

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}

^aNanjing University of Chinese Medicine, 138 Xianlin Road, Nanjing 210023, China.

^bDrug 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.

^cUniversity of Chinese Academy of Sciences, No. 19A Yuquan Road, Beijing 100049, China.

^dShanghaiTech University, Shanghai 201210, China.

^eLingang Laboratory, Shanghai 200031, China.

^fProtonUnfold 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
---	----

Original TxT Article:

.....
115.6, 99.3, 71.4, 71.2. Quaternary carbon missing. HRMS (ES+) m/z calc. for C₁₆H₁₃CIN₂O₃ [M + H]⁺: 317.0688; found: 317.0698.

2-(((4-Chlorobenzyl)oxy)methyl)-9-hydroxy-4H-pyrido [1,2-a]pyrimidin-4-one (36)

The title compound was obtained as a white solid (8 mg, 0.025 mmol, 10.7%). ¹H NMR (600 MHz, DMSO-d₆) δ 8.46 (dd, J = 6.9, 1.4 Hz, 1H), 7.45 (s, 4H), 7.25–7.15 (m, 2H), 6.43 (s, 1H), 4.66 (s, 2H), 4.56 (s, 2H). ¹³C NMR (151 MHz, DMSO-d₆) δ 162.8, 157.4, 144.9, 137.1, 132.1, 129.4, 128.4, 117.1, 116.2, 115.6, 99.3, 71.4, 71.2. Quaternary carbon missing. HRMS (ES+) m/z calc. for C₁₆H₁₃CIN₂O₃ [M + H]⁺: 317.0688; found: 317.0698.

9-Hydroxy-2-(((3-hydroxybenzyl)oxy)methyl)-4H-pyrido [1,2-a]pyrimidin-4-one (37)

NMR paragraph:

Steps 1-2

2-(((4-Chlorobenzyl)oxy)methyl)-9-hydroxy-4H-pyrido [1,2-a]pyrimidin-4-one (36) The title compound was obtained as a white solid (8 mg, 0.025 mmol, 10.7%). ¹H NMR (600 MHz, DMSO-d₆) δ 8.46 (dd, J = 6.9, 1.4 Hz, 1H), 7.45 (s, 4H), 7.25–7.15 (m, 2H), 6.43 (s, 1H), 4.66 (s, 2H), 4.56 (s, 2H). ¹³C NMR (151 MHz, DMSO-d₆) δ 162.8, 157.4, 144.9, 137.1, 132.1, 129.4, 128.4, 117.1, 116.2, 115.6, 99.3, 71.4, 71.2. Quaternary carbon missing. HRMS (ES+) m/z calc. for C₁₆H₁₃CIN₂O₃ [M + H]⁺: 317.0688; found: 317.0698.

Steps 3-5

NMR data:

"IUPAC": "2-(((4-Chlorobenzyl)oxy)methyl)-9-hydroxy-4H-pyrido [1,2-a]pyrimidin-4-one",
"SMILES": "O=C1C=C(COCC2=CC=C(C1)C=C2)N=C3N1C=CC=C3O",
"1H NMR conditions": "600 MHz, DMSO-d₆",
"1H NMR chemical shifts": "8.46 (dd, J = 6.9, 1.4 Hz, 1H), 7.45 (s, 4H), 7.25–7.15 (m, 2H), 6.43 (s, 1H), 4.66 (s, 2H), 4.56 (s, 2H)",
"13C NMR conditions": "151 MHz, DMSO-d₆",
"13C NMR chemical shifts": "162.8, 157.4, 144.9, 137.1, 132.1, 129.4, 128.4, 117.1, 116.2, 115.6, 99.3, 71.4, 71.2"

