Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry

Marvin Alberts[1,2,4*] **Oliver Schilter**[1,3,4*] **Federico Zipoli**[1,4]
Nina Hartrampf[2] **Teodoro Laino**[1,4]
[1]IBM Research [2]University of Zürich [3]EPFL [4]NCCR Catalysis
marvin.alberts@ibm.com
{oli, fzi, teo}@zurich.ibm.com
nina.hartrampf@chem.uzh.ch

Abstract

Spectroscopic techniques are essential tools for determining the structure of molecules. Different spectroscopic techniques, such as Nuclear magnetic resonance (NMR), Infrared spectroscopy, and Mass Spectrometry, provide insight into the molecular structure, including the presence or absence of functional groups. Chemists leverage the complementary nature of the different methods to their advantage. However, the lack of a comprehensive multimodal dataset, containing spectra from a variety of spectroscopic techniques, has limited machine-learning approaches mostly to single-modality tasks for predicting molecular structures from spectra. Here we introduce a dataset comprising simulated ^{1}H-NMR, ^{13}C-NMR, HSQC-NMR, Infrared, and Mass spectra (positive and negative ion modes) for 790k molecules extracted from chemical reactions in patent data. This dataset enables the development of foundation models for integrating information from multiple spectroscopic modalities, emulating the approach employed by human experts. Additionally, we provide benchmarks for evaluating single-modality tasks such as structure elucidation, predicting the spectra for a target molecule, and functional group predictions. This dataset has the potential automate structure elucidation, streamlining the molecular discovery pipeline from synthesis to structure determination. The dataset and code for the benchmarks can be found at `https://rxn4chemistry.github.io/multimodal-spectroscopic-dataset`.

1 Introduction

The rapid advancement of artificial intelligence (AI) and machine learning (ML) methods has ushered in a new era for the field of chemistry. Computational approaches have transformed various aspects of chemical research, including retrosynthesis planning [1, 2, 3, 4], reaction optimization through Bayesian optimization [5, 6, 7, 8], molecular design [9, 10, 11, 12] and more. Tasks that were previously laborious and time-consuming when performed manually are now being automated, accelerating the discovery process. Despite these advancements, one critical aspect of chemistry that remains heavily reliant on human expertise is structural elucidation – the process of determining the molecular structure from spectroscopic data.

While chemists often have an intuition about a molecule that was synthesized, the actual composition of the product needs to be verified using spectroscopic data. Different spectroscopic techniques yield different types of information. For instance, certain functional groups (e.g., alcohols) will exhibit characteristic peaks in specific regions of the infrared (IR) spectrum (e.g., 3200-3300 cm^{-1} [13]),

[†] Equal Contributions

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

while the mass spectrum (MS) can be used to find the molecular weight of a molecule in question. Similar to solving a complex puzzle, the more spectroscopic modalities a chemist has access to, the more information and hints they can gather to predict the molecular structure and explain the observed spectral peaks.

While AI/ML models have been developed for this task, they predominantly focus on single spectroscopic modalities, such as infrared (IR) [14, 15] or nuclear magnetic resonance (NMR) [16, 17] spectroscopy. In contrast, human experts leverage multiple modalities by combining information from various spectroscopic techniques to gain a better understanding of the molecular structure. To bridge this gap and enable the automation of structural elucidation, there is a need for a multimodal dataset containing spectra from a variety of spectroscopic techniques.

Multimodal datasets in other fields, such as computer vision and natural language processing [18, 19, 20, 21, 22], have enabled remarkable achievements like text-to-image generation [23, 24, 25], image captioning [26], object detection using bounding boxes [27, 28], and even multitask models [29, 30]. Similarly, we postulate that a multimodal dataset for chemical spectra could lead to significant advancements. Such a dataset would serve as a valuable resource for developing AI/ML models capable of integrating information from multiple spectroscopic modalities, emulating the approach employed by human experts in analyzing and interpreting spectral data.

In this paper, we introduce a dataset comprising simulated IR, ^{1}H-NMR, ^{13}C-NMR, Heteronuclear Single Quantum Coherence (HSQC)-NMR, positive-ion mass spectrometry (MS), and negative-ion MS spectra for a large set of 790k realistic molecules extracted from patent data. We specifically sample molecules from the United States Patent Office (USPTO) dataset, which is commonly used for reaction prediction [31, 32]. We also introduce initial baseline models for single-modality tasks, namely predicting molecular structures from spectral data, generating spectras from molecular structures, and identifying functional groups present in molecules based on spectral information. These models demonstrate the potential of our dataset for automated molecular structure elucidation and serve as benchmarks for evaluating other AI architectures on these tasks.

By leveraging AI/ML methodologies and the comprehensive information from multiple spectroscopic modalities, this dataset has the potential to close the loop between automated synthesis and automated structural elucidation, streamlining the molecular discovery cycle.

2 Related Work

USPTO Dataset: The USPTO dataset, by Lowe [33], has become a staple for machine learning based works in chemistry [2, 31, 32, 34]. It is sourced from patent data and in contrast to many other datasets in chemistry fully open source, making a popular choice for training and evaluating reaction prediction models. The main advantage of this dataset is that all molecules are sourced from patent data, i.e. their distribution is very similar to molecules common in industry and to a lesser extent academia.

NMR: Predicting the chemical structure from NMR spectra remains a largely unexplored subject. Jonas [35] first utilized imitation learning to predict the molecular structure from ^{13}C-NMR spectra. Sridharan et al. [36] approached the problem from a different angle using a reinforcement learning guided Monte Carlo tree search to generate molecules from ^{13}C-NMR spectra. The first work to combine both ^{1}H- and ^{13}C-NMR spectra employed 1D-CNNs to predict substructures contained in the parent molecule from both ^{1}H and ^{13}C-NMR spectra. Subsequently, a database search is employed to provide the closest match [16]. More recently Alberts et al. [14] demonstrated that Transformer models are capable of generating molecular structure from annotated NMR spectra.

However, very few other works have been published and comparison between the approaches is rare. A few studies have evaluated model performance on the experimental spectra available in the nmrshiftdb2 database [37] but most works utilise different private datasets. While some exclusively train on a limited amount of experimental spectra, most simulate a large number of spectra and pretrain on these simulated spectra. At the time of writing, none of the simulated datasets used in these works are publicly available hindering the transparent benchmarking of model architectures.

IR: Similar to NMR spectra few works investigate full structure elucidation from IR spectra. Alberts et al. [14] showed that it is possible to predict the chemical structure from IR spectra of small molecules. On the other hand, predicting the presence of certain functional groups from IR spectra

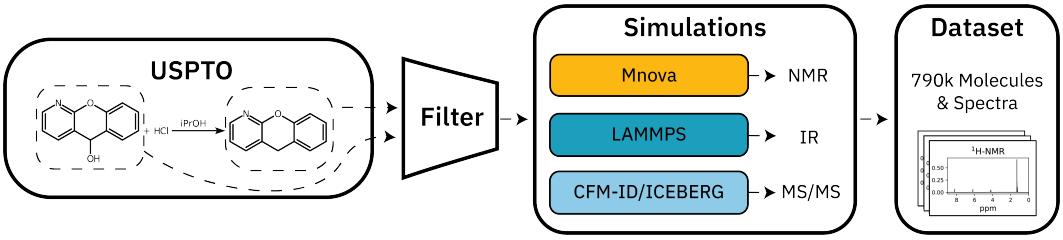


Figure 1: Overall workflow: Molecules are extracted from reaction data (USPTO), filtered to only contain certain atom types as well as minimum and maximum molecule size, then for each molecule the corresponding spectra are simulated resulting in a dataset of spectra for 790k molecules.

has been assessed extensively [38, 39, 40, 41]. For this application there are also no standardized datasets with each work using different spectra for training and evaluation. The closest standardised dataset, the NIST Gas-Phase IR Database [42] contains solely 5,228 spectra limiting the applicability for machine models.

MS/MS: Out of the three spectroscopic methods, structure elucidation from MS/MS spectra is the most explored and commonly used in laboratories. However, most approaches rely on matching a given MS/MS spectra to a large database inherently limited by the size and diversity of the database [43]. In another approach a fragmentation tree is derived from the MS/MS spectrum and matched to a database of fragmentation trees. While also relying on a database, this approach is less limited as fragmentation trees can be predicted with relatively high fidelity [44, 45, 46, 47]. To remove the need for database matching some works have proposed predicting the chemical structure directly from the MS/MS spectrum. Of these MSNovelist [48] relies on an LSTM whereas MassGenie [49] utilizes a Transformer model to predict the structure as Simplified molecular-input line-entry system (SMILES) [50]. As with NMR and IR spectra, the datasets used for training mass spectrometry models are often not publicly accessible. These datasets typically require a commercial license for access, such as the NIST MS database[51]. However, efforts to create open-access repositories of experimentally measured data are underway, with the GNPS database[52] being a notable example.

3 Dataset

Since in organic chemistry spectral data is often acquired during reactions to monitor progress or after completion, a dataset intended for inferring molecular structures should encompass a chemical space similar to that accessible through common organic chemistry reactions. Therefore, we chose to utilize the USPTO reaction dataset, mined by Lowe [33] who extracted chemical reactions from the US patent database. This dataset spans 1,435,481 chemical reactions across various reaction classes and as such only contains realistic molecular structures and commonly used chemicals such as solvents, reactant and reagent. We identified all unique molecules from these reactions and applied filtering criteria based on the heavy atom count (all atoms except Hydrogen), retaining only those molecules with more than five and fewer than 35 heavy atoms. Additionally, we filtered out molecules containing elements other than Carbon, Hydrogen, Oxygen, Nitrogen, Sulfur, Phosphorus, Silicon, Boron, and the halogens. This reduced the number of molecules from 1,675,439 to 1,416,499. We attempted to simulate all molecules; however, since some simulations failed for certain molecules, we opted to include only those molecules for which all spectra simulations were successful (see Figure1).

Overall, we ended up with 794.403 unique molecules and their corresponding IR, ^{1}H-NMR, ^{13}C-NMR, HSQC-NMR, and MS/MS spectra (for more details about the simulations, refer to Section 3.1). The molecular structures is represented as SMILES and additionally the molecular formula of each molecule (e.g. $C_6H_{12}O_6$) is provided. The distribution of SMILES lengths and heavy atom counts is visualized in Figure 2 (A), spanning the full range between 5 and 35 heavy atoms. Additionally, the chemical similarity between 200 randomly sampled molecules was investigated by calculating the Tanimoto similarity of their chemical fingerprints (see Figure 2 (C)). It can be seen that the dataset comprises a broad range of dissimilar chemical structures which is desired.

