

Lawrence Berkeley National Laboratory

LBL Publications

Title

Text-mined dataset of sol-gel synthesis protocols with phase purity outcomes using large language models

Permalink

<https://escholarship.org/uc/item/2fk398qd>

Journal

Scientific Data

ISSN

2052-4463

Authors

Chen, Wesley Y

Lee, Sanghoon

Kodalle, Tim

et al.

Publication Date

2026-10-15

DOI

10.1038/s41597-026-08492-6

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed



OPEN

DATA DESCRIPTOR

Text-mined dataset of sol-gel synthesis protocols with phase purity outcomes using large language models

Wesley Y. Chen^{1,2}, Sanghoon Lee^{1,3}, Tim Kodalle⁴, Philothei Sahinidis^{1,5}, Gerbrand Ceder^{1,3}, Carolin M. Sutter-Fella⁴  & Anubhav Jain¹ 

Sol-gel synthesis is a versatile method for producing inorganic materials that offers advantages such as atomic-level mixing and lower processing temperatures over traditional solid-state methods. Despite these benefits, achieving phase purity remains a challenge due to the process's sensitivity to synthesis parameters such as precursor choice, solvent selection, and heating conditions. Phase-impure outcomes are essential for understanding sol-gel reaction pathways, but are largely absent from existing curated datasets. This work leverages Gemini 3.0 Flash to construct a text-mined inorganic sol-gel database with phase purity information. The dataset contains ~148,000 synthesis records, including ~28,000 phase-impure entries. All extracted entities achieved F1 scores greater than 0.89. At the record level, 68% of extracted records match human annotation on all evaluated fields. Furthermore, to validate the scientific utility of our dataset, we demonstrate that the aggregated data reproduces previously reported phase-purity trends for BiFeO₃ thin films.

Background & Summary

Sol-gel synthesis is an established wet chemical technique¹ for the fabrication of inorganic materials, especially nanomaterials^{2,3} and thin films^{4,5}. Unlike conventional solid state methods, which usually require high temperatures and extensive mechanical processing (e.g., grinding, milling, etc.) to overcome diffusion limitations⁶, the sol-gel method can potentially enable lower processing temperatures, greater control over morphology, and higher phase purity through its atomic level inter-mixing of precursor material^{2–10}. These capabilities become critical in applications like flexible electronics by enabling crystalline metal oxide thin films to be formed at temperatures compatible with heat-sensitive substrates¹¹.

However, navigating the large synthesis parameter space (e.g., precursors, solvents, time, temperature, and pH), often makes sol-gel synthesis optimization a time and labor-intensive process. Careful control of these parameters is necessary as even slight variations can lead to impurity phase formation^{12–16}. In the absence of a general theory for materials synthesis, documenting and analyzing these phase-impure outcomes is crucial to understanding reaction pathways^{17–19}. Unfortunately, these details are often scattered across the scientific literature. To help capture these insights at scale, data processing pipelines that can systematically parse the text into structured synthesis records are necessary²⁰.

The early transition towards this data-driven framework was facilitated by the development of natural language processing (NLP) tools^{20–30} like ChemDataExtractor^{23–25}, BERT^{26–28}, and Word2Vec^{29,30}, along with annotated corpora such as the Materials Science Procedural Text Corpus³¹. These tools enabled the construction of various text-mined datasets^{30,32–36}, including for solid-state synthesis³⁴, solution-based synthesis³⁵, and chemical vapor deposition³⁶. Notably, Kononova *et al.*³⁴ also released a sol-gel dataset of 9,518 protocols in 2020 through the same public repository, though this dataset lacked phase-purity information. Most recently, Lederbauer *et al.*

¹Lawrence Berkeley National Lab, Energy Technologies Area, Berkeley, USA. ²Northwestern University, Department of Materials Science and Engineering, Evanston, USA. ³University of California, Berkeley, Department of Materials Science and Engineering, Berkeley, USA. ⁴Lawrence Berkeley National Lab, Molecular Foundry, Berkeley, USA. ⁵University of California, Berkeley, Department of Electrical Engineering and Computer Sciences, Berkeley, USA.  e-mail: ajain@lbl.gov

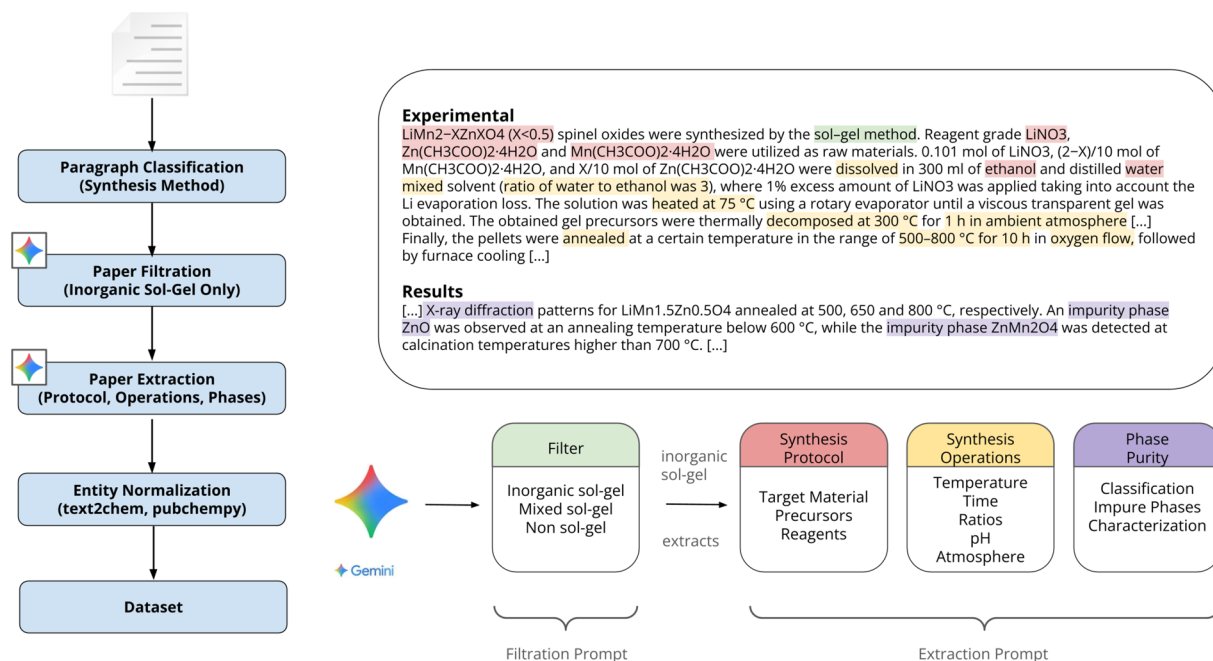


Fig. 1 Schematic representation of the data extraction pipeline. Gemini 3.0 Flash is used to filter for inorganic sol-gel papers and extract protocols, operations, and phase purity outcomes from text. Example text is adapted from Demidzu *et al.*⁶⁷.

released LeMat-Synth, a toolbox that uses LLMs and vision-language models to extract synthesis procedures from ~81,000 open-access papers, yielding ~58,000 procedures across 35 synthesis methods and 16 material classes³⁷.

Many of these datasets have been used to train machine learning (ML) models^{38–42}, including for precursor recommendations^{38,39} and synthesis conditions predictions^{40,41}. However, limitations of this method include anthropogenic biases^{43,44} and a lack of negative data, such as phase-impure syntheses, in training sets^{45,46}. The latter is crucial as studies have shown that the inclusion of negative data greatly enhances ML models for predictive synthesis^{47–51}. Recent efforts have sought to address some of these limitations through explicitly extracting phase-impure outcomes in new text-mining efforts^{52–54}. In particular, Lee *et al.* used GPT-4o to construct a new text-mined solid-state synthesis database with ~80,000 protocols, including ~19,000 with impurity phases⁵⁴.