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

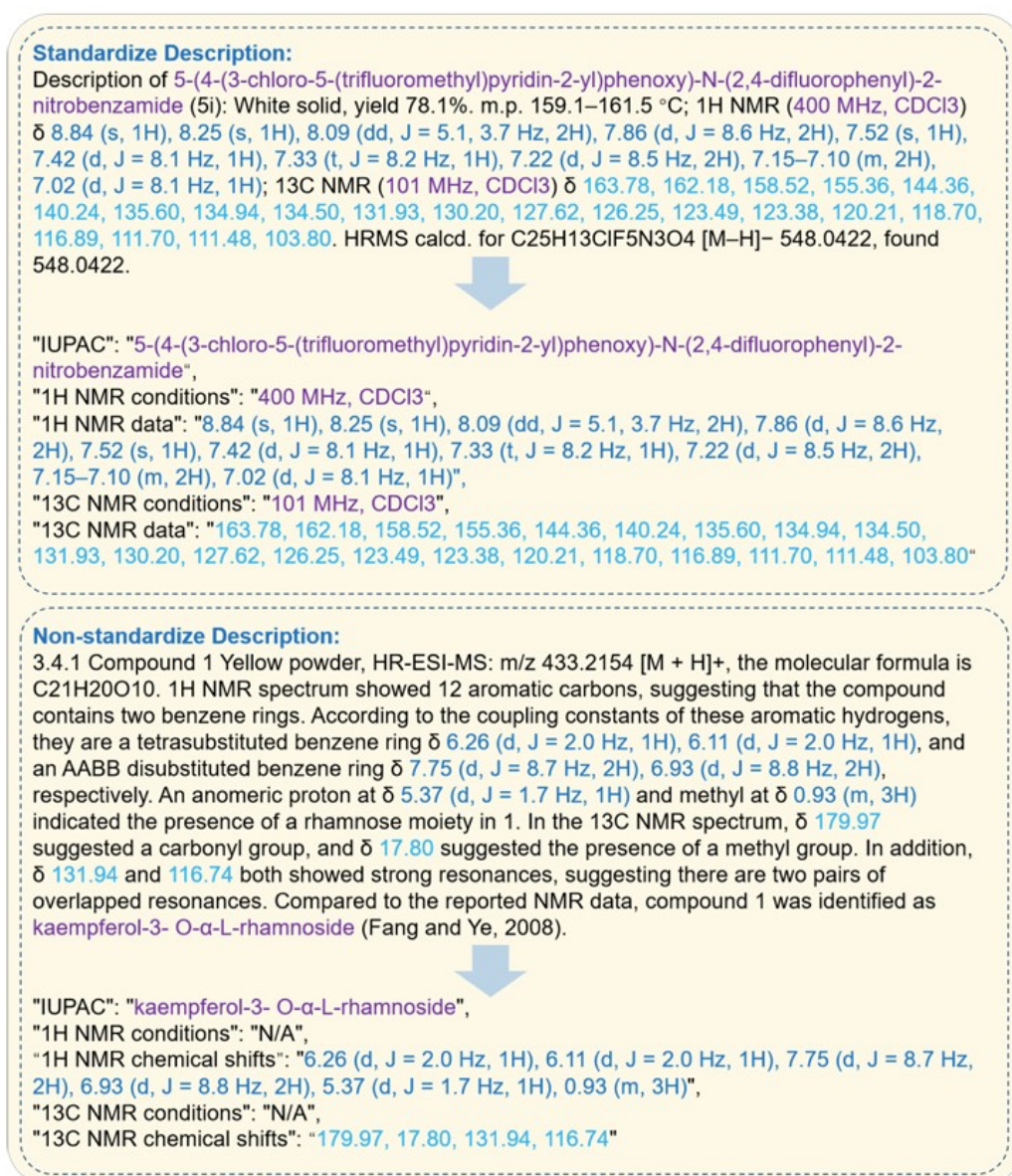


Fig. S3 Examples of text annotations in the test set 2 (1022)

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 **16** Colourless crystals, (93%) yield; m.p. 135 °C. IR (KBr): $\bar{\nu}$ 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: ν (cm⁻¹): 3235, 1682, 1517, 1468, 1320, 695. ¹H NMR (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 [(C₂₅H₃₂N₄O₅)+H]⁺: 437.2375; found: 437.2366 (Figure S29). Analysis calculated for C₂₅H₃₂N₄O₅, 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

