AutoIE-LLM: An Automated Information Extraction Framework from Scientific Literature Based on the LLM

Anonymous ACL submission

Abstract

Specialized research literature in PDF contains abundant domain-specific knowledge, yet extracting critical information from these documents remains a daunting challenge. To address this, we propose AutoIE-LLM, an innovative information extraction framework integrating Large Language Models (LLMs) with human-in-the-loop for domain-specific knowledge processing. The framework comprises layout analysis, key information extraction, and continuous learning modules. We introduce a novel dataset of 1,122 chemical molecular sieve documents to validate our approach. Experimental results demonstrate that AutoIE-LLM achieves 79% accuracy in named entity recognition and relation extraction tasks, a 10% improvement over the baseline AutoIE model. The framework handles complex terminology and non-standard document structures, demonstrating its effectiveness in specialized domains. This study enhances LLMs' capabilities in expert fields and provides a valuable resource for future molecular sieve information extraction studies.

1 Introduction

In recent years, the rapid digitalization of scientific publishing has led to exponential growth in literature volume and has posed a significant challenge for researchers and industry experts: efficiently extracting key information from complex documents. Traditional information extraction(IE) methods, such as rule-based systems and early machine learning models, require extensive manual effort and face difficulties in generalizing across different domains or adapting to new types of documents (Reichenpfader et al., 2023).

The advent of deep learning, particularly Transformer-based models like BERT and GPT, has demonstrated exceptional capabilities in understanding and generating human language, making them particularly suitable for tasks involving complex linguistic structures (Choi et al., 2023). Despite the success of Transformer-based models in general NLP tasks, they still face significant limitations when applied to scientific literature. These limitations include difficulties in processing long documents, understanding domain-specific terminology, and extracting structured information from unstructured text. Moreover, Large Language Models (LLMs) often lack the specialized knowledge required for specific scientific domains, potentially producing hallucinations or inaccurate outputs (Hong et al., 2021). Recent studies have shown that even state-of-the-art LLMs achieve only about 50-70% accuracy when dealing with highly specialized, complex scientific texts (Ghosh et al., 2024; Rasool et al., 2024; Hartmann et al., 2023).

This study proposes AutoIE-LLM, an automated IE framework that leverages Large Language Models with a human-in-the-loop mechanism to address these challenges. It enables efficient and accurate extraction of key information from complex domain-specific scientific literature, significantly reducing human effort and processing time. We validate the framework through rigorous comparisons with baseline models.

AutoIE-LLM comprises three modules: Layout Analysis Unit for accurate parsing of document structure; Key IE Unit leveraging large language models for precise domain-specific extraction; Human Feedback Unit integrating expert knowledge for continuous model refinement and bias reduction.

To validate the effectiveness of the AutoIE-LLM framework, we conducted rigorous testing in the specialised field of chemical molecular sieves. This domain was chosen due to its complexity and urgent need for precise IE. The results are compelling: AutoIE-LLM achieved an average accuracy of 79%. These metrics not only demonstrate the framework's robust capabilities but also mark a significant advancement in the field of specialised

IE.

In summary, the main contributions of this paper include:

(1) We propose **AutoIE-LLM**, an end-to-end information extraction framework that synergistically integrates large language models with human-in-the-loop feedback. This design effectively addresses the challenges of extracting structured knowledge from complex, domain-specific scientific literature.

(2) We construct and release a high-quality benchmark datasetthe molecular sieve literature dataset of 1,122 molecular sieve papers, filling a critical gap in domain-specific IE research and enabling systematic evaluation of scientific IE models.

(3) Through comprehensive experiments, we demonstrate that AutoIE-LLM significantly outperforms state-of-the-art baselines on this benchmark, achieving an average accuracy of 79% and showcasing strong adaptability to specialised scientific domains.

2 Related Work

2.1 Challenges and Advances in Scientific Literature Information Extraction

With the rapid increase in scientific publications, extracting key information efficiently and comprehensively has become a pressing issue. Traditional IE methods, such as rule-based systems and early machine learning models, face significant challenges when handling the complexity and diversity of scientific literature (Martsinkevich et al., 2023). These methods require substantial manual effort to create and maintain rule sets, and they struggle to generalize across domains or adapt to new types of documents. Recent deep learning approaches, particularly Transformer-based models such as BERT and GPT, have significantly enhanced IE performance (Xu et al., 2020). However, applying these models directly to specialized scientific literature—where terminology, structure, and domain knowledge are highly complex—still poses considerable challenges. Empirical evidence indicates that LLMs often achieve only 50-70% accuracy in specialized domains (Pan et al., 2024; Hasan et al., 2020; Zhang et al., 2024b), highlighting the need for improved domain adaptation and knowledge integration methods.

2.2 Applications of Large Language Models in Scientific Literature Information Extraction

Recent efforts to adapt LLMs for named entity recognition (NER) and relation extraction (RE) tasks in scientific documents have shown promise. For instance, (Zhang et al., 2024a) leveraged Gemini for pseudo-labeling and fine-tuned the GLM-4 model to address overlapping entities in engineering inspection data, offering insights for dealing with unstructured specialized texts. Likewise, (Uchida, 2024) demonstrated LLMs' capacity for corpus linguistics, indicating potential in handling complex linguistic structures. In the NER context, (Cheng et al., 2024) introduced a standardized prompting strategy that improves cross-domain and low-resource performance, complementing the "divide and transfer" paradigm from (Zhang et al., 2024c). Additionally, (Xu et al., 2024) proposed a Dual Contrastive Learning model to bolster LLMs' cross-domain extraction capabilities under limited data, demonstrating the effectiveness of token- and sentence-level contrastive learning. Despite these advances, challenges persist in highly specialized contexts like chemical zeolites, where complex documents and domain-specific terminology are prevalent.

2.3 The Role of Document Layout Analysis in Information Extraction

Document layout analysis is important in scientific literature IE, particularly for documents with complex structures, such as chemical zeolite research reports, which often contain tables, figures, and non-standard structures. LayoutLM (Xu et al., 2020) and DocFormer (Appalaraju et al., 2021) exemplify this trend by integrating visual, spatial, and textual signals, thus enhancing document understanding tasks. These advanced layout analysis techniques form a crucial foundation for the AutoIE-LLM framework, enabling more accurate recognition and processing of the complex structured information in chemical zeolite literature.

2.4 The Key Role of Human-Machine Collaboration in Information Extraction

While LLMs have made strides in IE, human expertise remains critical for improving accuracy and reliability. (Liu et al., 2024) introduced a human-in-the-loop strategy that incrementally refines model predictions with expert feedback, reducing man-

ual effort while boosting performance in specialized domains. Similarly, (Hsu et al., 2022) showed that incorporating human feedback helps models handle complex document layouts better. Active learning, online learning, and knowledge distillation collectively enable the model to identify and address uncertainties, continuously refine extraction patterns, and assimilate expert judgments. By incorporating these human-machine collaboration techniques, especially online learning, the AutoIE-LLM framework effectively addresses challenges posed by complex terminology and experimental data in chemical zeolite literature, enhancing system flexibility and precision.

2.5 Specific Challenges and Solutions for Information Extraction in the Chemical Zeolite Domain

IE in chemical zeolite research is complicated by specialized chemical structures, non-uniform data formats, and domain-specific terminology. (Meuschke et al., 2023) benchmarked text extraction tools for scientific documents, highlighting the difficulties of dealing with complex domain literature. Likewise, (Sharma, 2023) combined advanced OCR techniques with deep learning for chemical patents, underscoring the need for robust solutions to handle specialized content. Despite these efforts, existing methods struggle to balance accuracy, adaptability, and efficiency in highly specialized domains. Our proposed AutoIE-LLM framework integrates layout analysis, human-machine collaboration, and domain-adaptive learning to mitigate these limitations. AutoIE-LLM aims to deliver a more reliable and flexible solution for extracting information from chemical zeolite literature by focusing on complex terminologies, non-standard structures, and continuous feedback loops.

3 Framework

This section elaborates on the architecture of our proposed AutoIE-LLM framework, providing a detailed explanation of our methodologies and algorithms. We comprehensively overview the framework's processing pipeline (as illustrated in Figure 1) and systematically introduce each component.

3.1 Layout Analysis Unit

As discussed in Section 2.1, extracting text from PDF documents is fundamental to information retrieval. The inherent complexity of PDF layouts, including non-standard fonts, graphics, and diverse language and character encoding issues, poses significant challenges for current technologies in preserving the original logical structure during text extraction. To address these challenges, we introduce a sophisticated document layout analysis unit.

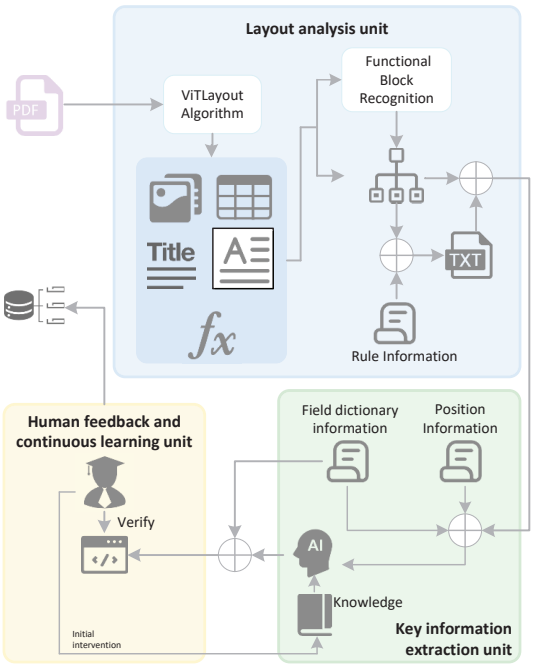

Figure 1: The overall architecture of the AutoIE-LLM framework comprises three core modules: the layout analysis unit, the key information extraction unit, and the Human-in-the-loop and continuous learning unit.

