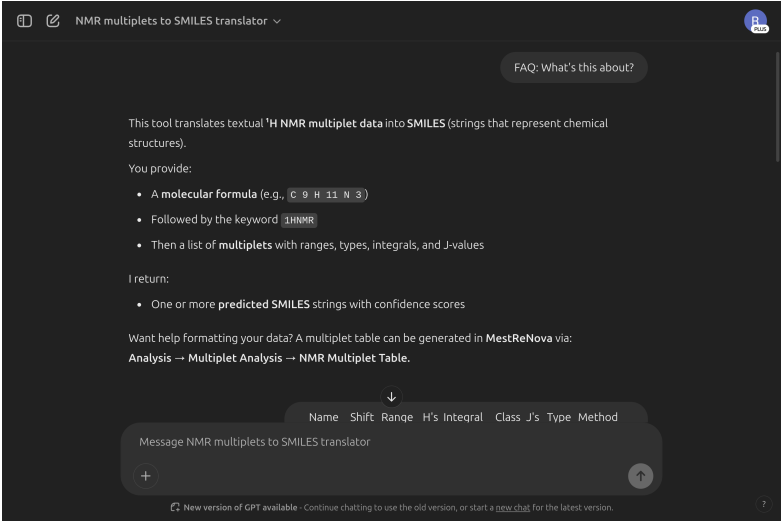
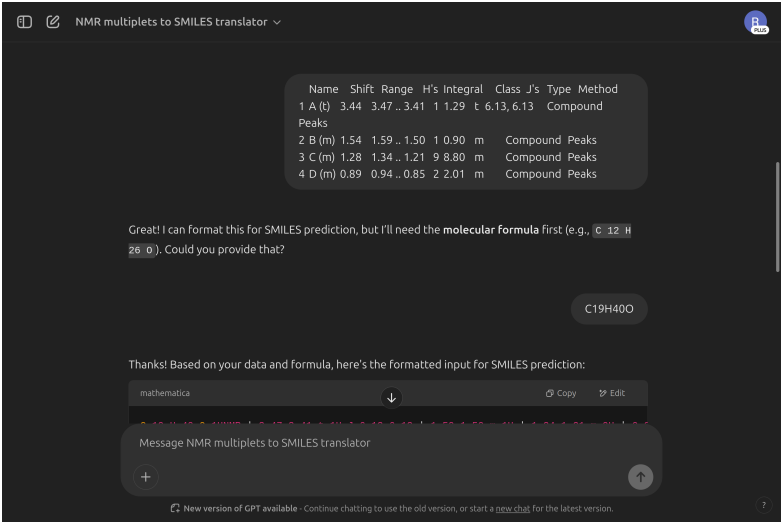
Molecular structure inference from molecular spectra

Custom ChatGPT interface to Alberts et al. [1]



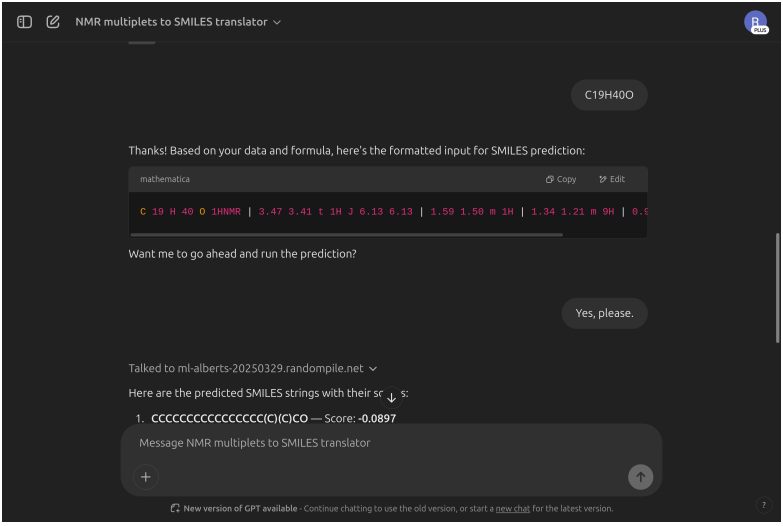
https://chatgpt.com/g/g-67e7f80c685881918d42729048cd2985-nmr-multiplets-to-smiles-translator

