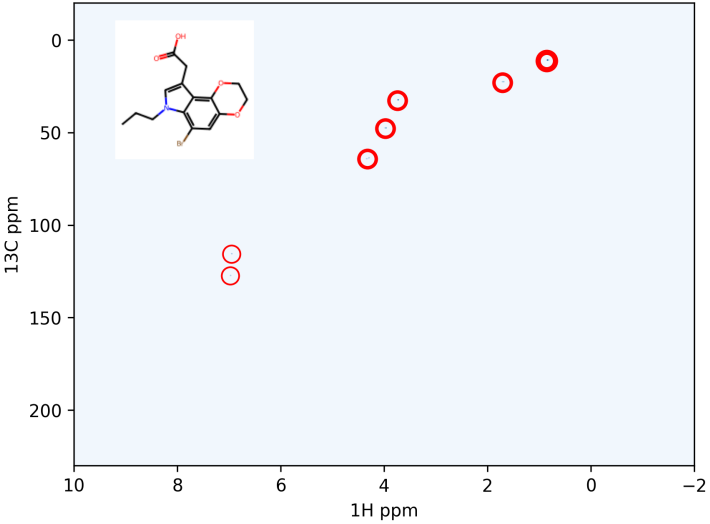
Molecular structure inference from molecular spectra

and vice versa

HSQC



HSQC peaklist + H/C multiplets → SMILES / functional groups

HSQC 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
 },
 ...
]
```

HSQC data in Alberts et al. [1]

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

HSQC 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
 },
 ...
]
```

HSQC data in Alberts et al. [1]

Issue: even accounting for multiplicity,

- HSQC peaks don't always correspond to C/H peaks
- C peaks don't always correspond to C atoms
- H peaks don't always correspond to H atoms

Reflex: train a structure-to-NMR model to disambiguate the peaks

Major assumption

Assumption:

- One atom $=$ one peak

In other words (on the NMR time-scale):

- Single chemical environment per C/H atom
- The energy landscape is not multimodal ($\approx$)
- No substantially distinct conformers (?)
- But rotamers are OK (?)

$\rightsquigarrow$ Fewer long-range effects

Major assumption

Assumption:

- One atom $=$ one peak

In other words (on the NMR time-scale):

- Single chemical environment per C/H atom
- The energy landscape is not multimodal ($\approx$)
- No substantially distinct conformers (?)
- But rotamers are OK (?)

$\rightsquigarrow$ Fewer long-range effects

Major assumption

Assumption:

- One atom $=$ one peak

In other words (on the NMR time-scale):

- Single chemical environment per C/H atom
- The energy landscape is not multimodal ($\approx$)
- No substantially distinct conformers (?)
- But rotamers are OK (?)

$\rightsquigarrow$ Fewer long-range effects

x:
molecule representation

Atom encoder

- Encodes RDKit atom object into a feature vector:
  - one-hot: atomic number (Z) from 1 to 118
  - one-hot: formal charge
  - one-hot: chiral tag (R/S/unspecified)
  - binary: chirality possible (based on RDKit flags)
  - binary: is aromatic
  - one-hot: total number of hydrogens (explicit + implicit)
  - scalar: isotope number
  - binary: has explicit Hs (i.e., no implicit Hs)
  - one-hot: number of implicit hydrogens

Bond encoder

- Encodes RDKit bond object into a feature vector:
  - one-hot: bond type (single, double, etc., as float)
  - binary: is conjugated (part of resonance structure)
  - binary: is aromatic
  - one-hot: stereo configuration (e.g., E/Z)
  - one-hot: bond direction (e.g., up/down in wedge bonds)

y:
peak targets
(MestReNova simulations [1])

Carbon peak targets

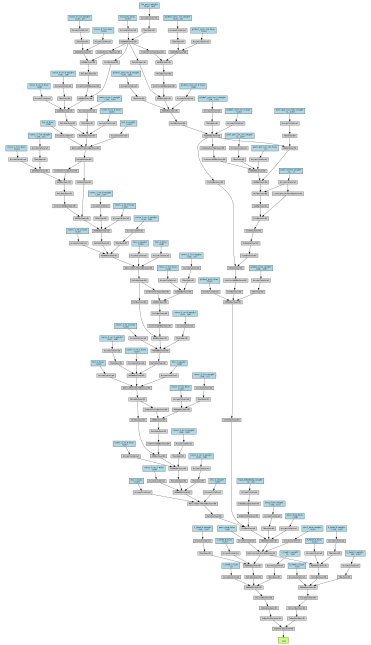
- Extracted from the C-NMR data of Alberts et al. [1]
- For each carbon peak:
  - `delta_ppm` $\rightarrow$ chemical shift
  - `integral` $\rightarrow$ relative signal area
- Stored as lists (but evaluated as sets):
  - `c_shifts`: list of float shifts
  - `c_integral`: list of float integrals
- Peaks are shuffled to remove positional bias (just in case)

Hydrogen peak targets

- Extracted from H-NMR data of Alberts et al. [1]
- For each hydrogen peak:
  - `delta` → chemical shift
  - `j_values` → J-coupling constants
  - `category` → multiplet category label
  - `nH` → number of equivalent Hs in the signal
- Stored as lists (but evaluated as sets):
  - `h_shifts`, `h_j_values`, `h_categories`, `h_nH`
- Peaks are shuffled to remove positional bias (just in case)

Molecule $\rightarrow$ NMR model

300k trainable parameters



Input projection

- Each atom starts with a feature vector of size `in_dim`
- We apply a linear projection `in_dim` $\rightarrow$ `hidden_dim`
- Purpose: embed atom features into a shared hidden space
- Edges have features, too

Local GNN encoder

- Stack of 4 GINEConv layers with BatchNorm, ReLU, and Dropout
- Each layer updates atom features based on neighbors:

$$h_i \leftarrow \text{MLP}\left(h_i + \sum_{j \in \mathcal{N}(i)} \phi(h_j, e_{ij})\right)$$

  where
  - $\text{MLP} = \text{Linear} \circ \text{ReLU} \circ \text{Linear}$
  - $\phi(h_j, e_{ij}) = \text{edge-aware message} = h_j + \text{Linear}(e_{ij})$
- Learns local chemical environments for each atom