Our unit employs the VTLayout algorithm (Li et al., 2021) to classify components within PDFs (such as headings, text, and figures), ensuring precise identification and segmentation. Subsequently, we implement scientific document functional block identification techniques (Xu et al., 2020) to analyze text components in-depth, describing the logical structure within the PDF and thus preserving the document's hierarchical organization. This approach facilitates the rearrangement of text in text documents according to the original logical order, significantly enhancing the accuracy of subsequent IE processes.

To precisely locate important information, we adopt a hybrid approach that combines traditional rule-based methods enhanced by domain expertise with logical layout analysis of the document. This refined text is then transmitted to the key IE unit for further processing.

Furthermore, our unit addresses several specific

challenges in scientific PDF conversion:

(1) Text formatting: We preserve the original text format by analyzing layout patterns in scientific literature and optimizing paragraph recognition processes. This mitigates issues such as improper line breaks or paragraph segmentation.

(2) Unique characters and noise: We tackle challenges posed by special characters, diacritics, and formulas prevalent in scientific literature. Our approach uses a rule-based denoising method tailored for molecular sieve literature, significantly reducing the impact of noise on subsequent models.

(3) Non-standard fonts: For older documents using non-standard fonts, we implement PaddleOCR technology combined with font recognition algorithms, thereby improving processing accuracy.

This comprehensive strategy significantly enhances the efficiency and accuracy of text extraction from complex scientific PDF documents through advanced layout analysis and functional block identification techniques.

3.2 Key Information Extraction Unit

While traditional transformer-based models have demonstrated excellence in many IE tasks, they face limitations in handling long-distance dependencies and generalizing across different contexts. Our research leverages Large Language Models (LLMs), renowned for their powerful generalization capabilities and proficiency in managing long-distance textual relationships to address these challenges.

Our key IE unit operates as follows: Firstly, refined text from the layout analysis unit is input into the LLM for initial processing. Secondly, we employ prompt engineering techniques supplemented with domain-specific datasets to offset potential deficiencies of general LLMs in domain-specific knowledge. We utilize the CRISPE framework to fine-tune the model's domain adaptability and IE precision. Thirdly, we collaborate with domain experts to compile fundamental principles and terminology related to molecular sieves. This domain-specific knowledge is integrated into the model through custom-designed knowledge injection mechanisms, enhancing the model's focus and enabling accurate identification and extraction of domain-relevant information. Finally, The fine-tuned and knowledge-enhanced LLM processes the input text, extracting key information based on predefined extraction tasks and domain-specific requirements.

By combining the powerful processing capabilities of LLMs with in-depth domain knowledge, our approach transcends the limitations of traditional models in managing long-distance dependencies and domain-specific tasks. This approach significantly improves the precision and efficiency of IE in specialized domains.

3.3 Human-in-the-loop and Continuous Learning Unit

To address the limitations of LLMs in specialized scientific domains and reduce the high cost of manual annotations, we design a Human-in-the-loop and Continuous Learning Unit. This module enables efficient expert intervention at the early stages and progressively reduces the need for human input as the system improves its domain-specific extraction capabilities. Specifically, once the model's performance on key domain-specific metrics (extraction precision rate) stabilizes above a predefined threshold, expert review becomes optional or triggered only by system uncertainty. Compared to traditional deep learning pipelines, this mechanism achieves higher accuracy while requiring significantly fewer labelled samples and computational resources. The workflow of this unit is as follows:

- Key IE: Upon uploading new scientific documents, the system autonomously extracts key information and displays it for review via a web interface.

- Domain experts validate and correct the extracted results through an intuitive interface, providing high-quality feedback with minimal effort.

- Knowledge base update: Verified information is integrated into a growing knowledge base, which is subsequently used to Knowledge base update: Verified information is integrated into a growing knowledge base, which is subsequently used to enrich fine-tuning datasets with high-quality domain-specific samples, refine prompt strategies and injection mechanisms based on emerging domain patterns, and adjust prompt engineering strategies to

accommodate new patterns or requirements identified during expert reviews.

- Adaptive Model Updating: The LLM is periodically fine-tuned using the enhanced knowledge base, leading to continual improvement in domain understanding and extraction precision.

This unit employs continuous learning, ensuring the framework enriches its domain knowledge from initial configuration through ongoing application. This strategy enhances the model's accuracy and reliability and optimizes resource allocation by minimizing the need for substantial human and material investment in data annotation. This design supports a dynamic, self-improving learning loop. As the model matures, it relies less on expert input and delivers increasingly accurate results, making the framework more scalable, resource-efficient, and adaptable to evolving scientific domains.

4 Chemical Molecular Sieve Literature Dataset Data Collection and Processing

This section introduces and examines a groundbreaking Chemical Molecular Sieve Literature Dataset that represents a significant advance in the field. This curated resource, with portions available on GitHub, fills a critical gap in molecular sieve research by serving as a specialized, high-quality benchmark for IE models. By supporting deeper exploration of structure-property relationships, the dataset has the potential to substantially enhance our understanding of these materials.

4.1 Data Source

To maintain the specificity and depth of our research, we focused exclusively on literature about molecular sieves(as shown in Figure E4 of 8.2), a critical class of materials in various industrial applications and scientific studies. Domain experts meticulously curated a dataset of 1,122 papers from 1993 to 2022 (the data distribution of each year is shown in Figure 3), ensuring a representative sample of key research within this specialized field. Our dataset covers 51 peer-reviewed journals with SJR scores ranging from 0.296 to 18.509. High-impact titles such as NATURE and SCIENCE and specialized journals like Microporous and Mesoporous Materials are included. We systematically extracted 1,575 unique data points from this corpus related to gel composition, a fundamental aspect of molecular sieve synthesis and performance. This extensive data extraction process provides a robust foundation for our analysis and represents one of the largest compilations of molecular sieve gel composition data to date, offering significant potential for advancing our understanding of structure-property relationships in these materials.

4.2 Annotation Process

Domain experts performed a comprehensive, multi-stage annotation process on the sampled documents to create a high-quality labelled dataset. Initially, a large language model extracted information based on a predefined dictionary (included terms and parameters specific to molecular sieve research, such as Zeolite types (e.g., ZSM-5, Beta, MOR), templates (e.g., TPAOH, TEAOH), silica sources (e.g., tetraethyl orthosilicate), Si/Al ratio, synthesis temperature, crystallization time) and rule set (as shown in Figure E6 of 8.2). Three domain experts then reviewed and corrected these initial extractions. Subsequently, two senior experts conducted a final review of the corrected results. The annotated information was stored as JSON files (exemplified in 8.1 in a temporary repository) if approved.

To ensure data quality, a Python script was employed to analyze all corrected JSON files, followed by manual verification to identify potential errors. After passing all quality checks, the JSON files were stored in a database for future model fine-tuning. This rigorous process ensured the identification and correct labelling of key information fields relevant to molecular sieve research, as demonstrated by the annotation tools in Figure E5 of 8.2.

4.3 Data Cleaning and Verification

We implemented a multi-stage data cleaning and verification process to ensure data quality and minimize noise, combining expert human judgment with automated analysis techniques. Our approach consisted of three key stages:

1. Stratified Expert Review: We employed a tiered annotation system involving junior and senior experts. Junior experts performed initial annotations in groups, followed by cross-validation among these groups. Senior experts then conducted a final review, ensuring comprehensive error detection and consistency across the dataset.

2. Custom Python scripts (detailed in Appendix

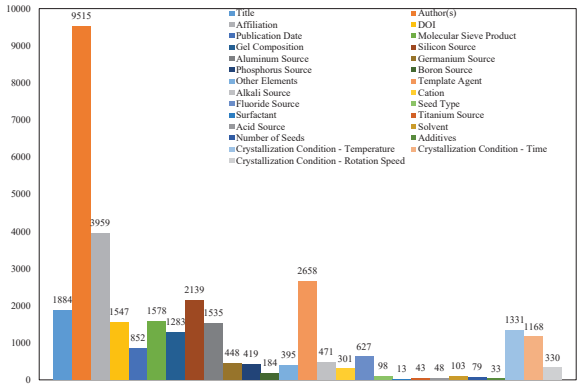

Figure 2: Data Distribution of Chemical Molecular Sieve Dataset.

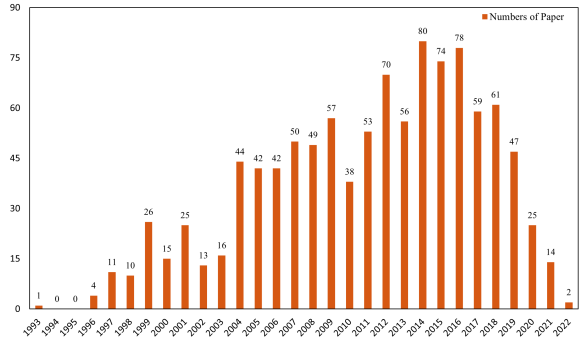

Figure 3: Data Distribution of Chemical Molecular Sieve paper from 1993 to 2022.

D, Section 8.4) performed advanced text analysis tasks, including:

- **Syntactic and formatting consistency checks:** Ensuring all data entries follow a uniform format, such as consistent units and standardized parameter names.
- **Semantic relevance analysis:** Verifying that extracted data is relevant to molecular sieve research and excluding irrelevant information.
- **Cross-document comparison:** Detecting potential duplicate data or conflicting information across different publications.
- **Statistical analysis of annotation distribution:** Evaluating the frequency and distribution of variables to identify outliers or biases (some statistical information is shown in Appendix E, Section 8.5).

3. **Iterative Refinement:** Anomaly reports generated by our automated system were returned to annotators for correction. This iterative process rectified errors and helped continuously optimize our annotation guidelines, improving overall data quality.

4.4 Data Distribution

We analyzed the frequency of both bibliographic and chemical-specific fields (Figure 2). Bibliographic information (Title, Author, Date) occurs most frequently, while fields like Molecular Sieve Product and Gel Composition also appear prominently. Certain specialized parameters (e.g., Phosphorus or Fluoride Source) are less common, reflecting narrower research scopes.

