Supplementary Material for ZeolBERT: A Domain-Adapted Language Model for Automated Extraction of Zeolite Synthesis Procedures

Data Annotation Process

To ensure high-quality annotations, a systematic and rigorous annotation protocol was developed involving domain experts. Scientific articles were initially retrieved from Elsevier's API in XML format and parsed into raw text using the *lxml* package. Text segmentation into sentences was performed using ChemDataExtractor to generate zeolite-specific training corpora. For manually collected literature in PDF format, PDFMiner was employed for text extraction and subsequent sentence segmentation.

Paragraphs explicitly describing zeolite synthesis procedures were labeled as "1", while other paragraphs detailing characteristics, properties, and analysis were labeled as "0". After deduplication and data cleaning, the final ZeolitePC dataset comprised 998 unique paragraphs.

Annotation was carried out according to 15 experimental operations and 12 entity types detailed in Tables 1 and 2.

Table 1: Experimental Operations and Descriptions

Operation	**Description**
Add	Addition of raw materials
Process	Execution of experimental steps
Stir	Mixing acceleration
Dry	Moisture removal
Wash	Removal of impurities
Recover	Sample retrieval
Calcine	Heat-induced decomposition
Heat	Temperature elevation
Crystallize	Crystal formation process
Transfer	Container relocation
Cool	Temperature reduction
React	Chemical interaction
Age	Material maturation
Seal	Airtight closure
Yield	Output production

Table 2: Entity Types and Descriptions

Entity Category	**Description**
Action	Procedure-describing verbs
Material	Raw chemical components
Temperature	Reaction temperature
Duration	Reaction duration
Container	Reaction vessel (e.g., autoclave)
Sample	Analytical specimen
Solvent	Substance facilitating dissolution
Product	Resultant substance
Revolution	Stirring rate
Times	Washing frequency
Yield	Output amount or purity
Rate	Temperature increment rate

Annotations were performed using the domain-specific annotation tool DoTAT, depicted in Figure 1. The annotation tasks were categorized into three stages: (1) paragraph classification, (2) action extraction, and (3) reactive element extraction, each executed by specialized annotator groups.

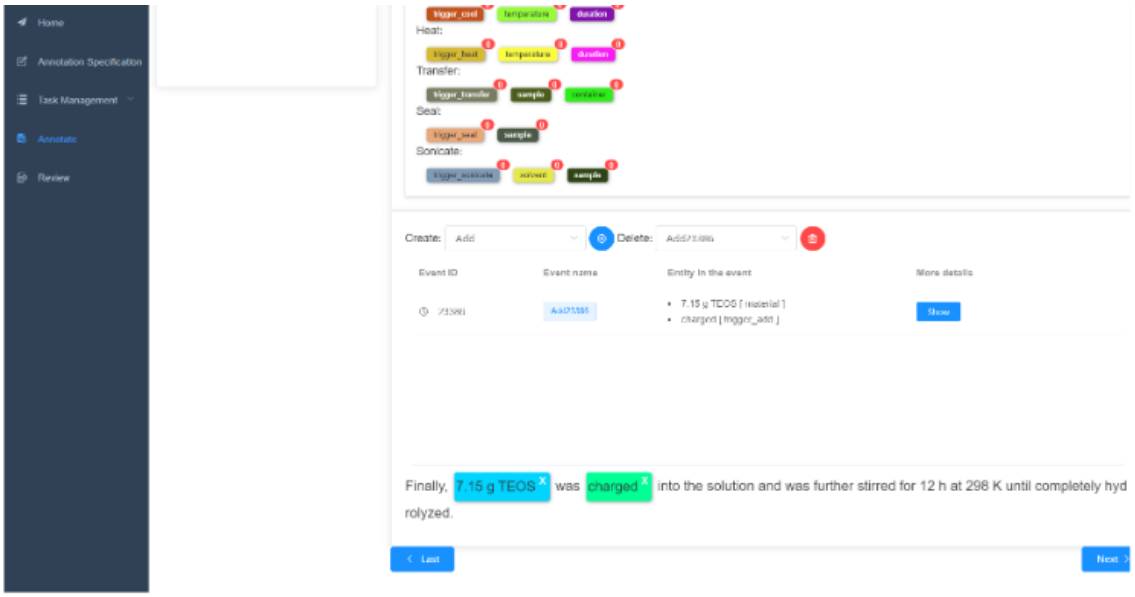

Figure 1: Interface of the domain-specific annotation tool (DoTAT) for the ZeolitePC dataset.

Upon completion, annotations were rigorously reviewed by three independent reviewers. The complete annotation process spanned approximately two months, involving ten annotators. Results were exported in JSON format and visualized in Figure 2.

Examples from annotated datasets (ZeoliteNER and ZeoliteSynPro) are provided in Figures 3 and 4.

Detailed statistics for dataset splits and entity spans are summarized in Tables 3 and 4. These statistics reveal distinct distribution patterns, with frequent entities such as Actions, Materials, and Temperatures typically exhibiting shorter spans compared to less frequent entities.

