Electronic Supplementary Material (ESI) for Digital Discovery.

Unveiling the Synthesis Patterns of Nanomaterials: A Text Mining and Meta-Analysis Approach with ZIF-8 as a Case Study

Joseph R.H. Manning*[a] and Lev Sarkisov*[a]

SUPPLEMENTARY INFORMATION

Validation corpus of papers

Validation data was gathered from the NIST database of emerging adsorbents in August 2022 using the search terms listed in Table S1. From this database, 69 isotherms were manually downloaded from 37 journal articles. 43 synthesis paragraphs were identified manually from the journal articles (as detailed in Table S2), from which synthesis sequences and bills of materials were extracted by the author for validation against the text mining used. In this supplementary information file, we provide in-depth analysis of the method's performance, and all data and code used to create these validation analyses are provided in the GitHub repository accompanying this publication for the interested reader to perform their own analyses.

Table S1 – search terms for the NIST ISODB

Field	Term
Adsorbent material	ZIF-8
Adsorbate gas	Nitrogen
Measurement	Experimental
Temperature (K)	77

Table S2 – details of each synthesis paragraph extracted as validation data, alongside: the paragraph index as identified by ChemDataExtractor; the paragraph's first 4 words which can be used as unique search strings; and the paragraph's unique identifier used within the Synthetic Oracle software (and elsewhere within this supplementary information document).

Paper DOI	Paragraph index[a]	Paragraph starts with…	Sequence identifier
10.1080/00958972.2013.797966	16	"The reagents employed were…"	009589722013797966.16
10.1002/Aic.13970	6	"Synthesis of ZIF-8 was…"	Aic13970.6
	7	"To further remove the…"	Aic13970.7
10.1002/Aic.14525	0[a]	"In a typical synthesis…"	Aic14525.0
10.1016/j.powtec.2013.09.013	44	"ZIF-8 was obtained by…"	Jpowtec201309013.44
10.1002/anie.201104383	0	"ZIF-8 nanoparticles were synthesized…"	anie201104383.0
10.1039/C1ce05780d	0	"Controllable synthesis of ZIF-8…"	c1ce05780d.0
10.1039/C2cc34893d	0	"ZIF-8 crystals: ZIF-8 crystals…"	c2cc34893d.0
10.1039/C2jm15685g	24	"ZIF-8 nanoparticles were prepared…"	c2jm15685g.24
10.1039/C3ta11483j	23	"In a typical synthesis…"	c3ta11483j.23
10.1039/C4ee01009d	0	"The ZIF-8 sample was…"	c4ee01009d.0
10.1039/c5ra01183c	11	"Pure ZIF-8 was prepared…"	c5ra01183c.11
10.1002/chem.201301461	32	"Preparation of ZIF-8 nanocrystals…"	chem201301461.32

10.1021/cm3006953	27	“A solution of 20…”	cm3006953.27
10.1016/j.electacta.2014.11.093	30	“ZIF8 crystals were synthetized…”	jelectacta201411093.30
10.1016/j.ijhydene.2015.10.038	40	“ZIF-8 was synthesized according…”	jijhydene201510038.40
10.1016/j.memsci.2014.11.038	43	“ZIF-8 nanoparticles were prepared…”	jmemsci201411038.43
10.1016/j.memsci.2015.05.015	55	“The ZIF-8 nanocrystals were…”	jmemsci201505015.55
	56[b]	“The ZIF-8 nanocrystals were…”	jmemsci201505015.56
10.1016/j.micromeso.2012.03.052	34	“ZIF-8 nanoparticles were synthesized…”	jmicromeso201203052.34
10.1016/j.micromeso.2012.11.012	32	“As a standard synthesis…”	jmicromeso201211012.32
	33	“The substrate mixture was…”	jmicromeso201211012.33
10.1021/Jp407792a	25	“All the reagents and…”	jp407792a.25
	26	“Nanoscale ZIF-8 was prepared…”	jp407792a.26
10.1021/Jp5081466	1	“10 nm ZIF-8 sample…”	jp5081466.1
	2	“18 nm ZIF-8 sample…”	jp5081466.2
	3	“52 nm ZIF-8 sample…”	jp5081466.3
	4	“92 nm ZIF-8 sample…”	jp5081466.4
	5	“540 nm ZIF-8 sample…”	jp5081466.5
	6	“1 µm ZIF-8 sample…”	jp5081466.6
	7	“3.4 µm ZIF-8 sample…”	jp5081466.7
	8	“7.6 µm ZIF-8 sample…”	jp5081466.8
	9	“15.8 µm ZIF-8 sample…”	jp5081466.9
	10	“324 µm ZIF-8 sample…”	jp5081466.10
10.1016/j.jssc.2014.06.017	48	“The preparation of pure…”	jssc201406017.48
10.1016/j.ultsonch.2017.04.030	47	“Zeolitic imidazole framework-8 was…”	jultsonch201704030.47
10.1021/Jz300855a	1	“The 26 nm ZIF-8…”	jz300855a.1
	2	“The 7.9 µm ZIF-8…”	jz300855a.2
	3	“The 162 µm ZIF-8…”	jz300855a.3
10.1021/La401471g	31	“For a typical ZIF-8…”	la401471g.31
10.1007/s10450-012-9407-1	18	“First, a solid mixture…”	s10450-012-9407-1.18
10.1007/s12274-014-0501-4	0	“Synthesis of ZIF-8: A…”	s12274-014-0501-4.0
10.1016/s1872-2067(14)60292-8	30	“ZIF-8 samples were synthesized…”	s1872-2067(14)60292-8.30