Notable observations from the data distribution include: Bibliographic information (Title, Author(s), Publication Date) is the most consistently available data across the dataset; Chemical-specific information, such as Molecular Sieve Product and Gel Composition, is also well-represented, though less frequent than bibliographic data; Some specialized fields like Phosphorus Source, Boron Source, and Fluoride Source have relatively low occurrence rates (395, 301, and 98 occurrences, respectively), indicating they may be relevant only for specific molecular sieves or synthesis methods.

This data distribution provides insights into the types of information most commonly reported in molecular sieve literature and highlights areas where data may be more scarce. Understanding this distribution is crucial for developing effective IE models and identifying potential gaps in reporting practices within the field.

To promote reproducibility and facilitate further research in this domain, we have made a subset of our dataset publicly available on GitHub (accessible at: `https://anonymous.4open.science/r/molecular-sieve-dataset-3CC6/`). This subset includes 500 data entries covering major synthesis parameters and product characteristics.

This initiative addresses a significant gap in the field, as, to our knowledge, no public IE dataset specific to molecular sieves existed prior to this work. While our dataset's scale may be smaller than general-purpose IE datasets, its high quality and domain specificity offer unique advantages. The specialized nature of our dataset enables superior generalization capabilities in molecular sieve-related applications, as demonstrated by our experimental results in Section 5.1.

The creation of this dataset represents a significant contribution to the molecular sieve research

Hyper-parameter	Value	Hyper-parameter	Value
Batch size	4	LoRA rank	16
Number of iterations	1	LoRA Alpha	32
Learning rate	1e-6	LoRA Dropout	0.1
Maximum sentence length	4096	Regularization coefficient	0.01
Warmup ratio	0.01		

Table 1: Fine-tuning parameter values.

Field	Baseline LLMs		AutoIE-LLMs
Title	**0.93**	0.77	**0.93**
Journal	**0.88**	**0.88**	**0.88**
DOI	0.95	0.84	**0.97**
Gel Composition	0.57	0.85	**0.92**
Crystallization Conditions-Time	0.77	0.56	**0.91**
Template	0.43	0.46	**0.69**
Alkali Source	0.54	0.59	**0.64**
Aluminum Source	0.53	0.50	**0.62**
Cation	0.60	0.42	**0.61**
Accuracy	0.69	0.65	**0.79**

Table 2: Accuracy comparison of baseline and LLM methods across various information extraction fields.

community. It provides a benchmark for evaluating IE models in this domain and opens up new avenues for developing and fine-tuning specialized natural language processing models for domain-specific scientific literature analysis. Additional system interfaces are shown in Appendix C, Section 8.3.

5 Experiment

This section presents a comprehensive overview of our experimental design, including dataset construction, experimental setup, evaluation metrics, and results analysis. Our experiments aim to rigorously assess the performance of the AutoIE-LLM framework in processing scientific literature, focusing on IE in the domain of molecular sieves.

5.1 Experimental Setup

Our framework supports flexible integration with different LLMs. In this study, we selected the Llama-2-13B-v2 model as our backbone, fine-tuning it with the LoRA (Low-Rank Adaptation) method. This technique introduces low-rank approximations in the linear layers to reduce parameter overhead and mitigate overfitting. The training process used a chemical zeolite domain dataset with 33,211 records, and key hyperparameters are listed in Table 1. To simulate the human-in-the-loop mechanism, we manually verified only 1/3 of the training data, after which high-confidence predictions (confidence ≥ 0.9) were progressively incorporated without expert intervention.

5.2 Evaluation Metrics

To comprehensively assess the performance of our model, we employed accuracy measures of the overall correctness of the model's predictions and is computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.

5.3 Results and Analysis

We evaluated five methods for information extraction: AutoIE, LLM, AutoIE-LLM (with LLM as the backbone), FINE_TUNE, and AutoIE-LLM(F) (with fine-tuned LLM as the backbone). We then selected ten key fields from the dataset—Title, Journal, DOI, Gel Composition, Crystallization Conditions Time, Template, Alkali Source, Aluminum Source, and Cation—based on high data completeness, direct relevance to zeolite synthesis, adequate sample size, and a balanced number of fields to mitigate overfitting. As Gel Composition already integrates related fields (e.g., Silicon Source), these were omitted to avoid redundancy. Although some factors may be overlooked, this selection ensures a concise analytical framework capturing core elements of zeolite synthesis while preserving parsimony and generalizability.

The experimental results, as presented in Tables 2 and 3, demonstrate that the AutoIE-LLM framework generally excels across various information extraction fields, particularly in areas requiring deep semantic understanding.

Comparing AutoIE-LLM with baseline and pure LLM models (Table 2), we observe that AutoIE-LLM outperforms both in most fields. Notable

Field	FINE TUNE	AutoIE-LLM(F)	AutoIE-LLMs
Title	0.85	0.88	**0.91**
Journal	0.78	0.80	**0.88**
DOI	0.94	**0.97**	**0.97**
Gel Composition	0.90	0.84	**0.92**
Crystallization Conditions-Time	0.63	0.81	**0.91**
Template	0.30	0.22	**0.69**
Alkali Source	0.53	0.63	**0.64**
Aluminum Source	0.41	0.49	**0.62**
Cation	0.34	0.35	**0.61**
Accuracy	0.63	0.67	**0.79**

Table 3: Ablation study: Accuracy comparison of AutoIE-LLM variants across various information extraction fields.

improvements are seen in DOI extraction (accuracy 0.97), gel composition recognition (0.92), and crystallization condition-time extraction (0.91). These findings highlight the framework's robust capabilities in processing complex scientific information. The success of AutoIE-LLM can be attributed to several key factors:

- Effective integration of AutoIE's structured information processing with LLMs' semantic understanding capabilities.
- A critical layout analysis module for accurately identifying document structures.
- Strong competence in handling complex scientific terminology.
- Continuous optimization through the learning module, enhancing performance in specific fields.

Pure LLM models exhibit strong but inconsistent performance across different fields. This indicates powerful semantic understanding capabilities but potentially insufficient comprehension of domain-specific structural nuances. Table 3 presents an ablation study comparing the performance of different AutoIE-LLM variants across information extraction fields.

The ablation study (Table 3) provides insights into the performance of fine-tuned models. FINE_TUNE models show improvements in some areas but do not consistently outperform AutoIE-LLM. This suggests that domain-specific fine-tuning can enhance performance but may not fully compensate for the lack of structured information processing capabilities.

AutoIE-LLM(F), a fine-tuned variant of AutoIE-LLM, shows mixed results. It performs well in some complex domains but exhibits instability in others. For instance, it achieves the highest accuracy for DOI extraction (0.97, tied with AutoIE-LLM) but underperforms in fields like Template and Cation extraction. This instability may be attributed to overfitting, highlighting the challenge of balancing generalization and task-specific performance in fine-tuned models.

In conclusion, these results emphasize the potential of integrated approaches like AutoIE-LLM in scientific information extraction tasks. AutoIE-LLM achieves significant performance improvements across multiple complex domains by combining structured information processing with deep semantic understanding. However, the results also point to some limitations and areas for improvement, particularly in handling highly structured information and ensuring consistent performance across different domains.

6 Conclusion

This paper introduces AutoIE-LLM, an automated information extraction framework grounded in large language models for extracting key information from scientific texts. By integrating layout analysis, key information extraction, and a human feedback loop for continuous learning, AutoIE-LLM effectively addresses challenges in processing complex scientific literature. Experiments on a zeolite-related chemical literature dataset demonstrate its robust performance, with the ERNIE BOT model—incorporating AutoIE-LLM—achieving notable metric improvements and confirming the framework's effectiveness and stability. Moreover, AutoIE-LLM exhibits strong adaptability and a modular design that facilitates processing domain-specific data. Future directions include expanding to broader scientific domains and more complex text types, integrating diverse datasets, and exploring advanced algorithms to further enhance the framework's capabilities. We believe AutoIE-LLM will play a significant role in scientific literature analysis, driving the development and application of automated information extraction technology.

7 Limitations

The AutoIE-LLM framework demonstrates significant strengths in integrating structured information processing with deep semantic understanding. The framework's exceptional performance in multiple domains, particularly in fields requiring deep semantic comprehension, can be attributed to several factors. The effective integration of AutoIE's structured information processing capabilities with the semantic understanding of Large Language Models (LLMs) plays a crucial role. The layout analysis unit's accuracy in identifying document structures and the robust handling of complex scientific terminology by the integrated LLM contribute significantly to the framework's success. Notably, the framework achieved remarkable accuracy in DOI extraction (0.97), gel composition identification (0.92), and crystallization conditions-time extraction (0.91). These results underscore the framework's ability to effectively process and interpret complex scientific information.

However, the framework exhibits limitations in accurately extracting information from certain complex fields such as Unit and Molecular Sieve. These challenges are likely due to the highly domain-specific semantics and inconsistent structural representations in the source texts. Future work could address this by incorporating domain-adaptive pretraining or structure-aware decoding mechanisms to enhance generalization across complex scientific fields.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R. Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003.

Qi Cheng, Liqiong Chen, Zhixing Hu, Juan Tang, Qiang Xu, and Binbin Ning. 2024. A novel prompting method for few-shot ner via llms. *Natural Language Processing Journal*, 8:100099.

Hyeon Seok Choi, Jun Yeong Song, Kyung Hwan Shin, Ji Hyun Chang, and Bum-Sup Jang. 2023. Developing prompts from large language models for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiation Oncology Journal*, 41(3):209.

Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, and Debasis Ganguly. 2024. Image 1 AlpaPICO: extraction of pico frames from clinical trial documents using LLMs. *Methods*, 226:78–88.

Benedict Hartmann, Philippe Tamla, Florian Freund, and Matthias Hemmje. 2023. Fine-tune it like i'm five: supporting medical domain experts in training NER models using cloud, LLM, and auto fine-tuning. In *Proceedings of the 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS)*, pages 1–8. IEEE.

Fatema Hasan, Arpita Roy, and Shimei Pan. 2020. Integrating text embedding with traditional nlp features for clinical relation extraction. In *Proceedings of the 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 418–425.

Zhi Hong, Logan Ward, Kyle Chard, Ben Blaiszik, and Ian Foster. 2021. Challenges and advances in information extraction from scientific literature: A review. *JOM*, 73(11):3383–3400.

