Supplementary Information

NMRExtractor: Leveraging Large Language Models to Construct an Experimental NMR Database from Open-source Scientific Publications†

Qinggong Wang,‡[ab] Wei Zhang,‡[bc] Mingan Chen,[bde] Xutong Li,[bc] Zhaoping Xiong,[f] Jiacheng Xiong,*[b] Zunyun Fu,*[d] Mingyue Zheng,*[abc]

[a.]Nanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China.

[b.]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, 555 Zuchongzhi Road, Shanghai 201203, China.

[c.]University of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China.

[d.]ShanghaiTech University, Shanghai 201210, China.

[e.]Lingang Laboratory, Shanghai 200031, China.

[f.]ProtonUnfold Technology Co., Ltd, Suzhou, China.

‡These authors contributed equally to this work.

*Correspondence should be addressed to:

Mingyue Zheng: myzheng@simm.ac.cn
Zunyun Fu: fuzy@shanghaitech.edu.cn
Jiacheng Xiong: s19-xiongjiacheng@simm.ac.cn

Contents

List of Figures

Fig. S1 An example of the NMRExtractor extraction process 3

Fig. S2 Prompt for fine-tuning the Mistral-7b-instruct-v-0.2 model 4

Fig. S3 Examples of text annotations in the test set 2 (1022) 5

Fig. S4 Error cases of NMR data extraction 6

Fig. S5 Error cases of NMR data extraction 7

Fig. S6 Standardization of SMILES Using RDKit 8

Fig. S7 The 115 journals and their data counts for test set 2 (1022) 9

Fig. S8 Some regular expressions used by rule-based methods 10

Fig. S9 618 common compound group words and atomic count 12

Fig. S10 The frequency distribution of these duplicate IUPAC names 13

Fig. S11 Accuracy of NMR chemical shift values and their order in NMRBank 14

Fig. S12 Regular expression missing case 15

Fig. S13 Unconvertible IUPAC names case 16

List of Tables

Table. S1 Hyperparameters for the models 4

Table. S2 Fine-tuning and Inferencing Cost 4

Table. S3 Exact match accuracy of open source LLMs on Test Set 1(300) 5

Table. S4 Exact match accuracy of open source LLMs on Test Set 2 (1022) 5

Table. S5 Exact match accuracy in three-fold cross-validation at different confidences 5

Table. S6 Table S6 Most frequent duplicate IUPAC names and their frequency 15

Supplementary Methods

The confidence of the model predictions 22

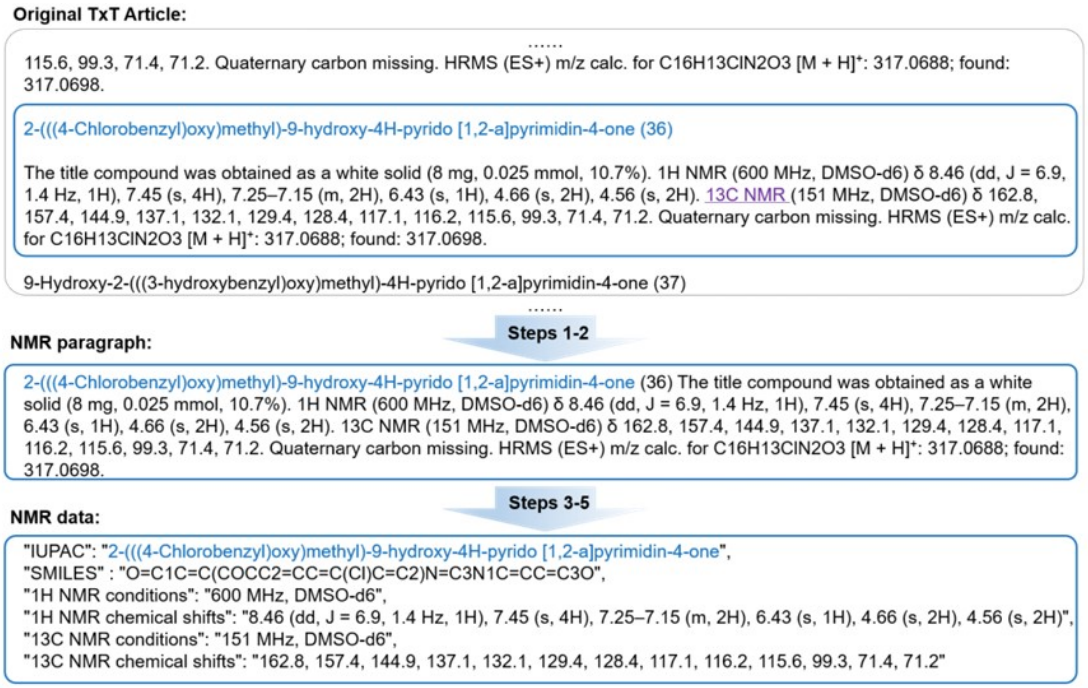


Fig. S1 An example of the NMRExtractor extraction process. Steps 1–2 use the expression "13C.{0,3}NMR" to locate paragraphs mentioning NMR. To ensure completeness, the matched paragraph is concatenated with its adjacent ones, capturing the compound's IUPAC name, 1H/13C NMR spectral data. In Steps 3–5, a fine-tuned LLM extracts structured information, including the compound's IUPAC name, 1H/13C NMR chemical shifts and conditions. The IUPAC name is then converted to SMILES and standardized to generate the final dataset.

Fine-tuning Prompt:

Extract text containing 1H NMR and 13C NMR data, remove interference information such as reactants, raw materials, solvents and other non-final product names based on text semantics, and then extract the name, code or number of the final product. Please delete the IUPAC name Alias, numbers and ordinal numbers before and after fields, such as '2.1.3.', '(HL4)', '(9)', '(4d)'. NMR text should contain complete information, such as instrument power and solvent information, For example, "13C NMR text": "13C NMR (400 MHz, acetone-d6) 174.0 (C), 157.7 (C). Then split the NMR text. The content in NMR conditions is NMR instrument power and solvent information, such as "13C NMR conditions": "400MHz, acetone-d6". The content in the 13C NMR data removes information such as the position and shape of the peak, such as "13C NMR data": "131.4-128.0, 157.7". The content in the 1H NMR data should include information such as the position and shape of the peak, such as "1H NMR data": "12.57 (s, 1H), 7.97-7.95 (d, J = 8.25 Hz, 2H)". Please keep the duplicate values of the original data and do not modify the number of decimal places. All responses must originate from information extracted from the given text, ensuring that the extracted content has not been modified or fragmented, and that capitalization and punctuation are exactly the same as the given text. Must end with {"IUPAC":"text","1H NMR text":"text","1H NMR conditions":"text","1H NMR data":"text","13C NMR text":"text","13C NMR conditions":"text","13C NMR data":"text"} format reply.

