Molecular Catalysis 601 (2026) 116097

Contents lists available at ScienceDirect

Molecular Catalysis

journal homepage: www.journals.elsevier.com/molecular-catalysis

Automated extraction of synthesis recipes from the literature using large language models: Application to zeolite ZSM-5☆

Tanvir Ahmed, Jeffrey D. Rimer, Mingjian Wen *, Jeremy C. Palmer *

Department of Chemical and Biomolecular Engineering, University of Houston, Houston, TX, USA

ARTICLE INFO

Dataset link: https://doi.org/10.5281/zenodo.18761174

Keywords:
Zeolites
ZSM-5
Synthesis recipe
Large language models
Data mining

ABSTRACT

Data-driven strategies for accelerating materials discovery and optimization are often constrained by the lack of large, reliable, and machine-readable synthesis data. Although numerous synthesis recipes have been reported in the scientific literature, these data are predominantly presented in unstructured and heterogeneous formats, which substantially constrain their direct integration into data-driven research workflows. Here, we present an automated, fully open-source approach that uses large language models (LLMs) to extract structured synthesis information directly from the scientific literature. We apply it to zeolites, focusing on the hydrothermal synthesis of ZSM-5, an industrially important catalyst with physicochemical properties that are highly sensitive to the selection of synthesis conditions. The workflow integrates large-scale document processing, prompt-guided extraction using a compact open-source LLM, and chemistry-aware post-processing to generate a database of ZSM-5 synthesis recipes. The resulting dataset consists of 1659 ZSM-5 synthesis entries spanning key parameters, including temperature, crystallization time, precursor selection, and synthesis gel composition among others. Validation is performed against a manually curated set of synthesis recipes. The automated approach demonstrates strong extraction accuracy across most categories, while statistical analysis of the dataset reveals synthesis trends consistent with established understanding of ZSM-5 crystallization. We discuss limitations arising from incomplete and inconsistent reporting in the literature and identify opportunities to enhance the reliability and robustness of automated synthesis data extraction as a foundation for the rational design of next-generation zeolite catalysts for a broad range of commercial applications.

1. Introduction

Zeolites are crystalline aluminosilicate materials containing ordered pore networks with molecular-scale dimensions [1,2]. Their high surface area, tunable acidity, shape and size selectivity, and thermal and chemical stability facilitate their use in myriad large-scale industrial applications, including heterogeneous catalysis, separations, adsorption, and ion exchange [1–5]. Despite their technological importance, optimizing zeolites for specific applications remains a formidable challenge. Zeolite performance is sensitive to framework topology, chemical composition, defect density, crystal size and morphology, and the incorporation of heteroatoms or extra-framework species [1,2,6–10]. These characteristics are governed by a high-dimensional synthesis space involving precursor chemistry, organic structure-directing agents (OSDAs), mineralizers, gel composition, temperature, time, and post-synthetic treatments [1,2,11–13]. Small variations in one or more of these synthesis parameters can lead to zeolite materials with different physiochemical properties, making systematic experimental exploration to identify optimal materials for targeted applications time-consuming, costly, and unpredictable [1,2,14,15]. Consequently, efforts to optimize zeolite materials have relied heavily on expert intuition and Edisonian experimentation [1,16,17]. An empirical approach to zeolite catalyst optimization is even more challenging due to the complex mechanisms of zeolite crystallization, [2] which emphasizes the importance of understanding the effects of synthesis conditions on material properties. Prior studies have demonstrated that subtle changes to zeolite growth conditions can significantly impact their physicochemical properties, and hence catalyst performance, which is particularly evident in attempts to achieve ultrasmall particle sizes, [18] hierarchical architectures, [19,20] or introduce special gradients in acid site density. [21]

Data-driven approaches have shown tremendous promise in navigating complex material design spaces [15,22–24]. However, a major

☆ This article is part of a Special issue entitled: 'In honor of Prof. Dion Vlachos' published in Molecular Catalysis.

* Corresponding authors.
E-mail addresses: wenxx151@gmail.com (M. Wen), jcpalmer@uh.edu (J.C. Palmer).

https://doi.org/10.1016/j.mcat.2026.116097
Received 2 April 2026; Received in revised form 22 May 2026; Accepted 30 May 2026
Available online 16 June 2026
2468-8231/© 2026 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

bottleneck to the application of such approaches is the accumulation of reliable synthesis data from the vast and heterogeneous literature. Zeolite synthesis information is distributed across thousands of research articles published over the past half-century and reported with varying levels of detail using inconsistent terminology. Further, synthesis conditions are scattered across text, tables, and figures, making manual curation labor-intensive and difficult to scale. These challenges have limited the size, completeness, and reproducibility of datasets available for modeling and machine-learning applications.

Early attempts to address this data curation problem relied on rule-based text-mining pipelines. Tools such as `OSCAR4` and `ChemicalTagger` used predefined dictionaries, grammatical rules, and pattern matching to identify chemical entities and experimental relationships [25,26]. Related approaches focused on extracting numerical data from tables using document-structure parsing [27], and similar rule-driven pipelines have been successfully applied to extract reaction steps and synthesis workflows from patents and journal articles [28,29]. For zeolites, these methods enabled the first automated compilations of hydrothermal synthesis recipes using tailored heuristics, table parsing, and regular expressions [30] and the subsequent development of the `ZeoSyn` dataset comprising nearly 24,000 syntheses across 233 framework types [31]. Related work also relied on keyword filtering, text matching, and natural language processing (NLP) algorithms to analyze reported synthesis data and identify trends linking OSDAs to zeolite framework formation [32]. Nonetheless, they require extensive manual rule development and exhibit limited robustness to variations in writing style and experimental reporting conventions.

More recently, large language models (LLMs) have fundamentally altered this landscape by enabling context-aware interpretation of published experimental protocols. Domain-adapted models such as `MatSciBERT` improve recognition of chemical entities and synthesis relationships in complex and variable sentence structures [33]. General-purpose LLMs now offer additional capabilities, including reasoning across sentences and sections, integrating information from text and tables, and adapting to diverse reporting styles with minimal manual rule design [34,35]. These advances are particularly well-suited to zeolite synthesis literature, where key parameters are often distributed across multiple sections of a manuscript. Recent approaches that combine domain-adapted encoders with structured extraction workflows have also shown promise in mining synthesis recipes at scale [36]. Several recent studies have extended this direction specifically within zeolite synthesis research. For example, zeolite synthesis event extraction has been formulated as a structured NLP task, with recent LLM-based evaluations showing that high-level synthesis events can be identified more reliably than fine-grained experimental arguments [37]. Other recent work has focused on automated extraction and validation of synthesis arguments from zeolite synthesis text [38]. Beyond extraction, LLMs have also been explored for zeolite synthesis design, including knowledge-informed OSDA selection and agent-based de novo OSDA generation [39,40].

Motivated by recent advances in general-purpose LLMs for scientific text mining, this work presents an automated pipeline for extracting structured materials synthesis information from the published literature using a fully open-source LLM. The proposed approach integrates document processing, prompt-guided extraction, and chemistry-aware post-processing to convert unstructured experimental descriptions into structured, machine-readable synthesis recipes. As a focused case study, the approach is applied to the hydrothermal synthesis of zeolite ZSM-5, an industrially important and extensively studied framework whose physicochemical properties are highly sensitive to synthesis conditions [1]. Despite relying on a compact, non-fine-tuned LLM, the extraction achieved strong accuracy across most synthesis parameter categories, consistent with the quality of reporting in the underlying literature. Beyond dataset construction, the extracted synthesis records enable systematic analysis of synthesis trends reported across the ZSM-5 literature. Limitations of the approach associated with incomplete and heterogeneous reporting are discussed, along with potential avenues for improving automated synthesis data extraction. Our study suggests that with further refinement to address some of these limitations, the proposed approach will be extensible to more complex zeolite systems and facilitate the synthesis of materials with specific, targeted properties tailored for different applications.

2. Methods

The construction of the ZSM-5 synthesis dataset follows a multi-staged workflow that integrates large-scale document processing, automated extraction via a LLM, and chemistry-aware post-processing (Fig. 1). The first stage, *Document processing*, encompasses PDF acquisition, text extraction, and preliminary formatting to convert research articles into machine-readable text. The second stage, *Data extraction*, leverages domain-specific knowledge to formulate instructions that guide the LLM in capturing synthesis-relevant entities and experimental conditions, and a language-model inference followed by structured parsing to get the output. The final stage, *Data cleaning*, includes cleaning, and standardization of outputs into a format suitable for downstream analysis. A full description of the methodology and a discussion of alternative approaches are presented in the Supporting Information (SI); the following describes the final implementation used to curate and analyze the ZSM-5 synthesis dataset.

2.1. Document collection and preprocessing

Research articles related to ZSM-5 were collected from the Web of Science database using a keyword-based search strategy designed to capture both synthesis and characterization studies. The initial corpus comprised 12,287 publications spanning multiple years and publishers and served as the input for subsequent processing steps. All PDF documents were converted into machine-readable text using the open-source tool `MinerU` [41], which extracts text, equations, tables, and other structural elements while preserving the logical reading order of each document. Although this preprocessing step retained text and many tabular elements, information present solely as image content was not comprehensively extracted. This limitation is relevant for information contained in microscopy images, graphical plots, or other figure-only content. In particular, morphology-related information such as crystal size, morphology, and faceting may be missed when it is shown only in SEM or TEM images and not described in the surrounding text or figure captions. This standardized text representation enabled reliable downstream parsing and analysis.

To retain only studies that report ZSM-5 synthesis protocols, an automated relevance filtering step was applied using the Mistral Instruct 2.0 LLM [42] with prompt engineering (see Section S2 in the SI for the prompts). Papers were retained if they contained at least one synthesis-relevant parameter, resulting in a filtered set of 3008 articles that formed the basis for structured data extraction.

2.2. Data extraction and error handling

The Mistral Instruct 2.0 LLM was employed for automated synthesis data extraction due to its open-source availability and its support for a 32,800-token context window. This extensive capacity allows for the analysis of an entire research article in a single pass, rather than restricting extraction to isolated sections or requiring the fragmentation of text into smaller segments. This capability is critical because synthesis parameters are typically scattered throughout the manuscript, appearing in the main text, experimental tables, and even figures. We also tested other open-source LLMs such as LLaMA. Mistral Instruct 2.0 showed higher accuracy in our preliminary test, and thus it was adopted.

