Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

The application of machine learning models in chemistry has made remarkable strides in recent years. Even though there is considerable interest in automating common procedure in analytical chemistry using machine learning, very few models have been adopted into everyday use. Among the analytical instruments available to chemists, Nuclear Magnetic Resonance (NMR) spectroscopy is one of the most important, offering insights into molecular structure unobtainable with other methods. However, most processing and analysis of NMR spectra is still performed manually, making the task tedious and time consuming especially for large quantities of spectra. We present a transformer-based machine learning model capable of predicting the molecular structure directly from the NMR spectrum. Our model is pretrained on synthetic NMR spectra, achieving a top–1 accuracy of 67.0% when predicting the structure from both the ^{1}H and ^{13}C spectrum. Additionally, we train a model which, given a spectrum and a set of likely compounds, selects the structure corresponding to the spectrum. This model achieves a top–1 accuracy of 98.28% when trained on both ^{1}H and ^{13}C spectra in selecting the correct structure.

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy is widely considered the most crucial tool in determining the structure of molecules [1]. Unlike other techniques such as infrared (IR) spectroscopy or mass spectroscopy (MS), NMR provides comprehensive and human-interpretable information about the molecule. It reveals details such as the number of NMR-active nuclei, the functional group to which a peak belongs, and, for some nuclei, information about its surrounding environment [2]. Typically the spectra of multiple NMR-active nuclei are used to definitely assign the structure. Most commonly, an ^{1}H NMR and a ^{13}C NMR are used for this purpose. In the literature, the combination of these two spectra has become the *de facto* proof that a compound has been synthesised [3]. Consequently, NMR spectroscopy has risen to prominence as the preferred analytical instrument in standard chemical laboratories.

Nevertheless, analyzing NMR spectra is not straightforward. Although there are various software tools available to assist chemists in this process, the majority of spectra are still processed manually. As a result, the analysis of NMR spectra, particularly in large quantities, becomes a time-consuming and tedious undertaking [4].

The increasing availability of computational power has ushered in a new era of statistical methods: machine learning and deep learning. These approaches have revolutionized fields such as image classification and language modeling by addressing previously unsolvable problems [5, 6]. In the realm of chemistry, machine learning, and particularly language modeling, has emerged as a highly

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

promising tool. Such models have diverse applications, spanning from predicting retrosynthetic routes over designing new drug candidates to assisting in the automation of experiments [7, 8, 9].

In addition to changes brought about by machine learning, chemistry is experiencing a paradigm shift due to the growing prominence of robotics and automation in laboratories [10, 11]. Advances in both fields have carried over into chemistry, enabling fully automated high-throughput experimental campaigns that generate vast volumes of data previously inaccessible. By operating at nanomolar scales, these techniques can conduct hundreds to thousands of reactions per day [12, 13, 14, 15]. However, one crucial step remains a limitation: the analysis of the reaction products.

Current high-throughput approaches are predominantly restricted to a limited number of reagents and reactants, largely due to their heavy reliance on high-performance liquid chromatography (HPLC) systems. Each reactant and product necessitates a separate calibration curve, imposing limitations on the chemical space that can be explored [16, 17]. Despite the automation of most physical handling steps, the analysis of the resulting data still predominantly relies on manual labor, demanding weeks to months of tedious work. Among these tasks, the analysis of NMR data obtained from high-throughput experiments can be particularly burdensome.

Even though the analysis of NMR spectra obtained from high-throughput experiments remains time consuming, advances have been made to alleviate the burden to some extent. Commercial NMR software offers options to automate peak picking, integration and multiplet assignement of the spectra [18, 19]. However, automatically determining a structure from the spectra without strong prior knowledge is currently not feasible. Approaches addressing this task using machine learning have so far been limited in the sense that they either limit the number of elements, the heavy atom count (all atoms other than hydrogen) or solely rely on one type of spectrum (e.g. ^{13}C) [20, 21, 22, 23, 24].

To close the loop between automated high throughput experiments and NMR spectroscopy, an automated NMR structure elucidation workflow is required. Here we propose to utilise language models trained on NMR spectra to directly predict the structure. We achieve a top–1 accuracy in predicting the correct molecular structure from simulated ^{1}H and ^{13}C NMR spectra of 67.0%. If the language model is provided with additional information such as the reagents and expected products of a reaction, the model is able to identify the correct structure in 98.28% of cases from the combination of both the ^{1}H and ^{13}C NMR spectrum.

2 Results and Discussion

We focus on two primary tasks. The first one involves predicting the molecular structure directly from the ^{1}H spectrum, ^{13}C spectrum, or the combination of both spectra. The second one focuses on exploring the effect of adding additional context to the NMR spectrum. This second task corresponds to a typical high-throughput scenario, where chemists are aware of the reaction that was conducted and, consequently, the potential molecules present in the spectrum. We task the model to match the correct molecule to a given spectrum.

2.1 Data

As the number of publicly available experimental NMR spectra is limited, we simulate a large training set using MestreNova [18]. We sample reactions from the Pistachio dataset [25] and simulate NMR spectra for both the reactants and products. In contrast to previous work, we do not exclude stereoisomers or restrict the heavy atom count drastically, opting for a range of 5 to 35, with an average heavy atom count of 22.7. We limit the elements to the ones most commonly found in organic chemistry, excluding molecules with elements other than carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorous and the halogens. In total we generate 1.94 million ^{1}H and ^{19}F decoupled ^{13}C NMR spectra as well as 1.10 million ^{1}H NMR spectra. Further details on the molecules can be found in Appendix A.1.

Instead of utilizing the raw ^{1}H NMR spectrum, as demonstrated previously by Huang et al. [20], we opt for a processed version of the spectrum. There are two main reasons behind this choice. Firstly, if starting from the raw vector, the model would need to learn concepts such as peak picking, peak integration, and multiplet assignment. Our approach reduces the learning demand on the model by preprocessing the spectra using MestreNova. Secondly, the wide availability of such processed experimental NMR spectra in papers and patents presents a potential avenue for validating our models on experimental data. Further information on the exact simulation details can be found in Appendix A.1.

2.2 Model

In this study, we adopt a sequence-to-sequence encoder-decoder transformer architecture, building upon the formulation utilized in our previous investigation of IR spectra [26]. More detailed information can be found in Appendix A.2.