{Input NMR paragraph}

Input:

Description of 5-(4-(3-chloro-5-(trifluoromethyl)pyridin-2-yl)phenoxy)-N-(2,4-difluorophenyl)-2-nitrobenzamide (5i): White solid, yield 78.1%. m.p. 159.1–161.5 °C; 1H NMR (400 MHz, CDCl3) δ 8.84 (s, 1H), 8.25 (s, 1H), 8.09 (dd, J = 5.1, 3.7 Hz, 2H), 7.86 (d, J = 8.6 Hz, 2H), 7.52 (s, 1H), 7.42 (d, J = 8.1 Hz, 1H), 7.33 (t, J = 8.2 Hz, 1H), 7.22 (d, J = 8.5 Hz, 2H), 7.15–7.10 (m, 2H), 7.02 (d, J = 8.1 Hz, 1H); 13C NMR (101 MHz, CDCl3) δ 163.78, 162.18, 158.52, 155.36, 144.36, 140.24, 135.60, 134.94, 134.50, 131.93, 130.20, 127.62, 126.25, 123.49, 123.38, 120.21, 118.70, 116.89, 111.70, 111.48, 103.80. HRMS calcd. for C25H13ClF5N3O4 [M–H]− 548.0422, found 548.0422.

Output:

{"IUPAC": "5-(4-(3-chloro-5-(trifluoromethyl)pyridin-2-yl)phenoxy)-N-(2,4-difluorophenyl)-2-nitrobenzamide",
"1H NMR conditions": "400 MHz, CDCl3",
"1H NMR data": "8.84 (s, 1H), 8.25 (s, 1H), 8.09 (dd, J = 5.1, 3.7 Hz, 2H), 7.86 (d, J = 8.6 Hz, 2H), 7.52 (s, 1H), 7.42 (d, J = 8.1 Hz, 1H), 7.33 (t, J = 8.2 Hz, 1H), 7.22 (d, J = 8.5 Hz, 2H), 7.15–7.10 (m, 2H), 7.02 (d, J = 8.1 Hz, 1H)",
"13C NMR conditions": "101 MHz, CDCl3",
"13C NMR data": "163.78, 162.18, 158.52, 155.36, 144.36, 140.24, 135.60, 134.94, 134.50, 131.93, 130.20, 127.62, 126.25, 123.49, 123.38, 120.21, 118.70, 116.89, 111.70, 111.48, 103.80"}

Fig. S2 Prompt for fine-tuning the Mistral-7b-instruct-v-0.2 model

Fig. S3 Examples of text annotations in the test set 2 (1022)

Error Cases

Example 1:

3-{4-[(2-(4-bromobenzoyl)-7-methylfuro[2,3-c]pyridin-4-yl)methyl]piperazin-1-yl}propanenitrile (PD10). Yield: 1.18 g (84%). Light beige crystals, m.p. 129–132 °C (EtOH). 1H NMR (DMSO-d6): δ ppm 2.44–2.61 (m, 14H, CH3, H-2′, H-3′, H-5′, H-6′, DMF), 2.75 (s, 4H, 2CH2), 3.78 (s, 2H, CH2), 7.80 (d, J 8.3, 2H, H-2″, H-6″), 7.86 (s, 1H, H-3), 8.01 (d, J 8.2, 2H, H-3″, H-5″), 8.23 (s, 1H, H-5).13C NMR (DMSO-d6) δ ppm: 14.877, 18.288, 52.117–52.220(d), 52.585, 56.554, 114.768, 119.833, 126.200, 127.585, 131.310, 131.817, 132.150, 135.164, 142.005, 142.672, 150.245, 152.175, 182.766 (DMSO-d6: 38.832–40.082). Anal.Calc. for C23H23BrN4O2;(%).: C, 59.11; H, 4.96; Br, 17.10; N, 11.99%. Found: C, 59.03; H, 4.89; Br, 17.29; N, 11.72%.

Ground Truth:

{"IUPAC":"3-{4-[(2-(4-bromobenzoyl)-7-methylfuro[2,3-c]pyridin-4-yl)methyl]piperazin-1-yl}propanenitrile",
"1H NMR conditions":"DMSO-d6",
"1H NMR data":"2.44–2.61 (m, 14H, CH3, H-2′, H-3′, H-5′, H-6′, DMF), 2.75 (s, 4H, 2CH2), 3.78 (s, 2H, CH2), 7.80 (d, J 8.3, 2H, H-2″, H-6″), 7.86 (s, 1H, H-3), 8.01 (d, J 8.2, 2H, H-3″, H-5″), 8.23 (s, 1H, H-5)",
"13C NMR conditions":"DMSO-d6",
"13C NMR data":"14.877, 18.288, 52.117–52.220, 52.585, 56.554, 114.768, 119.833, 126.200, 127.585, 131.310, 131.817, 132.150, 135.164, 142.005, 142.672, 150.245, 152.175, 182.766"}

Predicted Result:

{"IUPAC":"3-{4-[(2-(4-bromobenzoyl)-7-methylfuro[2,3-c]pyridin-4-yl)methyl]piperazin-1-yl}propanenitrile",
"1H NMR conditions":"DMSO-d6",
"1H NMR data":"2.44–2.61 (m, 14H, CH3, H-2′, H-3′, H-5′, H-6′, DMF), 2.75 (s, 4H, 2CH2), 3.78 (s, 2H, CH2), 7.80 (d, J 8.3, 2H, H-2″, H-6″), 7.86 (s, 1H, H-3), 8.01 (d, J 8.2, 2H, H-3″, H-5″), 8.23 (s, 1H, H-5)",
"13C NMR conditions":"DMSO-d6",
"13C NMR data":"14.877, 18.288, 52.117–52.220, 52.585, 56.554, 114.768, 119.833, 126.200, 127.585, 131.310, 131.817, 132.150, 135.164, 142.005, 142.672, 150.245, 152.175, 182.766 (DMSO-d6: 38.832–40.082)"}

Example 2:

3.1.11. 2-Amino-4-(5-bromo-1-ethyl-1H-indol-2-yl)-7,7-dimethyl-5-oxo-5,6,7,8-tetrahydro-4H-chromene-3-carbonitrile (4j) The product 4j was prepared from the substrates 1j, 2, and 3 using the general procedure. M.P: 280–282 °C; yield: 82%; IR (KBr): 3339 (broad, -NH), 2189 (C=N), 1709 (C=O) cm−1; 1H NMR δ (400 MHz; CDCl3): 0.9 (s, 6H, -CH3), 1.2 (q, 3H, -CH3), 1.8–2.2 (s, 4H, -CH2), 3.8 (t, 2H, -CH2), 4.3 (s, 1H, -CH), 7.2–8.4 (m, 6H, Ar-H and -NH2); 13C NMR δ (100 MHz; CDCl3): 14.1, 18.8, 20.3, 21.9, 28.3, 41.6, 48.7, 60.6, 121.5, 124.9, 126.6, 128.9, 130.1, 132.6, 138.9, 140.6, 142.8, 143.8, and 170.2; and [M+ and M+2]: 439 and 441..

