

# Molecular structure inference from molecular spectra

# Custom ChatGPT interface to Alberts et al. [1]

The screenshot shows a web interface for a custom ChatGPT model. At the top, the title 'NMR multiplets to SMILES translator' is displayed with a dropdown arrow. A blue 'PLUS' button is in the top right corner. Below the title, there is a button labeled 'FAQ: What's this about?'. The main text area contains the following information:

This tool translates textual **<sup>1</sup>H NMR multiplet data** into **SMILES** (strings that represent chemical structures).

You provide:

- A **molecular formula** (e.g., C9H11N3)
- Followed by the keyword **<sup>1</sup>H NMR**
- Then a list of **multiplets** with ranges, types, integrals, and J-values

I return:

- One or more **predicted SMILES** strings with confidence scores

Want help formatting your data? A multiplet table can be generated in **MestReNova** via:  
**Analysis** → **Multiplet Analysis** → **NMR Multiplet Table**.

Below the text, there is a table header with a dropdown arrow above it:

Name	Shift	Range	H's	Integral	Class	J's	Type	Method
------	-------	-------	-----	----------	-------	-----	------	--------

Below the table header is a text input field with the placeholder text 'Message NMR multiplets to SMILES translator'. To the left of the input field is a '+' button, and to the right is an '↑' button.

At the bottom of the interface, there is a footer message: 'New version of GPT available - Continue chatting to use the old version, or start a [new chat](#) for the latest version.' and a '?' icon in the bottom right corner.

# Custom ChatGPT interface to Alberts et al. [1]

📄 ✎ NMR multiplets to SMILES translator

	Name	Shift	Range	H's	Integral	Class	J's	Type	Method
1	A (t)	3.44	3.47 .. 3.41	1	1.29	t	6.13, 6.13	Compound	Peaks
2	B (m)	1.54	1.59 .. 1.50	1	0.90	m		Compound	Peaks
3	C (m)	1.28	1.34 .. 1.21	9	8.80	m		Compound	Peaks
4	D (m)	0.89	0.94 .. 0.85	2	2.01	m		Compound	Peaks

Great! I can format this for SMILES prediction, but I'll need the **molecular formula** first (e.g., `C 12 H 26 O`). Could you provide that?

C19H40O

Thanks! Based on your data and formula, here's the Formatted input for SMILES prediction:

mathematica

Message NMR multiplets to SMILES translator

New version of GPT available - Continue chatting to use the old version, or start a [new chat](#) for the latest version.

<https://chatgpt.com/g/g-67e7f80c685881918d42729048cd2985-nmr-multiplets-to-smiles-translator>

1 / 19

# Custom ChatGPT interface to Alberts et al. [1]

The screenshot shows a web interface for a custom ChatGPT model. At the top, the title bar reads 'NMR multiplets to SMILES translator'. A user input box contains the text 'C19H40O'. The model's response begins with 'Thanks! Based on your data and formula, here's the Formatted input for SMILES prediction:'. Below this, a code block is displayed with the text 'mathematica' and a copy/edit icon. The code block contains a line of NMR data: 'C 19 H 40 O 1HNMR | 3.47 3.41 t 1H J 6.13 6.13 | 1.59 1.50 m 1H | 1.34 1.21 m 9H | 0.5'. Below the code block, the model asks 'Want me to go ahead and run the prediction?'. A user response box contains 'Yes, please.'. The model's next response is 'Talked to ml-alberts-20250329.randompile.net'. Below this, it says 'Here are the predicted SMILES strings with their scores:'. A list item follows: '1. CCCCCCCCCCCCCC(C)(C)CO — Score: -0.0897'. At the bottom, there is a text input field with the placeholder 'Message NMR multiplets to SMILES translator' and a send button. A footer note states 'New version of GPT available - Continue chatting to use the old version, or start a new chat for the latest version.'

NMR multiplets to SMILES translator

C19H40O

Thanks! Based on your data and formula, here's the Formatted input for SMILES prediction:

```
mathematica
```

C 19 H 40 O 1HNMR | 3.47 3.41 t 1H J 6.13 6.13 | 1.59 1.50 m 1H | 1.34 1.21 m 9H | 0.5

Want me to go ahead and run the prediction?

Yes, please.