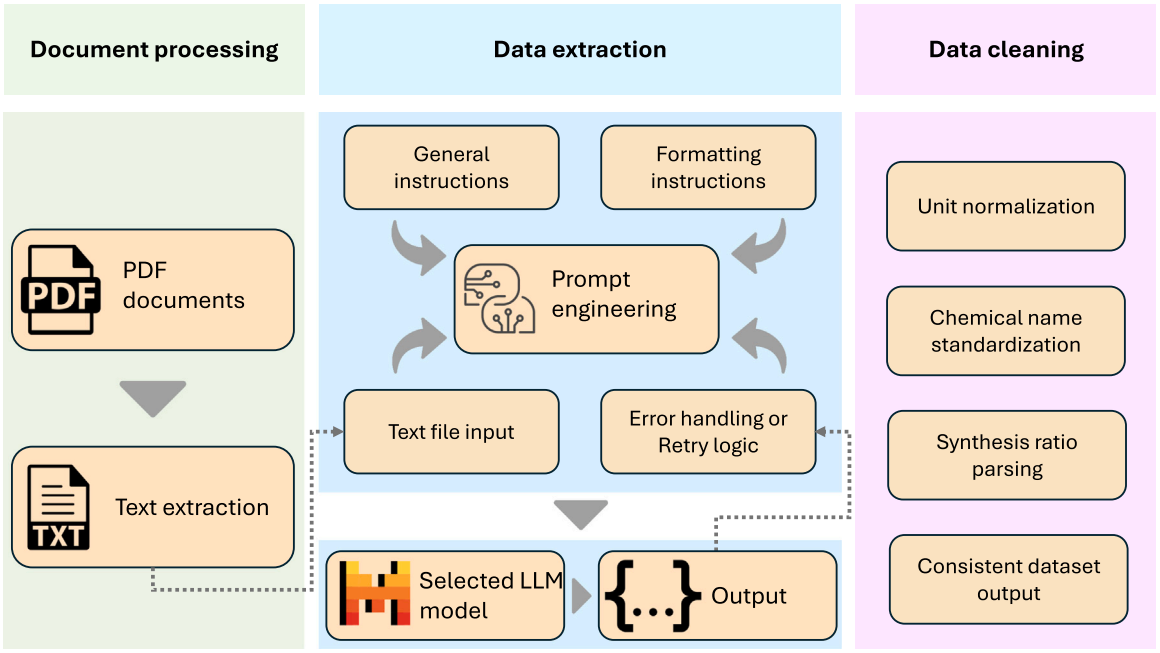

Fig. 1. LLM-based workflow for synthesis recipe extraction. The process consists of three stages: Document processing (left column) where PDF research articles are converted into machine-readable text; data extraction (middle column) based on prompt engineering using structured instructions and error-handling strategies; and synthesis information generation via the Mistral LLM, and data cleaning and standardization (right column), where extracted outputs are cleaned and converted into a consistent dataset.

Table 1
Categories of synthesis information extracted by the LLM and their descriptions.

Category	Description
Synthesis composition (molar ratios)	The molar ratios of different precursors used in the synthesis recipe (e.g., silica, alumina, OSDA, water, etc.), typically reported in the form of a chemical composition ratio.
Synthesis temperature(s)	The temperature(s) at which the hydrothermal synthesis or crystallization process was conducted.
Synthesis time(s)	The duration(s) of the synthesis or crystallization step, often reported in hours or days.
Silica source(s)	The chemical compounds or materials that supply silicon in the synthesis gel or solution (e.g., tetraethyl orthosilicate (TEOS), fumed silica, sodium silicate).
Aluminum source(s)	The chemical compounds or materials that provide aluminum for the zeolite framework (e.g., aluminum isopropoxide, sodium aluminate).
Organic structure-directing agent(s) (OSDAs)	The organic molecules that direct the crystallization of a specific zeolite framework during synthesis (e.g., TPAOH, TEAOH, TBAOH).
Bulk Si/Al ratio(s)	The silicon-to-aluminum molar ratio measured in the bulk zeolite sample after synthesis.
Surface Si/Al ratio(s)	The silicon-to-aluminum molar ratio measured on the external surface of the zeolite crystals.
Crystal size(s)	The reported size or size range of the synthesized zeolite crystals, typically in nanometers or micrometers.

A key task in using the LLM to extract synthesis information involves designing appropriate prompts to instruct the LLM to accomplish the work. To optimize the extraction, we developed a prompt engineering strategy focused on both information coverage and structural reliability. First, we designed prompts to specify the precise information categories to be extracted, which are detailed in Table 1. The LLM was instructed to act as a research assistant, scanning the entire manuscript to recover synthesis parameters that are often scattered in a scientific article. For each targeted field, the prompt explicitly requires either the specific value or the string "Not provided" to ensure completeness and minimize hallucinations. Second, we designed prompts to enforce a strict output format by mandating a structured JSON schema. By combining general extraction instructions with well-formatted schema definitions and representative examples, we guided the model toward a consistent, machine-readable structure that facilitates reliable downstream parsing. The full prompt templates used for this extraction are provided in Section S2 in the SI.

To maintain robustness, error-handling logic was integrated into the extraction workflow to ensure the reliability of the synthesis dataset. Although we utilized a Pydantic [43] data model to define a target JSON schema, the Mistral LLM occasionally failed to adhere to this structure, producing outputs with malformed syntax or extraneous text. As illustrated in Fig. 2, a custom post-processing pipeline was developed to address these issues by stripping markdown code blocks, filtering out nested schema definitions, and removing invalid or empty dictionaries.

Once valid JSON objects are isolated, the framework identifies the most relevant synthesis record. Because the keys generated by the LLM do not always match the standardized names specified in Table 1, we employed a fuzzy matching pipeline using RapidFuzz [44] (Fig. 3). For example, if the model outputs a key labeled `Silicon source` instead of the target key `Silica (Si) source`, fuzzy string matching calculates a similarity score based on token normalization to correctly map the data to the specified schema.

To further improve extraction robustness, a multi-pass retry mechanism was implemented. If the initial processing fails to yield a valid or complete JSON object, the prompt is automatically reissued to the model with minimal refinements. This iterative approach ensures formatting compliance and maximizes the recovery of synthesis parameters from the full-text corpus.

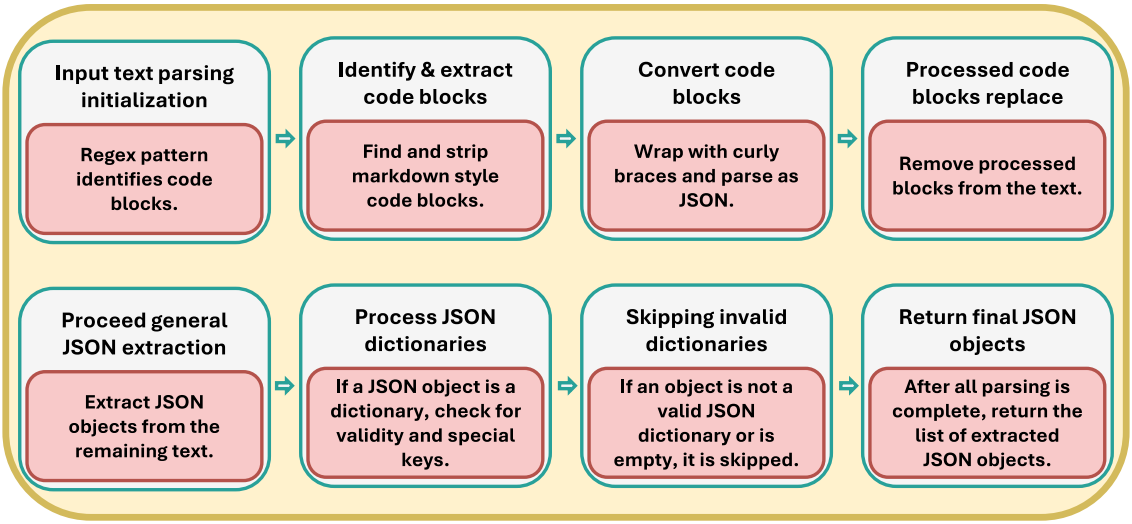

Fig. 2. Pipeline for extracting, cleaning, and validating JSON objects from LLM-generated outputs. The pipeline removes irrelevant code segments, filters nested schema structures, and isolates valid synthesis records for data consistency.

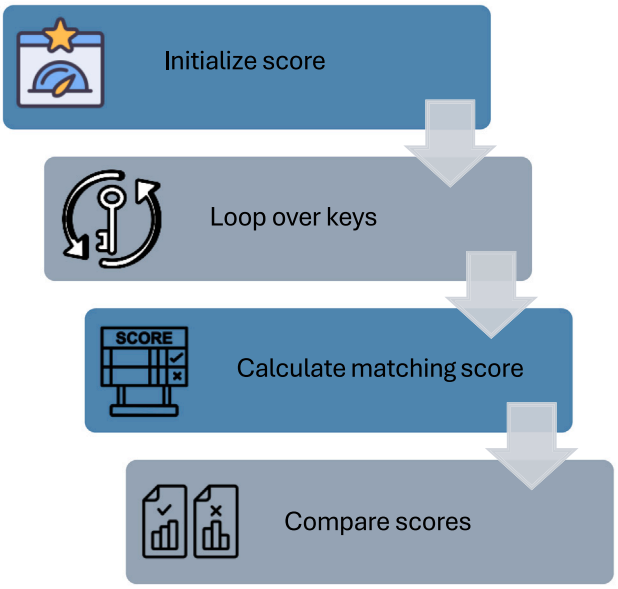

Fig. 3. Fuzzy matching pipeline for selecting the most relevant JSON object. Key normalization and scoring against predefined targets ensure that the best-structured synthesis data is retained.

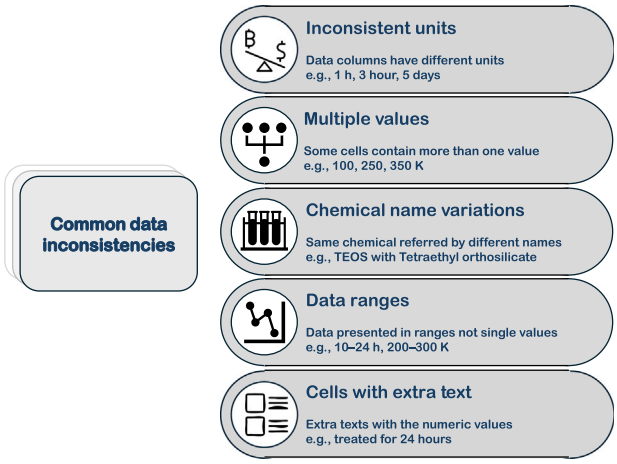

Fig. 4. Representative examples of data inconsistencies in the extracted dataset. Common issues included inconsistent units, multiple values reported in the same entry, interchangeable chemical names, use of ranges rather than discrete values, and extraneous descriptive text accompanying quantitative information.

2.3. Data cleaning and standardization

Following extraction, the dataset underwent a dedicated data cleaning and standardization procedure to ensure consistency for downstream machine learning analysis. Although the previously described error-handling logic in Section 2.2 ensures structural validity (i.e., that the model outputs a valid JSON object), a separate data cleaning and standardization procedure was developed to address content-level inconsistencies. This latter stage focuses on harmonizing the extracted information into a uniform format without changing the underlying data.

