Transmittance [%]
4000
3500
3000
2500
2000
1500
1000
500
Wavenumber cm^{-1}

4-Chloro-5-Iodo-7H-pyrrolo[2,3]dyrimidine

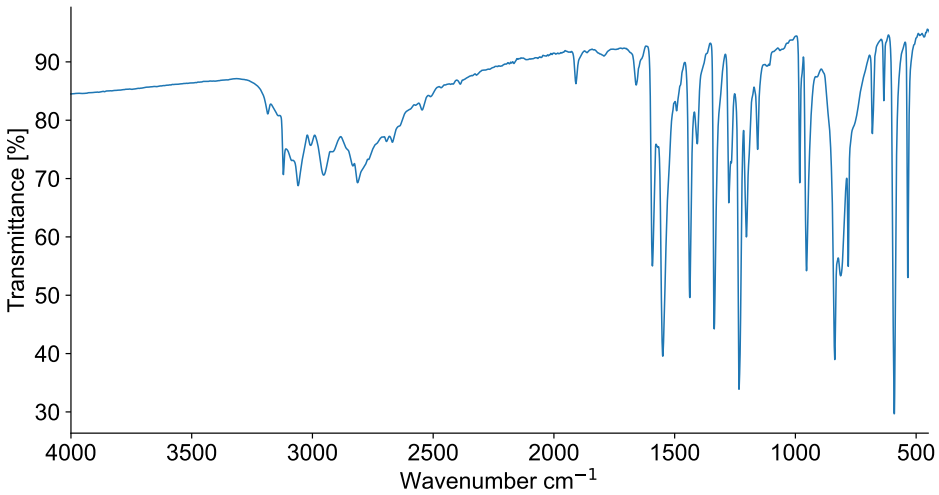
^{1}H-NMR



^{13}C-NMR

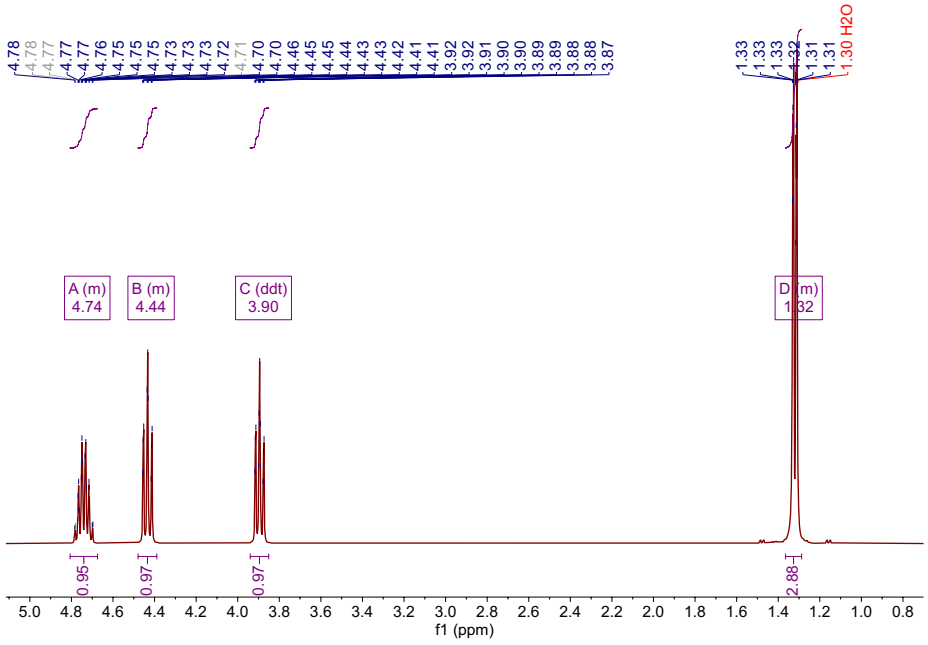


IR

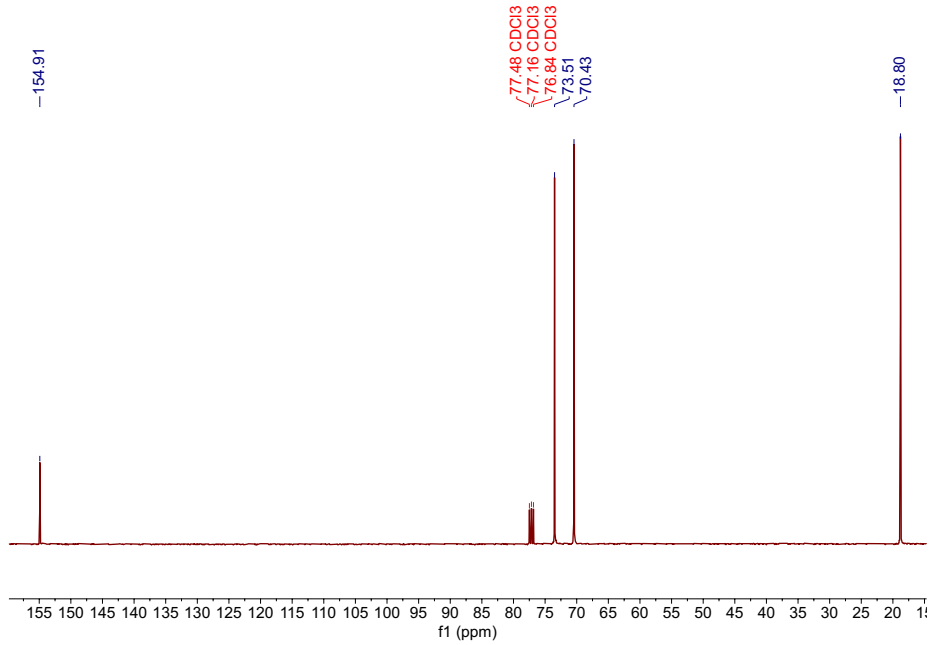


Propylene Carbonate

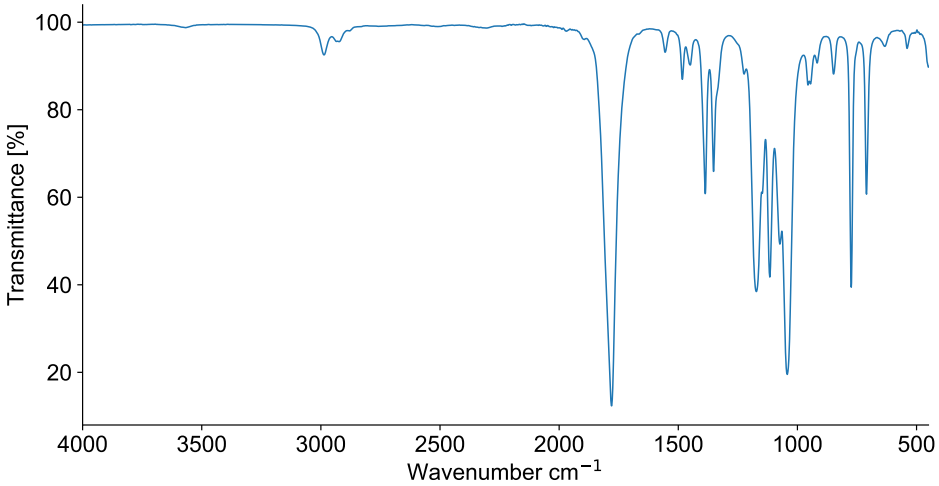
^{1}H-NMR



