Molecular structure inference from molecular spectra

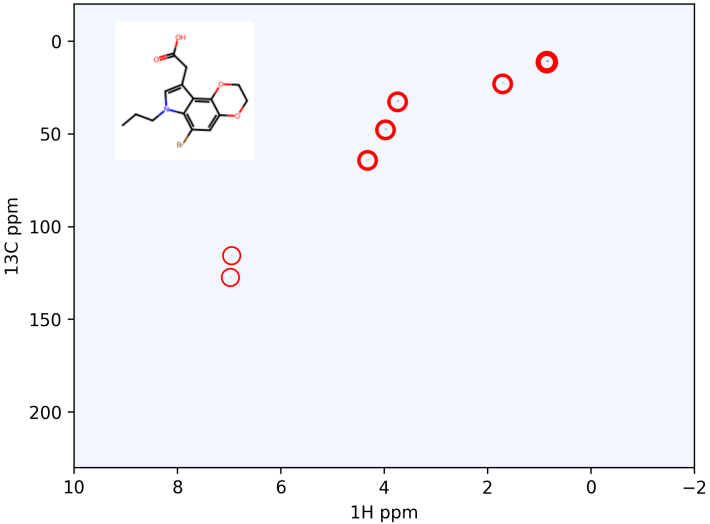
with LLMs

ROMAN ANDREEV (andreevr @ ethz ch)

HSQC $\rightarrow$ structure

based on

[1] Marvin Alberts et al. “Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry”. In: *NeurIPS 2024 Datasets and Benchmarks Track*. 2024. [url]


13C ppm
1H ppm
0
50
100
150
200
10
8
6
4
2
0
−2

HSQC data in Alberts et al. [1]

```
hsqc_peaks = [
 {
  "13C_centroid": 11.3,
  "13C_max": 12.0,
  "13C_min": 10.5,
  "1H_centroid": 0.857,
  "1H_max": 0.907,
  "1H_min": 0.807,
  "nH": 3.0
 },
 ...
]
```

HSQC data in Alberts et al. [1]

```
c_peaks = [
 {
  "delta (ppm)": 174,
  "integral": 0.000,
  "intensity": 0.0503,
  "width (ppm)": 0.0119
 },
 ...
]
```

HSQC data in Alberts et al. [1]

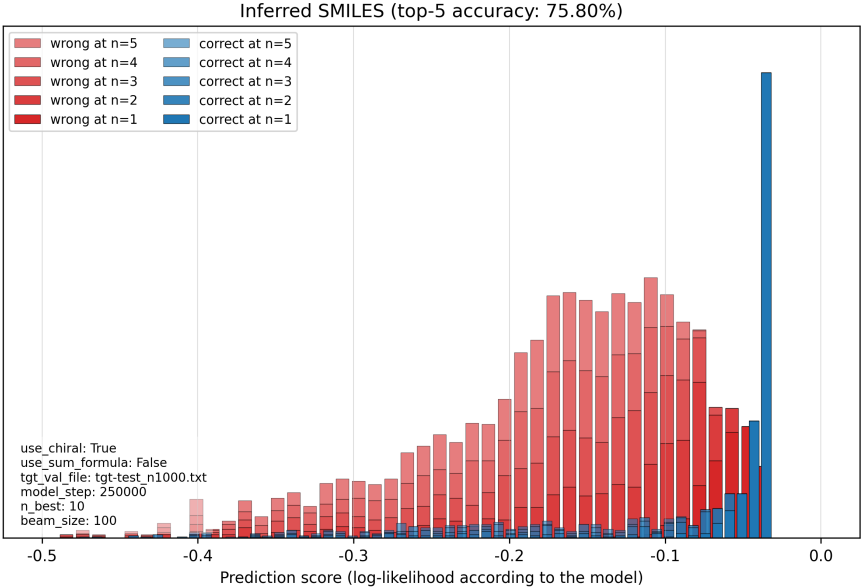
```
h_peaks = [
 {
  "category": "m",
  "centroid": 6.96,
  "delta": 6.97,
  "j_values": null,
  "nH": 2,
  "rangeMax": 6.99,
  "rangeMin": 6.94
 },
 ...
]
```

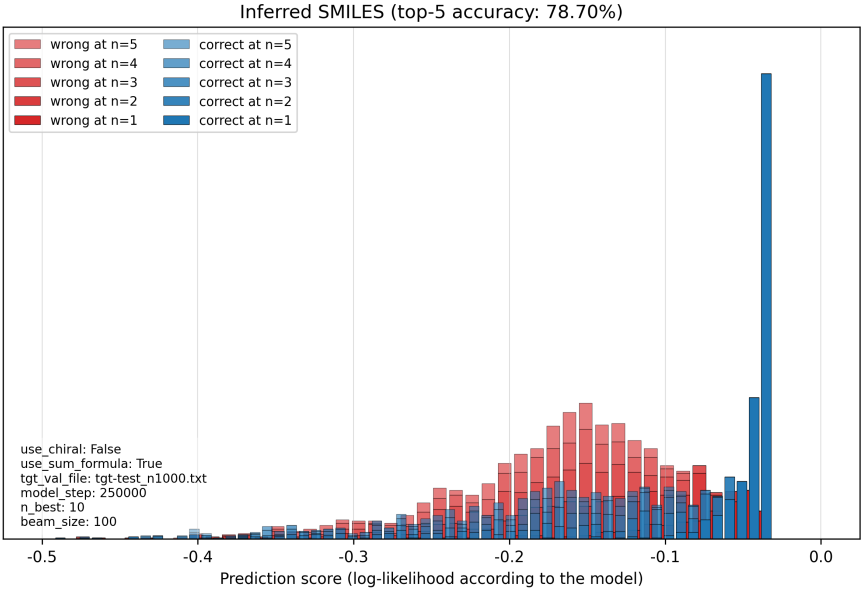
Modality	Formula	Format	Tokens	Top–1%	Top–5%	Top–10%
^{1}H NMR	✗	Center		38.29%	54.67%	58.43%
	✓	Center		53.34%	71.71%	75.09%
	✓	Adaptive		53.39%	71.84%	75.23%
	✓	Range		55.32%	73.59%	76.74%
^{1}H NMR (Aug.)	✓	Range		51.58%	70.52%	73.94%
^{1}H NMR (Ens.)	✓	Range		**57.99%**	**76.65%**	**80.04%**
^{13}C NMR	✗			37.21%	53.98%	57.45%
	✓			51.37%	70.74%	74.32%
^{13}C NMR (Aug.)	✓			49.02%	69.05%	72.90%
^{13}C NMR (Ens.)	✓			**53.91%**	**73.45%**	**77.72%**
^{1}H+^{13}C NMR	✗	Range	Own	56.88%	73.91%	76.89%
	✓	Range	Own	64.78%	81.74%	84.43%
	✓	Range	Shared	65.05%	82.07%	84.70%
^{1}H+^{13}C NMR (Aug.)	✓	Range	Shared	62.35%	80.15%	82.93%
^{1}H+^{13}C NMR (Ens.)	✓	Range	Shared	**66.99%**	**84.09%**	**86.59%**

Alberts et al. [2], Table 1

Alberts et al. [2] provided the code to train their models
(but not the weights),
which relies on the OpenNMT package [3]
(“neural machine translation”, vocab-to-vocab; now: Eole)

We closely reproduced some of their results, for example:


Inferred SMILES (top-5 accuracy: 75.80%)
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
use_chiral: True
use_sum_formula: False
tgt_val_file: tgt-test_n1000.txt
model_step: 250000
n_best: 10
beam_size: 100
-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: 78.70%)
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
use_chiral: False
use_sum_formula: True
tgt_val_file: tgt-test_n1000.txt
model_step: 250000
n_best: 10
beam_size: 100
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)

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

Interpretation:

- Peaks inferred from spectra – `"nmrAutoPeakPicking"`
- MestReNova simulates conformers (#peaks $\geq$ #atoms)
- From [1, §3.1]: The spectra were simulated using deuterated Chloroform as solvent. Default settings were used. For C-NMR spectra, H decoupled spectra were generated. [Use] in-built spectral analysis tools of MestreNova to annotate the spectra. For H-NMR spectra, the automultiplet analysis function yields1 a set of peaks, the type of each peak [..], and the normalized integration of the peak. [Same for] position and intensity of the peaks in the C-NMR spectra [and HSQC] spectra.