As discussed above, we employ the processed NMR representation of a spectrum instead of a vector. For the ^{1}H NMR this takes the form of a string containing the position of the peak in ppm, the multiplet type (‘s’, ‘d’, ‘t’, etc.), and the integration of the peak (i.e. the number of hydrogen atoms). All ^{1}H values are rounded to the nearest second decimal. On the other hand, ^{13}C NMR spectra are presented to the model as a simple list of peaks. All values in ppm are rounded to the nearest first decimal. Examples are illustrated in Figure 1. A detailed account of how NMR spectra are processed can be found in Appendix A.3.

All molecules are presented to the model as presented to the model as Simplified molecular-input line-entry system (SMILES) [27, 28].

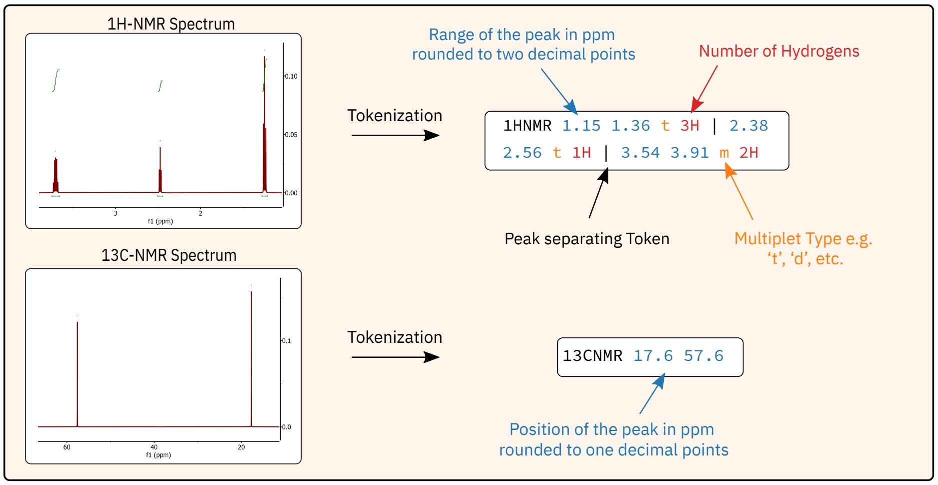
1H-NMR Spectrum
0.10
0.05
0.00
3
2
f1 (ppm)
Range of the peak in ppm rounded to two decimal points
Number of Hydrogens
Tokenization
1HNMR 1.15 1.36 t 3H | 2.38 2.56 t 1H | 3.54 3.91 m 2H
Peak separating Token
Multiplet Type e.g. ‘t’, ‘d’, etc.
13C-NMR Spectrum
0.1
0.0
60
40
20
f1 (ppm)
Tokenization
13CNMR 17.6 57.6
Position of the peak in ppm rounded to one decimal points

Figure 1: Summary of the tokenization process for NMR spectra. Top: Tokenization of an ^{1}H NMR spectrum following the Range representation. Bottom: Tokenization of a ^{13}C NMR spectrum.

2.3 Structure Prediction from NMR spectra

In the following we focus on predicting the molecular structure directly from the NMR spectrum. We assess three different scenarios: Predicting the structure solely from the ^{1}H NMR spectrum, solely from the ^{13}C NMR spectrum, and from the combined ^{1}H and ^{13}C NMR spectra.

2.3.1 Model optimisation

To explore the consequences of various data preparation methods, we examine the effects of supplementing the model with the chemical formula alongside the spectra, altering the formatting of ^{1}H NMR peaks, and investigate the effect of a shared or separate token space between the ^{1}H and ^{13}C NMR peaks. We train 13 models to assess the impact of these changes and evaluate the performance of the trained models based on the top–1, top–5, and top–10 accuracy metrics. These metrics indicate the percentage of cases in which the predicted structure matches the target structure within the first, first five, and first ten predictions, respectively. Molecules are defined as matching if their canonical SMILES are identical. The results of these experiments can be found in Table 1. In the following, we delve deeper into the different data preparation methods and their respective effects.

We trained a model for all the three scenarios (solely ^{1}H or ^{13}C and combined ^{1}H and ^{13}C) with and without the chemical formula. We observe an increase in performance of $\sim$8–14% in performance for

Table 1: Summary of experiments on simulated data and associated metrics.

	Formula	Format*	Tokens†	Top–1%	Top–5%	Top–10%
^{1}H NMR	✗	Center	N/A	38.29%	54.67%	58.43%
	✓	Center	N/A	53.34%	71.71%	75.09%
	✓	Adaptive	N/A	53.39%	71.84%	75.23%
	✓	Range	N/A	55.32%	73.59%	76.74%
^{1}H NMR (Augmented)	✓	Range	N/A	51.58%	70.52%	73.94%
^{1}H NMR (Ensemble)	✓	Range	N/A	**57.99%**	**76.65%**	**80.04%**
^{13}C NMR	✗	N/A	N/A	37.21%	53.98%	57.45%
	✓	N/A	N/A	51.37%	70.74%	74.32%
^{13}C NMR (Augmented)	✓	N/A	N/A	49.02%	69.05%	72.90%
^{13}C NMR (Ensemble)	✓	N/A	N/A	**53.91%**	**73.45%**	**77.72%**
^{1}H+^{13}C NMR	✗	Range	Separate	56.88%	73.91%	76.89%
	✓	Range	Separate	64.78%	81.74%	84.43%
	✓	Range	Shared	65.05%	82.07%	84.70%
^{1}H+^{13}C NMR (Augmented)	✓	Range	Shared	62.35%	80.15%	82.93%
^{1}H+^{13}C NMR (Ensemble)	✓	Range	Shared	**66.99%**	**84.09%**	**86.59%**

* The format used to represent the position of the ^{1}H NMR peaks
Center: Center of the peak
Range: Minimum and maximum ppm of the peak
Adaptive: If the range is larger than 0.15 ppm use the range format otherwise center format

† Whether the token space of the ^{1}H and ^{13}C NMR spectrum is shared or separate

all three models when including the formula. Adding the chemical formula constrains the chemical space that the model explores. This transforms the task from predicting the structure based solely on the spectrum to generating a set of isomers from the chemical formula and matching the best one to the spectrum. Consequently, we include the formula in all subsequent experiments. Experimentally the chemical formula is easily obtained via Liquid Chromatography – Mass Spectrometry (LC–MS). The integration of LC–MS into high-throughput workflows is common and as such this data would typically be obtained alongside the NMR spectra.