^{13}C-NMR

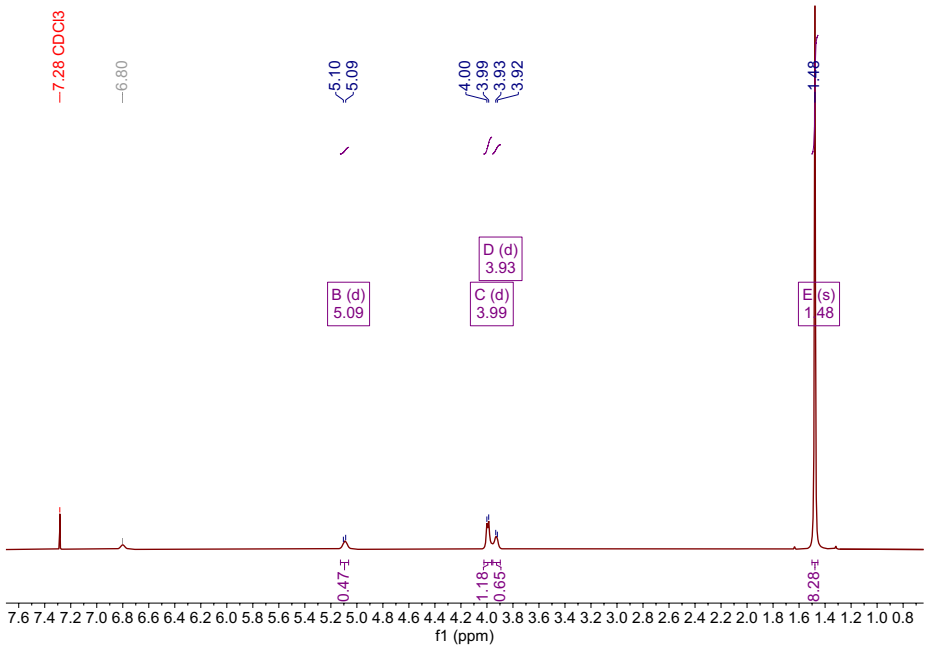

154.91
77.48 CDCl3
77.16 CDCl3
76.84 CDCl3
73.51
70.43
18.80
155 150 145 140 135 130 125 120 115 110 105 100 95 90 85 80 75 70 65 60 55 50 45 40 35 30 25 20
f1 (ppm)

IR

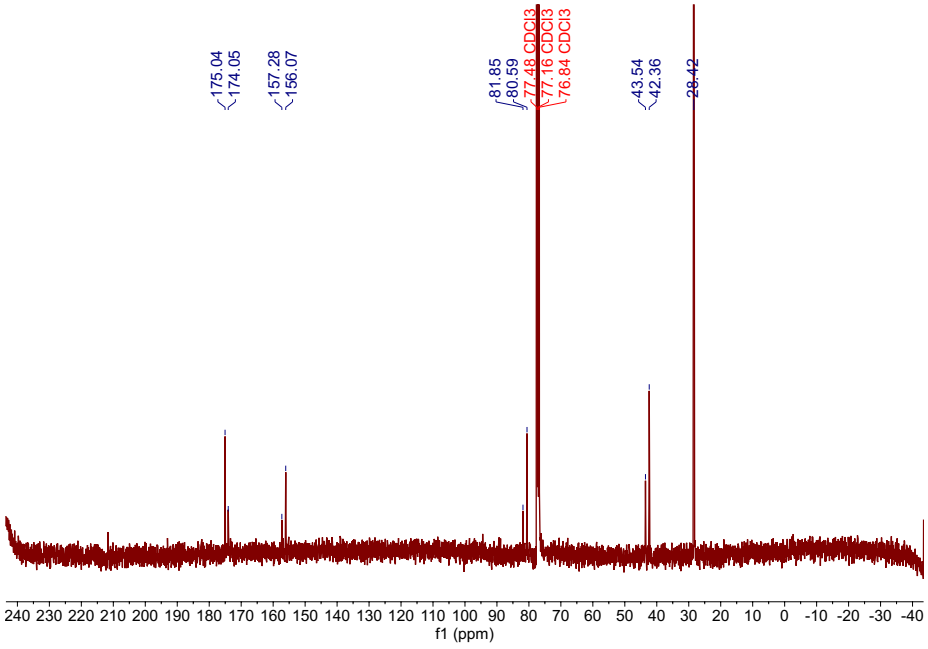

Transmittance [%]
100
80
60
40
20
4000
3500
3000
2500
2000
1500
1000
500
Wavenumber cm^{-1}