a) a paragraph index of 0 indicates the synthesis methodology section was contained within a supplementary information file

b) Multiple synthesis descriptions were manually split into two separate paragraphs for the purposes of validation

Cross-validation against GPT-powered grammar parsing

Since the publication of ChemicalTagger in 2011,[1] significant advancements have been made in the field of named entity recognition and grammar parsing as has been recently reviewed in supplementary reference 2. More recently, large language models (LLMs) powered by transformer-based machine learning architectures e.g. BERT[3,4] and GPT[5,6] represent an exciting new development in the field of natural language programming. Accordingly, to test the quality of ChemicalTagger-based sequence extraction versus LLM-based sequence extraction, we developed a 5-step series of prompts using ChatGPT model gpt-3.5-turbo (release 0613) to perform equivalent pre-processing to ChemicalTagger. These stages are described in Table S3.

Table S3 - ChatGPT prompts used for cross-validation against ChemicalTagger

Stage	System message	Input message
1	"Extract the synthesis actions from the following text into a JSON object with the keys 'Action type' and 'Text'. \n"	Full synthesis paragraph
2	Classify the chemical synthesis action provided as either 'Add', 'Apparatus action', 'Concentrate', 'Cool', 'Degas', 'Dissolve', 'Dry', 'Extract', 'Filter', 'Heat', 'Partition', 'Precipitate', 'Purify', 'Quench', 'Recover', 'Remove', 'Stir', 'Synthesize', 'Wait', 'Wash', or 'Yield'. Return the class name only, or 'invalid' if no class could be found. \n	Identified synthesis actions from stage 1
3	Identify the chemicals and solvents in the following statement and their quantities delimited by quotes. Return as a python list of json objects with the keys 'name' and 'quantity'. If any information is not provided or you are unsure, use 'N/A'. \n	Identified synthesis text from stage 1
4	Does the following statement mention a specific amount of time? If so, reply with the time mentioned as a python string, or a list of python strings if multiple times are mentioned. If no time values are mentioned or you are unsure, return an empty list. \n	Identified synthesis text from stage 1
5	Does the following statement mention a specific amount of temperature? If so, reply with the temperature mentioned as a python string, or a list of python strings if multiple temperatures are mentioned. If no temperature values are mentioned or you are unsure, return an empty list. \n"	Identified synthesis text from stage 1

Further details on validation of synthesis action parsing

The first action to be taken to break down a synthesis paragraph to its constituent actions is segmentation into individual phrases representing specific actions taken.[2] Accordingly, the length of each sequence extracted is dependent on the method used. Once segmented, the phrases must be categorised according to which action is being taken. To maintain consistency between each validation method using in this study, we defaulted to the list of actions defined by Hawizy et al.[1] This led to three ways that sequences could be compared – their length, the set of actions identified therein, and the specific sequence of actions identified. It should be noted that during manual labelling, repeated actions e.g. when a sample was "washed three times" were only counted once to ensure that the text mining algorithms were not being penalised for their inability to account for such implicit actions.

Sequence length

From manual processing each synthesis contained 5.5 ± 1.0 actions, contrasting with 7.7 ± 3.2 actions identified with either ChemicalTagger or ChatGPT. As can be seen in Figure S1A, manually segmented paragraphs were shorter on average than either automatic parsing method. As described in the main text, sequence actions were mapped to the labels "addition," "reaction," "extraction" and "other," reducing the overall length of each sequence. Once condensed, the sequence length was reduced to 2.8 ± 0.6 actions for manually-parsed sequences, 3.2 ± 1.3 actions for sequences parsed with ChemicalTagger, and 3.1 ± 1.2 actions for sequences parsed with ChatGPT (Figure S1B).