The chemical similarity is weakly correlated with the similarity in the IR spectra domain, as shown in Figure 2 (D), indicating that molecules with similar chemical compositions may also have similar IR spectra. For all similarity calculation refer to section A.4.

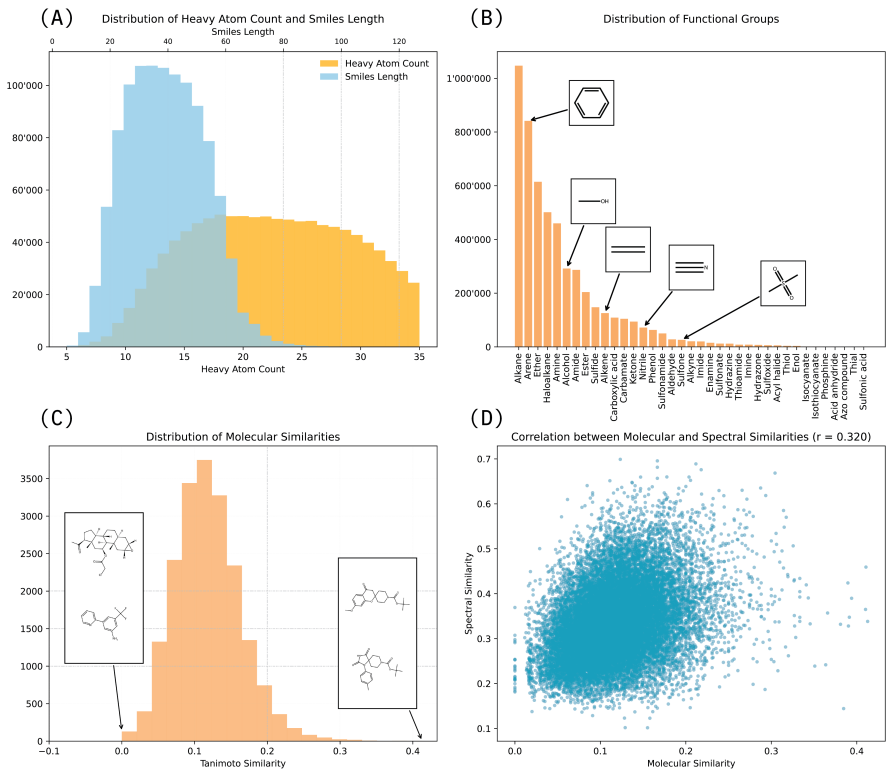


Figure 2: (A) Size and functional group (B) distribution of the full dataset. 200 randomly sampled molecules were investigated for their chemical similarity to each (C) as well as if the IR spectra similarity correlates to the chemical similarity (D).

As chemical functional groups are often distinctly responsible for patterns in certain areas of the spectra (e.g., aromatic rings causing peaks in the range of 6.0 – 8.7 ppm in the ^{1}H-NMR, as exemplified in Section A.3), we analyzed the functional group compositions of our collected dataset. In Figure 2 (B), the distribution is visualized. As can be seen, the most prevalent functional groups are Alkanes, Arenes, and Ethers, followed by Haloalkanes, and overall spanning a broad range of functional groups.

Overall our dataset comprises 790k unique molecules and their spectra, spanning a large diverse space regarding their chemical similarity, molecule size, as well as functional group composition.

3.1 Data Generation

An overview of the generated data can be found in Table 1. In addition, the spectra and annotations for two molecules are shown in Appenxix section A.1.

NMR Simulations: We employ MestReNova [53] to simulate ^{1}H-, ^{13}C- and HSQC-NMR spectra. The spectra were simulated using deuterated Chloroform as solvent. Default settings were used for all simulations. For ^{13}C-NMR spectra, ^{1}H decoupled spectra were generated.

We utilize the in-built spectral analysis tools of MestreNova to annotate the spectra. For ^{1}H-NMR spectra, we employ the automultiplet analysis function yielding a set of peaks, the type of each peak e.g. doublet, triplet, etc., and the normalized integration of the peak. The same method yields the position and intensity of the peaks in the ^{13}C-NMR spectra. Similarly, we obtain the position and integration for peaks in the HSCQC spectra.

IR Simulations: IR spectra can be simulated either by approximating the bonds in the molecule as harmonic oscillators and calculating their frequencies or by measuring the dipole-dipole moment of the molecule over time [54, 55]. While the first approach is computationally cheaper it only yields the position and intensity of each peak in the spectrum which can subsequently be broadened e.g. via

Table 1: Overview of the data available for the different modalities. For all modalities except IR we provide annotations in addition to the unprocessed spectrum.

Modality	Subtype	Data Description
IR	Spectrum	Vector of size 1.800
^{1}H-NMR	Spectrum	Vector of size 10.000
	Annotated Spectrum	Start, End, Centroid, Integration and Type of each peak
^{13}C-NMR	Spectrum	Vector of size 10.000
	Annotated Spectrum	Centroid and Intensity of each peak
HSQC-NMR	Spectrum	Matrix: 512x512
	Annotated Spectrum	X, Y coordinates and integration of each peak
Positive MS/MS	Spectrum	m/z & Intensity of each peak
	m/z Annotations	Chemical formula corresponding to the m/z of each peak
Negative MS/MS	Spectrum	m/z & Intensity of each peak
	m/z Annotations	Chemical formula corresponding to the m/z of each peak

a Gaussian function. Overtones and anharmonicities are neglected by this approach. On the other hand, a simulated IR spectrum derived from dipole-dipole data does contain these features at the expense of higher computation requirements.

We developed a high throughput pipeline to orchestrate molecular dynamics simulations and calculate the spectra from the molecule's dipole moment. Based on a molecule as a SMILES string we generate the corresponding Protein Data Bank (PDB) file and optimize the geometry of the molecule with the General AMBER Force Field (GAFF) [56]. We choose the same force field for the molecular dynamics simulation and generate the input files for a Large-scale Atomic-Molecular Massively Parallel Simulator (LAMMPS) [57] simulation using AMBER tools [58]. The system is allowed to equilibrate for 250 ns, before recording the dipole moment of the molecule for a further 250 ns. IR spectra are calculated from the dipole moment according to Braun [59]. The simulated spectra have a range from 400–4000 cm^{-1} with a resolution of 2 cm^{-1}.

MS/MS Simulations: The development MS/MS simulation tools is advancing rapidly, with new tools and approaches emerging frequently. To capture the current state of the art, we selected three distinct methods that represent different approaches to MS/MS simulation: Competitive Fragmentation Modeling for Metabolite Identification 4.0 (CFM-ID 4.0) [60], Subformula Classification for Autoregressively Reconstructing Fragmentations (SCARF) Goldman et al. [47], and ICEBERG Goldman et al. [46]. While SCARF and ICEBERG employ pure machine learning approaches, CFM-ID represents a hybrid methodology combining machine learning with rule-based systems. Important to note is that we use the publically available checkpoints for both SCARF and ICEBERG. The performance report in Goldman et al. [47] and Goldman et al. [46] was obtained using the closed source NIST20 database.

We simulate positive mode Electrospray Ionisation (ESI) MS/MS spectra using hydrogen adducts using all three methods. Additionally, we use CFM-ID to simulate negative mode spectra, generating spectra at three ionization energies (10eV, 20eV, and 40eV) in both positive and negative modes. All three tools provide chemical formula annotations for the fragments in their simulated MS/MS spectra.

3.2 Experimental vs Simulated Spectra

To evaluate the similarity of the simulated spectra to experimental ones, we compared them with a set of 251 molecules and their corresponding experimentally measured spectra from Van Bramer and Bastin [61]. Out of these 251 molecules, 96 had all spectroscopic techniques measured and were also simulated in the dataset introduced in this manuscript (excluding HSQC-NMR). Since each spectral technique has a different representation, multiple approaches were required for comparison. Table 2 presents the spectral similarity between the real-world measured spectra and our simulated approach. Additionally, as a comparison metric, the similarity of each the 96 simulated spectra versus

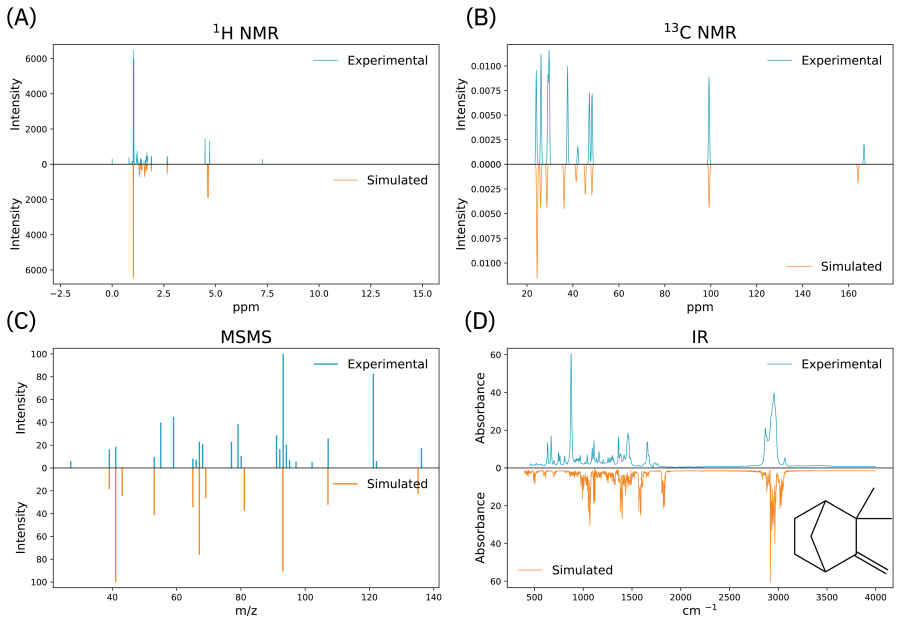


Figure 3: Simulated vs experimentally measured (A) ^{1}H-NMR (B) ^{13}C-NMR (C) MS/MS [+40 eV] and (D) IR of the molecule 2,2-dimethyl-3-methylenebicyclo[2.2.1]heptane (shown in the lower right).

Table 2: Similarity metrics between experimental and simulated spectra. For the MS/MS only positive modes are compared as the experimental spectra were measured in this mode.