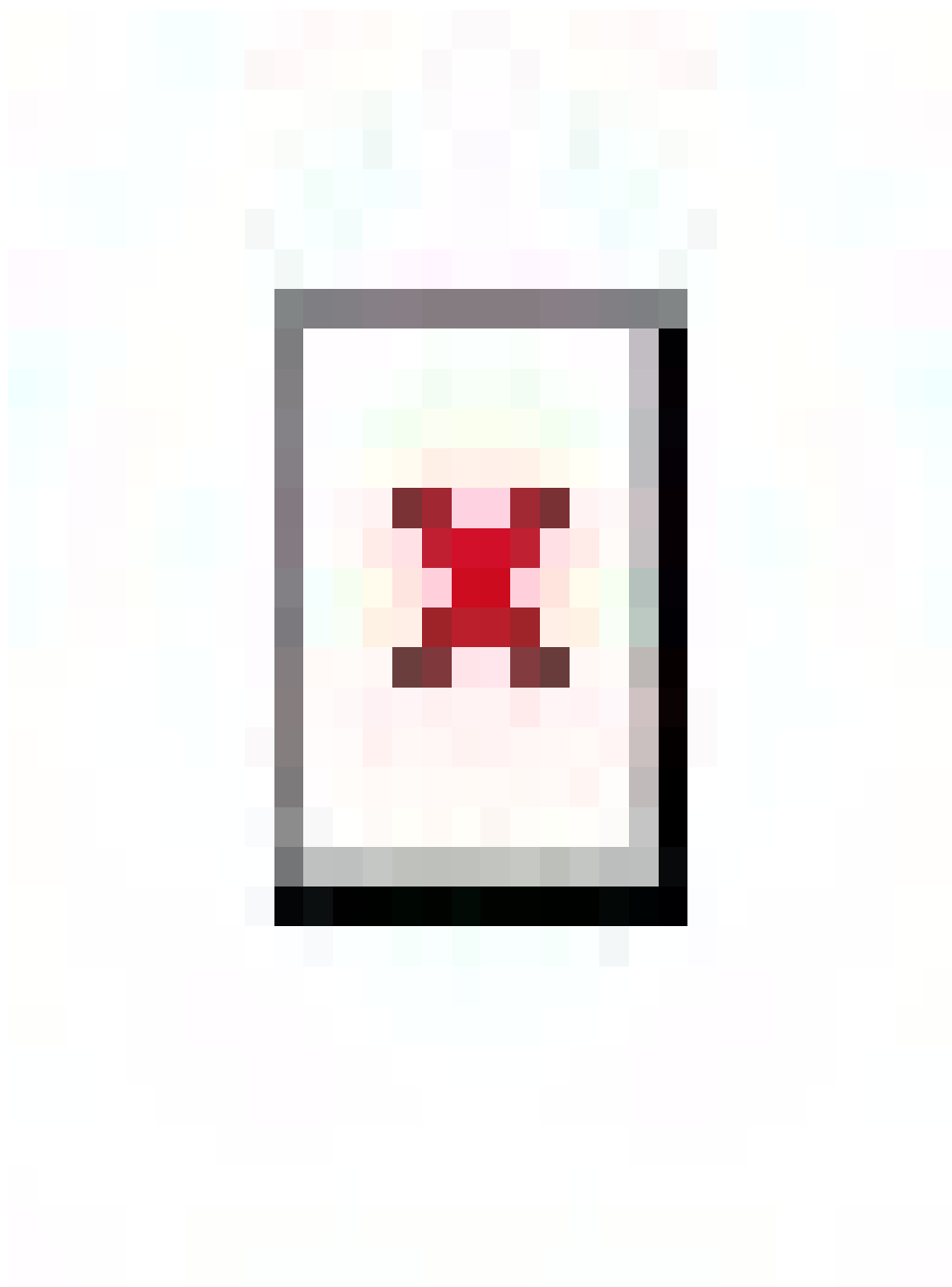
Journal name and count

Journal	Count	Journal	Count	Journal	Count
Molecules	349	Front Mol Biosci	3	Bioact Mater	1
Int J Mol Sci	137	Bioconjug Chem	3	RPS Pharm Pharmacol Rep	1
Pharmaceuticals (Basel)	39	Curr Issues Mol Biol	3	Front Immunol	1
Nat Commun	33	J Org Chem	3	Curr Radiopharm	1
Polymers (Basel)	33	Mar Life Sci Technol	3	ACS Appl Polym Mater	1
Sci Rep	31	IUCrdata	3	Chin Med	1
Pharmaceutics	25	Bioorg Med Chem	3	JACS Au	1
Mar Drugs	20	Viruses	3	ACS Appl Nano Mater	1
Front Pharmacol	19	Front Oncol	3	Cells	1
Antioxidants (Basel)	15	bioRxiv	3	Theranostics	1
Front Chem	12	Angew Chem Int Ed Engl	3	Arch Microbiol	1
Materials (Basel)	12	Microorganisms	3	Nucleic Acids Res	1
Front Microbiol	11	BMC Complement Med Ther	3	Pharmacol Rep	1
J Med Chem	11	Res Sq	3	Cancer Drug Resist	1
Plants (Basel)	9	Life (Basel)	2	Biomacromolecules	1
Acta Crystallogr E Crystallogr Commun	9	Pathogens	2	Microb Cell Fact	1
Cancers (Basel)	9	BMC Chem	2	Discov Nano	1
Adv Sci (Weinh)	9	PLoS Biol	2	ACS Catal	1
Antibiotics (Basel)	8	ACS Infect Dis	2	Nat Chem	1
J Nat Prod	7	J Nematol	2	Braz J Med Biol Res	1