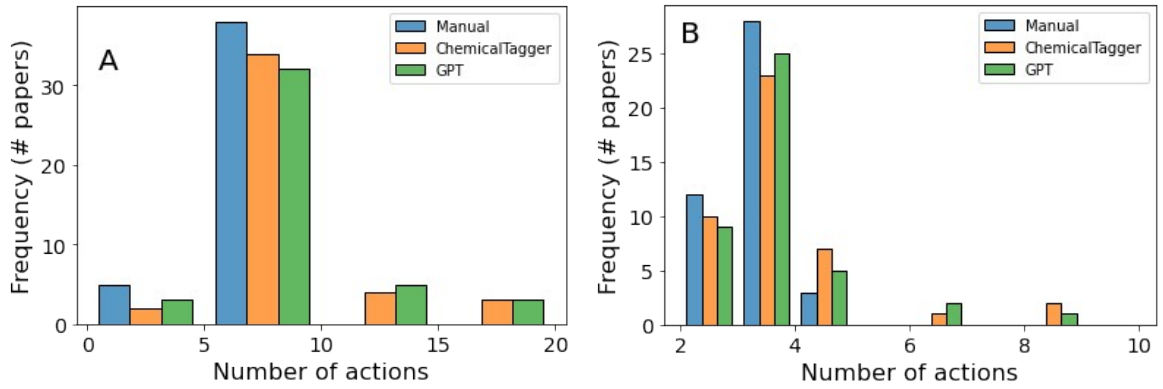

Figure S1 – Histograms of synthesis actions identified from each parsing method. (A) raw sequence actions and (B) condensed sequence actions

Actions identified

To more precisely compare the number of actions between each parsing methods, the specific frequency of action terms was compared against manual parsing to provide information about each text mining system's precision, recall and F1-score (as defined in the main text). Furthermore, actions identified with ChemicalTagger and ChatGPT were cross-validated against one another, providing insight into the consistency of the text mining software and juxtaposition against the manual identification.

From this analysis, both ChatGPT and ChemicalTagger performed almost equivalently when compared against manually-identified synthesis actions (Table S4, Figure S2). Both precision and recall showed F1-scores of approximated 65%, which jumped to ca. 85% when the two text-mining algorithms were compared against one another. This step change when changing from manual validation indicates that while the text mining methods were able to consistently categorise synthesis actions, manual categorisation is highly user-subjective and therefore has a low level or reliability; for example, introduction of reagents at the start of the reaction could reasonably be assigned the "add" or "dissolve" action categories due to their semantic similarities.

A final analysis performed was the Levenshtein sequence similarity test,[7] which identifies the number of modifications required to convert sequence A to sequence B. Here, the discrepancy between manual validation and cross-validation is starker – with sequence similarity increasing from 55% to 98% - supporting the idea that manually-labelled synthesis actions were too subjective to reliably match the text-mined data.

Table S4 – Average text mining parsing metrics for action identification. Error bars are one standard deviation around the mean (n=43)

Validation	Precision	Recall	F1 (%)	Levenshtein

method	(%)	(%)		similarity (%)
ChemicalTagger	59 ± 19	77 ± 20	66 ± 19	55 ± 19
ChatGPT	58 ± 20	76 ± 21	65 ± 20	54 ± 19
Cross-validation	100 ± 3	99 ± 8	99 ± 5	98 ± 8

