Enhancing Molecular Structure Elucidation: MultiModalTransformer for both simulated and experimental spectra

Martin Priessner,*[a] Richard J. Lewis,[b] Isak Lemurell,[a] Magnus J. Johansson,[a] Jonathan M. Goodman,[c] Jon Paul Janet,[d] Anna Tomberg*[a]

[a] Dr. M. Priessner (ORCID: 0009-0005-7349-283X), Dr. A. Tomberg (ORCID: 0000-0002-0718-5381), Prof. M. J. Johansson (ORCID: 0000-0001-8811-2629)
Medicinal Chemistry, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), BioPharmaceuticals R&D
AstraZeneca
Pepparedsleden 1, 43183 Mölndal (Sweden)
E-mail: martin.priessner@astrazeneca.com, anna.tomberg@astrazeneca.com

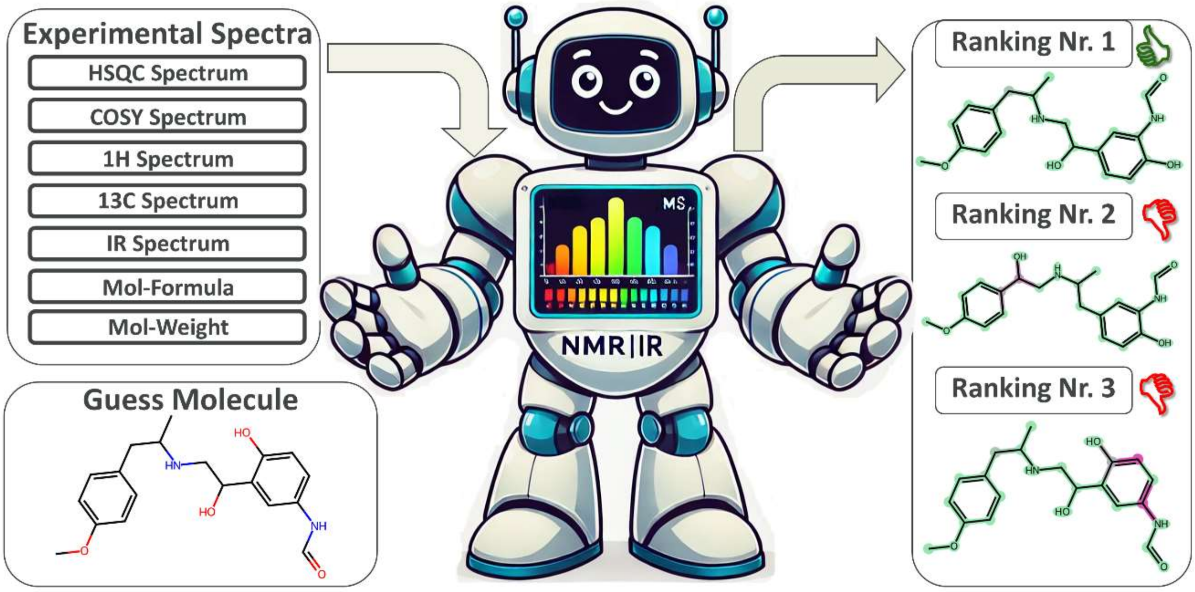
[b] Dr. R.J. Lewis (ORCID: 0000-0001-9404-8520)
Department of Medicinal Chemistry, Research & Early Development, Respiratory & Immunology, BioPharmaceuticals R&D
AstraZeneca
Pepparedsleden 1, 43183 Mölndal (Sweden)

[c] Prof. J.M. Goodman (ORCID: 0000-0002-8693-9136)
Centre for Molecular Informatics, Yusuf Hamied Department of Chemistry
University of Cambridge
Lensfield Road, Cambridge CB2 1EW, (UK)

[d] Dr. J.P. Janet (ORCID: 0000-0001-7825-4797)
Molecular AI, Discovery Sciences, R&D
AstraZeneca
Pepparedsleden 1, 43183 Mölndal (Sweden)

We present MultiModalTransformer (MMT), a novel deep learning architecture that directly predicts molecular structures from diverse spectroscopic data (^{1}H-NMR, ^{13}C-NMR, HSQC, COSY, IR, and mass spectrometry (MS). Utilizing a modified Transformer model with attention mechanisms, the MMT simultaneously processes multiple data modalities to focus on the most relevant spectral features. Our approach demonstrates significant advancements in automated structure determination, achieving up to 94% correct identifications for real experimental samples despite being trained solely on simulated spectra. To address the challenges of vast chemical space and limited experimental data we introduce an innovative improvement cycle that allows MMT to adapt to new chemical spaces. The model's robustness is evidenced by its ability to maintain substantial predictive power even when starting with slightly incorrect molecular structures, identifying 56% of experimental molecules correctly from modified initial guesses. MMT provides explainable predictions through token-based analysis, offering insights into its decision-making process. We also present a user-friendly GUI that integrates the full improvement cycle workflow, facilitating practical application in chemistry laboratories. By leveraging diverse spectral inputs and adaptive learning techniques, MMT represents a significant step towards fully automated structure elucidation, potentially accelerating drug discovery and natural product research while demonstrating that comprehensive chemical space coverage in training data is more critical than precise spectral accuracy.

Keywords: MultiModalTransformer, NMR, IR, MS, Structure Elucidation, Machine Learning


Experimental Spectra
HSQC Spectrum
COSY Spectrum
1H Spectrum
13C Spectrum
IR Spectrum
Mol-Formula
Mol-Weight
Guess Molecule
MS
NMR|IR
Ranking Nr. 1
Ranking Nr. 2
Ranking Nr. 3

Introduction

When we think about the future chemistry laboratory, we often envision a fully automated system: scientists enter a molecule they want to create, advanced software suggests a synthetic route, which is then executed by a robot. This is followed by a purification step to prepare the sample for analysis by a range of analytical instruments. Finally, an automated process determines the structure of the molecule based on the spectral data collected. Achieving this vision requires advancements in multiple areas, from route predictions to synthesis and purification automation. A critical component of this pipeline is the automated interpretation of spectroscopic data for molecular structure elucidation. In this work, we describe our contribution to solve automatic structure elucidation by proposing a flexible model that can translate spectroscopic data directly into molecules, while addressing several limitations of current Computer-Assisted Structure Elucidation (CASE) programs.

CASE programs use Nuclear Magnetic Resonance (NMR) and/or Infrared (IR) data to elucidate molecular structures. H and ^{13}C NMR are the most widely used spectroscopic techniques for characterizing molecules and provide detailed insights into hydrogen and carbon atoms. Furthermore, 2D techniques such as COSY (Correlation Spectroscopy) and HSQC (Heteronuclear Single Quantum Coherence) provide connectivity information.[1] Mass spectrometry (MS) offers molecular weight and formula information, crucial for confirming molecular structures.[1] IR spectroscopy is cost-effective and non-destructive allowing for quick identification of functional groups. The fingerprint region (400–1500 cm^{-1}) of IR spectra contains complex, molecule-specific absorption patterns that are challenging to interpret using traditional methods. This region holds detailed molecular information often underutilized in traditional analysis due to its complexity. However, modern machine learning algorithms show promise in extracting and interpreting this rich structural information from the fingerprint region.[2–4] Together, NMR, IR, and MS provide complementary information essential for structure elucidation.

Classic CASE software packages, such as ACD/Structure Elucidator (ACD Labs) and Mnova Structure Elucidator (Mestrelab Research), have been developed to aid in this process. Typically, the user inputs processed spectral data, including peak assignments and correlations, and the software generates candidate structures that fit the data. These programs often employ spectral comparisons as part of their structure elucidation process, comparing input spectra against databases of known compounds or simulated spectra to build up structural fragments. However, these programs often require substantial human input, especially in isolating relevant peaks within NMR spectra.[5–7] Additionally, they rely on databases of chemical structures and spectra, which can fail when experimental conditions alter spectra[8] or when the database does not cover the required chemical space.[7–10]

Spectral comparison methods, such as the Goodman DP4 method and its variants, along with newer neural network-based and grid-based approaches, have also emerged as valuable tools for verifying molecular structures among different options.[11–27] While these methods are not standalone structure elucidation techniques, they play a crucial role in structure verification and can be integrated into broader elucidation workflows. However, they often require an initial suggestion of the target molecule and may not be practical for de novo structure elucidation.