Biomedicines	7	Food Sci Nutr	2	Environ Sci Pollut Res Int	1
Biomolecules	7	Chemphyschem	2	ACS Sens	1
Sensors (Basel)	7	Front Plant Sci	2	Toxics	1
ACS Omega	6	J Am Chem Soc	2	Cell Death Dis	1
Nanomaterials (Basel)	6	Inorg Chem	2	J Chem Ecol	1
Metabolites	6	J Phys Chem B	2	ChemMedChem	1
Acta Crystallogr C Struct Chem	6	ACS Chem Neurosci	2	PLoS Negl Trop Dis	1
ACS Appl Mater Interfaces	6	Heliyon	2	Bioeng Transl Med	1
Foods	6	J Labelled Comp Radiopharm	2	Inflamm Res	1
Beilstein J Org Chem	5	ACS Appl Energy Mater	2	Commun Chem	1
Gels	5	J Biomed Sci	2	J Biol Inorg Chem	1
Chemistry	5	Chem Mater	2	Iran J Med Sci	1
PLoS One	5	J IRAN CHEM SOC	1	Chem Res Toxicol	1
J Fungi (Basel)	5	Nat Prod Bioprospect	1	ChemistryOpen	1
Biosensors (Basel)	5	Front Cell Infect Microbiol	1	BMC Cancer	1
Nutrients	5	Chem Asian J	1	Electrophoresis	1
Membranes (Basel)	4	J Phys Chem Lett	1	Micromachines (Basel)	1
Nature	3	Toxins (Basel)	1		
Front Nutr	3	Glob Chall	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((?!s*NMR).)*?[:.]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'[+-]?[0-9]+(\.[0-9]+)?\s?([^\s])', re.I)`