Spectrum	Sim. vs Exp.	Sim vs other Exp.	Similarity Metrics
IR	31.5	26.6	Cosine Similarity
MS/MS (CFM-ID, Positive) [10 eV]	30.1	17.9	CosineGreedy Similarity [62]
MS/MS (CFM-ID, Positive) [20 eV]	40.1	22.6	CosineGreedy Similarity [62]
MS/MS (CFM-ID, Positive) [40 eV]	48.9	26.6	CosineGreedy Similarity [62]
MS/MS (SCARF, Positive)	14.1	8.2	CosineGreedy Similarity [62]
MS/MS (ICEBERG, Positive)	17.0	10.9	CosineGreedy Similarity [62]
^{1}H-NMR	21.9	6.8	Cosine Similarity
^{13}C-NMR	48.4	8.6	CosineGreedy Similarity [62]

all experimental spectra was calculated. The similarity metrics used are listed in the table (their definitions can be found in Appendix section A.4).

Relying primarily on cosine similarity-based metrics has a significant limitation. For example, in the case of NMR spectra, even if the shape and integral are accurately simulated, a slight peak shift compared to the experimentally measured spectrum (a common effect caused by the solvent [63]) can result in a drastic reduction in cosine similarity. While a chemist would consider the two compared spectra similar, the cosine similarity score would be substantially lower than when the peaks are aligned. Despite this limitation, it can be seen that, on average, all simulated spectra have a higher similarity to their corresponding experimental spectra compared to average similarity against all other experimental spectra. This shows that the simulated data represents somewhat realistically experimentally measured spectra. For visual inspection of the similarity between simulated and experimental spectra see Figure 3. A more in-depth analysis of the similarity between experimental and simulated spectra on larger datasets is presented in appendix section A.5.

4 Benchmarks

In the following we will present benchmarks on predicting the correct structure, functional groups contained in a molecule and generating spectra from a given molecule. We only evaluate performance on single modalities and leave exploring multimodal tasks for future work. All experiments are conducted with five fold cross validation. An overview of the different tasks is shown in Figure 4.

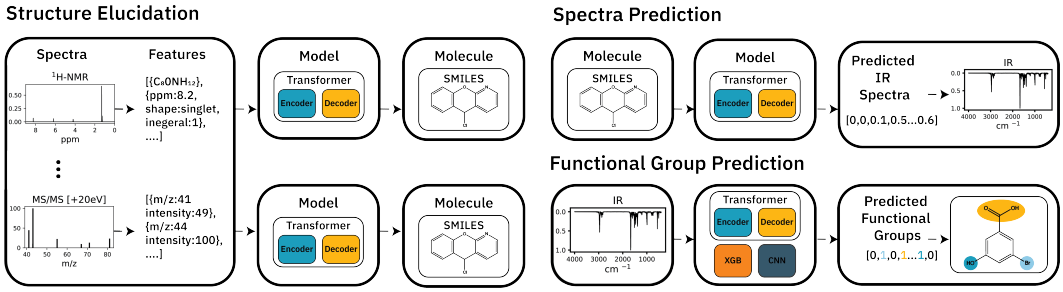


Figure 4: Overview of the benchmarks. Left: Structure elucidation using transformer models. We convert each spectra into a structured text representation to make it ingestible by the model. Top Right: Generation of spectra from molecules using a transformer model. We reuse the same structured text representation. Bottom right: Predicting functional groups from the spectra as a multilabel multiclass classification problem. We assess transformers, a 1D-CNN and gradient boosted trees.

4.1 Structure Elucidation from Spectra

As described in the introduction, we envision full structure elucidation from spectra as the primary use case for this dataset. To this end, we provide baseline results on predicting the exact chemical structure from spectra (see Table 3). We train a vanilla encoder-decoder transformer model [64, 65] on each individual modality and on the combination of ^{1}H- and ^{13}C-NMR. More information on the exact model and parameters used can be found in Appendix section A.6. In addition to the spectra, we provide the models with the chemical formula, i.e. the elements present in the molecule, as a prior. The chemical formula can be obtained experimentally via high resolution MS.

To train a transformer model on the spectra we convert the spectra into a structured text representation. For IR and NMR spectra we follow the representations described in earlier works by Alberts et al. [17, 14]. For IR spectra this representation converts the spectrum to a set of 400 tokens each sampled from a fixed position in the spectrum and bins the intensities to tokens. For ^{13}C-NMR spectra the representation provides the model with the position of each peak in the spectrum, whereas ^{1}H-NMR spectra we provide the integration and type of each peak in addition its beginning and end. We train a

Table 3: Top–1, Top–5, and Top–10 Accuracy (see Appendix A.7) of a Transformer model trained to predict the chemical structure (SMILES) from the different modalities.

	Top–1%	Top–5%	Top–10%
IR	9.97 ± 0.46	21.23 ± 0.33	24.01 ± 0.42
MS/MS (CFM-ID, Negative)	20.98 ± 0.23	39.32 ± 0.19	44.93 ± 0.29
MS/MS (CFM-ID, Positive)	23.53 ± 0.21	42.59 ± 0.14	47.53 ± 0.31
MS/MS (SCARF, Positive)	1.92 ± 0.11	5.26 ± 0.37	6.81 ± 0.48
MS/MS (ICEBERG, Positive)	15.52 ± 2.10	31.46 ± 3.28	36.22 ± 3.45
^{13}C-NMR	51.95 ± 0.29	70.01 ± 0.21	74.12 ± 0.30
^{1}H-NMR	64.99 ± 0.31	81.94 ± 0.31	84.07 ± 0.32
^{1}H-NMR + ^{13}C-NMR	73.38 ± 0.08	87.94 ± 0.14	89.98 ± 0.16

model on the spectra generated by SCARF, ICEBERG and CFM-ID. In addition, for CFM-ID we train a model both for the positive as well as the negative MS/MS spectra. Each peak in a spectra is described using the peaks m/z and intensity. Examples for all representations can be found in Appendix section A.8.

We observe the worst performance for models solely trained on IR spectra followed by MS/MS spectra. On the other hand, ^{1}H- and ^{13}C-NMR perform relatively well. Encouragingly the combination of both ^{1}H- and ^{13}C-NMR performs the best. These results can be explained by the information contained in each modality. While IR spectra can be leveraged easily to determine the functional groups present in a molecule, for larger and more complex molecules the peaks in the spectrum start to overlap rendering it difficult to extract information. The low performance on MS/MS spectra is caused by similar factors: The more complex the molecule, the larger the number of potential fragmentations, increasing the difficulty of assigning a definite structure. Models trained on ^{1}H-NMR spectra perform better than ^{13}C-NMR as ^{1}H-NMR spectra typically contain more information. However, as the two types of spectra probe different aspects of the molecule they complement each other resulting in a performance increase of 7.8% when combined. We also conducted zero-shot experiments of the models trained on simulated data on the Van Bramer and Bastin [61] dataset. These results are shown in appendix section A.9.

4.2 Functional Group prediction

Another task that can be explored using the dataset is predicting the functional groups present in the structure from the spectra. We extract functional groups from the molecules using the SMARTS [66] pattern defined in A.2. While not as useful to chemists as full structure elucidation, the success of a chemical reaction can in most cases be determined by a change in functional groups. We approach this task as a multiclass, multilabel classification problem. As such we evaluate the performance of three different models, a boosted tree classifier [67], a 1D-CNN as implemented by Jung et al. [41] and a transformer model, in predicting the functional groups present in the target molecule. The performance of the models on the modalities is shown in Table 4. We train the boosted gradient tree and 1D-CNN on the non processed vector form of each spectrum. In contrast we employ the same representations as used in for structure elucidation task for the transformer model. Unlike the previous task, we do not include the chemical formula as an input.

Across four modalities the transformer trained on the structured text representations outperforms both the 1D-CNN and the gradient boosted trees. Only on IR spectra is the performance of the 1D-CNN marginally better than the Transformer model. In contrast to MS/MS and the NMR spectra, IR spectra are not sparse explaining the good performance of the 1D-CNN.

4.3 Spectra prediction

We primarily conceived the dataset to explore structure elucidation. However, the dataset can also be used for the reverse, i.e. predicting the corresponding spectrum given a target molecule. To this end we train a transformer model to predict the from the molecule for each modality. We use the same

Table 4: F1 scores for predicting functional groups from the different spectra.

Spectrum	XGBoost	1D-CNN [41]	Transformer
IR	0.834 ± 0.001	$\mathbf{0.895 \pm 0.002}$	0.881 ± 0.021
MS/MS (CFM-ID, Positive)	0.725 ± 0.002	0.645 ± 0.001	$\mathbf{0.897 \pm 0.012}$
MS/MS (CFM-ID, Negative)	0.761 ± 0.001	0.648 ± 0.006	$\mathbf{0.905 \pm 0.009}$
MS/MS (SCARF, Positive)	0.763 ± 0.001	0.737 ± 0.003	$\mathbf{0.771 \pm 0.004}$
MS/MS (ICEBERG, Positive)	0.734 ± 0.001	0.677 ± 0.001	$\mathbf{0.885 \pm 0.007}$
^{13}C-NMR	0.804 ± 0.001	0.674 ± 0.056	$\mathbf{0.913 \pm 0.017}$
^{1}H-NMR	0.797 ± 0.003	0.839 ± 0.005	$\mathbf{0.935 \pm 0.031}$

Table 5: Cosine similarity and token accuracy of transformer models when predicting spectra from structure. We predict an individual MS/MS spectra for each ionisation energy

Spectrum	Cosine Similarity	Token Accuracy
IR	23.91 ± 0.14	13.55 ± 0.16
MS/MS (Positive) [10 eV]	83.94 ± 0.10	31.58 ± 0.09
MS/MS (Positive) [20 eV]	77.09 ± 0.18	11.05 ± 0.13
MS/MS (Positive) [40 eV]	66.35 ± 0.15	6.94 ± 0.16
MS/MS (Negative) [10 eV]	82.87 ± 0.25	33.92 ± 0.19
MS/MS (Negative) [20 eV]	75.86 ± 0.18	11.82 ± 0.11
MS/MS (Negative) [40 eV]	69.50 ± 0.23	8.95 ± 0.17
MS/MS (SCARF, Positive)	66.39 ± 0.03	5.04 ± 0.08
MS/MS (ICEBERG, Positive)	63.17 ± 0.01	4.62 ± 0.04
^{13}C-NMR	92.69 ± 0.31	35.7 ± 0.27
^{1}H-NMR	94.86 ± 0.29	17.93 ± 0.24

model architecture and representations as in section 4.1. This mean that while we predict the whole spectrum for IR spectra, for all other modalities a processed form of the spectrum is generated. In the case of the MS/MS spectra this consists of the m/z of each peak and it's intensity. Similarly for ^{13}C-NMR spectra the model predicts the position of each peak. However, for ^{1}H-NMR spectra we predict the start and end of each peak, it's type and integration.