Global: GNN, SAGPooling, mean pooling

- Apply another GINEConv layer to initial features
- Produces global-aware atom features before pooling
- This path runs in parallel with the local GNN
- Learnable pooling that drops 50% of the least important nodes
- Uses self-attention scores to retain key atoms
- Aggregates pooled node features into one vector per molecule:

$$\text{global context} = \frac{1}{|G|} \sum_{i \in G} x_i$$

- A “holistic” representation, projected using a linear layer

Global context broadcast

- The global vector is given to each atom in the original graph
- Now, every atom knows something about the full molecule
- Enables late fusion of global + local info

Atom type embedding and feature fusion

- Atom type is embedded (e.g., C or H)
- Embedding is learnable and provides type-specific context
- Other atoms are masked
- Probably redundant
- Concatenate:
  - Local GNN output
  - Global context
  - Atom type embedding
- Project to `hidden_dim` and normalize

Prediction Heads and Output Processing

- Two heads:
  C: shift, integral
  H: shift, J-values, multiplet category logits
- Only valid predictions are assigned using type masks
- Softplus ensures integrals and J-values are positive
- Carbon integrals are normalized to sum to *one*
- Final output is a dictionary with all per-atom predictions

Prediction Heads and Output Processing

- Two heads:
  C: shift, integral
  H: shift, J-values, multiplet category logits
- Only valid predictions are assigned using type masks
- Softplus ensures integrals and J-values are positive
- Carbon integrals are normalized to sum to *one*
- Final output is a dictionary with all per-atom predictions

Loss function

Total carbon loss

- Compares predicted and target 13C peak sets using Sinkhorn "earth mover distance" (EMD) – smooth & fast optimal transport of histograms with entropy regularization[1]
- Each peak has a chemical shift and an integral (used as weight)
- Peaks are grouped by molecule
- A barrier coordinate promotes matching only within molecule
- Integrals are normalized per graph to form proper distributions

[1] Feydy et al., 2019 [2]

Total hydrogen loss

- Total loss is a weighted sum of:
  - shift loss
  - J-coupling loss
  - category (multiplicity) loss
- Weights are defined in the training config
- All components are permutation-invariant and molecule-separated

Total hydrogen loss

- Total loss is a weighted sum of:
  - shift loss
  - J-coupling loss
  - category (multiplicity) loss
- Weights are defined in the training config
- All components are permutation-invariant and molecule-separated

Total hydrogen loss

Hydrogen shift loss

- Compares predicted and target 1H chemical shifts using EMD
- Each predicted shift is weighted uniformly
- Target shifts are weighted by the number of equivalent hydrogens
- Graph-wise separation is enforced using a barrier coordinate

Total hydrogen loss

J-coupling loss

- Compares predicted and target J-values using EMD
- Predicted J-values are flattened across all H atoms
- Target J-values are pooled per molecule
- Graph assignments are repeated to match flattened layout
- Weights are uniform and normalized per graph