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



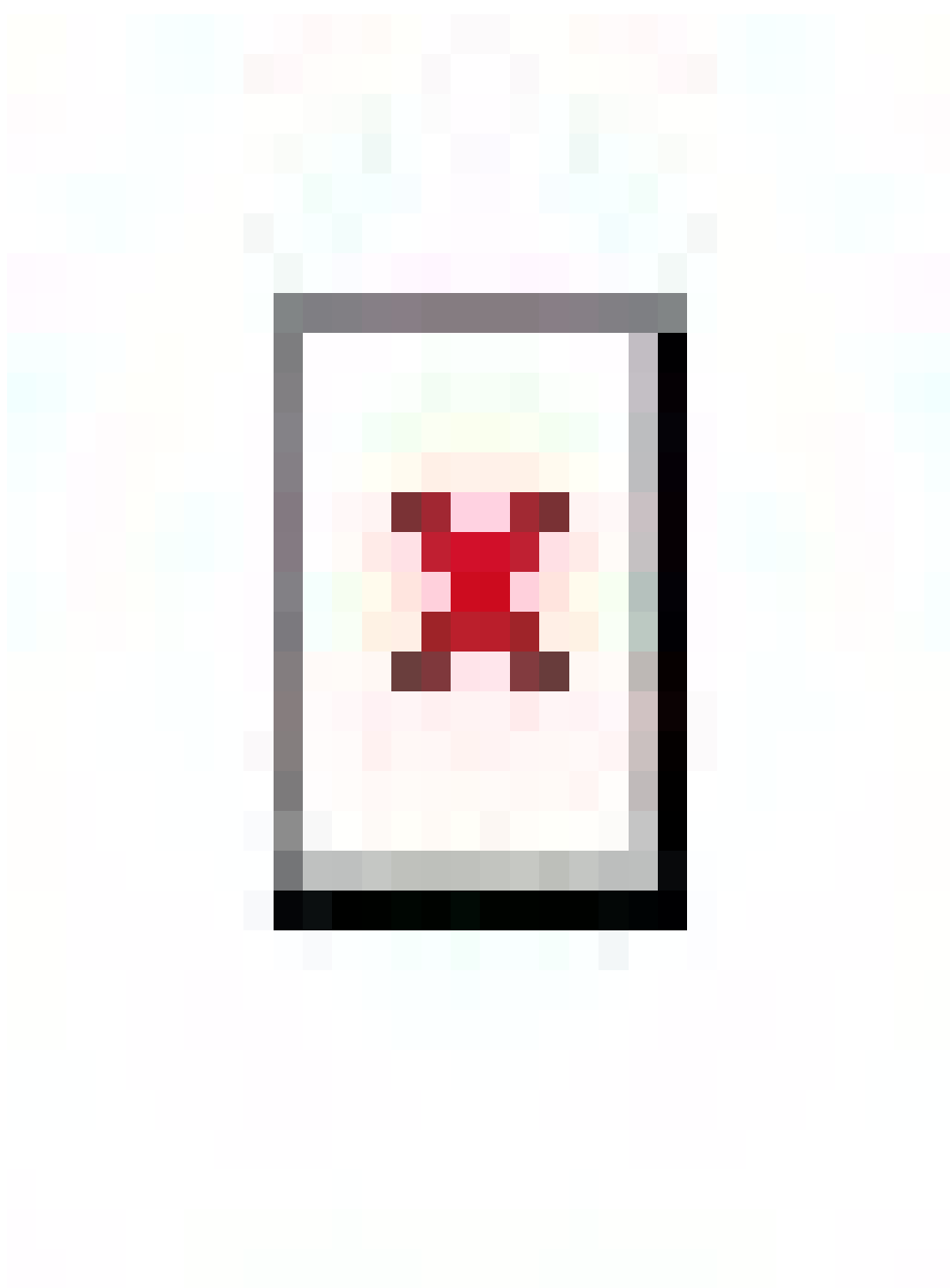


Fig. S9 618 common compound group words and atomic count

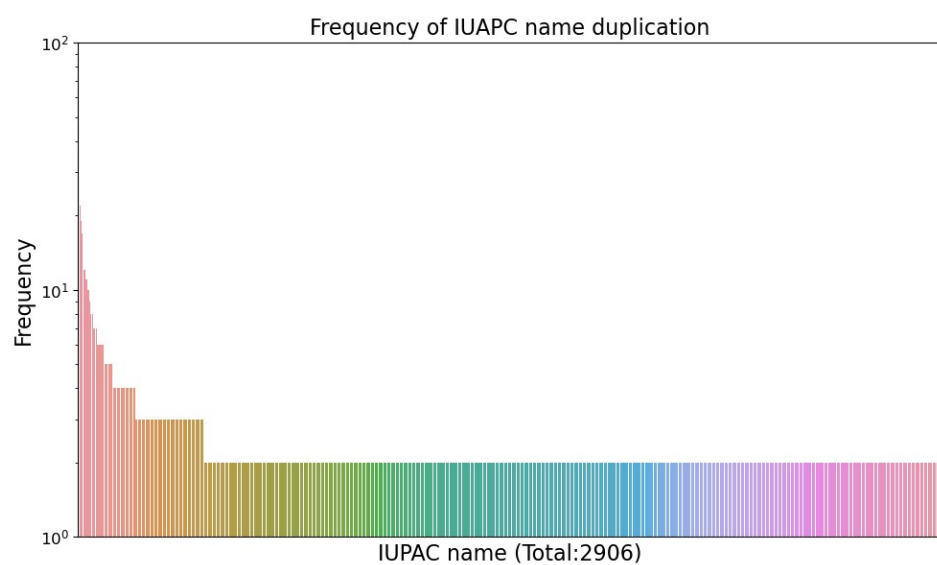


Fig. S10 The frequency distribution of these duplicate IUPAC names

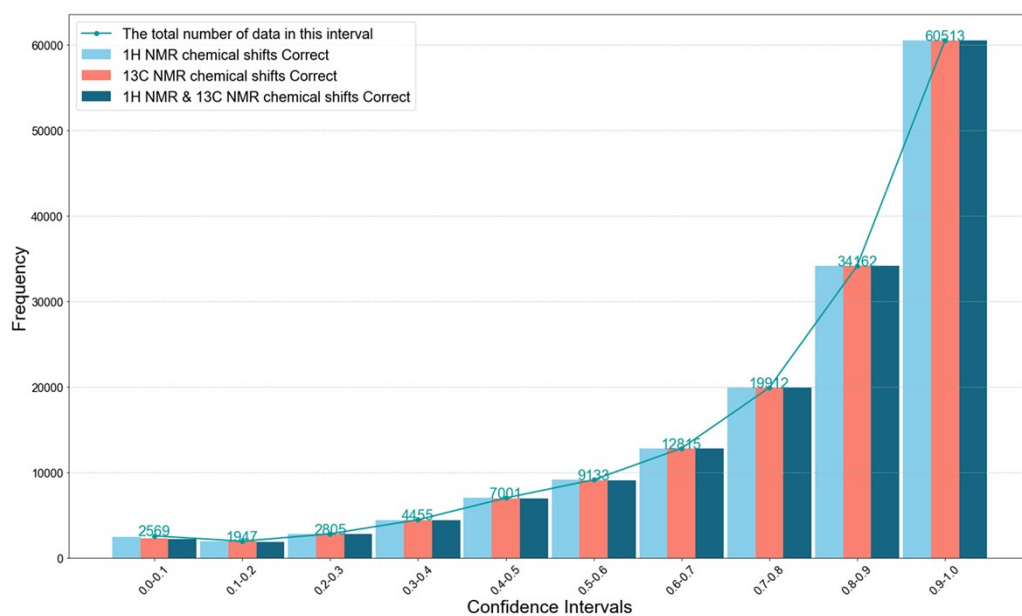


Fig. S11 Accuracy of NMR chemical shift values and their order in NMRBank

Regular expression missing case

Example 1:

2-(2-Fluoro-[1,1'-biphenyl]-4-yl)-N-(4-(N-(5-methylisoxazol-3-yl)sulfamoyl)phenyl)propanamide (17) White crystalline solid; yield (%): 72.6; m.p. (°C): 125–127; Rf:0.79; IR (ATR, ν cm⁻¹): 3461 (sulfonyl-NH), 3018 (aromatic, =C-H), 2921 (amide-NH), 1710 (-C=O), 