Ground Truth:

{"IUPAC":"2-Amino-4-(5-bromo-1-ethyl-1H-indol-2-yl)-7,7-dimethyl-5-oxo-5,6,7,8-tetrahydro-4H-chromene-3-carbonitrile",
"1H NMR conditions":"400 MHz, CDCl3",
"1H NMR data":"0.9 (s, 6H, -CH3), 1.2 (q, 3H, -CH3), 1.8–2.2 (s, 4H, -CH2), 3.8 (t, 2H, -CH2), 4.3 (s, 1H, -CH), 7.2–8.4 (m, 6H, Ar-H and -NH2)",
"13C NMR conditions":"100 MHz, CDCl3",
"13C NMR data":"14.1, 18.8, 20.3, 21.9, 28.3, 41.6, 48.7, 60.6, 121.5, 124.9, 126.6, 128.9, 130.1, 132.6, 138.9, 140.6, 142.8, 143.8, 170.2"}

Predicted Result:

{"IUPAC":"2-Amino-4-(5-bromo-1-ethyl-1H-indol-2-yl)-7,7-dimethyl-5-oxo-5,6,7,8-tetrahydro-4H-chromene-3-carbonitrile",
"1H NMR conditions":"400 MHz; CDCl3",
"1H NMR data":"0.9 (s, 6H, -CH3), 1.2 (q, 3H, -CH3), 1.8–2.2 (s, 4H, -CH2), 3.8 (t, 2H, -CH2), 4.3 (s, 1H, -CH), 7.2–8.4 (m, 6H, Ar-H and -NH2)",
"13C NMR conditions":"100 MHz; CDCl3",
"13C NMR data":"14.1, 18.8, 20.3, 21.9, 28.3, 41.6, 48.7, 60.6, 121.5, 124.9, 126.6, 128.9, 130.1, 132.6, 138.9, 140.6, 142.8, 143.8, 170.2"}

Fig. S4 Error cases of NMR data extraction

Error Cases

Example 3:

(E)-4-(hexyloxy)-N-(4-(benzyloxy)benzylidene)aniline I6 Colourless crystals, (93%) yield; m.p. 135 °C. IR (KBr): ῡ 3037 (SP2 =C-H), 2931 (SP3 -C-H) and 1614 (C=N) cm−1.1H NMR (DMSO-d6, 400 MHz): δ 8.54 (s, 1H, CH=N), 7.86 (d, J = 8.4 Hz, 2H, Ar-H), 7.48 (d, J = 7.4 Hz, 2H, Ar-H), 7.42 (t, J = 7.4 Hz, 2H, Ar-H), 7.38–7.32 (m, 1H, Ar-H), 7.24 (d, J = 8.6 Hz, 2H, Ar-H), 7.14 (d, J = 8.4 Hz, 2H, Ar-H), 6.95 (d, J = 8.6 Hz, 2H, Ar-H), 5.20 (s, 2H, OCH2Ph), 3.97 (t, J = 6.4 Hz, 2H, OCH2), 1.70 (m, 2H, CH2), 1.43 (m, 2H, CH2), 1.32 (m, 4H, 2 CH2) and 0.89 (t, J = 6.4 Hz, 3H, CH3) ppm. 13C NMR (DMSO-d6, 101 MHz): δ 161.14, 158.08, 144.80, 137.16, 130.60, 129.86, 128.97, 128.45, 128.28, 127.86, 122.69, 115.54, 115.39, 69.87, 68.12, 31.48, 29.15, 25.67, 22.55 and 14.43 ppm. C26H29NO2 requires: C, 80.57; H, 7.55; N, 3.61% found: C, 80.41; H, 7.72; N, 3.87%. MS M+ at m/z 387.44 (34%).

Ground Truth:

{"IUPAC":"(E)-4-(hexyloxy)-N-(4-(benzyloxy)benzylidene)aniline",
"1H NMR conditions":"DMSO-d6, 400 MHz",
"1H NMR data":"8.54 (s, 1H, CH=N), 7.86 (d, J = 8.4 Hz, 2H, Ar-H), 7.48 (d, J = 7.4 Hz, 2H, Ar-H), 7.42 (t, J = 7.4 Hz, 2H, Ar-H), 7.38– 7.32 (m, 1H, Ar-H), 7.24 (d, J = 8.6 Hz, 2H, Ar-H), 7.14 (d, J = 8.4 Hz, 2H, Ar-H), 6.95 (d, J = 8.6 Hz, 2H, Ar-H), 5.20 (s, 2H, OCH2Ph), 3.97 (t, J = 6.4 Hz, 2H, OCH2), 1.70 (m, 2H, CH2), 1.43 (m, 2H, CH2), 1.32 (m, 4H, 2 CH2), 0.89 (t, J = 6.4 Hz, 3H, CH3)",
"13C NMR conditions":"DMSO-d6, 101 MHz","13C NMR data":"161.14, 158.08, 144.80, 137.16, 130.60, 129.86, 128.97, 128.45, 128.28, 127.86, 122.69, 115.54, 115.39, 69.87, 68.12, 31.48, 29.15, 25.67, 22.55, 14.43"}

Predicted Result:

{"IUPAC":"(E)-4-(hexyloxy)-N-(4-(benzyloxy)benzylidene)aniline",
"1H NMR conditions":"DMSO-d6, 400 MHz",
"1H NMR data":"8.54 (s, 1H, CH=N), 7.86 (d, J = 8.4 Hz, 2H, Ar-H), 7.48 (d, J = 7.4 Hz, 2H, Ar-H), 7.42 (t, J = 7.4 Hz, 2H, Ar-H), 7.38–7.32 (m, 1H, Ar-H), 7.24 (d, J = 8.6 Hz, 2H, Ar-H), 7.14 (d, J = 8.4 Hz, 2H, Ar-H), 6.95 (d, J = 8.6 Hz, 2H, Ar-H), 5.20 (s, 2H, OCH2Ph), 3.97 (t, J = 6.4 Hz, 2H, OCH2), 1.70 (m, 2H, CH2), 1.43 (m, 2H, CH2), 1.32 (m, 4H, 2 CH2), 0.89 (t, J = 6.4 Hz, 3H, CH3)",
"13C NMR conditions":"DMSO-d6, 101 MHz","13C NMR data":"161.14, 158.08, 144.80, 137.16, 130.60, 129.86, 128.97, 128.45, 128.28, 127.86, 122.69, 115.54, 115.39, 69.87, 68.12, 31.48, 29.15, 25.67, 22.55 and 14.43"}