Other methodologies have been proposed like those by Pesek et al. that integrate data from IR, ^{1}H and ^{13}C NMR, and mass spectra to build molecular structures, simulating the process a spectroscopist would follow.[28] More recently, some models have been developed to process IR or ^{1}H and ^{13}C NMR spectra, transforming spectral data into tokenized text formats to predict molecular structures as SMILES.[29,30] Other frameworks assess structural connectivity by processing ^{1}H and ^{13}C NMR spectra, predicting substructures and assembling candidate isomers with probabilistic rankings.[31] Additionally, DeepSAT, a CNN-based system, uses HSQC spectrum data for scaffold prediction.[32] The NMR-TS method combines machine learning and density functional theory to automate molecule identification from NMR spectra.[33] However, this neural network generates candidate structures without directly considering the spectra, relying on chance to predict the correct molecules. In IR spectroscopy, advancements in deep and convolutional neural networks now enable functional group identification from FTIR spectra without relying on databases or rule-based methods.[34,35]

However, these current approaches still face limitations, such as

1) applicability domain & reliance on extensive databases
2) need for suggestion of target molecule
3) limited consideration of multiple data modalities

Our approach addresses these limitations by proposing an automated pipeline from spectra to molecular structure. This pipeline leverages the Transformer neural network architecture[36] that can simultaneously process multiple spectroscopic data types (^{1}H-NMR, ^{13}C-NMR, HSQC, COSY, IR, and MS). The Transformer's attention mechanism allows it to focus on the most relevant spectral features across different data types, enabling it to learn complex relationships between spectral inputs and molecular structures.

Furthermore, we introduce an innovative improvement cycle that allows the model to adapt to unseen chemical spaces. This iterative process enhances the model's ability to predict structures in novel domains, effectively expanding its applicability. Importantly, this improvement cycle enables our model to solve real experimental spectra despite being trained initially on simulated spectra. This capability demonstrates the model's robustness and potential for practical applications in real-world structure elucidation tasks.

Methodology

Spectral Data Generation and Preprocessing

This study utilizes simulated spectroscopic data across multiple modalities: ^{1}H NMR, ^{13}C NMR, COSY, HSQC, IR spectra and mass spectrometry (MS) information. The ^{1}H and ^{13}C chemical shifts are generated using the Scalable Graph Neural Network (SGNN).[37] These shifts serve as the basis for both 1D NMR information and the generation of 2D spectra. Two-dimensional NMR spectra (COSY and HSQC) were reconstructed from the SGNN-predicted ^{1}H and ^{13}C shifts. For HSQC, we employed a reconstruction logic validated against state-of-the-art simulation software.[14] A similar rule-based approach was implemented for COSY spectra, accounting for molecular structure and H-H coupling patterns. To simulate ^{1}H peak splitting patterns, we developed a rule-based algorithm considering J-couplings, producing spectra that mimic real-world ^{1}H NMR data. ^{13}C shifts were presented without intensity values, consistent with experimental practices. IR spectra were simulated using ChemProp-IR,[38] a directed message passing neural network. MS information, specifically molecular weights and formulas was derived from the SMILES structures using RDKit,[39] simulating the output of high-resolution LCMS measurements. While exact LCMS data was not directly processed in our workflow, our approach assumes accurate molecular weight and formula determination from MS analysis. For data handling, spectral data for each modality (except IR) is stored in CSV files containing SMILES strings, unique sample identifiers, and corresponding spectral information. IR spectral data is stored individually, with file names serving as molecular identifiers. In preprocessing, chemical shift data for ^{1}H and ^{13}C NMR spectra are normalized by factors of 10 and 200, respectively, with this normalization also applied to the relevant dimensions in the HSQC and COSY spectra. IR spectra are down sampled to 1,000 data points along the frequency dimension. Prior to training, all data is consolidated into a single .pkl file, with spectra, SMILES strings, and sample identifiers stored in a dictionary format, ensuring efficient data retrieval during the training process. Detailed information on the reconstruction logic, simulation of coupling constants and splitting patterns, and limitations of the spectral simulations can be found in **Supplementary Information Section 1**.

Generation of Synthetic Datasets

The initial training dataset was created by randomly selecting approximately 5 million molecules from the ZINC270 database (accessed via DeepChem Python package, https://chembl.gitbook.io/chembl-interface-documentation/downloads), within the 250-350 Dalton molecular weight range. This dataset was split into training and testing subsets at a 9:1 ratio. To evaluate model generalizability, a secondary dataset of 1.5 million molecules (0-500 Daltons) was extracted from PubChem. Molecules containing elements Se, Sn, As, Ge, Te, Al, Hg, Ga, Sb, Pb, Tl, Bi, Ti, Li, Zn, Na, and Pd were excluded to maintain consistency with the ZINC dataset composition. To ensure dataset uniqueness and prevent overlap, we compared the canonicalized SMILES representations of molecules from both ZINC and PubChem datasets, removing any duplicates.

The MMT model was first trained and tested on the ZINC dataset to establish baseline performance. Further validation utilized the PubChem dataset, for exploring the iterative improvement cycle. The PubChem dataset was stratified into three molecular weight ranges for comprehensive out-of-distribution testing:

1. *0-250 Daltons*: Testing lower range adaptability.
2. *250-350 Daltons*: Control group within similar weight range, but with diverse chemical structures.
3. *350-500 Daltons*: Assessing performance on larger molecules beyond the ZINC dataset's upper limit.

https://doi.org/10.26434/chemrxiv-2024-zmmnw **ORCID:** https://orcid.org/0009-0005-7349-283X Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

This stratified approach allowed for the assessment of the MMT model performance of previously unseen chemical spaces.

Neural Network Architecture and Verification Logic

The MMT model has a modified Transformer architecture,[36] that processes multiple spectroscopic inputs simultaneously. The structure verification pipeline consists of:

1. Modality-specific point/spectrum embedding layers
2. Transformer encoder cascade
3. Transformer decoder
4. HSQC & COSY matching for error analysis

Embedding Layers: Raw peaks from each modality (^{1}H, ^{13}C, HSQC, COSY, and IR) are transformed into a 128-dimensional latent space. NMR spectra are normalized (^{1}H dimensions divided by 10, ^{13}C by 200) and zero-padded to 64 inputs. ^{1}H NMR data is provided as 2D input (chemical shift and intensity), while HSQC and COSY data are provided as 2D inputs (x and y chemical shift coordinates without intensity). ^{13}C NMR data is provided as 1D input, using only chemical shifts without intensity information. IR spectra (400-4000 cm^{-1}) are discretized and down sampled to a total of 1000 datapoints. These datapoints are then embedded to match the 128-dimension input expected by the model. After embedding of each spectrum, a rectified linear unit (ReLU) activation is applied, and a mask is used to identify actual data points. Finally molecular weight and formula embeddings are concatenated to each spectrum embedding to be considered in each encoder type.
Encoder: Individual spectrum encoders, each comprising 6 encoder blocks with 16 attention heads and a 4x forward expansion, process the embeddings. The resulting feature vectors are then concatenated and fed into a cross-modality encoder with an identical structure. This cross-modality encoder employs self-attention mechanisms to mix and correlate information from different spectral inputs, allowing the model to identify inter-spectral relationships.
Decoder: A 6-block and 16-attention head decoder processes the encoder output along with target SMILES strings.
HSQC & COSY Matching: The HungDist-NN algorithm[14] is used for HSQC & COSY peak matching. It compares the HSQC & COSY spectra of generated analogs against the target molecules' spectra. In this process, each generated molecule's spectrum is compared to the input spectrum, and molecules are ranked according to their individual HSQC and COSY errors. These rankings are then combined to produce a final overall ranking. The analog with the lowest spectral discrepancy, as determined by the best combined ranking, is identified as the most likely correct structure.

Two illustrations of the full workflow are provided in the supplementary materials: **Supplementary Figure *1*** presents a general overview, while **Supplementary Figure 2** offers a more detailed illustration of the data flow through the model architecture.

Model Scaling and Data Volume Analysis

To optimize the performance and efficiency of our MMT model, we conducted a systematic study exploring various model configurations and training dataset sizes. This analysis was crucial for determining the most effective architecture and data volume for our specific task of molecular structure prediction from spectroscopic data. We conducted a systematic study exploring three model configurations:

1. *Small*: 2 encoder and decoder layers, 4 attention heads
2. *Medium*: 3 encoder and decoder layers, 8 attention heads

3. *Large*: 6 encoder and decoder layers, 16 attention heads

Each configuration was trained on spectral data from 1 million molecules from the ZINC dataset for 20 epochs, using Cross Entropy loss for SMILES prediction which allowed a direct comparison of architectural complexity under controlled conditions.
We also investigated the impact of training dataset size on model performance. Datasets of 100k, 1M, 2M, and 4M molecules were prepared, with training epochs adjusted to maintain a constant exposure of 20 million molecules:

- 4M molecules: 5 epochs
- 2M molecules: 10 epochs
- 1M molecules: 20 epochs
- 100k molecules: 200 epochs

This normalization allowed evaluation of larger datasets' intrinsic benefits on model accuracy and generalizability, independent of epoch count. These experiments systematically examined the effects of both model scale and training data size on the overall performance of the MMT.

Training Stages

The optimized MMT was designed to process ^{1}H, ^{13}C, HSQC, COSY, IR spectra, and mass spectrometry (MS) information. For MS, we utilized molecular weight data calculated from the SMILES using RDKit. The model underwent three progressive training stages:

1. *Initial Training on SMILES Loss*: The network was trained to generate SMILES strings from spectral data using teacher forcing. Loss was calculated using PyTorch's CrossEntropyLoss function, evaluating predictions against actual SMILES tokens.
2. *Integration of Molecular Weight Loss*: This stage refined the model's performance by incorporating molecular weight comparisons, leveraging the MS information. The model generates SMILES strings, from which molecular weights are calculated and compared to the target molecular weights using PyTorch's MSELoss function. This loss was normalized via min-max scaling to align with SMILES loss metrics. The molecular weight loss contribution was gradually integrated into the total loss, increasing by 1% every 50,000 steps. This gradual integration allowed the model to smoothly adapt to the new loss component without disrupting the learning process for SMILES prediction.
3. *Training with Spectral Data Dropout*: The model was trained under conditions of random spectral data omission to enhance resilience and performance in data-constrained environments. Each spectrum had a 50% chance of being omitted and replaced with zero-padded data to maintain input consistency. This stage continued to utilize SMILES and molecular weight losses.