1634 (imine -C=N-), 1594 (-CH=CH-), 1366 (asymmetric, -NH-S=O), 1141 (symmetric, -NH-S=O), 1130–1220 (C-F), 1028 (-S=O); ¹H NMR (400 MHz, DMSO-d₆): δ 10.91 (brs, 1H, NH), 7.55–7.19 (m, 8H, ArH), 6.57 (d, 2H, J = 8.0 Hz, ArH), 6.58 (d, 2H, J = 8.0 Hz, ArH), 6.56 (d, 1H, J = 8.0 Hz, CH=CH), 6.08 (brs, 1H, NH), 3.94 (q, 1H, J = 4.0 Hz, CH), 1.42 (d, 3H, J = 8.0 Hz, CH₃). ¹³C NMR (100 MHz, DMSO-d₆) δ 171.2, 166.8, 160.4, 156.8, 143.8, 141.3, 131.3, 129.2, 128.2, 127.7, 126.8, 124.6, 123.4, 121.3, 115.7, 112.8, 110.5, 43.4, 24.9, 19.7. Anal. calculated for C₂₅H₂₂FN₃O₄S (479.53 g/mol): C, 62.62; H, 4.62; N, 8.76; O, 13.35; S, 6.69% found; C, 62.52; H, 4.69; N, 8.66; O, 13.25; S, 6.76%.