Example 4:

3.2.8. N-(4-Ethylphenyl)-2-((5-(1-(4-Isobutylphenyl)Ethyl)-4-Methyl-4H-1,2,4-Triazol-3-yl)Thio)Acetamide (7c) Off-white, amorphous solid. Yield 75%, m.p 100–102 °C. IR: v (cm−1): 3235, 1682, 1517, 1468, 1320, 695. 1HNMR (500 MHz, DMSO) δ 10.18 (s, 1H), 7.43–7.42 (d, 2H, J = 5.0 Hz), 7.14–7.13 (d, 2H, J = 5.0 Hz), 7.05–7.01, **(m, 4H,)** 4.32–4.29 (q, 1H), 3.98 (s, 2H), 3.26 (s, 3H), 2.39(d, 2H, J = 5.0 Hz), 1.81–1.76 (m, 1H), 1.59 (d, J = 12, 3H), 1.17–1.14 (t, 3H) 0.84 (d, J = 5.0 Hz, 6H). 13C NMR (126 MHz, DMSO) δ 165.44, 157.97, 149.05, 139.46, 139.35, 138.87, 136.40, 129.25, 127.91, 126.68, 119.14, 44.15, 37.84, 35.49, 29.97, 29.52, 27.54, 22.14, 20.89, 15.60 (Figures S15–S17). HRMS (ESI+): m/z calculated for [(C25H32N4OS)+H]+: 437.2375; found: 437.2366 (Figure S29). Analysis calculated for C25H32N4OS, C, 68.77; H, 7.39; N, 12.83; S, 7.34.

Ground Truth:

{"IUPAC":"N-(4-Ethylphenyl)-2-((5-(1-(4-Isobutylphenyl)Ethyl)-4-Methyl-4H-1,2,4-Triazol-3-yl)Thio)Acetamide",
"1H NMR conditions":"500 MHz, DMSO",
"1H NMR data":"10.18 (s, 1H), 7.43–7.42 (d, 2H, J = 5.0 Hz), 7.14–7.13 (d, 2H, J = 5.0 Hz), **7.05–7.01, (m, 4H,),** 4.32–4.29 (q, 1H), 3.98 (s, 2H), 3.26 (s, 3H), 2.39(d, 2H, J = 5.0 Hz), 1.81–1.76 (m, 1H), 1.59 (d, J = 12, 3H), 1.17–1.14 (t, 3H) 0.84 (d, J = 5.0 Hz, 6H)",
"13C NMR conditions":"126 MHz, DMSO",
"13C NMR data":"165.44, 157.97, 149.05, 139.46, 139.35, 138.87, 136.40, 129.25, 127.91, 126.68, 119.14, 44.15, 37.84, 35.49, 29.97, 29.52, 27.54, 22.14, 20.89, 15.60"}

Predicted Result:

{"IUPAC":"N-(4-Ethylphenyl)-2-((5-(1-(4-Isobutylphenyl)Ethyl)-4-Methyl-4H-1,2,4-Triazol-3-yl)Thio)Acetamide",
"1H NMR conditions":"500 MHz, DMSO",
"1H NMR data":"10.18 (s, 1H), 7.43–7.42 (d, 2H, J = 5.0 Hz), 7.14–7.13 (d, 2H, J = 5.0 Hz), 7.05–7.01, **(m, 4H,)** 4.32–4.29 (q, 1H), 3.98 (s, 2H), 3.26 (s, 3H), 2.39(d, 2H, J = 5.0 Hz), 1.81–1.76 (m, 1H), 1.59 (d, J = 12, 3H), 1.17–1.14 (t, 3H) 0.84 (d, J = 5.0 Hz, 6H)",
"13C NMR conditions":"126 MHz, DMSO",
"13C NMR data":"165.44, 157.97, 149.05, 139.46, 139.35, 138.87, 136.40, 129.25, 127.91, 126.68, 119.14, 44.15, 37.84, 35.49, 29.97, 29.52, 27.54, 22.14, 20.89, 15.60"}

Fig. S5 Error cases of NMR data extraction

```
from rdkit import Chem

# Example SMILES string
smiles = "C[C@H](O)[C@H]1[C@@](C)([C@H]2OC(C)(C)OC2)CCC1"

# Convert SMILES to a molecule object
mol=Chem.MolFromSmiles(smiles)

# Standardize the SMILES string with stereochemistry preserved
canonical_smiles = Chem.MolToSmiles(mol, isomericSmiles=True)

print("Canonical SMILES with stereochemistry:",canonical_smiles)

Result: C[C@H](O)[C@@H]1CCC[C@]1(C)[C@@H]1COC(C)(C)O1
```

Fig. S6 Standardization of SMILES Using RDKit

Journal name and count

Journal	Count
Molecules	349
Int J Mol Sci	137
Pharmaceuticals (Basel)	39
Nat Commun	33
Polymers (Basel)	33
Sci Rep	31
Pharmaceutics	25
Mar Drugs	20
Front Pharmacol	19
Antioxidants (Basel)	15
Front Chem	12
Materials (Basel)	12
Front Microbiol	11
J Med Chem	11
Plants (Basel)	9
Acta Crystallogr E Crystallogr Commun	9
Cancers (Basel)	9
Adv Sci (Weinh)	9
Antibiotics (Basel)	8
J Nat Prod	7
Biomedicines	7
Biomolecules	7
Sensors (Basel)	7
ACS Omega	6
Nanomaterials (Basel)	6
Metabolites	6
Acta Crystallogr C Struct Chem	6
ACS Appl Mater Interfaces	6
Foods	6
Beilstein J Org Chem	5
Gels	5
Chemistry	5
PLoS One	5
J Fungi (Basel)	5
Biosensors (Basel)	5
Nutrients	5
Membranes (Basel)	4
Nature	3
Front Nutr	3

Journal	Count
Front Mol Biosci	3
Bioconjug Chem	3
Curr Issues Mol Biol	3
J Org Chem	3
Mar Life Sci Technol	3
IUCrdata	3
Bioorg Med Chem	3
Viruses	3
Front Oncol	3
bioRxiv	3
Angew Chem Int Ed Engl	3
Microorganisms	3
BMC Complement Med Ther	3
Res Sq	3
Life (Basel)	2
Pathogens	2
BMC Chem	2
PLoS Biol	2
ACS Infect Dis	2
J Nematol	2
Food Sci Nutr	2
Chemphyschem	2
Front Plant Sci	2
J Am Chem Soc	2
Inorg Chem	2
J Phys Chem B	2
ACS Chem Neurosci	2
Heliyon	2
J Labelled Comp Radiopharm	2
ACS Appl Energy Mater	2
J Biomed Sci	2
Chem Mater	2
J IRAN CHEM SOC	1
Nat Prod Bioprospect	1
Front Cell Infect Microbiol	1
Chem Asian J	1
J Phys Chem Lett	1
Toxins (Basel)	1
Glob Chall	1