To compare the predicted and target spectrum we use two similarity metrics: One one hand we employ greedy cosine similarity and on the other the exact token accuracy. For MS/MS, ^{13}C- and ^{1}H-NMR spectra we compute the cosine similarity by first aligning the peaks in the predicted and target spectrum before calculating the similarity. The results are shown in Table 5.

For IR spectra we observe both a low cosine and token similarity. This is likely caused by the represenation used for predicting the spectra as a sequence of 400 tokens has to be generated for each spectrum. Other approaches such as graph neural networks as proposed by McGill et al. [55] may show better performance. For both positive and negative MS/MS we observe a decrease in performance with an increase in the ionisation energy, likely a result of molecules fragmenting to a larger extent at higher ionisation energy and as such resulting in a more complex spectrum. Predicted ^{1}H- and ^{13}C-NMR spectra both exhibit a high cosine similarity while showing a small token accuracy. This is caused by two factors: One one hand only the position of the peak is used to calculate the similarity and on the other hand the token accuracy requires an exact match of the token, i.e. even if the predicted peak has an error of only 0.1ppm it would be deemed false.

4.4 Other tasks to explore

While we benchmark three different tasks that could be of interest to researchers, the dataset can also be used for various other ML applications. The following ideas serve as starting points for potential research opportunities to explore.

Including more information: Typically, when predicting the structure of molecules from spectra, chemists do not start from scratch. Instead, they leverage prior knowledge of the reaction performed to make informed initial guesses about the structure. These guesses may include the desired product, the starting material, or plausible side products. Chemists then use clues obtained from various spectra to confirm or eliminate these initial hypotheses. In this context, we propose a task where a model is provided with a set of molecules and a spectrum, and it predicts which molecule from the set corresponds to the given spectrum.

Mixtures: While the previous sections discussed structure elucidation from pure compounds, in reality chemists typically encounter mixtures far more often than pure compounds, e.g. a reaction is a mixture of a set of compounds. The mixtures commonly need to be separated into their constituent components before definite structure elucidation can be carried out. If the components of a mixture could identified accurately this would greatly aid chemists. The spectra of mixtures can be constructed as convex combinations of their constituent components for NMR and IR spectra. As such this dataset could be used to construct the spectra of complex mixture based on which the components of the mixture can be predicted.

Representations: While our study utilized SMILES as the primary form to represent molecules there are a variety of different chemical representations that could be explored, reaching from SELFIES[68], deep-SMILES[69] as well as generating graph instead of a text-based representation. Additionally, the spectral information could be represented in various ways. For instance, generating figures from vector representations and employing image-based models.

Multimodal Approaches: Combining the different types of spectra for a true multimodal model is significantly harder than e.g. predicting the structure from a single modality. Not only does the best representation for each modality need to be considered but also how to combine the different modalities. We suggest the following approach: First optimise the representation for each modality individually before considering how to combine them. Here the approaches range from early fusion, in practice often a simple concatenation of the different modalities, over medium fusion, embed each modality then fuse, to late fusion. Another consideration is how to process each modality and weigh each modality. Each of these factors would need to be considered when building a multimodal model.

5 Conclusion and Limitations

In this study we introduce the first multimodal spectroscopic dataset for structure elucidation including six different types of spectra for 790k molecules sampled from USPTO. We conduct a series of experiments on structure elucidation, generating a spectrum from a molecule and evaluate boosted gradient trees, 1D-CNNs and transformer models on classifying which functional groups are present in a molecule based on the spectra.

However, the dataset also has multiple limitations, the largest being that all data is simulated. As a result, there is a distribution shift between simulated and experimental spectra and models trained on the data in this set will likely benefit from further finetuning on experimental spectra. The efficacy of these models is inherently determined by the fidelity of the underlying simulation used to generate the spectra. Another limitation is the chemical space covered by USPTO. The USPTO dataset contains molecules sourced from patents and as a result, is biased towards synthesisable molecules with applications in industry. This means that models trained on our dataset may not perform as well on molecules outside of this scope, e.g. natural products.

We hope that this work can address the severe lack of openly available spectroscopic datasets and serve as a foundation to build models capable of automated structure elucidation for chemistry and with that streamline the molecular discovery pipeline from synthesis to structure determination.

Acknowledgments and Disclosure of Funding

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

Code and Data availability

The code for the benchmark and the dataset used in this study is publicly available under the following link: `https://github.com/rxn4chemistry/multimodal-spectroscopic-dataset`.

References

[1] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

[2] Connor W Coley, William H Green, and Klavs F Jensen. Machine learning in computer-aided synthesis planning. *Accounts of chemical research*, 51(5):1281–1289, 2018.

[3] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.

[4] Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020.

[5] Jonas B Mockus. *The Bayesian approach to global optimization*. Freie Univ., Fachbereich Mathematik, 1984.

[6] Connor J Taylor, Kobi C Felton, Daniel Wigh, Mohammed I Jeraal, Rachel Grainger, Gianni Chessari, Christopher N Johnson, and Alexei A Lapkin. Accelerated chemical reaction optimization using multi-task learning. *ACS Central Science*, 9(5):957–968, 2023.

[7] Jeff Guo, Bojana Ranković, and Philippe Schwaller. Bayesian optimization for chemical reactions. *Chimia*, 77(1/2):31–31, 2023.

[8] Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, 2022.

[9] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[10] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[11] Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Greta Markert, and María Rodríguez Martínez. Paccmannrl: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *Iscience*, 24(4), 2021.

[12] Matteo Manica, Jannis Born, Joris Cadow, Dimitrios Christofidellis, Ashish Dave, Dean Clarke, Yves Gaetan Nana Teukam, Giorgio Giannone, Samuel C Hoffman, Matthew Buchan, et al. Accelerating material design with the generative toolkit for scientific discovery. *npj Computational Materials*, 9(1):69, 2023.

[13] Jerry Workman Jr and Lois Weyer. *Practical guide to interpretive near-infrared spectroscopy*. CRC press, 2007.

[14] Marvin Alberts, Teodoro Laino, and Alain C Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. 2023.

[15] Wenwen Zhang, Liyanaarachchi Chamara Kasun, Qi Jie Wang, Yuanjin Zheng, and Zhiping Lin. A review of machine learning for near-infrared spectroscopy. *Sensors*, 22(24):9764, 2022.

[16] 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. *Chemical Science*, 12(46):15329–15338, 2021.

[17] Marvin Alberts, Federico Zipoli, and Alain C Vaucher. Learning the language of nmr: Structure elucidation from nmr spectra using transformer models. 2023.

[18] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021.

[19] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021.

[20] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

[21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[22] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. `https://github.com/kakaobrain/coyo-dataset`, 2022.

[23] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[24] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[26] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[27] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[29] David Mizrahi, Roman Bachmann, Oguzhan Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5): 434–443, 2017.

[32] 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 central science*, 5(9):1572–1583, 2019.

[33] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[34] Alessandra Toniato, Philippe Schwaller, Antonio Cardinale, Joppe Geluykens, and Teodoro Laino. Unassisted noise reduction of chemical reaction datasets. *Nature Machine Intelligence*, 3(6):485–494, 2021.

[35] Eric Jonas. Deep imitation learning for molecular inverse problems. *Advances in neural information processing systems*, 32, 2019.

[36] Bhuvanesh Sridharan, Sarvesh Mehta, Yashaswi Pathak, and U Deva Priyakumar. Deep reinforcement learning for molecular inverse problem of nuclear magnetic resonance spectra to molecular structure. *The Journal of Physical Chemistry Letters*, 13(22):4924–4933, 2022.

[37] Stefan Kuhn, Heinz Kolshorn, Christoph Steinbeck, and Nils Schlörer. Twenty years of nmrshiftdb2: A case study of an open database for analytical chemistry. *Magnetic Resonance in Chemistry*, 62(2):74–83, 2024.

[38] Jonathan A Fine, Anand A Rajasekar, Krupal P Jethava, and Gaurav Chopra. Spectral deep learning for prediction and prospective validation of functional groups. *Chemical science*, 11 (18):4618–4630, 2020.

[39] Abigail A Enders, Nicole M North, Chase M Fensore, Juan Velez-Alvarez, and Heather C Allen. Functional group identification for ftir spectra using image-based machine learning models. *Analytical Chemistry*, 93(28):9711–9718, 2021.

[40] Tianyi Wang, Ying Tan, Yu Zong Chen, and Chunyan Tan. Infrared spectral analysis for prediction of functional groups based on feature-aggregated deep learning. *Journal of Chemical Information and Modeling*, 63(15):4615–4622, 2023.

[41] Guwon Jung, Son Gyo Jung, and Jacqueline M Cole. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chemical Science*, 14(13):3600–3609, 2023.

[42] Stephen E Stein. Nist 35. nist/epa gas-phase infrared database-jcamp format. 2008.

[43] Armen G Beck, Matthew Muhoberac, Caitlin E Randolph, Connor H Beveridge, Prageeth R Wijewardhane, Hilkka I Kenttamaa, and Gaurav Chopra. Recent developments in machine learning for mass spectrometry. *ACS Measurement Science Au*, 2024.

[44] Sebastian Böcker and Kai Dührkop. Fragmentation trees reloaded. *Journal of cheminformatics*, 8:1–26, 2016.

[45] Arpana Vaniya and Oliver Fiehn. Using fragmentation trees and mass spectral trees for identifying unknown compounds in metabolomics. *TrAC Trends in Analytical Chemistry*, 69:52–61, 2015.

[46] Samuel Goldman, Janet Li, and Connor W. Coley. Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks. *Analytical Chemistry*, 96(8):3419–3428, 2024.

[47] Samuel Goldman, John Bradshaw, Jiayi Xin, and Connor W. Coley. Prefix-Tree Decoding for Predicting Mass Spectra from Molecules. *Advances in neural information processing systems*, 37, 2023.

[48] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. Msnovelist: de novo structure generation from mass spectra. *Nature Methods*, 19(7):865–870, 2022.

[49] Aditya Divyakant Shrivastava, Neil Swainston, Soumitra Samanta, Ivayla Roberts, Marina Wright Muelas, and Douglas B Kell. Massgenie: A transformer-based deep learning method for identifying small molecules from their mass spectra. *Biomolecules*, 11(12):1793, 2021.

[50] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28 (1):31–36, 1988.

[51] NIST. Nist mass spectrum. `https://www.nist.gov/mml/biomolecular-measurement/mass-spectrometry-data-center`, 2022.

[52] Mingxun Wang, Jeremy J Carver, Vanessa V Phelan, Laura M Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, Jeramie Watrous, Clifford A Kapono, Tal Luzzatto-Knaan, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nature biotechnology*, 34(8):828–837, 2016.

[53] MNova. `https://mestrelab.com/software/mnova/` (Accessed September 29, 2023).