Sol-gel synthesis lacks an analogous text-mined dataset, despite being highly sensitive to phase formation pathways. Efforts to address this gap include a manually annotated, text-mined database of BiFeO_3 sol-gel thin films that explicitly tracked impurity phases⁵⁵. The authors demonstrated the utility of this data in informing the development of ML and thermodynamic models to help rationalize experimental phase formation outcomes^{55–57}. To expand on these efforts, this work leverages Gemini 3.0 Flash to automate the construction of a text-mined dataset containing ~148,000 inorganic sol-gel synthesis records, including ~28,000 phase-impure records (Fig. 1). Furthermore, we demonstrate that our large language model driven framework can be complementary to specialized NLP tools like MatBERT²⁶ while also offering an alternative to tools like Synthesis Action Retriever²² used in prior works^{34,35,54}.

By documenting both phase-pure and impure syntheses, this work provides a foundation for applying data-driven methods to investigate the parameters that govern phase purity in sol-gel synthesis. This includes serving as a knowledge base for emerging scientific tools, such as integration with agentic systems⁵⁸ to facilitate evidence-based hypothesis generation and the rational planning of synthesis experiments. Ultimately, the continued development of text-mined datasets across diverse synthesis methods enables data-driven comparative analysis of why certain materials are favored or only accessible through certain synthesis routes.

Methods

Content Acquisition. Using similar methods as Lee *et al.*⁵⁴, over 8 million scientific papers were sourced from the following publishers and publication years: Elsevier (2000–2023), American Institute of Physics (2000–2023), Institute of Physics (2000–2024), Springer Nature (2000–2024), Royal Society of Chemistry (2000–2023), American Chemical Society (2000–2020), American Physical Society (2000–2020), and Wiley (2000–2020).

Along with the text, metadata associated with each publication, such as their DOI and section headings associated with each paragraph (e.g., abstract, introduction, results, etc.), were stored in an internal MongoDB database maintained by the Ceder group at Lawrence Berkeley National Lab. More details are provided in prior work^{34,35,54}.

Paragraph Classification and Paper Construction. Using a MatBERT²⁶ classifier, paragraphs from each paper were scored and classified by synthesis method into solid state, sol-gel, precipitation, hydrothermal, and “something else”. The DOIs of papers with at least one paragraph classified as sol-gel with an associated

classification score greater than 0.85 were retained. For each DOI, the paper text was reconstructed by merging all paragraphs from relevant sections (e.g., abstract, results, discussion, experiment, materials, methods, etc.). We exclude paragraphs from the introduction and conclusion sections because they usually contain repetitive content or commentary on other works. Through this classification process, ~102,000 candidate sol-gel papers were identified.

Paper Filtration. threshold tuning effort to evaluate the efficacy of the MatBERT classifier, 200 papers were randomly sampled from the ~102,000 candidate sol-gel papers in batches of 50 across increasing sol-gel classification score ranges (i.e. greater than 0.85, 0.85 to 0.90, 0.90 to 0.95, and 0.95 to 1.00). Manual inspection of these papers found that many were misclassified as sol-gel, but also that stricter score cutoffs excluded genuine sol-gel papers. To address this issue, we introduce an additional paper filtration step using a few-shot guided Gemini 3.0 Flash (LLM) to filter out these misclassified papers rather than increasing the threshold. Additionally, we used the LLM to further classify the papers it identifies as sol-gel into inorganic or mixed inorganic-organic synthesis based on the composition of the final target material. This allows us to filter out mixed inorganic-organic sol-gel synthesis papers, which are outside the scope of this dataset. For the purposes of this dataset, a synthesis is treated as sol-gel when a precursor solution undergoes hydrolysis and condensation (or chelation and polymerization, as in the Pechini and citrate-gel routes) to form a gel network before drying and heat treatment. Procedures in which precursors are simply dissolved and dried by evaporation, without a sol-to-gel transition, are classified as non-sol-gel. In total, ~33,000 non-sol-gel papers and ~16,000 mixed inorganic-organic sol-gel papers were filtered out, leaving ~53,000 inorganic sol-gel papers for extraction. The raw outputs of this LLM-based paper filtration are shared publicly (see Code Availability). The filtration prompt used by the LLM is provided in Supplementary Sec. 1.

Synthesis Data Extraction. Our prior methodology focused on extracting target materials, precursors, and impurity phases using GPT-4o⁵⁴, while synthesis operations were extracted using Synthesis Action Retriever²². Here, we unify the entire extraction using a few-shot guided Gemini 3.0 Flash (LLM). This includes target materials, precursors, reagents, synthesis conditions, and phase purity information described in the text of each inorganic sol-gel synthesis paper. Additionally, to achieve phase-pure synthesis, studies will often test multiple synthesis routes through variations in precursors, reagents, or synthesis conditions. To capture these nuances, the LLM is instructed to extract each experimental variation as a separate data record. Then, the LLM parses the experimental procedures described in the text into sequential synthesis operations, each associated with its own set of synthesis conditions. These sequences are then mapped to their respective phase purity outcomes, which the LLM classifies into “pure”, “impure”, or “insufficient info”. If impure, the specific impurity phases described in the text are also recorded. For example, in the text illustrated in Fig. 1, the LLM extracted a separate data record and phase purity outcome for each annealing temperature (i.e., 500 °C, 650 °C, and 800 °C) tested in the synthesis of $\text{LiMn}_{1.5}\text{Zn}_{0.5}\text{O}_4$ (see Technical Validation for more details). To provide additional insights, the LLM also extracts any methods used to characterize the target material, and provides some context associated with the phase purity classification in the phase_purity_details field. This comprehensive framework expands on our prior work⁵⁴, which only tracked the presence of impurity phases and paper-level synthesis operations. The extraction prompt was developed in consultation with practitioners of sol-gel synthesis and can be found in Supplementary Sec. 2.

Following the extraction, a multi-stage post-processing pipeline was implemented. First, Pydantic (<https://github.com/pydantic/pydantic>) was used to enforce a strict JSON schema, where LLM extracted data records that do not follow the schema outlined in the extraction prompt (e.g., missing fields or structural inconsistencies) were removed. Then, text2chem (<https://github.com/CederGroupHub/text2chem>) was used to normalize material entities to add chemical metadata such as material formulas and composition. PubChemPy (<https://github.com/mcs07/PubChemPy>) was also employed to map reagent entities to their PubChem compound ID, IUPAC names, SMILES, and other metadata. Custom entity mappings for the most commonly extracted strings were developed to normalize fields like atmosphere and characterization methods. The post-processing pipeline and custom entity mappings are shared publicly (see Code Availability).

Data Records

The full dataset containing ~148,000 inorganic sol-gel synthesis data records extracted from ~53,000 papers is provided as a JSONL file (<https://doi.org/10.6084/m9.figshare.33040070>)⁵⁹. At the top level, each data record contains a synthesis protocol, synthesis operations, and phase purity information.

- protocol: an object containing the DOI, target material, precursors, and reagents
- operations: a list of objects containing sequences of synthesis operations and conditions
- phase_purity: an object describing phase purity classifications, impurity phases, and characterization methods.

More details regarding the data structure are provided in Table 1. An example entry is provided in Supplementary Sec. 3.