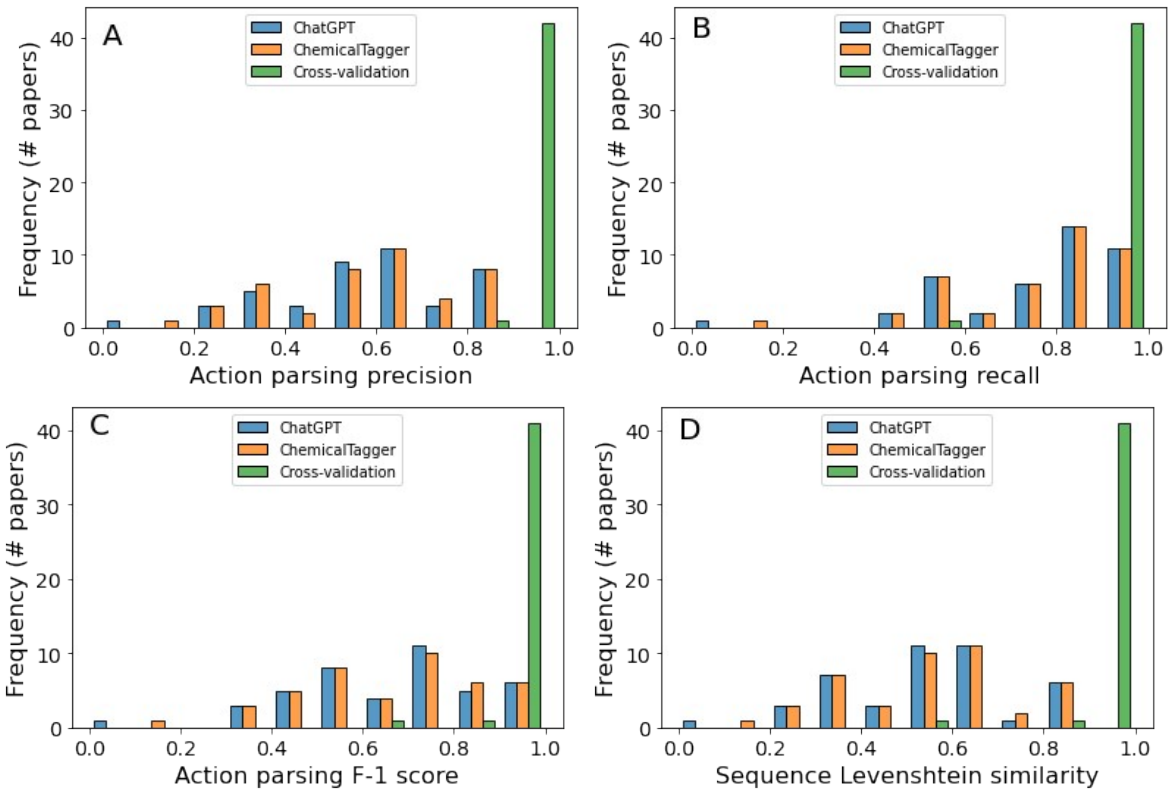

Figure S2 – Histograms of text mining performance metrics for raw synthesis sequences. (A) Precision, (B) recall, (C) F1-score, and (D) Levenshtein similarity

To reduce this subjectivity, we repeated this analysis on the sequence of condensed actions, where the number of categories was reduced to "addition," "reaction," "extraction," and "other," and neighbouring actions of the same type were condensed together (Table S5, Figure S3). Here, each of the text mining techniques significantly improved compared against manual parsing while remaining approximately equivalent when comparing ChemicalTagger against ChatGPT. From this, we conclude that the use of reduced synthesis action categories is significantly more reliable than the 15 categories proposed by Hawizy et al.[1] and therefore the use of a reduced synthesis action vocabulary like that proposed in supplementary reference [8] will significantly improve synthesis text mining in future studies.

Table S5 – Average text mining parsing metrics for condensed synthesis action identification. Error bars are one standard deviation around the mean (n=43)

Validation method	Precision (%)	Recall (%)	F1 (%)	Levenshtein similarity (%)
ChemicalTagger	83 ± 22	89 ± 16	84 ± 16	75 ± 24
ChatGPT	84 ± 21	89 ± 17	84 ± 17	76 ± 22