[54] Beatriz von der Esch, Laurens DM Peters, Lena Sauerland, and Christian Ochsenfeld. Quantitative comparison of experimental and computed ir-spectra extracted from ab initio molecular dynamics. *Journal of Chemical Theory and Computation*, 17(2):985–995, 2021.

[55] Charles McGill, Michael Forsuelo, Yanfei Guan, and William H Green. Predicting infrared spectra with message passing neural networks. *Journal of Chemical Information and Modeling*, 61(6):2594–2609, 2021.

[56] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25 (9):1157–1174, 2004.

[57] Aidan P Thompson, H Metin Aktulga, Richard Berger, Dan S Bolintineanu, W Michael Brown, Paul S Crozier, Pieter J In't Veld, Axel Kohlmeyer, Stan G Moore, Trung Dac Nguyen, et al. Lammps-a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, 2022.

[58] David A Case, Hasan Metin Aktulga, Kellon Belfon, David S Cerutti, G Andrés Cisneros, Vinícius Wilian D Cruzeiro, Negin Forouzesh, Timothy J Giese, Andreas W Götz, Holger Gohlke, et al. Ambertools. *Journal of chemical information and modeling*, 63(20):6183–6191, 2023.

[59] Efrem Braun. Calculating An IR Spectra From A Lammps Simulation, 2016. URL `https://zenodo.org/record/154672`. 10.5281/ZENODO.154672.

[60] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, and David S Wishart. Cfm-id 4.0: More accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Analytical chemistry*, 93(34):11692–11700, 2021.

[61] Scott E Van Bramer and Loyd D Bastin. Spectroscopy data for undergraduate teaching. *Journal of Chemical Education*, 100(10):3897–3902, 2023.

[62] Florian Huber, Stefan Verhoeven, Christiaan Meijer, Hanno Spreeuw, Efraín Manuel Villanueva Castilla, Cunliang Geng, Justin JJ van der Hooft, Simon Rogers, Adam Belloum, Faruk Diblen, et al. matchms-processing and similarity evaluation of mass spectrometry data. *bioRxiv*, pages 2020–08, 2020.

[63] AD Buckingham, T Schaefer, and WG Schneider. Solvent effects in nuclear magnetic resonance spectra. *The Journal of Chemical Physics*, 32(4):1227–1233, 1960.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

[66] Christiane Ehrt, Bennet Krause, Robert Schmidt, Emanuel SR Ehmki, and Matthias Rarey. Smarts. plus–a toolbox for chemical pattern design. *Molecular Informatics*, 39(12):2000216, 2020.

[67] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System, June 2016. arXiv:1603.02754.

[68] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020.

[69] Noel O'Boyle and Andrew Dalke. Deepsmiles: an adaptation of smiles for use in machine-learning of chemical structures. *Chemrxiv*, 2018.

[70] Daylight, online. Daylight chemical information systems. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, 2024.

[71] RDKit, online. RDKit: Open-source cheminformatics. http://www.rdkit.org, 2024.

[72] NextMove Software: Pistachio. URL https://www.nextmovesoftware.com/pistachio.html.

[73] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

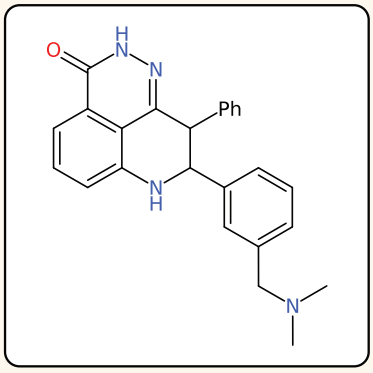
[74] François Chollet et al. Keras. https://keras.io, 2015.

[75] OpenNMT-py: Open-Source Neural Machine Translation, 2017. URL https://github.com/OpenNMT/OpenNMT-py. https://github.com/OpenNMT/OpenNMT-py (Accessed April 20, 2023).

A Appendix

A.1 Sample of the dataset

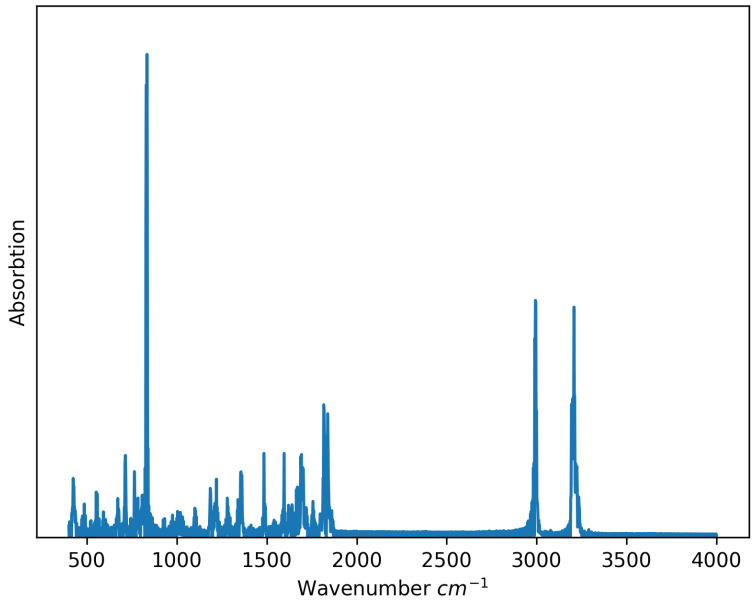
Molecule



Smiles String: `CN(C)Cc1cccc(C2Nc3cccc4c(=O)[nH]nc(c34)C2c2ccccc2)c1`

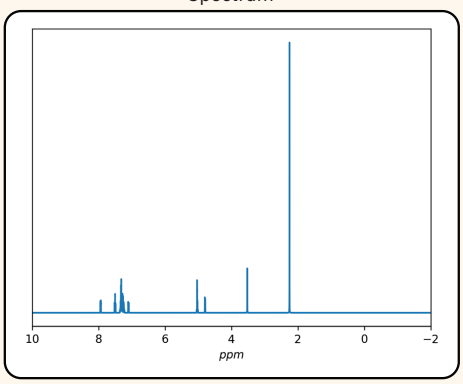
Chemical Formula: $C_{25}H_{24}N_4O$

Infrared Spectrum



^{1}H-NMR Spectrum

Spectrum

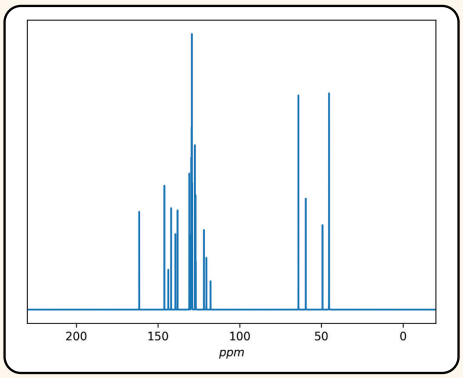


Annotations

	rangeMin	rangeMax	centroid	category	nH	j_values
0	7.915	7.915	7.938	dd	1	1.18, 8.14
1	7.473	7.473	7.504	t	1	8.24
2	7.205	7.205	7.300	m	11	N/A
3	7.076	7.076	7.100	dd	1	1.18, 8.51
4	4.996	4.996	5.032	m	2	N/A
5	4.772	4.772	4.795	dd	1	0.81, 6.56
6	3.508	3.508	3.524	d	2	0.92
7	2.239	2.239	2.254	s	5	N/A

^{13}C-NMR Spectrum

Spectrum

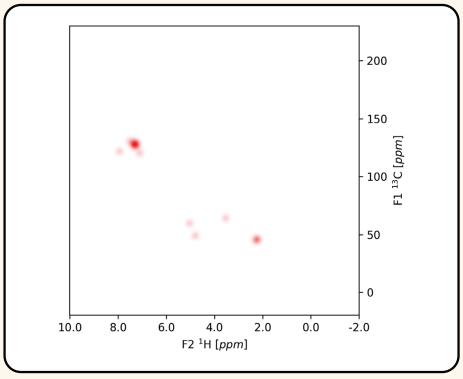


Annotations

	delta (ppm)	intensity	integral
0	161.487827	0.050	0.161
1	146.084370	0.050	0.161
2	143.694211	0.050	0.161
3	141.956434	0.050	0.161
4	139.327831	0.050	0.161
...	...	...	...
17	117.818310	0.050	0.161
18	64.048321	0.121	0.387
19	59.550320	0.085	0.274
20	49.301060	0.085	0.274
21	45.331450	0.312	1.000