Data Overview

The top five most synthesized ternary compounds and their corresponding fraction of phase-pure data records (i.e., pure / pure + impure) are provided in Table 2. The variety of metal sources, chelating, and combustion agents reproduces the breadth of sol-gel protocols documented across the literature. This includes the Pechini method (citric acid, ethylene glycol), glycine-nitrate combustion (metal nitrates, glycine), citrate-gel (citric acid, ammonia), metal alkoxide, and acetylacetonone-based sol-gel routes. Many of the impurity phases are also found

	Data Description	Data Key Label	Data Type
Synthesis Protocol	DOI of the original paper	doi	string
	Target material of synthesis	target	dict: material_string: string • material_formula: string • composition: list of objects • phase: string • material_form: string • is_mixture: bool • additives: list of string • oxygen_deficiency: string • oxidation_states: dict
	Metal precursors used in synthesis	metal_precursors	list of objects • see target
	Reagents used in synthesis (solvents, catalysts, additives, etc.)	reagents	list of dicts: • reagent_string: string • reagent_role: string • pubchem_CID: int • iupac_name: string • reagent_formula: string • smiles: string
Synthesis Operations	Type of operation (mixing, heating, deposition, grinding, other)	operation_type	string
	Time of operation	time	dict • raw_string: string • max_value: float • min_value: float • unit: string
	Temperature of operation	temperature	dict • see time
	Any ratios between reagents or precursors of the operation	ratios	list of dict • type: string • description: string • value: string
	pH of operation	pH	dict • raw_string: string • max_value: float • min_value: float
	Atmospheric condition of operation (in air, N ₂ , etc.)	atmosphere	dict: • raw_string: string • atmosphere_normalized: string
	Short description of operation step including any additional information not mentioned anywhere else	operation_details	string
Phase Purity Information	Impurity phase(s)	impurity_phase	list of objects: • see target
	Classification of phase purity based on text (i.e. impure, pure, insufficient info)	classification	string
	Any additional details regarding phase purity outcome mentioned in text	phase_purity_details	string
	Any characterization methods mentioned in text	characterization_methods	dict: • characterization_string: list of string • characterization_normalized: list of string

Table 1. Data records format. Each data record includes the synthesis protocol, synthesis operations, and phase purity information under the following top-level keys: protocol, operations, and phase_purity.

in solid state synthesis⁵⁴ (e.g., Bi₂Fe₄O₉, Bi₂₅FeO₄₀ for BiFeO₃), but the table also highlights some impurities specific to the solution-based and carbon-rich nature of sol-gel synthesis. For instance, the La₂O₂CO₃ impurity phase in LaFeO₃ may arise from local cation segregation caused by insufficient precursor homogeneity, which leaves lanthanum-rich intermediates susceptible to carbonation by CO₂ evolved during decomposition of the organic precursors^{60,61}.

Technical Validation

Validation of Paragraph Classification and Paper Filtration. Our synthesis classification and filtration pipeline first used a MatBERT paragraph classifier to screen for relevant papers, followed by LLM-based paper level verification to improve precision and separate the inorganic sol-gel papers from the mixed inorganic-organic. To evaluate this pipeline, 200 papers were randomly sampled from the pool of ~102,000 candidate sol-gel papers identified in the initial MatBERT classification step. Then, a human evaluator determined ground truth labels for each paper to construct our validation set (see Supplementary Sec. 4). This validation set is distinct from the 200 papers inspected during threshold tuning (see Methods). The prompts, few-shot examples, and filtration rules were developed on a preliminary dataset built with Llama 4 Scout and a separately sampled 100 paper development set. There is no overlap between the threshold set and the validation set, however, one paper from that development set also appears in the validation set, and excluding it changes no value in Tables 3 to 5 by more than 0.002.

Target Material(Fraction of Phase Pure)	Precursors	Reagents	Common Impurity Phases
BiFeO ₃ (789/1194)	Bi(NO ₃) ₃ , Fe(NO ₃) ₃ , Bi(CH ₃ COO) ₃ , Fe(CH ₃ COCHCOCH ₃) ₃	ethylene glycol, 2-methoxyethanol, nitric acid, citric acid	Bi ₂ Fe ₄ O ₉ , Bi ₂₅ FeO ₄₀ , Bi ₂ O ₃ , Fe ₂ O ₃
CoFe ₂ O ₄ (911/1054)	Fe(NO ₃) ₃ , Co(NO ₃) ₂ , Co(CH ₃ CO ₂) ₂ , FeCl ₃	citric acid, ammonia, ethanol, ethylene glycol	Fe ₂ O ₃ , CoO, Co ₃ O ₄ , Fe ₃ O ₄
BaTiO ₃ (631/803)	Ba(CH ₃ COO) ₂ , Ti(OCH(CH ₃) ₂) ₄ , Ba(NO ₃) ₂ , Ti(OC ₄ H ₉) ₄	acetic acid, ethanol, citric acid, acetylacetone	BaCO ₃ , TiO ₂ , Ba ₂ TiO ₄
NiFe ₂ O ₄ (587/709)	Fe(NO ₃) ₃ , Ni(NO ₃) ₂ , Ni(CH ₃ COO) ₂ , NiCl ₂	citric acid, ammonia, ethylene glycol, ethanol	Fe ₂ O ₃ , NiO, Ni, Fe ₃ O ₄
LaFeO ₃ (599/651)	Fe(NO ₃) ₃ , La(NO ₃) ₃ , FeC ₃ H ₅ O(COO) ₃	citric acid, ethylene glycol, glycine, ethanol, ammonia	La ₂ O ₃ , Fe ₂ O ₃ , La(OH) ₃ , La ₂ O ₂ CO ₃

Table 2. Most common ternary compounds with some of their top precursors, reagents, and impurity phases. Data records assigned with the “insufficient info” phase purity classification label are excluded from the analysis.

Method	Classification	Precision	Recall	F1	Support
MatBERT Only	Sol Gel	0.650	1.00	0.788	130
MatBERT and Gemini 3.0 Flash	Sol Gel	0.977	0.985	0.981	130

Table 3. Evaluation metrics for classification of papers by synthesis method.

Method	Classification	Precision	Recall	F1	Support
Gemini 3.0 Flash	Inorganic Sol Gel	0.981	0.911	0.944	112
	Mixed Sol Gel	0.667	1.00	0.800	18

Table 4. Evaluation metrics for classification of sol-gel papers into inorganic or mixed organic-inorganic based on the composition of the final target material.

Entity		Precision	Recall	F1	Support
Target	Formula	0.965	0.954	0.959	259
	Form	0.948	0.948	0.948	286
Precursors	Formula	0.950	0.936	0.943	624
Reagents	PubChem CID	0.916	0.956	0.935	727
Operations	Temperature	0.969	0.968	0.968	774
	Time	0.969	0.969	0.969	752
	Ratios	0.955	0.969	0.962	262
	pH	0.980	0.980	0.980	50
	Atmosphere	0.950	0.947	0.948	243
Phase Purity	Classification	0.899	0.899	0.899	286
	Impurity Phase	0.977	0.896	0.935	48

Table 5. Evaluation metrics for synthesis data extraction by Gemini 3.0 Flash. Normalized fields are used to evaluate (e.g., material_formula, pubchem_CID).

The legacy MatBERT-only pipeline achieved a low precision of 0.65 for sol-gel synthesis classification (Table 3). This outcome was anticipated because the priority in this stage was to maximize recall by including any paper containing at least one paragraph with a sol-gel classification score greater than 0.85. Additionally, MatBERT was designed for paragraph-level classification and not the paper level evaluated here⁶². This can result in false positives for papers that include paragraphs mentioning key words associated with sol-gel, but in a non-synthetic context, such as Sohail *et al.*⁶³ and Xue *et al.*⁶⁴ Neither of these is a sol-gel paper, but both contain a paragraph discussing chemicals or laboratory procedures similar to those found in sol-gel. This led MatBERT to assign high sol-gel classification scores for these paragraphs that are not representative of the paper's overall contents. In contrast, the context-aware LLM analyzed the overall intent of these papers and correctly classified them as ‘something else’ rather than sol-gel. By incorporating this additional paper-level LLM verification, we improved the precision to 0.977 while maintaining a high recall of 0.985 (see Supplementary Fig. S1 for confusion matrix). This led to an improvement in F1 scores from 0.788 to 0.981 (Table 3).