Cross-validation	100 ± 0	97 ± 10	98 ± 6	97 ± 10

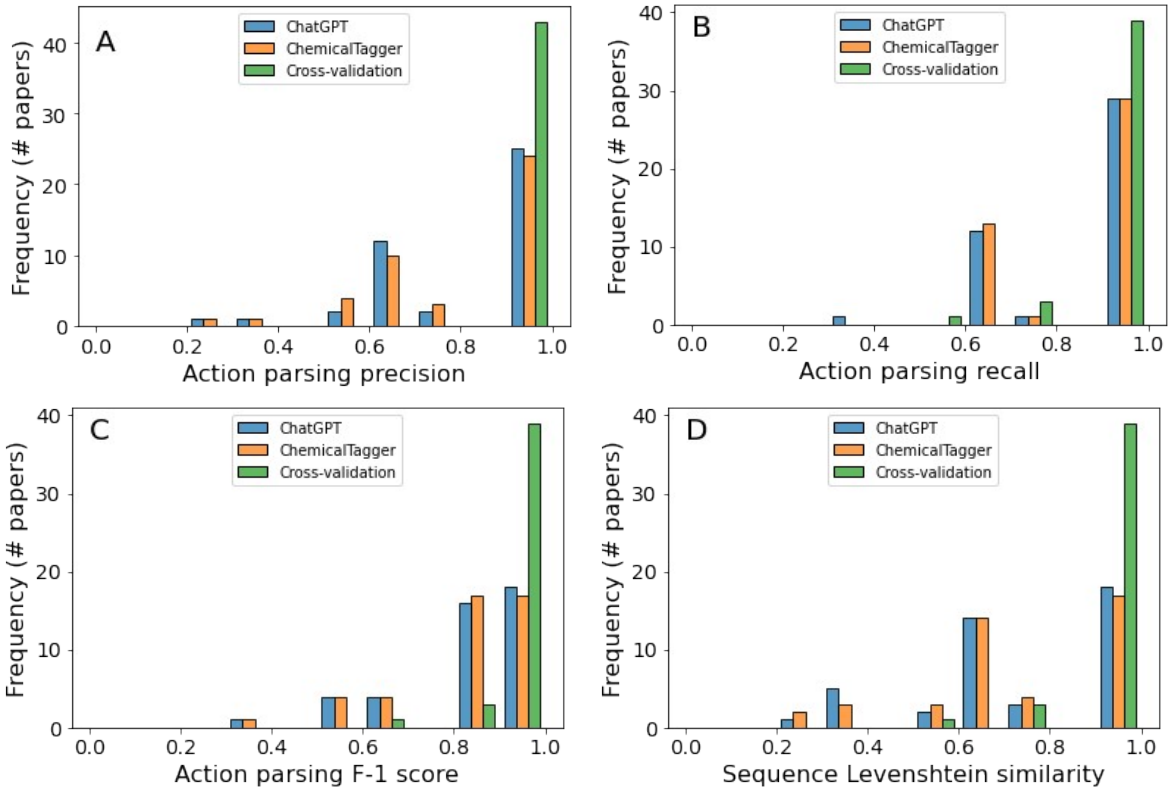

Figure S3 – Histograms of text mining performance metrics for condensed synthesis sequences. (A) Precision, (B) recall, (C) F1-score, and (D) Levenshtein similarity

Further details on chemical identification and quantity matching

Once sequences had been segmented into specific phrases, the next step to be taken was the identification of chemicals and their associated quantities in each phrase. In the ChemicalTagger software, chemicals are identified using the OSCAR4 named entity recognition algorithm[9] which combines regular expression matching, ontological parsing, and feature-matching to identify chemicals. Conversely ChatGPT used no such rule-based entity recognition, simply identifying chemical-like tokens through its general language parsing structure. This meant that a large number of tokens were falsely flagged by ChatGPT as chemical entities e.g. “100 mL autoclave” or “filtration”. Once tokens were identified, their structures were cross-referenced against the PubChem database of chemical entities and a manually-compiled hash table. This had the benefit of discarding the obviously incorrect chemical entities recognised by ChatGPT. As with synthesis actions, parsing quality was judged by comparing against manual chemical identification after cross-referencing (Figure S4, Table S6). This led to significantly better parsing accuracy for both text mining techniques when compared against action categorisation and approximately identical performance between the algorithms, indicating that large language models are equally adept at identifying chemicals chem compared against dedicated algorithms.