As summarized in Fig. 4, common inconsistencies in the raw outputs included unit mismatches, multiple values within single entries, interchangeable chemical names, and descriptive text interleaved with numerical values. To address these issues, reported numerical parameters were normalized to common units using rule-based parsing and regular expressions. Specifically, temperatures were converted to degrees Celsius (°C), synthesis times to hours (h), and crystal sizes to nanometers (nm), while extraneous qualifiers and non-numeric descriptors were removed. Chemical precursor names were standardized to canonical forms to resolve inconsistent naming across studies. We utilized an LLM-based normalization strategy to preserve meaningful distinctions among different silicon, aluminum, and OSDA sources while maintaining commercial formulation details where available.

Synthesis composition entries were further normalized using a specialized LLM prompt designed to rewrite heterogeneous textual recipes into a fixed reagent–ratio representation. These reported recipes were transformed into structured key–value pairs suitable for quantitative analysis, removing extraneous descriptors in the process. Finally, the parsed composition dictionaries were expanded into separate columns, ensuring that individual chemical components appear as distinct fields in the final dataset.

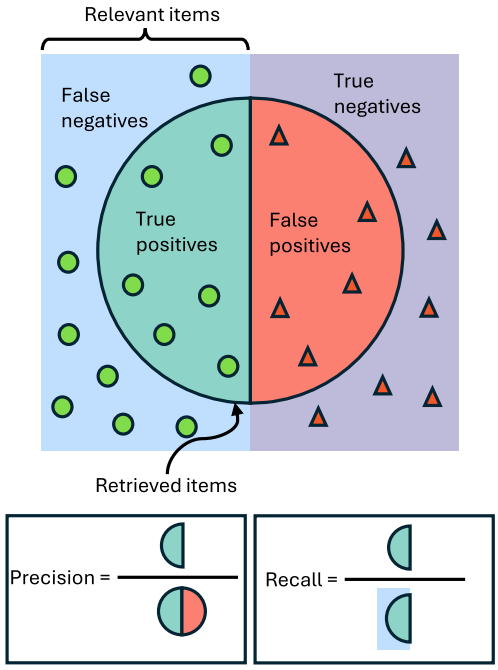

Fig. 5. Schematic illustration of precision and recall. Retrieved and relevant items are shown, with their overlap representing true positives; false positives, false negatives, and true negatives are explicitly labeled to illustrate the performance of the extraction model.

2.4. Dataset benchmarking metrics

Extraction performance was evaluated using precision, recall, and F_1 score. Precision and recall are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{\text{Relevant retrieved items}}{\text{All retrieved items}}. \quad (1)$$

and

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{Relevant retrieved items}}{\text{All relevant items}}. \quad (2)$$

where TP stands for true positives, FP for false positives, FN for false negatives, and TN for true negatives. A schematic illustration of precision and recall is shown in Fig. 5. F_1 score is the harmonic mean of precision and recall,

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (3)$$

Precision measures the fraction of retrieved values that are correct, whereas recall measures the fraction of relevant values successfully captured. F_1 score provides a balanced measure of accuracy and completeness.

3. Results and discussion

3.1. Validation of extracted synthesis data

We applied the automated pipeline described in Section 2 to the extraction of ZSM-5 zeolite synthesis protocols, resulting in the successful recovery of 1659 synthesis records. To assess the reliability of this automated process, we performed a benchmark evaluation against a manually curated reference dataset. A subset of 140 papers, representing approximately 10% of the total extracted records, was randomly selected for verification by human annotators with domain expertise. For these articles, synthesis information was manually re-extracted to establish a ground-truth reference across all nine synthesis parameters listed in Table 1.

Table 2
Benchmarking results for synthesis parameter extraction.

Category	Precision	Recall	F_1 score
Synthesis temperature	0.95	0.93	0.94
Surface Si/Al	0.95	0.93	0.94
Synthesis time	0.93	0.90	0.92
OSDA	0.95	0.88	0.91
Aluminum source	0.85	0.83	0.84
Crystal size	0.91	0.71	0.80
Synthesis composition	0.78	0.76	0.77
Silica source	0.76	0.69	0.73
Bulk Si/Al	0.70	0.67	0.68

The quality of the automated extraction was then quantified by computing the precision, recall, and F_1 score for each synthesis parameter. The results of this evaluation are summarized in Table 2. The quality of the extracted data varies across categories. High-performing fields include *Synthesis temperature*, *Surface Si/Al*, *Synthesis time*, and *OSDA*, each achieving an F_1 score above 0.9. We found that these synthesis parameters are often reported in consistent and standardized formats in the literature (e.g., temperatures in °C, and explicit chemical names for OSDAs), enabling the LLM to reliably identify and extract them. We note that, unlike the other three synthesis parameters, *Surface Si/Al* is often not measured and thus is infrequently reported in the literature. Because surface Si/Al ratios are reported much less frequently than other categories, the reliability of this field was interpreted differently from commonly reported synthesis variables. In this work, a surface Si/Al value was considered valid only when the paper explicitly reported a surface-sensitive measurement, such as XPS. Values obtained from bulk elemental analysis, gel composition, ICP, XRF, or EDS were not treated as surface Si/Al ratios. Therefore, for this sparsely reported field, the main reliability concern is not only whether a numerical value is extracted correctly, but also whether the model avoids incorrectly assigning bulk or nominal Si/Al ratios as surface Si/Al values. We therefore interpret this field with caution and identify it as a higher-uncertainty descriptor in the extracted dataset. In the extracted dataset, 92% of entries for this category were labeled as "Not provided", indicating that only a small subset of papers reported this value. As a result, its high scores reflect performance on a limited number of reported cases. Nevertheless, whenever the field was reported, the model extracted it accurately, leading to high scores within that subset.

Mid-performing fields include *Synthesis composition*, *Aluminum (Al) source*, and *Crystal size*. The moderate score for *Synthesis composition* reflects the wide variability in how molar ratios are presented across papers; some use compact notations like "x SiO_2: y Al_2O_3", whereas others embed quantities in prose like "X g of SiO_2 was mixed with Y g of water". The model tends to favor regular patterns, missing composition details presented in free-text form or in unconventional formats, thus resulting in reduced scores. The Al source category showed lower performance primarily due to misclassification errors rather than missing information. In several cases, valid aluminum precursors were incorrectly assigned as silica sources, particularly in complex synthesis descriptions where Si and Al reagents were discussed in close proximity. Additional errors arose in heteroatom substituted syntheses, where metal precursors were mistakenly identified as Al sources. Crystal size showed high precision but lower recall, indicating that extracted values were usually correct, but many valid values were missed. This behavior is likely due to the way crystal size is reported in the literature, where measurements are frequently embedded in figure captions, microscopy image annotations, or image-based characterization rather than in the main text. Because the extraction pipeline operates purely on textual content, such non-textual references remain inaccessible, leading to systematic false negatives in this category.

The low-performing fields are *Silica (Si) source* and *Bulk Si/Al*. In the case of *Si source*, the model occasionally confused it with *Al source*, especially when both precursors were listed adjacently with

similar sentence structures. This category-level confusion was a recurring issue: for example, the model sometimes extracted TEOS as the Al source, due to their co-occurrence in similar linguistic contexts. For *Bulk Si/Al*, the low performance reflects variability in how such ratios are reported (e.g., explicitly embedded in figures/tables or inferred from indirect references), which poses a challenge for the LLM's pattern-based extraction.

Although the Mistral model employed in this work is a relatively small LLM with 7 billion parameters, the extraction performance achieved here is comparable to that reported in prior studies that relied on substantially larger models for related materials domains [34,45,46]. This result highlights that careful prompt design, coupled with domain-informed post-processing and validation, can compensate for model scale in complex scientific extraction tasks. The observed accuracy trends are consistent with growing evidence that optimized prompt strategies, even without model fine-tuning, can yield reliable performance in scientific information text mining [47]. At the same time, the remaining errors mainly arise from domain-specific terminology, unclear descriptions, and parameters that are reported inconsistently across studies, indicating that future improvements could benefit from stronger use of domain knowledge and hybrid retrieval-augmented generation (RAG) frameworks [48]. Several of the error types identified in this work can be reduced through additional domain-specific filters. For example, temperature and time values can be filtered using nearby keywords to distinguish crystallization from aging, drying, calcination, and ion-exchange steps. Similarly, surface Si/Al values can be restricted to text regions containing surface-sensitive terms such as 'XPS' or 'surface', which helps prevent confusion with bulk Si/Al values from ICP, XRF, EDS, or nominal gel composition. Precursor dictionaries can also be used to normalize common silica sources, alumina sources, OSDAs, and inorganic SDAs, while additional filters can flag cases where commercial ZSM-5 was used rather than synthesized. These filters would not remove all ambiguity, but they provide a practical route for reducing systematic extraction errors in future versions of the workflow.

3.2. Domain-knowledge-guided refinement

To improve the reliability of the extracted ZSM-5 synthesis dataset, domain-specific knowledge was applied systematically to evaluate and refine the initial model outputs. Although the model successfully identified synthesis-related information in most cases, several recurring errors were observed that required chemistry-aware intervention to resolve. A second round of benchmarking was performed after these post-processing consistency checks. Five categories on (*Synthesis temperature*, *Synthesis time*, *Silica source*, *Aluminum source*, and *OSDA*) were re-evaluated, and clear improvements were observed across all metrics, with the most substantial effect observed in precursor classification fields. Table 3 summarizes the before and after comparison.

The *Synthesis temperature* and *Synthesis time* categories already exhibited relatively high accuracy in the initial benchmark, reflecting their relatively structured and consistently reported nature. Even so, inspection of the extracted data revealed a subset of implausible values. These errors primarily resulted from confusion between hydrothermal crystallization conditions and related experimental steps such as calcination, drying, or aging, which are often reported alongside synthesis details. To address this, domain knowledge regarding typical ZSM-5 synthesis windows was used to flag unusually high temperatures or anomalous durations. The corresponding articles were then selectively reprocessed to recover the correct hydrothermal conditions. This intervention led to modest but consistent improvements, with the F_1 score increasing from 0.94 to 0.95 for temperature and from 0.92 to 0.93 for synthesis time.

The largest improvement occurred in the *Silica source* field, where the F_1 score increased from 0.73 to 0.83. Several systematic issues were identified in the initial extraction. In particular, in heteroatom-substituted ZSM-5 syntheses involving Ga, Fe, B, or Zn, the corresponding metal precursors were frequently misclassified as Si sources. Additional errors arose when *Silica source* was interchanged with *Aluminum source* or, in rarer cases, with OSDAs, especially when these species were discussed in close proximity within the text. A smaller subset of errors stemmed from hallucinated outputs, in which the model proposed chemically unreasonable or nonexistent precursors. These issues were corrected through element-based validation using molecular formulas retrieved using the PubChemPy library [49]. Entries labeled as *Silica source* were required to contain Si atoms in their molecular formula, and compounds failing this criterion were flagged as incorrect. The incorrectly labeled articles were then reprocessed to refine classification. This chemistry-aware domain knowledge validation eliminated the majority of heteroatom-related and cross-field misassignments, accounting for the largest performance gain observed among all evaluated categories.

Table 3
Benchmark scores before and after post-processing.