Abraham and Mobli [4], §8.3.1:

Once the energies of a number of minimized structures are calculated, the relative populations of these is given by the Boltzmann distribution (Equation (8.4)) where f is the fractional population of the higher energy state and ΔE is the energy difference ($E_a - E_{min}$):

$$f = \exp\left(-\Delta E / RT\right) \tag{8.4}$$

If chemical shifts are calculated for the nuclei in all of these structures, then a weighted average can be obtained which should be a better fit of the experimental data obtained at any given temperature (see also Section 8.5). This procedure is used in the accompanying `NMRPredict` *software package.*

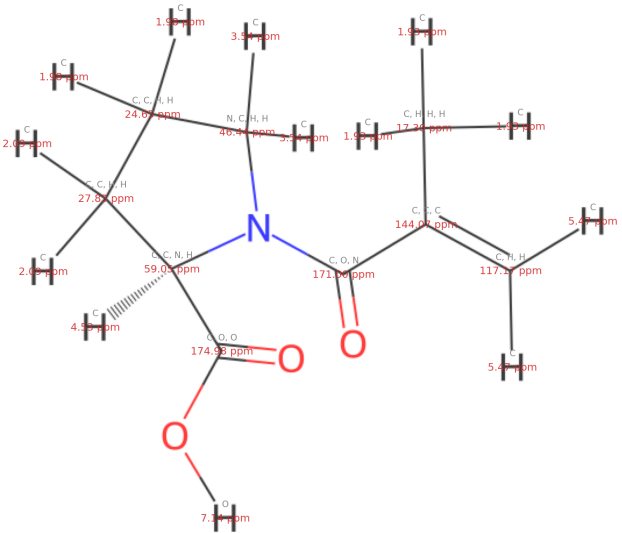
structure $\rightarrow$ shifts
predictor

where each atom has one peak

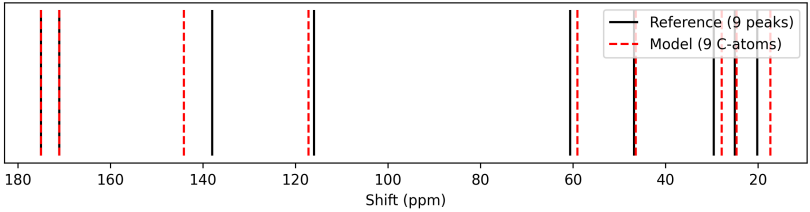
trained to match the given H and C shifts

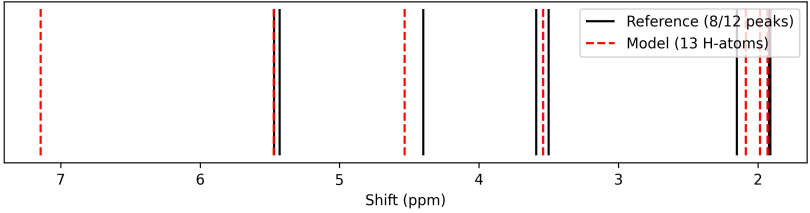
graph neural net with 300k trainable parameters

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

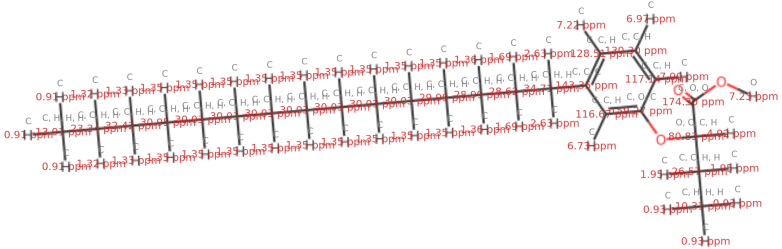


C=C(C)C(=O)N1CCC[C@@H]1C(=O)O

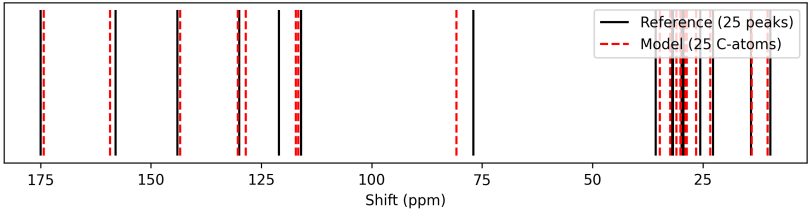


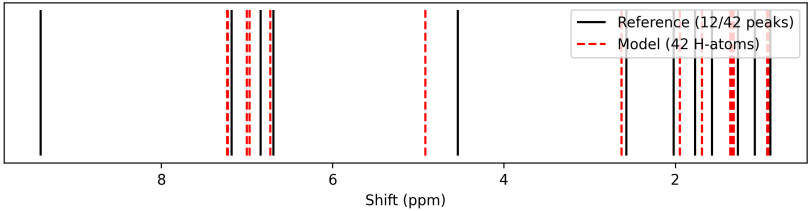


CCCCCCCCCCCCCCCc1cccc(OC(CC)C(=O)O)c1

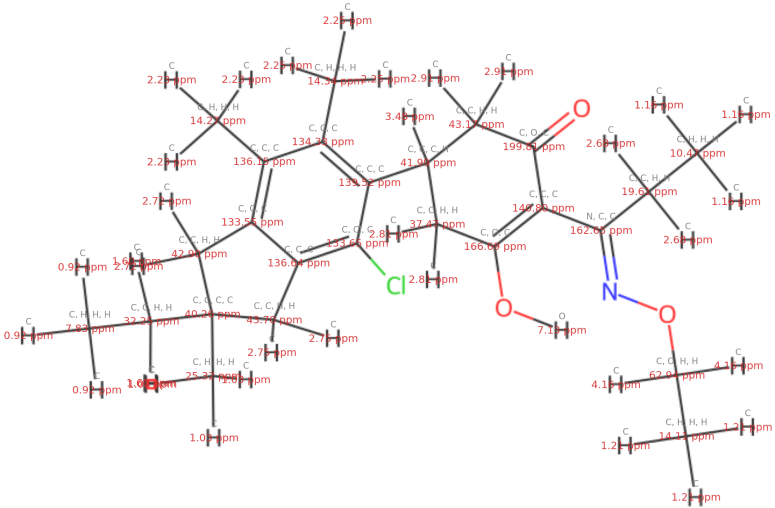


CCCCCCCCCCCCCCCc1cccc(OC(CC)C(=O)O)c1

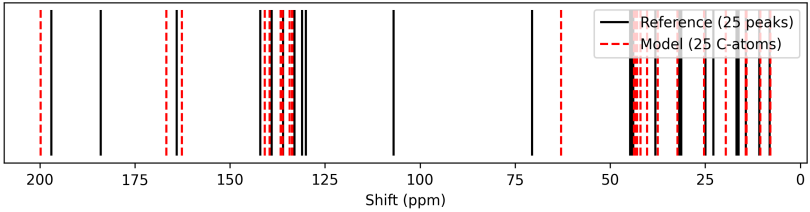


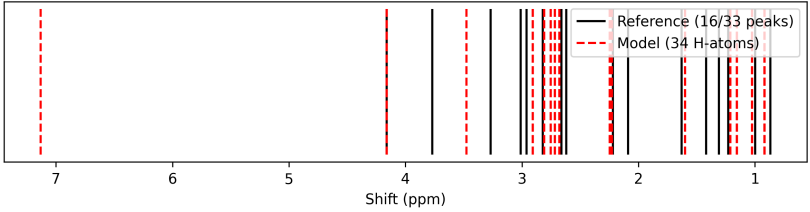


CCON=C(CC)C1=C(O)CC(c2c(C)c(C)c3c(c2Cl)CC(C)(CC)C3)CC1=O

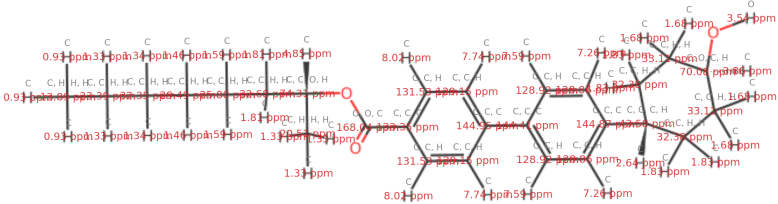


CCON=C(CC)C1=C(O)CC(c2c(C)c(C)c3c(c2Cl)CC(C)(CC)C3)CC1=O

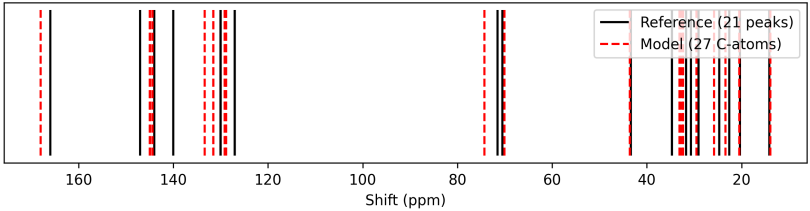


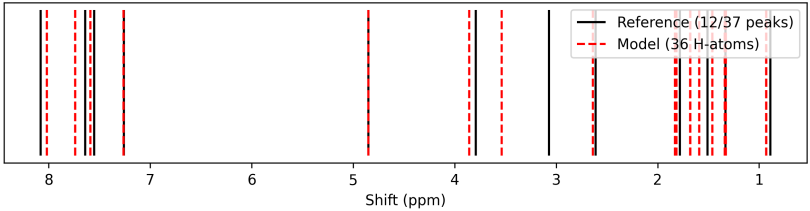


CCCCCC[C@@H](C)OC(=O)c1ccc(-c2ccc([C@@H]3CC[C@@H](O)CC3)cc2)cc1



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





“One atom, one peak” is a simplified model of NMR

Is it easier to invert?..

back to HSQC:

The code based on the legacy OpenNMT is awkward and limited.