Talked to ml-alberts-20250329.randompile.net

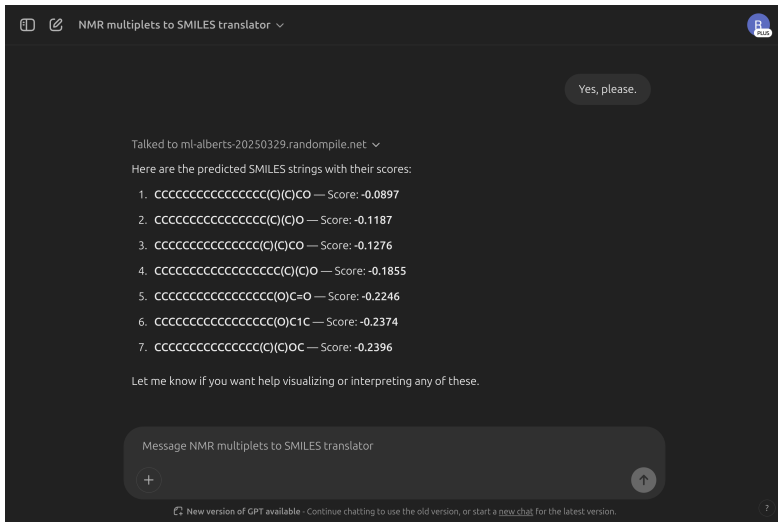
Here are the predicted SMILES strings with their scores:

1. CCCCCCCCCCCCCC(C)(C)CO — Score: -0.0897

Message NMR multiplets to SMILES translator

New version of GPT available - Continue chatting to use the old version, or start a new chat for the latest version.

# Custom ChatGPT interface to Alberts et al. [1]



# Training dataset and transformers model

- **Dataset:** USPTO reaction dataset<sup>1</sup>
  - 1,435,481 reactions
  - Realistic molecules & common chemicals
- **Unique molecules:**
  - Initially: 1,675,439
  - Filter:  $5 < \# \text{ heavy atoms} < 35$
  - Allowed elements: C, H, O, N, S, P, Si, B, halogens
  - After filtering: 1,416,499
  - Simulation w/ MestReNova (all spectra): **794,403** molecules
- **Spectra simulated:** <sup>1</sup>H-NMR, IR, <sup>13</sup>C-NMR, HSQC-NMR, MS/MS
- “... **deuterated chloroform** as solvent. Default settings were used”
- **Representation:** SMILES

---

<sup>1</sup>Lowe, 2017 [2]

# Training dataset 001\_big

## SRC-TRAIN.TXT

- C 15 H 24 N 2 1HNMR | 6.86 6.83 d 1H J 0.68 | 5.98  
5.96 s 1H | 4.22 4.19 s 2H | 3.83 3.74 p 1H J 6.74 |  
3.55 3.50 m 2H | 2.24 2.21 s 3H | 2.05 1.99 m 2H |  
1.35 1.32 s 5H | 1.26 1.22 d 6H J 6.67
- C 30 H 28 F N O 3 1HNMR | 7.39 7.30 m 4H | 7.30 7.05 m  
11H | ... | 2.60 2.46 m 2H
- ...

## TGT-TRAIN.TXT

- C c 1 c c 2 c ( c c 1 N ) N ( C ( C ) C ) ... ( C ) C
- O = C ( C C c 1 c c c c c 1 ) N C ( C c 1 c c c ( O c  
2 c c c c c 2 ) c c 1 ) C ( O ) c 1 c c c ( F ) c c 1
- ...

# Training dataset 002\_big\_explicit-h

## SRC-TRAIN.TXT

- C 15 H 24 N 2 1HNMR | 6.86 6.83 d 1H J 0.68 | 5.98 5.96 s 1H | 4.22 4.19 s 2H | 3.83 3.74 p 1H J 6.74 | 3.55 3.50 m 2H | 2.24 2.21 s 3H | 2.05 1.99 m 2H | 1.35 1.32 s 5H | 1.26 1.22 d 6H J 6.67
- C 30 H 28 F N O 3 1HNMR | 7.39 7.30 m 4H | 7.30 7.05 m 11H | ... | 2.60 2.46 m 2H
- ...

## TGT-TRAIN.TXT with explicit hydrogens (new)

- [CH3] [c] 1 [cH] [c] 2 [c] ( [cH] [c] 1 [NH2] ) [N] ( [CH] ( [CH3] ) [CH3] ) [CH2] [CH2] [C] 2 ( [CH3] ) [CH3 ]
- ...