Journal	Count
Bioact Mater	1
RPS Pharm Pharmacol Rep	1
Front Immunol	1
Curr Radiopharm	1
ACS Appl Polym Mater	1
Chin Med	1
JACS Au	1
ACS Appl Nano Mater	1
Cells	1
Theranostics	1
Arch Microbiol	1
Nucleic Acids Res	1
Pharmacol Rep	1
Cancer Drug Resist	1
Biomacromolecules	1
Microb Cell Fact	1
Discov Nano	1
ACS Catal	1
Nat Chem	1
Braz J Med Biol Res	1
Environ Sci Pollut Res Int	1
ACS Sens	1
Toxics	1
Cell Death Dis	1
J Chem Ecol	1
ChemMedChem	1
PLoS Negl Trop Dis	1
Bioeng Transl Med	1
Inflamm Res	1
Commun Chem	1
J Biol Inorg Chem	1
Iran J Med Sci	1
Chem Res Toxicol	1
ChemistryOpen	1
BMC Cancer	1
Electrophoresis	1
Micromachines (Basel)	1

Fig. S7 The 115 journals and their data counts for test set 2 (1022)

1. Regular expression for extracting H spectrum

- pattern_H = re.compile('(1\s*H.{0,3}?NMR((?!\s*NMR).)*?[;.]\s+)(.{0,8}?(NMR|IR|ESI|MS|Anal|Mass|Found|UV|HMBC|NOESY|31P|/n))', re.I)

2. Regular expression for extracting C spectrum

- pattern_C = re.compile('(13\s*C.{0,7}?NMR((?!\sNMR).)*?[;.]\s+)(.{0,8}?(NMR|IR|ESI|MS|Anal|Mass|Found|UV|HMBC|NOESY|31P|/n))', re.I)

3. To ensure the quality of the extracted spectrum, it is necessary to filter shorter sentences

- pattern_filter = re.compile('([^0-9]{30,})', re.I)

4. Used to find complete sentences

- pattern_compound_1 = re.compile('^.*?[;.:]', re.I)

5. Find compound sentences with brackets like xxx (x) and truncate them with (

- pattern_filter_again = re.compile('^.*\s+\(.{0,4}\)[\s:.]', re.I)

6. Determine whether the extracted H NMR is followed by brackets, which are usually solvents

- find_solution_H = re.compile('1\s*H.{0,3}?NMR\s*(\(.*?\))', re.I)

7. Determine whether the extracted C NMR is followed by brackets, which are usually solvents

- find_solution_C = re.compile('13\s*C.{0,7}?NMR\s*(\(.*?\))', re.I)

8. Delete(), extract() contents

- pattern_delete = re.compile(r'\s+\([^)]*\)$',re.I)

9. Only match integers or decimals

- pattern_number = re.compile(r'^[+-]?[^()a-zA-Z]*$',re.I)

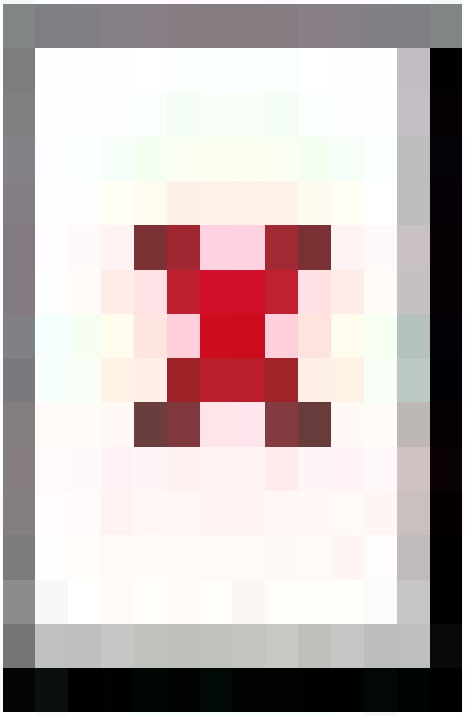
10. Matches a string that starts with a single digit and can be followed by any number of characters

- pattern_H_NMR_data_first = re.compile(r'[0-9].*', re.I)

11. Used to match a number (with a decimal part) that begins with an optional sign.

- pattern_singular = re.compile(r'[+-]?\d+(\.\d+)?\s?\([^)]*\)',re.I)