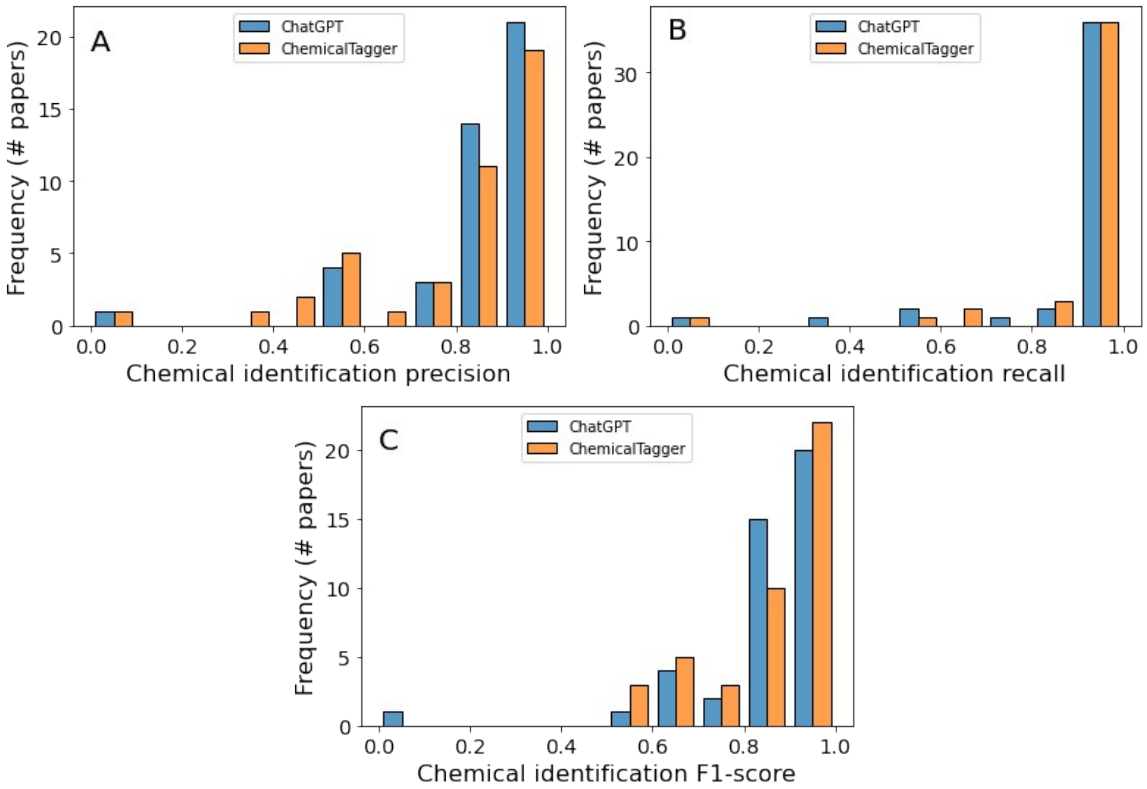

Figure S4 - Histograms of text mining performance metrics for chemical identification. (A) Precision, (B) recall, (C) F1-score

Table S6 – Average text mining parsing metrics for chemical entity identification. Error bars are one standard deviation around the mean (n=43)

Validation method	**Precision (%)**	**Recall (%)**	**F1 (%)**
ChemicalTagger	81 ± 23	94 ± 18	85 ± 20
ChatGPT	86 ± 20	92 ± 21	87 ± 18

Alongside identificaiton of the chemicals, associated quantities such as the chemical's mass or volume must be connected to the chemical structure. In ChemicalTagger this is performed by sentence grammar parsing algorithms, where chemical names and quantities are stored within individual "Molecule" XML tags. Conversely, these actions are performed implicitly with ChatGPT. However, when quantity identification was compared between methods, the text mining performance was approximately equivalent in all cases (Figure S5, Table S7). Therefore, despite the preponderance of false positive chemical names identified with ChatGPT, the parsing quality is approximately equivalent.

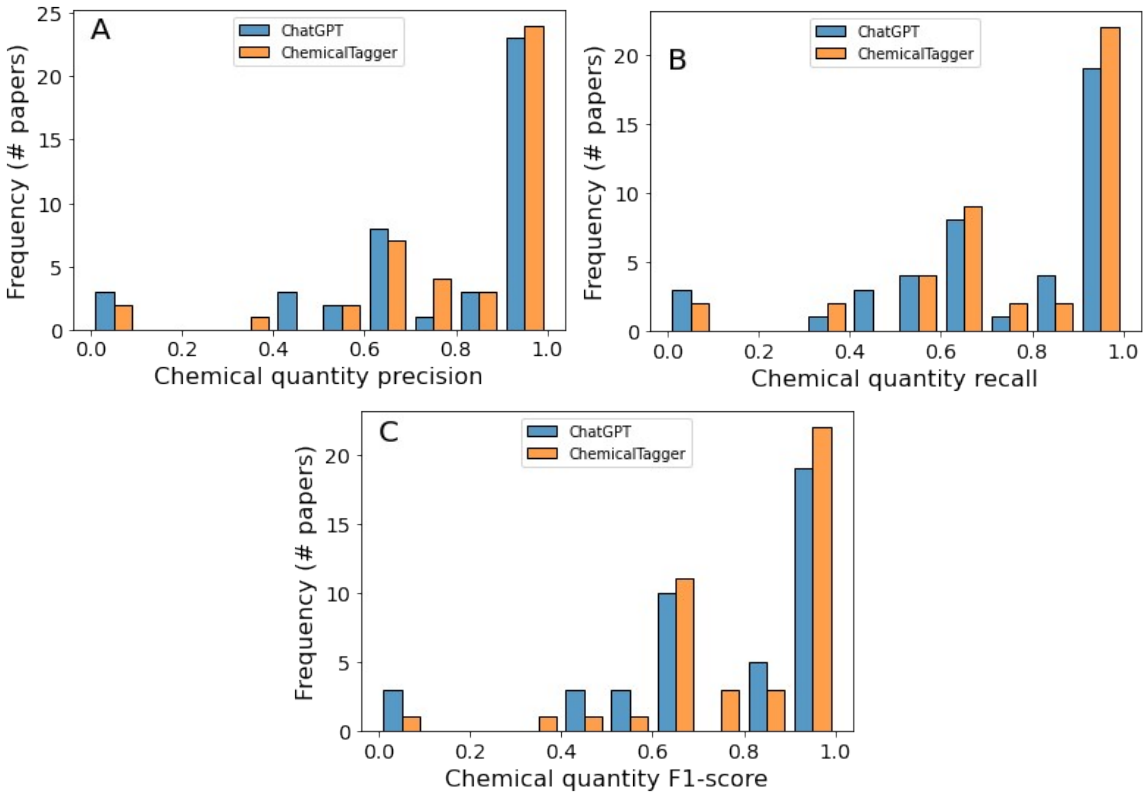

Figure S5 - Histograms of text mining performance metrics chemical quantity matching for those chemicals which were correctly identified. (A) Precision, (B) recall, (C) F1--score