Custom ChatGPT interface to Alberts et al. [1]



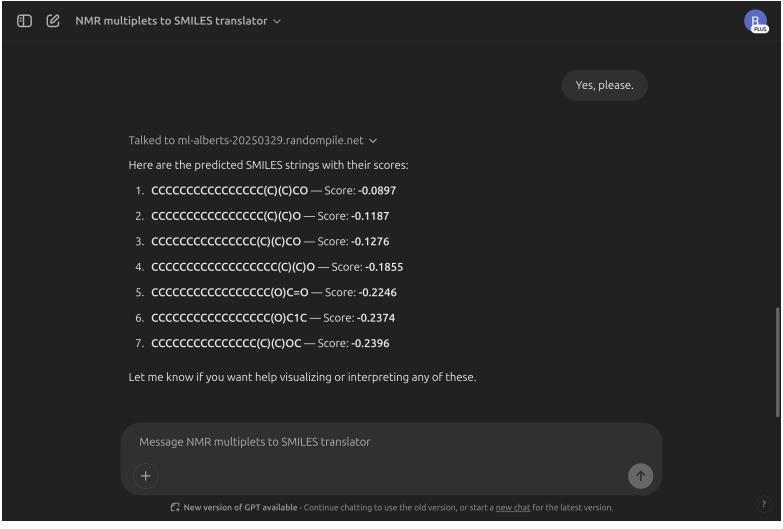
https://chatgpt.com/g/g-67e7f80c685881918d42729048cd2985-nmr-multiplets-to-smiles-translator

Custom ChatGPT interface to Alberts et al. [1]



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

Custom ChatGPT interface to Alberts et al. [1]



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

Training dataset and transformers model

- **Dataset:** USPTO reaction dataset[1]
  - 1,435,481 reactions
  - Realistic molecules & common chemicals
- **Unique molecules:**
  - Initially: 1,675,439
  - Filter: 5 < # 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:** ^{1}H-NMR, IR, ^{13}C-NMR, HSQC-NMR, MS/MS
- "... **deuterated chloroform** as solvent. Default settings were used"
- **Representation:** SMILES

[1] 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)

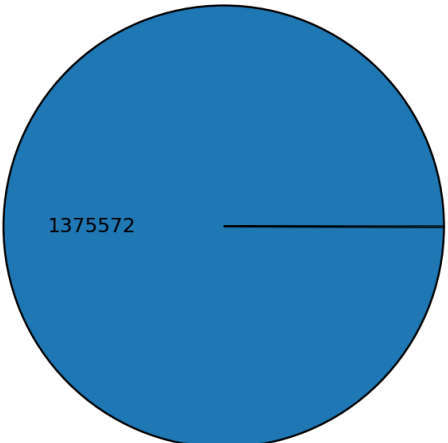
- `[CH3] [c] 1 [cH] [c] 2 [c] ( [cH] [c] 1 [NH2] ) [N] ( [CH] ( [CH3] ) [CH3] ) [CH2] [CH2] [C] 2 ( [CH3] ) [ CH3 ]`

...

Technical note (new)

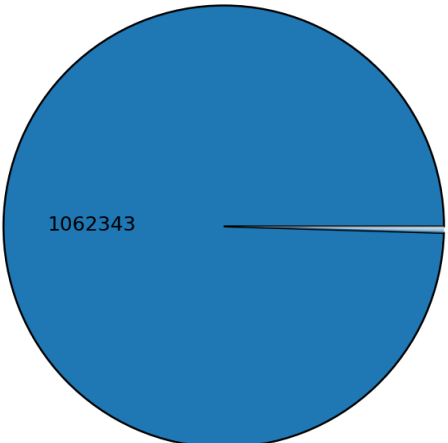
- Molecules from the same patent are more similar
- Thus: samples are not independent
- The train/test dataset split should preserve groups (patents)
- But: the bipartite graph smiles–patents is (almost) fully connected
- Heuristic: disconnect top nodes until largest component is small