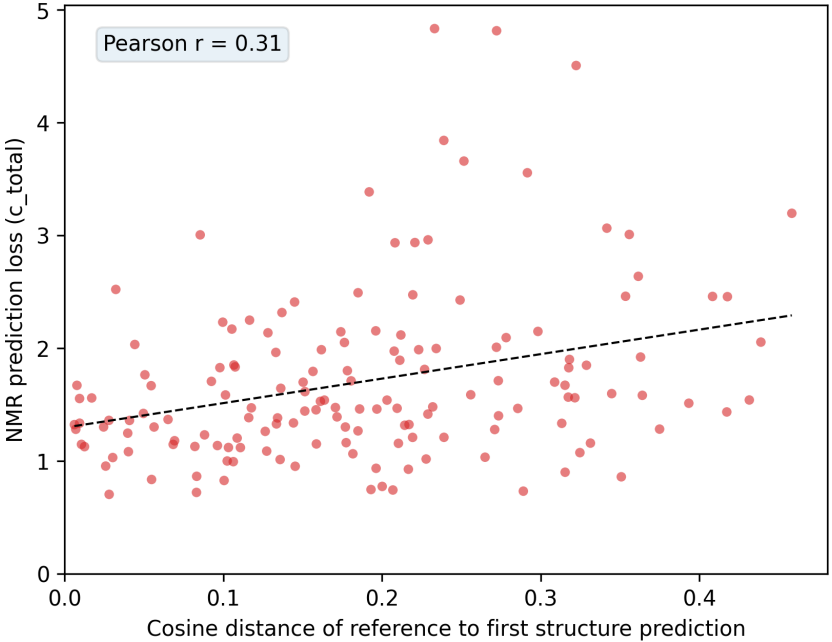
Total hydrogen loss

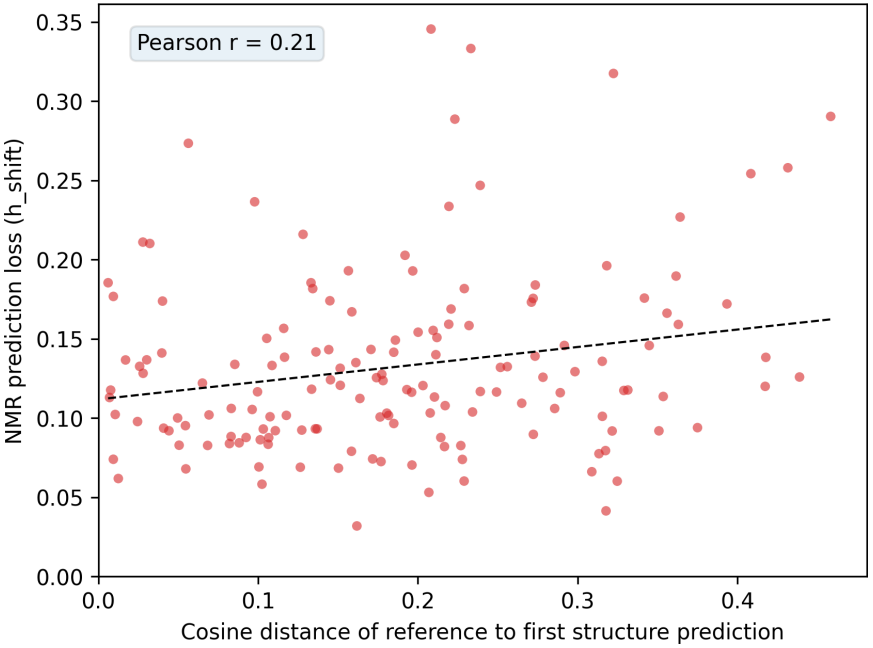
Category loss

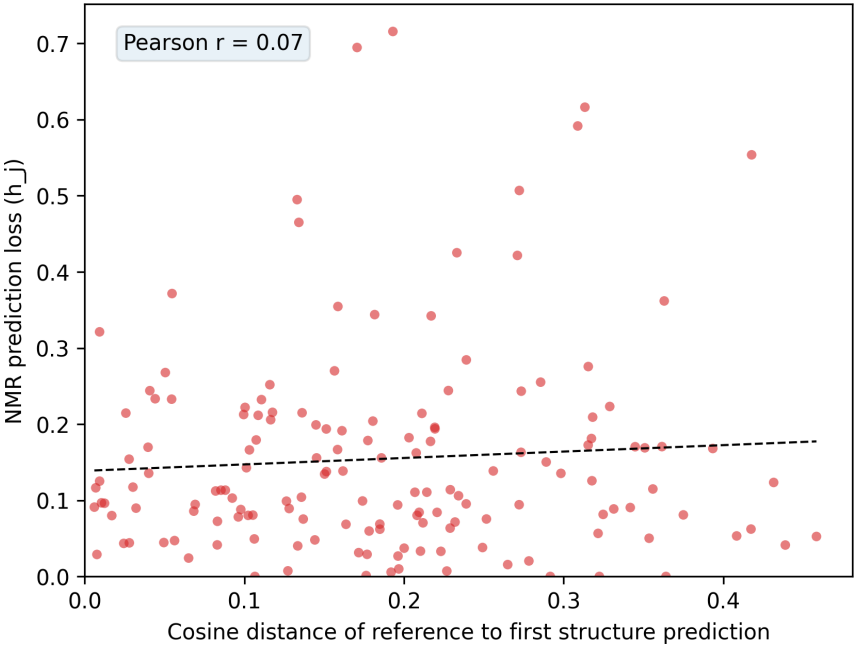
- Computes L1 loss between predicted and target class histograms
- For each molecule, the model predicts probabilities over multiplicity classes (softmax)
- Histograms are built by summing predictions per molecule
- Histograms are normalized to sum to one (per molecule)
- Class weights are applied to the L1 loss based on dataset stats

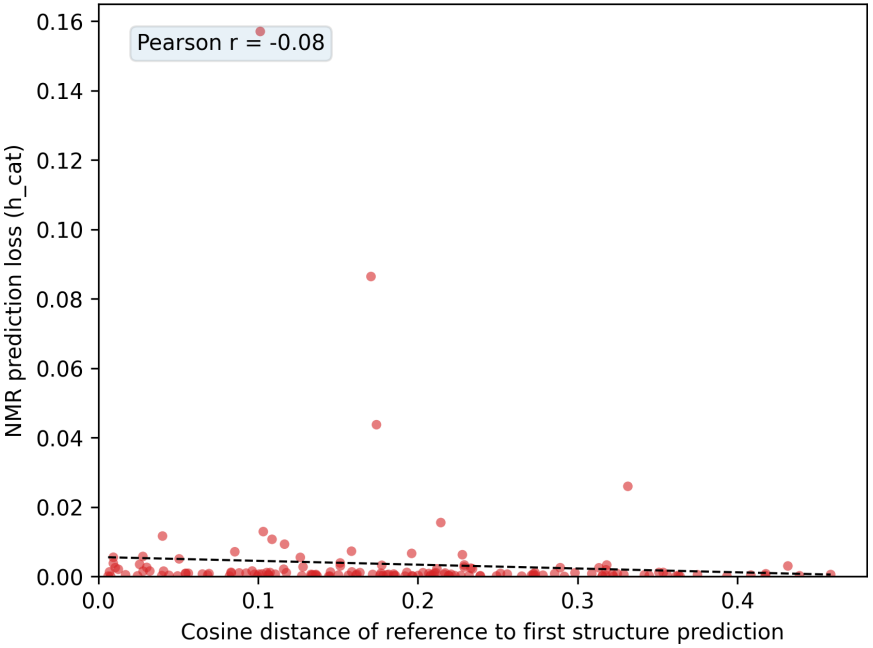
Evaluation

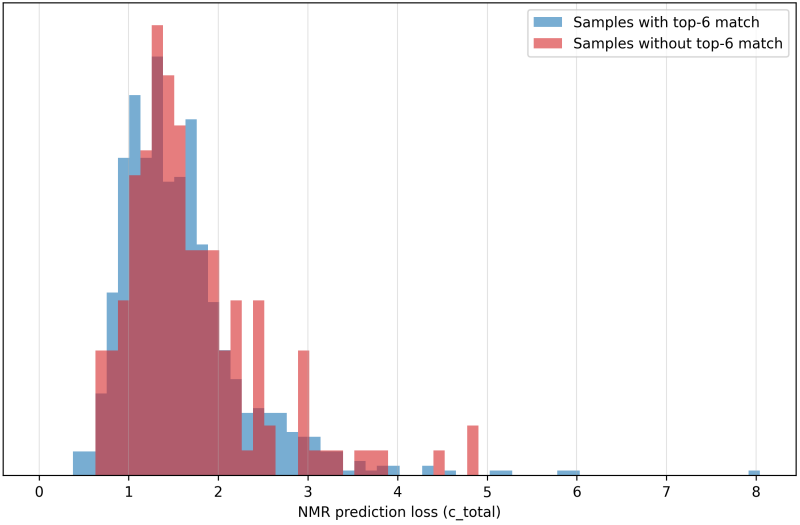
Structure-to-NMR loss
vs
NMR-to-Structure loss