Another point of interest is the format in which ^{1}H NMR peaks are presented to the model. In the literature, two formats are commonly used to describe ^{1}H NMR peaks. For smaller, narrower peaks, the center of the peak is typically used. Conversely, for larger, broader peaks, the peak is described as a range by indicating the minimum and maximum values at which the peak begins and ends. Here, we investigate three cases: (1) providing the model only with the center of the peak, (2) using a range by specifying the start and end values of each peak, and (3) employing an adaptive approach inspired by the format found in the literature with thinner peaks using the center and broader peaks the range representation. We define broad peaks as those with a width greater than 0.15 ppm. The results are presented in Table 1 within the ^{1}H NMR section. We find that the range representation yields the best performance, likely due to the additional information on the width of the peak. Therefore, for all subsequent experiments involving ^{1}H NMR spectra, we utilize the range representation.

Next, we shift our focus to the combination of ^{1}H and ^{13}C spectra. To assign a structure from NMR spectra, it is common practice to rely on both the ^{1}H and ^{13}C spectra, as opposed to analysing a single spectrum on its own. In these experiments, we investigate the impact of providing the model with both the ^{1}H and ^{13}C NMR spectra. Following our earlier experiments we reuse the best representations for ^{1}H spectra and concatenate it with the ^{13}C spectrum. More detailed information regarding the data format utilized to feed the model can be found in Appendix A.3. Additionally, we examine whether the model performs better when the tokens representing the position of the peaks fall into a shared space or a separate one. This is achieved by dividing the position of the ^{13}C NMR peaks by 10 causing a significant overlap in tokens describing the position of peaks between the two modalities. The advantage of sharing tokens is a decreased vocabulary size. However, when the tokens are shared the model has to learn to differentiate between ^{1}H and ^{13}C NMR tokens. The results, presented in Table 1 under the ^{1}H+^{13}C NMR section, demonstrate that the shared tokenization scheme outperforms the separate one by $\sim$0.25%.

To enhance the models' performance and promote generalization, we augment the training data. Specifically, we utilize jitter augmentation with a range of 0.5 ppm, as outlined in Appendix A.4. This augmentation approach generates two augmented spectra for each original spectrum. When training the models on the combined augmented and original spectra, we observe a noticeable decline in performance across all scenarios (^{1}H, ^{13}C, and the combined ^{1}H and ^{13}C). We hypothesize that this is caused by the reliance of the models on the high homogeneity of the data, its consistency in peak position and width, and the lack of noise. Introducing noise through augmentation disrupts the learning process and results in decreased performance on the simulated test set. However, should the models be evaluated on experimental spectra, which naturally contain noise, we expect that the augmented models would likely perform better.

Ensembling was used to further increase the performance of the models. We used an ensemble of the five best performing checkpoints for each model trained on non-augmented data. Across the three scenarios this increases performance on average by $\sim$2.4%. Results of the best performing models can be seen in Table 1. Ultimately, our final top–1 accuracy reaches 58.0% for ^{1}H NMR, 53.9% for ^{13}C NMR, and 67.0% for the combined ^{1}H and ^{13}C NMR spectra.

2.3.2 Model Analysis

In the following we analyse the performance of the model across the three tasks. We use the best ensembled model from above and evaluate how the performance of the model changes with respect to the heavy atom count and in relation to the presence of specific functional groups. In addition, we also demonstrate that even if the model makes mistakes, most predicted molecules are relatively similar to the ground truth by evaluating the Tanimoto similarity of all predicted molecules[29].

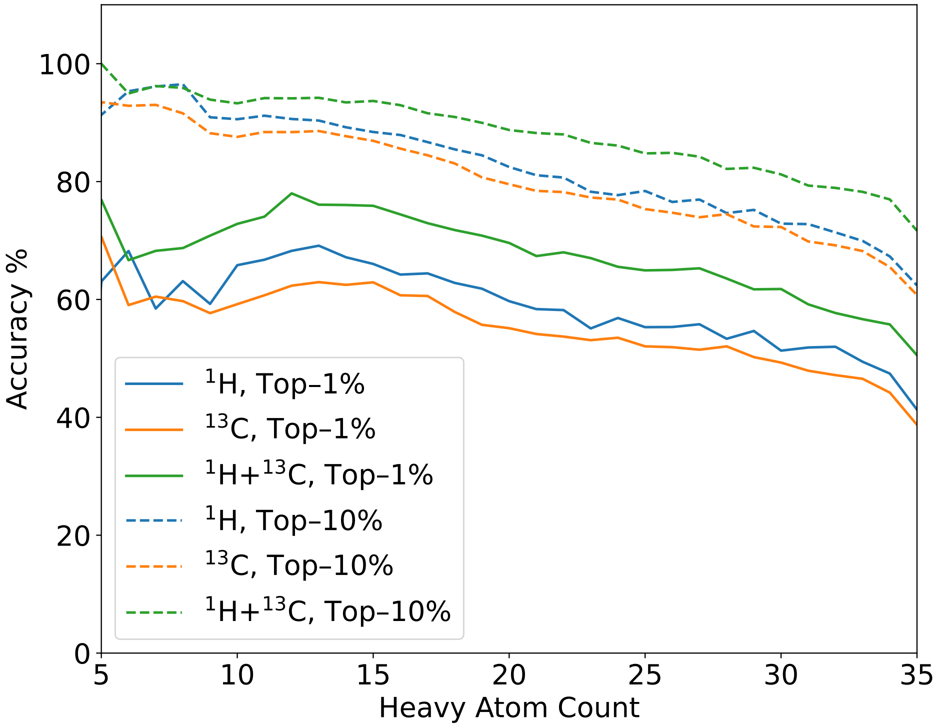

Figure 2: Heavy atom count vs accuracy. Results for ^{1}H spectra are shown in blue, for ^{13}C in orange and in green for the combination of both.