Additionally, to quantify each modality's contribution to the predictive accuracy, we performed an ablation study. The fully trained MMT model underwent single epoch fine-tuning iterations, each omitting one spectral modality. Performance comparisons between these ablated models and the complete MMT model elucidated the relative importance of individual spectral data types in molecular structure prediction.

Training Configuration

The dataset underwent a 9:1 train-test split, with the training data further divided 9:1 to create a validation set. Training was executed in multiple five-epoch stages using four Nvidia V100 GPUs, with a batch size of 256 molecules (64 per GPU), spanning approximately seven days per stage. The PyTorch

Lightning framework was utilized for efficient multi-GPU training management. The AdamW optimizer was employed with an initial learning rate of $1e^{-4}$. Learning rate adjustment was managed by the ReduceLROnPlateau scheduler, which reduced the rate by a factor of 0.5 following two consecutive epochs without loss improvement. For the improvement cycle fine-tuning, the learning rate was adjusted to $2e^{-4}$ when using the simulated PubChem dataset. When working with the ACD Labs generated data and the experimental dataset, the learning rate was further increased to $3e^{-4}$, a 50% spectral dropout was applied, and the molecular weight loss contribution was set to 100%. These modifications in learning rate, dropout, and loss contribution facilitated more effective fine-tuning on specific datasets.

Evaluation Metrics

We established specific criteria to assess the performance of our molecular generation models across the different configurations:

Correct SMILES Sample Probability: This primary metric quantifies the likelihood of generating the correct molecular structure during inference. It is calculated as the multiplicative probability of selecting the correct next token in the SMILES sequence under teacher forcing conditions.

Tanimoto Similarity: Calculated using RDKit for valid molecules generated via greedy sampling, employing 1024-bit fingerprints.

Validity of Generated Molecules: We employ the RDKit library to assess the validity of generated molecules. A SMILES string is considered valid if RDKit successfully constructs a molecule object.

These metrics were systematically applied across all model configurations to enable quantitative performance comparisons.

Sequence Generation Methods

Two sampling methods are employed for generating SMILES strings from the MMT's learned representations.

Greedy Sampling: This deterministic approach selects the most probable token at each step of the sequence generation. It builds the SMILES string by appending the highest probability token until reaching the end token or maximum sequence length. While efficient, this method may lack diversity in generated sequences.

Multinomial Sampling: This stochastic method samples from the entire probability distribution at each step, allowing exploration of a broader chemical space. It utilizes a temperature parameter to adjust the sharpness of the sampling distribution. Lower temperatures yield results closer to greedy sampling, while higher temperatures enhance molecule diversity.

The multinomial sampling process incorporates an adaptive mechanism to generate a specified number of unique molecules. It employs an iterative approach that dynamically adjusts the temperature parameter. Starting from an initial value of 1, the temperature is incrementally increased by 0.1 for each iteration if the desired number of unique molecules is not achieved. This process continues until either the target number of unique molecules is generated, the temperature reaches a maximum value of 3, or the iteration count hits a limit of 500. This adaptive strategy ensures a balance between diversity and computational efficiency, while guaranteeing the termination of the sampling process.

Improvement Cycle Methodology and Cross-Dataset Evaluation

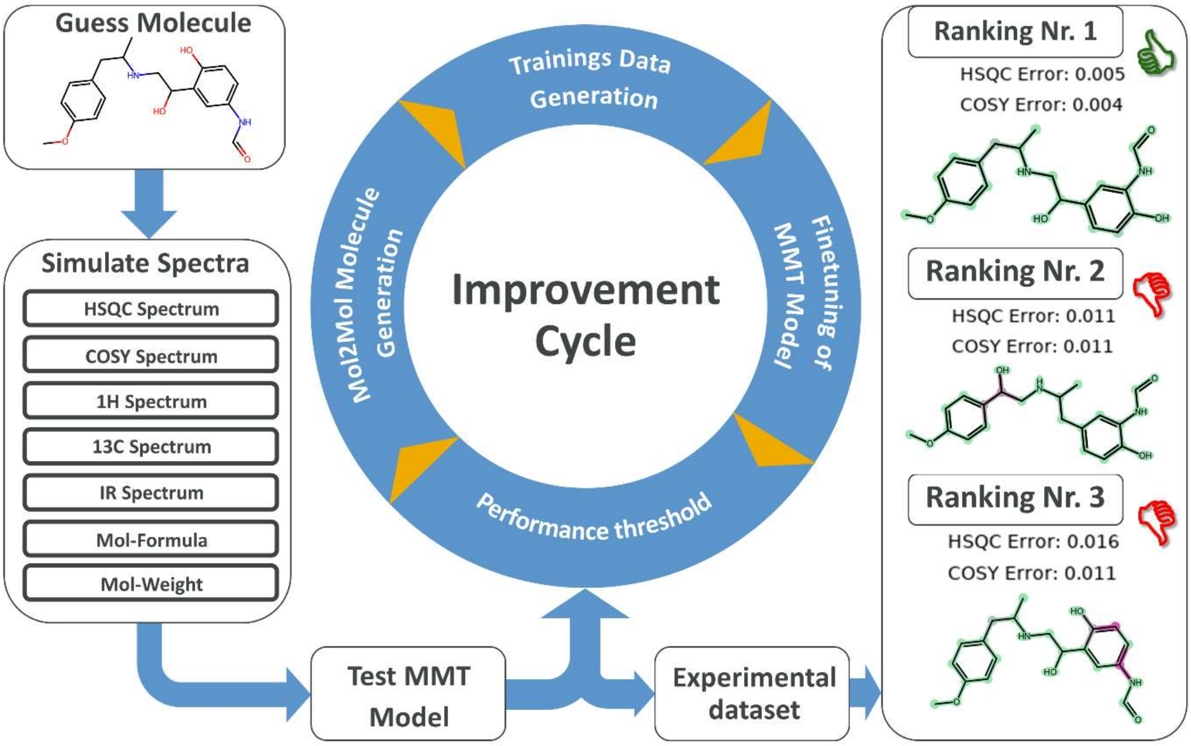


Figure 1: Schematic representation of the Improvement Cycle for enhancing the MMT model. It shows the key steps from spectrum selection, testing the MMT model, then performing the improvement cycle with the data generation and fine-tuning pipeline to running MMT on experimental data and ranking of generated molecular structures based on HSQC and COSY error ranking.

The improvement cycle is a crucial component of our approach, designed to enhance the MMT model's adaptability to new chemical spaces. By iteratively generating analogs, simulating their spectra, and fine-tuning the model, we aim to improve its performance on previously unseen molecular structures. This process is particularly important for bridging the gap between simulated training data and real-world applications, where novel molecular structures are frequently encountered.

The process was initially applied to 100 molecules from the ZINC test dataset and subsequently extended to 100 molecules from the PubChem dataset. The methodology comprised:

Analog Generation: Using the Mol2Mol model,[40–43] we generated 10, 30, 50, or 100 analogs for each target molecule, exploring related chemical spaces without producing exact duplicates.
Spectral Simulation: NMR and IR spectra were simulated for all generated analogs.
Fine-tuning: The MMT model underwent fine-tuning on these datasets for 50 epochs to enhance performance on the targeted chemical space.

The improvement cycle allows the model to adapt to new chemical spaces by incorporating structurally similar molecules into the training process. By fine-tuning the model on these augmented datasets, we enhance its ability to predict structures for novel compounds not represented in the original training data.

The performance quantification utilized averaged correct sample probability and averaged Tanimoto similarity (via greedy sampling). Post-fine-tuning, multinomial sampling generated three candidates per target molecule, from which the highest Tanimoto similarity was used for assessment.

The PubChem evaluation expanded testing to three molecular weight ranges: 0-250, 250-350, and 350-500 Daltons, with 100 molecules in each category. This allowed assessment of model performance on molecular weights varying from the initial training data. The same improvement cycle was applied to the PubChem dataset. Additionally, we fine-tuned the ZINC-trained MMT model on the entire PubChem training set for 18 epochs, comparing its performance to that achieved through targeted improvement cycles on smaller subsets. A second improvement iteration was conducted using the checkpoint from the smallest dataset (10 analogs) testing for the iterative improvement capabilities of the model. The fine-tuning learning rate was increased to 2e^{-4} for this iteration, applied across all molecular weight settings in the PubChem dataset.