Lastly, the LLM is tasked to further classify each sol-gel paper into inorganic and mixed inorganic-organic based on the final composition of the target material. Only papers classified as inorganic sol-gel are included in our dataset, in which the LLM achieved precision, recall, and F1 scores of 0.981, 0.911, and 0.944

respectively (Table 4). The lower scores for this sub-classification could be attributed to the LLM mislabeling some carbon-containing inorganic composites^{65,66} as mixed inorganic-organic. To maintain a more homogeneous dataset, we filter out these mixed inorganic-organic target compounds because they can involve complex sol-gel processing to fabricate encapsulated or core-shell architectures that our extraction schema is not designed to track. Furthermore, our downstream normalization pipeline relies on packages like text2chem, which were designed to parse inorganic compounds.

Validation of Paper Extraction. From the set of 200 randomly sampled papers, 104 were classified by the hybrid MatBERT and LLM filtration pipeline as inorganic sol-gel papers and processed through the synthesis data extraction pipeline, yielding 286 data records of which 279 were true sol-gel synthesis records. To evaluate this pipeline, two human evaluators independently annotated the same set of 104 papers, yielding 279 ground truth data records after discussion. The two annotators worked independently in separate spreadsheets and the ground truth was produced by discussion of the disagreements. Agreement statistics were not recorded at the time. The LLM and human annotated datasets were subsequently normalized through the same post-processing pipeline (see Methods). Each data record extracted by the LLM was then manually matched and compared to its respective ground truth data record. Exact string matches were required, with normalized entities used for the matching whenever possible (e.g., material_formula, pubchem_CID). For cases where string matching failed, the entities were further reviewed manually to determine whether the discrepancy was due to an extraction error or a post-processing normalization failure (e.g., text2chem). For the latter, we excluded them from our evaluation when aggregating true positive (TP), false positive (FP), and false negative (FN) labels across data records to compute micro-averaged precision, recall and F1 scores. An example data record with TP, FP, and FN labels is provided in Supplementary Table S1. To quantify the extent of these exclusions, Supplementary Table S2 reports the number of normalization failures and the normalization rate for each affected field. The phase_purity_details, operation_details, reagent_role, and characterization_methods fields were excluded from our evaluation given that these are intended solely to provide context.

Overall, the LLM achieved an F1 score of at least 0.935 for all evaluated entities except for phase purity classification with an F1 score of 0.899 (Table 5). Per-class metrics for this field are given in Supplementary Table S3 and Fig. S2. The impure class, which is the most consequential for downstream use, is classified with a precision of 0.96 and a recall of 0.94 ($n = 50$), whereas the insufficient-info class has a recall of 0.37 ($n = 27$) because Gemini 3.0 Flash labeled 17 of these records as pure. This lower score could be attributed to our text-only approach where phase purity information embedded in figures (e.g., X-ray diffraction, XRD) is not accessible to our LLMs, but were available and used by the annotators. The lower recall score of 0.896 for impurity phases could also be attributed to this. Another cause of error is instances in which the LLM confuses intentionally doped compounds as impurity phases. Ultimately, our data extraction is limited by how thoroughly authors describe synthesis procedures and phase purity outcomes within the main text of the paper.

In many cases where the text comprehensively describes the synthesis, the LLM demonstrated the ability to track and resolve how varying synthesis conditions lead to different phase purity outcomes. For example, in Demidzu *et al.*, which explored synthesis routes for $\text{LiMn}_{2-x}\text{Zn}_x\text{O}_4$ spinel oxides (also illustrated in Fig. 1), the LLM correctly extracted and distinguished the phase pure routes (*i.e.*, 700 °C for $X < 0.4$ and 650 °C for $X = 0.5$) from the impure routes (*i.e.*, 500 °C and 800 °C for $X = 0.5$) where ZnO or ZnMn_2O_4 impurities were detected⁶⁷. Critical annealing temperatures and dopant concentrations influencing phase purity were also explicitly noted in the phase_purity_details field by the LLM (see Supplementary Sec. 3).

In addition to the field-level metrics provided in Table 5, we also evaluate record-level accuracy (correct records / total records). From the 286 data records extracted by the LLM, 193 matched the ground truth exactly on all 11 evaluated fields, resulting in an overall record-level accuracy of 0.675. Record-level accuracy is higher for subsets of fields. For example, 235 of the 286 records (0.822) have the target and all precursors extracted correctly and linked in the same record (Supplementary Table S4, Fig. S3, and Fig. S4).

Lastly, we also evaluate paper-level accuracy (correct paper / total papers). A paper is considered “correct” only if every experimental variant was extracted from the paper, each parsed into its own data record, and contain no field-level errors. Using this strict definition, the overall paper-level data extraction accuracy is 0.490. Most paper-level failures are papers with at least one record containing a field-level error (34 papers) or with two or more experimental variants merged into a single record (15 papers), which occurs most often when all variants share the same phase-purity outcome. A looser definition that accepts merged records gives a paper-level accuracy of 0.635. The breakdown of the paper-level errors is provided in Supplementary Table S5. A comparison of field, record, and paper level accuracy is also provided in Supplementary Table S6.

Dataset Coverage. The chemical space covered by our dataset is visualized in Fig. 2. The count of data records (red) reveals that oxides, especially transition metal oxides, are the most synthesized target compounds. For each element, the average fraction of phase pure outcomes (blue) of synthesized ternary compounds containing the respective element was also calculated. Here, we observe that the average fraction of phase pure outcomes for 3d transition metal-based ternaries ranges between 73% and 87%. Transition metals can adopt multiple oxidation states and selectively stabilizing certain oxidation states under synthetic conditions can be thermodynamically challenging^{68,69}. This can lead to the formation of impurity phases^{70,71} such as Fe_2O_3 , Fe_3O_4 , CoO, and Co_3O_4 during the synthesis of compounds like BiFeO_3 and CoFe_2O_4 . To facilitate future redox-aware analysis, material entities (*i.e.*, target, precursors, and impurity phases) are also assigned predicted oxidation states using pymatgen’s `oxi_state_guesses()`⁷². Mixed valence cases such as Fe in Fe_3O_4 are reported as integer pairs (e.g., Fe: [2, 3]) rather than fractional.

Overall, Fig. 2 indicates that certain chemistries (e.g., Cu- and Ti- based) have a larger fraction of documented impure records than others. However, there are potential biases in the data that need to be accounted for.



Fig. 2 Target materials elemental frequency and phase purity trends. Upper left triangle: total data records (10^x scale). Lower right triangle: average fraction of phase-pure records calculated from ternary compounds. Elements with less than three ternary compounds with at least ten data records each are excluded from phase purity analysis (dark gray). The figure includes 117,831 of the 148,024 records (79.6%). The remainder are excluded because the target formula could not be normalized by text2chem or parsed by pymatgen.

For example, the frequency of phase pure outcomes is likely also correlated with how frequently certain targets (“difficult” or “easy”) are studied for a system. Given impurity phase formation is highly complex and dependent on the specific thermodynamic landscape of each material system, careful examination is necessary to decouple inherent biases from the underlying thermodynamics.