Enshuo Hsu, Ioannis Malagaris, Yong-Fang Kuo, Rizwana Sultana, and Kirk Roberts. 2022. Deep learning-based nlp data pipeline for ehr-scanned document information extraction. *JAMIA Open*, 5(2):ooac045.

Shoubin Li, Xuyan Ma, Shuaiqun Pan, Jun Hu, Lin Shi, and Qing Wang. 2021. Vtlayout: Fusion of visual and text features for document layout analysis. In *Proceedings of PRICAI 2021: Trends in Artificial Intelligence*, pages 308–322, Cham. Springer International Publishing.

Yangyang Liu, Shoubin Li, Kai Huang, and Qing Wang. 2024. Autoie: An automated framework for information extraction from scientific literature. In *Proceedings of Knowledge Science, Engineering and Management*, pages 424–436, Singapore. Springer Nature Singapore.

Viacheslav Martsinkevich, Andrei Berezhkov, Vladislav Tereshchenko, Natalia Gorlushkina, and Violetta Tretjakova. 2023. Algorithms for extracting lines, paragraphs with their properties in pdf documents. In *E3S Web of Conf.*, volume 389, page 08024.

Norman Meuschke, Apurva Jagdale, Timo Spinde, Jelena Mitrović, and Bela Gipp. 2023. A benchmark of pdf information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity*, pages 383–405, Cham. Springer Nature Switzerland.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Zafaryab Rasool, Stefanus Kurniawan, Sherwin Balugo, Scott Barnett, Rajesh Vasa, Courtney Chesser, Benjamin M. Hampstead, Sylvie Belleville, Kon Mouzakis, and Alex Bahar-Fuchs. 2024. Evaluating llms

on document-based qa: Exact answer selection and numerical extraction using cogtale dataset. *Natural Language Processing Journal*, 8:100083.

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2023. Protocol: large language model-based information extraction from free-text radiology reports: a scoping review protocol. *BMJ Open*, 13(12).

Parikshit Sharma. 2023. Advancements in ocr: A deep learning algorithm for enhanced text recognition. *International Journal of Inventive Engineering and Sciences*, 10(8).

Satoru Uchida. 2024. Using early llms for corpus linguistics: Examining chatgpt's potential and limitations. *Applied Corpus Linguistics*, 4(1):100089.

Jingyun Xu, Junnan Yu, Yi Cai, and Tat-Seng Chua. 2024. Dual contrastive learning for cross-domain named entity recognition. *ACM Transactions on Information Systems*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, New York, NY, USA. Association for Computing Machinery.

Chenhong Zhang, Xiaoming Lei, Ye Xia, and Limin Sun. 2024a. Automatic bridge inspection database construction through hybrid information extraction and large language models. *Developments in the Built Environment*, 20:100549.

Jiaxin Zhang, Lingxue Zhang, Yuxuan Sun, Wei Li, and Ruge Quhe. 2024b. Named entity recognition in the perovskite field based on convolutional neural networks and matbert. *Computational Materials Science*, 240:113014.

Xinghua Zhang, Bowen Yu, Xin Cong, Taoyu Su, Quangang Li, Tingwen Liu, and Hongbo Xu. 2024c. Cross-domain ner under a divide-and-transfer paradigm. *ACM Transactions on Information Systems*, 42(5):1–32.

8 Appendix

8.1 Labelled data content


```
{
    "content": "\nMicroporous and Mesoporous
    ↪  Materials 78 (2005) 181–188\nwww.el⌋
    ↪  sevier.com/locate/micromeso\n
    ↪  Meso/macroporous AlPO-5 spherical
    ↪  macrostructures tailoredby resin
    ↪  templating  Valeri Naydenov a,*,
    ↪  Lubomira Tosheva a,1, Oleg N.
    ↪  Antzutkin b,*, Johan Sterte a,2\na
    ↪  Division of Chemical Technology,
    ↪  Lulea  University of Technology,
    ↪  S-971 87 Lulea  , Swedenb Division of
    ↪  Chemistry, Lulea  University of
    ↪  Technology, S-971 87 Lulea  ,
    ↪  Sweden\nReceived 22 June 2004;
    ↪  received in revised form 4 October
    ↪  2004; accepted 5 October
    ↪  2004Available online 30 November
    ↪  2004\n Abstract\n A multi-step
    ↪  procedure for the preparation of
    ↪  meso/macroporous AlPO-5 spherical
    ↪  macrostructures using cation
    ↪  exchange resinbeads as
    ↪  macrotemplates is presented. Firstly,
    ↪  aluminum species were introduced
    ↪  into the resin beads by ion exchange
    ↪  resulting in aresin-aluminum
    ↪  composite. Thereafter, the
    ↪  resin-aluminum composite was mixed
    ↪  with TEAOH, H3PO4 and distilled water
    ↪  andhydrothermally treated at 150  C
    ↪  to yield resin-AlPO-5 composite.
    ↪  Finally, the resin was removed by
    ↪  calcination leaving
    ↪  behindself-bonded AlPO-5 spheres. The
    ↪  product AlPO-5 macrostructures were
    ↪  thoroughly characterized by SEM, XRD,
    ↪  nitrogen adsorp-tion measurements,
    ↪  31P and 27Al solid state NMR
    ↪  spectroscopy. The influence of
    ↪  various components of the synthesis
    ↪  mixture on thecrystallinity, phase
    ↪  purity and stability of the AlPO-5
    ↪  spheres was systematically studied.
    ↪  Samples prepared for different
    ↪  treatmenttimes using the initial
    ↪  synthesis composition that gives
    ↪  spheres of the highest quality were
    ↪  used to study the crystallization
    ↪  processwithin the resin.Ó 2004
    ↪  Elsevier Inc. All rights reserved.\n
    ↪  Keywords: AlPO-5; Hierarchical
    ↪  porosity; Spheres; Macrotemplate;
    ↪  Ion-exchange resin\n1.
    ↪  Introduction\n Molecular sieve
    ↪  materials have found wide use in
    ↪  alarge number of industrially
    ↪  important areas such
    ↪  aschemicalseparation, adsorption and
    ↪  heterogeneouscatalysis. For certain
    ↪  applications however, the smallpore
    ↪  size of zeolites (micropores) may
    ↪  cause diffusionlimitations. Also,
    ↪  zeolites are usually synthesized
    ↪  aspowders, which are difficult to
    ↪  handle and post-synthetic\n*
    Corresponding authors. Tel.: +46 920
    ↪  492524; fax: +46 920491199.E-mail
    ↪  address: oleg.antzutkin@ltu.se (O.N.
    ↪  Antzutkin).1 Present address: LMM,
    ↪  UMR-7016 CNRS, ENSCMu, Universite´de
    ↪  Haute Alsace, rue Alfred Werner,
    ↪  F-68093 Mulhouse Cedex,France.2
    ↪  Present address: Va¨ xjo¨ University,
    ↪  Universitetplatsen 1, S-351 95Va¨
    ↪  xjo¨ , Sweden.\n 1387-1811/$ - see
    ↪  front matter Ó 2004 Elsevier Inc. All
    ↪  rights reserved.doi:10.1016/j.micro⌋
    ↪  meso.2004.10.008\n modifications to
    ↪  obtain the zeolites in
    ↪  macroscopicforms are needed. During
    ↪  the last years, a lot of researchhas
    ↪  been directed towards the
    ↪  development of syntheticprocedures
    ↪  that tailor the pore structure and
    ↪  the macro-scopic shape of zeolites. A
    ↪  number of molecular sievebodies with
    ↪  hierarchical porosity providing fast
    ↪  trans-port to and from the zeolite
    ↪  pores as well as withmacro-shapes
    ↪  that meet the operating conditions
    ↪  fora particular application have
    ↪  been synthesized usingmacrotemplates.
    ↪  The macrotemplate acts as a
    ↪  molddetermining the macroscopic
    ↪  shape of the product mate-rial,
    ↪  whereas the removal of the
    ↪  macrotemplate aftersynthesis creates
    ↪  a secondary porosity in the meso
    ↪  and/or macropore range. Thus,
    ↪  zeolites in forms of mono-liths
    ↪  [1–4], fibers [5], hollow capsules
    ↪  [6], sponge-likearchitectures [7]
    ↪  and self-standing tissues [8] have
    ↪  beensynthesized using starch [1],
    ↪  latex beads [2], mesoporoussilica
    ↪  spheres [3,6], polyurethane foams
    ↪  [4], bacterial\n182\n V. Naydenov et
    ↪  al. / Microporous and Mesoporous
    ↪  Materials 78 (2005) 181–188\nthreads
    ↪  [5], cellulose acetate membranes [7]
    ↪  and woodcellular structures [8] as
    ↪  templates.Microporous
    ↪  aluminophosphate solids are
    ↪  anotherimportant class of molecular
    ↪  sieve materials that havea
    ↪  three-dimensional framework built up
    ↪  of alternating(AlO4) and (PO4) units.
    ↪  AlPO-5 (AFI type structure) isthe
    ↪  most studied member of this family.
    ↪  The influenceof synthesis
    ↪  composition, chemicals used and the
    ↪  hea-ting procedure on the AlPO-5
    ↪  crystallization have beendiscussed
    ↪  in a number of papers [9–18]. However
    ↪  tothe best of our knowledge, the
    ↪  accessibility of alumino-phosphate
    ↪  molecular sieves for preparation of
    ↪  self-bonded macro-shaped bodies have
    ↪  not yet beenexplored.Here, we report
    ↪  on a procedure for tailoring
    ↪  sphericalmeso/macroporous AlPO-5
    ↪  macrostructures using cat-ion
    ↪  exchange resins as macrotemplates.
```

```
The procedureis a further development of
↪ the resin-templating methodused for
↪ the preparation of silicalite-1 [19]
↪ and zeolite[20,21] molecular sieve
↪ macrostructures using anionexchange
↪ resins.  However, this type of resin
↪ is not appli-cable for AlPO-5
↪ macrostructure synthesis since
↪ posi-tively charged aluminophosphate
↪ species are present inAlPO-5
↪ synthesis solutions [22]. In
↪ addition, the crystal-lization
↪ mechanism of AlPO-5 within the resin
↪ was stud-ied in detail by 31P and
↪ 27Al solid state NMR.\n 2.
↪ Experimental section\n Fig. 1 shows
↪ a schematic representation of the
↪ wholeprocedure used in this work for
↪ the preparation of
↪ themeso/macroporous AlPO-5 spherical
↪ macrostructuresthrough resin
↪ templating.\n 2.1. Preparation of
↪ the resin-Al composites\n A
↪ macroreticular Amberlite IRA-200
↪ cation ex-change resin (mesh size
↪ 16–50, Sigma) was used in
↪ allexperiments. The ionic form of the
↪ resin was reversedfrom Na+ to H+ by
↪ passing a 10 wt.% HCl solutionthrough
↪ an ion exchange column loaded with
↪ the resin.A large batch of resin-Al
↪ composites was prepared bymixing the
↪ resin (H+ form) with a 0.1 M
↪ tetraethyl-ammonium chloride hydrate
↪ (TEACl, Aldrich), dis-solved in an
↪ aluminum chlorohydrate solution
↪ diluted\n 10 times (Locron L, 23.4
↪ wt.% Al2O3, OH/Al = 2.5,Hoechst) in a
↪ weight ratio resin to solution of
↪ 1:10,followed by a treatment in an
↪ oil bath at 100  C underreflux for 24
↪ h. The resin-Al composite was
↪ separatedafter the synthesis by
↪ decanting, rinsed repeatedly
↪ withdistilled water and dried at room
↪ temperature. Theamount of aluminum
↪ (22.9 wt.%) exchanged into theresin
↪ (calculated as Al2O3) was determined
↪ gravimetri-cally by the weight
↪ difference between resin-Al
↪ compos-ites dried at 105  C and the
↪ beads calcined at 600  C.\n 2.2.
↪ Preparation of AlPO-5
↪ macrostructures\n Synthesis mixtures
↪ containing ortho-phosphoric acid(85%,
↪ Merck), distilled
↪ water,tetraethylammoniumhydroxide
↪ (TEAOH, 20 or 35 wt.% aqueous
↪ solutions,Sigma) and resin-Al
↪ composites in quantities to
↪ yieldmolarcompositionsxTEAOH:Al2O3:⌋
↪ yP2O5:zH2O,where x = 1.5\u00003.5, y
↪ = 0.9\u00001.3 and z = 50\u0000500,
↪ wereprepared. The water gained
↪ during storage or the
↪ tetra-ethylammonium ion present in
↪ the resin-Al compositefrom the
↪ initial treatment of the resin with
↪ alumina, werenot taken into account
↪ in the calculations.

The mixtureswere hydrothermally treated
↪ in autoclaves at 150  C fortreatment
↪ times between 2 and 24 h. After
↪ hydrothermaltreatment, the
↪ resin-AlPO-5 composite was
↪ separatedfrom the mother liquor by
↪ decanting, rinsed repeatedlywith
↪ distilled water and dried at room
↪ temperature.The dried composite was
↪ finally calcined at 600  C for20 h
↪ after heating to this temperature at
↪ a rate of1  C min\u00001.\n 2.3.
↪ Characterization\n A Philips XL 30
↪ scanning electron microscopy
↪ (SEM)was used to study the morphology
↪ of the samples. Nitro-a Micromeritics
↪ ASAP 2010 instrument at \u0000196  C
↪ aftergen adsorption/desorption
↪ isotherms were obtained withdegassing
↪ the samples at 300  C overnight prior
↪ to anal-ysis. Specific surface area
↪ was calculated with the BETequation.
↪ Pore size distributions were
↪ determined fromthe desorption branch
↪ of the isotherms using the BJHmethod.
↪ Micropore surface areas and
↪ micropore vol-umes were determined
↪ by the t-plot method and totalpore
↪ volumes were obtained from the volume
↪ adsorbedat a relative pressure of
↪ 0.995. Crystalline phases were\nFig.
↪ 1. Schematic representation of
↪ procedure for the preparation of
↪ self-bonded AlPO-5 spheres using
↪ macroporous cation exchange resin
↪ asmacrotemplate.\n V. Naydenov et al.
↪ / Microporous and Mesoporous
↪ Materials 78 (2005) 181–188\n183\n
↪ identified with a Siemens D 5000
↪ X-ray powder diffracto-meter (XRD)
↪ using CuKa radiation. The pH of
↪ themother liquors was measured with
↪ a pH meter 691(Metrohm).Solid-state
↪ 31P magic-angle-spinning (MAS)
↪ NMRspectra were recorded on a
↪ Varian/Chemagnetics Infi-nity
↪ CMX-360 (B0 = 8.46 T) spectrometer
↪ using the sin-gle-pulse experiment
↪ with proton decoupling. The
↪ 31Poperating frequency was 145.73
↪ MHz. In the single pulseexperiment,
↪ the 31P 90  pulse duration was 5.0 ls
↪ and thenutation frequency of protons
↪ during decoupling wasxnut/2p = 100
↪ kHz. 16 signal transients spaced by
↪ arelaxation delay of 60 s were
↪ accumulated.Calcined powder samples
↪ (ca. 30–35 mg, additionallydried at
↪ 350  C for 3 days in order to allow
↪ quantitativemeasurements of
↪ phosphorus) were packed in
↪ zirconiumdioxide double bearing 4 mm
↪ rotors. All 31P solid stateMAS NMR
↪ spectra were recorded at room
↪ temperature.
```