Boc-Gly-OH

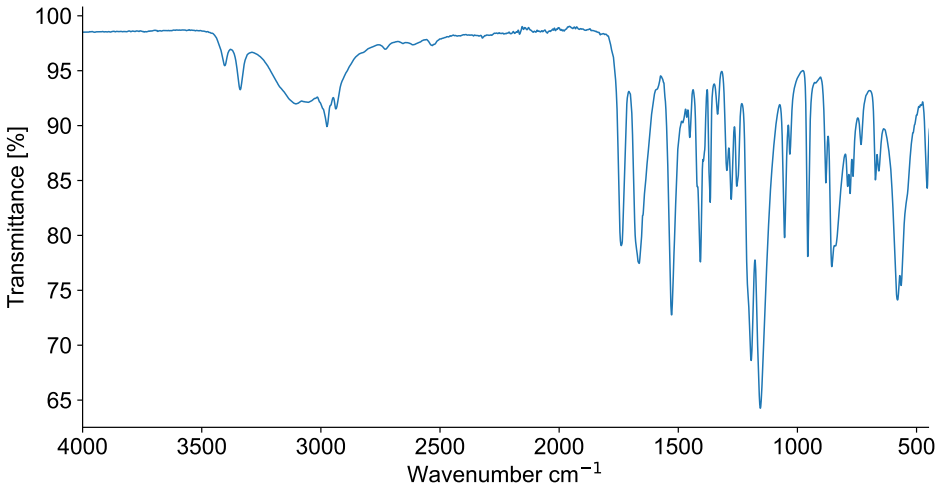
^{1}H-NMR



^{13}C-NMR

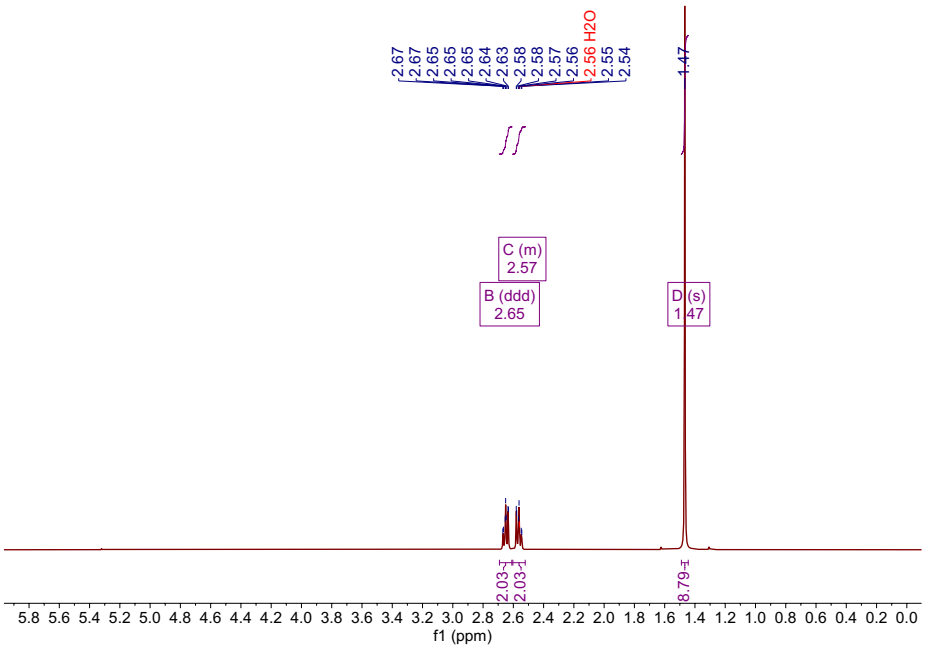


IR

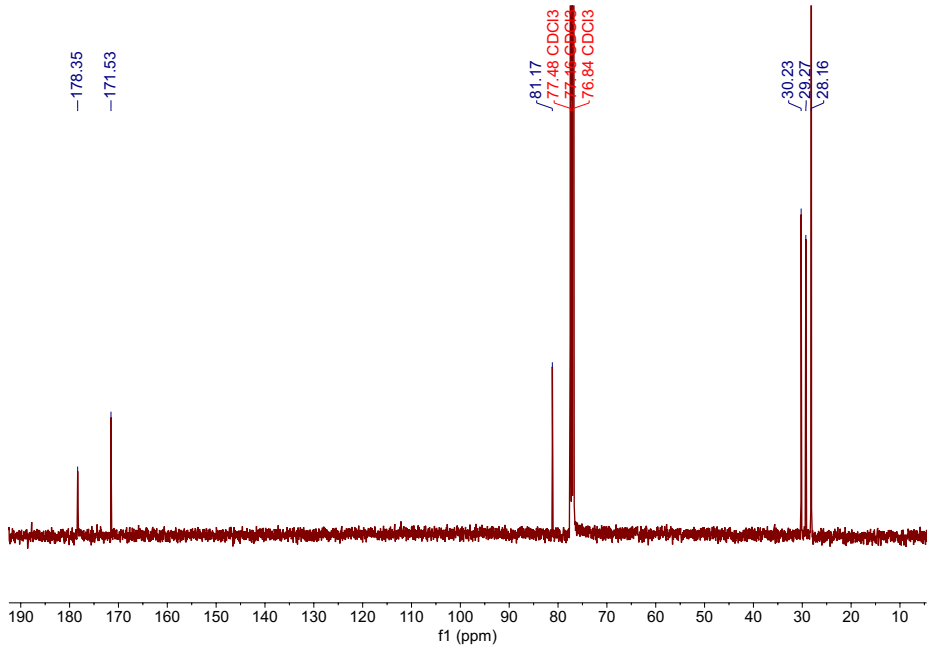


Mono-tert-butyl-succinate

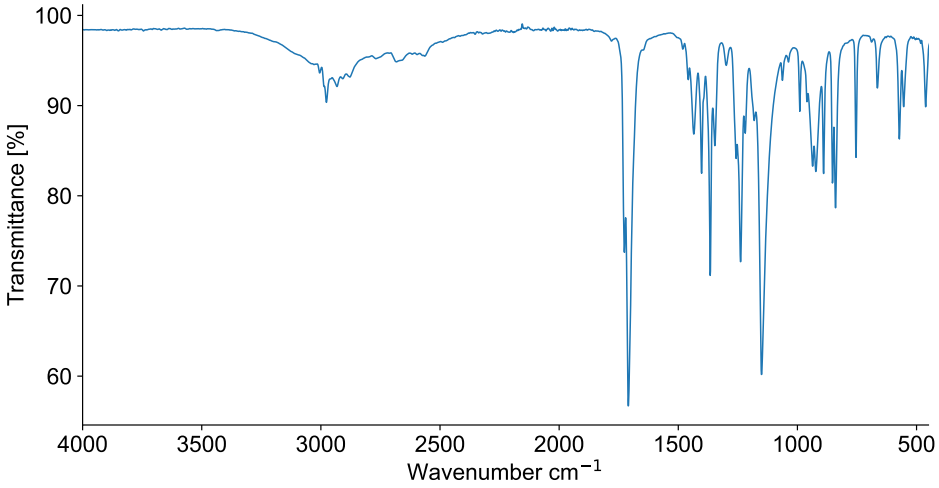
^{1}H-NMR



^{13}C-NMR

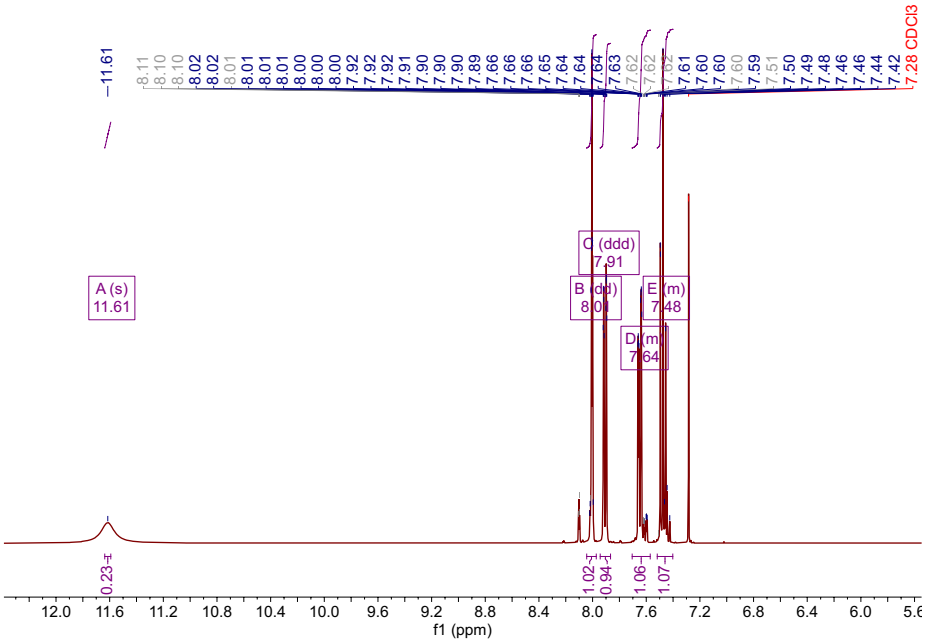

178.35
171.53
81.17
77.48 CDCl3
77.16 CDCl3
76.84 CDCl3
30.23
29.27
28.16
190 180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20 10
f1 (ppm)