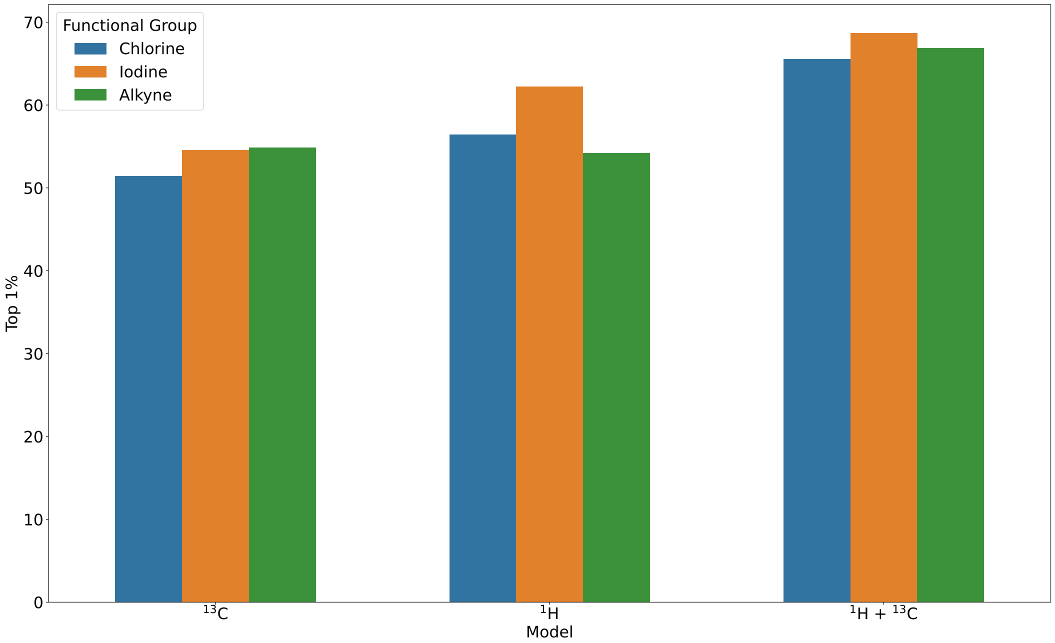

Figure 3: The models ability to correctly predict the molecular structure plotted against the presence of certain functional groups: a) ^{1}H NMR, b) ^{13}C NMR, c) ^{1}H+^{13}C NMR.

Heavy Atom count

In order to assess the model's performance, we evaluate its accuracy in relation to the heavy atom count. Figure 2 shows a negative correlation between the heavy atom count and the model's accuracy. The model trained on both ^{1}H and ^{13}C spectra outperforms the models trained on a sole spectrum, highlighting the complementary information that can be extracted from both types of spectra. As expected, the ^{1}H model demonstrates better performance compared to the ^{13}C model, albeit by a relatively small margin of $\sim$5%. The relatively high variability in performance for molecules with a heavy atom count ranging from 5 to 10 can be attributed to the limited training data available in this particular range, comprising only around 2.5% of the total training dataset.

The negative correlation of the model's performance with the heavy atom count can be attributed to two factors. Firstly, as the heavy atom count increases, molecules tend to become more complex, resulting in longer SMILES strings. Since the model generates predictions autoregressively, even a single incorrect token prediction can lead to a significantly different structure. This sensitivity to errors becomes more pronounced with an increase in the complexity of the molecules. Secondly, as the heavy atom count rises, the chemical space expands exponentially, giving rise to a greater number of potential isomers that the model must differentiate, making the prediction more challenging.

Functional Group to Structure

We analyse the model's ability to generate the correct structure depending on the presence of certain functional groups by calculating the top–n metrics for subsets containing a specific functional group in the test set. The scores are shown for each of the selected functional groups in Figure 3. The functional group definitions used and full performance across all evaluated functional groups can be found in Appendix B and C respectively. As with the heavy atom count, the model trained on the combined spectra outperforms both models trained on a sole spectrum, demonstrating the synergy that can be obtained by using both.

When comparing between the model trained on ^{1}H and ^{13}C specta, the ^{1}H NMR model's performance is notably higher when predicting molecules containing halogens. This divergence can be attributed to the fundamental differences between the two modalities. While ^{13}C NMR offers some insight into the presence of halogens, ^{1}H NMR spectra provide substantially more information, enabling conclusions to be drawn regarding the presence and even quantity of halogens on adjacent atoms.

Conversely, we find that the ^{1}H NMR model performs worse compared to the ^{13}C NMR model when predicting molecules containing alkynes. This can be attributed to two factors. Firstly, carbon NMR alkyne peaks are relatively distinctive and easily identifiable. Secondly, in many cases, there are simply no hydrogen atoms directly attached to the alkynes. As a result, alkynes become a potential blind spot for ^{1}H NMR.

When both ^{1}H and ^{13}C NMR spectra are provided to the model, we observe an improvement for all functional groups. This is especially apparent for both halogens and alkynes compared to the individual models. In fact, these functional groups now perform above average in the combined model. This highlights the the model's capacity to effectively utilize and integrate information from both modalities, thereby harnessing the complementary strengths of the two types of spectra enhancing its predictive capabilities.

Similarity

We compute the Tanimoto similarity [29] to the ground truth for all predicted molecules using Morgan fingerprints with a radius of 2 and a bit vector size of 1024 [30]. The average Tanimoto similarity is 0.534, 0.537, and 0.553 when the prediction relies on ^{1}H NMR, ^{13}C NMR, and combined spectra, respectively. Examples of molecules predicted by the combined model are shown in Figure 4. Even when the model makes incorrect predictions, most of them still exhibit a high degree of similarity to the ground truth. The similarity distribution for all three models can be found in Appendix D.

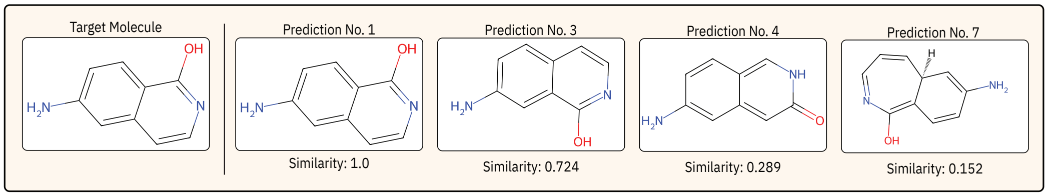

Figure 4: Four predictions of the model trained on the combined data. Illustrated are the target molecule on the left and four predictions on the right, including their rank and similarity to the target molecule.

2.4 NMR Matching

In this task, our objective is to evaluate the model's ability to accurately match the correct structure to an NMR spectrum based on a set of potential molecules and a spectrum. In practical terms, this task resembles a situation in which, after a reaction has been completed and NMR spectra have been obtained for each fraction, these fractions must be assigned to a potential molecule. For this task, we train models on ^{1}H, ^{13}C, and their combination. We compare these models to a baseline which randomly picks a molecule from the set.

We provide the model with a set of molecules along with an NMR spectrum. The input of the model consists of the SMILES of the potential molecules separated by "." and the spectrum in the optimal data format as developed above. With this input, the model is tasked to either generate the SMILES of the correctly matching molecule or, if no molecule in the set matches the spectrum, a non matching token.

We develop two methods to generate the input sets. Firstly, a reaction dataset in which we provide the model with both the reagents and products of a reaction and either a matching or not matching NMR spectrum. Secondly, a dataset consisting of molecules randomly picked from all molecules present in Pistachio. We vary the number of molecules from two to seven to evaluate the models performance as the size of the set increases.