Example 2:

3.3.3. 4-(4-Pentenyl)benzoic Acid (1) [27] Ethyl 4-(4-pentenyl)benzoate (19.2 g, 82 mmol) and potassium hydroxide (15 g, 0.27 mol), dissolved in ethanol (50 mL) and water (100 mL), were refluxed for 6 h. After being cooled to room temperature, the solution was acidified to pH 2 with concentrated HCl. The resulting precipitate was collected and recrystallized from ethanol, affording 15.2 g (90%) of 4-(4-pentenyl)benzoic acid (1) as white shiny crystals. ¹H-NMR (δ ppm TMS, CDCl₃): 1.92 (m, CH₂CH₂O), 2.24 (q, CH₂CH=), 4.04 (t, OCH₂), 5.05 (m, =CH₂), 5.86 (m, CH=), 6.93 (d, 2 aromatic H ortho to OR), 8.05 (d, 2 aromatic H ortho to CO₂H); ¹³C-NMR (δ ppm TMS, CDCl₃): δ 28.2, 30.0, 67.3, 114.0, 115.2, 121.3, 132.1, 137.3, 163.3, 171.8.

Example 3:

2,2'-((1E,8E)-3,3,6,6-Tetramethyl-9-phenyl-3,4,6,7,9,10-hexahydroacridine-1,8(2H,5H)-diylidene)bis(N-(2-fluorophenyl)hydrazine-1-carbothioamide) (3b). Yellow solid, yield 85% (over three steps), MP 275–278°C. IR (ν max, cm⁻¹): (KBr disc) 3749, 3446, 3271, 3222, 3093, 2955, 2926, 1646, 1620, 1527, 1478, 1404, 1367, 1239, 1223, 1157, 1077, 1029, 811, 752, 709, 658. ¹H-NMR (400 MHz, DMSO-d₆): δ 10.32 (2H, brs, 2NH), 9.23 (2H, brs, 2NH), 8.80 (1H, brs, NH), 7.64 (2H, t, J = 6.8 Hz, ArH), 7.31 (2H, d, J = 7.2 Hz, ArH), 7.22–7.18 (2H, m, ArH), 7.10 (2H, t, J = 7.6 Hz, ArH), 6.96–6.91 (5H, m, ArH), 5.27 (1H, s, CH), 2.49 (obscured by DMSO signal, H-2/7), 2.31 (2H, d, J = 16.4 Hz, H-4/5), 2.23 (2H, d, J = 16.4 Hz, H-2/7), 2.16 (2H, d, J = 16.4 Hz, H-4/5), 1.03 (6H, s, (CH₃)₂), 0.82 (6H, s, (CH₃)₂). ¹³C-NMR (100 MHz, DMSO-d₆): δ 193.7 (C), 175.8 (C), 157.2/154.8 (C), 152.1 (C), 148.4 (C), 140.1 (C), 128.2 (CH), 128.0 (CH), 127.2 (CH), 126.9/126.8 (CH), 126.6/126.5 (CH), 125.2 (CH), 123.6 (CH), 115.3/115.1 (CH), 39.6 (CH₂ × 2), 29.2 (CH₂ × 2), 35.9 (CH), 30.3 (C), 29.2 (CH₃ × 2), 26.5 (CH₃ × 2). ESI-MS: m/z 684.1 (M + H).