Fig. S8 Some regular expressions used by rule-based methods



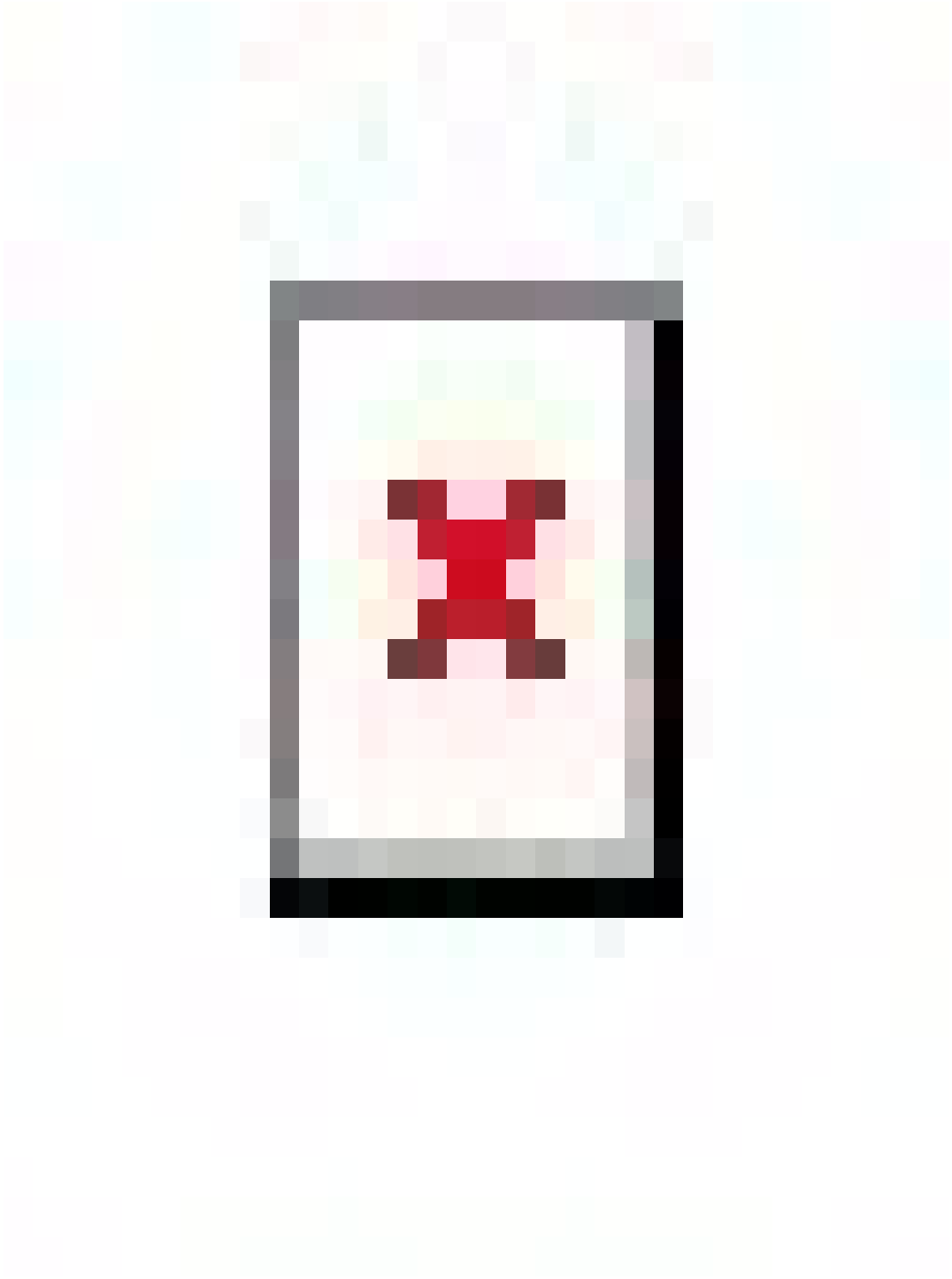


Fig. S9 618 common compound group words and atomic count

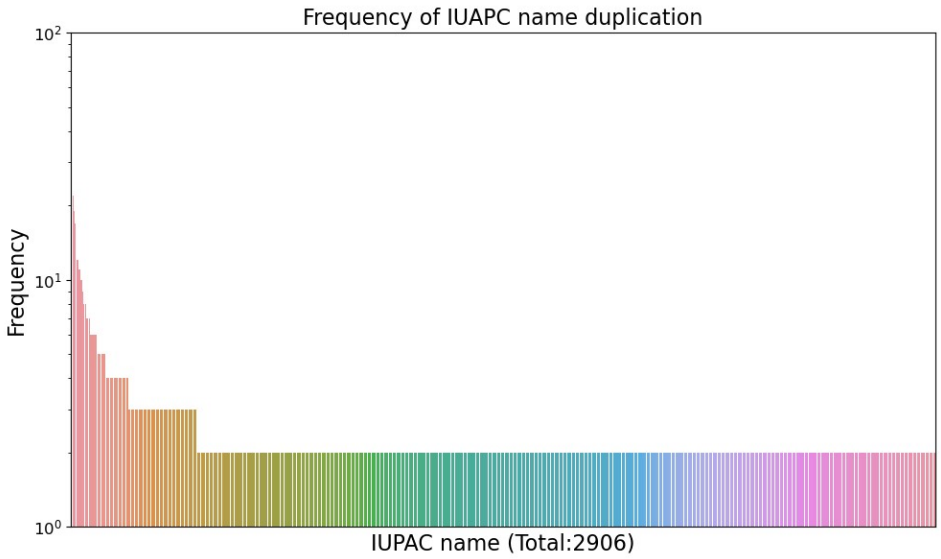


Fig. S10 The frequency distribution of these duplicate IUPAC names

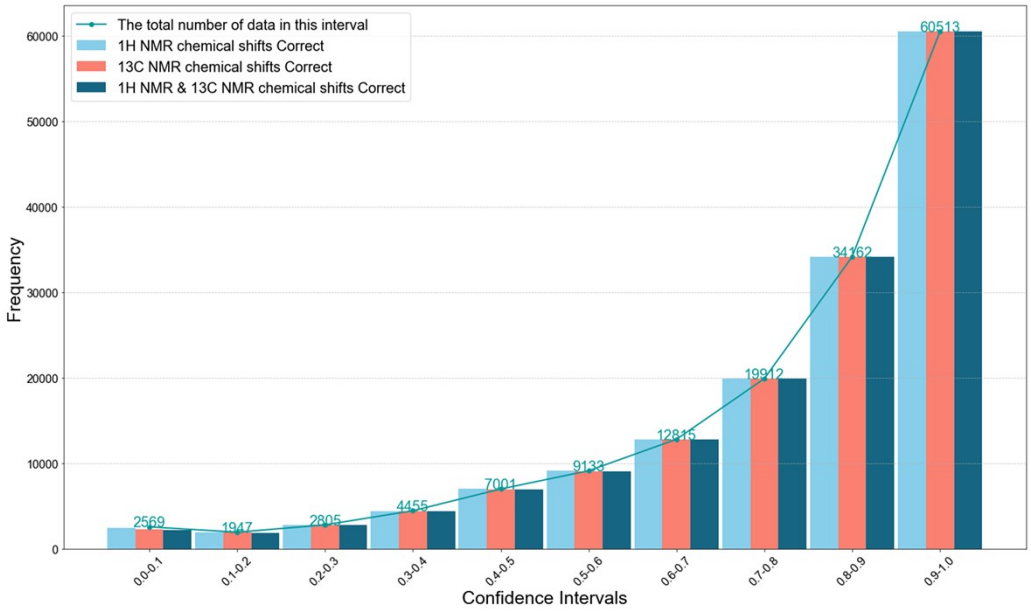


Fig. S11 Accuracy of NMR chemical shift values and their order in NMRBank

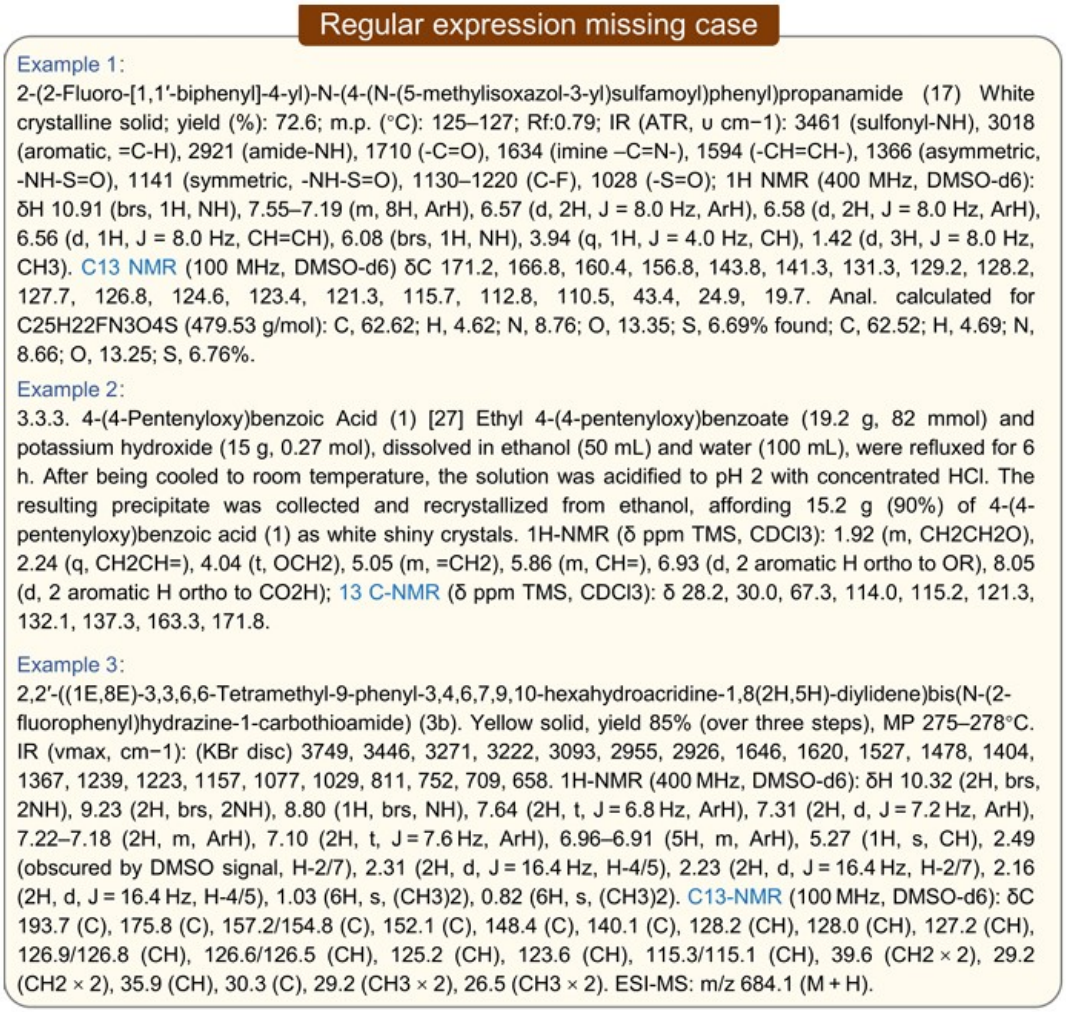


Fig. S12 Regular expression missing case

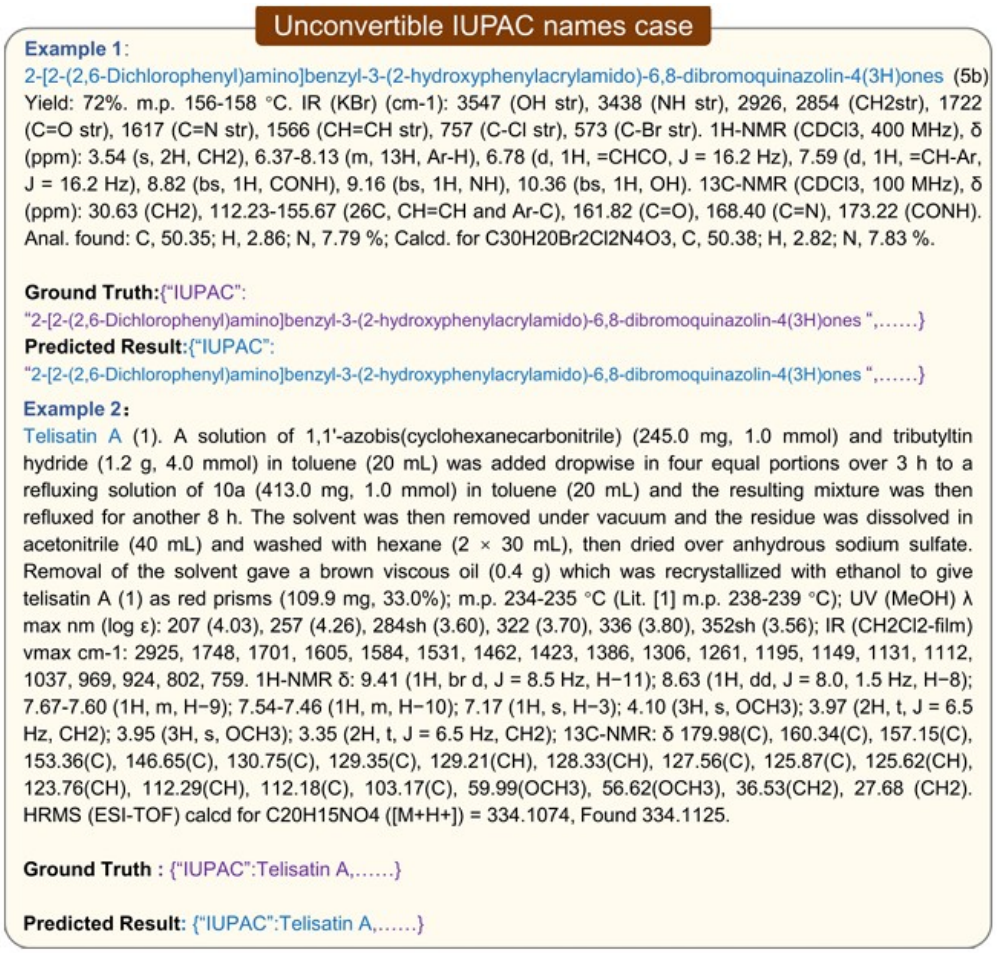


Fig. S13 Unconvertible IUPAC names case

Table. S1 Hyperparameters for the models

Model	Hyperparameters
Llama2-13b-chat (q-lora fine-tune)	epoch; lora_r 64; lora_alpha 128; learning rate 1e-4;
Llama3-8b-instruct (full fine-tune)	epoch; batch size 2; learning rate 5e-6;
Mistral-7b-instruct-v0.2 (full fine-tune)	epoch; batch size 2; learning rate 5e-6;