Mol2Mol Molecule Augmentation

The Mol2Mol[40–43] is a sequence to sequence transformer that produces close analogues to input molecules. This approach is essential for expanding the training data into unexplored chemical spaces, enhancing the model's ability to learn and predict in these regions. The process begins with SMILES standardization using RDKit, converting molecules to a canonical form and extracting molecular scaffolds. New molecules are then generated by modifying side chains or functional groups while retaining the core scaffold structure. To ensure practical relevance and drug-likeness of the synthesized molecules, several constraints are imposed: generation within a range of ± 100 Daltons from the template's molecular weight, a modified Lipinski's Rule of Five allowing molecules up to 550 Daltons, and a configurable Tanimoto similarity filter (default 0.3) to maintain a desired level of similarity to the target molecule. The system incorporates a scaffold hopping mechanism that shifts to a new molecular scaffold if the current one fails to produce viable candidates after a set number of iterations or if too many molecules with the same scaffold have been generated.

Experimentation with Simulated and Experimental Data Samples

We designed an experiment to evaluate our models' performance across various data types, progressing from our own simulations to ACD Labs simulations and finally to real experimental data. We curated a set of 34 diverse in-house collected and publicly available molecules (see **Supplementary Figure 3**) with all experimental spectral modalities (^{1}H-NMR, ^{13}C-NMR, HSQC, COSY, IR, and MS) available. To maintain a focused fine-tuning process, we handled each molecule individually in the improvement cycle and combined the sampling performance of 3 individual runs. This individual approach was chosen to allow the model to adapt more precisely to the specific chemical features of each molecule, potentially improving its performance on challenging or unique structures at the expense of broader generalization. We first established a baseline performance using our pretrained MMT model on our simulated data. For each molecule, we conduct the improvement cycle, generating 50 analogues per target molecule (±100 Da range, max. 50 per scaffold), fine-tuning for 15 epochs with a learning rate of 3e^{-4}, employing a loss function with fixed molecular weight contribution of 100% and 50% spectral dropout. We reassessed performance on our simulated data, then tested on ACD Labs simulated spectra, which provided an intermediate challenge due to different underlying algorithms and error profiles. For IR simulations, we used ChemProp-IR in both our simulation pipeline and for the ACD Labs data. Finally, we evaluated the models on manually curated experimental data. Throughout all phases, we used multinomial sampling (rate 3x20) and our HSQC/COSY matching logic for molecule identification. This experiment allows us to assess the model's performance across different data sources and levels of complexity. By progressively challenging the model with data that increasingly deviates from the training simulations, we evaluate its robustness and adaptability to real-world applications.

Robustness Assessment of the Improvement Cycle

To evaluate the robustness of our improvement cycle approach, we conducted an additional experiment simulating potential errors in initial structure assumptions. We tested the improvement cycle using slightly modified versions of the actual target molecules as starting points (shown in **Supplementary Figure 4**). This scenario mimics real-world situations where chemists might begin with an incorrect assumption about the target molecule's structure, which is common in structure elucidation tasks. For instance, a chemist might misinterpret initial spectral data, leading to an incorrect initial structure guess. By demonstrating that our model can overcome these initial inaccuracies, we show its potential to assist in real-world structure elucidation tasks where the exact structure is unknown and initial hypotheses may be flawed. We applied this modified approach to simulated, ACD, and experimental spectra, assessing the model's ability to overcome initial structural inaccuracies and still often predict the correct molecular structure.

Results and Discussion

Base Model Architecture and Optimization

Our initial experiments focused on optimizing the MMT model, which processes various spectral data types including NMR (^{1}H, ^{13}C, HSQC, and COSY), IR, and MS. We evaluated different model configurations and training strategies using metrics such as *SMILES prediction accuracy*, *structural similarity*, and *SMILES validity* of generated molecules. Larger models and datasets consistently improved performance across all metrics, leading us to select the largest model configuration and a 4 million molecule dataset for further analysis. We implemented a three-stage training strategy, progressively incorporating SMILES prediction, molecular weight loss, and spectral data dropout, which enhanced the model's overall performance. For more detailed information on the optimization process and experimental results, please refer to **Supplementary Information Section 2**. Additionally, we evaluated the model's molecule identification accuracy using HSQC spectral matching, achieving an 89.9% accuracy with multinomial sampling, significantly outperforming greedy sampling which achieved 50.0% accuracy. These optimizations establish a robust foundation for the MMT model's application in molecular structure elucidation from spectral data. More detailed information on this experiment can be found in **Supplementary Information Section 3**.

Impact of Spectral Modalities on Model Performance

To understand the relative importance of different spectral data types on model performance, we conducted an ablation study by omitting each spectral modality in turn during single-epoch fine-tuning iterations of the fully trained MMT model. We evaluated the impact on three key metrics: averaged correct SMILES sample probability, average greedy sampled Tanimoto similarity, and number of invalid molecules generated. The results, illustrated in **Figure 2**, reveal that 2D NMR data (HSQC and COSY) contribute most significantly to the model's performance. Omitting HSQC data led to the most substantial drops in correct SMILES probability (0.51 to 0.04) and Tanimoto similarity (0.82 to 0.43), while also resulting in the highest number of invalid molecules (44,847). COSY omission showed the second-largest impact, particularly evident in the notable increase of invalid molecules (38,798). It's important to note that the model's interpretation of spectral importance may differ from that of human spectroscopists. Each spectral embedding in our model includes molecular weight and formula information, enhancing its information content beyond what's visually apparent in the spectrum alone. This additional context influences the model's prioritization of different modalities. Interestingly, while ^{13}C NMR showed less impact on model performance, its time-consuming acquisition process in practice might make it a candidate for deprioritization in time-sensitive scenarios. Conversely, IR spectroscopy, despite showing minimal effect on performance in this study, offers rapid data collection, potentially making it valuable in practical applications where speed is crucial.

Interestingly, while 2D NMR techniques (HSQC and COSY) showed the most significant impact on model performance, the time-consuming acquisition process of ^{13}C NMR in practice might make it a candidate for deprioritization in time-sensitive scenarios, given its relatively lower impact on model performance. Conversely, IR spectroscopy, despite showing minimal effect on performance in this study, offers rapid data collection, potentially making it valuable in practical applications where speed is crucial. The substantial influence of 2D NMR data suggests that prioritizing HSQC and COSY spectra acquisition could significantly enhance structure elucidation accuracy, especially when balanced against time and resource constraints.

These findings can guide spectroscopists in optimizing data collection strategies, potentially reducing experimental time and costs while maintaining high accuracy. However, it's crucial to balance these machine learning-derived insights with traditional human-centered elucidation approaches, emphasizing the complementary nature of AI and human expertise in structural analysis tasks.

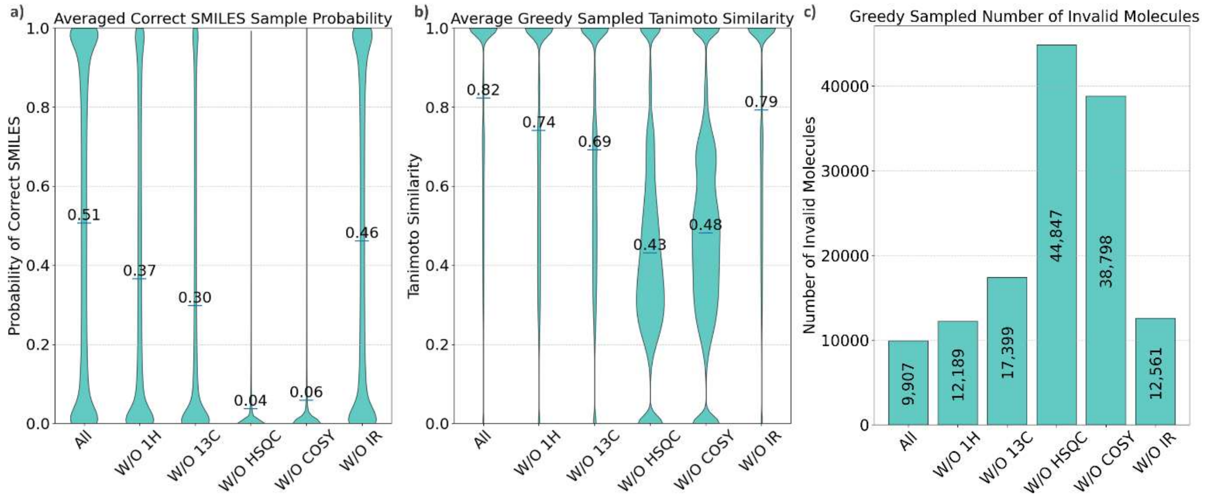


Figure 2*: Ablation Study Results for MMT model.* Impacts of omitting individual spectral modalities on (a) averaged correct SMILES sample probability, (b) average greedy sampled Tanimoto similarity, and (c) number of invalid molecules generated. Results highlight the critical importance of 2D NMR data (HSQC and COSY) for model performance and structural accuracy.