Enter **GPT-2**: text-to-text decoder-only autoregressive transformer [5].

Original training:

- Unsupervised next-token prediction on the WebText corpus
- Causal language modeling: $\mathbb{P}(x_t \mid x_1, \ldots, x_{t-1})$

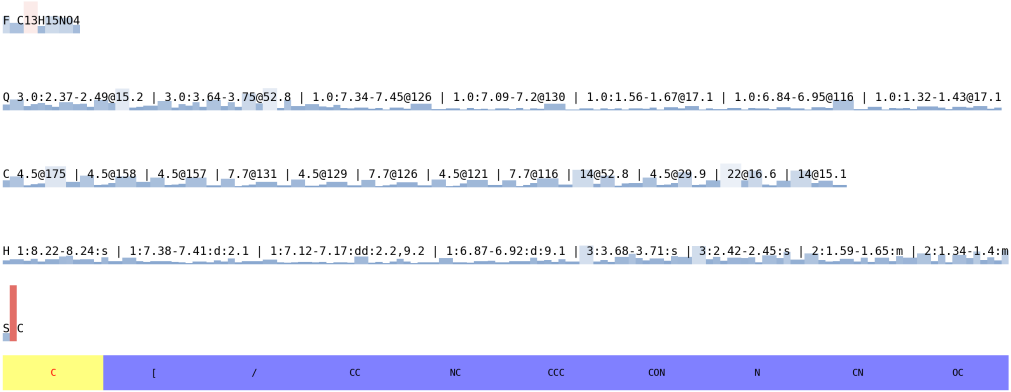
Architecture:

- Byte-Level BPE tokenizer (iteratively merge frequent pairs, 50k)
- Stack of masked self-attention blocks (12–48 layers)
- Learned token and positional embeddings

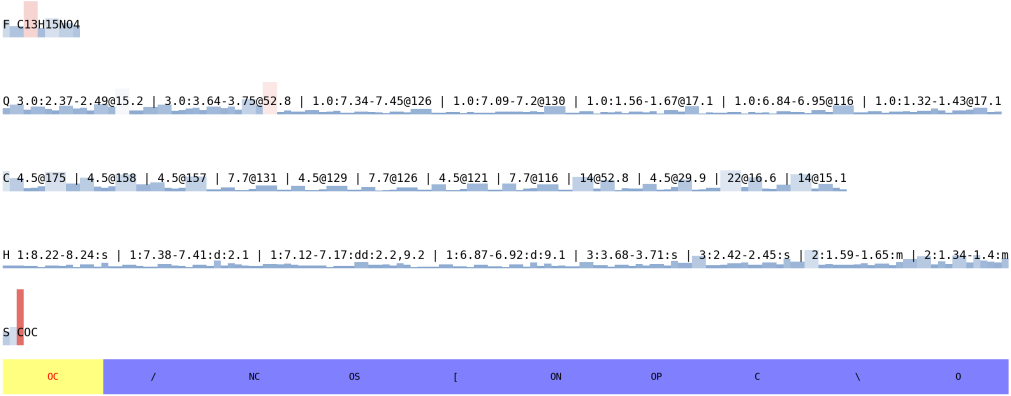
Key models:

- GPT-2: Up to 1.5B parameters (XL), strong zero-shot performance
- **DistilGPT-2:** “distilled” to 82M (cf. [6]), 1024 token context

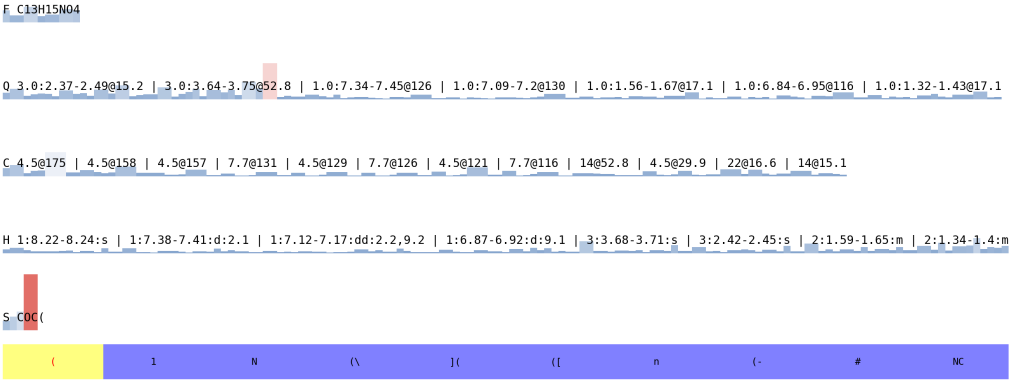
Saliency/sensitivity maps (teacher forcing)



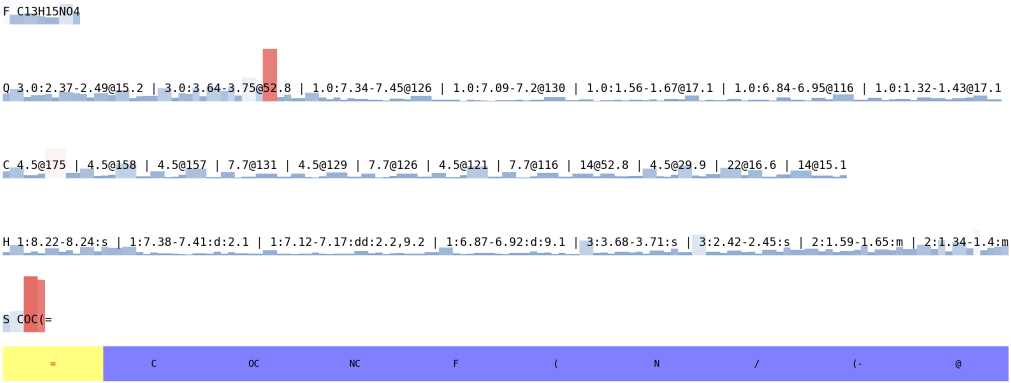
Saliency/sensitivity maps (teacher forcing)



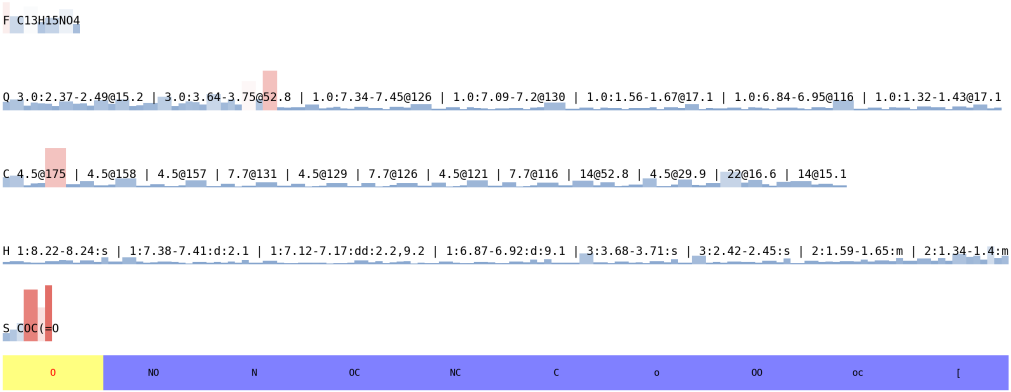
Saliency/sensitivity maps (teacher forcing)



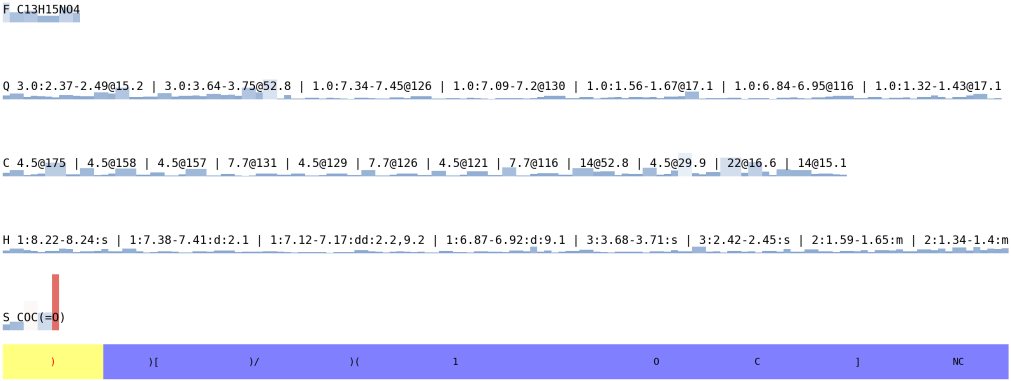
Saliency/sensitivity maps (teacher forcing)