Table. S2 Fine-tuning and Inferencing Cost

Model	Strategy	Number of Fine-tuning Data	Fine-tuning Cost	Hyperparameters	Test Set 1 (300) Inferencing Cost	Test Set 2 (1022) Inferencing Cost
Mistral-7b-Instruct-v0.2	full fine-tuning	800	about 1 epoch × 12 min/epoch on 4×40GB A100 (~38 GB/GPU)	lr = 5e-6, bs = 1, max_length = 4096	2 min on 1×40GB A100 (using vllm)	9 min on 1×40GB A100 (using vllm)
Llama-3-8b-Instruct	full fine-tuning	800	about 1 epoch × 12 min/epoch on 4×40GB A100 (~33 GB/GPU)	lr = 5e-6, bs = 1, max_length = 4096	2 min on 1×40GB A100 (using vllm)	9 min on 1×40GB A100 (using vllm)
Lama-2-13b-chat-hf	q-lora fine-tuning	800	about 1 epoch × 26 min/epoch on 1×40GB A100 (~20 GB/GPU)	lr = 1e-4, bs = 2, max_length = 4096	14 min on 1×40GB A100 (using vllm)	50 min on 1×40GB A100 (using vllm)

Use 4×40GB A100 for full parameter fine-tuning and 1×40GB A100 for Q-LoRA fine-tuning.