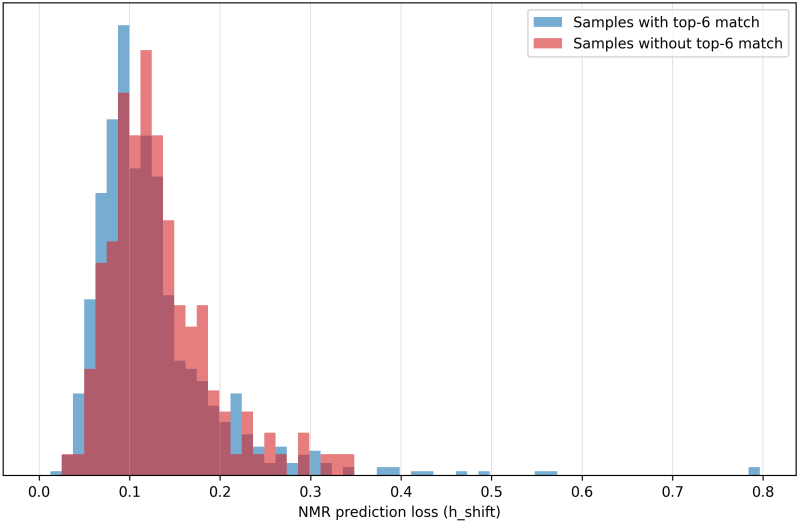

Pearson r = 0.31
NMR prediction loss (c_total)
Cosine distance of reference to first structure prediction
0
1
2
3
4
5
0.0
0.1
0.2
0.3
0.4

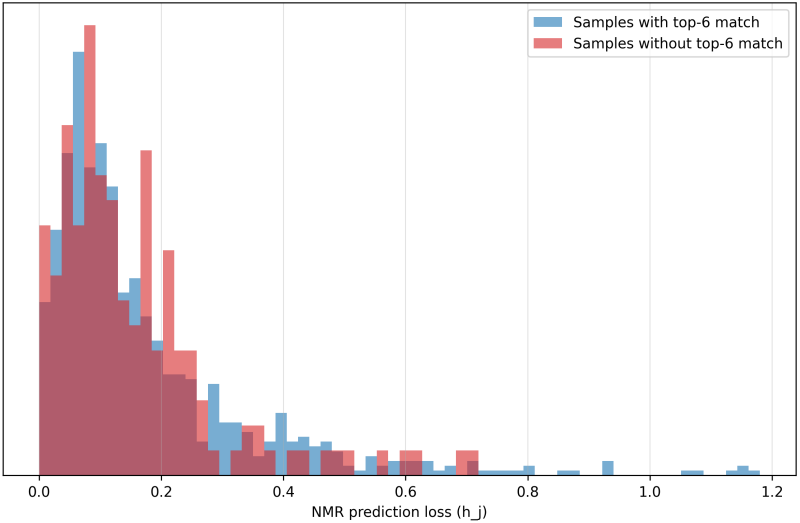

Pearson r = 0.21
NMR prediction loss (h_shift)
Cosine distance of reference to first structure prediction
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.0
0.1
0.2
0.3
0.4

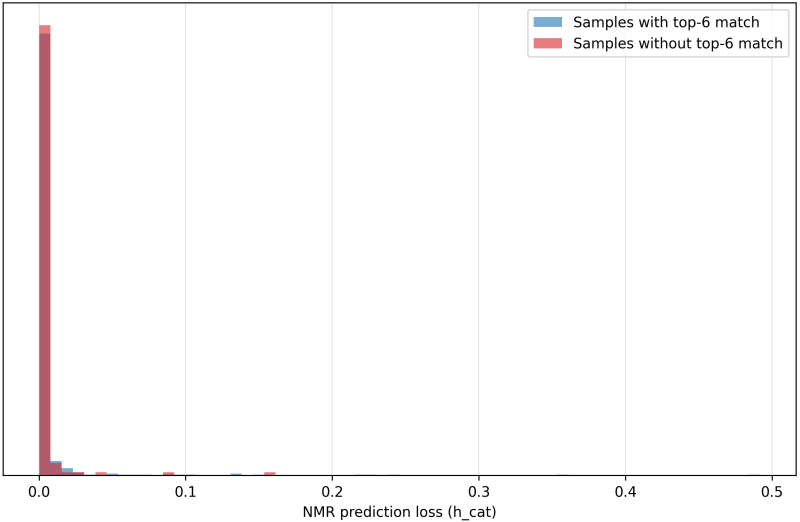

Pearson r = 0.07
NMR prediction loss (h_j)
Cosine distance of reference to first structure prediction
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.0
0.1
0.2
0.3
0.4


Pearson r = -0.08
NMR prediction loss (h_cat)
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.0
0.1
0.2
0.3
0.4
Cosine distance of reference to first structure prediction


Samples with top-6 match
Samples without top-6 match
0
1
2
3
4
5
6
7
8
NMR prediction loss (c_total)