Table S7 – Average text mining parsing metrics for chemical quantity identification. Error bars are one standard deviation around the mean (n=43)

Validation method	**Precision (%)**	**Recall (%)**	**F1 (%)**
ChemicalTagger	81 ± 23	94 ± 18	85 ± 20
ChatGPT	86 ± 20	92 ± 21	87 ± 18

Further details on time and temperature parsing

The final synthesis aspects considered in this study were times and temperatures associated with specific synthesis steps. These quantities were identified in much the same way as chemical names and amounts – named entity recognition in the case of ChemicalTagger and general language parsing in the case of ChatGPT. In both cases, parsing quality was very high (Table S8, Table S9), with the majority of synthesis protocols being parsed in with perfect fidelity (Figure S6, Figure S7).

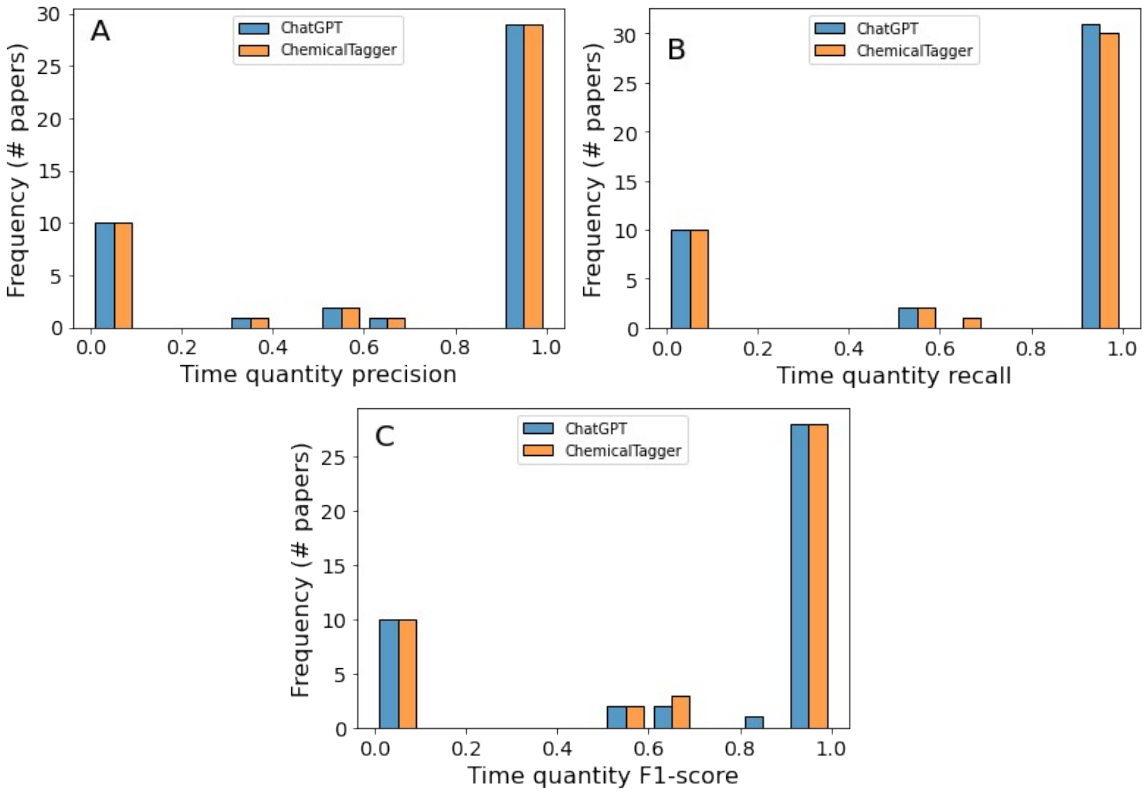

Figure S6 - Histograms of text mining performance metrics for time quantity identification. (A) Precision, (B) recall, (C) F1-score

Table S8 – Average text mining parsing metrics for time quantity identification. Value agreement is the number of syntheses where the total time agreed between manual and text-mined values. Error bars are one standard deviation around the mean (n=43)

Validation method	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Value agreement (%)**
ChemicalTagger	72 ± 43	74 ± 43	72 ± 42	69
ChatGPT	72 ± 43	74 ± 43	72 ± 42	69