Top-100 connected components (#nodes)



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

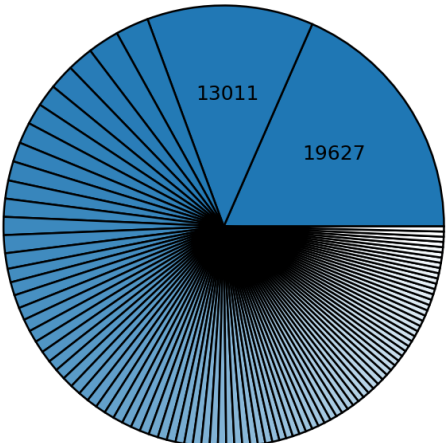
Top-100 connected components (#nodes)



#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**, $\sim$ 31M parameters
- 582'903 training samples, batch size 4k
- Trained on NVIDIA 1$\times$A10 GPU[2] 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

[2]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

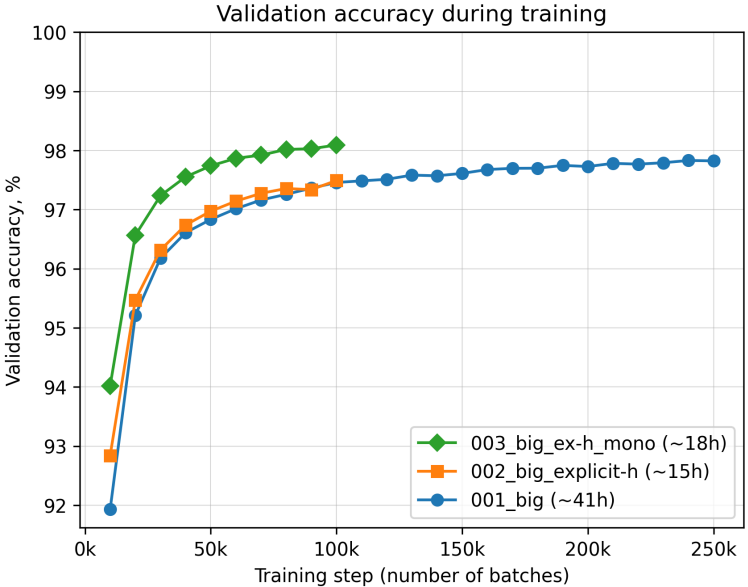
- How to generate all structures compatible with a given spectrum?
  - Molecules with indistinguishable spectra
  - Mixture of molecules from mixed spectra
- The spectrum $\approx$ a mixture of molecular fragments
- The model should have a built-in sense for additivity

Evaluation

Validation accuracy during training –

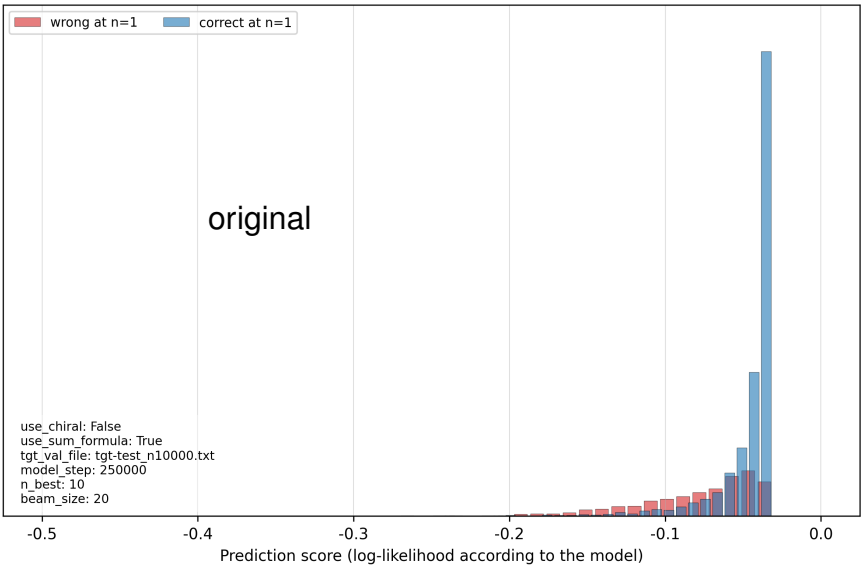
next-token prediction accuracy
over the unused validation dataset
with “teacher forcing”