IR

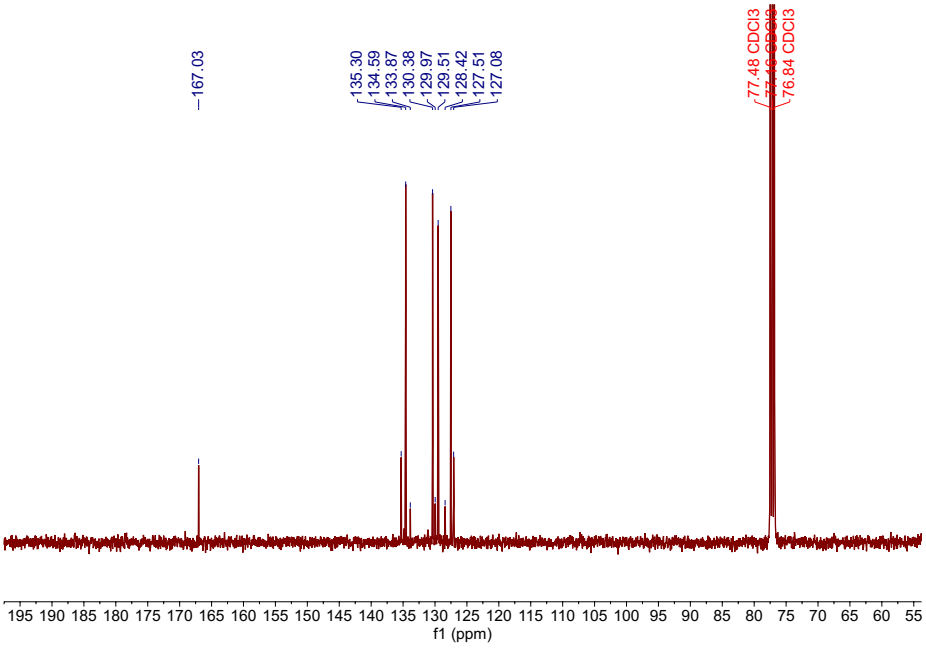

100
90
80
70
60
Transmittance [%]
4000 3500 3000 2500 2000 1500 1000 500
Wavenumber cm^{-1}