Improvement Cycle Evaluation

To address the challenge of the vast chemical space that no model can be trained on entirely, we integrated an improvement cycle that activates when the model encounters an unfamiliar region not covered in the training data. This cycle employs a generative model designed to suggest structurally similar molecules within the unexplored chemical space, allowing for the creation of a fine-tuned dataset tailored to these novel regions. Coupled with this process is a data generation pipeline, which includes the SGNN network[37] for generating ^{1}H and ^{13}C NMR spectra and rule-based algorithms for reconstructing HSQC and COSY spectra, as well as for calculating coupling constants in ^{1}H NMR spectra. Furthermore, IR spectra are generated using a message-passing neural network.[38] For mass spectrometry (MS) data, we calculate the exact molecular weight from the SMILES representation of each molecule using RDKit, simulating the molecular ion peak that would be observed in high-resolution MS. This comprehensive approach ensures that all relevant spectral modalities, including MS data, are represented in the fine-tuning dataset, enhancing the model's ability to adapt to new chemical spaces.

ZINC Dataset Evaluation

We initially tested this improvement cycle on a test set from the ZINC dataset to determine if further improvements could be achieved beyond the pretrained network. **Figure 3** presents the averaged Tanimoto similarity (**a**) and averaged correct sample probability (**b**) results for the ZINC test data before and after fine-tuning with different numbers of generated analogs. The multinomial sampling (MNS) approach, generating three candidates per target molecule, demonstrates remarkable effectiveness, identifying up to 96% of correct molecules within the top 3 candidates. For the MNS, we employed a molecular weight filter for the sampling process, accepting only molecules that fulfill these requirements. Greedy sampling also shows robust performance, correctly identifying up to 78% of molecules after fine-tuning. Model performance improves with up to 30 training analogs but plateaus or slightly declines with 50 or 100 analogs. This may result from the Mol2Mol model's tendency to generate up to 30 analogs per scaffold before switching, potentially impacting analog quality and fine-tuning effectiveness. For this experiment, we set the parameter for the number of samples per scaffold to 30. While the number of samples per scaffold is adjustable, we did not further investigate this parameter in the current study.

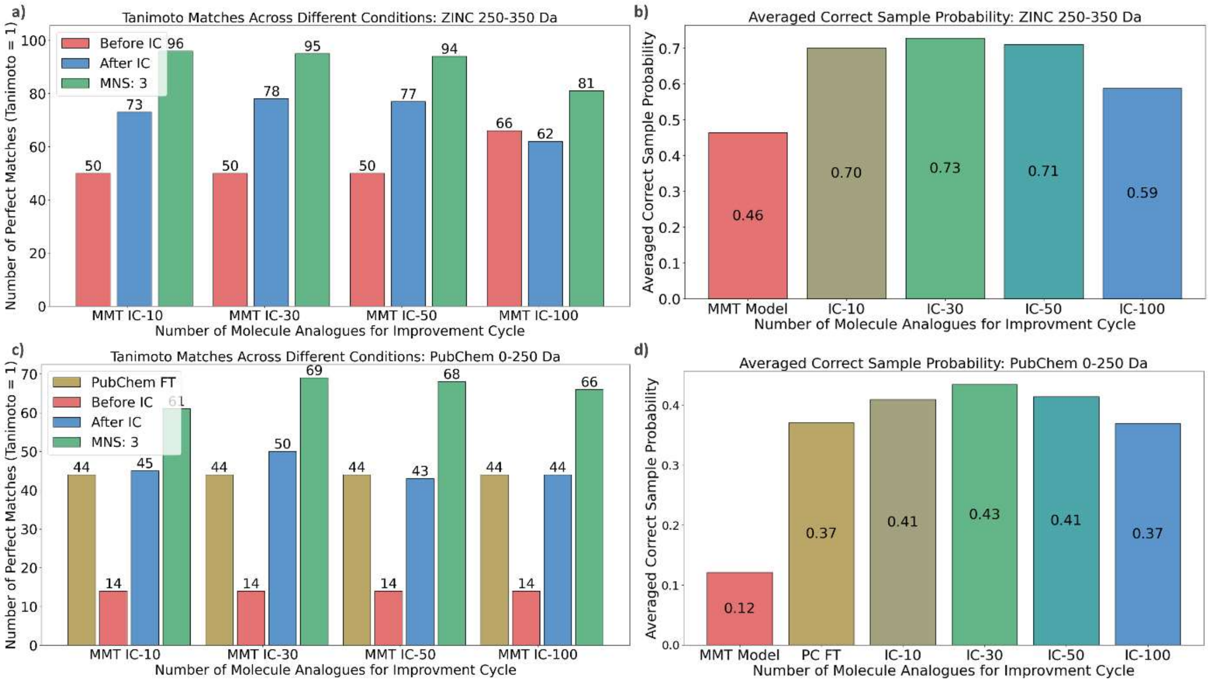


Figure 3: Performance Analysis of Improvement Cycle on ZINC (250-350 Da) and PubChem (0-250 Da) Datasets. a) ZINC Tanimoto Matches: Perfect matches (similarity = 1) for varying analog numbers (10-100) before/after Improvement Cycle (IC) and with multinomial sampling (MNS). b) ZINC Averaged Correct Sample Probability: Performance across IC stages (initial MMT, fine-tuned with 10-100 analogs). c) PubChem Tanimoto Matches: Perfect matches for varying analog numbers, including PubChem Fine-Tuned model comparison. d) PubChem Averaged Correct Sample Probability: Probabilities across IC stages, including PubChem fine-tuned model. Charts a) and c) compare pre-IC (red), post-IC (blue), PubChem finetuned(gold) and MNS (green, 3 samples) results. All bar charts show performance trends with increasing analog numbers.

PubChem Test Dataset Evaluation

We expanded our testing to the PubChem dataset to explore the model's capabilities across varied molecular weights and chemical structures beyond the initial training set. We curated three sets of 100 molecules from PubChem, categorized into molecular weight ranges: 0-250 Da (see **Figure 3c, d**), 250-350 Da, and 350-500 Da (see **Supplementary Figure 12**).

We applied the improvement cycle methodology, previously used for the ZINC dataset, to PubChem. This involved generating molecular analogs via the Mol2Mol model, simulating spectral data, and fine-tuning the model (details in Methods section). For comparison, we also fine-tuned the ZINC-trained MMT model on the entire PubChem training set (referred to as PC-FT), allowing us to assess the effectiveness of our targeted improvement cycles against a more comprehensive fine-tuning approach.

Results across all molecular weight ranges (0-250 Da, 250-350 Da, and 350-500 Da) demonstrated the effectiveness of the improvement cycle. Key findings include:

1. Consistent outperformance of the base model by all IC fine-tunings.

2. Peak performance generally observed with 30 analogues, often surpassing the PC-FT model even with just 10 analogs.

3. For larger molecules (350-500 Da), performance improved with higher numbers of analogs, with 100 analogs yielding the best results.

Iterative application of the cycle on the 10-analog generation model showed further improvements in model accuracy (**Supplementary Figure 13**). Comprehensive visualizations of the model's improvement and chemical space exploration are provided in **Supplementary Information Section 5**, including t-SNE plots and examples of molecules from different weight ranges and their generated analogs.

Our evaluation across molecular weight ranges demonstrates the model's adaptability and the improvement cycle's effectiveness in enhancing performance on diverse structures. The cycle significantly improves identification accuracy, with notable results even when using just 10 analogs for weight ranges covered in initial training. For the ZINC dataset (250-350 Da), the IC increased perfect Tanimoto matches from 50% to 73% with 10 analogs, and up to 78% with 30 analogs. In the PubChem dataset (0-250 Da), the IC improved perfect matches from 14% to 45% with 10 analogs, surpassing the 44% achieved by the model fine-tuned on the entire PubChem dataset. Multinomial sampling consistently outperforms greedy sampling in identifying correct molecules, achieving up to 96% and 68% accuracy within the top 3 candidates for the ZINC and PubChem dataset, respectively. However, selecting the single most accurate candidate remains a challenge. The following section explores an additional step to address this.

Enhancing Structure Prediction Accuracy through Targeted Fine-Tuning of Simulation and Real Data

This experiment evaluates the performance of the MMT model using a selected set of 34 molecules for which all experimental data modalities (^{1}H, ^{13}C, HSQC, COSY, IR, and MS) were available. The peaks of all real experimental data used in this study was manually peak picked to ensure accuracy and consistency. A preliminary study (**Supplementary Section 2**) revealed that multinomial sampling (MNS) combined with peak matching significantly outperformed greedy sampling, increasing identification accuracy from 50% to 90% (see **Supplementary Figure 7**). Despite challenges in differentiating similar structures (examples in **Supplementary Figure 8**), we adopted MNS with spectral error ranking, replacing greedy sampling for this experiment.

For each target, we sampled molecules using multinomial sampling (MNS with sample size 3x 20) and applied molecular weight filters to ensure the generated molecules matched these key properties of the target compound. After sampling we employed the HSQC matching methodology developed in our previous research,[14] using the HungDist-NN matching algorithm to score and rank the set of sampled molecules. Additionally, we applied the same point matching methodology for the COSY spectrum and investigated its impact on molecule ranking. Evaluation uses three ranking methods: COSY, HSQC, and combined HSQC & COSY, with top-k accuracy calculated for k = 1, 3, 5, 10, 20 and "Total" representing all generated molecules out of the maximum of 60 sampling options.