Validating Data-Driven Insights. This dataset can enable broad investigations into synthesis trends, allowing researchers to transition from trial-and-error approaches to data-driven exploration of the sol-gel reaction space. Beyond supporting the development of ML and thermodynamic models, this sol-gel dataset offers immediate utility through simple querying of the aggregated experimental data. This allows researchers to quickly map out previously explored experimental spaces and identify established precedents that are otherwise fragmented across thousands of publications. To validate this, we use BiFeO_3 (BFO) thin films as an example, given its abundance of data in our dataset (372 data records). Here, we evaluate the fidelity of our LLM-based text-mined dataset by attempting to replicate the findings in prior works to test if our extracted data set provides similar information to previous dedicated studies.

In our prior work^{55–57} based on a human-annotated BFO dataset, the combination of 2-methoxyethanol (2ME) and citric acid (CA), along with 2ME-only were identified as two solvent systems that consistently led to phase-pure outcomes. Using this human-annotated dataset⁵⁵ as a benchmark, we follow similar methods in Baibakova *et al.*⁵⁶ to see if our LLM-derived dataset reproduces this manually compiled trend. This is illustrated in Fig. 3a where we track how precursor and reagent choice influences phase purity outcomes. In our dataset, protocols using 2ME alone achieve a phase purity fraction (pure / pure + impure) of 89% (24/27), while those utilizing 2ME and CA achieved 100% (12/12). The remaining synthesis routes achieve a phase purity fraction of 77% (252/328). Consistent with Baibakova *et al.*⁵⁶, both these routes having a higher fraction of phase-pure outcomes than the remaining routes indicates that our extracted data supports the same conclusions as manual annotation. The 27 records using 2ME alone come from 23 different papers and the 12 records using 2ME and CA from 10, with no paper contributing more than three records to either group. A leave-one-paper-out analysis (see Code Availability) preserves the ordering of the three route groups.

Similarly, our dataset also reproduces the narrow 500°–650 °C temperature window associated with phase-pure BFO as discussed in Cruse *et al.*⁵⁵ As shown in Fig. 3b, the highest density of phase-pure syntheses occurs between 500 °C and 650 °C, where 81% (240/298) of data records result in phase-pure films. At temperatures above 700 °C, the fraction of phase-pure outcomes drops to 63% (15/24). These findings are consistent with the manually annotated data⁵⁵ and experimental observations^{9,57}, especially that higher temperatures often trigger bismuth volatilization and the formation of bismuth-poor impurity phases like $\text{Bi}_2\text{Fe}_4\text{O}_9$. This is reflected in Fig. 3c which illustrates the most common impurity phases across all BFO thin film records. Overall, these validations against prior works reinforce that our LLM-based data extraction pipeline and resulting data set

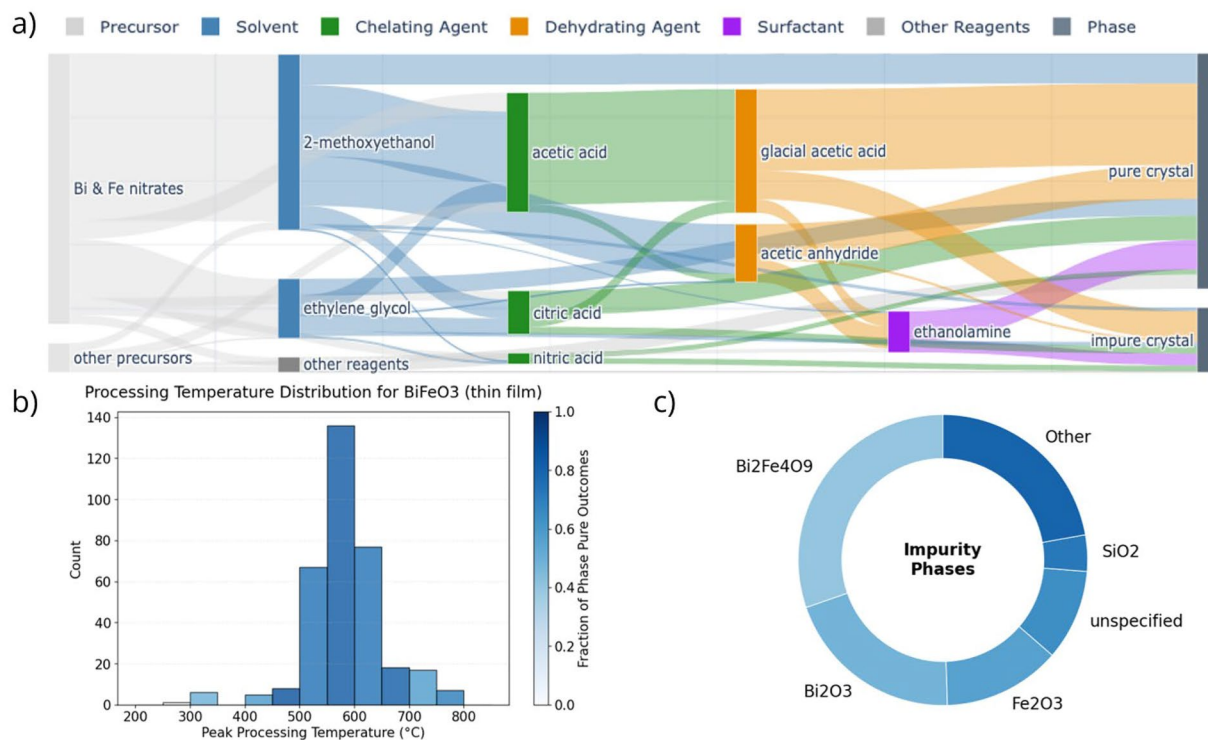


Fig. 3 Graphical examples of dataset queries for thin film BiFeO_3 . (a) Sankey diagram connecting precursor and reagent choice to phase purity outcomes. Analysis was adapted from Baibakova *et al.*⁵⁶ and applied onto our LLM-derived dataset. (b) Histogram of peak processing temperatures with phase purity heatmap. (c) Breakdown of most common impurity phases.

achieves a level of fidelity that enables findings comparable to those from labor-intensive human research, article search, and manual annotation.

Given the diversity of sol-gel routes and phase formation pathways, a text-mined inorganic sol-gel synthesis dataset requires both breadth of coverage and depth of annotation to facilitate large-scale analysis. This BFO thin film example illustrates the potential depth of analysis enabled by this dataset and can be replicated across many other compounds to identify precursors, reagents, and temperature windows associated with phase-pure synthesis. Furthermore, the identification of anomalous phase-pure routes and recurrent impurity phases also allows researchers to narrow down specific systems for further study. This can deepen our understanding of the thermodynamic and kinetic factors governing phase formation, particularly for systems where existing knowledge remains incomplete or poorly understood. Such findings from text-mined datasets can help motivate further studies, as demonstrated in Abdelsamie *et al.*⁵⁷ Beyond BFO, this dataset consists of ~148,000 more synthesis records from ~53,000 papers, spanning various compounds and sol-gel protocols. Notably, ~28,000 of these records document phase-impure syntheses, which are essential for learning the conditions that govern phase formation. Together, the breadth and depth of this dataset make it a large-scale resource for data-driven inorganic sol-gel synthesis planning.