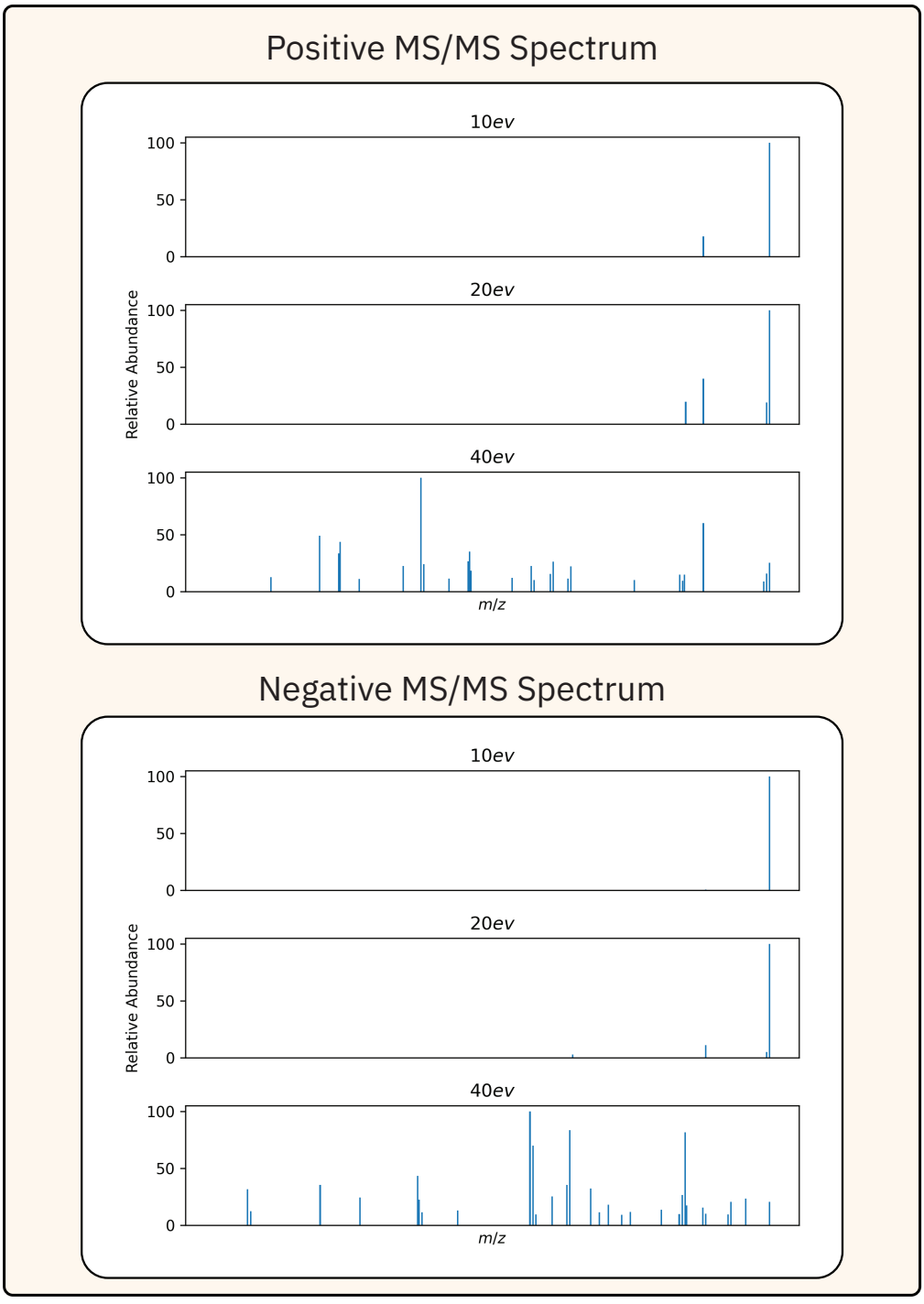
HSQC-NMR Spectrum

Spectrum

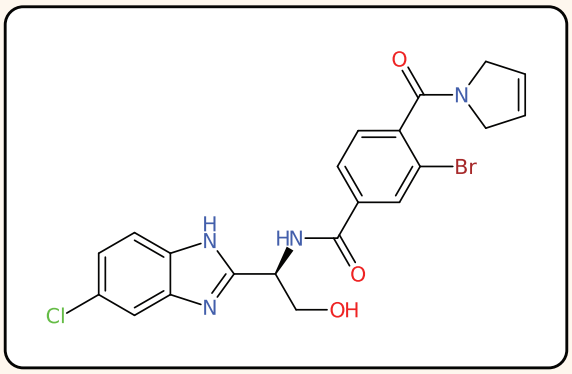


Annotations

	13C_centroid	13C_max	13C_min	1H_centroid	1H_max	1H_min	nH
0	45.434	45.434	45.434	45.434	45.434	45.434	3.0
1	121.614	121.614	121.614	121.614	121.614	121.614	1.0
2	48.853	48.853	48.853	48.853	48.853	48.853	1.0
3	63.991	63.991	63.991	63.991	63.991	63.991	1.0
4	59.596	59.596	59.596	59.596	59.596	59.596	1.0
5	127.273	127.273	127.273	127.273	127.273	127.273	6.0
6	128.939	128.939	128.939	128.939	128.939	128.939	3.0
7	130.404	130.404	130.404	130.404	130.404	130.404	1.0
8	120.149	120.149	120.149	120.149	120.149	120.149	1.0


Positive MS/MS Spectrum
10ev
20ev
40ev
100
50
0
Relative Abundance
m/z
Negative MS/MS Spectrum
10ev
20ev
40ev
100
50
0
Relative Abundance
m/z

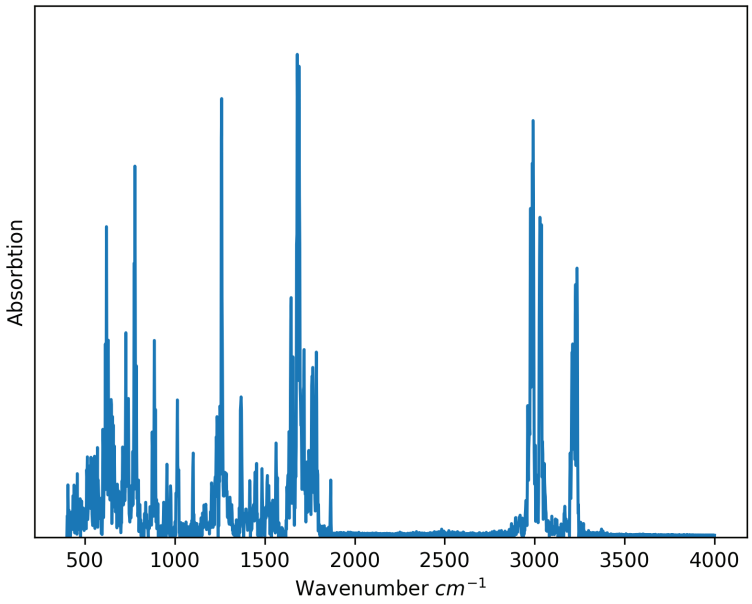
Molecule



Smiles String: `O=C(N[C@@H](CO)c1nc2cc(Cl)ccc2[nH]1) (continued)`
`c1ccc(C(=O)N2CC=CC2)c(Br)c1`

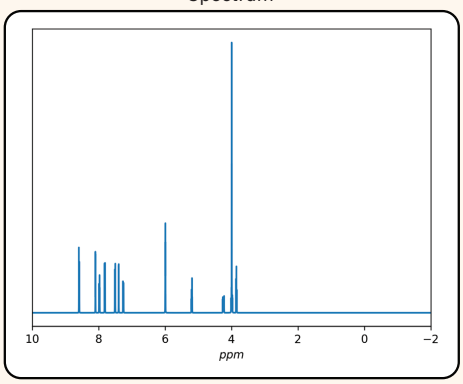
Chemical Formula: $C_{21}H_{18}BrClN_4O_3$

Infrared Spectrum



^{1}H-NMR Spectrum

Spectrum

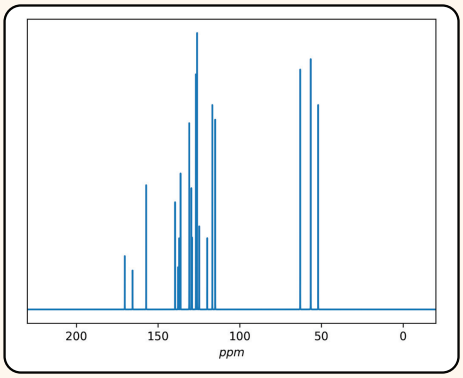


Annotations

	rangeMin	rangeMax	centroid	category	nH	j_values
0	8.561	8.561	8.586	d	1	7.51
1	8.075	8.075	8.094	d	1	2.1
2	7.949	7.949	7.977	dd	1	2.2, 8.79
3	7.786	7.786	7.814	d	1	8.87
4	7.479	7.479	7.504	d	1	7.79
...	...	...	...	...	...	...
7	5.967	5.967	5.990	p	2	1.48
8	5.164	5.164	5.195	dt	1	3.11, 7.51
9	4.205	4.205	4.243	ddd	1	3.11, 5.95, 13.18
10	3.952	3.952	3.994	m	5	N/A
11	3.822	3.822	3.852	t	1	5.95

^{13}C-NMR Spectrum

Spectrum

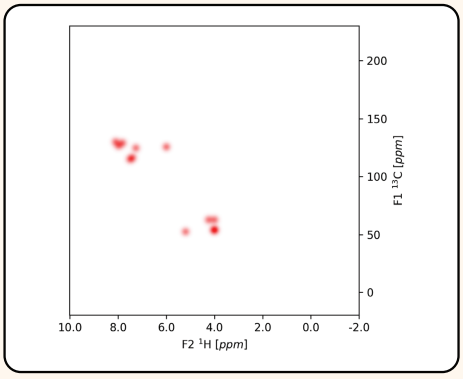


Annotations

	delta (ppm)	intensity	integral
0	170.348392	0.050	0.208
1	165.550906	0.050	0.208
2	157.239716	0.050	0.208
3	139.528124	0.050	0.208
4	137.746473	0.050	0.208
...	...	...	...
14	116.757712	0.085	0.354
15	114.974154	0.085	0.354
16	62.974370	0.121	0.500
17	56.528763	0.241	1.000
18	51.969721	0.085	0.354