# Training dataset 003\_big\_ex-h\_mono

## SRC-TRAIN.TXT only fully-specified stereo (new)

- C 15 H 24 N 2 1HNMR | 6.86 6.83 d 1H J 0.68 | 5.98 5.96 s 1H | 4.22 4.19 s 2H | 3.83 3.74 p 1H J 6.74 | 3.55 3.50 m 2H | 2.24 2.21 s 3H | 2.05 1.99 m 2H | 1.35 1.32 s 5H | 1.26 1.22 d 6H J 6.67
- C 30 H 28 F N O 3 1HNMR | 7.39 7.30 m 4H | 7.30 7.05 m 11H | ... | 2.60 2.46 m 2H
- ...

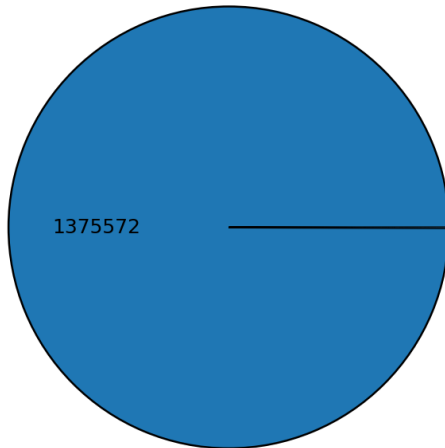
## TGT-TRAIN.TXT with explicit hydrogens and stereo-free (new)

- [CH3] [c] 1 [cH] [c] 2 [c] ( [cH] [c] 1 [NH2] ) [N] ( [CH] ( [CH3] ) [CH3] ) [CH2] [CH2] [C] 2 ( [CH3] ) [CH3 ]
- ...

## Technical note (new)

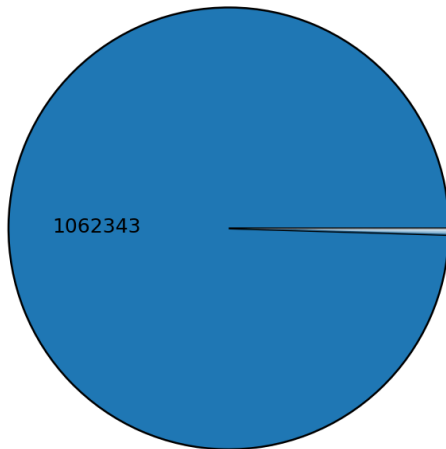
- Molecules from the same patent are more similar
- Thus: samples are not independent
- The train/test dataset split should preserve groups (patents)
- But: the bipartite graph smiles–patents is (almost) fully connected
- Heuristic: disconnect top nodes until largest component is small

## Top-100 connected components (#nodes)



#nodes: 1376543, #edges: 4940383, #connected components: 156

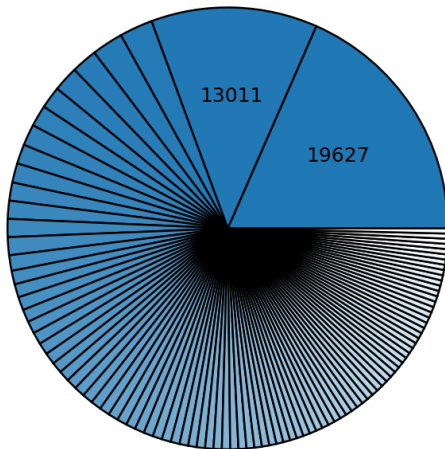
## Top-100 connected components (#nodes)



#nodes: 1376543, #edges: 1937663, #connected components: 228708

...

## Top-100 connected components (#nodes)



#nodes: 1376543, #edges: 987491, #connected components: 661984

617'347 molecules from Alberts et al. [1] found in USPTO [2]  
Not found: 177'039  
(even canonicalized and ignoring stereo info)

Removed samples with incomplete stereo info:  
MestReNova behavior not clearly documented [3, Ch. 13]

Training set: 582'903  
Validation set: 30'680  
Test set: 82'655

– constructed preserving the groups

# Training and model predictions

- Model like Alberts et al. [1], **32 heads**,  $\sim 31\text{M}$  parameters
- 582'903 training samples, batch size 4k
- Trained on NVIDIA 1  $\times$  A10 GPU<sup>2</sup> up to 100k batches
- The model generates several hypotheses ranked by “score”, i.e., log-likelihood of the prediction according to the model
- Top-N accuracy evaluated on 10'000 unseen samples
- “Beam search” is a heuristic to find the *most likely* hypotheses; use beams size 20, keep top 10

---

<sup>2</sup>Thanks to Lambda Cloud for compute credits



# Evaluation

What is a *correct* prediction by the trained model?