27Al MAS NMR experiments were performed atroom temperature on a Varian/Chemagnetics Infinity-600 (University of Warwick, UK) spectrometer at 27Alcarrier frequency of 156.37 MHz in 3.2 mm zirconiumdioxide rotors at a spinning frequency of 18,000 ±10 Hz. The duration of the 30 -excitation pulse was0.5 ls. 128 signal transitients were accumulated with arepetition time of 1 s. Samples were externally referencedon a powder YAG sample [24], which was also usedfor the tuning of the magic angle.\n 3. Results and discussion\n Fig. 2 shows a SEM image of the initial resin beadsused as macrotemplates (Fig. 2a) and images of theproduct particles (Fig. 2b–f) obtained at the differentsteps of the procedure depicted in Fig. 1. According toSEM, the first step of the procedure, namely the intro-duction of aluminum within the resin does not causechanges into e.g. the size and/or shape of macrotem-plates (Fig. 2b). The resin-AlPO-5 composite particlesobtained in the next step of the procedure were also simi-lar in size and shape to the original resin beads but withsomewhat rougher surfaces due to AlPO-5 agglomeratesexposed on the surface (Fig. 2c). The calcined AlPO-5spheres were similar in shape with a slightly reduced sizecompared to the original resin beads (Fig. 2d). Some ofthe particles were cracked and even broken. The spheresurfaces were rough with micrometer-sized voids andcavities (Fig. 2e). No such voids were observed in thesphere interiors. The AlPO-5 spheres were built up offine nano-particles as shown in Fig. 2f. This observation\nis not surprising considering the fact that AlPO-5 crys-tallizes within the resin pores, which have an averagesize of ca. 100 nm [25]. This may explain the absenceof micrometer-sized crystals with well defined AlPO-5hexagonal morphology.The described procedure for the preparation ofAlPO-5 spherical macrostructures differs significantlyfrom the conventional AlPO-5 syntheses [9] reportedin literature, as well as from the procedures used forthe synthesis of silicalite-1 [19] or zeolite [20,21] macro-structures by resin templating. The experiments to pre-pare AlPO-5 macrostructures by direct treatment ofcation exchange resins with AlPO-5 synthesis solutionsdid not give satisfactory results.

Particles of limited crys-tallinity often accompanied by a loss of macroshapewere obtained. The problem of disintegration of themacroshape was solved by insertion of aluminum pre-cursor within the resin prior to AlPO-5 synthesis. Thispreliminary step ensures that the AlPO-5 crystallizationis realized only within the resin pore structure and there-fore an easy recovery of the product spheres due to theabsence of bulk crystallization. Further, the presence ofTEA+ in the solution used for Al ion exchange wasessential for the AlPO-5 synthesis. The exact role ofthe TEA+ is not clear at this point of the study. However, when a resin-Al composite was prepared in the ab-sence of TEACl and used for AlPO-5 synthesis theproducts obtained upon calcinations were powdersrather than beads. This indicates that the TEA+ might‘‘fix’’ the aluminum species within the resin to ensurehomogeneous AlPO-5crystallization withinthemacrotemplate.\n Further, structural and macroscopic characteristicsof the product samples prepared with different initialcompositions were studied and results are given in Table1. The duration of the hydrothermal treatment for allsamples was 10 h at 150 C. The objective of the presentwork was to synthesize stable AlPO-5 spheres of highcrystallinity. Therefore, the quality of the samples wasevaluated by the degree of crystallinity, purity ofAlPO-5 and by the mechanical stability of the productAlPO-5 spheres. Although the mechanical stability ofthe macrostructures was not tested, obtained particleswere stable and could withstand various laboratorymanipulations. However it should be mentioned, thatAlPO-5 spheres were easier to grind (e.g. prior XRDor NMR studies) compared to Silicalite-1 and zeoliteBeta macrostructures prepared by resin templating[19,20]. A possible explanation might be the differencein the amount of solid material within resin-molecularsieve composites obtained after synthesis, which de-creases from zeolite Beta (56 wt.%) through Silicalite-1(44 wt.%) to AlPO-5 (26 wt.%). As evident from the datapresented (Table 1) the best results in terms of crystalli-nity, purity and mechanical stability were obtained forsample 4 and this synthesis mixture was used to prepare\n184\n V. Naydenov et al. / Microporous and Mesoporous Materials 78 (2005) 181–188\n