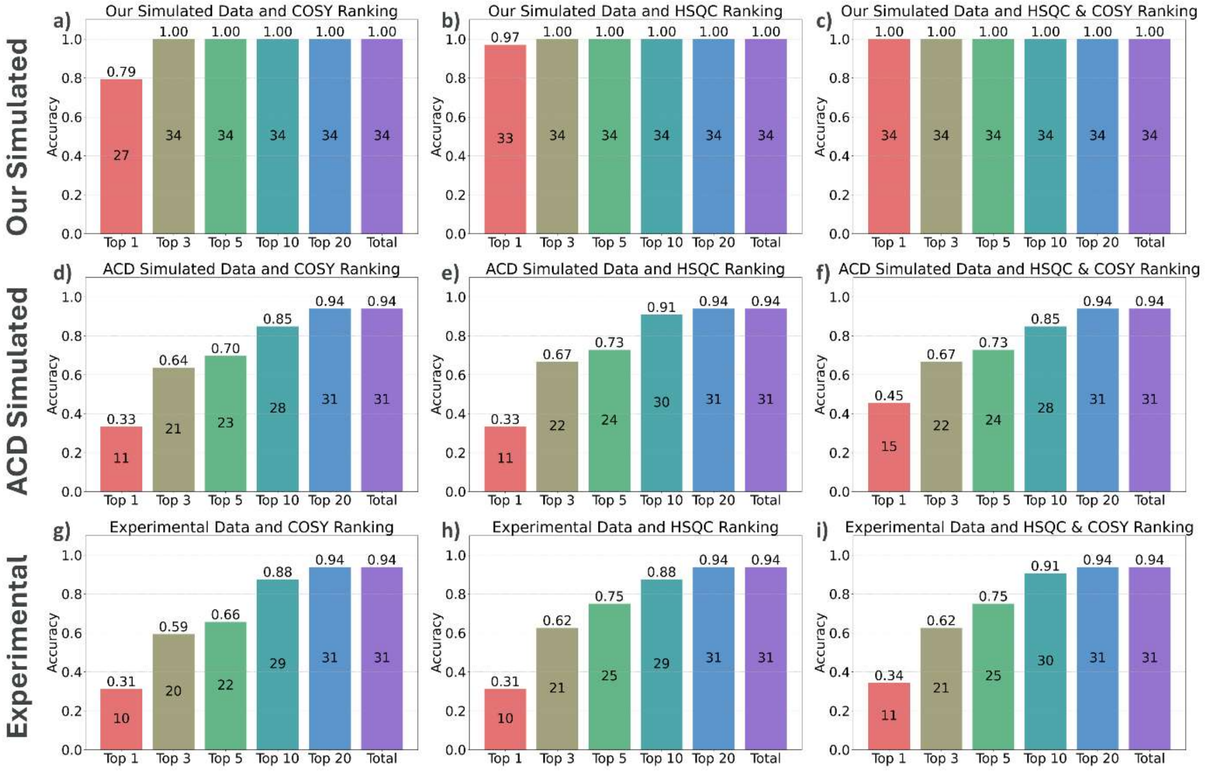


Figure 4: Top-k accuracy of the MMT model across different data types and ranking methods using the correct starting guess molecule. (a-c) Results for our simulated data, (d-f) ACD simulated data, and (g-i) experimental data. For each data type, results are shown using COSY, HSQC, and combined HSQC & COSY ranking methods, respectively. Each subplot displays accuracy for top 1, 3, 5, 10, 20 predictions and total number of correct hits, with the number of correct predictions indicated within each bar. N=34 for all experiments.

Results for the MMT model, illustrated in **Figure 4**, show near-perfect accuracy on our simulated data (**a-c**), with already top-3 accuracy reaching 100% across all ranking methods. Performance on ACD Labs simulated data (**d-f**) remains strong, with top-3 accuracy of 67% and reaching 94% for top-20. Experimental data (**g-i**) shows similar performance, with top-3 accuracy between 62% and a total top performance of 94% over all sampled molecules. Notably, the combined HSQC & COSY ranking outperforms individual rankings across all data types, suggesting enhanced prediction accuracy. For instance, in experimental data, the combined ranking achieves 91% top-10 accuracy compared to 88% for COSY or HSQC alone.

The baseline performance of the pretrained MMT model without the improvement cycle, detailed in **Supplementary Information Section 6**, demonstrates the model's initial limitations: solving only 58% of our simulated data, 16% of ACD Labs simulated data, and a mere 3% of real experimental data. In contrast, the improved model achieves high accuracy on experimental data, highlighting the effectiveness of our improvement cycle.

This dramatic enhancement, despite training solely on simulated spectra, underscores that comprehensive chemical space coverage is more critical than precise training data accuracy. The model's ability to adapt to various spectral simulation methods and real experimental data showcases its robustness and generalizability, crucial for diverse research environments. These results demonstrate the model's practical applicability in real-world structure elucidation tasks, with the potential to significantly accelerate and improve the structure determination process in chemistry laboratories. Furthermore, the model's adaptability to discrepancies between simulated and experimental data suggests that additional fine-tuning with real experimental data could yield even greater improvements, opening avenues for further enhancing its performance in practical settings.

Model Robustness and Adaptability

Our assessment of the improvement cycle's robustness revealed that the model maintains good performance even when starting with slightly incorrect molecular structures, as shown in **Figure 5**. Comparing the total number of correctly identified molecules, we observed the following: For our simulated data, the model identified 34 out of 34 (100%) molecules with correct starting structures, and 29 out of 34 (85%) with modified starting structures. With ACD simulated data, 31 out of 34 (94%) molecules were found using correct starting structures, compared to 22 out of 34 (65%) with modified starting structures. For experimental data, the model identified 31 out of 34 (94%) molecules correctly with accurate initial guesses, versus 19 out of 34 (56%) with modified starting points. These results demonstrate that while there is a decrease in performance when starting with modified structures, the model still maintains a substantial ability to identify correct molecules. This robustness is valuable in real-world applications where initial structural assumptions may not always be entirely accurate, showcasing the Mol2Mol network's ability to explore relevant chemical space and allow the MMT model to overcome initial inaccuracies in many cases. A figure showing all the wrong starting guesses of the molecules is presented in **Supplementary Figure 4**.

The model's ability to maintain good performance even with slightly incorrect initial structures has significant implications for real-world applications. In practice, chemists often begin structure elucidation with incomplete or partially incorrect hypotheses. Our model's robustness in these scenarios suggests it can serve as a powerful tool to refine and correct initial structural guesses, potentially reducing the iterative cycles typically required in structure determination. This capability could be particularly valuable in analyzing complex natural products or in drug discovery processes where rapid and accurate structure elucidation is crucial.

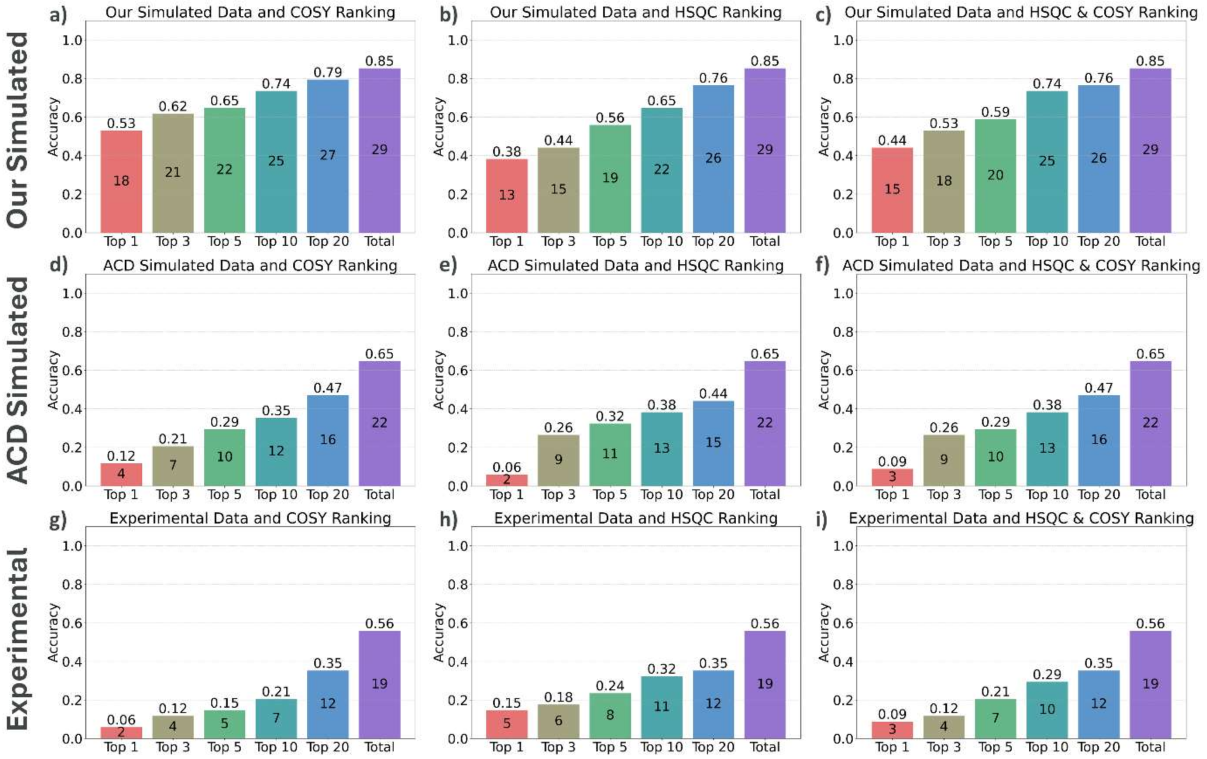


Figure 5: Top-k accuracy of the MMT model across different data types and ranking methods using an incorrect starting guess molecule. (a-c) Results for our simulated data, (d-f) ACD simulated data, and (g-i) experimental data. For each data type, results are shown using COSY, HSQC, and combined HSQC & COSY ranking methods, respectively. Each subplot displays accuracy for top 1, 3, 5, 10, 20 predictions and total number of correct hits, with the number of correct predictions indicated within each bar. N=34 for all experiments.