Category	Precision		Recall		F_1 score	
	Before	After	Before	After	Before	After
Synthesis temperature	0.95	0.97	0.93	0.94	0.94	0.95
Synthesis time	0.93	0.94	0.90	0.92	0.92	0.93
Silica source	0.76	0.86	0.69	0.81	0.73	0.83
Aluminum source	0.85	0.92	0.83	0.86	0.84	0.89
OSDA	0.95	0.96	0.88	0.92	0.91	0.94

Similar misclassification patterns were observed for the *Aluminum source* field, though with a smaller initial error rate. As with *Silica source*, *Aluminum source* was validated using molecular formulas obtained from PubChemPy. Only compounds containing Al atoms were retained as a valid *Aluminum source*, and misclassified entries were removed or corrected through reprocessing of the original texts. This intervention strongly benefited this category, improving from an F_1 score of 0.84 to 0.89, demonstrating the effectiveness of element-specific validation for framework-forming precursors.

The *OSDA* category exhibited high accuracy in the initial benchmark, but a limited number of misclassifications remained. These errors arose primarily from misclassification, like inorganic precursors or heteroatom-containing compounds being incorrectly labeled as OSDAs when multiple reagents were described in close proximity within the synthesis procedure. To identify such cases, molecular formulas were retrieved, and entries were flagged as incorrect when they contained framework-forming elements (e.g., Si or Al) or heteroatom metals (e.g., Fe, Ga, or B). The affected articles were then reprocessed to recover the correct entry, and this post-processing increased the F_1 score from 0.91 to 0.94.

The improvements observed across precursor-related fields highlight a broader limitation of LLM-based extraction when applied to complex chemical systems. This is particularly challenging in zeolite synthesis articles, where auxiliary reagents, dopants, and post-treatment chemicals appear alongside the true gel components. Without domain-specific constraints, models may assign functional roles prematurely or incorrectly. Incorporating chemistry-aware validation rules grounded in synthesis knowledge proved essential for filtering out such errors and refining the extracted dataset. These observations also suggest that further gains could be achieved by first extracting all chemical entities into a single aggregated field and subsequently re-categorizing them using chemoinformatic tools. Providing a more complete chemical context before role assignment may further reduce misclassification and represents a promising direction for future workflow refinement.

3.3. Patterns in synthesis parameters

To complement the extraction and benchmarking analyses, the distribution of key synthesis parameters was examined across the curated

ZSM-5 dataset. While this study does not attempt to derive mechanistic models of zeolite crystallization, statistical trends in synthesis temperature, synthesis time, precursor selection, and synthesis gel compositional ratios provide meaningful insight into commonly adopted experimental practices within the literature. By analyzing these patterns, it becomes possible to identify standard synthesis windows, preferred reagent choices, and composition ranges that have been repeatedly validated across independent research efforts.

3.3.1. Temperature and time distributions

Fig. 6 presents the distribution of *Synthesis temperature* and *Synthesis time* reported across the collected ZSM-5 literature. The histogram of the *Synthesis temperature* shows a pronounced concentration in the range of approximately 420–480 K (150–210 °C), with the mean centered near 450 K. This temperature interval corresponds to the widely recognized hydrothermal crystallization window for ZSM-5, where silicate and aluminate species possess sufficient mobility to reorganize into the MFI framework without promoting competing phase formation [50]. At temperatures below this regime, nucleation proceeds slowly, often yielding amorphous aluminosilicate gels or partially ordered intermediate phases. Conversely, temperatures significantly above 500 K increase the likelihood of dense silica phases, MOR-type impurities, or structural collapse during growth due to accelerated dissolution–reprecipitation dynamics [51,52]. Thus, the clustering of reported temperatures around 450 K is consistent with a balance between promoting effective crystal growth and suppressing undesired topological transformations.

The histogram of the *Synthesis time* also exhibits a well-defined mode, with most reported crystallization times lying between 24–48 h. These durations reflect the characteristic nucleation–growth kinetics of ZSM-5 formation under hydrothermal conditions [53]. Shorter times (< 12 h) are typically only successful in the presence of seeding or microwave-assisted heating, where nucleation barriers are substantially reduced. Longer crystallization periods (>72 h), while occasionally beneficial for developing hierarchical porosity or larger crystal domains, can also promote secondary grain growth or Ostwald ripening [51]. The statistical prominence of the 24–48 h range therefore represents a practical convergence across the field: sufficiently long to allow full MFI network assembly, but not so long as to induce overgrowth or structural densification.

Taken together, the temperature and time distributions observed in this dataset reinforce a consistent synthetic consensus across decades of ZSM-5 research: hydrothermal crystallization is most robust when carried out near 450 K for at least 24 h, conditions that reliably enable formation of the MFI framework while minimizing impurity formation and morphological instability.

3.3.2. Si, Al, and OSDA source frequencies

Fig. 7A illustrates the distribution of *Silica source* used in the reported ZSM-5 syntheses. A clear trend emerges: TEOS is the most widely used *Silica source*, followed by sodium silicate, while all remaining precursors with individual counts below a threshold of 10 were grouped into the "Other" category for clarity. The dominance of TEOS reflects its well-known role in producing homogeneous synthesis gels with tunable hydrolysis and condensation rates. Because TEOS hydrolyzes to monomeric silanol species while releasing ethanol, it enables controlled silicate speciation, which in turn promotes the assembly of the MFI framework in the presence of OSDAs such as tetrapropylammonium (TPA^+) [54]. Use of TEOS also avoids the introduction of additional alkali cations, which are known to affect framework selection and may shift crystallization toward competing phases under strongly basic conditions [55].

In contrast, sodium silicate remains the second most common Si source primarily for readily available silicate anions in alkaline media, accelerating nucleation and shortening crystallization times which is practical for scale-up and lowering production costs [53]. When templated with TPAOH or related quaternary ammonium cations, sodium silicate-based gels can yield highly crystalline MFI provided that the Na^+ concentration is balanced to suppress undesired phases such as MOR or dense silicates [56].

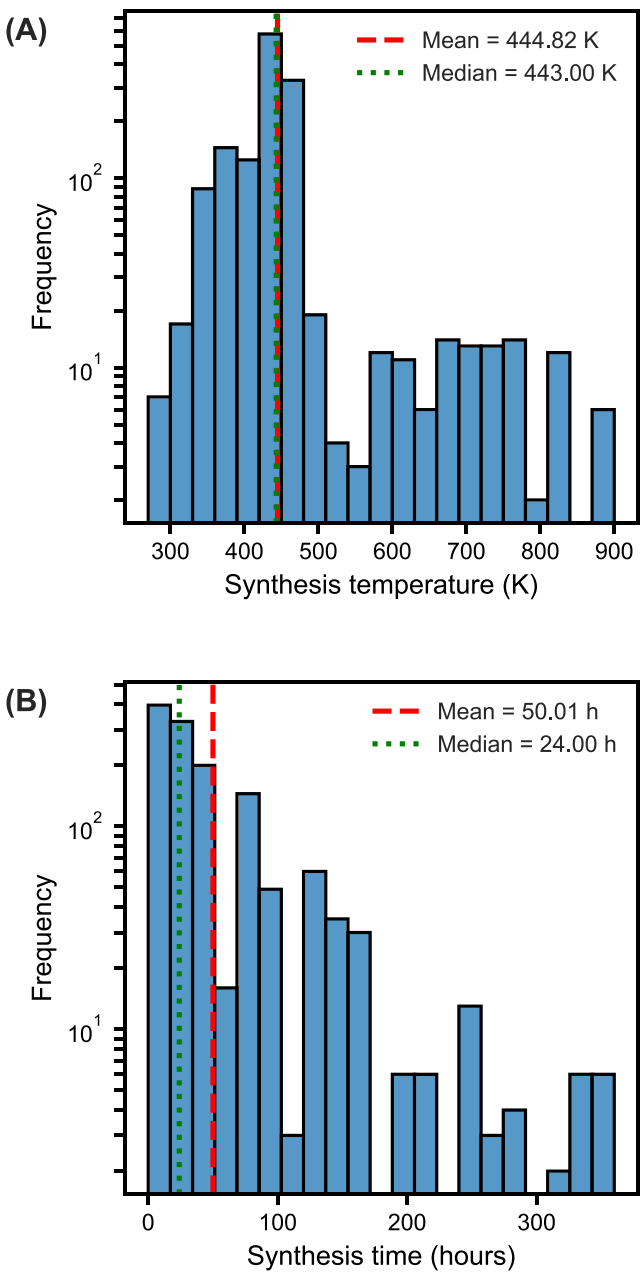

Fig. 6. Histogram of the ZSM-5 *Synthesis temperature* (A) and *Synthesis time* (B). Dashed red and dotted green lines denote mean and median values, respectively.

The dataset shows that sodium aluminate ($NaAlO_2$), aluminum isopropoxide (AIP), and aluminum sulfate ($Al_2(SO_4)_3$) are the most frequently used as the *Aluminum source* in ZSM-5 synthesis, together accounting for the majority of reported formulations. All remaining compounds with a frequency below five occurrences were grouped into an "Other" category (Fig. 7B). $NaAlO_2$ appears most frequently in the corpus. Under the alkaline conditions typical of ZSM-5 synthesis gels, $NaAlO_2$ dissolves readily to generate tetrahydroxoaluminate species that mix homogeneously with silicate oligomers and promote direct

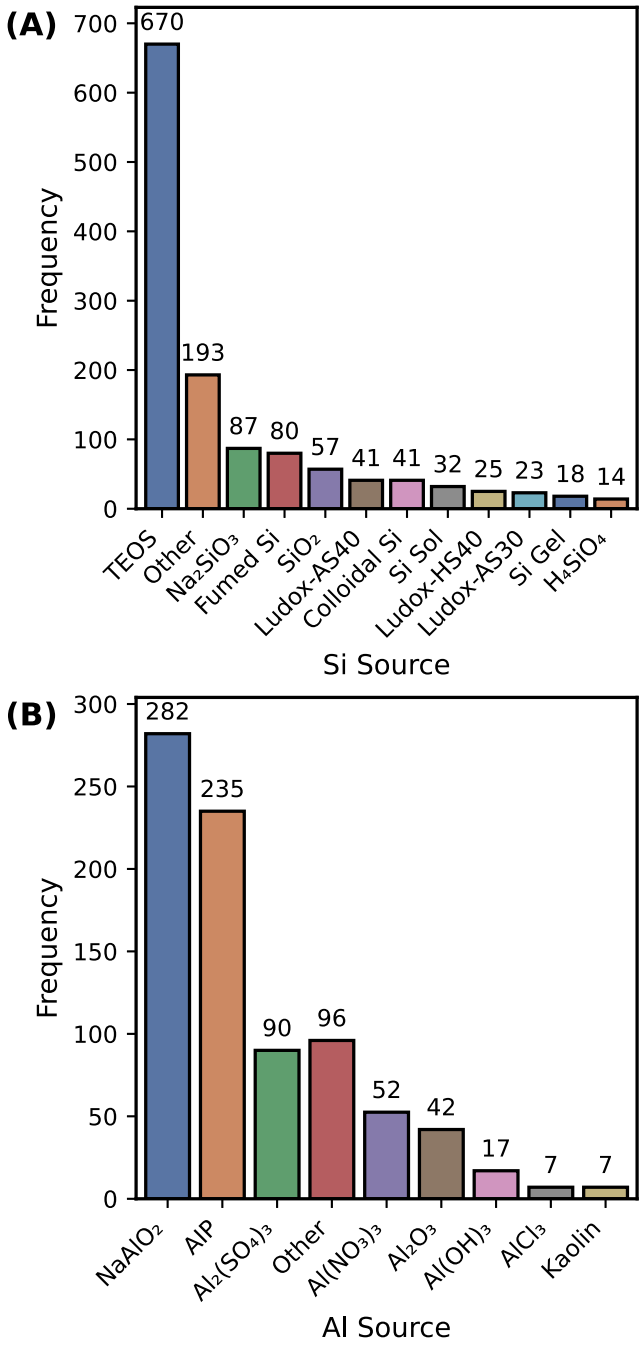

Fig. 7. Reported precursor distributions in ZSM-5 synthesis protocols. (A) Frequency of Si sources and (B) Al sources compiled from the literature dataset. Major contributors are shown individually, while low-frequency precursors were combined into an "Other" category for clarity. Short labels denote chemical shorthand: TEOS = tetraethyl orthosilicate; Na_2SiO_3 = sodium silicate; H_4SiO_4 = silicic acid; AIP = aluminum isopropoxide; $Al_2(SO_4)_3$ = aluminum sulfate; $NaAlO_2$ = sodium aluminate; $Al(NO_3)_3$ = aluminum nitrate; Al_2O_3 = aluminum oxide; $Al(OH)_3$ = aluminum hydroxide; $AlCl_3$ = aluminum chloride.