The performance of the models is evaluated on the two test sets sampled using the methods above (reaction set (rxn), molecule set (mol)). We measure the overall performance of the model using the top–1 accuracy metrics. Top-1 Accuracy is an appropriate metric in this case as the matching and non-matching set are balanced. Results of the experiments can be found in Table 2. Table 2 shows that the random baseline achieves an average top–1 accuracy of 27.05%, which is consistent with an average set size of 3.44.

Table 2: Top–1 Accuracy of the model in predicting either the correct matching SMILES or a non-matching token. We show the performance of both the best and an ensemble of the five best checkpoints.

Model	Rxn (Top–1 Acc. %)		Mol (Top–1 Acc. %)	
	Matching	Non-Matching	Matching	Non-Matching
Random Baseline	25.05	30.54	24.75	27.68
^{1}H–Model	95.66	98.99	**95.86**	97.73
^{1}H–Model Ensemble of 5	**96.08**	**99.12**	95.55	**97.97**
^{13}C–Model	**97.08**	99.18	96.10	97.81
^{13}C–Model Ensemble of 5	96.34	**99.23**	**96.30**	**98.03**
^{1}H $+^{13}$C–Model	97.56	99.45	**96.13**	98.55
^{1}H $+^{13}$C–Model Ensemble of 5	**98.28**	**99.58**	95.16	**99.49**

All models demonstrate a high performance in both matching a spectrum to a molecule and detecting if a spectrum does not match any of the molecules in the set. The accuracy is notably higher when evaluated on non-matching test sets. This can potentially be attributed to the relative ease of identifying a mismatch between a spectrum and the molecules contained in the set, compared to accurately pinpointing the correct match especially if the similarity between some molecule is high.

The models also show high proficiency in correctly predicting the SMILES of the matching molecule. We investigate the model mistakes in this task by assigning the wrong predictions into three categories: 1) “Non-Matching”: the model predicts a non-matching token instead of the expected SMILES, 2) “Other Molecule in the set”: the model predicts a SMILES found in the input set which does not correspond to the target SMILES and 3) “Incorrect SMILES”: the model predicts a SMILES sequence that is either not not found in the input set or incorrect. Results for the three ensemble model are shown in Table 3.

Table 3: Analysis of the incorrect predictions on the matching testsets.

Model	Testset	Incorrect Predictions (%)		
		Non-Matching	Other Molecule in Set	Incorrect SMILES
^{1}H–Model	Mol	96.81	2.56	0.64
	Rxn	10.85	88.22	0.93
^{13}C–Model	Mol	95.46	3.67	0.86
	Rxn	37.72	60.92	1.37
^{1}H $+^{13}$C–Model	Mol	99.14	0.51	0.34
	Rxn	6.69	92.20	1.11

Across all three models we observe a distinctly different distribution between the molecule and reaction set. For the former, most mistakes occur via the model incorrectly predicting a non-matching token. On the other hand for the reaction-testset the most common mistake consists of the model predicting a different molecule that can also be found in the set of molecules provided to the model. These differences can be attributed to the higher similarity found between molecules in the reaction sets causing the model to incorrectly assign the NMR to another molecule in the set. Reassuringly, all three models can reliably generate accurate SMILES strings from the input set. Errors arising from incorrect SMILES or SMILES that are not contained in the input set account for less than 2% of all mistakes.

We do not observe a significant different in between the different modalities. Models trained on ^{1}H, ^{13}C, and the combination perform within $\sim$1-2% of top–1 accuracy. This is points to all three modalities containing enough information to correctly match an NMR spectrum to a molecule.

Overall, our findings demonstrate that a transformer model can accurately assign a molecule to an NMR spectrum when provided with a set of reactants and products from a reaction, achieving a high level of accuracy.

2.5 Limitations

One of the key limitations of our methodological approach lies in the availability of large NMR datasets. While these datasets exist, licenses for their use are often expensive and restrict machine learning applications, limiting their use. Consequently, we opt to simulate NMR spectra using MestreNova. While this approach is not inherently limiting, it is important to note that the resulting spectra are highly coherent and consistent. Experimental spectra likely exhibit greater variability and inconsistencies.

3 Conclusions

NMR spectroscopy is a very powerful tool routinely used by chemists. The analysis of spectra, or rather their use for structure elucidation, remains a primarily manual task. Taking in consideration the number of spectra analyzed every day in the world, it is surprising that few data-driven approaches to aid in this process have been adopted so far. In this work, we explored ways to change that.

To this end, we presented a transformer model capable of predicting the molecular structure directly from NMR spectra. We trained and optimised the transformer model to predict the molecular structure from the ^{1}H, ^{13}C, and combined ^{1}H/^{13}C NMR spectra. We report a top–1 accuracy of 58.0%, 53.9% and 67.0% for the tasks on simulated spectra, respectively. In different experiments, we observe that weaknesses present in models trained on a single modality can be eliminated by combining the two modalities. Erroneous model predictions are very similar to the target molecules, with an average Tanimoto similarity of 0.55 for the model trained ^{1}H and ^{13}C spectra. This demonstrates that the model predictions, even when incorrect, provide chemists with structure guesses that are close to the correct compound.

In another task, we train models to select, among potential candidates, the molecule corresponding to an NMR spectrum. We find that for all three modalities the model is able to accomplish this task with a top–1 accuracy above 95%.

The models trained on simulated data in this work will provide a basis for fine-tuning on experimental datasets, allowing the models to leverage the fundamentals learned from the simulated spectra while adapting to the variability and noise found in experimental ones.

These advancements hold the potential to transform the analysis of NMR spectra, enabling faster and more accurate identification and characterization of compounds. As a result, the integration of automated NMR analysis into the workflow of high-throughput experiments promises to enhance efficiency and accelerate discoveries in the field of chemistry.

References

[1] Qingxin Li and CongBao Kang. A Practical Perspective on the Roles of Solution NMR Spectroscopy in Drug Discovery. *Molecules*, 25(13):2974, 2020.

[2] David R. Klein. *Organic Chemistry*. Wiley, 2013.

[3] Gregory M. Banik, Grace Baysinger, Prashant V. Kamat, and Norbert Pienta. *The ACS Guide to Scholarly Communication*. American Chemical Society, 2020.

[4] Gayathri Dev Ammini, Jordan P. Hooker, Joren Van Herck, Anil Kumar, and Tanja Junkers. Comprehensive high-throughput screening of photopolymerization under light intensity variation using inline NMR monitoring. *Polym. Chem.*, 14(22):2708–2716, 2023.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. arXiv:1706.03762.