Additional Model Evaluation Experiments

In addition to the improvement cycle experiments, we conducted further evaluations of our model's performance under various conditions. These experiments included assessing prediction accuracy with and without molecular weight constraints and evaluating the number of attempts required for correct structure generation. These additional experiments provided valuable insights into the model's capabilities and limitations. For a detailed description of these experiments and their results, please refer to **Supplementary Information Section 4**.

Model Explainability and Practical Application

The transformer's token-based predictions provide explainable suggestions, offering insights into the model's decision-making process. By analyzing the token-level confidence scores, we can identify which structural features the model is most certain about and which areas might require further refinement. **Figure 6** demonstrates this concept using an experimental dataset example. The target molecule (a) is compared with four model predictions (b-e), sampled using multinomial sampling from our fine-tuned MMT model. SMILES strings are color-coded to represent the model's token-level confidence: green for high and pink for low.

We observe that all suggested molecules share core structural features (two aromatic rings and one aliphatic piperidine ring) predicted with high confidence. Variations in substituent positions or specific functional groups, such as chlorine atom and amino group placements, are predicted with lower confidence. Notably, the correct molecule is the most probable suggestion, and there's good correlation between spectra reconstruction errors and overall sample probability.

This visualization allows us to identify which structural features the model is most certain about and which areas might require further refinement, providing a window into the model's intuition process. Similar analyses for our simulated data and ACD Labs predictions are shown in **Supplementary Figure 22** and **Supplementary Figure *23***.

These explainability features provide practical benefits for structure elucidation tasks. By visualizing the model's confidence in structural components, chemists can focus on uncertain aspects, guiding targeted experimental work like selective 2D NMR or chemical derivatization. For multiple suggested structures, confidence visualization aids in prioritizing hypotheses, streamlining the elucidation process. The correlation between prediction probabilities and spectral reconstruction errors offers a metric for assessing prediction reliability, helping chemists decide when to trust the model's suggestions or seek additional experimental evidence.

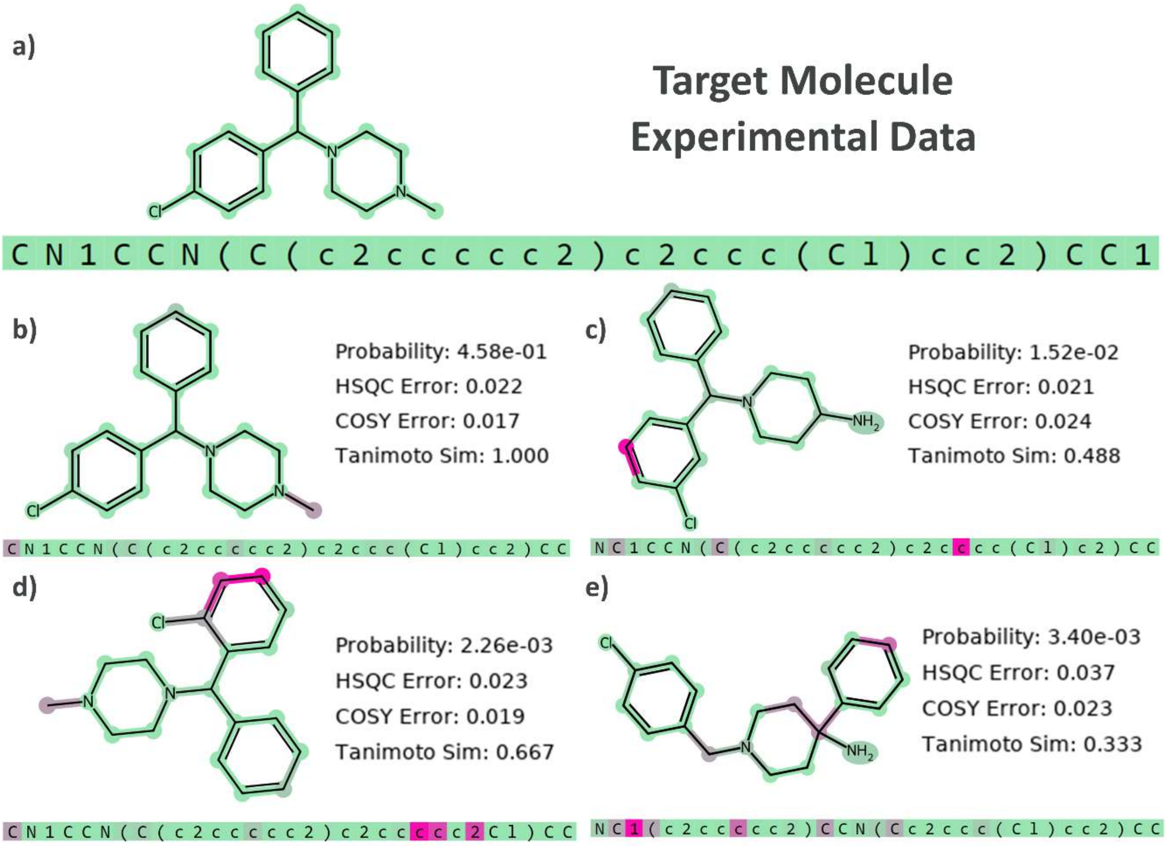


Figure 6: Explainability analysis of model predictions for an experimental molecule. (a) Target molecule from experimental data with its SMILES representation. (b-e) Predictions from the model, showing the molecular structure, SMILES string, prediction probability, HSQC and COSY errors, and Tanimoto similarity to the target. Color-coding on SMILES strings indicates the model's token-level confidence: green for high confidence, red for low confidence. This visualization demonstrates the model's ability to capture core structural features while highlighting areas of uncertainty in predicting specific substituents or functional groups.

To facilitate practical application of these insights, we have implemented the full improvement cycle workflow as a GUI in the form of an HTML website. This interface includes probability plotting for suggested molecules and allows users to compare simulated spectra of generated molecules with experimental data. The code and user manual are provided in the **Supplementary Information Section 7** with explanatory screenshots in **Supplementary Figure 24**-**Supplementary Figure *28***, offering chemists a powerful tool to elucidate structures based on our MMT model.

To leverage these insights in practical applications, we have implemented the full improvement cycle workflow together with the probability plottings of the suggested molecules. This implementation also includes an option for users to compare the simulated spectra of the generated molecules with the experimental spectra. This feature should aid in investigating potential wrong assignments and mistakes of the model, providing a powerful tool for chemists to critically evaluate and refine the model's predictions.

Conclusion

In this study, we developed the MultiModalTransformer (MMT) model, an innovative architecture for molecular structure elucidation that integrates multiple spectroscopic modalities including NMR, IR, and MS data. Our research demonstrates enhanced molecular structure prediction accuracy through the integration of diverse spectral data, achieving up to 94% correct identifications for experimental samples. We implemented a robust simulated data generation pipeline and an iterative improvement cycle, enabling effective performance on experimental data despite training only on simulations. The model's robustness is evidenced by its ability to overcome initial structural inaccuracies, maintaining substantial predictive power even with incorrect starting guesses. Additionally, the MMT provides explainable predictions through token-based analysis, offering insights into its decision-making process. While there's potential for further enhancement through increased experimental data training and automated peak-picking, the current model already serves as a powerful tool for chemists, effectively bridging the gap between simulated and experimental data. Its adaptability makes it particularly valuable for real-world applications where perfect initial structural information may not be available, representing a significant advancement in automated structure elucidation.

Data and Code Availability

The code used in this study will be made available upon reasonable request to foster further academic collaboration and verification. Additionally, the simulated data used for training the networks is made available under the following Zenodo link (https://zenodo.org/uploads/13221541). Github: https://github.com/knlr326_azu/MultiModalTransformer

Acknowledgements

We gratefully acknowledge AstraZeneca for their support and funding of the Postdoctoral position, instrumental in the success of this research.