HSQC-NMR Spectrum

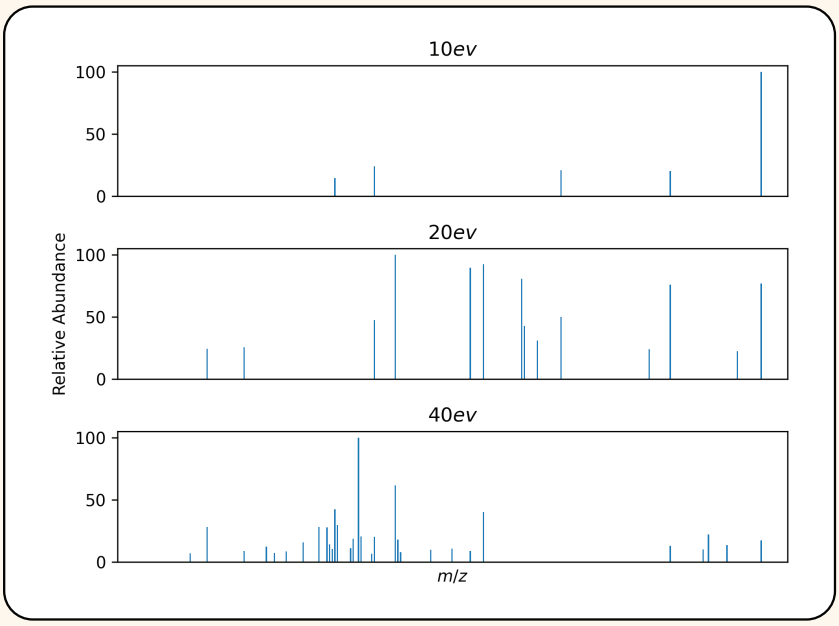
Spectrum



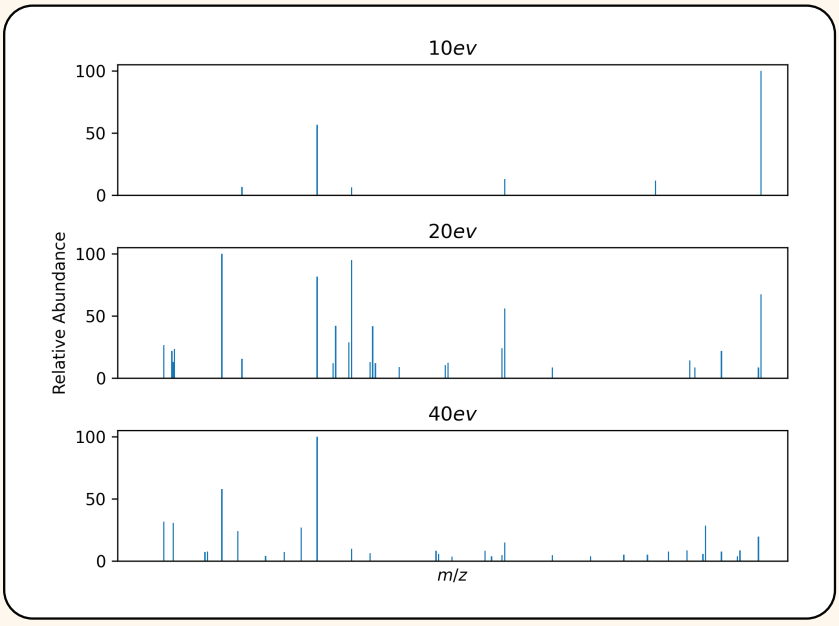
Annotations

	13C_centroid	13C_max	13C_min	1H_centroid	1H_max	1H_min	nH
0	53.736	53.736	53.736	53.736	53.736	53.736	2.0
1	125.521	125.521	125.521	125.521	125.521	125.521	1.0
2	115.733	115.733	115.733	115.733	115.733	115.733	1.0
3	129.916	129.916	129.916	129.916	129.916	129.916	1.0
4	124.544	124.544	124.544	124.544	124.544	124.544	1.0
5	52.271	52.271	52.271	52.271	52.271	52.271	1.0
6	128.939	128.939	128.939	128.939	128.939	128.939	1.0
7	126.498	126.498	126.498	126.498	126.498	126.498	1.0
8	62.526	62.526	62.526	62.526	62.526	62.526	1.0
9	62.526	62.526	62.526	62.526	62.526	62.526	1.0

Positive MS/MS Spectrum



Negative MS/MS Spectrum



A.2 Functional Group Analysis

The count of functional groups in the molecular structures was analyzed by SMARTS [70] pattern matching using RdKit [71]. The applied SMARTS patterns were derived from Jung et al. [41] and listed in Table 6. Each molecule can have only one occurrence of a particular functional group. For instance, if a molecule has two alcohol groups, the occurrence counts only once. However, each molecule can have multiple occurrences of different functional groups.

Table 6: SMARTS pattern used for functional group analysis.

Functional Group	SMARTS Pattern
Acid anhydride	[CX3] (= [OX1]) [OX2] [CX3] (= [OX1])
Acyl halide	[CX3] (= [OX1]) [F, Cl, Br, I]
Alcohol	[#6] [OX2H]
Aldehyde	[CX3H1] (= O) [#6, H]
Alkane	[CX4; H3, H2]
Alkene	[CX3] = [CX3]
Alkyne	[CX2] # [CX2]
Amide	[NX3] [CX3] (= [OX1]) [#6]
Amine	[NX3; H2, H1, H0; !$ (NC = O)]
Arene	[cX3] 1 [cX3] [cX3] [cX3] [cX3] [cX3] 1
Azo compound	[#6] [NX2] = [NX2] [#6]
Carbamate	[NX3] [CX3] (= [OX1]) [OX2H0]
Carboxylic acid	[CX3] (= O) [OX2H]
Enamine	[NX3] [CX3] = [CX3]
Enol	[OX2H] [#6X3] = [#6]
Ester	[#6] [CX3] (= O) [OX2H0] [#6]
Ether	[OD2] ([#6]) [#6]
Haloalkane	[#6] [F, Cl, Br, I]
Hydrazine	[NX3] [NX3]
Hydrazone	[NX3] [NX2] = [#6]
Imide	[CX3] (= [OX1]) [NX3] [CX3] (= [OX1])
Imine	[$ ([CX3] ([#6]) [#6]) , $ ([CX3H] [#6])] = [$ ([NX2] [#6]) , $ ([NX2H])]
Isocyanate	[NX2] = [C] = [O]
Isothiocyanate	[NX2] = [C] = [S]
Ketone	[#6] [CX3] (= O) [#6]
Nitrile	[NX1] # [CX2]
Phenol	[OX2H] [cX3] : [c]
Phosphine	[PX3]
Sulfide	[#16X2H0]
Sulfonamide	[#16X4] ([NX3]) (= [OX1]) (= [OX1]) [#6]
Sulfonate	[#16X4] (= [OX1]) (= [OX1]) ([#6]) [OX2H0]
Sulfone	[#16X4] (= [OX1]) (= [OX1]) ([#6]) [#6]
Sulfonic acid	[#16X4] (= [OX1]) (= [OX1]) ([#6]) [OX2H]
Sulfoxide	[#16X3] = [OX1]
Thial	[CX3H1] (= S) [#6, H]
Thioamide	[NX3] [CX3] = [SX1]
Thiol	[#16X2H]

A.3 Functional Group influence on ^{1}H-NMR spectra

To investigate if the simulated ^{1}H-NMR spectra exhibit the behavior described in the literature of having peaks in the region between 6.0 – 8.7 ppm for aromatic compounds, we divided the spectra into two groups: one containing molecules with aromatic atoms and one with molecules without aromatic atoms. Then, we averaged all the spectra within each group and plotted the averaged spectra with the corresponding standard deviations (see Figure 5). It can be clearly seen that the non-aromatic molecules lack a signal in the literature-reported aromatic range, as expected. Conversely, the averaged spectra of the aromatic molecules show a distinct peak region in the aromatic range. This indicates that the simulations tend to correctly predict the aromatic regions of the ^{1}H-NMR spectra.

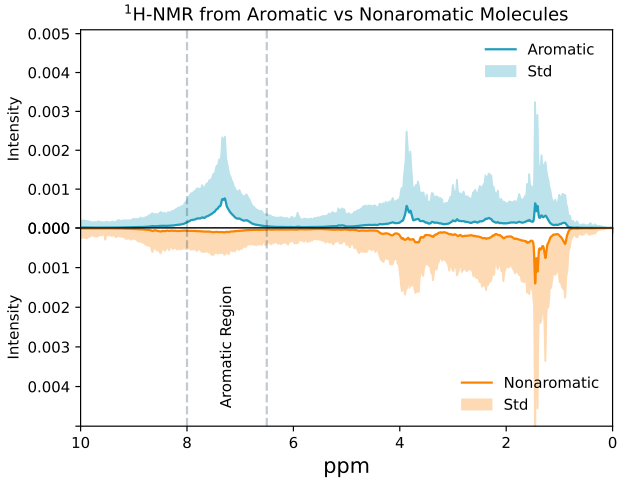


Figure 5: Aromatic molecules ^{1}H-NMR spectra averaged and plotted against the average nonaromatic molecules ^{1}H-NMR spectra

A.4 Metrics

Top–N We define accuracy as the exact character match of the predicted vs the true SMILES string. All SMILES in the dataset are canonicalized and in addition, we canonicalize all predicted SMILES strings after generation. Top–N accuracies are calculated as the fraction of samples in which the correct SMILES string is found within the Top–N predictions.

Cosine Similarity:
The cosine similarity, $S_C(A, B)$, is defined as the cosine of the angle θ between two vectors A and B, representing the continuous spectra. It is calculated as:

$$S_C(A, B) := \cos(\theta) = \frac{A \cdot B}{|A||B|} \tag{1}$$

Chemical Similarity:
The chemical similarity between two molecules is calculated by computing the Tanimoto similarity between their Morgan fingerprints (built using RDKit [71]). The Morgan fingerprints are generated with a length of 1024 and radius of 2. The Tanimoto similarity, $T(A, B)$, between two bit vectors A and B, representing the molecular fingerprints, is defined as:

$$T(A, B) = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B} \tag{2}$$

CosineGreedy: The CosineGreedy class from the MatchMS package [62] calculates a modified version of the cosine similarity between two mass spectra. Instead of treating the spectra as vectors across all possible m/z values, it identifies pairs of peaks between the two spectra that have m/z values within a specified tolerance. It then uses a greedy approach to find the best set of matched peak pairs, rather than solving the optimal assignment problem. The cosine similarity score is calculated based on the intensities of these matched peak pairs, optionally weighted by the m/z values raised to a specified power. It can be adapted to a list of NMR peaks by giving instead of the m/z value the ppm, and instead of the intensity giving the integral value.

A.5 Sim To Real Gap

In the following section, we conduct additional experiments to assess the similarity between the computed and experimental spectra. We use larger databases and also measure the similarity between a simulated spectra and the experimental spectra of the molecule with the closest Tanimoto similarity. Below we will outline the experimental datasets used for each modality followed by the similarities presented in Table 7. We did not simulate additional spectra but used spectra of molecules contained in both our database and the simulated one.

- **IR** (NIST Gas Phase IR Database [51]): 2.375 Spectra
- **MS/MS** (GNFPS as prepared by Goldman et al. [47]): 640 Spectra
- ^{13}**C-NMR** (nmrshiftdb2 [37]): 6.627 Spectra
- ^{1}**H-NMR** (Pistachio [72]): We extract the experimental ^{1}H-NMR from the patent texts of the entries in Pistachio. In total we compare 10.000 ^{1}H-NMR from Pistachio to the text representation (i.e. centroid, integration and type of each peak) of our simulated ^{1}H-NMR spectra

Table 7: Additional similarity experiments: Column *Exp.* compares the similarity of the simulated spectrum to the same experimental one, *Tanimoto* compares the simulated spectrum of the molecule to the experimental spectrum of the molecule with the next closest Tanimoto similarity. *All Others* compares the similarity of a particular simulated spectra to all other experimental spectra.

Modality	Metric	Exp.	Tanimoto	All Others	Avg. Tanimoto Similarity
IR	Cosine Similarity	0.366±0.149	0.195±0.141	0.190±0.113	0.814±0.112
MSMS (CFM-ID, 10 eV)	Greedy Cosine	0.486±0.342	0.083±0.208	0.011±0.009	0.714±0.153
MSMS (Scarf)	Greedy Cosine	0.148±0.154	0.092±0.125	0.043±0.019	0.714±0.153
MSMS (Iceberg)	Greedy Cosine	0.812±0.184	0.215±0.265	0.044±0.019	0.714±0.153
^{1}H-NMR	Greedy Cosine	0.941±0.069	0.826±0.135	0.664±0.094	0.687±0.105
^{13}C-NMR	Greedy Cosine	0.915±0.137	0.534±0.218	0.175±0.051	0.795±0.108

A.6 Model

For all models, we employ the same 90/10 train test split. The seeds are fixed for all runs.