Fig. 2. SEM images of the cation exchange resin beads (a) and the product particles (b–f) synthesized from system with molar composition2TEAOH:Al2O3:1.2P2O5:100H2O after hydrothermal treatment at 150 C for 10 h.\n Table 1Synthesis of AlPO-5 macrostructures from different system compositions\nSample\n Molar composition\n TEAOH\n Al2O3\n P2O5\n H2O\n 12345678910f11121314\n 2.02.02.02.02.01.52.53.03.52.02.02.02.0\n1.001.001.001.001.001.001.001.001.001.001.001.00\n0.901.001.101.201.301.201.201.201.201.201.201.20\n100100100100100100100100100100100100\n a According to XRD; Am—amorphous; T—tridymite; Im—impurities.b Visually.c Yellowish.d Slightly gray.e Deteriorated.f Sample prepared with 35 wt.% TEAOH.\n Product characteristics\n Structurea\n AmAmAm + AlPO-5AlPO-5TIm + AmAm + traces of AlPO-5AmAmAlPO-5 + TAlPO-5 + ImAlPO-5 + ImAlPO-5 + Im + AmIm + Am\nMacro-shapeb\n SpherescSpherescSpherescSpheresSpheresSpheresdSphereselPowder + spheresdPowder + spheresdPowder + spheresSpheresSpheresSpheresSpheres\nsamples for different treatment times to study the AlPO-5 crystallization mechanism within the resin.\n Fig. 3 shows XRD patterns of series of calcined sam-ples synthesized using the same synthesis mixture as\n V. Naydenov et al. / Microporous and Mesoporous Materials 78 (2005) 181–188\n185\n .\n u\n .\n a\n /\n y\n t\n i\n s\n ne\n t\n n\nI\n (24 h)\n (12 h)\n (10 h)\n (8 h)\n (6 h)(5 h)\n (4 h)(2 h)\n 5\n10 15 20 25 30 35 40 45 50\n 2\theta/o\n Fig. 3. XRD patterns of the calcined samples obtained from thesystem with molar composition 2TEAOH:Al2O3:1.2P2O5:100H2O at150 C for various treatment times.\n sample 4 (Table 1) and prepared with duration of thehydrothermal treatment between 2 and 24 h. The initialresin-Al composite beads used in these experiments werefrom the same batch. The sample prepared using 2 h oftreatment was completely amorphous. AlPO-5 peaksof very low intensity were detected in the sample ob-tained after 4 h oftreatment. The intensity oftheAlPO-5 peaks further increased with an increase in thetime of hydrothermal treatment and the sample pre-pared for 10 h was highly crystalline pure AlPO-5. Inaddition to the AlPO-5 phase,impurities of anotherunidentified crystalline phase (designated with arrowson the Figure) were also present in the samples treatedfor 5, 6 and 8 h. Surprisingly, prolonging the hydrother-mal treatment to 12 and 24 h lead to AlPO-5 spherescontaining impurities of crystalline phases most likelythe same as detected for shorter treatment times.Sincethe trend shown in Fig. 3 is rather unexpected the repro-ducibility of the results was confirmed by the repetitionof the synthesis.The changes in the pore structure of product macro-structures with the time of treatment were studied bynitrogen adsorption, BET surface areas, micropore sur-face areas and volumes and total pore volumes are listedin Table 2. The values for the calcined initial resin-alu-mina composites are also included in the Table for acomparison. The size of those spheres was substantiallyreduced upon calcination and this might explain thelower total pore volume of this sample. Though the val-ues vary, the general trend is an increase in the BET sur-\nTable 2Porosity of the calcined samples synthesized from system with molarcomposition 2TEAOH:Al2O3:1.2 P2O5:100H2O at 150 C altering theduration of the hydrothermal treatment\n Sample\n BET surfacearea, SBET\u00001)(m2 g\n Microporesurface area,\u00001)Sl (m2 g\n Microporevolume,\u00001)Vl (cm3 g\n Al spheres2 h4 h5 h6 h8 h10 h12 h24 h\n 571108510230417228929626 5\n 881028183\n98214227184\n 0.0030.0020.0040.0130.0870.0460.102 0.1080.087\nTotal porevolume,Vp\u00001)(cm3 g\n 0.2380.8160.8010.8640.4370.6970.426 0.4010.409\n face area, micropore surface area and volume and adecrease in the total pore volume with increased timeof treatment. The samples prepared using 6 and 8 h oftreatment show values that do not fit into this trend.The reason might be that according to XRD, AlPO-5becomes the predominant phase in the 6 h sample andthus the changes in the structure of the inorganic mate-rial within the resin are most intense at this time intervalof treatment (6–8 h). Dif-ferent complex species containing Al or binuclear Alcenters with H3PO4 or H2PO4 and H2O molecules as li-gands showing P chemical shifts in this range have beenreported in literature [27–29]. However, the specificationof the short range solid structure judging only from the31P solid state NMR would be highly speculative. After5 h of hydrothermal treatment a new peak (with a linewidth of \u00185.6 ppm) at ca. \u000029 to \u000030 ppm appearedon the top of the broad peak (at \u000027 ppm).

This indicates\n\u0000\n that a new material with more ordered structure (P siteswith more uniform local chemical environment) is form-ing in this sample. Chemical shift \u000029 to \u000030 ppm is inthe range characteristic for tetrahedral P sites in theframework of microporous materials [30,31]. In supportto this, AlPO-5 peaks were also observed in the XRDpattern of sample 5 h (Fig. 3(5 h)). In the spectra of thesamples treated for 6 (not shown) and 8 h the peak at\u000029 to \u000030 ppm dominates over the broad componentthat corresponds to the initial amorphous solid. Accord-ing to the 31P NMR, after 10 h of hydrothermaltreatment, the conversion from amorphous to a poly-crystalline material is completed. Only slight changes inthe line shapes can be noticed in the spectra of the sam-ples prepared for 12 and 24 h (not shown) of hydrother-mal treatment. The spectra of samples (10 h), (12 h) and(24 h) are characterized by a resonance line, which mightactually be composed of two overlapping peaks, ca. \u000030and \u000031 ppm with total line-width of ca. \u00185.6 ppm. Thisresonance line width is somewhat broader than what isusually observed for highly polycrystalline phosphoruscontaining materials. A large part of the line broadeningis probably due to interaction of P-spins with 27Al nuclei(spin 5/2, 100% natural abundance). However,theassignment of the 31P peak(s) at \u000030/\u000031 ppm (togetherwith the line-widths) to microporous AlPO-5 is in agree-ment with the results previously reported in the literaturefor the dehydrated AlPO-5 phase [30–32].Integral intensities of the 31P resonance lines in Fig. 5can be used for quantitative estimation of the phospho-rus content (in wt %) for these samples. In the calcu-lations, all 31P integral peak areas (centre band wastween 20 and \u000070 ppm, see Fig. 5) were normalized tointegrated together with its spinning side bands, i.e. be-the area of the peak of commercial AlPO4 (Fig. 5 bot-tom) with known phosphorus content (25.4 wt%). Thetotal amount of P introduced as H3PO4 in the systemduring the sample preparation was 26.9 wt%. The P con-tent as well as the pH of the mother liquor as a functionof the duration of hydrothermal treatment are plotted inFig. 6.