3-Chloroperbenzoic acid

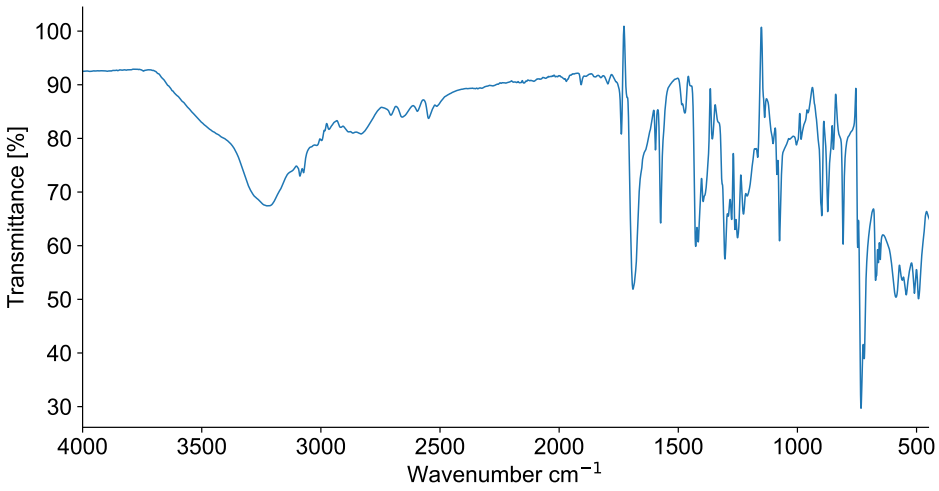
^{1}H-NMR



^{13}C-NMR

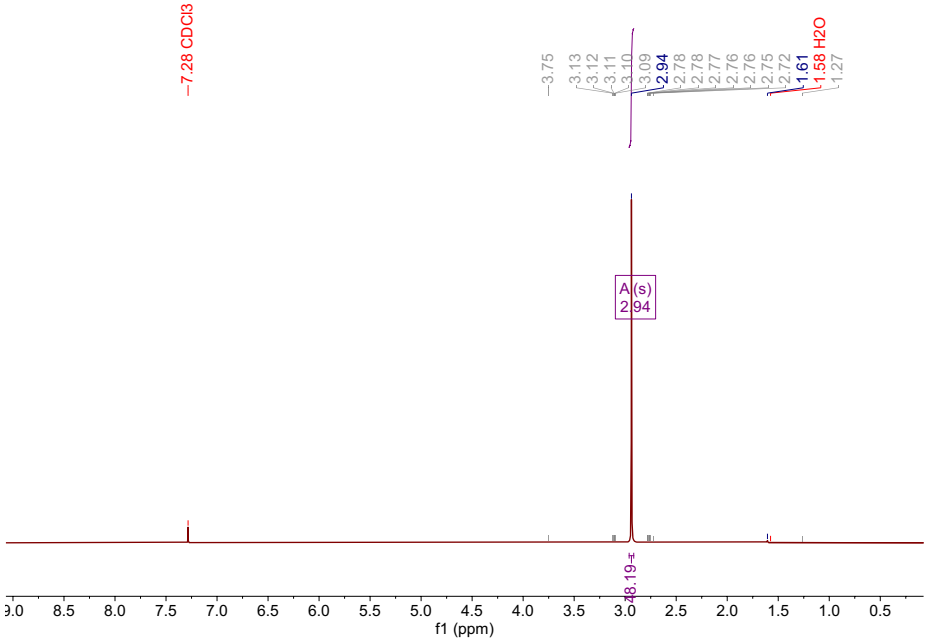


IR

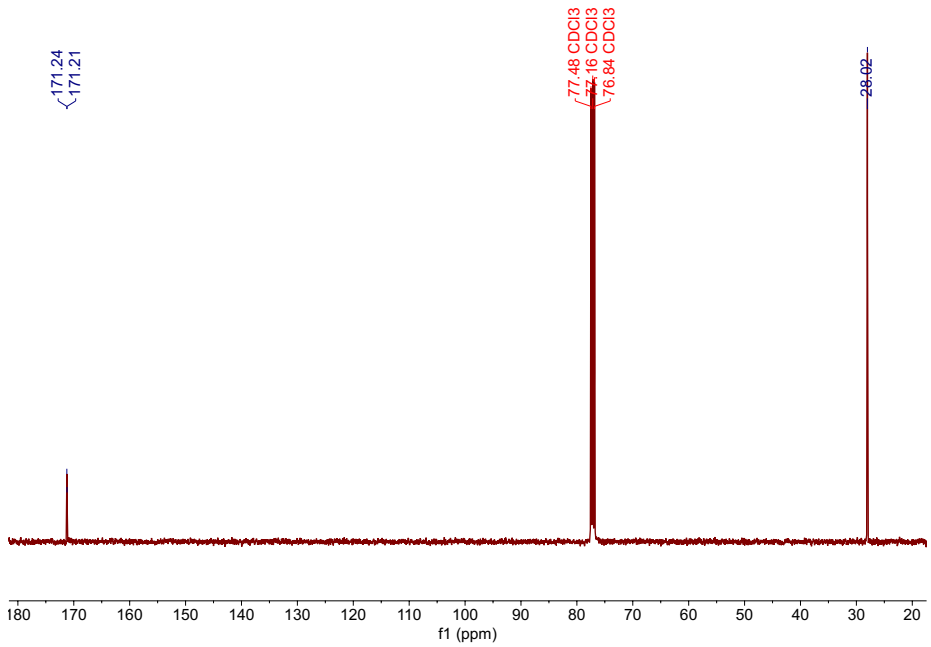


N-Chlorosuccinimidine

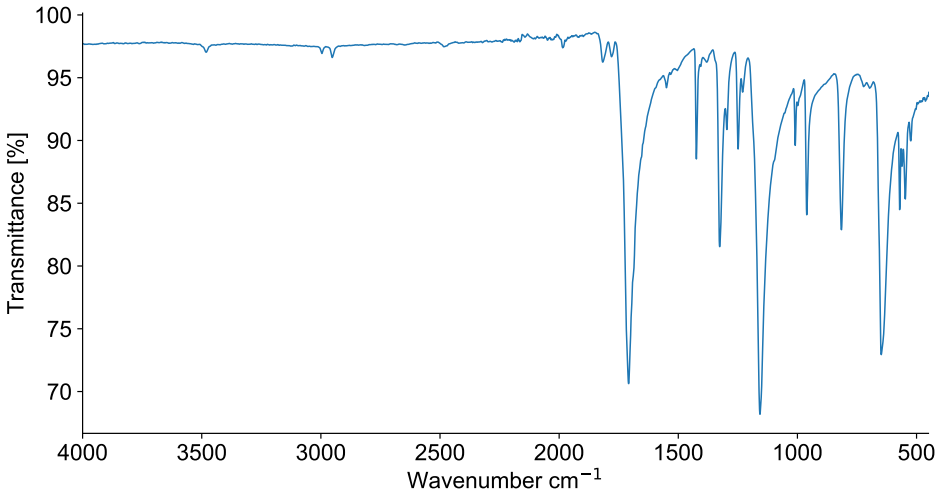
^{1}H-NMR



^{13}C-NMR

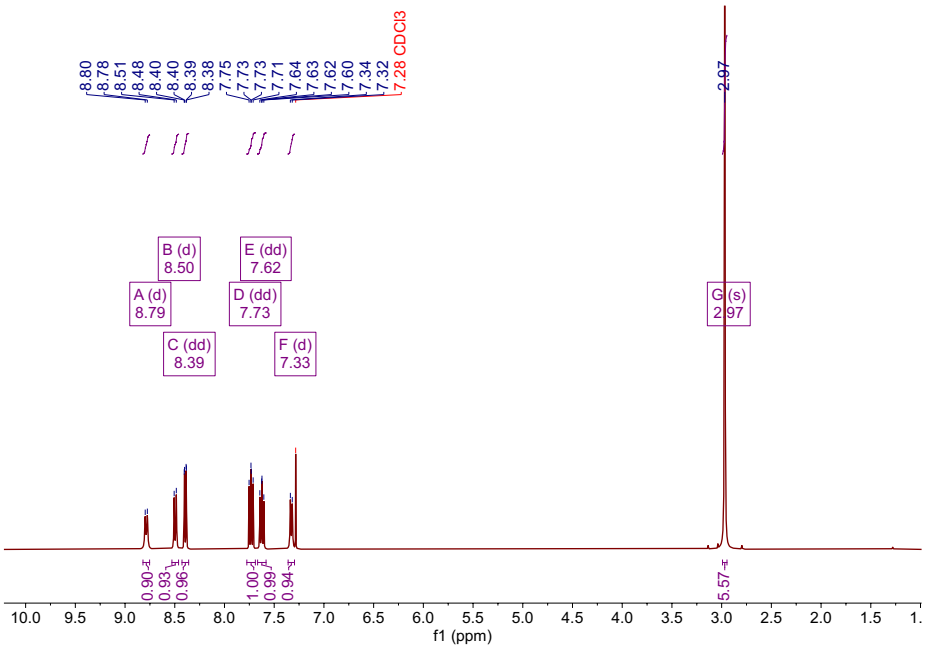

171.24
171.21
77.48 CDCl3
77.16 CDCl3
76.84 CDCl3
28.02
180 170 160 150 140 130 120 110 100 90 80 70 60 50 40 30 20
f1 (ppm)