- ✓ Canonicalize before comparison
- ✓ Remove stereo information

In an achiral medium/solvent, NMR is blind to mirror symmetry.  
Not well-captured in the model architecture or the dataset.

- How to generate all structures compatible with a given spectrum?
  - Molecules with indistinguishable spectra
  - Mixture of molecules from mixed spectra
- The spectrum  $\approx$  a mixture of molecular fragments
- The model should have a built-in sense for additivity

- How to generate all structures compatible with a given spectrum?
  - Molecules with indistinguishable spectra
  - Mixture of molecules from mixed spectra
- The spectrum  $\approx$  a mixture of molecular fragments
- The model should have a built-in sense for additivity

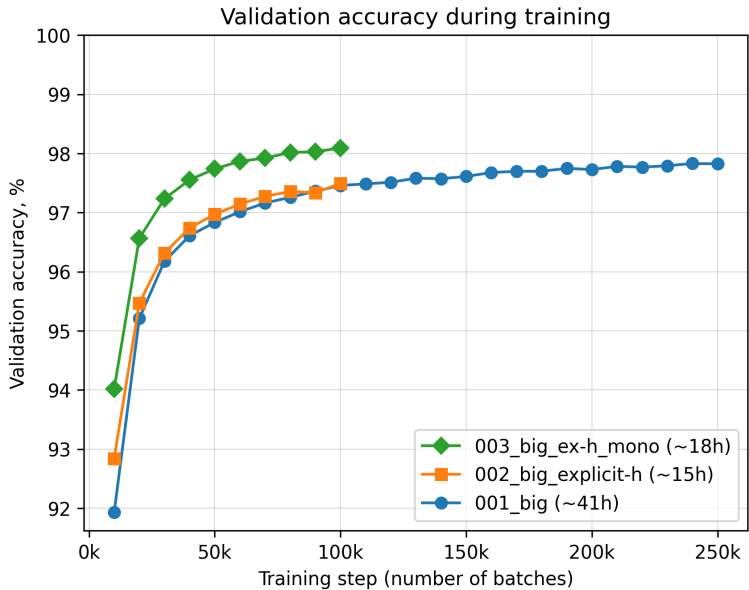
- How to generate all structures compatible with a given spectrum?
  - Molecules with indistinguishable spectra
  - Mixture of molecules from mixed spectra
- The spectrum  $\approx$  a mixture of molecular fragments
- The model should have a built-in sense for additivity

# Evaluation

Validation accuracy during training –

next-token prediction accuracy  
over the unused validation dataset  
with “teacher forcing”

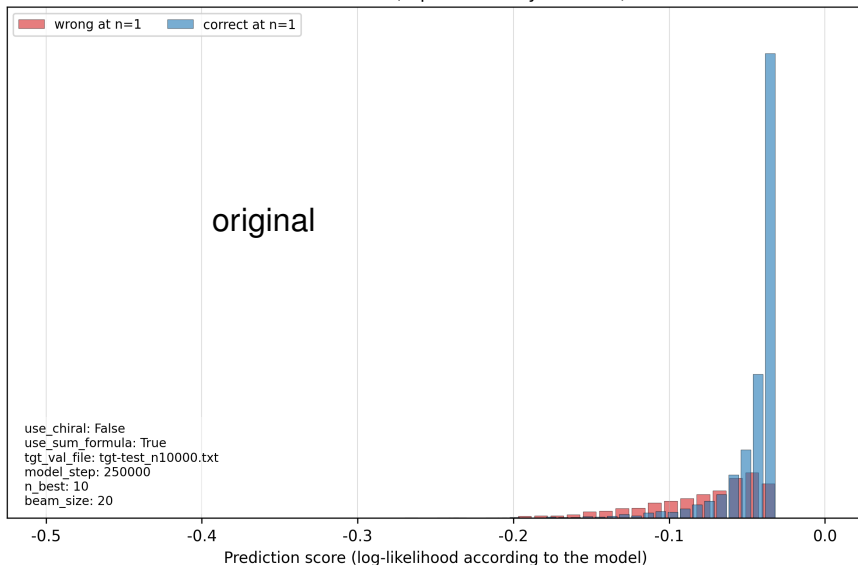
– indicates training progress.