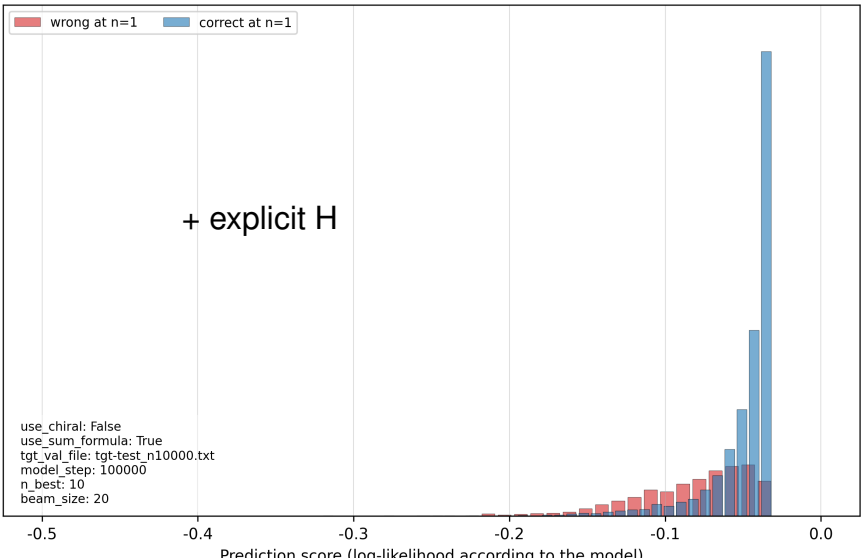
– indicates training progress.

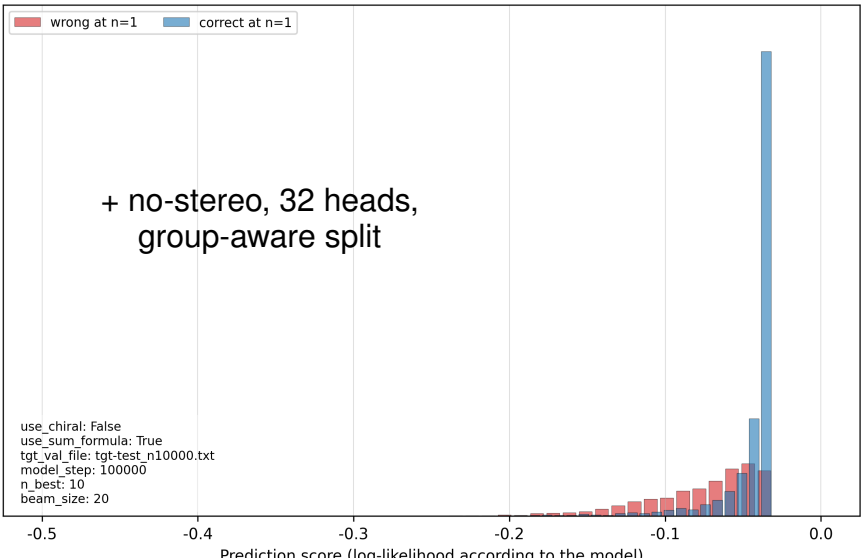

Validation accuracy during training
Validation accuracy, %
Training step (number of batches)
100
99
98
97
96
95
94
93
92
0k
50k
100k
150k
200k
250k
003_big_ex-h_mono (~18h)
002_big_explicit-h (~15h)
001_big (~41h)