From the Figure it is seen that most of the phos-phorus (ca. 25.0 wt.%) is present within the resin afterthe first 2 h of treatment. With an increase in the dura-tion of the hydrothermal treatment, the P content inthe spheres slowly increases and reaches a plateau atca. 29.1 wt.% after the tenth hour of treatment. The re-sults for P content are slightly higher than the calculatedvalue on H3PO4 basis but nevertheless give values ofsatisfactory accuracy. These errors in P content mea-surements might be caused by variation of the sampleweight due to water adsorbed from air by the samplesduring the rotor packing. It is known that anhydrousAlPO4, which was used as a reference sample in thecalculations, is a highly hydroscopic material.The pH is an important parameter in aluminophos-phate syntheses, which may alter crystal size and product\n V. Naydenov et al. / Microporous and Mesoporous Materials 78 (2005) 181–188\n187\n 1098765432\nr\n i\n ou\nq\n i\nl\n r\n e\nh\n t\n o\nm\n e\nh\n t\n n\ni\n H\n p\n P content pH\n 048121620\n24Duration of hydrothermal treatment/h\n 40\n 35\n 30\n 25\n 20\n 15\n 10\n%\n t\n w\n /\nt\n ne\n t\n noc\n P\n Fig. 6. P content in the product spheres (estimated from 31P MASNMR) and pH of the mother liquors collected after synthesis ofsamples obtained from system with molar composition 2TEAOH:Al2O3:1.2P2O5:100H2O at 150 C as a function of the duration of thehydrothermal treatment.\n yield [10,13,16–18]. In this study, the pH values plotted inFig. 6 are values measured outside the resin in the motherliquor, therefore they should be used with some precau-tion since the pH of the mother liquor might significantlydiffer compared to that within resin interior, whereAlPO-5 actually crystallizes. The pH monitoring mightgive information about the transportation of e.g. P andTEA+ in the resin. The initial pH of the H3PO4, TEAOHand distilled water mixture was 3.1. During the first 2 h oftreatment, the pH of the mother liquor sharply increasedto 5.4, slightly decreased after 5 h of treatment andreached ca. 4.5 upon further prolongation of the treat-ment. It should be emphasized that although this is thepH of the mother liquor, the pH remains acidic duringthe whole time interval explored, whereas for conven-tional syntheses the final pH is about neutral [10].

```
Inthe literature the initial increase in
↪ the pH during the syn-thesis is
↪ considered as an indication that the
↪ amount ofthe free phosphoric acid is
↪ decreasing [13,16]. This ismost
↪ likely the case in our approach as
↪ well, since the ini-tial increase of
↪ the pH (2 h of treatment) correlates
↪ wellwith the large amount of P for
↪ this sample measuredby quantitative
↪ 31P NMR. In addition, the fact that
↪ thepH remains in the acidic range,
↪ suggests that the TEA+is also
↪ transported into the resin already
↪ at the beginningof the synthesis.Fig.
↪ 7 shows single-pulse 27Al MAS NMR
↪ spectraof the calcined resin-Al
↪ composites (denoted as Alspheres),
↪ commercial AlPO4 and selected
↪ spectra fromseries of samples
↪ obtained varying the duration
↪ ofhydrothermal treatment discussed
↪ in Fig. 5 and Fig. 6.These spectra
↪ were recorded for as received
↪ samplesi.e. without additional
↪ drying prior to measurementsto
↪ remove adsorbed water. Aluminum
↪ sites in the cal-cined resin-Al
↪ composite have three predominant
↪ typesof chemical environment,
↪ characterized by broad peaksat ca.
↪ 60, 40 ppm (tetrahedral coordination)
↪ and ca.\n (24 h)\n (10 h)\n (6 h)\n(2
↪ h)\ncommercial AlPO4\n Al
↪ spheres\n100\n75\n 50\n25\n 0\n-25\n
↪ -50\n Chemical shift / ppm\n Fig. 7.
↪ Single pulse 27Al MAS NMR spectra of
↪ the samples obtainedfrom thesystem
↪ with molarcomposition2TEAOH:Al2O3:1
↪ .2P2O5:100H2O at 150  C for various
↪ treatment times.\n 5 ppm (octahedral
↪ coordination) (Al spheres). All
↪ thesepeaks completely disappeared
↪ after 2 h of hydrothermal(Al (V)) and
↪ about \u000013 ppm (Al (VI)) (Fig.
↪ 7(2 h)) ap-treatment, and new peaks
↪ at 45 ppm (Al (IV)), 10 ppmpeared.
↪ These new peaks are almost identical
↪ to thosein the spectrum of
↪ commercial AlPO4 (Fig. 7). The
↪ weakbut discernible signal at about
↪ 10 ppm can be assigned tothe
↪ five-coordinated aluminum according
↪ to previous re-ports [33]. With an
↪ increase in the duration of the
↪ hydro-thermal treatment all three
↪ resonance peaks becomenarrower with
↪ the peaks being sharpest for the
↪ samplesobtained after 10 h and 24 h
↪ of treatment. Also, a consi-derable
↪ shift of the 27Al-resonance peak at
↪ 45 ppm to atabout 37 ppm is noticed
↪ for the tetrahedral aluminumsites in
↪ the sample (Fig. 7(10 h)),
which was previouslyassigned to the
↪ AlPO-5 based on XRD (Fig. 4 (10
↪ h))and 31P (Fig. 5(10 h)) results.
↪ The small fraction of octa-hedral
↪ aluminum sites in AlPO-5 (at ca.
↪ \u000012 ppm) canalso be observed,
↪ since a certain amount of water is
↪ ad-sorbed by the sample. In previous
↪ reports it has beenshown by a number
↪ of 2D 27Al-31P NMR
↪ correlationexperiments performed on
↪ water-AlPO-5 system thatthese Al
↪ (VI)-sites, which additionally
↪ coordinate twowater molecules can be
↪ correlated with 31P-signals forthe
↪ latter system, i.e. these Al (VI)
↪ sites are actually inthe AlPO-5
↪ framework positions [34]. Therefore,
↪ both31P and 27Al NMR as well as XRD
↪ studies prove thatthe sample
↪ prepared for 10 h of hydrothermal
↪ treatmentcontains a high quality
↪ AlPO-5 phase.\n 4. Conclusions\n
↪ Highly crystalline and mechanically
↪ stable AlPO-5spheres were prepared
↪ using a cation exchange resin
↪ as\n188\n V. Naydenov et al. /
↪ Microporous and Mesoporous Materials
↪ 78 (2005) 181–188\na macrotemplate.
↪ The AlPO-5 phase crystallized in
↪ thepore structure of a cation
↪ exchange resin loaded withAl
↪ precursor species under a
↪ hydrothermal treatmentwith a mixture
↪ of H3PO4, TEAOH and distilled
↪ water.The overall molar composition
↪ of the synthesis mixtureinfluences
↪ both phase purity and sphere
↪ appearance.Best results, highly
↪ crystalline AlPO-5 spheres, were
↪ ob-tained using the mixture with the
↪ molar composition2TEAOH:Al2O3:1.2P2
↪ O5:100H2O and 10 h of hydro-thermal
↪ treatment at 150  C. The spheres
↪ synthesizedfor treatment times other
↪ than 10 h were contaminatedwith
↪ amorphous and/or other crystalline
↪ phases.The pore structure of the
↪ AlPO-5 spheres preparedfor 10 h of
↪ treatment was complex containing
↪ micro-meso- and macropores. The
↪ micropores are due to thepresence of
↪ AlPO-5, whereas the meso and
↪ macroporesemanate from the resin
↪ removal.The crystallization
↪ mechanism of AlPO-5 within theresin
↪ was extensively studied by solid
↪ state NMR.The quantitative
↪ determination of the P content
↪ withinthe solid spheres by 31P NMR
↪ indicated that P is takenup by the
↪ resin from the external solution at
↪ the begin-ning of the hydrothermal
↪ treatment (2 h). Further
↪ pro-longation ofthetreatmentleadsto
↪ structuralrearrangements of the
↪ system resulting in the
↪ crystalliza-tion of AlPO-5.
```

```
The quality of the AlPO-5 phase
↪ wasevaluated by XRD, nitrogen
↪ adsorption, 31P, and 27AlMAS NMR
↪ measurements.The procedure
↪ presented contributes to the
↪ currenttrends directed towards the
↪ preparation of self-bondedmaterials
↪ with hierarchical pore structures.
↪ The macro-scopic spherical shape
↪ and the complex pore structureof
↪ the AlPO-5 spheres prepared makes
↪ them interestingfor direct
↪ applications in e.g. fixed bed
↪ reactors.\n Acknowledgment\n We
↪ thank Prof. R. Dupree and D.
↪ Rusanova-Nayde-nova for both
↪ instrument time and assistance in
↪ operat-ing with Varian/CMX-600 NMR
↪ instrument (WarwickUniversity, UK).
↪ O.N.A. acknowledges Swedish
↪ Concilfor planning and coordination
↪ of research (FRN) for agrant for the
↪ Varian/CMX-360 spectrometer and
↪ Foun-dation to the memory of J.C.
↪ and Seth M. Kempe for agrant for the
↪ TR-4 mm-MAS probe and other
↪ equip-ment. The partial financial
↪ support from the SwedishResearch
↪ Councilfor Engineering Sciences
↪ (VR)isgratefully acknowledged.\n
↪ References\n [1] B. Zhang, S.A.
↪ Davis, S. Mann, Chem. Mater. 14
↪ (2002) 1369.\n [2] K.H. Rhodes, S.A.
↪ Davis, F. Caruso, B. Zhang, S. Mann,
↪ Chem.Mater. 12 (2000) 2832.\n[3] A.
↪ Dong, Y. Wang, Y. Tang, Y. Zhang, N.
↪ Ren, Y. Yue, Z. Gao,Adv. Mater. 14
↪ (2002) 1506.\n[4] Y.-J. Lee, J.S.
↪ Lee, Y.S. Park, K.B. Yoon, Adv.
↪ Mater. 13 (2001)\n1259.\n[5] B.
↪ Zhang, S.A. Davis, N.H. Mendelson,
↪ S. Mann, Chem.Commun. (2000)
↪ 781.\n[6] A. Dong, Y. Wang, Y. Tang,
↪ N. Ren, Y. Zhang, Z. Gao,
↪ Chem.Mater. 14 (2002) 3217.\n[7] Y.
↪ Wang, Y. Tang, A. Dong, X. Wang, N.
↪ Ren, W. Shan, Z.Gao, Adv. Mater. 14
↪ (2002) 994.\n [8] A. Dong, Y. Wang,
↪ Y. Tang, N. Ren, Y. Zhang, Y. Yue, Z.
↪ Gao,Adv. Mater. 14 (2002) 926.\n[9]
↪ S.T. Wilson, B.M. Lok, C.A. Messina,
↪ T.R. Cannan, E.M.Flanigen, J. Am.
↪ Chem. Soc. 104 (1982) 1146.\n [10]
↪ B.L. Newalkar, B.V. Kamath, R.V.
↪ Jasra, S.G.T. Bhat, Zeolites\n18
↪ (1997) 286.\n[11] M.Z. Yates, K.C.
↪ Ott, E.R. Birnbaum, T.M. McCleskey,
↪ Angew.Chem. Int. Ed. 41 (2002)
↪ 476.\n[12] G. Finger, J.
↪ Richter-Mendau, M. Bu¨ low, J.
↪ Kornatowski,Zeolites 11 (1991)
↪ 443.\n [13] X. Ren, S. Komarneni,
↪ D.M. Roy, Zeolites 11 (1991)
↪ 142.\n[14] O¨ . Weiß, G. Ihlein, F.
↪ Schu¨ th, Micropor. Mesopor. Mater.
↪ 35-36\n(2000) 617.\n[15] S. Mintova,
↪ S. Mo, T. Bein, Chem. Mater. 10
↪ (1998) 4030.\n
[16] T. Kodaira, K. Miyazawa, T. Ikeda,
↪ Y. Kiyozumi, Micropor.Mesopor. Mater.
↪ 29 (1999) 329.\n[17] I. Girnus, K.
↪ Jancke, R. Vetter, J. Richter-Mendau,
↪ J. Caro,Zeolites 15 (1995) 33.\n[18]
↪ H. Du, M. Fang, W. Xu, X. Meng, W.
↪ Pang, J. Mater. Chem. 7\n(1997)
↪ 551.\n[19] L. Tosheva, V. Valtchev,
↪ J. Sterte, Micropor. Mesopor.
↪ Mater.\n35-36 (2000) 621.\n[20] L.
↪ Tosheva, B. Mihailova, V. Valtchev,
↪ J. Sterte, Micropor.Mesopor. Mater.
↪ 48 (2001) 31.\n [21] L. Tosheva, J.
↪ Sterte, R. Fiello, G. Giorduo, F.
↪ Testa (Eds.), in:Impact of zeolites
↪ and other Porous Materials on the
↪ newTechnologies at the Beginning of
↪ the new Millenium,
↪ Elsevier,Amsterdam. Stud. Surf. Sci.
↪ Catal. 142A (2002) 183.\n[22] S.
↪ Prasad, S.-B. Liu, Micropor. Mater. 4
↪ (1995) 391.\n[23] K. Karaghiosoff,
↪ Encyclopedia of Nuclear Magnetic
↪ Resonance,Wiley, New York, 6 (1996)
↪ 3612.\n[24] D. Massiot, C. Bessada,
↪ J.P. Coutures, F. Taulelle, J.
↪ Magn.Reson. 90 (2) (1990) 231.\n[25]
↪ C.E. Harland, Ion Exchange: Theory
↪ and Practice, The RoyalSociety of
↪ Chemistry, Cambridge, 1994, p.
↪ 28.\n[26] B.B. Johnson, A.V. Ivanov,
↪ O.N. Antzutkin, W. Forsling,Langmuir
↪ 18 (2002) 1104.\n[27] J.W. Akitt,
↪ N.N. Greenwood, G.D. Lester, J. Chem.
↪ Soc. (A)\n(1971) 2450.\n[28] R.F.
↪ Mortlock, A.T. Bell, C.J. Radke, J.
↪ Phys. Chem. 97 (1993)\n767.\n[29]
↪ R.F. Mortlock, A.T. Bell, C.J. Radke,
↪ J. Phys. Chem. 97 (1993)\n775.\n[30]
↪ D. Mu¨ ller, E. Jahn, G. Ladwig, U.
↪ Haubenreisser, Chem. Phys.Lett. 109
↪ (1984) 332.\n[31] D. Mu¨ ller, E.
↪ Jahn, B. Fahlke, G. Ladwig, U.
↪ Haubenreisser,Zeolites 5 (1985)
↪ 53.\n[32] R.H. Meinhold, N.J. Tapp,
↪ J. Chem. Soc. Chem. Commun.\n(1990)
↪ 219.\n[33] G. Kunath-Fandrei, T.J.
↪ Bastow, J.S. Hall, C. Jaeger,
↪ M.E.Smith, J. Phys. Chem. 99 (1995)
↪ 15138.\n[34] C.A. Fyfe, K.C.
↪ Wong-Moon, Y. Huang, Zeolites 16
↪ (1996) 50.\n",
                "connections": [
{
        "toId": 114,
        "id": 1,
        "text": "belong to",
        "fromId": 116
},
{
        "toId": 114,
        "id": 2,
        "text": "belong to",
        "fromId": 134
},
{
        "toId": 114,
        "id": 4,
        "text": "belong to",
        "fromId": 136
```

```
},
{
        "toId": 114,
        "id": 6,
        "text": "belong to",
        "fromId": 123
},
{
        "toId": 114,
        "id": 8,
        "text": "belong to",
        "fromId": 124
},
{
        "toId": 114,
        "id": 9,
        "text": "belong to",
        "fromId": 126
},
{
        "toId": 114,
        "id": 10,
        "text": "belong to",
        "fromId": 139
},
{
        "toId": 114,
        "id": 11,
        "text": "belong to",
        "fromId": 142
},
{
        "toId": 137,
        "id": 7,
        "text": "belong to",
        "fromId": 114
}
],
"others": [],
"labels": [
{
        "startIndex": 167,
        "endIndex": 182,
        "id": 87,
        "text": "author"
},
{
        "startIndex": 1,
        "endIndex": 37,
        "id": 97,
        "text": "Jornal"
},
{
        "startIndex": 91,
        "endIndex": 168,
        "id": 100,
        "text": "Title"
},
{
        "startIndex": 191,
        "endIndex": 207,
        "id": 102,
        "text": "author"
},
{
        "startIndex": 213,
        "endIndex": 230,
        "id": 103,
        "text": "author"
},
{
        "startIndex": 236,
        "endIndex": 248,
        "id": 104,
        "text": "author"
},
{
        "startIndex": 532,
        "endIndex": 548,
        "id": 109,
        "text": "Publish Date"
},
{
        "startIndex": 347,
        "endIndex": 401,
        "id": 138,
        "text": "Unit"
},
{
        "startIndex": 255,
        "endIndex": 319,
        "id": 140,
        "text": "Unit"
},
{
        "startIndex": 2512,
        "endIndex": 2543,
        "id": 141,
        "text": "doi"
},
{
        "startIndex": 6402,
        "endIndex": 6425,
        "id": 114,
        "text": "Gel composition"
},
{
        "startIndex": 6742,
        "endIndex": 6748,
        "id": 124,
        "text": "Crystallization
        ↪  conditions - Temperature #1"
},
{
        "startIndex": 6768,
        "endIndex": 6786,
        "id": 126,
        "text": "Crystallization
        ↪  conditions - Time #1"
},
{
        "startIndex": 6159,
        "endIndex": 6165,
        "id": 137,
        "text": "Zeolite"
},
{
        "startIndex": 6565,
        "endIndex": 6583,
        "id": 123,
        "text": "Al#1"
},
{
        "startIndex": 6459,
        "endIndex": 6468,
        "id": 134,
        "text": "H2O Number"
},
{
        "startIndex": 6432,
```

```
        "endIndex": 6442,
        "id": 116,
        "text": "Template Number"
    },
    {
        "startIndex": 6213,
        "endIndex": 6238,
        "id": 136,
        "text": "P"
    },
    {
        "startIndex": 6445,
        "endIndex": 6452,
        "id": 139,
        "text": "P Number"
    },
    {
        "startIndex": 6293,
        "endIndex": 6298,
        "id": 142,
        "text": "Template#1"
    }
    ]
    }
```