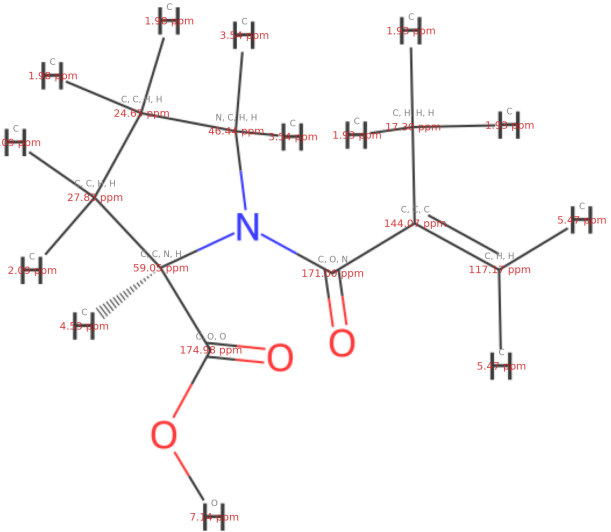

Samples with top-6 match
Samples without top-6 match
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
NMR prediction loss (h_shift)


Samples with top-6 match
Samples without top-6 match
0.0
0.2
0.4
0.6
0.8
1.0
1.2
NMR prediction loss (h_j)

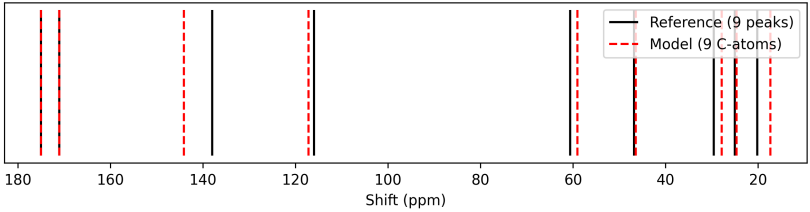

Samples with top-6 match
Samples without top-6 match
0.0
0.1
0.2
0.3
0.4
0.5
NMR prediction loss (h_cat)