# Evaluation

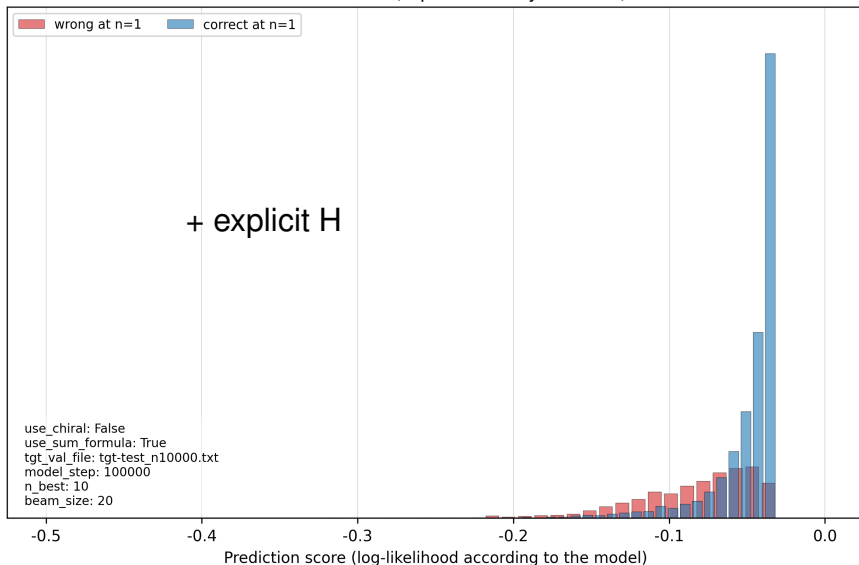
top-n accuracy of three models

# Inferred SMILES (top-1 accuracy: 71.02%)

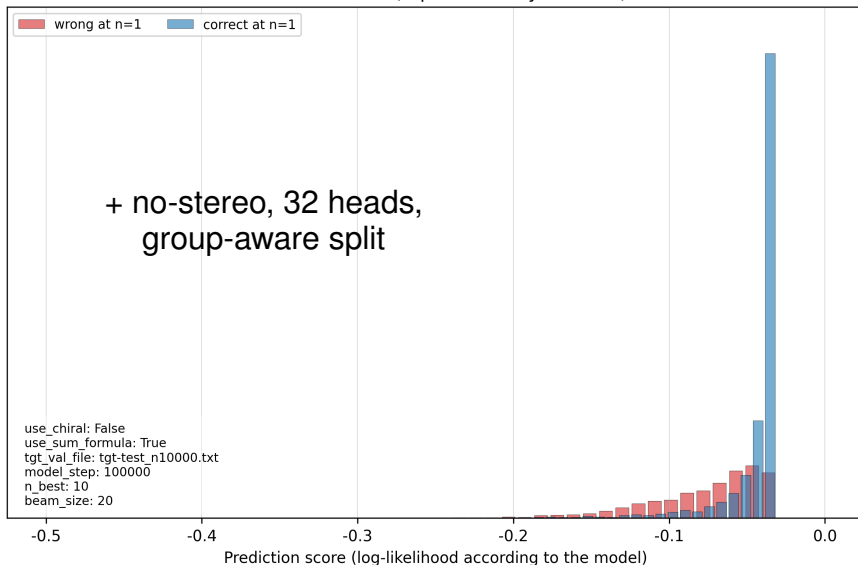




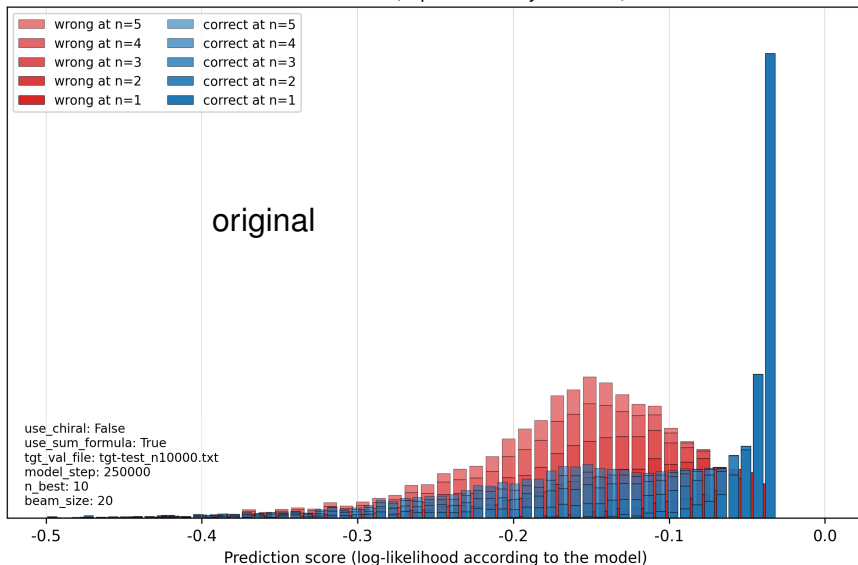
# Inferred SMILES (top-1 accuracy: 67.54%)