8.2 Images in data management

Figure E4: Source of data PDF sample.

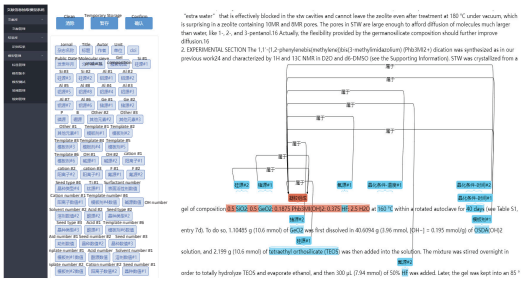

Figure E5: The annotation tools.

Figure E6: An example of dictionary and rule base.

8.3 System Interface

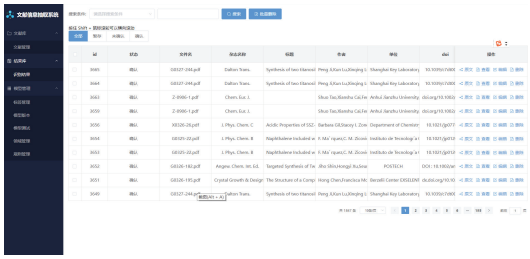

Figure E7: Recognition Result Interface.

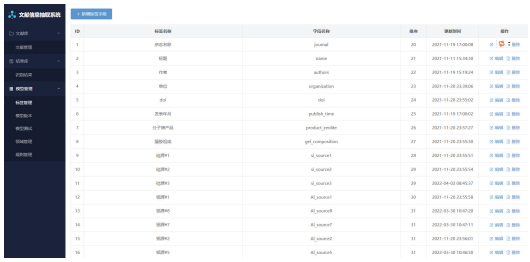

Figure E8: Label Management Page.

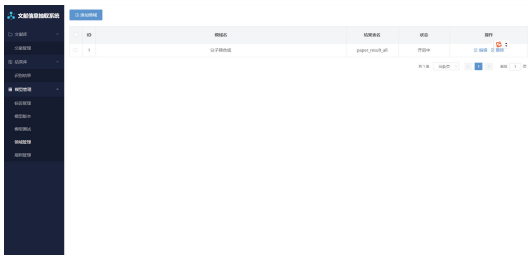

Figure E9: Domain Management Interface.

8.4 Data cross-check code

```
import os
import json

def read_labels(label_path):
with open(os.path.join(label_path, 'label.txt'),
↪  'r', encoding='utf8') as f_txt:
return [label.strip() for label in
↪  f_txt.readlines()]

def process_json_file(json_path, json_name,
↪  labels):
with open(os.path.join(json_path, json_name),
↪  'r', encoding='utf8') as fp:
json_data = json.load(fp)

extracted_data = {label: [] for label in labels}

for json_item in json_data['labels']:
for label in labels:
if label in json_item['text']:
start = json_item['startIndex']
end = json_item['endIndex']
content = json_data['content'][start:end]
extracted_data[label].append(content)

return extracted_data

def write_output(new_json_path, extracted_data):
for label, data in extracted_data.items():
with open(os.path.join(new_json_path,
↪  f"{label}.txt"), 'w', encoding='utf8') as
↪  json_file:
json_file.write(str(data))

def main():
label_path = ''
json_path = ''
new_json_path = ''

labels = read_labels(label_path)

all_extracted_data = {label: [] for label in
↪  labels}

for json_name in os.listdir(json_path):
extracted_data = process_json_file(json_path,
↪  json_name, labels)
for label, data in extracted_data.items():
all_extracted_data[label].extend(data)

write_output(new_json_path, all_extracted_data)

if __name__ == "__main__":
main()
```

8.5 Label information statistics

Figure E10: Label the DOI information statistics in the data.

Figure E11: Label the Si information statistics in the data.

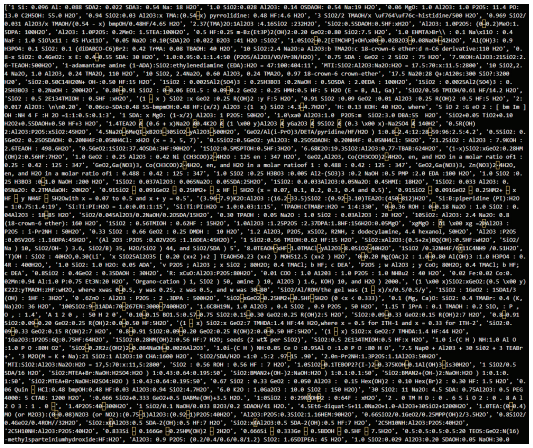

Figure E12: Label the Gel composition information statistics in the data.