AI-Assisted Development and Manuscript Preparation

For the development of the MultiModalTransformer and the preparation of this manuscript, we utilized AI technologies, including OpenAI's ChatGPT and Claude for code development support and Grammarly for text refinement. These tools served as supplementary aids, providing assistance in editing and optimizing both code and content. The AI suggestions were rigorously reviewed, tested, and selectively implemented by the authors to ensure the integrity and functionality of the code, as well as the accuracy, consistency, and clarity of the manuscript. Despite the involvement of AI technologies, the responsibility for the final content, its validation, and the overall quality of the paper rests solely with the authors.

References:

1. Klein, D. R. *Organic Chemistry*. (Wiley, 2020).

2. Noor, P., Khanmohammadi, M., Roozbehani, B. & Bagheri Garmarudi, A. Evaluation of ATR-FTIR spectrometry in the fingerprint region combined with chemometrics for simultaneous determination of benzene, toluene, and xylenes in complex hydrocarbon mixtures. *Monatsh Chem* **149**, 1341–1347 (2018).

3. Lewis, R. R., Rowlands, B., Jonsson, L. R., Goodman, J. M. & Howe, P. Towards automatically verifying chemical structures: the powerful combination of ^{1}H NMR and IR spectroscopy. *Res Sq* (2024) doi:10.21203/RS.3.RS-4719113/V1.

4. Coates, J. *Interpretation of Infrared Spectra, A Practical Approach. In Encyclopedia of Analytical Chemistry*. (John Wiley & Sons Ltd, 2020).

5. Griffiths, L. & Horton, R. Towards the automatic analysis of NMR spectra: Part 6. Confirmation of chemical structure employing both 1H and 13C NMR spectra. *Magnetic Resonance in Chemistry* **44**, 139–145 (2006).

6. Burns, D. C., Mazzola, E. P. & Reynolds, W. F. The role of computer-assisted structure elucidation (CASE) programs in the structure elucidation of complex natural products. *Nat Prod Rep* **36**, 919–933 (2019).

7. Valli, M. *et al.* Computational methods for NMR and MS for structure elucidation I: software for basic NMR. *Physical Sciences Reviews* **4**, (2019).

8. Valli, M. *et al.* Computational methods for NMR and MS for structure elucidation II: Database resources and advanced methods. *Physical Sciences Reviews* **4**, (2019).

9. Elyashberg, M. *et al.* Computer-assisted methods for molecular structure elucidation: Realizing a spectroscopist's dream. *J Cheminform* **1**, 1–26 (2009).

10. Steinbeck, C. Recent developments in automated structure elucidation of natural products. *Nat Prod Rep* **21**, 512–518 (2004).

11. Smith, S. G. & Goodman, J. M. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: The DP4 probability. *J Am Chem Soc* **132**, 12946–12959 (2010).

12. Ermanis, K., Parkes, K. E. B., Agback, T. & Goodman, J. M. Doubling the power of DP4 for computational structure elucidation. *Org Biomol Chem* **15**, 8998–9007 (2017).

13. Howarth, A., Ermanis, K. & Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem Sci* **11**, 4351–4359 (2020).

14. Priessner, M. *et al.* HSQC Spectra Simulation and Matching for Molecular Identification. *J Chem Inf Model* **64**, 34 (2023).

15. Howarth, A. & Goodman, J. M. The DP5 probability, quantification and visualisation of structural uncertainty in single molecules. *Chem Sci* **13**, 3507–3518 (2022).

16. Grimblat, N., Zanardi, M. M. & Sarotti, A. M. Beyond DP4: An Improved Probability for the Stereochemical Assignment of Isomeric Compounds using Quantum Chemical Calculations of NMR Shifts. *Journal of Organic Chemistry* **80**, 12526–12534 (2015).

17. Grimblat, N., Gavín, J. A., Hernández Daranas, A. & Sarotti, A. M. Combining the Power of J Coupling and DP4 Analysis on Stereochemical Assignments: The J-DP4 Methods. *Org Lett* **21**, 4003–4007 (2019).

18. Yang, Z., Vegh, V., Reutens, D. C. & Pierens, G. K. A rapid procedure for spectral similarity matching of heteronuclear single quantum coherence spectra. *Proceedings - 2011 International Conference on Digital Image Computing: Techniques and Applications, DICTA 2011* 302–307 (2011) doi:10.1109/DICTA.2011.57.

19. Pierens, G. K., Mobli, M. & Vegh, V. Effective protocol for database similarity searching of heteronuclear single quantum coherence spectra. *Anal Chem* **81**, 9329–9335 (2009).

20. Bodis, L., Ross, A., Bodis, J. & Pretsch, E. Automatic compatibility tests of HSQC NMR spectra with proposed structures of chemical compounds. *Talanta* **79**, 1379–1386 (2009).

21. Xin, D., Jones, P. J. & Gonnella, N. C. D iCE: Diastereomeric in Silico Chiral Elucidation, Expanded DP4 Probability Theory Method for Diastereomer and Structural Assignment. *Journal of Organic Chemistry* **83**, 5035–5043 (2018).

22. Novitskiy, I. M. & Kutateladze, A. G. DU8+ Computations Reveal a Common Challenge in the Structure Assignment of Natural Products Containing a Carboxylic Anhydride Moiety. *Journal of Organic Chemistry* **86**, 17511–17515 (2021).

23. Tsai, Y. H. *et al.* ML-J-DP4: An Integrated Quantum Mechanics-Machine Learning Approach for Ultrafast NMR Structural Elucidation. *Org Lett* **24**, 7487–7491 (2022).

24. Zanardi, M. M. & Sarotti, A. M. Sensitivity Analysis of DP4+ with the Probability Distribution Terms: Development of a Universal and Customizable Method. *Journal of Organic Chemistry* **86**, 8544–8548 (2021).

25. Marcarino, M. O., Zanardi, M. M., Cicetti, S. & Sarotti, A. M. NMR calculations with quantum methods: Development of new tools for structural elucidation and beyond. *Acc Chem Res* **53**, 1922–1932 (2020).

26. Sarotti, A. M. Successful combination of computationally inexpensive GIAO 13C NMR calculations and artificial neural network pattern recognition: a new strategy for simple and rapid detection of structural misassignments. *Org Biomol Chem* **11**, 4847–4859 (2013).

27. Zanardi, M. M. & Sarotti, A. M. GIAO C-H COSY Simulations Merged with Artificial Neural Networks Pattern Recognition Analysis. Pushing the Structural Validation a Step Forward. *Journal of Organic Chemistry* **80**, 9371–9378 (2015).

28. Pesek, M. *et al.* Database Independent Automated Structure Elucidation of Organic Molecules Based on IR, 1H NMR, 13C NMR, and MS Data. *J Chem Inf Model* **61**, 756–763 (2021).

29. Alberts, M., Zipoli, F. & Vaucher, A. C. Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models. (2023) doi:10.26434/CHEMRXIV-2023-8WXCZ.

30. Alberts, M., Laino, T. & Vaucher, A. C. Leveraging Infrared Spectroscopy for Automated Structure Elucidation. *ChemRxiv* (2023) doi:10.26434/CHEMRXIV-2023-5V27F.

31. Huang, Z., Chen, M. S., Woroch, C. P., Markland, T. E. & Kanan, M. W. A framework for automated structure elucidation from routine NMR spectra. *Chem Sci* **12**, 15329–15338 (2021).

32. Kim, H. W. *et al.* DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data. *J Cheminform* **15**, 1–12 (2023).

33. Zhang, J. *et al.* NMR-TS: de novo molecule identification from NMR spectra. *Sci Technol Adv Mater* 552–561 (2020) doi:10.1080/14686996.2020.1793382.

34. Fine, J. A., Rajasekar, A. A., Jethava, K. P. & Chopra, G. Spectral deep learning for prediction and prospective validation of functional groups. *Chem Sci* **11**, 4618–4630 (2020).

35. Jung, G., Jung, S. G. & Cole, J. M. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chem Sci* **14**, 3600–3609 (2023).

36. Vaswani, A. *et al.* Attention Is All You Need. *Adv Neural Inf Process Syst* **2017-December**, 5999–6009 (2017).

37. Han, J. *et al.* Scalable graph neural network for NMR chemical shift prediction. *Physical Chemistry Chemical Physics* **24**, 26870–26878 (2022).

38. McGill, C., Forsuelo, M., Guan, Y. & Green, W. H. Predicting Infrared Spectra with Message Passing Neural Networks. *J Chem Inf Model* **61**, 2594–2609 (2021).

39. Greg Landrum. RDKit: Open-source cheminformatics. *http://www.rdkit.org* (2022).

40. He, J. *et al.* Transformer-based molecular optimization beyond matched molecular pairs. *J Cheminform* **14**, 1–14 (2022).

41. Tibo, A., He, J., Janet, J. P., Nittinger, E. & Engkvist, O. Exhaustive local chemical space exploration using a transformer model. (2023) doi:10.26434/CHEMRXIV-2023-V25XB.

42. He, J. *et al.* Molecular optimization by capturing chemist's intuition using deep neural networks. *J Cheminform* **13**, 1–17 (2021).

43. Loeffler, H. H. *et al.* Reinvent 4: Modern AI–driven generative molecule design. *J Cheminform* **16**, 1–16 (2024).