IR

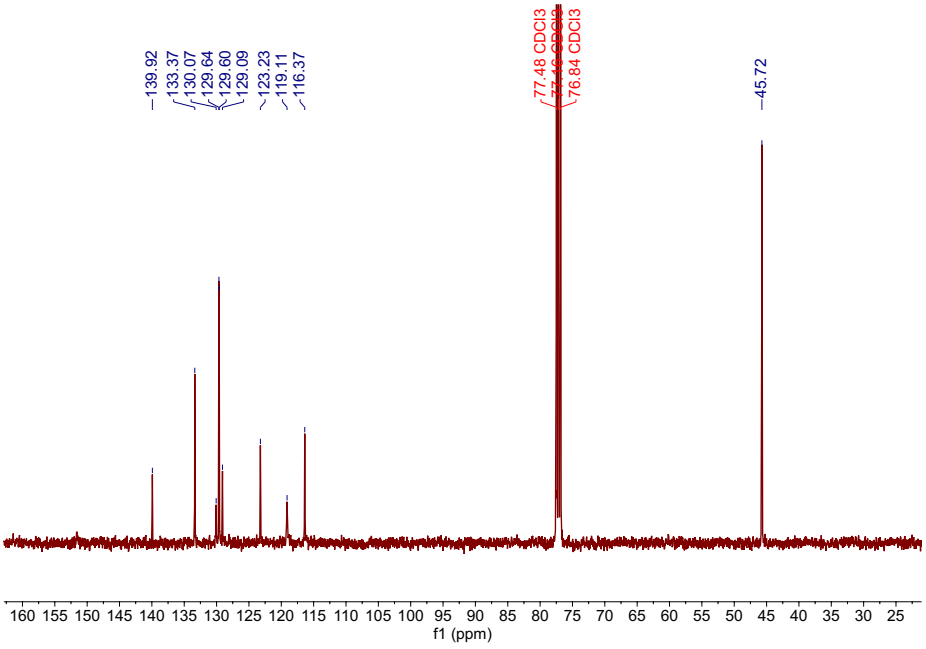

Transmittance [%]
100
95
90
85
80
75
70
4000 3500 3000 2500 2000 1500 1000 500
Wavenumber cm^{-1}