[7] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.*, 5(9):1572–1583, 2019.

[8] Zhichao Liu, Ruth A. Roberts, Madhu Lal-Nag, Xi Chen, Ruili Huang, and Weida Tong. AI-based language models powering drug discovery and development. *Drug Discovery Today*, 26(11):2593–2607, 2021.

[9] Andres M. Bran, Sam Cox, Andrew D. White, and Philippe Schwaller. ChemCrow: Augmenting large-language models with chemistry tools, 2023. arXiv:2304.05376.

[10] Melodie Christensen, Lars P. E. Yunker, Parisa Shiri, Tara Zepel, Paloma L. Prieto, Shad Grunert, Finn Bork, and Jason E. Hein. Automation isn't automatic. *Chemical Science*, 12(47):15473–15490, 2021.

[11] Milad Abolhasani and Eugenia Kumacheva. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth*, 2(6):483–492, 2023.

[12] Steven M. Mennen, Carolina Alhambra, C. Liana Allen, Mario Barberis, Simon Berritt, Thomas A. Brandt, Andrew D. Campbell, Jesús Castañón, Alan H. Cherney, Melodie Christensen, David B. Damon, J. Eugenio de Diego, Susana García-Cerrada, Pablo García-Losada, Rubén Haro, Jacob Janey, David C. Leitch, Ling Li, Fangfang Liu, Paul C. Lobben, David W. C. MacMillan, Javier Magano, Emma McInturff, Sebastien Monfette, Ronald J. Post, Danielle Schultz, Barbara J. Sitter, Jason M. Stevens, Iulia I. Strambeanu, Jack Twilton, Ke Wang, and Matthew A. Zajac. The Evolution of High-Throughput Experimentation in Pharmaceutical Development and Perspectives on the Future. *Org. Process Res. Dev.*, 23(6):1213–1242, 2019.

[13] Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin Bateman, Louis-Charles Campeau, Jonathan Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim Cernak, and Spencer D. Dreher. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53, 2015.

[14] Damith Perera, Joseph W. Tucker, Shalini Brahmbhatt, Christopher J. Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W. Sach. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434, 2018.

[15] Michael Shevlin. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.*, 8(6):601–607, 2017.

[16] Babak Mahjour, Rui Zhang, Yuning Shen, Andrew McGrath, Ruheng Zhao, Osama G. Mohamed, Yingfu Lin, Zirong Zhang, James L. Douthwaite, Ashootosh Tripathi, and Tim Cernak. Rapid planning and analysis of high-throughput experiment arrays for reaction discovery. *Nat Commun*, 14(1):3924, 2023.

[17] Adam Cook, Roxanne Clément, and Stephen G. Newman. Reaction screening in multiwell plates: high-throughput optimization of a Buchwald–Hartwig amination. *Nat Protoc*, 16(2):1152–1169, 2021.

[18] MestreLab, MNova. `https://mestrelab.com/software/mnova/` (Accessed July 24, 2023).

[19] ACD Labs, NMR Workbook Suite. `https://www.acdlabs.com/products/spectrus-platform/nmr-workbook-suite/` (Accessed July 24, 2023).

[20] 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, 2021.

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

[22] Weiwei Wei, Yuxuan Liao, Yufei Wang, Shaoqi Wang, Wen Du, Hongmei Lu, Bo Kong, Huawu Yang, and Zhimin Zhang. Deep Learning-Based Method for Compound Identification in NMR Spectra of Mixtures. *Molecules*, 27(12):3653, 2022.

[23] Iván Cortés, Cristina Cuadrado, Antonio Hernández Daranas, and Ariel M. Sarotti. Machine learning in computational NMR-aided structural elucidation. *Frontiers in Natural Products*, 2, 2023.

[24] Jinzhe Zhang, Kei Terayama, Masato Sumita, Kazuki Yoshizoe, Kengo Ito, Jun Kikuchi, and Koji Tsuda. NMR-TS: de novo molecule identification from NMR spectra. *Science and Technology of Advanced Materials*, 21(1):552–561, 2020.

[25] NextMove Software, Pistachio. `https://www.nextmovesoftware.com/pistachio.html` (Accessed July 24, 2023).

[26] Marvin Alberts, Teodoro Laino, and Alain C. Vaucher. Leveraging Infrared Spectroscopy for Automated Structure Elucidation, 2023. DOI: 10.26434/chemrxiv-2023-5v27f.

[27] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988.

[28] David Weininger, Arthur Weininger, and Joseph L. Weininger. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, 29(2):97–101, May 1989.

[29] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, 2015.

[30] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010.

[31] OpenNMT-py: Open-Source Neural Machine Translation, 2017. `https://github.com/OpenNMT/OpenNMT-py` (Accessed July 24, 2023).

[32] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation, 2017. arXiv:1701.02810.

[33] RDKit. `https://www.rdkit.org/` (Accessed July 24, 2023).

[34] Daylight: SMARTS Examples. `https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html` (Accessed April 20, 2023).

Appendix

A Methods

A.1 Synthetic Data

Before generating spectra, 1,029,381 reactions were sampled from the Pistachio patent dataset [25]. A set of molecules was assembled from the precursors and products of these reactions. Molecules were filtered out if they contain atoms other than carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorous and the halogens. In addition, all molecules with a heavy atom count outside the range of 5–35, charged molecules or containing isotope information were filtered out.

From this set, 1,120,390 ^{1}H and 1,943,950 ^{13}C NMR spectra were generated using MestreNova. Standard simulation settings were used for ^{1}H NMRs. For ^{13}C NMRs, ^{1}H and ^{19}F decoupled spectra were generated. For ^{13}C NMR, the position of all peaks was recorded. On the other hand ^{1}H NMR were further processed. First peak-picking was applied, followed by the autointegration and automultiplet assignment. All three processing steps were carried out using built-in MestreNova functions with standard settings. For each peak in an ^{1}H NMR, the range of the peak, its centroid, the number of hydrogen atoms and the multiplet was recorded.

A.2 Model

We base our model architecture on the Molecular Transformer [7]. The model takes the formatted NMR spectrum with the chemical formula as input, and outputs a molecular structure encoded as SMILES. This can be formulated as a translation task from the spectrum to the molecular structure. The model is a vanilla transformer as implemented in the OpenNMT-py library [31, 32] with the following hyperparameters deviating:

```
word_vec_size: 512
hidden_size: 512
layers: 4
batch_size: 4096
```

All models are trained for 350k steps amounting to approximately 35h on a A100 GPU.