A.6.1 Gradient Boosted Tree

All experiments using gradient boosted trees employ the XGBoost library [73] using the default settings on 32cpu cores:

```
n estimators: 100
base score: 0.5
gamma: 0
learning rate: 0.1
max delta step: 0
max depth: 10
```

A.6.2 1D-CNN Model

We use the implementation by Jung et al. [41] for all 1D-CNN experiments and reuse the same hyperparameters. This 1D-CNN employs three convolutional followed by three fully connected layers. The model is trained with an Adam optimiser for 41 epochs on A100 GPU using Keras [74].

A.6.3 Transformer Model

We employ a vanilla encoder-decoder transformer as implemented in the OpenNMT-py library [75, 65] with four layers each for the encoder and decoder and a hidden dimension of 512. We train all transformer models for 250k steps amounting to approximately 35h on a A100 GPU. Further hyperparameters can be found below:

```
layers: 4
heads: 8
word_vec_size: 512
hidden_size: 512
transformer_ff: 2048
optim: adam
adam_beta1: 0.9
```

```
adam_beta2: 0.998
decay_method: noam
learning_rate: 2.0
batch_size: 4096
activation function: ReLu
dropout: 0.1
```

A.7 Evaluation metrics

F1 Score :
The F1 score is a measure of a model's performance that combines precision and recall into a single score. It is calculated as the harmonic mean of precision and recall, given by:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

where precision is the fraction of true positives among the predicted positives, and recall is the fraction of true positives among the actual positives. The F1 score ranges from 0 to 1, with higher values indicating better performance.

Top-k Accuracy:
Top-k Accuracy measures the fraction of instances where the correct answer is among the top-k ranked predictions made by the model.

$$\text{Top-}k \text{ Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(y_i \in \hat{y}_i^{(k)}\right) \tag{4}$$

N:Total number of samples,y_i True label for the i-th sample $\hat{y}_i^{(k)}$Set of the top k predicted labels for the i-th sample $\mathbb{1}(\cdot)$ Indicator function, returns 1 if $y_i \in \hat{y}_i^{(k)}$ otherwise 0

A.8 Representations for Transformers

A.8.1 Molecules

All molecules in this study are treated as SMILES, a string based molecular representation. We tokenize the molecules following Schwaller et al. [32] using the following regex:

```
(\[[^\]]+]|Br?|Cl?|N|O|S|P|F|I|b|c|n|o|s|p|\(|\)|\.|=|#|\|\+|\\\\\/|:||@|
\?|>|\*|\$|\%[0-9]{2}|[0-9])
```

A.8.2 Infrared Spectra Representation

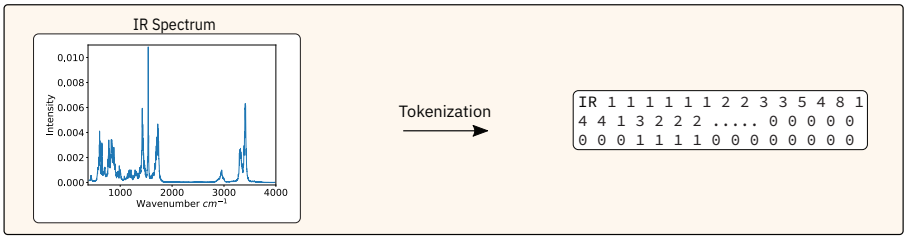


Figure 6: Tokenization of IR spectra.

The IR spectra generated via molecular dynamics range from 400 to 4000cm^{-1} with a resolution of 2cm^{-1}, i.e. it is a vector of size 1800. To convert this vector into a structured text representation ingestible by a transformer model we first downsample the spectrum to a vector of size 400 via linear interpolation. We subsequently scale the spectrum to a range of 0 to 100 and round the values to integers. In effect this converts a vector of size 1800 to a string of 400 integers. An example is shown in Figure 6.

A.8.3 NMR Spectra

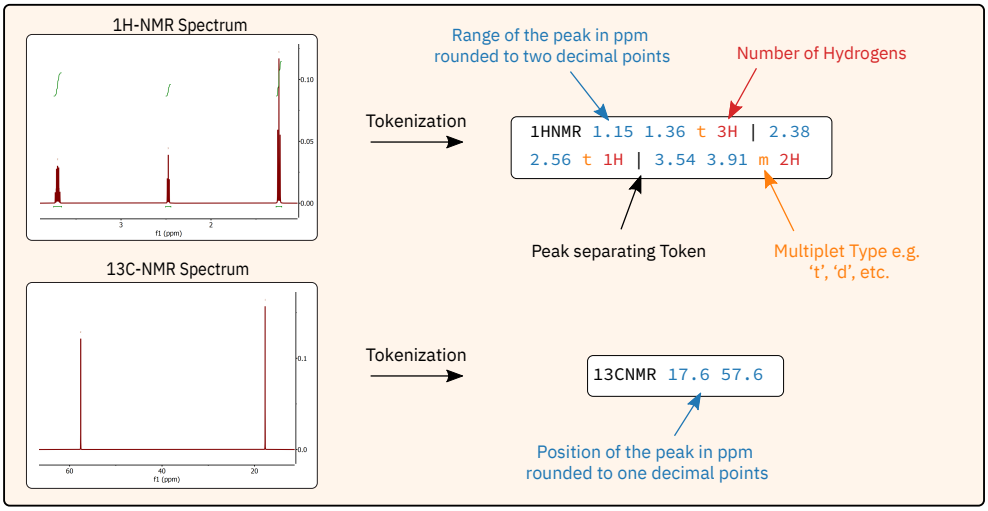


Figure 7: Summary of the tokenization process for NMR spectra. Top: Tokenization of an ^{1}H-NMR spectrum following the Range representation. Bottom: Tokenization of a ^{13}C-NMR spectrum

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`".

Both methods are illustrated in Figure 7. In case both the ^{1}H- and ^{13}C-NMR are used as input both are concatenated.

A.8.4 MS/MS Spectra

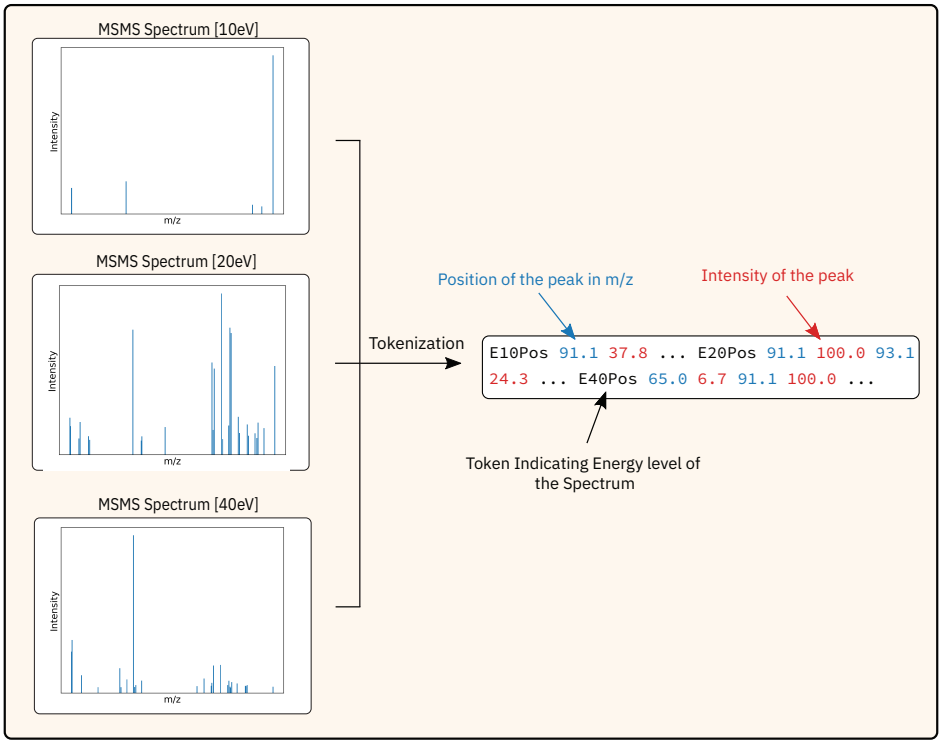


Figure 8: Tokenization of the MS/MS spectra: We include the position and intensity of each peak in the structured text represenation and concatenate the spectra resultant from the different ionisation energies.

The generated MS/MS consist of stick spectra, i.e. a list of the position in m/z and the intensity of each peak. We convert this stick spectrum into a structured text representation by listing each peak followed by it's intensity. Both the m/z and intensity are rounded to one decimal point. We concatenate the MS/MS generated at the three different ionisation energies as shown in Figure 8.

A.9 Zero Shot on van Bramer et. al.

We evaluate the zero-shot performance of the models trained to predict the structure from the spectra on the experimental van Bramer dataset. The results are shown in Table 8 below:

Table 8: Zero shot accuracy (Top–1, Top–5, and Top–10) of transformer models trained on simulated data to predict the chemical structure from different spectra evaluated on the experimental van Bramer et. al. dataset.

	Top–1%	Top–5%	Top–10%
IR	5.03	11.96	17.61
MS/MS (Negative, CFM-ID)	0.0	0.0	0.0
MS/MS (Positive, CFM-ID)	0.0	0.0	0.0
MS/MS (Positive, Scarf)	8.81	25.79	36.47
MS/MS (Positive, Iceberg)	5.66	11.94	13.83
^{13}C-NMR	58.49	69.81	72.96
^{1}H-NMR	42.77	55.35	59.75
^{1}H-NMR + ^{13}C-NMR	50.31	63.98	67.08

A.10 Compute Resources used for simulating spectra

In the following, we outline the compute resources used to generate the dataset. All simulations were run on AMD EPYC 7452 CPUs.

- **IR**: 500 CPU cores with 1TB RAM for ca. 46 days
- **^{1}H-NMR:** 200 CPU cores with 400GB RAM for ca. 15 days.
- **^{13}C-NMR:** 200 CPU cores with 400GB RAM for ca. 14 days.
- **MSMS-CFM ID:** 80 CPU cores with 160GB RAM for ca. 7 days.
- **MSMS-Iceberg:** 16 CPU cores with 32GB RAM for ca. 3 days.
- **MSMS-Scarf:** 16 CPU with 32GB RAM cores for ca. 1.5 days

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [NA]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [NA]
   (b) Did you include complete proofs of all theoretical results? [NA]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [NA] : The USPTO dataset has "CC0" license
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [NA]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [NA]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [NA]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [NA]