incorporation of tetrahedrally coordinated Al(IV) into the framework. This facilitates the formation of Brønsted acid sites upon ammonium exchange and calcination, and reliably yields the characteristic MFI powder X-ray diffraction (XRD) pattern [57,58].

AIP is frequently chosen for reduced-alkalinity and enabling faster crystallization and enhanced aluminum framework incorporation [59]. Controlled hydrolysis of AIP allows gradual release of Al species, improving uniformity of Al distribution in high-silica ZSM-5 and enabling reproducible acidity and catalytic performance [60,61]. This makes AIP particularly common in formulations targeting shape-selective catalysts.

Although $Al_2(SO_4)_3$ requires careful pH adjustment, it is valued for high solubility and scalability, especially in combination with aqueous sodium silicate systems [62]. When the synthesis gel composition and pH are properly adjusted, $Al_2(SO_4)_3$ can yield phase-pure ZSM-5 with crystallinity and acid site properties comparable to those obtained using sodium aluminate as the Al source [63,64].

The *OSDA* frequency distribution (Fig. 8) demonstrates a strong predominance of TPA^+, most commonly supplied in the hydroxide (TPAOH) or bromide (TPABr) form. Entries with frequency below 5 are grouped under Other. The dominance of TPA^+ aligns with its well-established role as the structure-directing agent for the MFI framework, where the molecular size, shape, and charge distribution of TPA^+ closely match the intersectional voids of the ZSM-5 channel system (straight and sinusoidal 10-membered-ring pores). This host–guest geometric compatibility stabilizes early-stage nuclei and guides the crystallization pathway toward ZSM-5 rather than competing topologies [65,66].

TPAOH is frequently employed in ZSM-5 synthesis because it simultaneously fulfills the roles of OSDA and alkaline mineralizer, delivering both TPA^+ cations for framework direction and OH^- ions to facilitate silica depolymerization and Al incorporation. This dual functionality streamlines the synthesis process, resulting in faster nucleation, enhanced crystallinity, and improved reproducibility by limiting variability in synthesis gel composition [67–69]. In contrast, TPABr supplies only the templating cation; thus, an external base such as NaOH must be added to achieve the requisite alkalinity [70,71]. TPABr-based formulations are popular, especially in industrial contexts, due to their lower reagent costs and the flexibility to independently tailor OSDA and pH levels [72]. In organic-free syntheses where charge balance is provided by inorganic cations (e.g. Na^+), these species were captured only as part of the reported synthesis compositions, without any explicit inorganic structure-directing agent (ISDA) assignment.

3.3.3. Molar composition ratios

The violin plots in Fig. 9 summarize the distributions of four key synthesis ratios, namely Si/Al, Si/Water, Si/OSDA, and Si/ISDA, as extracted from the compiled ZSM-5 literature dataset. The Si/Al ratio is most frequently reported within the range of 20 to 80, while significantly higher values appear only sporadically. This ratio is a fundamental descriptor of the ZSM-5 framework because it directly influences Brønsted acidity, catalytic behavior, hydrothermal stability, and the likelihood of forming phase-pure MFI crystals. Proper adjustment and homogeneous distribution of framework Al are essential, as inappropriate Si/Al ratios or spatially heterogeneous Al distributions within the framework can promote extra-framework alumina or amorphous byproducts [7,73].

The Si/Al ratio can influence crystallization behavior, phase purity, and product properties in ZSM-5 synthesis, particularly under specific gel compositions and synthesis conditions. But this relationship should not be interpreted as a universal decrease in crystallinity at high Si/Al ratios, since high-silica and fully siliceous MFI materials, such as silicalite-1, are well-established [74]. Therefore, the effect of Si/Al ratio on crystallinity is synthesis-system dependent and must be interpreted together with other variables such as alkalinity, OSDA content, water content, temperature, and synthesis time [32]. Higher Si/Al ratios decrease Brønsted acidity but often improve hydrothermal stability and in contrast, lower Si/Al enhances acid site density but can lead to increased framework defects and coke deposition during catalysis [73,75,76]. These effects are particularly pronounced in the synthesis of nanosized ZSM-5 at lower temperatures, where incomplete

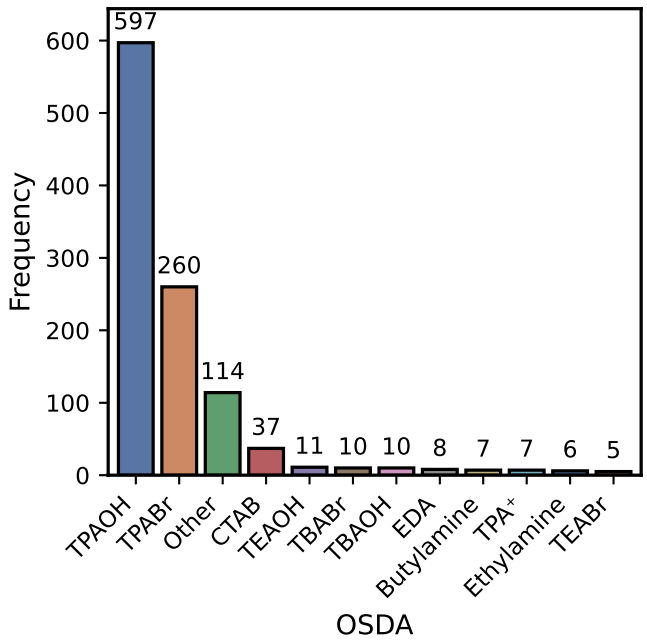

Fig. 8. Frequency distribution of OSDAs used in ZSM-5 synthesis reported in the literature. Low-frequency templates were grouped under "Other". Abbreviations: TPA = tetrapropylammonium, TEA = tetraethylammonium, TBA = tetrabutylammonium; OH = hydroxide, Br = bromide; CTAB = cetyltrimethylammonium bromide; EDA = ethylenediamine.

or non-uniform incorporation of Al into the framework can promote defect formation and extra-framework species [18].

The Si/Water ratio shows a relatively narrow, skewed distribution, highlighting the importance of water content in regulating gel viscosity and nucleation dynamics. Excess water dilutes reactive species (i.e. reduces supersaturation) and delays condensation, whereas too little water limits precursor mobility and leads to heterogeneous gels [77].

For the Si/OSDA ratio, a broader tail toward higher values is observed, reflecting variability in the amount of organic template relative to silica. TPA^+-based OSDAs (mainly TPAOH or TPABr) dominate this dataset because of their strong geometrical and electrostatic compatibility with the MFI framework. Moderate OSDA concentrations facilitate templated nucleation while minimizing excess organics and by-product removal challenges. At the same time, an excessively low OSDA content can weaken structural direction and lead to mixed phases [66,78]. The Si/ISDA ratio exhibits the broadest spread, reflecting the range of alkali concentrations used as inorganic structure-directing agents (typically Na^+ or K^+) [66,79]. A relatively low alkali content stabilizes the MFI topology by preventing excessive silica dissolution, whereas overly high alkalinity can shift crystallization toward competing frameworks such as zeolite MOR or LTA [80,81].

3.4. Limitations and outlook

While the automated extraction pipeline demonstrates competitive benchmark performance, several conceptual and technical limitations highlight the challenges of deploying general-purpose LLMs for domain-specific scientific extraction. LLMs remain imperfect tools for structured data recovery, and reliability issues were observed in this work despite the use of a custom output parser to constrain generation. The base Mistral 7B model occasionally produced hallucinated outputs, such as fabricated entries for missing quantitative data, and struggled with persistent process awareness, leading to the occasional mixing of unrelated statements or the misassignment of Si and Al precursors. Even with a custom JSON parser to validate and sanitize responses, outputs often contained misplaced keys or irrelevant structures, increasing the burden of downstream cleaning. Although the pipeline utilized a base model without fine-tuning to avoid the high computational costs associated with domain-specific training, the results indicate that smaller, open-source models can remain competitive when supported by robust prompt engineering. While larger frontier models such as the GPT-4, Gemini, Claude, DeepSeek, and Qwen series have demonstrated higher accuracy in related materials extraction tasks [82], our findings suggest that structured and modular inference may be better addressed by emerging lightweight reasoning architectures. Specifically, frameworks such as the Hierarchical Reasoning Model (HRM) [83] and Recursive Reasoning with Tiny Networks [84] represent promising alternatives for synthesis pathway interpretation and context-aware extraction due to their emphasis on modular inference with significantly fewer parameters.

Challenges in automated extraction arise from both the fragmented reporting of synthesis protocols and the inherent semantic ambiguity of scientific text. A significant structural challenge is that experimental procedures are frequently relegated to supplementary information files or external references (e.g., "followed from Ref. [X]"), which remain inaccessible to extraction pipelines limited to the main article body. Consequently, synthesis conditions documented outside the primary manuscript are systematically missed, introducing unavoidable gaps that single-article workflows cannot resolve. Furthermore, the dataset is affected by context-dependent language that leads to occasional misclassification; for example, a phrase such as "treated at 500 °C for 6 h" may describe hydrothermal crystallization, calcination, or drying depending on the specific synthesis step. Similarly, temporal descriptions like "aged overnight" and size-related terms such as "under 200 nm" often lack the specificity or clear attribution required for consistent interpretation, challenging models that lack a deep procedural understanding of synthesis workflows and resulting in periodic errors in temperature, time, and crystal size parameters.