A.3 Tokenization

To tokenize ^{1}H NMR peaks, we proceed as follows. The position of the peak is rounded to the second decimal point, the type of multiplet (singlet, doublet, triplet, etc.) and the number of hydrogens are appended as second and third token respectively. All peaks are separated with a separating token (“`|`”). As an example a singlet at 1.239 ppm with an integral of 3 would become “`1.24 s 3H |`”, with tokens separated by whitespaces. A string of the ^{1}H NMR spectrum is built accordingly by concatenating the peaks starting with the lowest ppm and ending at the highest one. In addition, a prefix token is used to differentiate ^{1}H from ^{13}C NMR spectra. As an example an ^{1}H NMR with two peaks would be formatted as follows: “`1HNMR 1.24 t 3H | 1.89 q 3H |`”.

^{13}C NMR are formatted according to a simpler scheme. As the multiplet type and integration is not relevant for this type of spectrum the position of the peaks are rounded to one decimal point and tokenized accordingly. To illustrate this, a typical NMR spectrum is tokenized as follows: “`13CNMR 12.1 27.8 63.5`”.

In addition to the spectra, the model is provided the chemical formula in addition to the NMR spectrum. The formula is calculated using RDKit [33] and prepended to the spectrum.

When both ^{1}H and ^{13}C NMR are used, the tokenized string consists first of the chemical formula, followed by the ^{1}H NMR spectrum and finally the ^{13}C NMR. To have the ^{1}H and ^{13}C NMR share the same token space, the ppm values of the ^{13}C NMR peaks are divided by 10.

A.4 Data augmentation

The spectra are augmented using jitter augmentation as used previously by Jonas et. al. [21]. This involves adding a random distortion sampled from a range of 0.5 ppm for ^{1}H NMR and 5 ppm for

^{13}C NMR. The random noise is added to each of the peaks in the spectra. In total, two augmented spectra are produced for each original one.

B Functional group definitions

Functional groups are defined in SMARTS as shown in Table 4. Using these SMARTS and RDKit the presence of a certain function group is determined by invoking `<RDKit molecule>.GetSubstrucMatches(<RDKit molecule from SMARTS pattern>)`

Table 4: Functional group definitions used.

	Definition
Alcohol	`[OX2H][CX4;!$(C([OX2H])[O,S,#7,#15])]`
Carboxylic Acid	`[CX3](=O)[OX2H1]`
Ester	`[#6][CX3](=O)[OX2H0][#6]`
Ether	`[OD2]([#6])[#6]`
Aldehyde	`[CX3H1](=O)[#6]`
Ketone	`[#6][CX3](=O)[#6]`
Alkene	`[CX3]=[CX3]`
Alkyne	`[$([CX2]#C)]`
Benzene	`c1ccccc1`
Primary Amine	`[NX3;H2;!$(NC=[!#6]);!$(NC#[!#6])][#6]`
Secondary Amine	`[NH1,nH1])`
Tertiary Amine	`[NH0,nH0])`
Amide	`[NX3][CX3](=[OX1])[#6]`
Cyano	`[NX1]#[CX2]`
Fluorine	`[#6][F]`
Chlorine	`[#6][Cl]`
Iodine	`[#6][I]`
Bromine	`[#6][Br]`
Sulfonamide	`[#16X4]([NX3])(=[OX1])(=[OX1])[#6]`
Sulfone	`[#16X4](=[OX1])(=[OX1])([#6])[#6]`
Sulfide	`[#16X2H0]`
Phosphoric Acid[†]	`[$(P(=[OX1])([$([OX2H]),$([OX1-]),$([OX2]P)])([$([OX2H]),$([OX1-]),$([OX2]P)])[$([OX2H]),$([OX1-]),$([OX2]P)]),$([P+]([OX1-])([$([OX2H]),$([OX1-]),$([OX2]P)])([$([OX2H]),$([OX1-]),$([OX2]P)])[$([OX2H]),$([OX1-]),$([OX2]P)])]`

[†] Adapted from [34]

C Performance on molecules containing specific functional groups

In Tables 5, 6, and 7, the accuracy of the model solely trained on ^{1}H, ^{13}C, and combined ^{1}H /^{13}C NMR data, respectively, is shown depending on the presence of specific functional groups in the target molecule. “Count” represents the number of molecules with this functional group in the test set. Additionally, the average heavy atom count (“Avg. HAC” in the table) is calculated to rule out bias.

Table 5: The model trained on ^{1}H NMR spectra’s ability to predict the correct molecular structure based on if a specific functional group is present in the target molecule.

	Count	Avg. HAC	Top–1%	Top–5%	Top–10%
Phosphoric Acid	76	27.09	31.58	47.37	48.68
Alkene	12727	22.55	46.94	66.87	70.46
Cyano	7691	23.58	53.54	71.92	75.83
Alkyne	2071	23.39	54.23	71.61	74.89
Alcohol	17214	22.86	54.23	74.83	78.49
Sulfide	15214	23.85	55.06	73.41	77.06
Primary Amine	12504	21.30	55.42	75.99	79.57
Amide	31834	26.10	56.13	74.26	77.77
Chlorine	23685	23.59	56.42	75.31	78.95
Tertiary Amine	83118	24.01	56.74	74.85	78.30
Carboxylic Acid	13838	23.26	56.79	77.03	80.60
Ketone	8100	22.35	56.91	73.10	76.35
Secondary Amine	56201	24.50	56.96	75.16	78.65
Fluorine	30166	25.16	57.70	75.59	78.97
Ether	34926	24.98	58.75	76.93	80.16
Sulfone	2428	26.03	58.86	75.41	78.46
Benzene	86972	24.08	58.86	76.92	80.18
Sulfonamide	5758	26.44	59.48	76.55	79.63
Ester	16344	23.20	59.50	79.08	82.13
Aldehyde	2208	19.09	60.19	79.71	83.02
Bromine	9687	20.11	60.48	80.21	83.47
Iodine	1728	19.93	62.21	82.52	85.30

Table 6: The model trained on ^{13}C NMR spectra's ability to predict the correct molecular structure based on if a specific functional group is present in the target molecule.