Dansyl Chloride

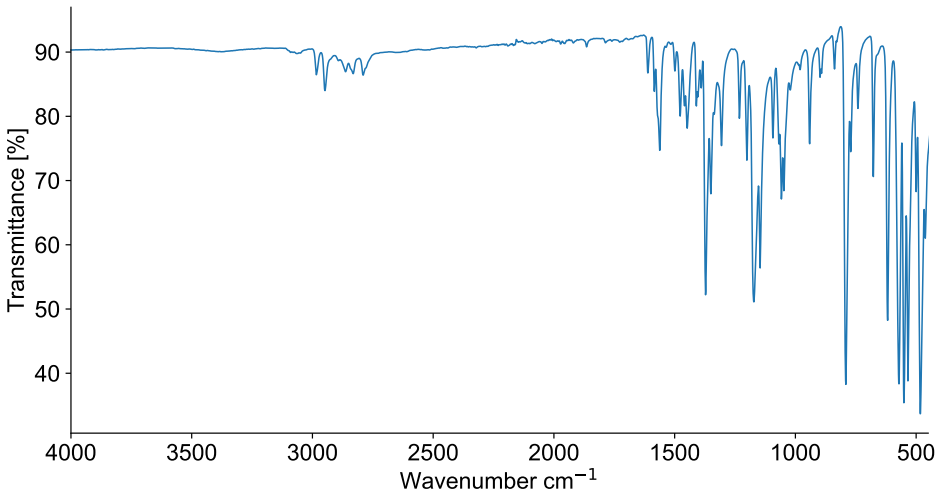
^{1}H-NMR



^{13}C-NMR

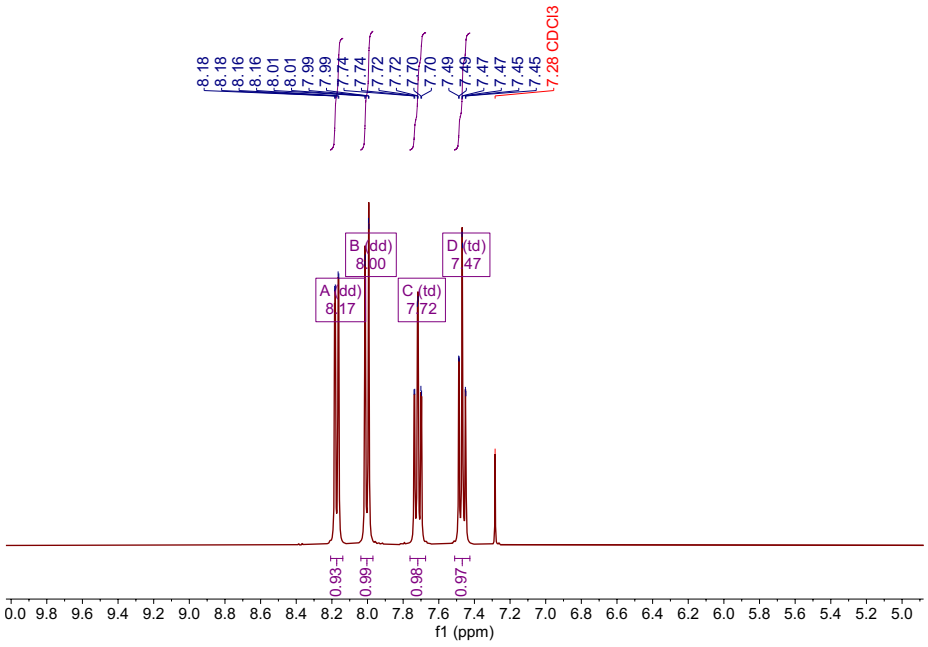


IR

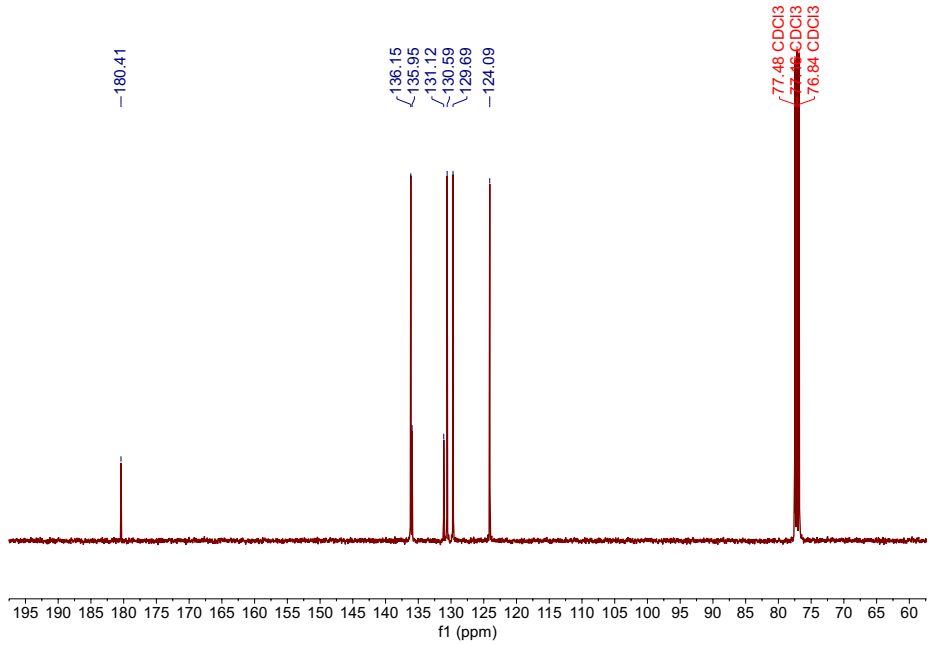


Phenanthraquinone

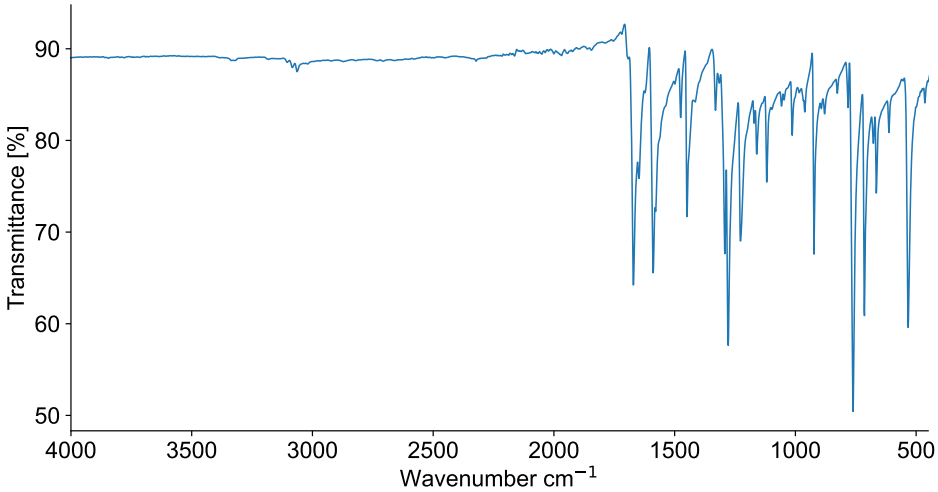
^{1}H-NMR



^{13}C-NMR

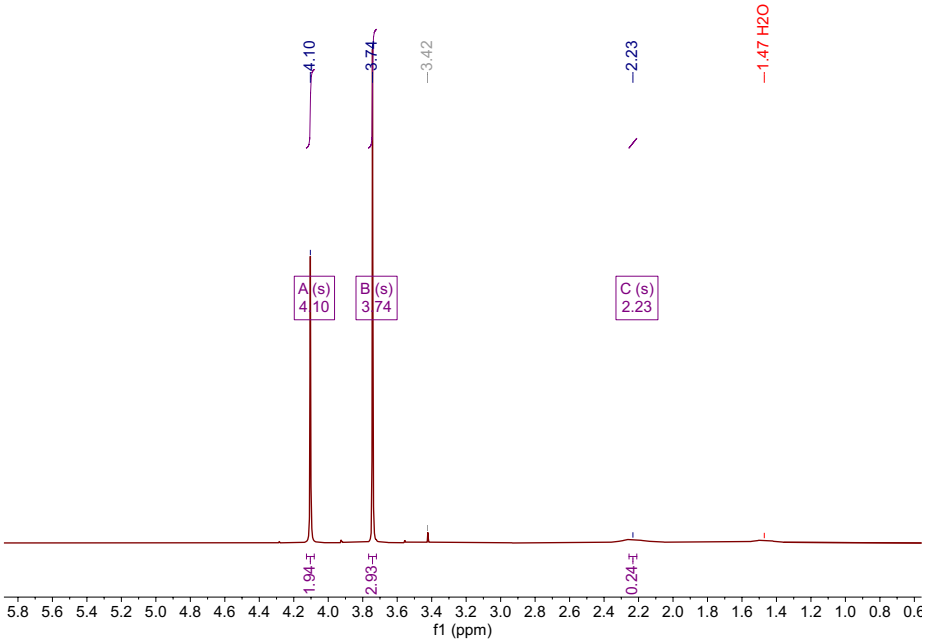

180.41
136.15
135.95
131.12
130.59
129.69
124.09
77.48 CDCl3
77.16 CDCl3
76.84 CDCl3
195 190 185 180 175 170 165 160 155 150 145 140 135 130 125 120 115 110 105 100 95 90 85 80 75 70 65 60
f1 (ppm)

IR

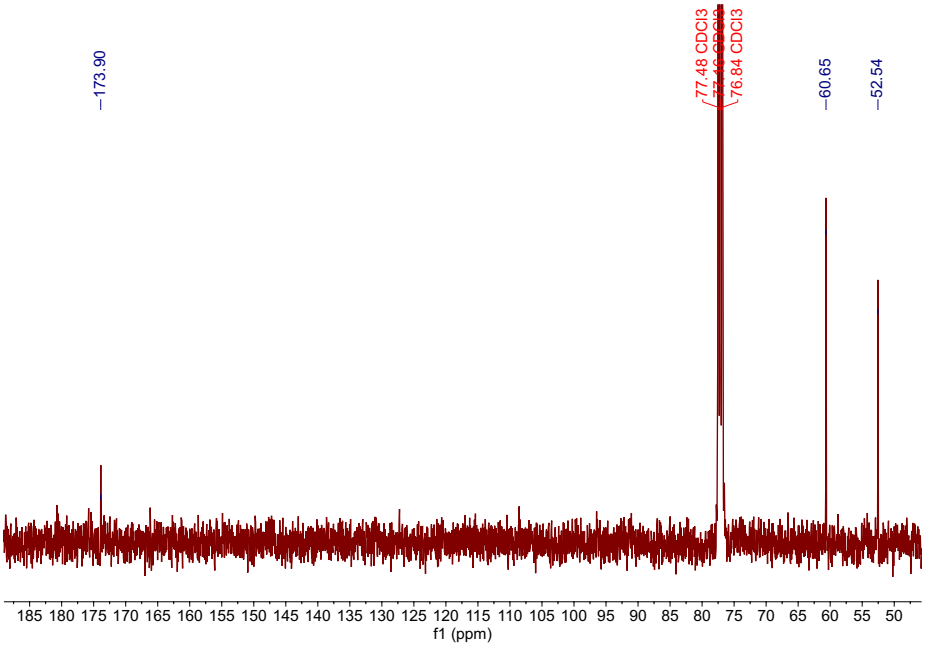

Transmittance [%]
90
80
70
60
50
4000
3500
3000
2500
2000
1500
1000
500
Wavenumber cm^{-1}