One limitation of the present dataset is that it does not assign new experimental uncertainties to extracted values. Uncertainty in this dataset can arise from two sources: the uncertainty of the experimental measurements reported in the original papers and the uncertainty associated with extracting those values using an LLM. For experimental uncertainty, the present workflow does not assign new error bars or standard deviations when they are not reported by the original authors. Instead, values are retained as reported in the source text. When a paper reports a range, standard deviation, or error bar for a quantity such as crystal size or synthesis time, that information is preserved in the extracted field whenever possible. When no experimental uncertainty is reported, the corresponding extracted value should be interpreted as a literature-reported value without an associated measurement uncertainty. Extraction uncertainty was evaluated separately by comparing the LLM-generated outputs against manually curated labels and by analyzing field-specific error types. These validation metrics indicate how reliably the workflow identifies and extracts a given field. Future versions of the dataset could include separate metadata fields for reported experimental uncertainty. Such metadata would improve compatibility with FAIR uncertainty-aware data practices and would allow downstream machine learning models to distinguish between measurement uncertainty and extraction uncertainty [85].

3.5. Future directions

To overcome these limitations and challenges and scale this approach to broader applications, several future extensions are envisioned. One direction involves incorporating retrieval-augmented generation (RAG) into the extraction pipeline. By incorporating document chunking, embedding-based indexing, and similarity-driven retrieval into the inference pipeline, RAG-based architectures enable model outputs to be conditioned on semantically relevant synthesis passages,

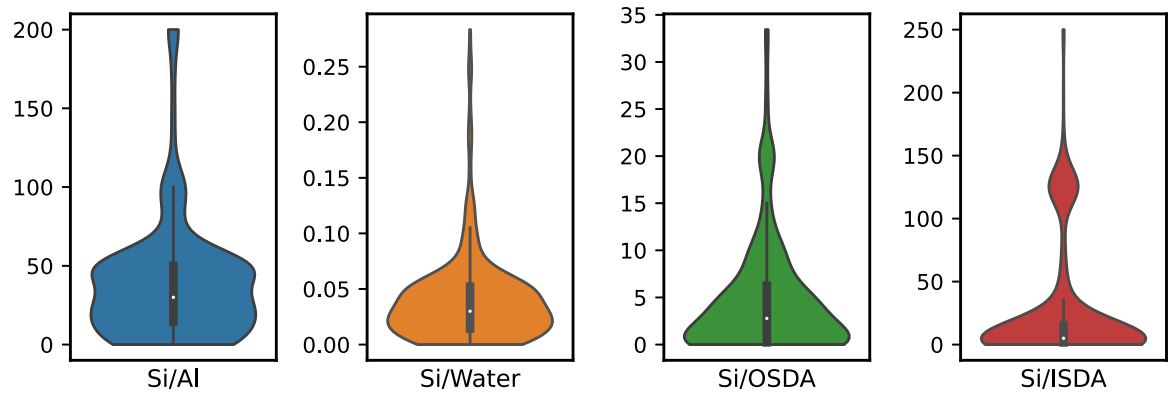

Fig. 9. Distribution of key molar composition ratios reported for ZSM-5 synthesis across the curated dataset. Violin plots summarize the empirical distributions of Si/Al, Si/Water, Si/OSDA, and Si/ISDA ratios, with embedded box plots indicating median values and interquartile ranges.

thereby reducing hallucinated outputs and improving extraction reliability compared with purely generative models. In contrast to model fine-tuning, this strategy requires relatively modest labeled data and computational infrastructure, and has already demonstrated strong performance in domain-specific tasks [48]. Also, future workflows could combine text-based extraction with multimodal figure analysis or vision-language models to extract morphology, faceting, and particle-size information directly from microscopy images, rather than relying only on text and captions.

As extraction accuracy is improved, synthesis conditions across papers can be projected into low-dimensional latent spaces using PCA, t-SNE, or UMAP. This would enable the identification of synthesis "clusters" or sparsely explored parameter regions which might be useful to design new synthesis regimes or optimize existing ones for ZSM-5 crystallization. While this study focused on ZSM-5 synthesis, the pipeline could be readily extensible to CHA, AEI, FAU, and other zeolite topologies. A comprehensive synthesis dataset across frameworks would enable new scientific questions to be addressed, such as the effect of SDA combinations on framework selection, the mapping of synthesis regimes to pore topology, or even the prediction of optimal structure-directing agents for catalytic applications.

4. Conclusion

We presented an automated, fully open-source workflow for extracting structured materials synthesis information from the scientific literature using large language models. The approach integrated large-scale document processing, prompt-guided extraction with a compact open-source LLM, and chemistry-aware post-processing to convert unstructured experimental descriptions into a machine-readable synthesis dataset. Application to the hydrothermal synthesis of zeolite ZSM-5 showed that the workflow could accurately recover several key synthesis parameters, including crystallization temperature, synthesis time, organic structure-directing agent identity, and overall framework composition (e.g., Si/Al ratios). These parameters exhibited good agreement with manually curated data, reflecting their relatively standardized reporting across the ZSM-5 literature. In contrast, parameters such as precursor roles, auxiliary reagents, and secondary additives showed lower raw extraction accuracy and required domain-informed post-processing to resolve common ambiguities. Statistical analysis of the curated dataset identified synthesis trends consistent with established experimental understanding of ZSM-5 crystallization, including commonly reported temperature and time windows and composition ranges.

Although these findings demonstrate the feasibility of automated synthesis data extraction, several limitations of the approach were identified. Important experimental details were often described incompletely or indirectly, with key parameters omitted, distributed across supplementary materials, or referenced from prior studies, which limited the completeness and reliability of extracted synthesis records. For ZSM-5 in particular, ambiguity in precursor nomenclature and overlapping chemical roles hindered consistent identification of framework sources, mineralizers, and structure-directing agents. In addition, procedural details that were frequently reported qualitatively or as narratives (e.g. aging protocols, mixing sequences, and pH adjustment steps) were difficult to capture in a fully standardized form. We expect that similar challenges will be faced when extending the approach to other zeolite materials. Nonetheless, these challenges point to clear opportunities for improving automated synthesis extraction through more systematic handling of incomplete descriptions and closer integration of domain-specific chemical context. Advanced data-driven strategies are essential for the rational design of materials, including the development of next-generation catalysts to meet the growing demand for diversifying chemical, energy, and fuel sources. We anticipate that further refinement of our LLM framework to address some of these challenges will enable analysis of more complex zeolite systems and ultimately the development of new synthesis strategies to design zeolite materials with specific, targeted properties tailored for different applications.

CRediT authorship contribution statement

Tanvir Ahmed: Writing – original draft, Methodology, Formal analysis, Data curation. **Jeffrey D. Rimer:** Writing – review & editing, Funding acquisition, Conceptualization. **Mingjian Wen:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Formal analysis, Conceptualization. **Jeremy C. Palmer:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was primarily funded by the Department of Energy, Office of Basic Energy Sciences, Materials Science Division (Award DE-SC0021384). JCP and JDR received additional financial support from the Welch Foundation, United States (Awards E-1882 and E-1794, respectively). For this special issue in honor of Prof. Dion Vlachos, JDR would like to express his profound appreciation for Dion's guidance and support over the years, both as a PhD advisor and colleague, which has played a pivotal role in his academic growth and development.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mcat.2026.116097.

Data availability

The dataset generated in this study is publicly available on Zenodo in JSON format at https://doi.org/10.5281/zenodo.18761174.

References

[1] M.E. Davis, Zeolites from a materials chemistry perspective, Chem. Mater. 26 (1) (2014) 239–245, http://dx.doi.org/10.1021/cm401914u, tex.eprint: https://doi.org/10.1021/cm401914u.

[2] A.J. Mallette, K. Shilpa, J.D. Rimer, The current understanding of mechanistic pathways in zeolite crystallization, Chem. Rev. 124 (6) (2024) 3416–3493, http://dx.doi.org/10.1021/acs.chemrev.3c00801.

[3] H. Zhang, I.B. Samsudin, S. Jaenicke, G.-K. Chuah, Zeolites in catalysis: sustainable synthesis and its impact on properties and applications, Catal. Sci. Technol. 12 (19) (2022) 6024–6039, http://dx.doi.org/10.1039/D2CY01325H.

[4] H. Xu, P. Wu, New progress in zeolite synthesis and catalysis, Natl. Sci. Rev. 9 (9) (2022) nwac045, http://dx.doi.org/10.1093/nsr/nwac045.

[5] E. Pérez-Botella, S. Valencia, F. Rey, Zeolites in adsorption processes: state of the art and future prospects, Chem. Rev. 122 (24) (2022) 17647–17695, http://dx.doi.org/10.1021/acs.chemrev.2c00140.

[6] P.J. Bereciartua, Á. Cantín, A. Corma, J.L. Jordá, M. Palomino, F. Rey, S. Valencia, E.W. Corcoran Jr., P. Kortunov, P.I. Ravikovitch, et al., Control of zeolite framework flexibility and pore topology for separation of ethane and ethylene, Science 358 (6366) (2017) 1068–1071, http://dx.doi.org/10.1126/science.aao0092.

[7] J. Li, M. Gao, W. Yan, J. Yu, Regulation of the Si/Al ratios and Al distributions of zeolites and their impact on properties, Chem. Sci. 14 (8) (2023) 1935–1959, http://dx.doi.org/10.1039/D2SC06010H.

[8] R. Jain, J.D. Rimer, Seed-assisted zeolite synthesis: the impact of seeding conditions and interzeolite transformations on crystal structure and morphology, Microporous Mesoporous Mater. 300 (2020) 110174, http://dx.doi.org/10.1016/j.micromeso.2020.110174.

[9] M.K. Choudhary, R. Jain, J.D. Rimer, In situ imaging of two-dimensional surface growth reveals the prevalence and role of defects in zeolite crystallization, Proc. Natl. Acad. Sci. 117 (46) (2020) 28632–28639, http://dx.doi.org/10.1073/pnas.2011806117.

[10] A.J. Mallette, S. Hong, E.E. Freeman, S.A. Saslow, S. Mergelsberg, R.K. Motkuri, J.J. Neeway, G. Mpourmpakis, J.D. Rimer, Heteroatom manipulation of zeolite crystallization: stabilizing Zn-FAU against interzeolite transformation, JACS Au 2 (10) (2022) 2295–2306, http://dx.doi.org/10.1021/jacsau.2c00325.

[11] K. Muraoka, Y. Sada, D. Miyazaki, W. Chaikittisilp, T. Okubo, Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials, Nat. Commun. 10 (1) (2019) 4459, http://dx.doi.org/10.1038/s41467-019-12394-0.

[12] M. Kumar, Z.J. Berkson, R.J. Clark, Y. Shen, N.A. Prisco, Q. Zheng, Z. Zeng, H. Zheng, L.B. McCusker, J.C. Palmer, B.F. Chmelka, J.D. Rimer, Crystallization of mordenite platelets using cooperative organic structure-directing agents, J. Am. Chem. Soc. 141 (51) (2019) 20155–20165, http://dx.doi.org/10.1021/jacs.9b09697.