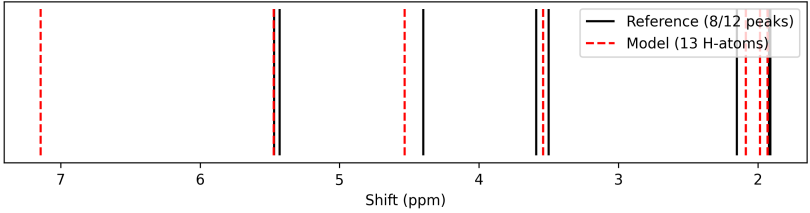
Examples

C=C(C)C(=O)N1CCC[C@@H]1C(=O)O

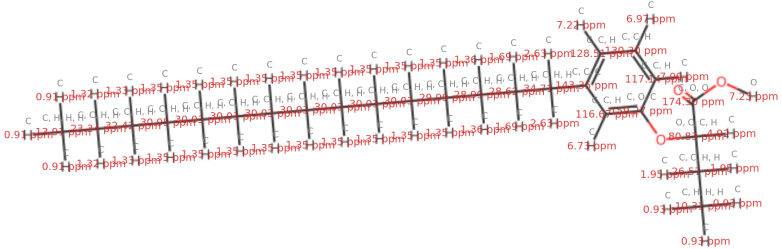


C=C(C)C(=O)N1CCC[C@@H]1C(=O)O

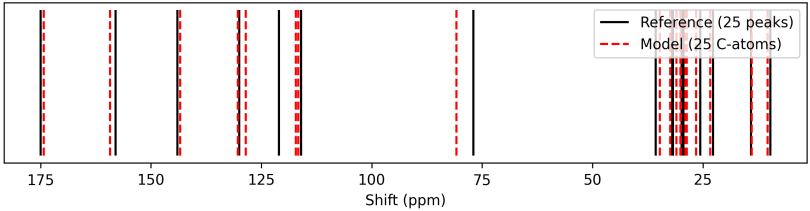


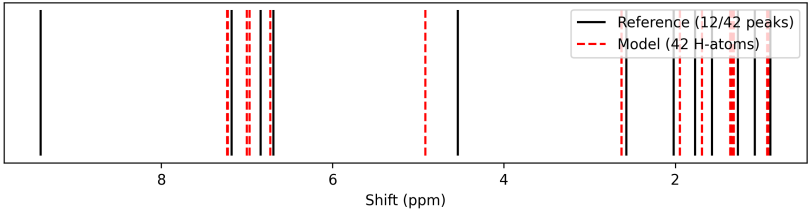


CCCCCCCCCCCCCCCc1cccc(OC(CC)C(=O)O)c1

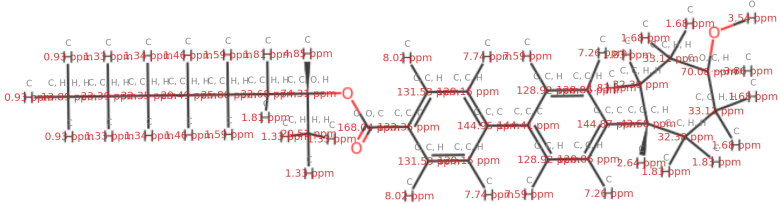


CCCCCCCCCCCCCCCc1cccc(OC(CC)C(=O)O)c1

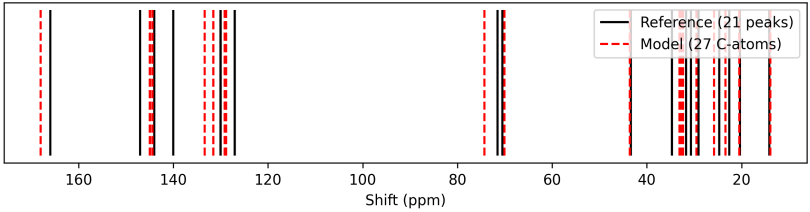


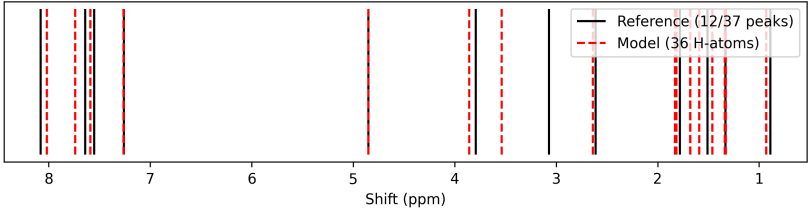


CCCCCC[C@@H](C)OC(=O)c1ccc(-c2ccc([C@@H]3CC[C@@H](O)CC3)cc2)cc1

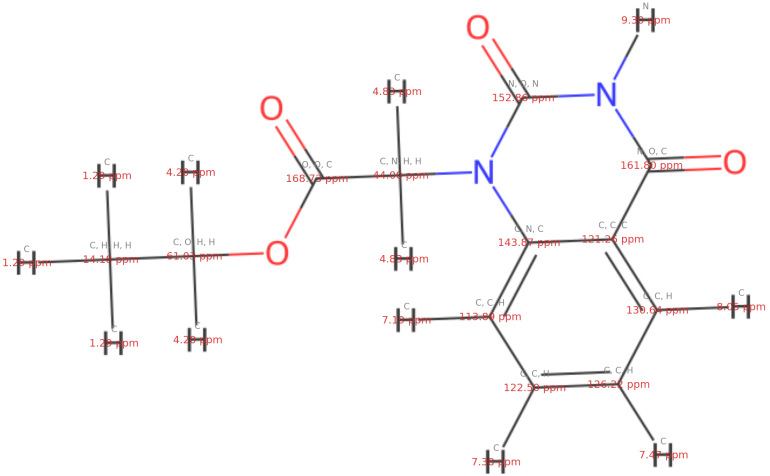


CCCCCC[C@@H](C)OC(=O)c1ccc(-c2ccc([C@@H]3CC[C@@H](O)CC3)cc2)cc1

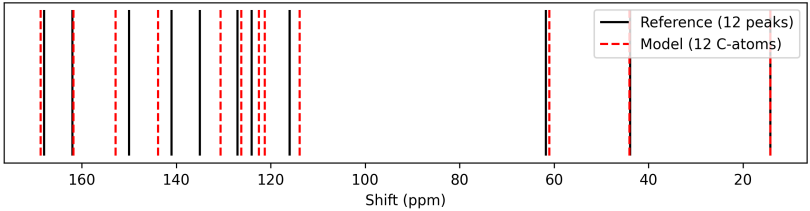


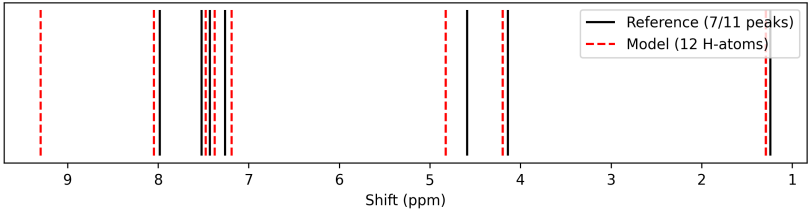


CCOC(=O)Cn1c(=O)[nH]c(=O)c2ccccc21