Table 3: Statistics of the Zeolite-related Datasets

Dataset	**Train**	**Validation**	**Test**
Corpus	174,499	–	19,427
ZeolitePC	598	149	251
ZeoliteNER	4,104	305	400
ZeoliteSynPro	1,997	303	399

The annotated datasets, annotation guidelines, and supplementary materials are publicly available on our GitHub repository: https://github.com/BoldHu/ZeolBERT.

Extraction Performance and System Throughput

The extraction performance was assessed on a computational setup featuring Linux 6.5.11-7pve (Debian 12.4), Python 3.7, CUDA 12.2, PyTorch 1.12.0, and dual NVIDIA Quadro RTX 8000 GPUs (48 GB each). Training employed a batch size of 16, a learning rate of 1×10^{-4}, and a maximum sequence length of 512 tokens.

Figure 5 illustrates structured extraction results in JSON format, enabling seamless downstream processing and analysis.

The system achieved an average throughput of approximately **120 documents per hour** during evaluation.

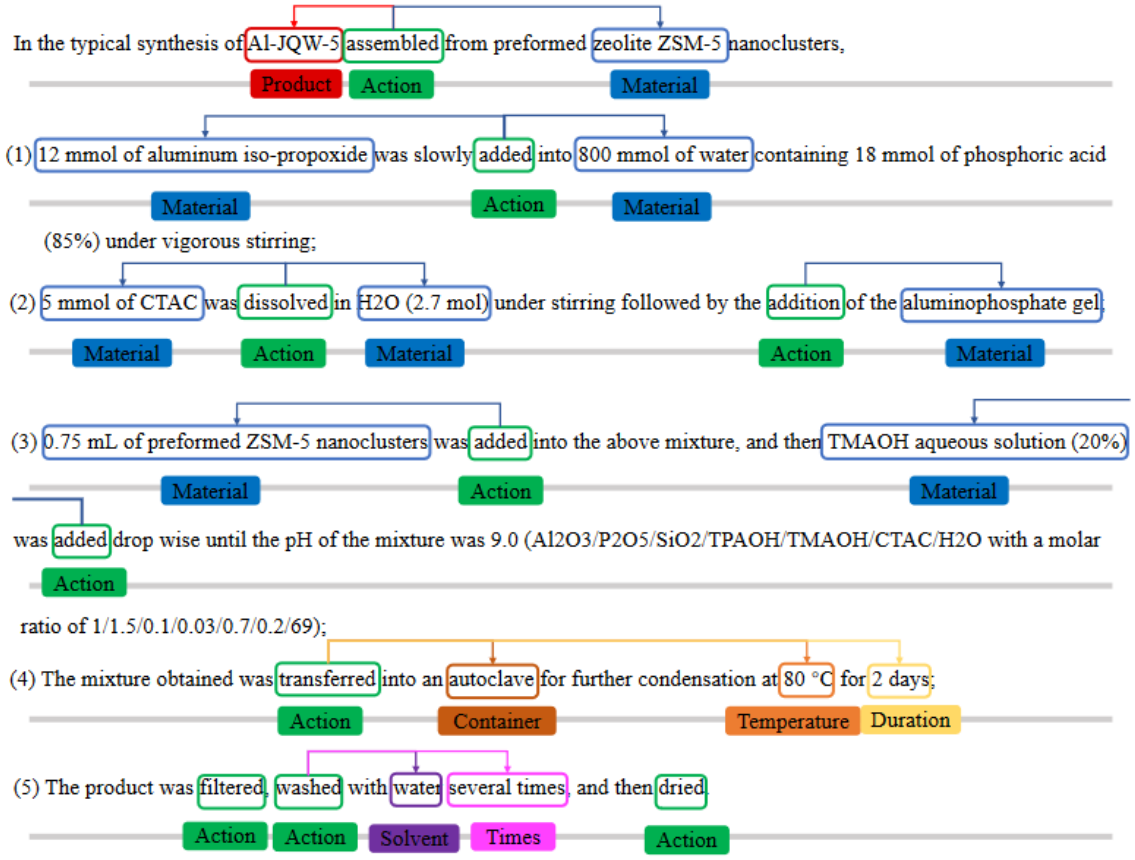

Figure 2: Visualization of annotation results for the ZeolitePC dataset.

Words: "Finally", ",", "a", "heat-treatment", "at", "900", "°C", "for", "4", "h", "under", "nitrogen", "flow", "was", "performed", "."

Labels: "O", "O", "O", "B-action", "O", "B-temperature", "I-temperature", "O", "B-duration", "I-duration", "O", "O", "O", "O", "O", "O"

Figure 3: Annotated example from the ZeoliteNER dataset.