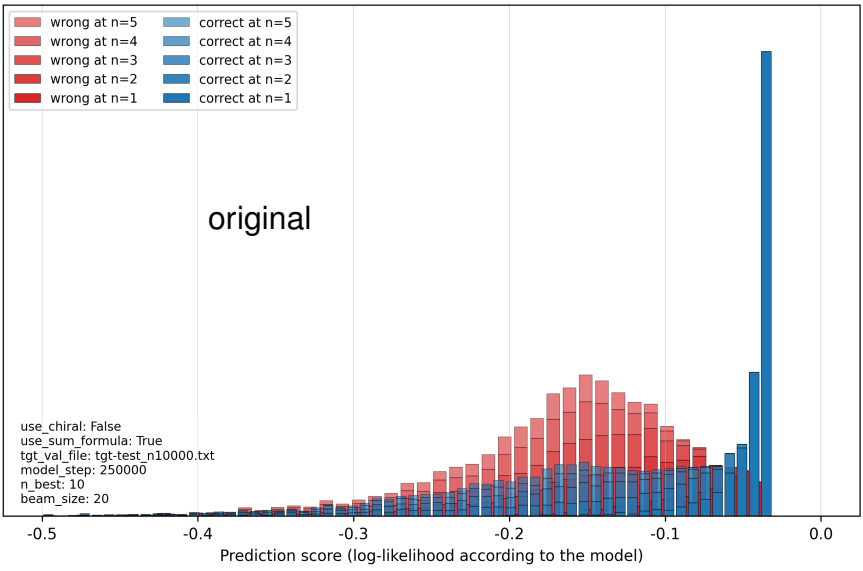
Evaluation

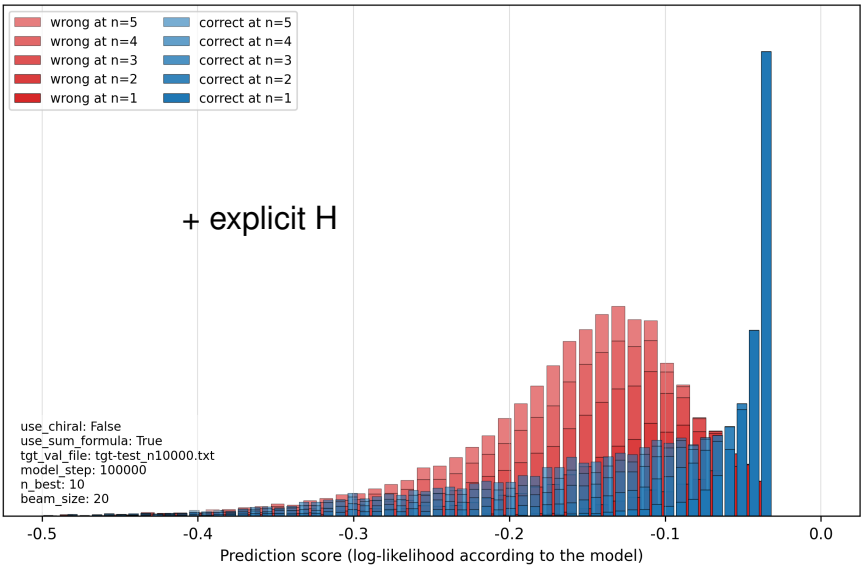
top-n accuracy of three models

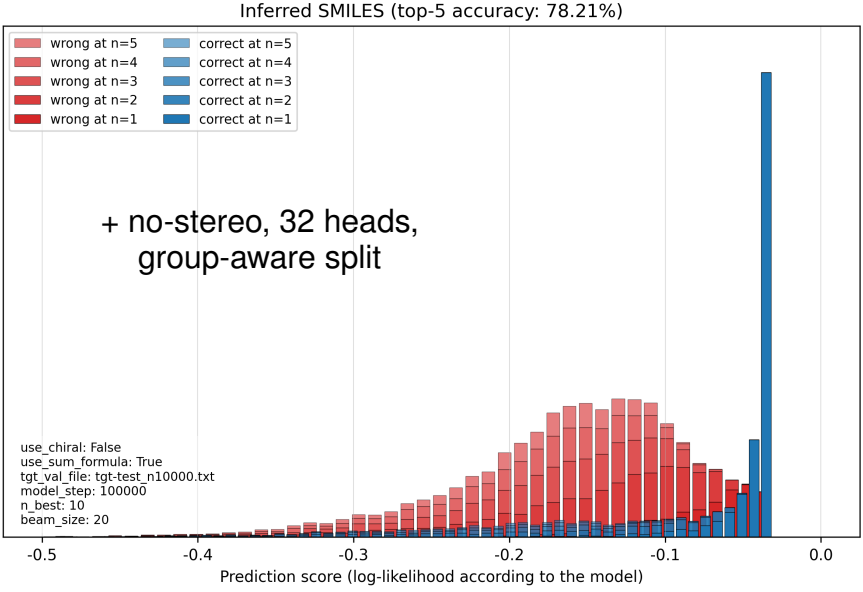

Inferred SMILES (top-1 accuracy: 71.02%)
wrong at n=1
correct at n=1
original
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n10000.txt
model_step: 250000
n_best: 10
beam_size: 20
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


Inferred SMILES (top-1 accuracy: 67.54%)
wrong at n=1
correct at n=1
+ explicit H
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n10000.txt
model_step: 100000
n_best: 10
beam_size: 20
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


Inferred SMILES (top-1 accuracy: 65.01%)
wrong at n=1
correct at n=1
+ no-stereo, 32 heads,
group-aware split
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n10000.txt
model_step: 100000
n_best: 10
beam_size: 20
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


Inferred SMILES (top-5 accuracy: 81.56%)
wrong at n=5
wrong at n=4
wrong at n=3
wrong at n=2
wrong at n=1
correct at n=5
correct at n=4
correct at n=3
correct at n=2
correct at n=1
original
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n10000.txt
model_step: 250000
n_best: 10
beam_size: 20
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