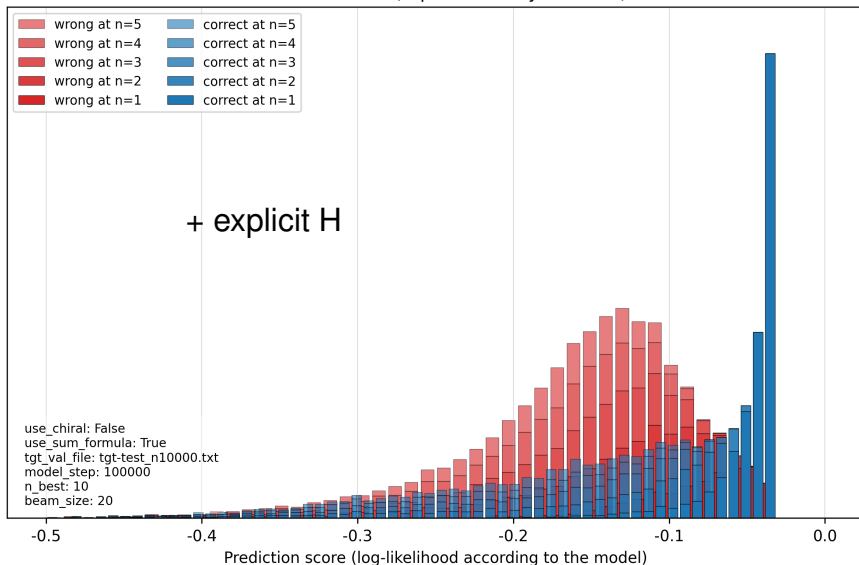
# Inferred SMILES (top-1 accuracy: 65.01%)



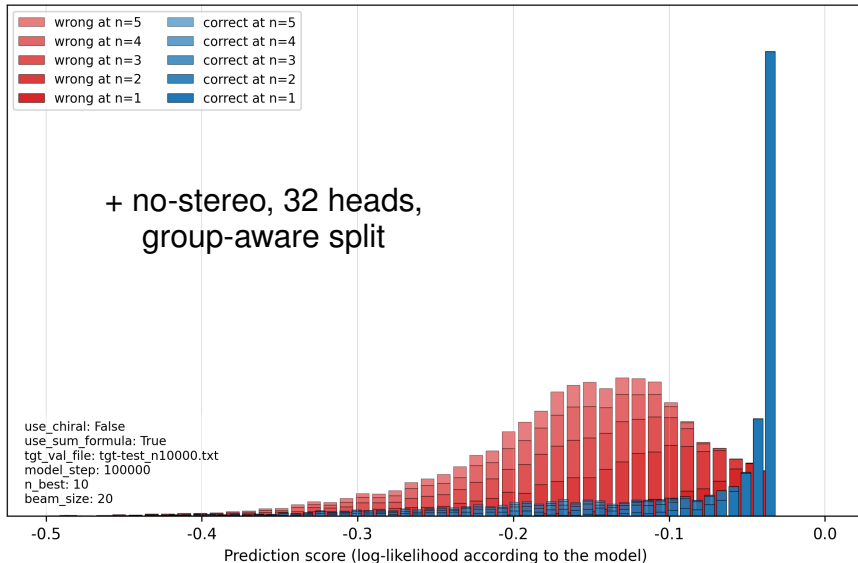
# Inferred SMILES (top-5 accuracy: 81.56%)