Words: "Then", "the", "zeolite", "was", "washed", "with", "deionized", "water", "five", "times", "and", "dried", "in", "air", "at", "333", "K", "for", "10", "h", ".“

Labels: "O", "O", "B-sample", "O", "B-action", "O", "B-solvent", "I-solvent", "B-times", "I-times", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O", "B-action", "O", "O", "O", "B-temperature", "I-temperature", "O", "B-duration", "I-duration", "O"

Figure 4: Annotated example from the ZeoliteSynPro dataset.

Table 4: Number of entities in different datasets and length of entity spans

Entity Type	**Entity Count**			**Entity Span Length**		
	Train	**Valid**	**Test**	**Train**	**Valid**	**Test**
Action	7227	538	728	1	1	1
Container	688	53	82	3	3	3
Duration	1731	120	190	2	2	2
Material	3182	270	412	5	5	4
Product	213	21	35	2	2	2
Rate	21	3	4	2	2	2
Revolution	41	3	13	2	2	2
Sample	596	42	56	2	2	2
Solvent	555	42	52	2	2	2
Temperature	1883	149	177	2	2	2
Times	78	6	5	2	2	2
Yield	18	2	2	1	1	1

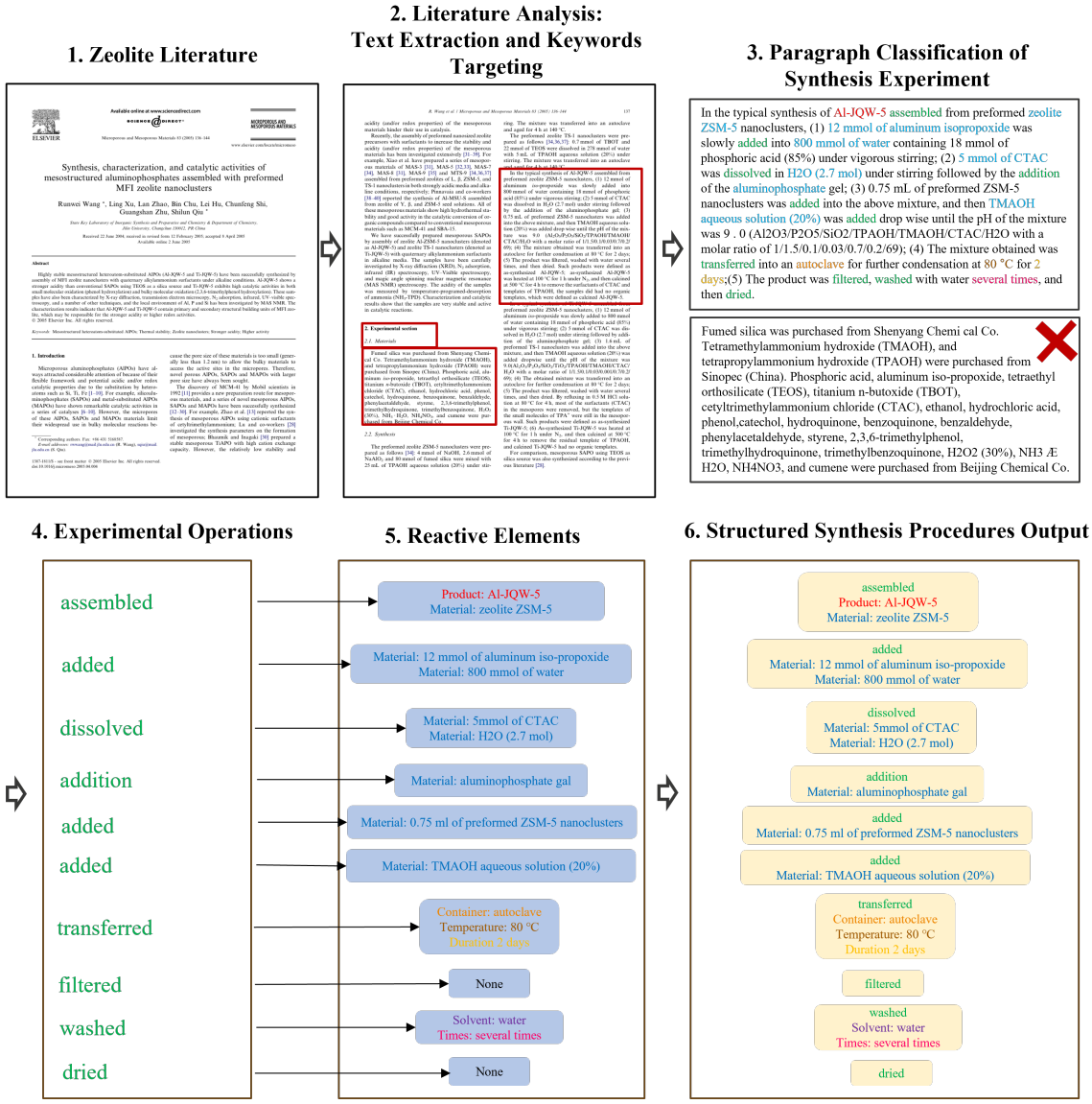

Figure 5: Example extraction results from the ZeolBERT pipeline, highlighting extracted entities and structured outputs.