Fig. S12 Regular expression missing case

Unconvertible IUPAC names case

Example 1:

2-[2-(2,6-Dichlorophenyl)amino]benzyl-3-(2-hydroxyphenylacrylamido)-6,8-dibromoquinazolin-4(3H)ones (5b)
Yield: 72%. m.p. 156-158 °C. IR (KBr) (cm⁻¹): 3547 (OH str), 3438 (NH str), 2926, 2854 (CH₂str), 1722 (C=O str), 1617 (C=N str), 1566 (CH=CH str), 757 (C-Cl str), 573 (C-Br str). ¹H-NMR (CDCl₃, 400 MHz), δ (ppm): 3.54 (s, 2H, CH₂), 6.37-8.13 (m, 13H, Ar-H), 6.78 (d, 1H, =CHCO, J = 16.2 Hz), 7.59 (d, 1H, =CH-Ar, J = 16.2 Hz), 8.82 (bs, 1H, CONH), 9.16 (bs, 1H, NH), 10.36 (bs, 1H, OH). ¹³C-NMR (CDCl₃, 100 MHz), δ (ppm): 30.63 (CH₂), 112.23-155.67 (26C, CH=CH and Ar-C), 161.82 (C=O), 168.40 (C=N), 173.22 (CONH). Anal. found: C, 50.35; H, 2.86; N, 7.79 %; Calcd. for C₃₀H₂₀Br₂Cl₂N₄O₃, C, 50.38; H, 2.82; N, 7.83 %.

Ground Truth: {"IUPAC":

"2-[2-(2,6-Dichlorophenyl)amino]benzyl-3-(2-hydroxyphenylacrylamido)-6,8-dibromoquinazolin-4(3H)ones ",}

Predicted Result: {"IUPAC":

"2-[2-(2,6-Dichlorophenyl)amino]benzyl-3-(2-hydroxyphenylacrylamido)-6,8-dibromoquinazolin-4(3H)ones ",}

Example 2:

Telisatin A (1). A solution of 1,1'-azobis(cyclohexanecarbonitrile) (245.0 mg, 1.0 mmol) and tributyltin hydride (1.2 g, 4.0 mmol) in toluene (20 mL) was added dropwise in four equal portions over 3 h to a refluxing solution of 10a (413.0 mg, 1.0 mmol) in toluene (20 mL) and the resulting mixture was then refluxed for another 8 h. The solvent was then removed under vacuum and the residue was dissolved in acetonitrile (40 mL) and washed with hexane (2 × 30 mL), then dried over anhydrous sodium sulfate. Removal of the solvent gave a brown viscous oil (0.4 g) which was recrystallized with ethanol to give telisatin A (1) as red prisms (109.9 mg, 33.0%); m.p. 234-235 °C (Lit. [1] m.p. 238-239 °C); UV (MeOH) λ max nm (log ε): 207 (4.03), 257 (4.26), 284sh (3.60), 322 (3.70), 336 (3.80), 352sh (3.56); IR (CH₂Cl₂-film) ν max cm⁻¹: 2925, 1748, 1701, 1605, 1584, 1531, 1462, 1423, 1386, 1306, 1261, 1195, 1149, 1131, 1112, 1037, 969, 924, 802, 759. ¹H-NMR δ: 9.41 (1H, br d, J = 8.5 Hz, H-11); 8.63 (1H, dd, J = 8.0, 1.5 Hz, H-8); 7.67-7.60 (1H, m, H-9); 7.54-7.46 (1H, m, H-10); 7.17 (1H, s, H-3); 4.10 (3H, s, OCH₃); 3.97 (2H, t, J = 6.5 Hz, CH₂); 3.95 (3H, s, OCH₃); 3.35 (2H, t, J = 6.5 Hz, CH₂); ¹³C-NMR: δ 179.98(C), 160.34(C), 157.15(C), 153.36(C), 146.65(C), 130.75(C), 129.35(C), 129.21(CH), 128.33(CH), 127.56(C), 125.87(C), 125.62(CH), 123.76(CH), 112.29(CH), 112.18(C), 103.17(C), 59.99(OCH₃), 56.62(OCH₃), 36.53(CH₂), 27.68 (CH₂). HRMS (ESI-TOF) calcd for C₂₀H₁₅NO₄ ([M+H]⁺) = 334.1074, Found 334.1125.

Ground Truth : {"IUPAC":Telisatin A,.....}

Predicted Result: {"IUPAC":Telisatin A,.....}

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)
Llama-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
galocatechin	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:

$$F = \frac{1}{n} \sum_{i=1}^n \log(P(w_i | w_1, w_2, \dots, w_{i-1}))$$

Where $P(w_i | w_1, w_2, \dots, w_{i-1})$ 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.