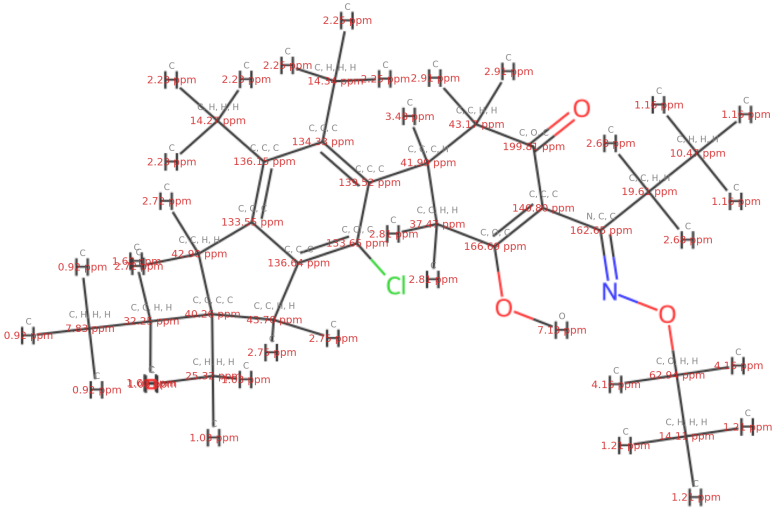


CCOC(=O)Cn1c(=O)[nH]c(=O)c2ccccc21

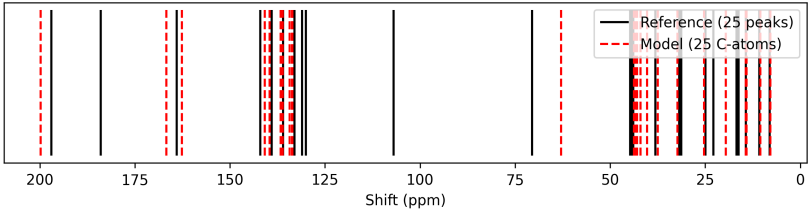


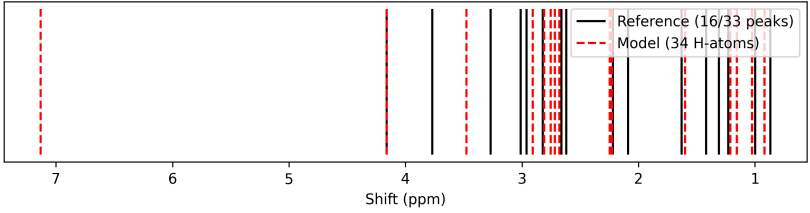


CCON=C(CC)C1=C(O)CC(c2c(C)c(C)c3c(c2Cl)CC(C)(CC)C3)CC1=O

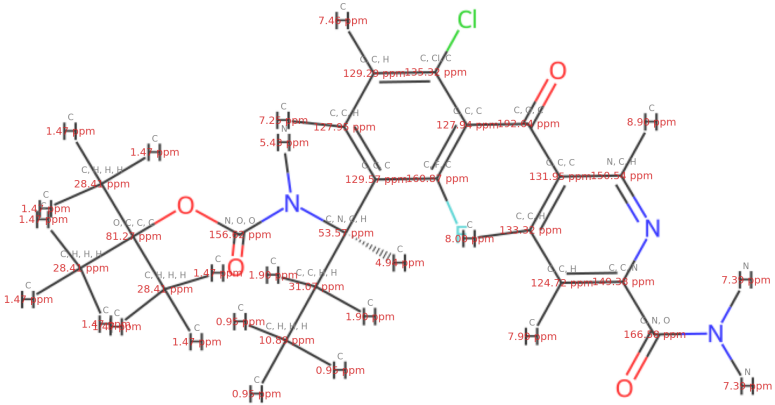


CCON=C(CC)C1=C(O)CC(c2c(C)c(C)c3c(c2Cl)CC(C)(CC)C3)CC1=O

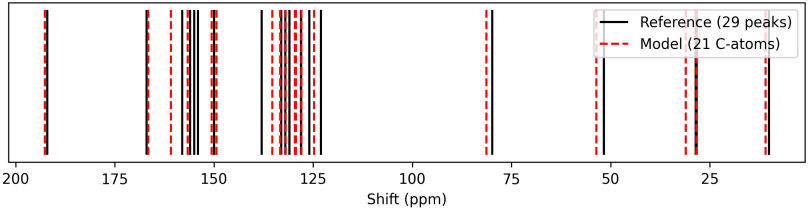


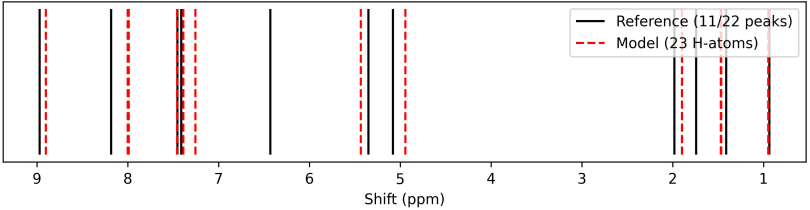


CC[C@@H](NC(=O)OC(C)(C)C)c1ccc(Cl)c(C(=O)c2ccc(C(N)=O)nc2)c1F

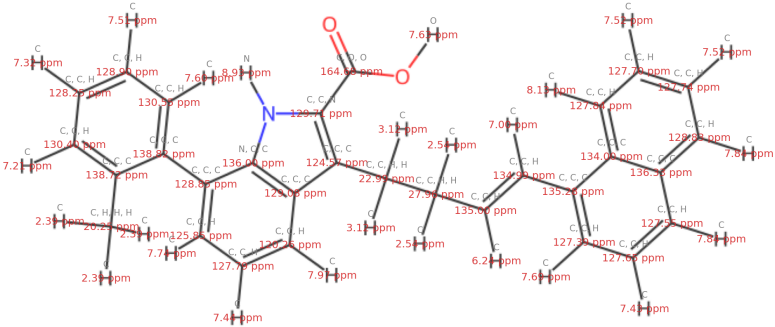


CC[C@@H](NC(=O)OC(C)(C)C)c1ccc(Cl)c(C(=O)c2ccc(C(N)=O)nc2)c1F

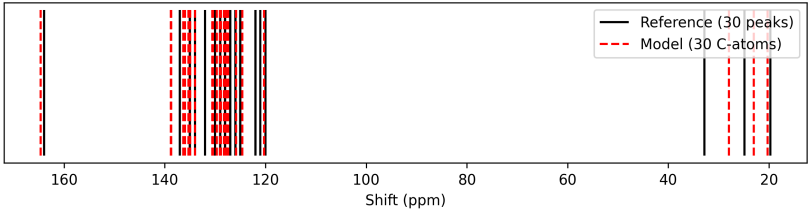


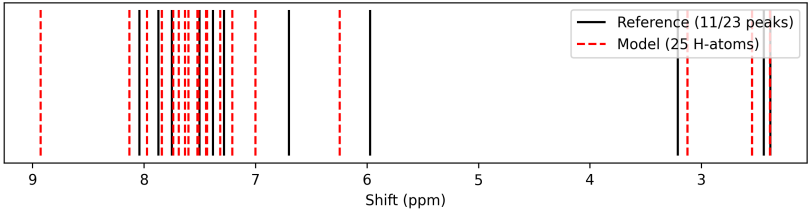


```
Cc1ccccc1-c1cccc2c(CCC=Cc3cccc4ccccc34)c(C(=O)O)[nH]c12
```