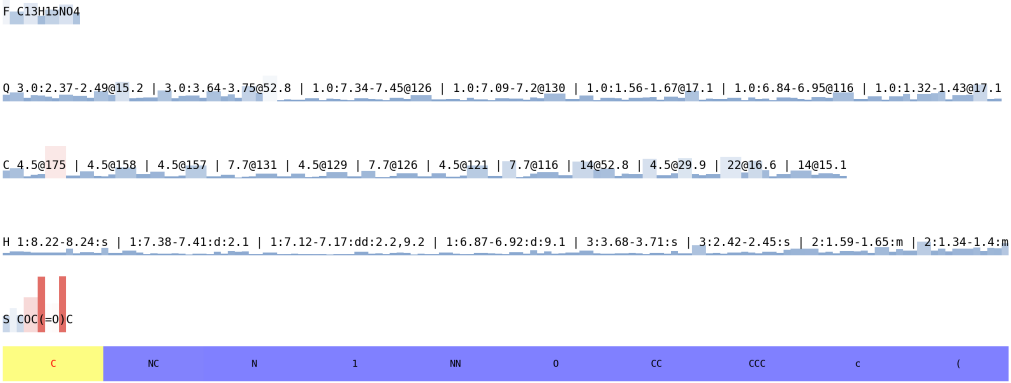
Saliency/sensitivity maps (teacher forcing)



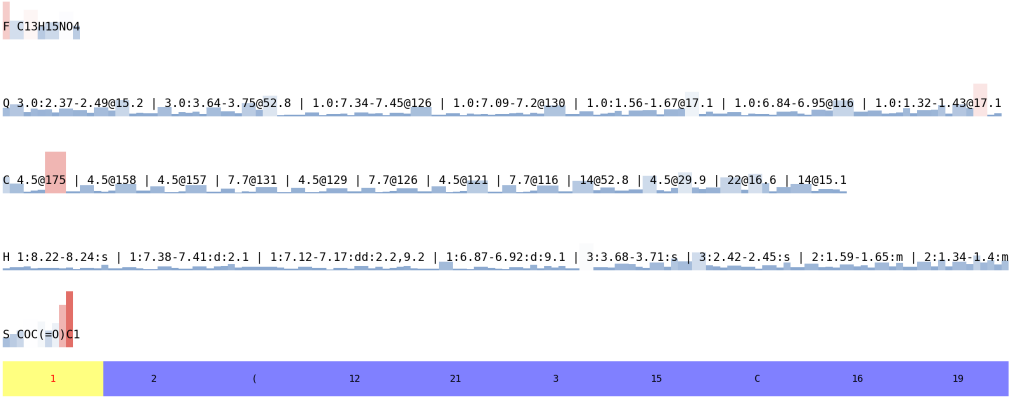
Saliency/sensitivity maps (teacher forcing)



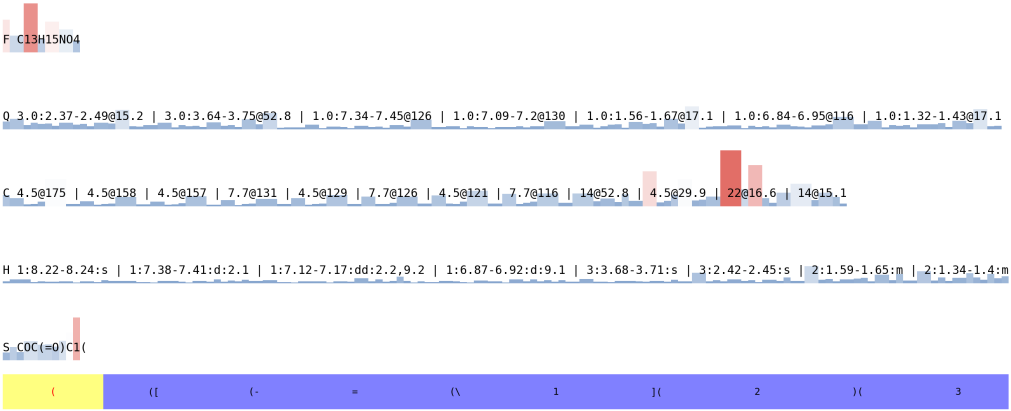
Saliency/sensitivity maps (teacher forcing)



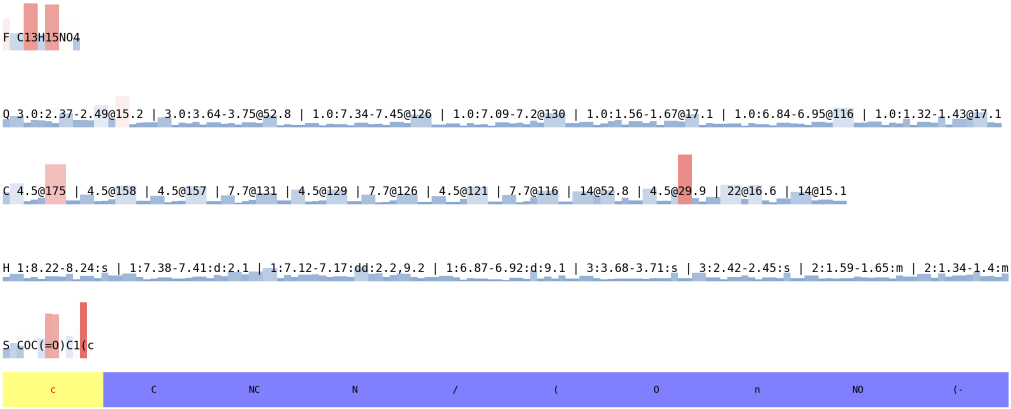
Saliency/sensitivity maps (teacher forcing)



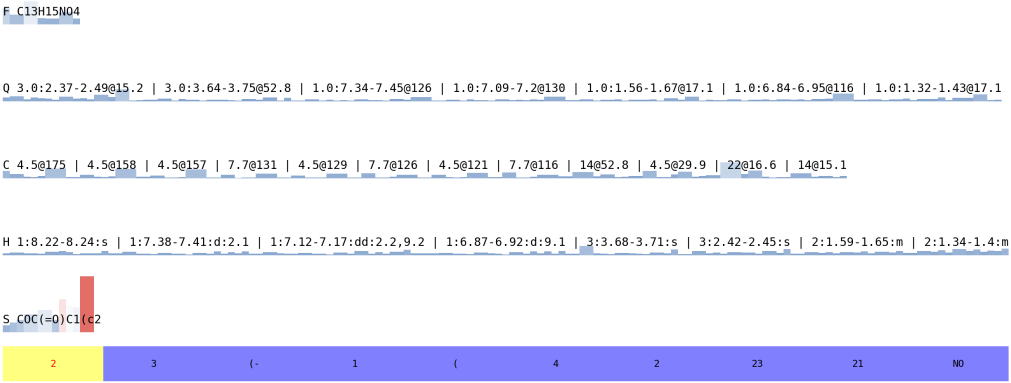
Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)

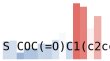
F C13H15NO4

Q 3.0:2.37-2.49@15.2 | 3.0:3.64-3.75@52.8 | 1.0:7.34-7.45@126 | 1.0:7.09-7.2@130 | 1.0:1.56-1.67@17.1 | 1.0:6.84-6.95@116 | 1.0:1.32-1.43@17.1

C 4.5@175 | 4.5@158 | 4.5@157 | 7.7@131 | 4.5@129 | 7.7@126 | 4.5@121 | 7.7@116 | 14@52.8 | 4.5@29.9 | 22@16.6 | 14@15.1

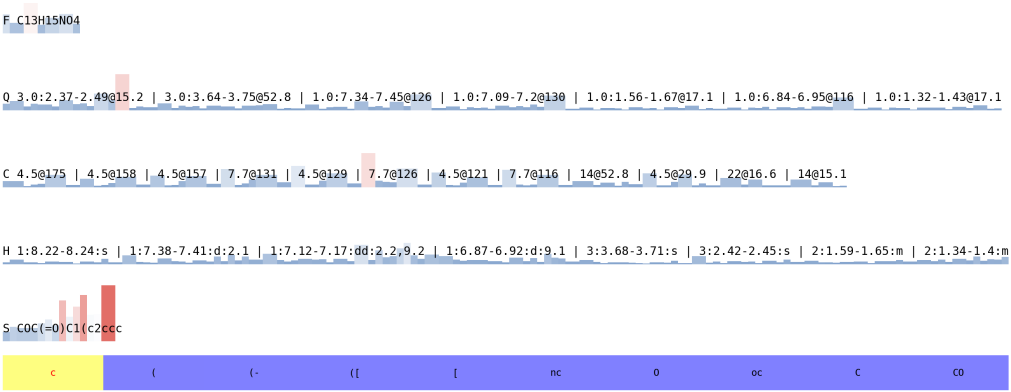
H 1:8.22-8.24:s | 1:7.38-7.41:d:2.1 | 1:7.12-7.17:dd:2.2,9.2 | 1:6.87-6.92:d:9.1 | 3:3.68-3.71:s | 3:2.42-2.45:s | 2:1.59-1.65:m | 2:1.34-1.4:m