## Inferred SMILES (top-5 accuracy: 79.56%)



# Inferred SMILES (top-5 accuracy: 78.21%)



# Multitask ML from 1D NMR from Hu et al. [5]

## Model Inputs & Outputs

- Substructure-to-Structure:
  - *Input*: Binary vector (957 substructures)
  - *Output*: SMILES string (tokenized)
- Multitask Model:
  - *Input*: <sup>1</sup>H NMR (28,000-point interpolated), <sup>13</sup>C NMR (80-bin)
  - *Output*: SMILES string + substructure probabilities

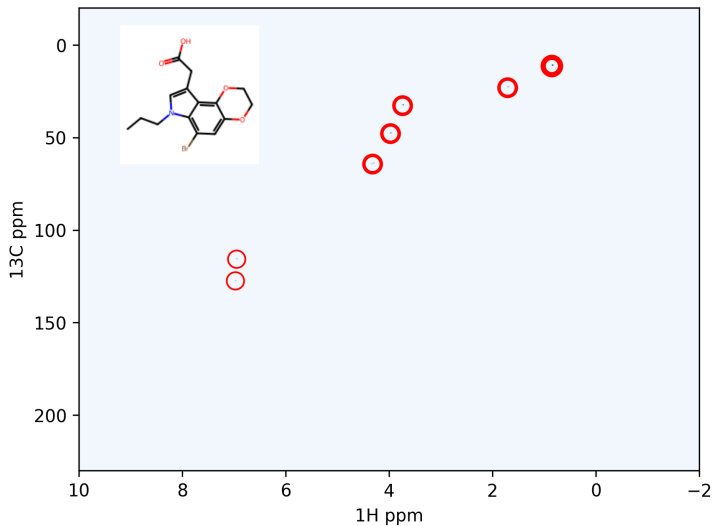
## Dataset

- ~3.1M molecules total (GDB-17 + SpectraBase)
- ~143k with simulated NMR spectra (SpectraBase)

## Accuracy

- Simulated NMR (test set): 69.6% (top-15 exact match)
- Experimental NMR (106 spectra): 33.0% (from Huang et al. [4])

HSQC



HSQC peaklist + H/C multiplets  $\rightarrow$  SMILES / functional groups



# HSCQ data in Alberts et al. [1]

```
hsqc_peaks = [  
    {  
        "13C_centroid": 11.251181374044274,  
        "13C_max": 11.98383639612529,  
        "13C_min": 10.518526351963258,  
        "1H_centroid": 0.8571500560320375,  
        "1H_max": 0.9073983348931892,  
        "1H_min": 0.8069017771708858,  
        "nH": 3.0  
    },  
    ...  
]
```

# HSCQ data in Alberts et al. [1]

```
c_peaks = [  
  {  
    "delta (ppm)": 173.71330825674195,  
    "integral": 0.0009437201155302,  
    "intensity": 0.0502866931546309,  
    "width (ppm)": 0.0119373346904897  
  },  
  ...  
]
```

# HSCQ data in Alberts et al. [1]

```
h_peaks = [  
  {  
    "category": "m",  
    "centroid": 6.9619175681911205,  
    "delta": 6.965392578832866,  
    "j_values": null,  
    "nH": 2,  
    "rangeMax": 6.988256203901225,  
    "rangeMin": 6.937767800903287  
  },  
  ...  
]
```

thx.

# References I

- [1] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. “Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry”. In: *NeurIPS 2024 Datasets and Benchmarks Track*. 2024.  
[10.48550/arXiv.2407.17492](https://arxiv.org/abs/2407.17492).
- [2] Daniel Lowe. *Chemical reactions from US patents (1976–Sep2016)*. figshare, 2017.  
[10.6084/m9.figshare.5104873.v1](https://figshare.com/figures/articles/10.6084/m9.figshare.5104873.v1).
- [3] Mestrelab Research. *MestReNova 15.1 Manual*. Mestrelab Research, 2024.

## References II

- [4] 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”. In: *Chemical Science* 12.46 (2021), pp. 15329–15338. [10.1039/d1sc04105c](https://doi.org/10.1039/d1sc04105c).
- [5] Frank Hu, Michael S. Chen, Grant M. Rotskoff, Matthew W. Kanan, and Thomas E. Markland. “Accurate and Efficient Structure Elucidation from Routine One-Dimensional NMR Spectra Using Multitask Machine Learning”. In: *ACS Central Science* 10.11 (2024), pp. 2162–2170. [10.1021/acscentsci.4c01132](https://doi.org/10.1021/acscentsci.4c01132).