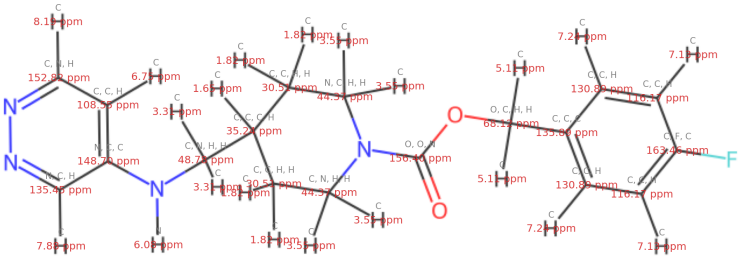


Cc1ccccc1-c1cccc2c(CCC=Cc3cccc4ccccc34)c(C(=O)O)[nH]c12

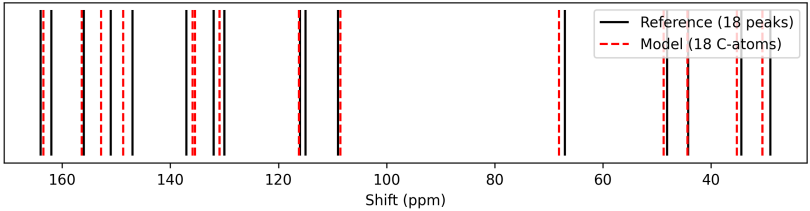


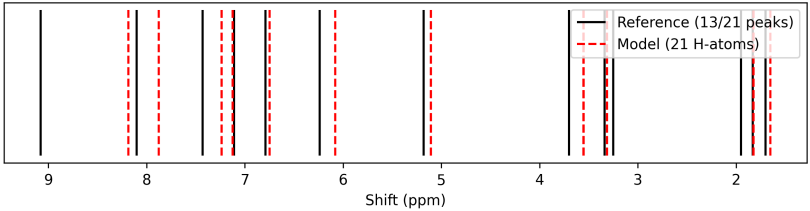


O=C(OCc1ccc(F)cc1)N1CCC(CNc2ccnnc2)CC1

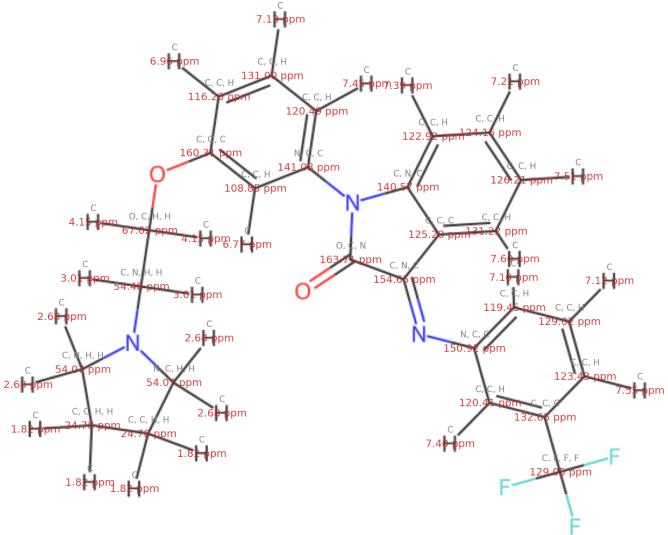


O=C(OCc1ccc(F)cc1)N1CCC(CNc2ccnnc2)CC1

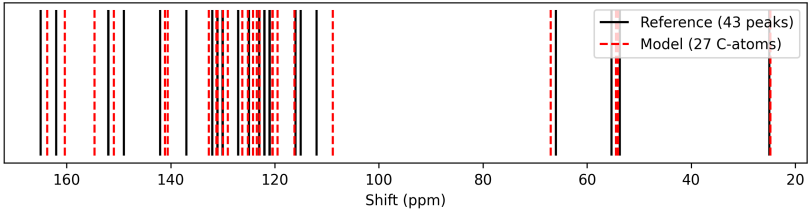


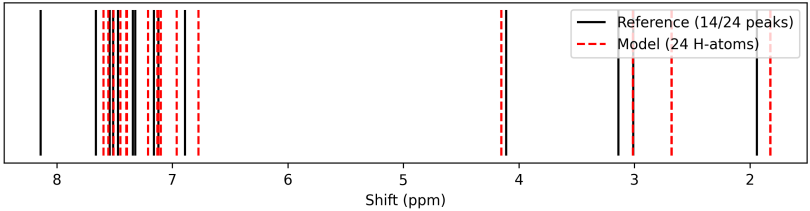


O=C1C(=Nc2cccc(C(F)(F)F)c2)c2ccccc2N1c1cccc(OCCN2CCCC2)c1

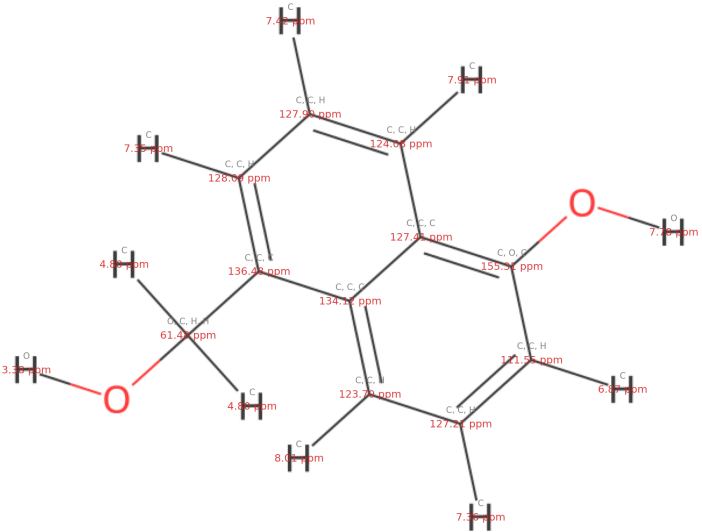


O=C1C(=Nc2cccc(C(F)(F)F)c2)c2ccccc2N1c1cccc(OCCN2CCCC2)c1

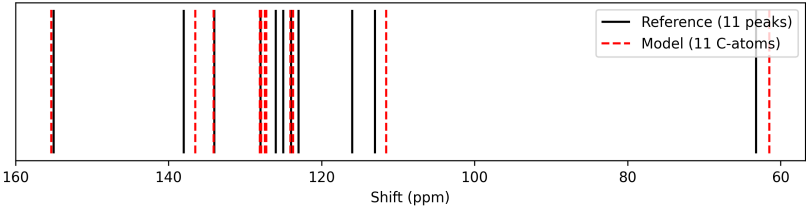


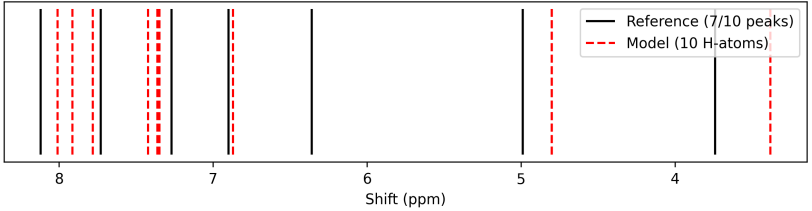


OCc1cccc2c(O)cccc12



OCc1cccc2c(O)cccc12





This ML model provides a simlified model of C/H NMR

Is it easier to invert?

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] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. “Interpolating between Optimal Transport and MMD using Sinkhorn Divergences”. In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 2681–2690.

this slide is intentionally left blank