S COC(=O)C1(c2cc

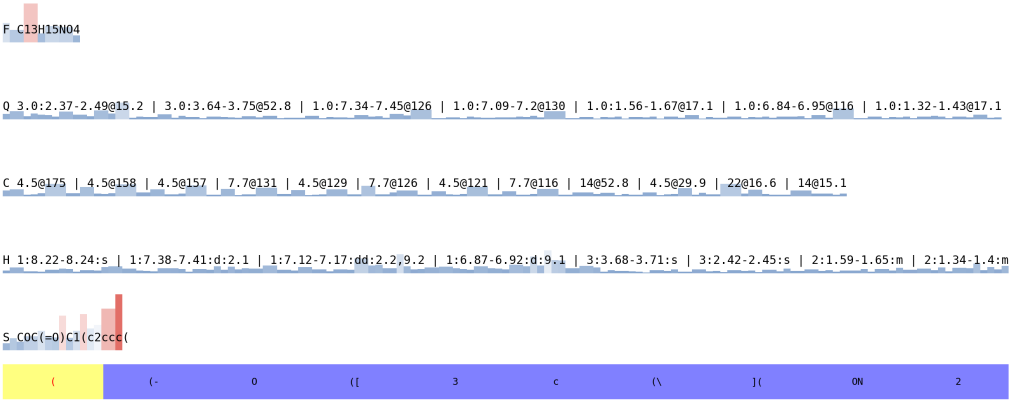


cc	on	n	oc	nc	c	ON	116	O	115

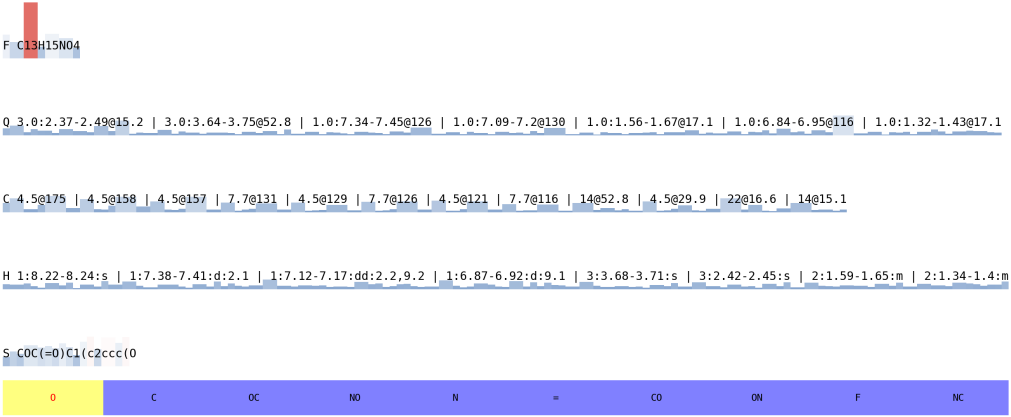
Saliency/sensitivity maps (teacher forcing)



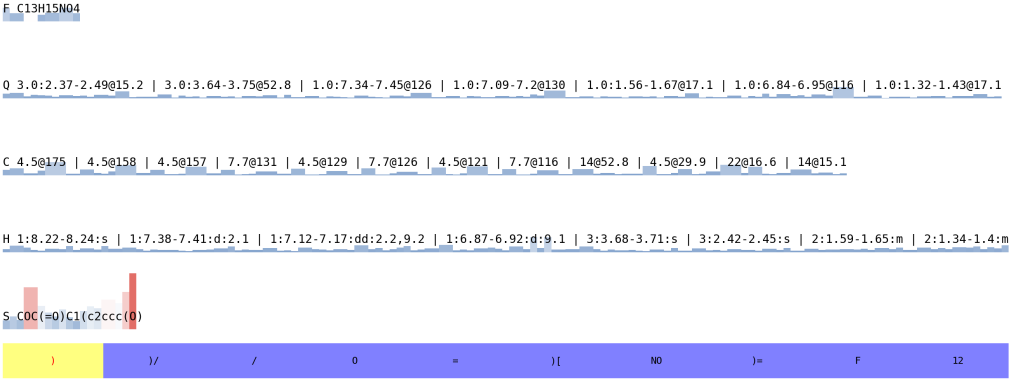
Saliency/sensitivity maps (teacher forcing)



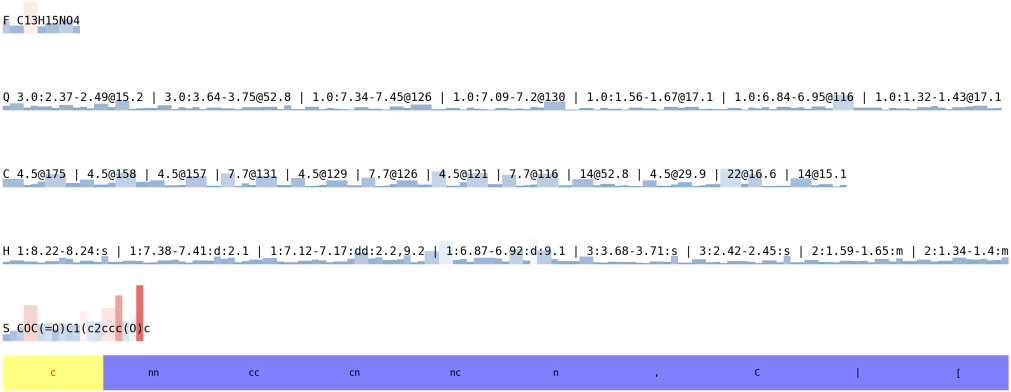
Saliency/sensitivity maps (teacher forcing)



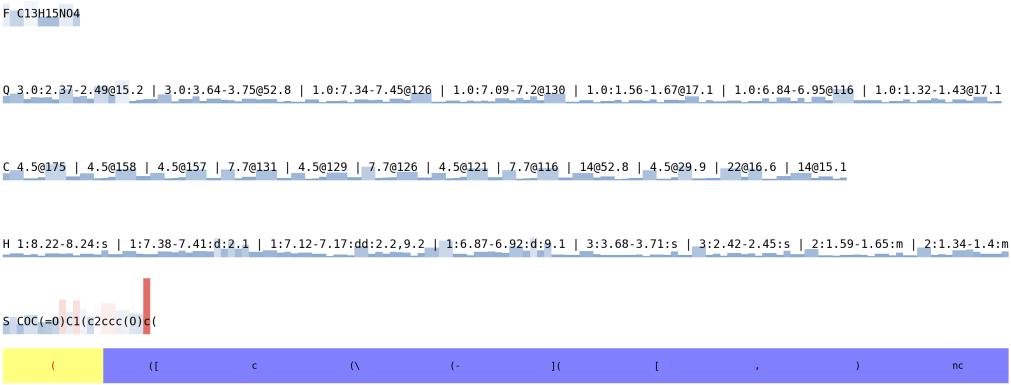
Saliency/sensitivity maps (teacher forcing)



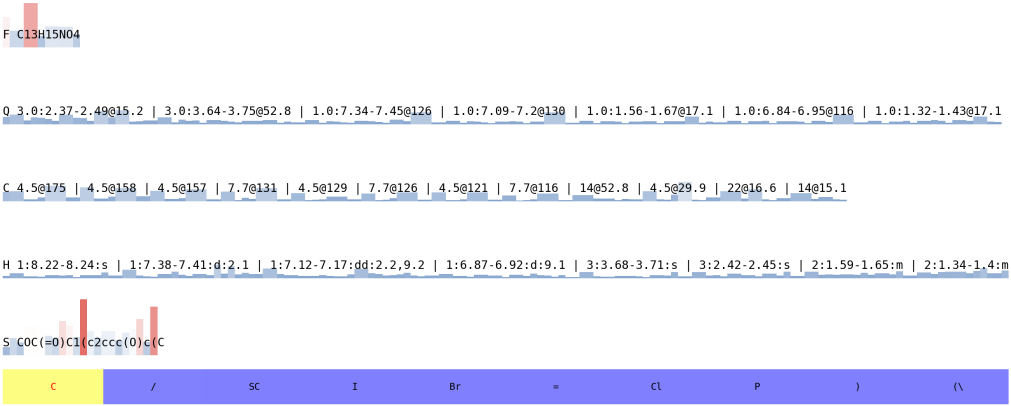
Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)

F C13H15N04

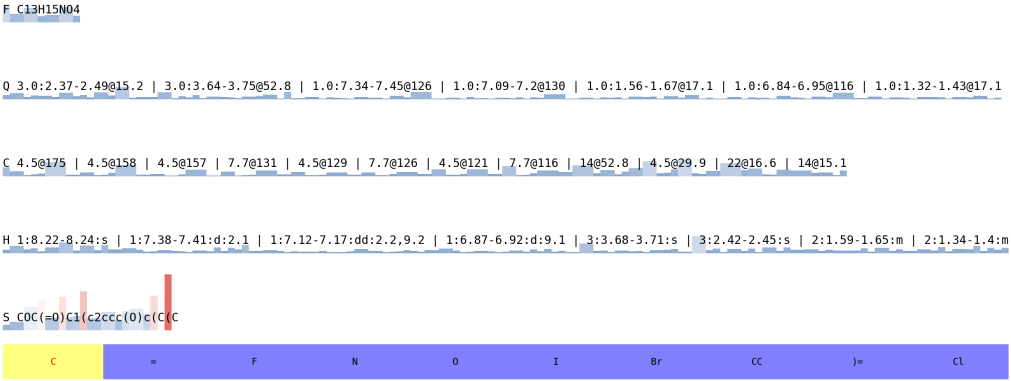
Q 3.0:2.37-2.49@15.2 | 3.0:3.64-3.75@52.8 | 1.0:7.34-7.45@126 | 1.0:7.09-7.2@130 | 1.0:1.56-1.67@17.1 | 1.0:6.84-6.95@116 | 1.0:1.32-1.43@17.1