Use vllm on the 1×40GB A100 for inference acceleration.

The average inference speed of Llama3-8b-instruct and Mistral-7b-instruct-v0.2 is 1 second for 2 items.

Table. S3 Exact match accuracy of open source LLMs on Test Set 1(300)

Model	IUPAC	^{1}H NMR conditions	^{1}H NMR shifts	^{13}C NMR conditions	^{13}C NMR shifts
Llama2-13b-insturct	**96.6**	**99.3**	96.3	**99.3**	98.6
Llama3-8b-insturct	95.3	99.0	97.0	99.3	94.0
Mistral-7b-insturct-v0.2	96.3	99.0	**97.3**	99.0	**99.3**

Table. S4 Exact match accuracy of open source LLMs on Test Set 2 (1022)

Model	IUPAC	^{1}H NMR conditions	^{1}H NMR shifts	^{13}C NMR conditions	^{13}C NMR shifts
Llama2-13b-insturct	82.9	93.9	87.6	94.2	91.9
Llama3-8b-insturct	78.7	93.4	84.4	93.7	82.5
Mistral-7b-insturct-v0.2	**86.2**	**94.9**	**88.4**	**95.0**	**92.0**

Table. S5 Exact match accuracy in three-fold cross-validation at different confidences

Confidence Intervals	IUPAC name	^{1}H NMR conditions	^{1}H NMR shifts	^{13}C NMR conditions	^{13}C NMR shifts
0-0.2	0.734±0.035	0.874±0.055	0.456±0.05	0.952±0.034	0.609±0.05
0.2-0.4	0.774±0.036	0.951±0.035	0.840±0.042	0.951±0.035	0.811±0.139
0.4-0.6	0.795±0.044	0.949±0.041	0.848±0.096	0.949±0.008	0.949±0.016
0.6-0.8	0.878±0.014	0.958±0.023	0.872±0.025	0.945±0.024	0.977±0.001
0.8-1	0.956±0.018	0.968±0.004	0.967±0.013	0.973±0.008	0.996±0.002
All	0.906±0.008	0.953±0.01	0.907±0.01	0.964±0.007	0.963±0.006

Table. S6 Table S6 Most frequent duplicate IUPAC names and their frequency

Duplicate IUPAC names	Frequency
quercetin	73
lupeol	36
ursolic acid	36
kaempferol	33
gallic acid	30
stigmasterol	28
scopoletin	22
luteolin	20
pyrimidine	19
protocatechuic acid	19
oleanolic acid	19
methyl gallate	18
caffeic acid	17
apigenin	16
benzoic acid	16
4-hydroxybenzoic acid	16
ferulic acid	15
catechin	15
betulinic acid	15
vanillic acid	12
emodin	12
curcumin	12
quercetin-3-o-β-d-glucopyranoside	11
betulin	11
palmitic acid	11
lupeol acetate	11
ergosterol	10
daidzein	10
ellagic acid	10
biphenyl	10
p-hydroxybenzoic acid	9
linoleic acid	9
p-hydroxybenzaldehyde	9
naringenin	8
genistein	8
triazine	8
isorhamnetin	8
epicatechin	8
p-coumaric acid	8
(+)-catechin	8
2,3-diaminopropanol	8
quercetin 3-o-β-d-glucopyranoside	7
acacetin	7
benzamide	7

2-methyl-5-nitrobenzamide	7
(-)-epicatechin	7
4-methoxy-1,1'-biphenyl	7
pinostrobin	7
quercetin-3-o-β-d-galactopyranoside	7
imidazopyrazine	7
ether	6
caffeine	6
(3'-(benzo-1,4-dioxan-6-yl)-2'-methyl-[1,1'-biphenyl]-4-yl)methanol	6
benzyl alcohol	6
benzyl acetate	6
n-trans-feruloyltyramine	6
ethyl gallate	6
lupenone	6
3-o-methylquercetin	6
hexadecanoic acid	6
gallocatechin	6
methyl oleate	6
triazole	6
cholesterol	6
luteolin-7-o-glucoside	6
pinocembrin	6
catechol	6
acetophenone	6
benzophenone	6
4-methylbiphenyl	6
methyl benzoate	6
benzoxazolyl-alanine	5
myricetin	5
benzaldehyde	5
4-cyanobiphenyl	5
piperonal	5
cyclohexylamine	5
quinazolin-6(7 h)-one	5
penicillic acid	5
sucrose	5
1,2,3,4,6-penta-o-galloyl-β-d-glucopyranose	5
2-(4-chlorophenyl)-2,3-dihydroquinazolin-4(1h)-one	5
cinnamic acid	5
2-(4-methoxyphenyl)imidazo[1,2-a]pyridine	5
kaempferol-3-o-α-l-rhamnopyranoside	5
3-hydroxy-6h-benzo[c]chromen-6-one	5
allylic alcohol	5
silyl ether	5
3,3'-((4-nitrophenyl)methylene)bis(1h-indole)	5
kaempferol-3-o-β-d-glucopyranoside	5
linalool	5

benzothiazole	5
2-phenylquinazolin-4(3h)-one	5
chalcone	5
taxifolin	5
5-hydroxymethylfurfural	5
phenylamine	5
dibutyl phthalate	5
guanidine	5
4-phenylphenol	5
methyl cinnamate	5
eriodictyol	5
oleic acid	5

The confidence of the model predictions

The confidence of the model predictions is computed using the cumulative log probability of the predicted tokens.

Specifically, the Large Language Model (LLM) calculates the probability accumulation using the following formula:

$$F = \frac{1}{n}\sum_{i=1}^{n} log\left(P\left(w_i|w_{1,}\ w_2,\ldots,w_{i-1}\right)\right)$$

Where $P\left(w_i|w_{1,}\ w_2,\ldots,w_{i-1}\right)$ represents the probability of generating the current word w_i given the previous words (or subwords).

After obtaining the cumulative log probability, we convert it to a confidence score using:

$$P = e^F$$

This transformation results in a confidence value between 0 and 1, indicating the model's confidence in its predictions.