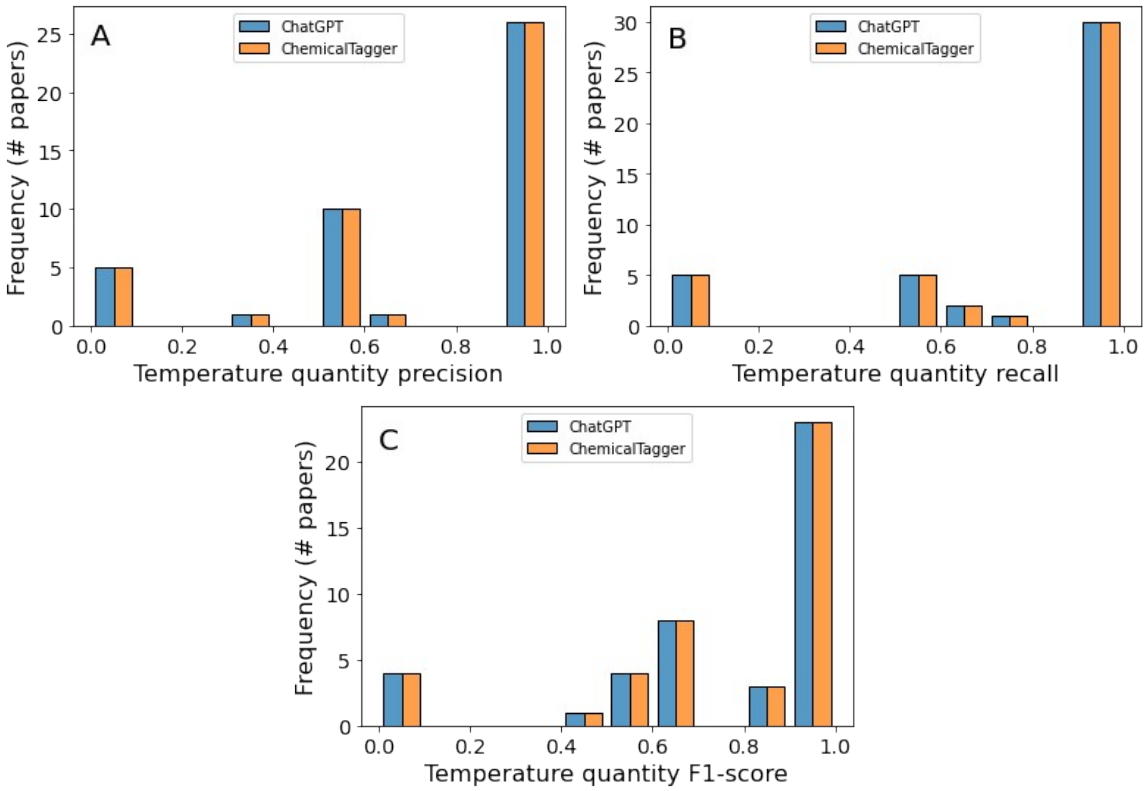

Figure S7 - Histograms of text mining performance metrics for temperature quantity identification. (A) Precision, (B) recall, (C) F1-score

Table S9 – Average text mining parsing metrics for temperature quantity identification. Value agreement is the number of syntheses with temperature values identified between text mining and manual parsing. Error bars are one standard deviation around the mean (n=43)

Validation method	**Precision (%)**	**Recall (%)**	**F1 (%)**	**Value agreement (%)**
ChemicalTagger	76 ± 34	83 ± 32	77 ± 31	74
ChatGPT	76 ± 34	83 ± 32	77 ± 31	74

Supplementary references

1 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, *J. Cheminform.*, 2011, **3**, 1–13.

2 O. Kononova, T. He, H. Huo, A. Trewartha, E. A. Olivetti and G. Ceder, *iScience*, 2021, **24**, 102155.

3 S. Huang and J. M. Cole, *Chem. Sci.*, 2022, **13**, 11487–11495.

4 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *Arxiv*, , DOI:10.48550/arXiv.1810.04805.

5 Z. Zheng, O. Zhang, C. Borgs, J. T. Chayes and O. M. Yaghi, *J. Am. Chem. Soc.*, 2023, **145**, 18048–18062.

6 K. M. Jablonka, P. Schwaller and A. Ortega-guerrero, *ChemRxiv*, 2023, 1–32.

7 V. I. Levenshtein, *Sov. Phys. Dokl.*, 1966, **10**, 707–710.

8 Z. Wang, K. Cruse, Y. Fei, A. Chia, Y. Zeng, H. Huo, T. He, B. Deng, O. Kononova and G. Ceder, *Digit. Discov.*, 2022, **1**, 313–324.

9 D. M. Jessop, S. E. Adams, E. L. Willighagen, L. Hawizy and P. Murray-Rust, *J. Cheminform.*, 2011, **3**, 1–12.