C 4.5@175 | 4.5@158 | 4.5@157 | 7.7@131 | 4.5@129 | 7.7@126 | 4.5@121 | 7.7@116 | 14@52.8 | 4.5@29.9 | 22@16.6 | 14@15.1

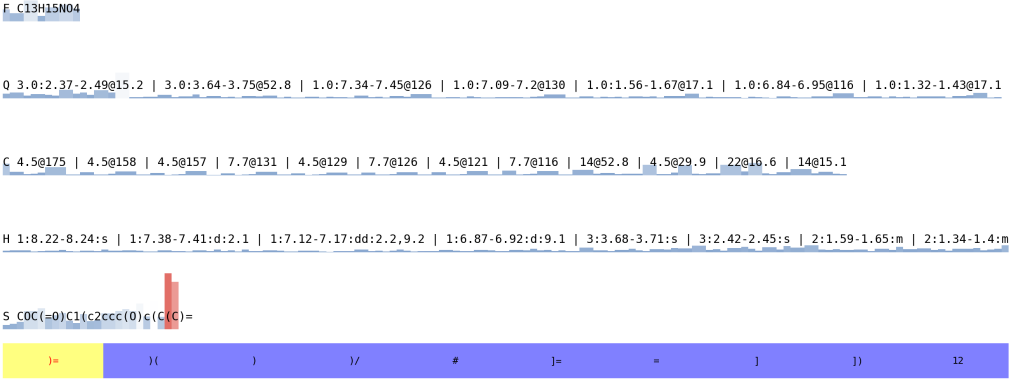
H 1:8.22-8.24:s | 1:7.38-7.41:d:2.1 | 1:7.12-7.17:dd:2.2,9.2 | 1:6.87-6.92:d:9.1 | 3:3.68-3.71:s | 3:2.42-2.45:s | 2:1.59-1.65:m | 2:1.34-1.4:m

S COC(=O)C1(c2ccc(O)c(C)

Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)

F C13H15NO4

Q 3.0:2.37-2.49@15.2 | 3.0:3.64-3.75@52.8 | 1.0:7.34-7.45@126 | 1.0:7.09-7.2@130 | 1.0:1.56-1.67@17.1 | 1.0:6.84-6.95@116 | 1.0:1.32-1.43@17.1

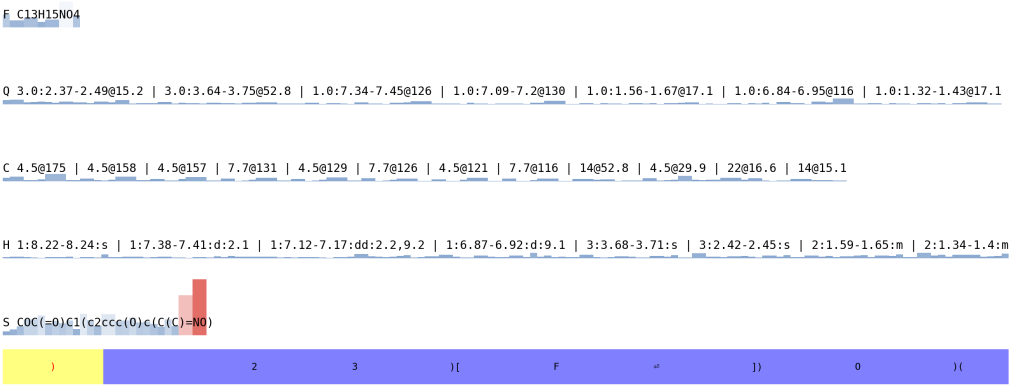
C 4.5@175 | 4.5@158 | 4.5@157 | 7.7@131 | 4.5@129 | 7.7@126 | 4.5@121 | 7.7@116 | 14@52.8 | 4.5@29.9 | 22@16.6 | 14@15.1

H 1:8.22-8.24:s | 1:7.38-7.41:d:2.1 | 1:7.12-7.17:dd:2.2,9.2 | 1:6.87-6.92:d:9.1 | 3:3.68-3.71:s | 3:2.42-2.45:s | 2:1.59-1.65:m | 2:1.34-1.4:m

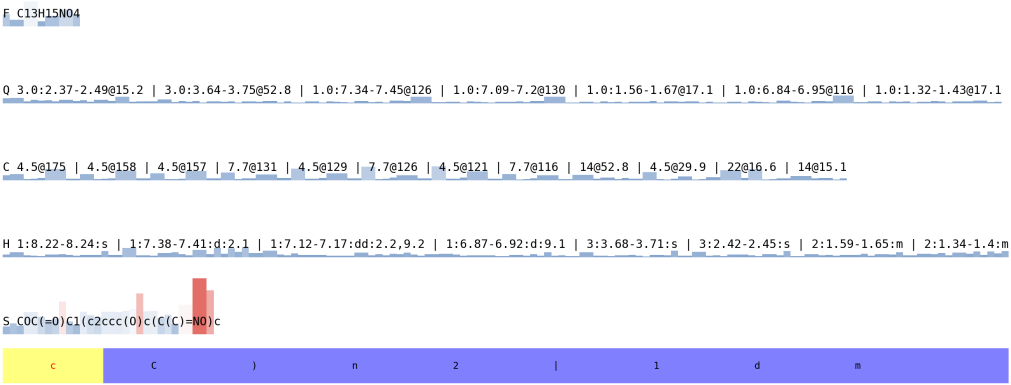
S COC(=O)C1(c2ccc(O)c(C(C)=NO

NO	N	NN	O	NC	ON	no	C	NS	OO

Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)

F C13H15NO4

Q 3.0:2.37-2.49@15.2 | 3.0:3.64-3.75@52.8 | 1.0:7.34-7.45@126 | 1.0:7.09-7.2@130 | 1.0:1.56-1.67@17.1 | 1.0:6.84-6.95@116 | 1.0:1.32-1.43@17.1

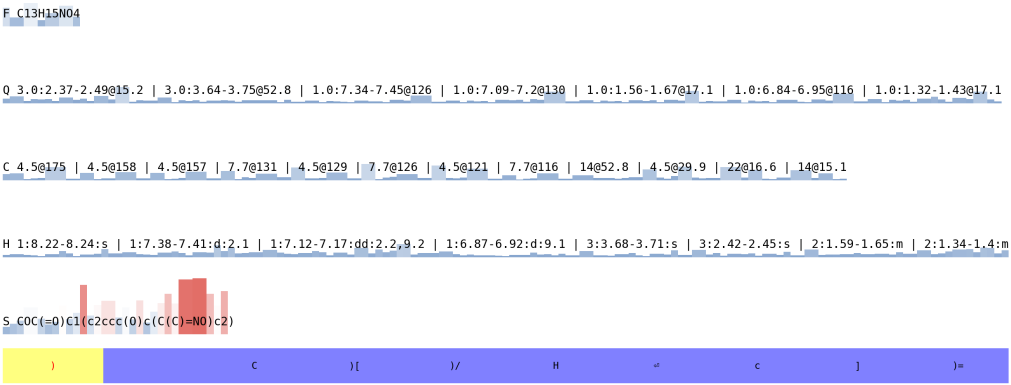
C 4.5@175 | 4.5@158 | 4.5@157 | 7.7@131 | 4.5@129 | 7.7@126 | 4.5@121 | 7.7@116 | 14@52.8 | 4.5@29.9 | 22@16.6 | 14@15.1

H 1:8.22-8.24:s | 1:7.38-7.41:d:2.1 | 1:7.12-7.17:dd:2.2,9.2 | 1:6.87-6.92:d:9.1 | 3:3.68-3.71:s | 3:2.42-2.45:s | 2:1.59-1.65:m | 2:1.34-1.4:m

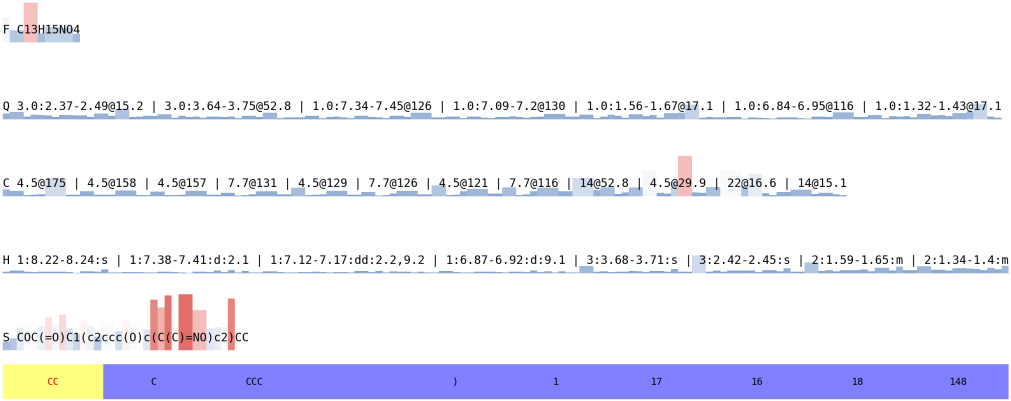
S COC(=O)C1(c2ccc(O)c(C(C)=NO)c2

2	)	c	3	1	4	21	23	24	.

Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)

F C13H15NO4

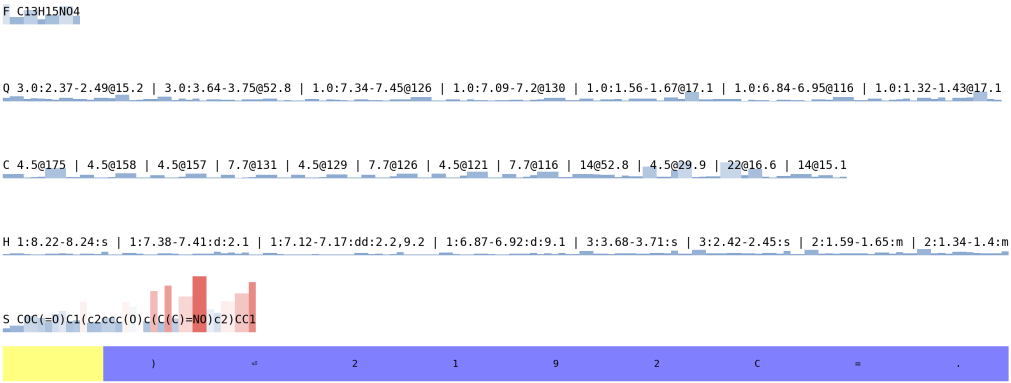
Q 3.0:2.37-2.49@15.2 | 3.0:3.64-3.75@52.8 | 1.0:7.34-7.45@126 | 1.0:7.09-7.2@130 | 1.0:1.56-1.67@17.1 | 1.0:6.84-6.95@116 | 1.0:1.32-1.43@17.1

C 4.5@175 | 4.5@158 | 4.5@157 | 7.7@131 | 4.5@129 | 7.7@126 | 4.5@121 | 7.7@116 | 14@52.8 | 4.5@29.9 | 22@16.6 | 14@15.1

H 1:8.22-8.24:s | 1:7.38-7.41:d:2.1 | 1:7.12-7.17:dd:2.2,9.2 | 1:6.87-6.92:d:9.1 | 3:3.68-3.71:s | 3:2.42-2.45:s | 2:1.59-1.65:m | 2:1.34-1.4:m

S COC(=O)C1(c2ccc(O)c(C(C)=NO)c2)CC1

Saliency/sensitivity maps (teacher forcing)



Saliency/sensitivity maps (teacher forcing)

F C13H15NO4

Q 3.0:2.37-2.49@15.2 | 3.0:3.64-3.75@52.8 | 1.0:7.34-7.45@126 | 1.0:7.09-7.2@130 | 1.0:1.56-1.67@17.1 | 1.0:6.84-6.95@116 | 1.0:1.32-1.43@17.1

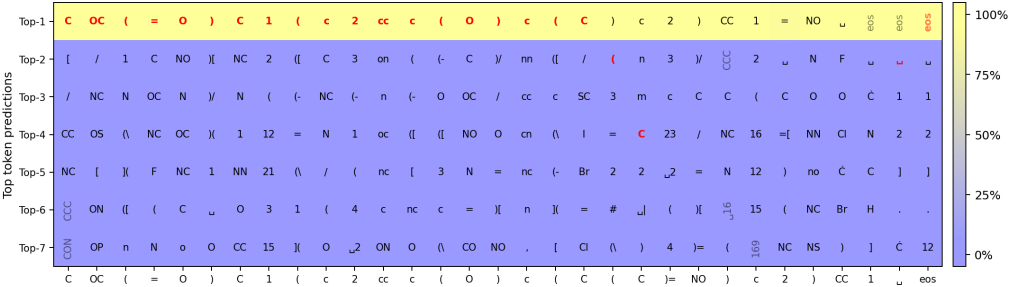
C 4.5@175 | 4.5@158 | 4.5@157 | 7.7@131 | 4.5@129 | 7.7@126 | 4.5@121 | 7.7@116 | 14@52.8 | 4.5@29.9 | 22@16.6 | 14@15.1

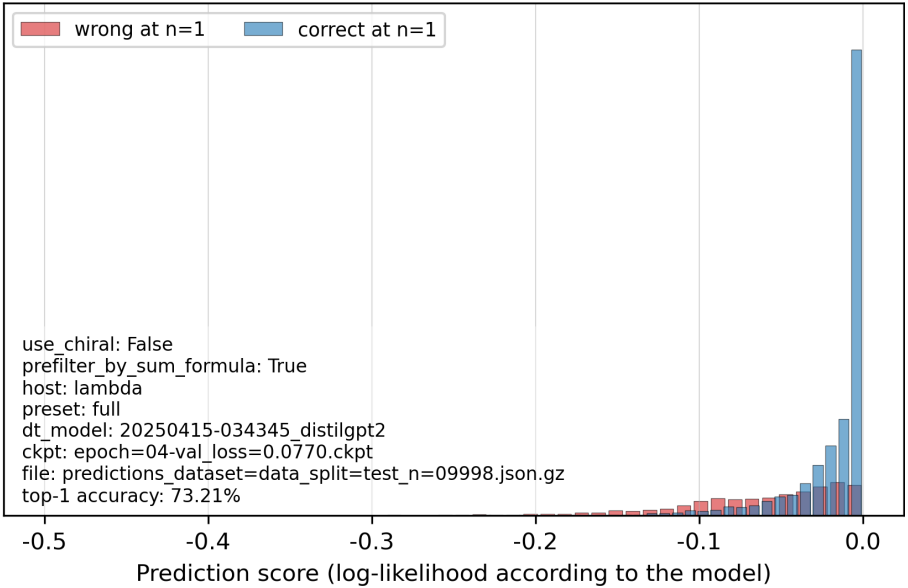
H 1:8.22-8.24:s | 1:7.38-7.41:d:2.1 | 1:7.12-7.17:dd:2.2,9.2 | 1:6.87-6.92:d:9.1 | 3:3.68-3.71:s | 3:2.42-2.45:s | 2:1.59-1.65:m | 2:1.34-1.4:m