Usage Notes

For more thorough analysis, the dataset can be further filtered into powder and thin film sol-gel synthesis (protocol.target.material_form). This distinction is crucial because powder and thin-film sol-gel processing pathways are not interchangeable. Additionally, users can filter out data records where the LLM failed to determine a phase purity outcome through the “insufficient info” label in the phase purity classification field (phase_purity.classification). This can include papers in which the text does not explicitly discuss phase purity outcomes because relevant characterization (e.g. XRD) was never performed. When using this field, it is important to note that the labels are based on what is reported rather than what is necessarily true of the sample. A “pure” label means that authors explicitly claimed phase purity or report no impurities in the text. The LLM does not interpret any XRD patterns or data from any figures. Therefore, amorphous content, impurities below the detection limits of the reported characterization methods, and impurities not discussed in the main text may not be captured. Users are encouraged to check the phase purity details (phase_purity.phase_purity_details) and characterization methods (phase_purity.characterization_methods) for additional context when assessing classification labels. The phase_purity_details and operation_details fields hold the text fragments on which Gemini 3.0 Flash based each record and point back to the supporting passages of the source paper, allowing individual records to be audited. The characterization_methods field lists every characterization method mentioned in the paper text and is not linked to individual records. Lastly, each data record is also associated with a DOI (protocol.doi) to help facilitate

paper-level analysis. There is significant value in exploring the dataset in this way, particularly by filtering for publications that report both phase-pure and impure records for the same target material. Analyzing these data records together provides a more comprehensive picture, making it easier to narrow down the chemical or thermal processing windows required to achieve phase purity.

When utilizing the data for downstream tasks, users should also account for the limitations of automated text-mining and normalization tools. Downstream packages will occasionally fail to parse the raw strings extracted by the LLM, resulting in empty data fields. In this dataset, 91% of material entities (i.e. targets, precursors, and impurity phases) were successfully normalized using text2chem, of which 86% had oxidation states from pymatgen, while 90% of reagents were resolved using pubchempy. Accordingly, when integrating this data with other text-mined datasets, users should note differences in the normalization tools used. For instance, our prior works^{34,35,54} used legacy tools like MaterialParser rather than text2chem. Resolving these discrepancies is critical for ML applications. Training models across disparate datasets requires consistent normalization schemes and labels. Because NLP tools are constantly evolving, the raw LLM-extracted strings are fully preserved in the post-processed dataset (i.e., material_string, raw_string fields). Additionally, the raw LLM extracted strings before post-processing is also released publicly (see Code Availability). Users are encouraged to leverage this raw data to apply their own custom normalization schemes as needed. The released schema-checking script (semantic_inconsistency.py) also reports four semantic consistency checks. In the current release, 1,114 records (0.75%) list the target as its own impurity phase, 846 (0.57%) are labeled pure but list an impurity phase, 804 (0.54%) are labeled impure with no impurity phase listed, and 2,108 (1.42%) list an impurity phase containing an element absent from both the target and the precursors (excluding carbonates and the C, H, N, and O of organic residues). These records are retained because their remaining fields are often correct. Users can exclude them with the script.

Finally, users should be aware of some inherent biases of this dataset. First, the corpus consists only of papers published between 2000 and 2024 by the eight publishers listed in the Methods, so earlier foundational sol-gel papers and practices may not be represented in this dataset. As a result, the frequency and choice of targets, precursors, reagents, and phase-purity outcomes may be skewed by publication year and publisher. Second, because the LLM attempts to extract each experimental variation as a separate record, a single paper reporting many conditions may contribute proportionally more data records than papers that report only a single synthesis. As such, analysis based on record count (e.g. fraction of phase-pure) may be biased towards a small number of papers with many data extractions. For ML applications, we recommend splitting data by DOI rather than individual records, since closely related records from the same paper distributed across training and test sets may lead to unintentional data leakage.

Data availability

The dataset is in a single JSONL file that is publicly available (<https://doi.org/10.6084/m9.figshare.33040070>)⁵⁹.

Code availability

The MatBERT synthesis paragraph classifier used to assist in the paper filtration is available at <https://github.com/CederGroupHub/MatBERT-synthesis-classifier>. Gemini 3.0 Flash was used to filter papers and extract data via the LBNL CBORG API (<https://cborg.lbl.gov/>). Other libraries used include pymatgen, pubchempy, pydantic, and text2chem (<https://github.com/CederGroupHub/text2chem>). The paper filtration, post-processing code and a tutorial is also publicly available (<https://github.com/hackingmaterials/sol-gel-data>). The repository also includes a scripts/ directory that regenerates every table and figure in the manuscript and the Supplementary Information from the Figshare version of the dataset cited in this paper, with a pinned environment, and a scripts/validation/ directory containing the validation labels.

Received: 22 July 2026; Accepted: 5 October 2026;

Published: xx xx xxxx

References

- Hench, L. L. & West, J. K. The sol-gel process. *Chem. Rev.* **90**, 33–72 <https://doi.org/10.1021/cr00099a003> (1990).
- Danks, A. E., Hall, S. R. & Schnepf, Z. The evolution of ‘sol-gel’ chemistry as a technique for materials synthesis. *Mater. Horiz.* **3**, 91–112 <https://doi.org/10.1039/c5mh00260e> (2016).
- Navas, D., Fuentes, S., Castro-Alvarez, A. & Chavez-Angel, E. Review on Sol-Gel Synthesis of Perovskite and Oxide Nanomaterials. *Gels* **7**, 275 <https://doi.org/10.3390/gels7040275> (2021).
- Biswas, P. K. Sol-gel thin films for optics and photonics. *J Sol-Gel Sci Technol* **59**, 456–474 <https://doi.org/10.1007/s10971-010-2368-5> (2011).
- Kaliyannan, G. V. *et al.* Investigation on sol-gel based coatings application in energy sector – A review. *Materials Today: Proceedings* **45**, 1138–1143 <https://doi.org/10.1016/j.matpr.2020.03.484> (2021).
- Chamorro, J. R. & McQueen, T. M. Progress toward Solid State Synthesis by Design. *Acc. Chem. Res.* **51**, 2918–2925 <https://doi.org/10.1021/acs.accounts.8b00382> (2018).
- Yarborough, R., Davis, K., Dawood, S. & Rathnayake, H. A sol-gel synthesis to prepare size and shape-controlled mesoporous nanostructures of binary (II–VI) metal oxides. *RSC Adv.* **10**, 14134–14146 <https://doi.org/10.1039/d0ra01778g> (2020).
- Esposito, S. Traditional Sol-Gel Chemistry as a Powerful Tool for the Preparation of Supported Metal and Metal Oxide Catalysts. *Materials* **12**, 668 <https://doi.org/10.3390/ma12040668> (2019).
- Suresh, P. & Srinath, S. Effect of synthesis route on the multiferroic properties of BiFeO₃: A comparative study between solid state and sol-gel methods. *Journal of Alloys and Compounds* **649**, 843–850 <https://doi.org/10.1016/j.jallcom.2015.07.152> (2015).
- Paulus, A. *et al.* Sol Gel vs Solid State Synthesis of the Fast Lithium-Ion Conducting Solid State Electrolyte Li₇La₃Zr₂O₁₂ Substituted with Iron. *J. Electrochem. Soc.* **166**, A5403 <https://doi.org/10.1149/2.0641903jes> (2019).
- Bretos, I. *et al.* Low-temperature sol-gel methods for the integration of crystalline metal oxide thin films in flexible electronics. *J Sol-Gel Sci Technol* **107**, 269–277 <https://doi.org/10.1007/s10971-023-06065-2> (2023).