[13] K.N. Bozhilov, T.T. Le, Z. Qin, T. Terlier, A. Palčić, J.D. Rimer, V. Valtchev, Time-resolved dissolution elucidates the mechanism of zeolite MFI crystallization, Sci. Adv. 7 (25) (2021) eabg0454, http://dx.doi.org/10.1126/sciadv.abg0454.

[14] D. Schwalbe-Koda, S. Kwon, C. Paris, E. Bello-Jurado, Z. Jensen, E. Olivetti, T. Willhammar, A. Corma, Y. Román-Leshkov, M. Moliner, R. Gómez-Bombarelli, A priori control of zeolite phase competition and intergrowth with high-throughput simulations, Science 374 (6565) (2021) 308–315, http://dx.doi.org/10.1126/science.abh3350.

[15] M. Moliner, Y. Román-Leshkov, A. Corma, Machine learning applied to zeolite synthesis: the missing link for realizing high-throughput discovery, Acc. Chem. Res. 52 (10) (2019) 2971–2980, http://dx.doi.org/10.1021/acs.accounts.9b00399.

[16] H. Thirumalai, J.D. Rimer, L.C. Grabow, Quantification and statistical analysis of errors related to the approximate description of active site models in metal-exchanged zeolites, ChemCatChem 11 (20) (2019) 5055–5067, http://dx.doi.org/10.1002/cctc.201901229.

[17] R. Millini, Beyond trial and error for zeolite catalysts, Science 355 (6329) (2017) 1028, http://dx.doi.org/10.1126/science.aam8037.

[18] W. Qin, Y. Zhou, J.D. Rimer, Deleterious effects of non-framework Al species on the catalytic performance of ZSM-5 crystals synthesized at low temperature, React. Chem. Eng. 4 (11) (2019) 1957–1968, http://dx.doi.org/10.1039/C9RE00231F.

[19] K. Shilpa, H. Dai, P. Lu, X. Ye, C. Rieg, S. Luo, H. Chen, O. Abdelrahman, W. Fan, B.M. Weckhuysen, et al., Hydrothermal annealing of hierarchical ZSM-5 zeolites improves catalytic performance, Angew. Chem. (2026) e15008.

[20] A.J. Mallette, S. Seo, J.D. Rimer, Synthesis strategies and design principles for nanosized and hierarchical zeolites, Nat. Synth. 1 (7) (2022) 521–534, http://dx.doi.org/10.1038/s44160-022-00091-8.

[21] T.T. Le, W. Qin, A. Agarwal, N. Nikolopoulos, D. Fu, M.D. Patton, C. Weiland, S.R. Bare, J.C. Palmer, B.M. Weckhuysen, J.D. Rimer, Elemental zoning enhances mass transport in zeolite catalysts for methanol to hydrocarbons, Nat. Catal. 6 (3) (2023) 254–265, http://dx.doi.org/10.1038/s41929-023-00927-2.

[22] L. Himanen, A. Geurts, A.S. Foster, P. Rinke, Data-driven materials science: status, challenges, and perspectives, Adv. Sci. 6 (21) (2019) 1900808.

[23] R. Pollice, G. Dos Passos Gomes, M. Aldeghi, R.J. Hickman, M. Krenn, C. Lavigne, M. Lindner-D'Addario, A. Nigam, C.T. Ser, Z. Yao, A. Aspuru-Guzik, Data-driven strategies for accelerated materials design, Acc. Chem. Res. 54 (4) (2021) 849–860, http://dx.doi.org/10.1021/acs.accounts.0c00785.

[24] A. Mannodi-Kanakkithodi, M.K. Chan, Computational data-driven materials discovery, Trends Chem. 3 (2) (2021) 79–82, http://dx.doi.org/10.1016/j.trechm.2020.12.007.

[25] L. Hawizy, D.M. Jessop, N. Adams, P. Murray-Rust, ChemicalTagger: a tool for semantic text-mining in chemistry, J. Cheminf. 3 (1) (2011) 17, http://dx.doi.org/10.1186/1758-2946-3-17.

[26] D.M. Jessop, S.E. Adams, E.L. Willighagen, L. Hawizy, P. Murray-Rust, OSCAR4: a flexible architecture for chemical text-mining, J. Cheminf. 3 (1) (2011) 41, http://dx.doi.org/10.1186/1758-2946-3-41.

[27] M.C. Swain, J.M. Cole, ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature, J. Chem. Inf. Model. 56 (10) (2016) 1894–1904, http://dx.doi.org/10.1021/acs.jcim.6b00207.

[28] D.M. Lowe, Extraction of Chemical Structures and Reactions from the Literature (Doctoral dissertation), University of Cambridge, 2012.

[29] J. He, D.Q. Nguyen, S.A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa, A. Albahem, L. Cavedon, T. Cohn, T. Baldwin, K. Verspoor, ChEMU 2020: natural language processing methods are effective for information extraction from chemical patents, Front. Res. Metr. Anal. 6 (2021) 654438, http://dx.doi.org/10.3389/frma.2021.654438.

[30] Z. Jensen, E. Kim, S. Kwon, T.Z.H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, A machine learning approach to zeolite synthesis enabled by automatic literature data extraction, ACS Cent. Sci. 5 (5) (2019) 892–899, http://dx.doi.org/10.1021/acscentsci.9b00193.

[31] E. Pan, S. Kwon, Z. Jensen, M. Xie, R. Gómez-Bombarelli, M. Moliner, Y. Román-Leshkov, E. Olivetti, ZeoSyn: a comprehensive zeolite synthesis dataset enabling machine-learning rationalization of hydrothermal parameters, ACS Cent. Sci. 10 (3) (2024) 729–743, http://dx.doi.org/10.1021/acscentsci.3c01615.

[32] Z. Jensen, S. Kwon, D. Schwalbe-Koda, C. Paris, R. Gómez-Bombarelli, Y. Román-Leshkov, A. Corma, M. Moliner, E.A. Olivetti, Discovering relationships between OSDAs and zeolites through data mining and generative neural networks, ACS Cent. Sci. 7 (5) (2021) 858–867, http://dx.doi.org/10.1021/acscentsci.1c00024.

[33] T. Gupta, M. Zaki, N.M.A. Krishnan, Mausam, MatSciBERT: a materials domain language model for text mining and information extraction, Npj Comput. Mater. 8 (1) (2022) 102, http://dx.doi.org/10.1038/s41524-022-00784-w.

[34] M.P. Polak, D. Morgan, Extracting accurate materials data from research papers with conversational language models and prompt engineering, Nat. Commun. 15 (1) (2024) 1569, http://dx.doi.org/10.1038/s41467-024-45914-8.

[35] J. Dagdelen, A. Dunn, S. Lee, N. Walker, A.S. Rosen, G. Ceder, K.A. Persson, A. Jain, Structured information extraction from scientific text with large language models, Nat. Commun. 15 (1) (2024) 1418, http://dx.doi.org/10.1038/s41467-024-45563-x.

[36] S. He, W. Du, X. Peng, X. Li, ZeoReader: automated extraction of synthesis steps from zeolite synthesis literature for autonomous experiments, Chem. Eng. Sci. 302 (2025) 120916, http://dx.doi.org/10.1016/j.ces.2024.120916.

[37] C.P. Rathore, S. Ray, D. Kumar, Evaluating llms for zeolite synthesis event extraction (ZSEE): a systematic analysis of prompting strategies, 2025, arXiv Prepr. arXiv:2512, 15312.

[38] Z. Yu, S. He, W. Du, A framework for automated extraction and validation of experimental arguments in zeolite synthesis, Comput. Mater. Sci. 264 (2026) 114495, http://dx.doi.org/10.1016/j.commatsci.2026.114495.

[39] S. Ito, K. Muraoka, A. Nakayama, Knowledge-informed molecular design for zeolite synthesis using general-purpose pretrained large language models toward human-machine collaboration, Chem. Mater. 37 (7) (2025) 2447–2456, http://dx.doi.org/10.1021/acs.chemmater.4c02726.

[40] Z. Hu, Y. Zhou, Z. Wang, X. Li, W. Yang, H. Fan, Y. Yang, OSDA agent: leveraging large language models for de novo design of organic structure directing agents, in: The Thirteenth International Conference on Learning Representations, 2025.

[41] OpenDataLab, MinerU: a toolkit for converting pdfs into machine-readable formats, 2025, URL https://github.com/opendatalab/MinerU. (Accessed 24 September 2025).

[42] Mistral Ai, Mistral-7B-instruct-v0.2, 2025, URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2. [Online; (Accessed 3 December 2025)].

[43] Pydantic Team, Pydantic, 2025, URL https://github.com/pydantic/pydantic. [Online; (Accessed 3 December 2025)].