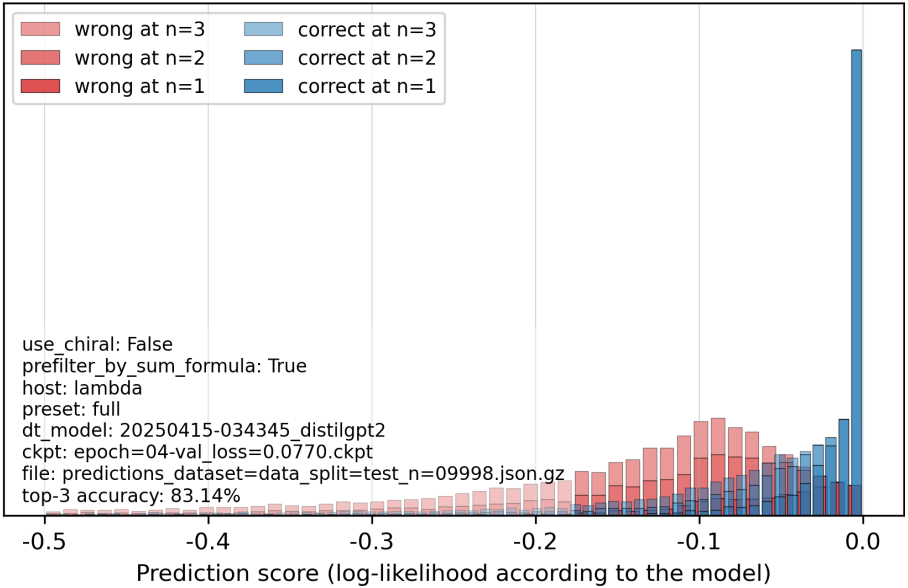
S COC(=O)C1(c2ccc(O)c(C(C)=NO)c2)CC1

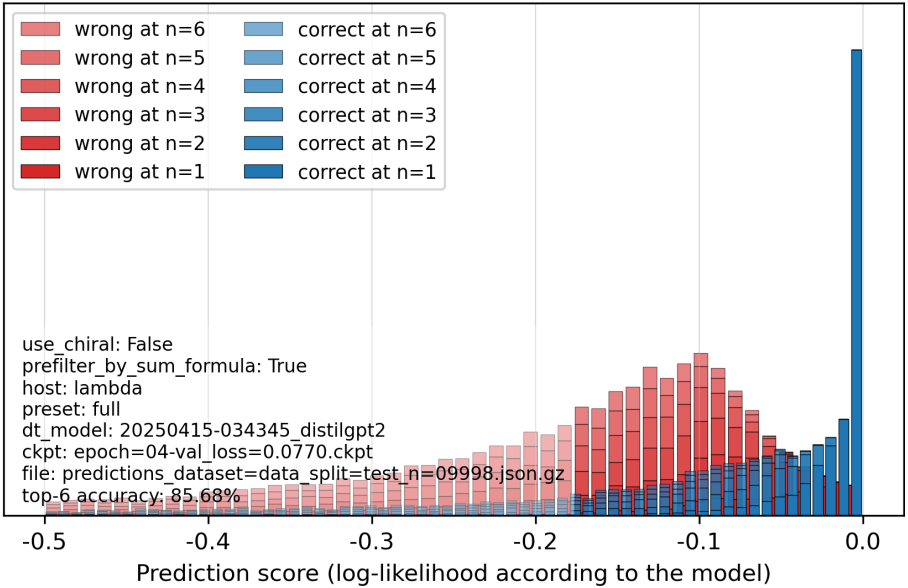
<|endoftext|>	2	H	↵	C	1	10	tt	3

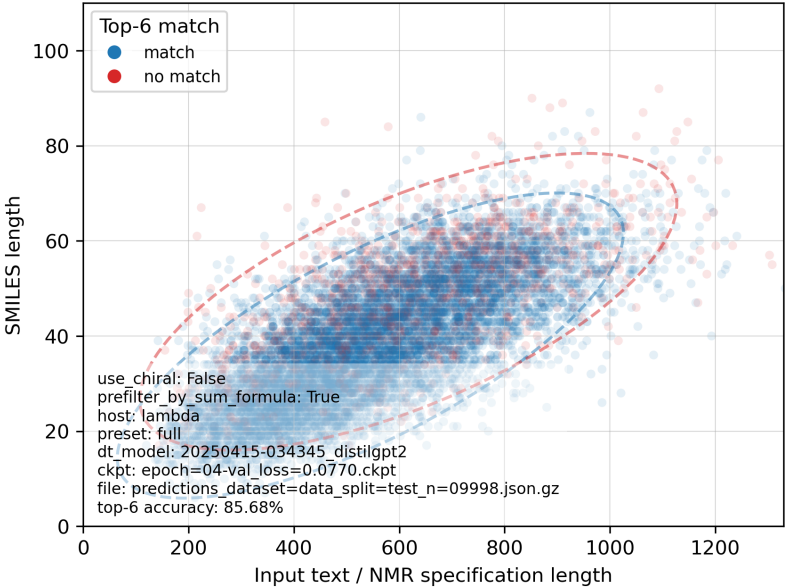
Token choice (free form)




wrong at n=1
correct at n=1
use_chiral: False
prefilter_by_sum_formula: True
host: lambda
preset: full
dt_model: 20250415-034345_distilgpt2
ckpt: epoch=04-val_loss=0.0770.ckpt
file: predictions_dataset=data_split=test_n=09998.json.gz
top-1 accuracy: 73.21%
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


wrong at n=3
wrong at n=2
wrong at n=1
correct at n=3
correct at n=2
correct at n=1
use_chiral: False
prefilter_by_sum_formula: True
host: lambda
preset: full
dt_model: 20250415-034345_distilgpt2
ckpt: epoch=04-val_loss=0.0770.ckpt
file: predictions_dataset=data_split=test_n=09998.json.gz
top-3 accuracy: 83.14%
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


wrong at n=6
wrong at n=5
wrong at n=4
wrong at n=3
wrong at n=2
wrong at n=1
correct at n=6
correct at n=5
correct at n=4
correct at n=3
correct at n=2
correct at n=1
use_chiral: False
prefilter_by_sum_formula: True
host: lambda
preset: full
dt_model: 20250415-034345_distilgpt2
ckpt: epoch=04-val_loss=0.0770.ckpt
file: predictions_dataset=data_split=test_n=09998.json.gz
top-6 accuracy: 85.68%
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
Prediction score (log-likelihood according to the model)


Top-6 match
match
no match
SMILES length
Input text / NMR specification length
0
20
40
60
80
100
0
200
400
600
800
1000
1200
use_chiral: False
prefilter_by_sum_formula: True
host: lambda
preset: full
dt_model: 20250415-034345_distilgpt2
ckpt: epoch=04-val_loss=0.0770.ckpt
file: predictions_dataset=data_split=test_n=09998.json.gz
top-6 accuracy: 85.68%

On “structural observability”

- Saliency is spread across modalities, with occasional emphasis
- Optimistically/tentatively: HSQC adds $\sim$ 5-10% accuracy
- To check, would have to train the same model on C/H only
- . . . or with mixted training samples (missing various inputs)

Other what-ifs

- Pre-fine-tune on SMILES grammar
- Train distilgpt2 architecture from zero
- Train on the one-atom-one-peak quasi-NMR
- More careful split; explicit stereo; explicit H (or mixed)
- Fine-tuning OpenAI gpt-4o-mini via API: $\sim$\$5’000

DiffMS [7]

[7] Montgomery Bohde et al. “DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra”. In: *arXiv* (2025). [url]

DiffMS [7]: generative diffusion from fingerprint

Generate molecular structures from MS/MS spectra and chemical formula. Key components:

- Encoder: transformer model (MIST-CF [8]) spectrum $\rightarrow$ fingerprint
- Conditioning: atom types from formula and spectrum embedding
- Decoder: discrete graph diffusion on the connectivity

Main contributions:

- Formula-constrained graph generation using discrete diffusion
- Decoder pretrained on 2.8M fingerprint–structure pairs
- End-to-end fine-tuning on paired spectrum–structure data
- Strong performance on out-of-distribution benchmarks

Morgan fingerprint $\mathbf{f} \in \{0,1\}^d$

- Molecular graph $G = (V, E)$
- Radius r: max number of hashing iterations (e.g., $r = 2$)
- Length d: size of fingerprint (e.g., 2048)
- Initialize atom identifiers:

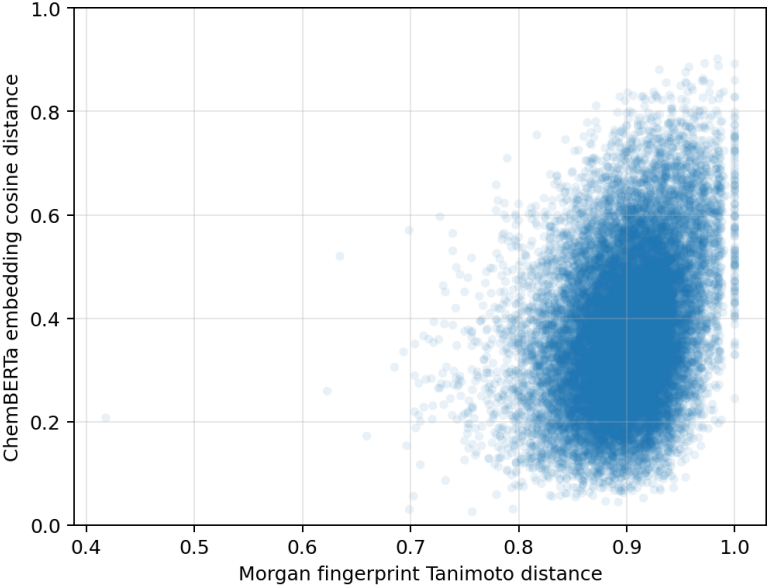
$$h_v^{(0)} := \text{hash}(\text{features}(v)) \quad \text{for all} \quad v \in V$$

- Iterate for $t = 1$ to r:

$$h_v^{(t)} := \text{hash}(h_v^{(t-1)} \text{ and all its neighbors})$$

- For each v:

$$i := \text{hash}(h_v^{(t)}) \bmod d \quad \text{and set} \quad \mathbf{f}[i] := 1$$


1.0
0.8
0.6
0.4
0.2
0.0
ChemBERTa embedding cosine distance
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Morgan fingerprint Tanimoto distance

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. [url].

[2] Marvin Alberts, Federico Zipoli, and Alain C. Vaucher. *Learning the Language of NMR: Structure Elucidation from NMR spectra using Transformer Models*. 2023. [url].

[3] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. “OpenNMT: Open-source toolkit for neural machine translation”. In: *Proceedings of ACL 2017, System Demonstrations* (2017), pp. 67–72. [url].

References II

[4] Raymond J. Abraham and Mehdi Mobli. *Modelling 1H NMR spectra of organic compounds: Theory, applications and NMR prediction software*. Wiley, 2008. [url].

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. *Language models are unsupervised multitask learners*. Tech. rep. 2019.

[6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *Proceedings of EMC^2 : Efficient Machine Learning, Co-located with NeurIPS 2019* (2019).

[7] Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W. Coley. "DiffMS: Diffusion Generation of Molecules Conditioned on Mass Spectra". In: *arXiv* (2025). [url].

References III

[8] Samuel Goldman, Jiayi Xin, Joules Provenzano, and Connor W. Coley. “MIST-CF: Chemical formula inference from tandem mass spectra”. In: *Journal of Chemical Information and Modeling* (2024), pp. 2421–2431. [url].

this slide is intentionally left blank