Methyl Glycolate

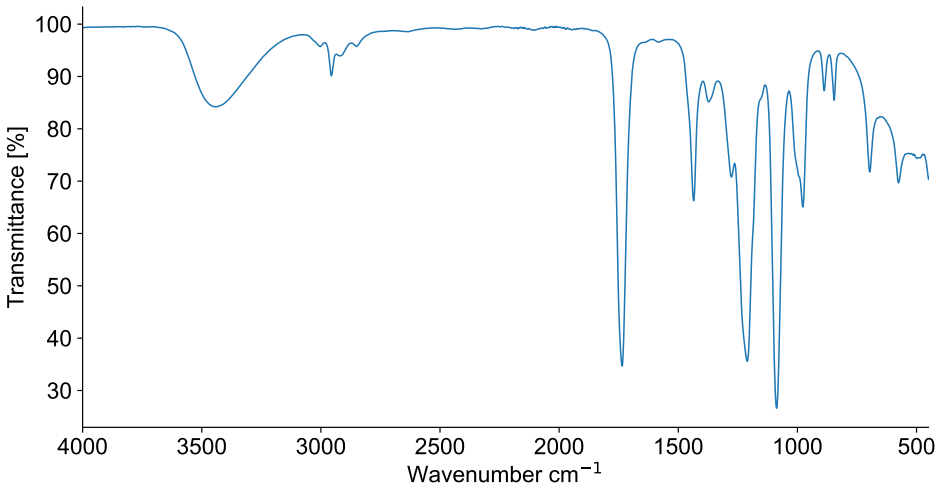
^{1}H-NMR



^{13}C-NMR

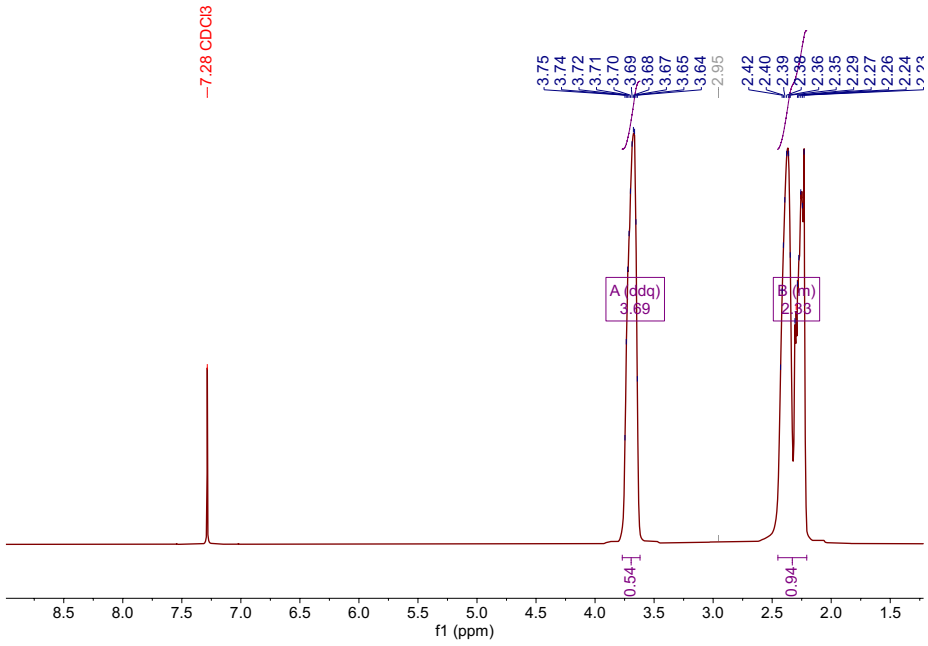


IR

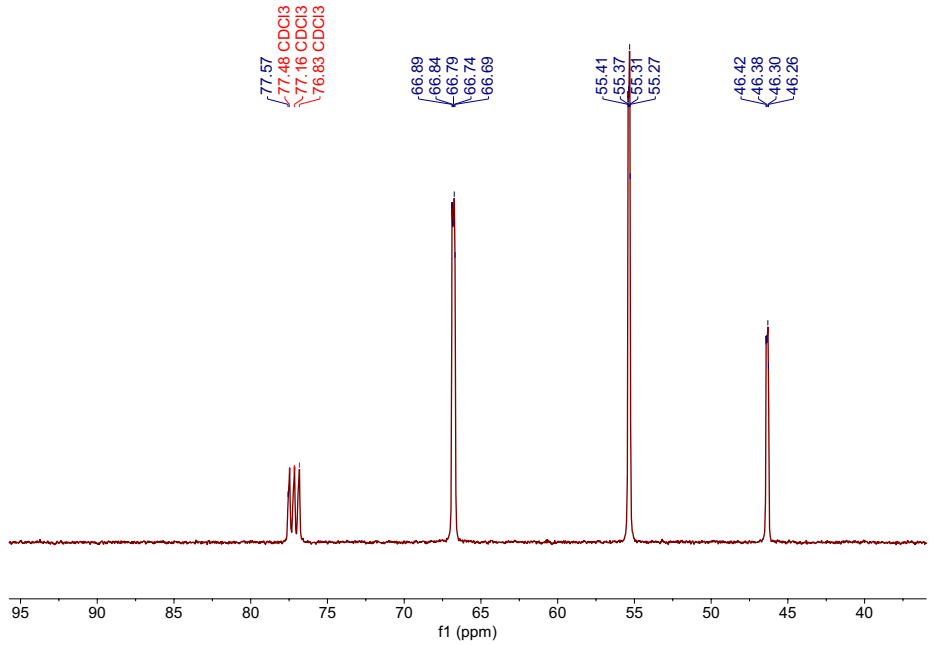


N-Methylmorpholine

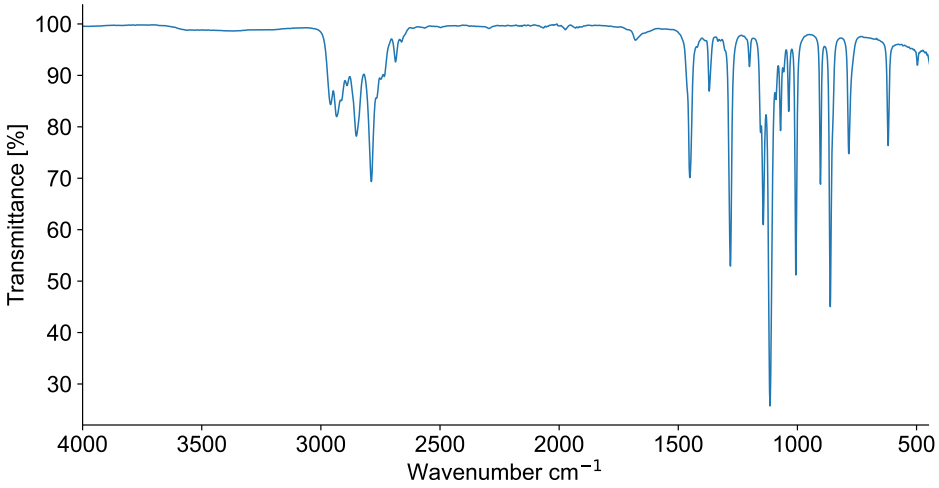
^{1}H-NMR



^{13}C-NMR


77.57
77.48 CDCl3
77.16 CDCl3
76.83 CDCl3
66.89
66.84
66.79
66.74
66.69
55.41
55.37
55.31
55.27
46.42
46.38
46.30
46.26
95
90
85
80
75
70
65
60
55
50
45
40
f1 (ppm)

IR


Transmittance [%]
100
90
80
70
60
50
40
30
4000
3500
3000
2500
2000
1500
1000
500
Wavenumber cm^{-1}