Inferred SMILES (top-5 accuracy: 79.56%)
wrong at n=5
wrong at n=4
wrong at n=3
wrong at n=2
wrong at n=1
correct at n=5
correct at n=4
correct at n=3
correct at n=2
correct at n=1
+ explicit H
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n10000.txt
model_step: 100000
n_best: 10
beam_size: 20
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


Inferred SMILES (top-5 accuracy: 78.21%)
wrong at n=5
wrong at n=4
wrong at n=3
wrong at n=2
wrong at n=1
correct at n=5
correct at n=4
correct at n=3
correct at n=2
correct at n=1
+ no-stereo, 32 heads,
group-aware split
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n10000.txt
model_step: 100000
n_best: 10
beam_size: 20
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)

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:* 1H NMR (28,000-point interpolated), 13C NMR (80-bin)
  - *Output:* SMILES string + substructure probabilities

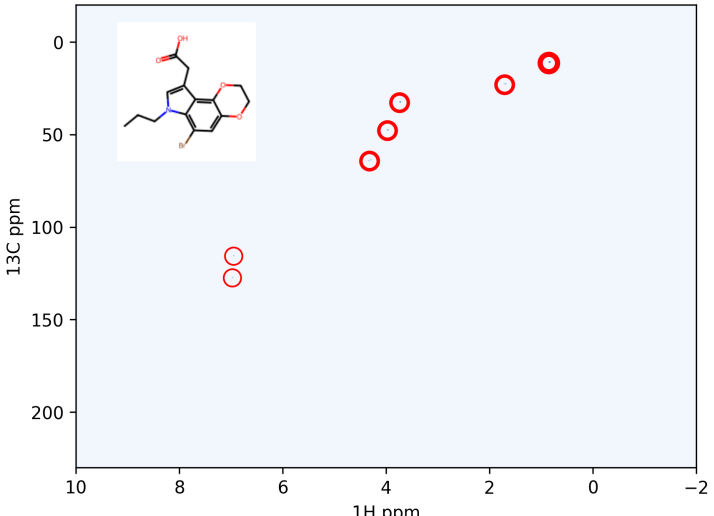
Dataset

- $\sim$3.1M molecules total (GDB-17 + SpectraBase)
- $\sim$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.

[2] Daniel Lowe. *Chemical reactions from US patents (1976–Sep2016)*. figshare, 2017. 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.

[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.

this slide is intentionally left blank