[44] RapidFuzz Developers, RapidFuzz: rapid fuzzy string matching in python and C++ using the levenshtein distance, 2025, URL https://github.com/rapidfuzz/RapidFuzz. [Online; (Accessed 15 December 2025)].
[45] S. Gupta, A. Mahmood, P. Shetty, A. Adeboye, R. Ramprasad, Data extraction from polymer literature using large language models, Commun. Mater. 5 (1) (2024) 269, http://dx.doi.org/10.1038/s43246-024-00708-9.
[46] J. Choi, B. Lee, Accelerating materials language processing with large language models, Commun. Mater. 5 (1) (2024) 13, http://dx.doi.org/10.1038/s43246-024-00449-9.
[47] Z. Zheng, O. Zhang, C. Borgs, J.T. Chayes, O.M. Yaghi, ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis, J. Am. Chem. Soc. 145 (32) (2023) 18048–18062, http://dx.doi.org/10.1021/jacs.3c05819.
[48] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, P. Kuksa, A. Fan, M. Lewis, S. Bhosale, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, Vol. 33, 2020.
[49] PubChemPy Developers, PubChemPy, 2025, URL https://docs.pubchempy.org/en/latest/. [Online; (Accessed 15 December 2025)].
[50] B. Bensafi, N. Chouat, F. Djafri, The universal zeolite ZSM-5: structure and synthesis strategies. A review, Coord. Chem. Rev. 496 (2023) 215397, http://dx.doi.org/10.1016/j.ccr.2023.215397.
[51] G. Xiong, F. Meng, J. Liu, L. Liu, L. Zhao, Rapid hydrothermal synthesis of hierarchical ZSM-5/beta composite zeolites, RSC Adv. 11 (35) (2021) 21235–21247, http://dx.doi.org/10.1039/D1RA03064G.
[52] G. Liang, Y. Li, C. Yang, X. Hu, Q. Li, W. Zhao, Synthesis of ZSM-5 zeolites from biomass power plant ash for removal of ionic dyes from aqueous solution: equilibrium isotherm, kinetic and thermodynamic analysis, RSC Adv. 11 (36) (2021) 22365–22375, http://dx.doi.org/10.1039/D1RA03847H.
[53] Z. Sun, Q. Shu, Q. Zhang, S. Li, G. Zhu, C. Wang, J. Zhang, H. Li, Z. Huang, A hydrothermal synthesis process of ZSM-5 zeolite for VOCs adsorption using desilication solution, Separations 11 (2) (2024) 39, http://dx.doi.org/10.3390/separations11020039.
[54] C.E.A. Kirschhock, R. Ravishankar, F. Verspeurt, P.J. Grobet, P.A. Jacobs, J.A. Martens, Reply to the comment on "identification of precursor species in the formation of MFI zeolite in the TPAOH-TEOS-H_2 O system", J. Phys. Chem. B 106 (12) (2002) 3333–3334, http://dx.doi.org/10.1021/jp0146521.
[55] J. Mei, A. Duan, X. Wang, A brief review on solvent-free synthesis of zeolites, Materials 14 (4) (2021) 788, http://dx.doi.org/10.3390/ma14040788.
[56] R.M. Mohamed, H.M. Aly, M.F. El-Shahat, I.A. Ibrahim, Effect of the silica sources on the crystallinity of nanosized ZSM-5 zeolite, Microporous Mesoporous Mater. 79 (1–3) (2005) 7–12, http://dx.doi.org/10.1016/j.micromeso.2004.10.031.
[57] M. Xing, L. Zhang, J. Cao, Y. Han, F. Wang, K. Hao, L. Huang, Z. Tao, X. Wen, Y. Yang, Y. Li, Impact of the aluminum species state on Al pairs formation in the ZSM-5 framework, Microporous Mesoporous Mater. 334 (2022) 111769, http://dx.doi.org/10.1016/j.micromeso.2022.111769.
[58] A.A. Ismail, R.M. Mohamed, O.A. Fouad, I.A. Ibrahim, Synthesis of nanosized ZSM-5 using different alumina sources, Cryst. Res. Technol. 41 (2) (2006) 145–149, http://dx.doi.org/10.1002/crat.200510546.
[59] L. Bing, J. Liu, K. Yi, F. Li, D. Han, F. Wang, G. Wang, Rapid synthesis of hierarchical SSZ-13 zeolite microspheres *via* a fluoride-assisted *in situ* growth route using aluminum isopropoxide as aluminum source, RSC Adv. 10 (6) (2020) 3566–3571, http://dx.doi.org/10.1039/C9RA08895D.
[60] K. Saenluang, T. Imyen, W. Wannapakdee, D. Suttipat, P. Dugkhuntod, M. Ketkaew, A. Thivasasith, C. Wattanakit, Hierarchical nanospherical ZSM-5 nanosheets with uniform Al distribution for alkylation of benzene with ethanol, ACS Appl. Nano Mater. 3 (4) (2020) 3252–3263, http://dx.doi.org/10.1021/acsanm.9b02568.
[61] L.A. Gallego-Villada, J. Cueto, M.D.M. Alonso-Doncel, P. Mäki-Arvela, E.A. Alarcón, D.P. Serrano, D.Y. Murzin, Dendritic ZSM-5 zeolites as highly active catalysts for the valorization of monoterpene epoxides, Green Chem. 26 (20) (2024) 10512–10528, http://dx.doi.org/10.1039/D4GC04003A.
[62] C. Auepattana-aumrung, V. Márquez, S. Wannakao, B. Jongsomjit, J. Panpranot, P. Praserthdam, Role of Al in Na-ZSM-5 zeolite structure on catalyst stability in butene cracking reaction, Sci. Rep. 10 (1) (2020) 13643, http://dx.doi.org/10.1038/s41598-020-70568-z.
[63] Y.-J. Wang, J.-P. Cao, X.-Y. Ren, X.-B. Feng, X.-Y. Zhao, Y. Huang, X.-Y. Wei, Synthesis of ZSM-5 using different silicon and aluminum sources nature for catalytic conversion of lignite pyrolysis volatiles to light aromatics, Fuel 268 (2020) 117286, http://dx.doi.org/10.1016/j.fuel.2020.117286.
[64] Y. Zhou, H. Liu, X. Rao, Y. Yue, H. Zhu, X. Bao, Controlled synthesis of ZSM-5 zeolite with an unusual Al distribution in framework from natural aluminosilicate mineral, Microporous Mesoporous Mater. 305 (2020) 110357, http://dx.doi.org/10.1016/j.micromeso.2020.110357.
[65] S.L. Burkett, M.E. Davis, Mechanisms of structure direction in the synthesis of pure-silica zeolites. 1. Synthesis of TPA/Si-ZSM-5, Chem. Mater. 7 (5) (1995) 921–928, http://dx.doi.org/10.1021/cm00053a017.
[66] A. Chawla, R. Li, R. Jain, R.J. Clark, J.G. Sutjianto, J.C. Palmer, J.D. Rimer, Cooperative effects of inorganic and organic structure-directing agents in ZSM-5 crystallization, Mol. Syst. Des. Eng. 3 (1) (2018) 159–170, http://dx.doi.org/10.1039/C7ME00097A.
[67] M.A. Sanhoob, A. Khan, A.C. Ummer, ZSM-5 catalysts for MTO: effect and optimization of the tetrapropylammonium hydroxide concentration on synthesis and performance, ACS Omega 7 (25) (2022) 21654–21663, http://dx.doi.org/10.1021/acsomega.2c01539.
[68] X. Niu, Y. Bai, Y.-e. Du, H. Qi, Y. Chen, Size controllable synthesis of ZSM-5 zeolite and its catalytic performance in the reaction of methanol conversion to aromatics, R. Soc. Open Sci. 9 (3) (2022) 211284, http://dx.doi.org/10.1098/rsos.211284.
[69] Z. Liu, K. Okabe, C. Anand, Y. Yonezawa, J. Zhu, H. Yamada, A. Endo, Y. Yanaba, T. Yoshikawa, K. Ohara, T. Okubo, T. Wakihara, Continuous flow synthesis of ZSM-5 zeolite on the order of seconds, Proc. Natl. Acad. Sci. 113 (50) (2016) 14267–14271, http://dx.doi.org/10.1073/pnas.1615872113.
[70] M. Hamidzadeh, M. Nazari, S. Shifteh, A. Abdolali, Optimizing H-B-ZSM-5 catalyst for efficient methanol-to-propylene conversion via precise control of NaOH-TPABr interaction, Microporous Mesoporous Mater. 381 (2025) 113344, http://dx.doi.org/10.1016/j.micromeso.2024.113344.
[71] S. Chen, K. Huddersman, D. Keir, L. Rees, Synthesis of large uniform crystals of ZSM-5, Zeolites 8 (2) (1988) 106–109, http://dx.doi.org/10.1016/S0144-2449(88)80074-5.
[72] T. Ma, L. Zhang, Y. Song, Y. Shang, Y. Zhai, Y. Gong, A comparative synthesis of ZSM-5 with ethanol or TPABr template: distinction of brønsted/lewis acidity ratio and its impact on *n* -hexane cracking, Catal. Sci. Technol. 8 (7) (2018) 1923–1935, http://dx.doi.org/10.1039/C7CY02418E.
[73] R. Simancas, A. Chokkalingam, S.P. Elangovan, Z. Liu, T. Sano, K. Iyoki, T. Wakihara, T. Okubo, Recent progress in the improvement of hydrothermal stability of zeolites, Chem. Sci. 12 (22) (2021) 7677–7695, http://dx.doi.org/10.1039/D1SC01179K.
[74] H. Hazrati, M. Rostamizadeh, M.R. Omidkhah, Z. Sadeghian, Influence of synthesis and operating parameters on silicalite-1 membrane properties, C. R. Chim. 21 (1) (2017) 19–26, http://dx.doi.org/10.1016/j.crci.2017.11.008.
[75] H. Yuan, X. Han, Z. Sheng, J. Zhou, Z. Wang, Y. Ye, Y. Hu, W. Yang, A strategy for efficiently enhancing the hydrothermal stability of ZSM-5 zeolites: TMPO selective adsorption, J. Phys. Chem. C 129 (33) (2025) 14761–14772, http://dx.doi.org/10.1021/acs.jpcc.5c02340.
[76] N.-Y. Yao, J.-P. Cao, X.-Y. Zhao, X.-B. Pang, J.-P. Zhao, C.-X. Chen, S.-J. Cai, X.-B. Feng, D. Dung Le, Coke formation on different metal-modified (Co, Mo and Zr) ZSM-5 in the catalytic pyrolysis of cellulose to light aromatics, J. Anal. Appl. Pyrolysis 179 (2024) 106516, http://dx.doi.org/10.1016/j.jaap.2024.106516.
[77] K.A. Sashkina, Z. Qi, W. Wu, A.B. Ayupov, A.I. Lysikov, E.V. Parkhomchuk, The effect of H2O/SiO2 ratio in precursor solution on the crystal size and morphology of zeolite ZSM-5, Microporous Mesoporous Mater. 244 (2017) 93–100, http://dx.doi.org/10.1016/j.micromeso.2017.02.060.
[78] X. Tang, O.F. Altundal, F. Daeyaert, Z. Liu, G. Sastre, Computational insights on the role of structure-directing agents (SDAs) in the synthesis of zeolites, Chem. Soc. Rev. 54 (15) (2025) 7067–7092, http://dx.doi.org/10.1039/D5CS00306G.
[79] M.D. Oleksiak, J.D. Rimer, Synthesis of zeolites in the absence of organic structure-directing agents: factors governing crystal selection and polymorphism, Rev. Chem. Eng. 30 (1) (2014) 1–49, http://dx.doi.org/10.1515/revce-2013-0020.
[80] L. Wang, H. Pan, J. Qian, K. Yan, X. Yang, L. Liu, G. Song, H. Zhu, Dual organic-structure directing agents in the synthesis of high-silica MOR zeolite: cooperation or competition? Microporous Mesoporous Mater. 354 (2023) 112569, http://dx.doi.org/10.1016/j.micromeso.2023.112569.
[81] Y. Awoke, M. Sánchez-Sánchez, I. Arnaiz, I. Diaz, Synthesis of ZSM-5 from natural mordenite from Spain, Microporous Mesoporous Mater. 385 (2025) 113463, http://dx.doi.org/10.1016/j.micromeso.2024.113463.
[82] L. Foppiano, G. Lambard, T. Amagasa, M. Ishii, Mining experimental data from materials science literature with large language models: an evaluation study, Sci. Technol. Adv. Mater.: Methods 4 (1) (2024) 2356506, http://dx.doi.org/10.1080/27660400.2024.2356506.
[83] G. Wang, J. Li, Y. Sun, X. Chen, C. Liu, Y. Wu, M. Lu, S. Song, Y.A. Yadkori, Hierarchical reasoning model, 2025, arXiv:2506.21734 [cs.AI].
[84] A. Jolicoeur-Martineau, Less is more: recursive reasoning with tiny networks, 2025, URL https://arxiv.org/abs/2510.04871. arXiv:2510.04871 [cs.LG].
[85] L.M. Ghiringhelli, C. Baldauf, T. Bereau, S. Brockhauser, C. Carbogno, J. Chamanara, S. Cozzini, S. Curtarolo, C. Draxl, S. Dwaraknath, et al., Shared metadata for data-centric materials science, Sci. Data 10 (1) (2023) 626, http://dx.doi.org/10.1038/s41597-023-02501-8.