12. Veber, A., Kunej, Š., Cerc Korošec, R. & Suvorov, D. The effects of solvents on the formation of sol–gel-derived Bi₁₂SiO₂₀ thin films. *Journal of the European Ceramic Society* **30**, 2475–2480 <https://doi.org/10.1016/j.jeurceramsoc.2010.04.026> (2010).
13. Zayat, M. & Levy, D. Blue CoAl₂O₄ Particles Prepared by the Sol–Gel and Citrate–Gel Methods. *Chem. Mater.* **12**, 2763–2769 <https://doi.org/10.1021/cm001061z> (2000).
14. Kodalle, T. *et al.* Solvent Determines the Formation Pathway in Sol–Gel Synthesized Disordered Rock Salt Material for Lithium Ion Battery Application. *Nano Lett.* **25**, 11647–11654 <https://doi.org/10.1021/acs.nanolett.5c02618> (2025).
15. Livage, J., Babonneau, F., Chatry, M. & Coury, L. Sol–gel synthesis and NMR characterization of ceramics. *Ceramics International* **23**, 13–18 [https://doi.org/10.1016/0272-8842\(95\)00133-6](https://doi.org/10.1016/0272-8842(95)00133-6) (1997).
16. Motta, M., Deimling, C. V., Saeki, M. J. & Lisboa-Filho, P. N. Chelating agent effects in the synthesis of mesoscopic-size superconducting particles. *J Sol-Gel Sci Technol* **46**, 201–207 <https://doi.org/10.1007/s10971-007-1673-0> (2008).
17. Park, H., Mastej, K. O., Dettrattawichai, P., Nduma, R. & Walsh, A. Closing the synthesis gap in computational materials design. *Nature Synthesis* **5**, 651–659 <https://doi.org/10.1038/s44160-026-01027-2> (2026).
18. Williamson, E. M. & Brutchey, R. L. Using Data-Driven Learning to Predict and Control the Outcomes of Inorganic Materials Synthesis. *Inorg. Chem.* **62**, 16251–16262 <https://doi.org/10.1021/acs.inorgchem.3c02697> (2023).
19. Kovnir, K. Predictive Synthesis. *Chem. Mater.* **33**, 4835–4841 <https://doi.org/10.1021/acs.chemmater.1c01484> (2021).
20. Olivetti, E. A. *et al.* Data-driven materials research enabled by natural language processing and information extraction. *Appl. Phys. Rev.* **7**, 041317 <https://doi.org/10.1063/5.0021106> (2020).
21. Wilary, D. M. & Cole, J. M. ReactionDataExtractor 2.0: A Deep Learning Approach for Data Extraction from Chemical Reaction Schemes. *J. Chem. Inf. Model.* **63**, 6053–6067 <https://doi.org/10.1021/acs.jcim.3c00422> (2023).
22. Wang, Z. *et al.* ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols. *Digital Discovery* **1**, 313–324 <https://doi.org/10.1039/d1dd00034a> (2022).
23. Swain, M. C. & Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **56**, 1894–1904 <https://doi.org/10.1021/acs.jcim.6b00207> (2016).
24. Mavračić, J., Court, C. J., Isazawa, T., Elliott, S. R. & Cole, J. M. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *J. Chem. Inf. Model.* **61**, 4280–4289 <https://doi.org/10.1021/acs.jcim.1c00446> (2021).
25. Dong, Q. & Cole, J. M. Snowball 2.0: Generic Material Data Parser for ChemDataExtractor. *J. Chem. Inf. Model.* **63**, 7045–7055 <https://doi.org/10.1021/acs.jcim.3c01281> (2023).
26. Trewartha, A. *et al.* Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. *Patterns* **3**, 100488 <https://doi.org/10.1016/j.patter.2022.100488> (2022).
27. Gupta, T., Zaki, M., Krishnan, N. M. A. & Mausam MatSciBERT: A materials domain language model for text mining and information extraction. *npj Comput Mater* **8**, 102 <https://doi.org/10.1038/s41524-022-00784-w> (2022).
28. Huang, S. & Cole, J. M. BatteryBERT: A Pretrained Language Model for Battery Database Enhancement. *J. Chem. Inf. Model.* **62**, 6365–6377 <https://doi.org/10.1021/acs.jcim.2c00035> (2022).
29. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 <https://doi.org/10.1038/s41586-019-1335-8> (2019).
30. Kim, E. *et al.* Machine-learned and codified synthesis parameters of oxide materials. *Sci. Data* **4**, 170127 <https://doi.org/10.1038/sdata.2017.127> (2017).
31. Mysore, S. *et al.* The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. in *Proceedings of the 13th Linguistic Annotation Workshop* 56–64 (Association for Computational Linguistics, 2019). <https://doi.org/10.18653/v1/W19-4007>
32. Jensen, Z. *et al.* A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction. *ACS Cent. Sci.* **5**, 892–899 <https://doi.org/10.1021/acscentsci.9b00193> (2019).
33. Kelly, C. R. & Cole, J. M. A dataset of Curie and Néel temperatures auto-generated with ChemDataExtractor and the Snowball algorithm. *Sci. Data* **12**, 2000 <https://doi.org/10.1038/s41597-025-06244-6> (2025).
34. Kononova, O. *et al.* Text-mined dataset of inorganic materials synthesis recipes. *Sci. Data* **6**, <https://doi.org/10.1038/s41597-019-0224-1> (2019).
35. Wang, Z. *et al.* Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Sci. Data* **9**, 231 <https://doi.org/10.1038/s41597-022-01317-2> (2022).
36. Lu, A.-Y. *et al.* Text Mining of CVD Synthesis Recipes for 2D Materials. *Advanced Materials* **38**, e09132 <https://doi.org/10.1002/adma.202509132> (2026).
37. Lederbauer, M. *et al.* LeMat-Synth: a multi-modal toolbox to curate broad synthesis procedure databases from scientific literature. Preprint at <https://doi.org/10.48550/arXiv.2510.26824> (2025).
38. He, T. *et al.* Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science Advances* **9**, eadg8180 <https://doi.org/10.1126/sciadv.adg8180> (2023).
39. He, T. *et al.* Similarity of Precursors in Solid-State Synthesis as Text-Mined from Scientific Literature. *Chemistry of Materials* **32**, 7861–7873 <https://doi.org/10.1021/acs.chemmater.0c02553> (2020).
40. Huo, H. *et al.* Machine-Learning Rationalization and Prediction of Solid-State Synthesis Conditions. *Chemistry of Materials* **34**, 7323–7336 <https://doi.org/10.1021/acs.chemmater.2c01293> (2022).
41. Karpovich, C., Pan, E., Jensen, Z. & Olivetti, E. Interpretable Machine Learning Enabled Inorganic Reaction Classification and Synthesis Condition Prediction. *Chemistry of Materials* **35**, 1062–1079 <https://doi.org/10.1021/acs.chemmater.2c03010> (2023).
42. Kim, E. *et al.* Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks. *J. Chem. Inf. Model.* **60**, 1194–1201 <https://doi.org/10.1021/acs.jcim.9b00995> (2020).
43. Jia, X. *et al.* Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **573**, 251–255 <https://doi.org/10.1038/s41586-019-1540-5> (2019).
44. Prein, T. *et al.* Language Models Enable Data-Augmented Synthesis Planning for Inorganic Materials. *ACS Applied Materials & Interfaces* **17**, 69221–69233 <https://doi.org/10.1021/acsaami.5c11229> (2025).
45. Sun, W. & David, N. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. *Faraday Discussions* **256**, 614–638 <https://doi.org/10.1039/d4fd00112e> (2025).
46. Schlesinger, J., Hjaltason, S., Szymanski, N. J. & Bartel, C. J. Thermodynamic assessment of machine learning models for solid-state synthesis prediction. Preprint at <https://doi.org/10.48550/arXiv.2602.04075> (2026).
47. Toniato, A., Vaucher, A. C., Laino, T. & Graziani, M. Negative chemical data boosts language models in reaction outcome prediction. *Science Advances* **11**, eadt5578 <https://doi.org/10.1126/sciadv.adt5578> (2025).
48. Strieth-Kalthoff, F. *et al.* Machine Learning for Chemical Reactivity: The Importance of Failed Experiments. *Angewandte Chemie International Edition* **61**, e202204647 <https://doi.org/10.1002/anie.202204647> (2022).
49. Maloney, M. P. *et al.* Negative Data in Data Sets for Machine Learning Training. *J. Org. Chem.* **88**, 5239–5241 <https://doi.org/10.1021/acs.joc.3c00844> (2023).
50. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 <https://doi.org/10.1038/nature17439> (2016).
51. Moosavi, S. M. *et al.* Capturing chemical intuition in synthesis of metal-organic frameworks. *Nat. Commun.* **10**, 539 <https://doi.org/10.1038/s41467-019-08483-9> (2019).