	Count	Avg. HAC	Top–1%	Top–5%	Top–10%
Phosphoric Acid	142	26.54	36.62	55.63	59.86
Alkene	21149	23.09	40.65	60.35	64.87
Alcohol	21781	23.11	48.25	69.14	74.11
Sulfide	26917	23.90	50.54	69.80	74.08
Primary Amine	22672	21.39	51.14	72.07	76.61
Amide	51806	26.33	51.29	69.90	74.21
Cyano	13327	23.34	51.30	70.18	74.57
Chlorine	40757	23.39	51.43	71.69	76.30
Secondary Amine	85969	24.94	51.97	71.35	75.62
Tertiary Amine	146144	24.10	52.21	71.32	75.66
Fluorine	49707	25.00	52.33	72.09	76.49
Carboxylic Acid	18879	23.21	54.47	75.34	79.46
Iodine	3193	19.40	54.56	76.20	80.74
Sulfone	4928	25.85	54.61	71.25	75.59
Benzene	149174	24.31	54.79	73.83	77.95
Alkyne	3700	23.31	54.89	73.14	76.89
Bromine	17680	19.99	55.71	76.84	81.46
Sulfonamide	9319	26.70	55.77	73.00	76.97
Ketone	14910	22.41	56.32	73.66	77.94
Ether	65246	25.10	56.60	75.00	78.97
Aldehyde	4452	19.25	57.46	78.23	82.88
Ester	33632	23.47	58.11	78.01	81.80

Table 7: The model trained on both ^{1}H and ^{13}C NMR spectra's ability to predict the correct molecular structure based on if a specific functional group is present in the target molecule.

	Count	Avg. HAC	Top–1%	Top–5%	Top–10%
Phosphoric Acid	71	25.82	38.03	50.70	54.93
Alkene	12799	22.36	54.68	74.48	77.40
Alcohol	16967	22.77	62.32	82.14	85.04
Primary Amine	12378	21.36	63.94	83.16	85.85
Sulfide	15219	24.17	64.02	80.92	83.68
Amide	32013	26.12	64.90	82.31	85.11
Chlorine	23849	23.58	65.57	82.79	85.53
Secondary Amine	56290	24.50	65.67	82.79	85.55
Cyano	7767	23.61	65.91	82.17	84.97
Fluorine	30724	25.09	66.00	82.98	85.66
Sulfone	2537	26.08	66.30	81.75	84.15
Tertiary Amine	83173	24.01	66.31	82.98	85.59
Carboxylic Acid	13719	23.30	66.80	85.41	87.82
Alkyne	2070	23.49	66.86	83.24	85.89
Ketone	8241	22.29	67.10	82.55	85.01
Ester	16499	23.25	67.66	85.24	87.50
Benzene	87374	24.05	67.85	84.40	86.86
Ether	34823	24.86	67.87	84.35	86.75
Sulfonamide	5663	26.58	68.20	83.68	86.39
Iodine	1705	19.88	68.68	85.34	87.21
Bromine	9838	20.19	69.66	86.91	89.04
Aldehyde	2152	19.11	70.77	87.04	89.22

D Tanimoto Similarity Distribution

In Figure 5 the Tanimoto similarity distribution for all three models is illustrated. The distribution shows a peak around 0.55 for all three models.

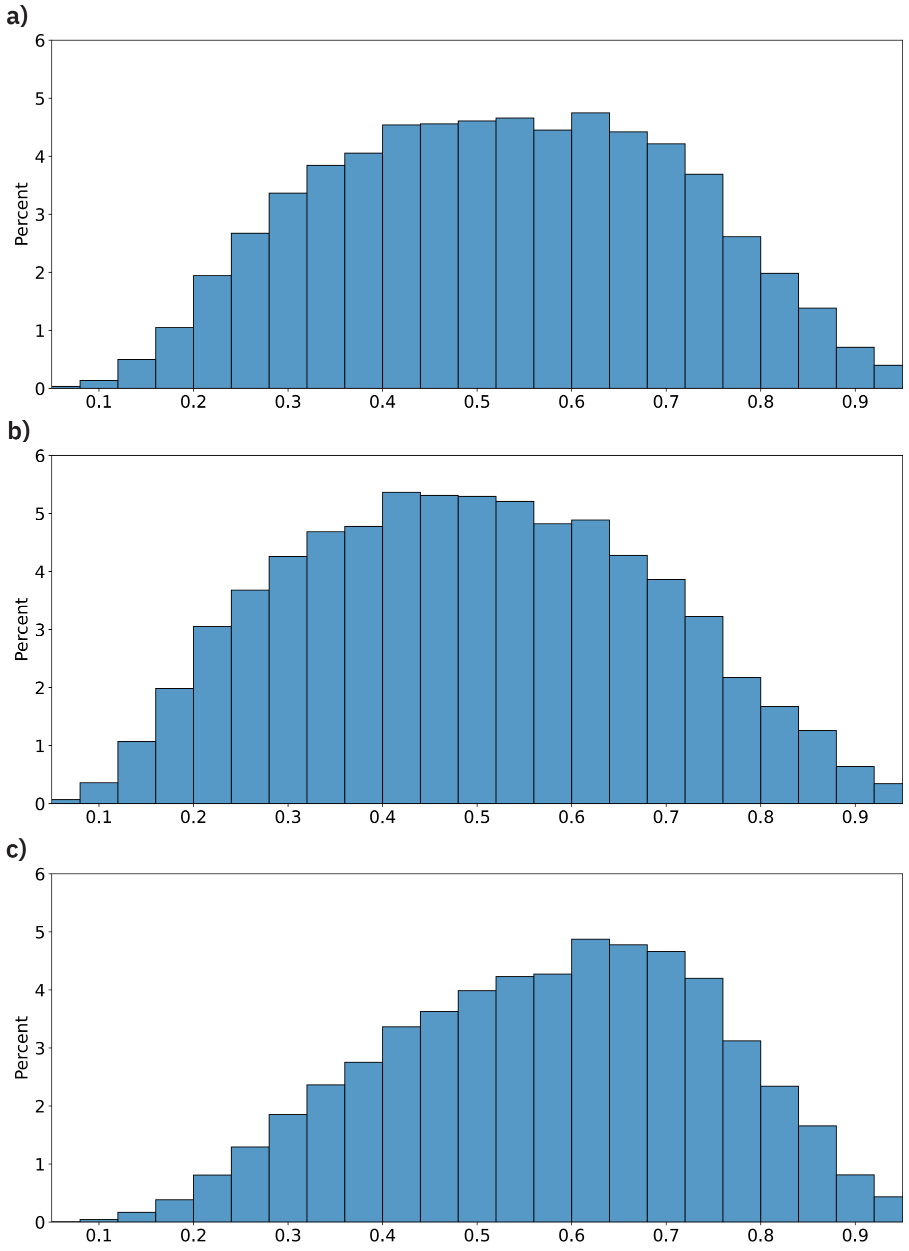

Figure 5: The Tanimoto distribution of three models: a) ^{1}H NMR, b) ^{13}C NMR, c) ^{1}H+^{13}C NMR. All correct molecules were excluded.