52. Pan, E. *et al.* ZeoSyn: A Comprehensive Zeolite Synthesis Dataset Enabling Machine-Learning Rationalization of Hydrothermal Parameters. *ACS Cent. Sci.* **10**, 729–743 <https://doi.org/10.1021/acscentsci.3c01615> (2024).
53. Thway, M. *et al.* Harnessing GPT-3.5 for text parsing in solid-state synthesis – case study of ternary chalcogenides. *Digital Discovery* **3**, 328–336 <https://doi.org/10.1039/d3dd00202k> (2024).
54. Lee, S., Cruse, K., Baibakova, V., Ceder, G. & Jain, A. Text-mined dataset of solid-state syntheses with impurity phases using Large Language Model. *Sci. Data* **12**, 1969 <https://doi.org/10.1038/s41597-025-06222-y> (2025).
55. Cruse, K. *et al.* Text Mining the Literature to Inform Experiments and Rationalize Impurity Phase Formation for BiFeO₃. *Chem. Mater.* **36**, 772–785 <https://doi.org/10.1021/acs.chemmater.3c02203> (2024).
56. Baibakova, V. *et al.* Precursor reaction pathway leading to BiFeO₃ formation: insights from text-mining and chemical reaction network analyses. *Digital Discovery* **4**, 1602–1611 <https://doi.org/10.1039/d5dd00160a> (2025).
57. Abdelsamie, M. *et al.* Combining text mining, *in situ* characterization, and *ab initio* calculations to rationalize BiFeO₃ crystallization pathways. *Matter* **6**, 4291–4305 <https://doi.org/10.1016/j.matt.2023.10.002> (2023).
58. Xin, H., Kitchin, J. R. & Kulik, H. J. Towards agentic science for advancing scientific discovery. *Nat Mach Intell* **7**, 1373–1375 <https://doi.org/10.1038/s42256-025-01110-x> (2025).
59. Chen, W. *et al.* Text-mined dataset of sol-gel synthesis protocols with phase purity outcomes using large language models. <https://doi.org/10.6084/m9.figshare.33040070> (2026).
60. Liu, T. & Xu, Y. Synthesis of nanocrystalline LaFeO₃ powders via glucose sol–gel route. *Materials Chemistry and Physics* **129**, 1047–1050 <https://doi.org/10.1016/j.matchemphys.2011.05.054> (2011).
61. Gabal, M. A. *et al.* Auto-combustion synthesis and characterization of perovskite-type LaFeO₃ nanocrystals prepared via different routes. *Ceramics International* **45**, 16530–16539 <https://doi.org/10.1016/j.ceramint.2019.05.187> (2019).
62. Huo, H. *et al.* Semi-supervised machine-learning classification of materials synthesis procedures. *npj Comput Mater* **5**, 62 <https://doi.org/10.1038/s41524-019-0204-1> (2019).
63. Sohail, M., De Marco, R., Lamb, K. & Bakker, E. Thin layer coulometric determination of nitrate in fresh waters. *Analytica Chimica Acta* **744**, 39–44 <https://doi.org/10.1016/j.aca.2012.07.026> (2012).
64. Xue, S. *et al.* Boron-doped activated carbon derived from *Zoysia sinica* for Rhodamine B adsorption: The crucial roles of defect structures. *FlatChem* **34**, 100390 <https://doi.org/10.1016/j.flatc.2022.100390> (2022).
65. Song, C., Chen, P., Wang, C. & Zhu, L. Photodegradation of perfluorooctanoic acid by synthesized TiO₂–MWCNT composites under 365 nm UV irradiation. *Chemosphere* **86**, 853–859 <https://doi.org/10.1016/j.chemosphere.2011.11.034> (2012).
66. Cen, H., Wu, C. & Chen, Z. N. S Co-doped carbon coated MnS/MnO/Mn nanoparticles as a novel corrosion inhibitor for carbon steel in CO₂-saturated NaCl solution. *Colloids and Surfaces A: Physicochemical and Engineering Aspects* **630**, 127528 <https://doi.org/10.1016/j.colsurfa.2021.127528> (2021).
67. Demidzu, H., Nakamura, T. & Yamada, Y. Synthesis and magnetic properties of Li–Mn–Zn spinel oxides. *Journal of Magnetism and Magnetic Materials* **322**, 1816–1821 <https://doi.org/10.1016/j.jmmm.2009.12.033> (2010).
68. Kiselev, Y. M. & Tretiyakov, Y. D. The problem of oxidation state stabilisation and some regularities of a Periodic system of the elements. *Russ. Chem. Rev.* **68**, 365–379 <https://doi.org/10.1070/rc1999v068n05abeh000496> (1999).
69. Kiselev, Y. M. Stabilization of oxidation states in transition metals. *Russ. J. Inorg. Chem.* **52**, 1717–1725 <https://doi.org/10.1134/s0036023607110149> (2007).
70. Fawaeer, S. H. *et al.* Role of Buffer Layers in Defect Chemistry and Parasitic Phase Formation of BiFeO₃ Films on Silicon. *ACS Omega* **11**, 7782–7795 <https://doi.org/10.1021/acsomega.5c08852> (2026).
71. Cao, Z. & Zuo, C. Direct Synthesis of Magnetic CoFe₂O₄ Nanoparticles as Recyclable Photo-Fenton Catalysts for Removing Organic Dyes. *ACS Omega* **5**, 22614–22620 <https://doi.org/10.1021/acsomega.0c03404> (2020).
72. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319 <https://doi.org/10.1016/j.commatsci.2012.10.028> (2013).

Acknowledgements

We thank Rohan Ukhade, Raphael Moral, and Anika Promi for assisting in the manual data validation and feedback on the data extraction pipeline. We also thank Viktoriia Baibakova and Kevin Cruse for developing the source dataset of human-annotated BiFeO₃ thin film synthesis. Finally, we thank Anna Sackmann, Rachael Samberg, and Timothy Vollmer (Science Data and Engineering Librarians at UC Berkeley) for helping us obtain Text and Data Mining agreements with publishers.

Author contributions

W.Y.C developed and validated the data extraction pipeline, analyzed the data, and wrote the manuscript. S.L. co-developed the data extraction pipeline, and supervised the development of preliminary datasets. T.K. inspected the data, provided sol-gel expertise, and revised the manuscript. P.S. inspected the data and contributed to the post-processing pipeline. G.C. supervised the project and revised the manuscript. C.M.S.-F. provided sol-gel expertise, supervised the project, and revised the manuscript. A.J. developed the approach, supervised the project, and revised the manuscript. All authors discussed results and reviewed the manuscript.

Funding

This work was intellectually led by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division under Contract No. DE-AC02-05CH11231 (D2S2 program KCD2S2). This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award BES-ERCAP0028631. This work was also supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internships (SULI) Program.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41597-026-08492-6>.

Correspondence and requests for materials should be addressed to A.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.